**Федеральное государственное образовательное бюджетное учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»
Ярославский филиал**

**Жаров А.Н.
Минеичева И.Г.**

АНАЛИЗ ДАННЫХ

Учебное пособие

Ярославль 2020

УДК 33
ББК 22.161
Ж 34

Рецензент:
Д.С. Вахрушев, доктор экономических наук, профессор, профессор кафедры «Финансов и кредита» ФГБОУ ВО «Ярославский государственный университет им. П.Г. Демидова».

Ж34 Жаров А.Н. Анализ данных: учебное пособие [Текст] / Жаров А.Н., Минеичева И.Г. – Ярославль: ООО «ПКФ «СОЮЗ-ПРЕСС», 2020. – 148 с.

ISBN 978-5-6044447-4-0

В учебном пособии изложены основные понятия и теоретические положения курса «Анализ данных». Приведены основные теоретические понятия, определения, теоремы, следствия необходимые для анализа статистической информации. Приводится достаточно большое количество практических примеров, необходимых для освоения курса.

Учебное пособие предназначено для студентов, обучающихся по направлениям подготовки 38.03.01 «Экономика» и 38.03.02 «Менеджмент», а также для преподавателей, проводящих лекционные и практические занятия по дисциплине «Анализ данных».

УДК 33
ББК 22.161

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ...5
Глава 1. ТЕОРИЯ ВЕРОЯТНОСТЕЙ. ОСНОВНЫЕ ПОНЯТИЯ И ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ 7
1.1. Классификация событий 7
1.2. Классическое определение вероятности 8
1.3. Статистическое определение вероятности....................................9
1.4. Элементы комбинаторики...10
1.5. Теорема сложения вероятностей...15
1.6. Теорема умножения вероятностей 16
1.7. Формула полной вероятности. Формула Байеса. 18
1.8. Повторные независимые испытания...19
1.9. Формула Пуассона 21
Глава 2. СЛУЧАЙНЫЕ ВЕЛИЧИИНЫ..23
2.1. Понятие случайной величины 23
2.2. Математическое ожидание случайной величины...........................25
2.3. Дисперсия дискретной случайной величины...................................26
2.4. Функция распределения случайной величины 28
2.5. Непрерывные случайные величины. Функция плотности вероятности ..30
Глава 3. ОСНОВНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ...............................33
3.1. Биномиальный закон распределения 33
3.2. Закон распределения Пуассона 35
3.3. Равномерный закон распределения...37
3.4. Показательный закон распределения...39
3.5. Нормальный закон распределения 41
3.6. Локальная и интегральная теоремы Муавра-Лапласа.......................43
Вопросы для самопроверки по разделам 1-3 45
Глава 4. ОСНОВЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ ВЫБОРОЧНОГО МЕТОДА 47
4.1. Общие сведения о выборочном методе. 47
4.2. Выборка и ее представление. 48
4.3. Числовые характеристики вариационных рядов 52
4.4. Понятие оценки параметров 59
4.5. Метод моментов для нахождения оценок параметров. 61
4.6. Метод наибольшего правдоподобия для нахождения оценок параметров. 63
4.7. Оценка параметров генеральной совокупности по собственно-случайной выборке..66
4.8. Интервальное оценивание параметров генеральной совокупности 68
4.8.1. Доверительный интервал для оценки математического ожидания нормального распределения при известной генеральной дисперсии 69
4.8.2. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии...............................70

4.8.3. Доверительные интервалы для оценки среднего квадратического отклонения σ нормального распределения 73
Глава 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ 75
5.1. Общая схема проверки гипотез 78
5.2. Сравнение двух дисперсий нормальных генеральных совокупностей в случае независимых выборок 78
5.3. Сравнение генеральной дисперсии с гипотетической генеральной дисперсией нормальной совокупности 81
5.4. Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам одинакового объема. Критерий Кохрена 84
5.5. Сравнение двух средних нормальных генеральных совокупностей с известными дисперсиями в случае независимых выборок 85
5.6. Сравнение двух средних произвольно распределенных генеральных совокупностей при больших независимых выборках 88
5.7. Сравнение двух средних нормальных генеральных совокупностей с неизвестными, но одинаковыми дисперсиями в случае независимых выборок 89
5.8. Сравнение выборочной средней с гипотетической генеральной средней нормальной совокупности 93
5.9. Проверка гипотезы о законе распределения генеральной совокупности. Критерий согласия Пирсона. 96
Глава 6. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ 100
6.1. Функциональная, статистическая и корреляционная зависимости 100
6.2. Линейная парная регрессия 101
6.3. Коэффициент корреляции 103
6.4. Проверка гипотезы о значимости коэффициента корреляции 105
6.5. Методика построения корреляционной таблицы 107
6.6. Выборочное корреляционное отношение и индекс корреляции 110
6.7. Ранговая корреляция 118
Глава 7. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ 126
7.1. Общая, факторная и остаточная суммы квадратов отклонений 127
7.2. Общая, факторная и остаточная дисперсии 128
7.3. Сравнение нескольких средних методом дисперсионного анализа 128
Вопросы для самопроверки по разделам 4-7 131
СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТРЫ 133
ПРИЛОЖЕНИЯ 137

ВВЕДЕНИЕ

Данное учебное пособие содержит в себе материал, посвященный основам теории вероятностей и математической статистики.

В первом разделе пособия излагаются основные понятия и теоремы теории вероятностей. Здесь приводится классификация событий, дается классическое и статистическое определения вероятности события, определяются основные понятия комбинаторики. Здесь же формулируются теоремы сложения и умножения вероятностей, формула полной вероятности, формула Байеса, рассматриваются повторные независимые испытания, а также обосновывается формула Пуассона.

Вторая глава пособия посвящена случайным величинам. В данном разделе дается понятие дискретных и непрерывных случайных величин, а также описываются такие важные понятия как математическое ожидание и дисперсия случайной величины. Здесь формулируются такие фундаментальные понятия как функция распределения случайной величины и функция плотности вероятности. А также формулируются основные фундаментальные свойства всех перечисленных величин.

Третий раздел пособия посвящен рассмотрению основных законов распределения. Так в этом разделе рассматриваются такие законы распределения как: биномиальный; Пуассона; равномерный; показательный; нормальный. Для каждого закона проверяется основное свойство – равенство единице суммы вероятностей событий, а также определяются математическое ожидание, дисперсия, функция распределения, обосновываются формулы для вычисления вероятностей событий.

Четвертый раздел пособия посвящен основам математической теории выборочного метода. В данном разделе собраны общие сведения о выборочном методе, рассмотрены виды выборок, вариационных рядов, графическое представление выборок, числовые характеристики вариационных рядов. Весьма значительное внимание уделено точечным и интервальным оценкам параметров генеральной совокупности. В частности, подробно описан метод моментов и наибольшего правдоподобия для нахождения оценок параметров распределений. В завершении раздела приводятся алгоритмы построения доверительных интервалов для генерального среднего и генеральной дисперсии.

В пятом разделе пособия изложен материал, посвященный проверке статистических гипотез. Достаточно большое внимание уделено проверке статистических гипотез о равенстве дисперсий и математических ожиданий в случае независимых выборок, а также вопросам сравнения дисперсий и математических ожиданий с номинальными значениями. Здесь же излагается вопрос о виде закона распределения.

Шестой раздел пособия содержит материал, посвященный корреляционному анализу. Здесь весьма большое место уделено вопросам линейной парной регрессии, рассматриваются вопросы проверки гипотезы о значимости коэффициента корреляции, обсуждается методика построения корреляционной таблицы. Для случая, когда два объекта можно характеризовать только рангами, рассматривается понятие коэффициентов ранговой корреляции Кендала и Спирмена. Рассматриваются гипотезы о проверке значимости данных коэффициентов.

В седьмом разделе учебного пособия рассмотрен однофакторный дисперсионный анализ.

Пособие может быть использовано для широкого круга читателей: специалистов, бакалавров, магистров.

Глава 1. ТЕОРИЯ ВЕРОЯТНОСТЕЙ. ОСНОВНЫЕ ПОНЯТИЯ И ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

1.1. Классификация событий

Под *испытанием* или *опытом* будем понимать осуществление определенного комплекса действий или условий. Каждое испытание заканчивается некоторым результатом. Такие результаты будем называть *исходами* или *событиями*. Таким образом, под *событием* понимается явление, которое наступает в результате данного испытания.

Приведем примеры событий:

1. Появление герба или цифры (решки) при однократном подбрасывании монеты.
2. Появление четной цифры при подбрасывании игрального куба.
3. Выигрыш приза в денежно-вещевой лотерее.

Достоверным (обозначаем символом Ω) называют событие, которое обязательно произойдет в данном испытании.

Невозможным (обозначаем символом $\emptyset$) называют событие, которое не произойдет в данном испытании.

Случайным или возможным называют событие, которое в данном испытании может произойти, либо не произойти.

События обозначаются прописными латинскими буквами $A, B, C, \ldots$

Если при каждом испытании, при котором происходит событие A, происходит и событие B, то говорят, что A *влечет за собой событие* В (входит в B) или B *включает событие* A и обозначают $A \subset B$.

Например, событие A = «появилось число 4», – выпадение числа 4 при однократном подбрасывании куба, B = «появилось 2 или 4 или 6» – выпадение четного числа, то $A \subset B$.

Если одновременно $A \subset B$ и $B \subset A$, то в этом случае события A и B называются *равносильными*. В этом случае записывают $A = B$.

Например, при подбрасывании монеты события A = «выпадет герб», B = «не выпадет цифра» равносильны.

События A и B называются *несовместными (несовместимыми)*, если в результате данного испытания появление одного из них исключает появление другого.

Например, события A – «экзамен сдан на 5» и B – «экзамен сдан на 4» будут несовместными, если студент сдает один экзамен. Если же рассматривать ситуацию, в которой студент сдает два экзамена по разным дисциплинам, то события A – «экзамен сдан на 5» и B – «экзамен сдан на 4» будут совместными.

События называются *равновозможными*, если в результате испытания по условиям симметрии ни одно из этих событий не является объективно более возможным.

Например, при подбрасывании симметричного кубика равновозможными исходами являются исходы, в которых на верхней грани появилась одна из цифр 1, 2, 3, 4, 5, 6.

Несколько событий называются *единственно возможными*, если в результате испытания обязательно должно произойти хотя бы одно из них.

Например, если в семье два ребенка, то единственно возможными вариантами являются следующие: A – «в семье две девочки», B – «в семье два мальчика», C – «в семье мальчик и девочка».

Несколько событий образуют *полную группу*, если они являются единственно возможными и несовместными исходами испытаний. Это означает, что в результате испытания обязательно должно произойти одно и только одно из этих событий.

Например, если студент сдает экзамен по анализу данных, то полная группа событий: A – «экзамен сдан на 5»; B – «экзамен сдан на 4»; C – «экзамен сдан на 3»; D – «экзамен сдан на 2»; E – «студент не явился на экзамен».

Частным случаем событий, образующих полную группу, являются противоположные события. *Противоположными* называются два несовместных события, из которых одно должно обязательно произойти. Событие противоположное событию A обозначают $\bar{A}$.

Например, если событие A – «при однократном подбрасывании куба выпадет число 6», противоположное событие $\bar{A}$ – «при однократном подбрасывании куба выпадет число не равное 6».

1.2. Классическое определение вероятности

Элементарными исходами некоторого испытания называются равновозможные, несовместные, единственно возможные события, образующие полную группу. Множеством элементарных исходов Ω называется множество, содержащее все возможные события данного случайного испытания, из которых в испытании происходит ровно одно событие.

Например, при однократном подбрасывании монеты множество элементарных исходов $\Omega = \{\text{герб}, \text{цифра}\}$.

Событие B называется *благоприятным* событию A, если появление события B влечет за собой появление события A.

Например, при однократном подбрасывании кубика возможны элементарные события: A_1 – «выпадет цифра 1»; A_2 – «выпадет цифра 2»; A_3 – «выпадет цифра 3»; A_4 – «выпадет цифра 4»; A_5 – «выпадет цифра 5»; A_6 – «выпадет цифра 6». Множество элементарных исходов в данном случае имеет вид $\Omega = \{A_1, A_2, A_3, A_4, A_5, A_6\}$. Если рассмотреть событие A – «выпадет нечетная цифра», то события A_1, A_3, A_5 будут благоприятными к событию A. События же A_2, A_4, A_6 будут неблагоприятными для события A.

Вероятностью события A называется отношение числа случаев m – благоприятных событию A к общему числу случаев n:

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

Пример 1.2.1. Если рассмотреть событие A – «появление простого числа при однократном подбрасывании кубика», то множество элементарных исходов, как указывалось выше, будет $\Omega = \{A_1, A_2, A_3, A_4, A_5, A_6\}$. Общее число случаев $n = 6$. Из чисел 1, 2, 3, 4, 5, 6 простыми числами являются 1, 2, 3, 5. Следовательно, события A_1, A_2, A_3, A_5 благоприятны для события A, а их количество $m = 4$. Вероятность события A будет равна:

$$P(A) = \frac{m}{n} = \frac{4}{6} = \frac{2}{3}.$$

Ответ $P(A) = \frac{2}{3}$.

Отметим основные свойства вероятности события.

1. Вероятность невозможного события равна нулю:

$$P(\emptyset) = \frac{0}{n} = 0.$$

2. Вероятность достоверного события равна единице:

$$P(A) = \frac{n}{n} = 1.$$

3. Вероятность любого события заключена между нулем и единицей:

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

1.3. Статистическое определение вероятности

Формула для расчета вероятности, рассмотренная в параграфе 1.2, может быть применена для расчета вероятностей, если все события, образующие множество элементарных исходов, равновозможные.

Например, если грани куба изготовлены из материалов разной плотности, тогда события $A_1, A_2, A_3, A_4, A_5, A_6$ не будут равновозможными и расчет вероятностей по формуле параграфа 1.2 будет невозможен.

В этом случае вероятность событий можно найти только экспериментальным или опытным путем. Для этого нужно изготовить нужный нам куб. Подбросить его достаточно большое число раз n и подсчитать число раз m, в которых наблюдалось событие A. Затем рассчитать *статистическую вероятность* $W(A)$ события A:

$$W(A) = \frac{m}{n}.$$

Заметим, что для расчета вероятности события $P(A)$ не нужно проводить натурный эксперимент. Достаточно «мысленно» представить множество элементарных исходов данного испытания, определить число благоприятных событий, общее число событий и рассчитать вероятность $P(A)$. Статистическая же вероятность $W(A)$ – это величина экспериментальная, полученная как результат некоторого натурного эксперимента.

Статистическое определение вероятности можно применить к событиям, которые обладают следующими свойствами:

1.Рассматриваемые события должны быть исходами только тех испытаний, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий.

2.События должны обладать так называемой статистической устойчивостью, или устойчивостью относительных частот. т.е. в разных сериях испытаний относительная частота события должна изменяться незначительно.

3.Число испытаний, в результате которых появляется событие A, должно быть достаточно велико, поскольку только в этом случае можно считать вероятность события $P(A)$ приближенно равной ее относительной частоте $W(A)$.

1.4. Элементы комбинаторики

Пусть $1 \leq i \leq n$, A_i – элементы конечного множества. Сформулируем правила комбинаторики.

Правило суммы. Если элемент A_1 может быть выбран n_1 способами, элемент A_2 – другими n_2 способами, A_3 – отличными от первых двух n_3 способами и т.д., A_k - n_k способами, отличными от первых $(k-1)$, то выбор одного из элементов или A_1 или A_2, …, или A_k может быть осуществлен $n_1 + n_2 + \cdots + n_k$ способами.

Пример 1.4.1. На книжной полке стоят 5 книг по математике, 6 книг по анализу данных и 3 книги по экономике. Сколько существует способов извлечь с полки одну книгу по математике или анализу данных? Какова вероятность этого события?

Число способов, которыми можно извлечь с полки одну книгу по математике равно $n_1 = 5$, а по анализу данных $n_2 = 6$. Число способов, которыми можно извлечь одну книгу по математике или анализу данных равно $n_1 + n_2 = 5 + 6 = 11 = m$. Общее число способов, с помощью которых можно извлечь одну книгу с полки равно $n = 5 + 6 + 3 = 14$. Вероятность события A – «извлечь с полки одну книгу по математике или анализу данных» будет:

$$P(A) = \frac{m}{n} = \frac{11}{14}.$$

Ответ $11, \frac{11}{14}$.

Правило произведения. Если элемент A_1 может быть выбран n_1 способами, после каждого такого выбора элемент A_2 может быть выбран n_2 способами и т.д., после каждого $(k-1)$ выбора элемент A_k может быть выбран n_k способами, то выбор всех элементов $A_1, A_2, \dots, A_k$ в указанном порядке может быть осуществлен $n_1 n_2 \dots n_k$ способами.

Пример 1.4.2. Сколько существует трехзначных чисел, которые делятся на 5? Какова вероятность того, что любое наудачу взятое трехзначное число делится на 5?

Первая цифра трехзначного числа может быть одной из цифр 1,2,3,4,5,6,7,8,9. Таких цифр $n_1 = 9$. Вторая цифра может быть 0,1,2,3,4,5,6,7,8,9. Значит $n_2 = 10$. Чтобы число делилось на 5 нужно, чтобы последняя цифра была 0 или 5. Значит $n_3 = 2$. Количество трехзначных чисел, которые делятся на 5, будет равно $n_1 n_2 n_3 = 9 \cdot 10 \cdot 2 = 180$. Общее же количество трехзначных чисел будет равно $9 \cdot 10 \cdot 10 = 900$. Следовательно, вероятность того, что любое наудачу взятое трехзначное число делится на 5 будет равна:

$$P = \frac{180}{900} = \frac{1}{5} = 0{,}2.$$

Пусть даны n элементов $x_1, x_2, \dots, x_n$ некоторого множества. Если из этих элементов выбрать m произвольных элементов $(m \le n)$, тогда можно образовать *размещения из n по m* – такие наборы элементов, которые отличаются либо самими элементами, либо их порядком. Если в размещениях каждый элемент встречается только один раз, то они называются размещениями без повторения. Если же элемент может встречаться в размещениях любое количество раз, то такие размещения называются размещениями с повторениями.

Например, если взять три элемента x_1, x_2, x_3 и составить размещения по два из них, то размещения без повторений будут иметь вид:

$$x_1\ x_2;\quad x_2\, x_1;\quad x_1\, x_3;\quad x_3\, x_1;\quad x_2\, x_3;\quad x_3\, x_2.$$

Количество таких размещений будет равно 6. Размещения с повторениями будут:

$$x_1\ x_1;\quad x_1\, x_2;\quad x_2\, x_1;\quad x_1\, x_3;\quad x_3\, x_1;\quad x_2\, x_2;\quad x_2\, x_3;\quad x_3\, x_2;\quad x_3\, x_3.$$

Количество таких размещений равно 9.

Количество размещений из n по m без повторений обозначается A_n^m, количество размещений из n по m с повторениями обозначается $\tilde{A}_n^m$. Так из примера следует, что $A_3^2 = 6 = 3 \cdot 2$, а $\tilde{A}_3^2 = 9 = 3 \cdot 3$.

Используя правило произведения комбинаторики, можно найти общее выражение для чисел размещения из n по m:

$$A_n^m = n \cdot (n-1) \cdots (n-m+1) = \frac{n!}{(n-m)!};$$

$$\tilde{A}_n^m = n \cdot n \cdots n = n^m.$$

Перестановками из m элементов называются такие наборы элементов, которые отличаются порядком следования данных элементов. Например, из трех элементов x_1, x_2, x_3 можно построить следующие перестановки:

$$x_1\, x_2\, x_3;\quad x_1\, x_3\, x_2;\quad x_2\, x_1\, x_3;\quad x_2, x_3, x_1;\quad x_3\, x_1\, x_2;\quad x_3\, x_2\, x_1.$$

Число таких перестановок равно 6. В общем же случае количество перестановок из m элементов равно:

$$P_m = A_m^m = \frac{m!}{(m-m)!} = \frac{m!}{0!} = \frac{m!}{1} = m!.$$

Сочетаниями из n элементов по m называются такие наборы из m элементов, которые отличаются только самими элементами. Причем, если в сочетании каждый элемент встречается только один раз, то такие сочетания называются сочетаниями без повторений. Если же в сочетании элемент может повторяться, то сочетания называются сочетания с повторениями.

Например, если рассмотреть три элемента x_1, x_2, x_3 и составить сочетания по два из них, то сочетания без повторений будут иметь вид:

$$x_1\ x_2;\quad x_1\, x_3;\quad x_2\, x_3.$$

Их количество равно 3. Сочетания с повторениями будут:

$$x_1\ x_1;\quad x_2\ x_2;\quad x_3\, x_3;\quad x_1\ x_2;\quad x_1\, x_3;\quad x_2\, x_3.$$

Количество таких сочетаний равно 6. Число сочетаний из n по m без повторения обозначается C_n^m, а с повторениями $\tilde{C}_n^m$.

В общем случае числа сочетаний из n по m можно вычислить по формулам:

$$C_n^m = \frac{A_n^m}{P_m} = \frac{n!}{m!\,(n-m)!};$$

$$\tilde{C}_n^m = C_{n+m-1}^m = \frac{(n+m-1)!}{m!\ (n-1)}.$$

Если в перестановках из общего числа n элементов есть k различных элементов, при этом 1-й элемент повторяется n_1 раз, 2-й элемент - n_2 раз, -й элемент n_k раз, причем $n_1 + n_2 + \cdots + n_k = n$, то такие перестановки называются *перестановками с повторениями из* n *элементов*. Число перестановок с повторениями из n элементов равно:

$$P_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!\,n_2! \cdots n_k!}.$$

Пример 1.4.3. Сколько существует шестизначных чисел, содержащих только цифры 1,2,3, в которых цифра 1 встречается 3 раза, цифра 2 – 2 раза, цифра 3 – 1 раз.

В данном случае имеем $n_1 = 3,\ n_2 = 2,\ n_3 = 1,\ n_1 + n_2 + n_3 = 6.$ Количество шестизначных чисел будет равно:

$$P_n(3,2,1) = \frac{6!}{3!\,2! \cdots 1!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 1} = \frac{4 \cdot 5 \cdot 3}{1} = 60.$$

Ответ 60.

Пример 1.4.4. На отдельных карточках написаны буквы Ф, И, К, Е, Р. Ребенок в случайном порядке прикладывает карточки одна к другой.

а) Какова вероятность того, что, сложив произвольные три карточки, он получит слово КИР?

б) Какова вероятность того, что, сложив все 5 карточек, он получит слово КЕФИР?

а) Пусть событие A – «получено слово КИР». Складывая 3 карточки из 5 получим размещения без повторения из 5 по 3. Число таких размещений будет:

$$A_5^3 = \frac{5!}{(5-3)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2} = 60.$$

Это число есть общее число испытаний $n = 60$. Количество благоприятных случаев $m = 1$. Вероятность события A будет равна:

$$P(A) = \frac{m}{n} = \frac{1}{60}.$$

б) Пусть событие B – «получено слово КЕФИР». Число благоприятных событий $m = 1$. Общее число событий представляет собой число перестановок из 5 букв $n = P_5 = 5! = 120$. Вероятность события B будет равна:

$$P(B) = \frac{m}{n} = \frac{1}{120}.$$

Ответ а) $\frac{1}{60}$, б) $\frac{1}{120}$.

Пример 1.4.5. На отдельных карточках написаны буквы М, А, Т, Е, М, А, Т, И, К, А. Ребенок в случайном порядке прикладывает карточки одна к другой.

Какова вероятность того, что, сложив все десять карточек, он получит слово МАТЕМАТИКА?

Пусть событие C – «получилось слово МАТЕМАТИКА». В слове математика буква А встречается 3 раза, буквы М и Т – по 2 раза, Е, И, К по одному разу. Число благоприятных случаев:

$$m = P_3 \cdot P_2 \cdot P_2 = 3! \cdot 2! \cdot 2!.$$

Общее число случаев $n = P_{10} = 10!$. Вероятность события C равна:

$$P(C) = \frac{m}{n} = \frac{3! \cdot 2! \cdot 2!}{10!} = \frac{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 1 \cdot 2}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10} = \frac{1}{151200}.$$

Ответ $\frac{1}{151200}$.

Пример 1.4.6. В цехе работают 10 рабочих. У 5-ти из рабочих высшая квалификация. Наудачу отобраны 3 рабочих. Найти вероятность того, что все трое отобранных имеют высшую квалификацию.

Пусть событие A – «трое отобранных имеют высшую квалификацию». В данном случае порядок отобранных рабочих нам не важен, поэтому общее число испытаний найдем:

$$n = C_{10}^3 = \frac{10!}{3!\,(10-3)!} = 120.$$

Число событий, благоприятных для события A равно:

$$m = C_5^3 = \frac{5!}{3!\,(5-3)!} = 10.$$

Вероятность события A равна:

$$P(A) = \frac{m}{n} = \frac{10}{120} = \frac{1}{12}.$$

Ответ $\frac{1}{12}$.

1.5. Теорема сложения вероятностей

Суммой $A+B$ *двух событий* A *и* B называют событие, состоящее в появлении события A, или события B, или обоих этих событий. Например, если студент сдавал два экзамена и событие A, состоит в том, что он сдал первый экзамен, а событие B – сдал второй экзамен, то сумма $A+B$ – это событие, состоящее в том, что он сдал хотя бы один из экзаменов. Т.е. либо сдал только первый, либо только второй, либо сдал оба экзамена. Если события A *и* B несовместные, то сумма $A+B$ – это событие, состоящее в появлении одного из этих событий A или B.

Произведением двух событий $A\cdot B$ называют событие, состоящее в совместном появлении этих событий. Например, если событие A – мяч волейбольный, событие B – предмет белый, то $A\cdot B$ – белый волейбольный мяч.

Теорема. Вероятность суммы двух событий $A+B$ равна сумме вероятностей этих событий без вероятности их произведения, т.е.:

$$P(A+B)=P(A)+P(B)-P(A\cdot B).$$

Если события A *и* B несовместные, то событие $A\cdot B$ является невозможным, а вероятность $P(A\cdot B)=0$. Значит, справедлива теорема

Теорема. (Сложения вероятностей для несовместных событий)

Вероятность суммы двух несовместных событий $A+B$ равна сумме вероятностей этих событий:

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

Пример 1.5.1. В урне 15 красных, 10 синих и 25 белых шаров. Из урны извлекается один шар. Найти вероятность того, что этот шар будет цветным.

Решение. Обозначим, A – шар красный, B – шар синий. Тогда $A+B$ – шар цветной. Вычислим вероятности:

$$P(A)=\frac{15}{50}=0{,}3,\ P(B)=\frac{10}{50}=0{,}2,$$

$$P(A+B)=P(A)+P(B)=0{,}3+0{,}2=0{,}5.$$

Ответ 0,5

Примем, что событие $B=\overline{A}$ представляет собой событие, противоположное для события A. В данном случае событие $A+B=A+\overline{A}$ является достоверным, а значит вероятность $P(A+B)=P\left(A+\overline{A}\right)=1$. Теорема сложения вероятностей в этом случае примет вид:

$$P\left(A+\overline{A}\right)=P(A)+P\left(\overline{A}\right) \quad \rightarrow \quad 1=P(A)+P\left(\overline{A}\right)$$

Выражая из последнего равенства вероятность $P\left(\overline{A}\right)$ получим:

Следствие. Вероятность события $P(\overline{A})$ противоположного для события A равна разности единицы и вероятности события A:

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

Например, если с вероятностью $P(A) = 0{,}7$ завтрашний день будет солнечным, то с вероятностью $P(\overline{A}) = 1 - P(A) = 1 - 0{,}7 = 0{,}3$ он будет пасмурным.

Предположим, что события $A_1, A_2, \dots, A_n$ образуют полную группу событий. События, образующие полную группу событий, несовместны, значит:

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2 + \dots + A_n) =$$
$$= P(A_1) + P(A_2) + P(A_3 + \dots + A_n) = \dots = P(A_1) + P(A_2) + \dots + P(A_n)$$

Событие $A_1 + A_2 + \dots + A_n$ является достоверным, а значит:

$$P(A_1 + A_2 + \dots + A_n) = 1.$$

Эти рассуждения позволяют сформулировать ещё одно следствие из теоремы сложения вероятностей.

Следствие. Сумма вероятностей событий $A_1, A_2, \dots, A_n$, образуют полную группу событий, равна единице:

$$P(A_1) + P(A_2) + \dots + P(A_n) = 1.$$

1.6. Теорема умножения вероятностей

Условной вероятностью $P_B(A)$ называют вероятность события A, вычисленную при условии, что событие B уже наступило.

Сформулируем теорему умножения вероятностей.

Теорема. Вероятность совместного появления двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило:

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

Данную теорему можно обобщить на любое количество сомножителей. Так, например, для произведения трех событий:

$$P(A \cdot B \cdot \mathrm{C}) = P(A) \cdot P_A(B \cdot \mathrm{C}) = P(A) \cdot P_A(B) \cdot P_{AB}(C).$$

События A и B называются независимыми, если:

$$\begin{cases} P_B(A) = P(A); \\ P_A(B) = P(B). \end{cases}$$

Теорема умножения для независимых событий формулируется так:

Вероятность совместного появления двух независимых событий равна произведению вероятностей этих событий:

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

Пример 1.6.1. К зачету студенту нужно выучить 60 вопросов. Он успел выучить только 40 вопросов. Зачет состоит в том, что преподаватель задает студенту три вопроса. Если студент отвечает на эти три вопроса, то студенту выставляется зачет. Какова вероятность того, что студент сдаст зачет.

Решение

Обозначим события A – студент ответит на первый вопрос, B – студент ответит на второй вопрос, C – студент ответит на третий вопрос. Для сдачи зачета нужно, чтобы студент ответил на три вопроса:

$$P(A \cdot B \cdot \mathrm{C}) = P(A) \cdot P_A(B) \cdot P_{AB}(C) = \frac{40}{60} \cdot \frac{39}{59} \cdot \frac{38}{58} = 0{,}289$$

Ответ 0,289

Пример 1.6.2. Банк выдает два крупных кредита двум крупным градообразующим предприятиям. Вероятность того, что первое предприятие вовремя погасит кредит равна $p_1 = 0{,}8$, для второго предприятия эта вероятность равна $p_1 = 0{,}9$. Найти вероятность того, что а) ни одно предприятие не погасит кредит вовремя; б) только одно предприятие погасит кредит вовремя; в) оба предприятия погасят кредит вовремя?

Решение

Введем события A – первое событие вовремя погасит кредит, B – второе предприятие вовремя погасит кредит.

а) $\overline{A} \cdot \overline{B}$- событие, состоящее в том, что оба предприятия не погасят кредит вовремя:

$$P\left(\overline{A} \cdot \overline{B}\right) = P\left(\overline{A}\right) \cdot P\left(\overline{B}\right) = \left(1 - P(A)\right) \cdot \left(1 - P(B)\right) =$$
$$= (1 - 0{,}8) \cdot (1 - 0{,}9) = 0{,}02.$$

б) $A \cdot \overline{B} + \overline{A} \cdot B$- событие, состоящее в том, что только одно предприятие погасят кредит вовремя:

$$P\left(A \cdot \overline{B} + \overline{A} \cdot B\right) = P\left(A \cdot \overline{B}\right) + P\left(\overline{A} \cdot B\right) = P(A) \cdot P\left(\overline{B}\right) + P\left(\overline{A}\right) \cdot P(B) =$$
$$= P(A) \cdot \left(1 - P(B)\right) + \left(1 - P(A)\right) \cdot P(B) = 0{,}8 \cdot 0{,}1 + 0{,}2 \cdot 0{,}9 = 0{,}26.$$

в) $A \cdot B$- событие, состоящее в том, что оба предприятия погасят кредит вовремя:

$$P(A \cdot B) = P(A) \cdot P(B) = 0{,}8 \cdot 0{,}9 = 0{,}72.$$

Ответ а) 0,02; б) 0,26; в) 0,72.

1.7. Формула полной вероятности. Формула Байеса

Предположим, что некоторое случайное событие A может произойти только совместно с одним из событий $H_1, H_2, \dots, H_n$ образующих полную группу. В этом случае событие A можно представить в виде:

$$A = A \cdot H_1 + A \cdot H_2 + \dots + A \cdot H_n$$

Применяя к этому событию теоремы сложения и умножения вероятностей найдем:

$$P(A) = P(A \cdot H_1 + A \cdot H_2 + \dots + A \cdot H_n) =$$
$$= P(A \cdot H_1) + P(A \cdot H_2) + \dots + P(A \cdot H_n) =$$
$$= P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + \dots + P(H_n)P_{H_n}(A)$$

События $H_1, H_2, \dots, H_n$ называют гипотезами. Формулу

$$P(A) = P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + \dots + P(H_n)P_{H_n}(A)$$

называют *формулой полной вероятности*.

Пример 1.7.1. В пирамиде пять винтовок, три из которых снабжены оптическим прицелом. Вероятность того, что стрелок поразит мишень при выстреле из винтовки с оптическим прицелом, равна 0,95; для винтовки без оптического прицела эта вероятность равна 0,7. Найти вероятность того, что мишень будет поражена, если стрелок произведет один выстрел из наудачу взятой винтовки.

Решение

Пусть A – событие, состоящее в том, что мишень поражена. Введем две гипотезы: H_1 – выстрел произведен из винтовки без оптического прицела; H_2 – выстрел произведен из винтовки с оптическим прицелом. По формуле полной вероятности:

$$P(A) = P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) = \frac{2}{5} \cdot 0{,}7 + \frac{3}{5} \cdot 0{,}95 = 0{,}85$$

Ответ 0,85

Рассмотрим формулу Байеса. Для этого найдем вероятность одновременного осуществления события A и какой - либо гипотезы H_i:

$$P(A \cdot H_i) = P(A) \cdot P_A(H_i) = P(H_i) \cdot P_{H_i}(A), \quad i = 1, \dots, n.$$

Из последнего равенства выразим вероятность осуществления гипотезы H_i при условии выполнения условия A:

$$P_A(H_i) = \frac{P(H_i) \cdot P_{H_i}(A)}{P(A)}, \quad i = 1, \dots, n.$$

В знаменателе этой формулы вероятность $P(A)$ заменим по формуле полной вероятности:

$$P_A(H_i) = \frac{P(H_i) \cdot P_{H_i}(A)}{P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + \cdots + P(H_n)P_{H_n}(A)}, \quad i = 1, \dots, n.$$

Данная формула называется *формулой Байеса*. Она позволяет найти вероятности осуществления гипотез $H_1, H_2, \dots, H_n$ при условии выполнения события A.

Отметим так же, что вероятности входящие в формулу Байеса получили особое название. Так вероятности $P(H_1), P(H_2), \dots, P(H_n)$ называются априорными, а вероятности $P_A(H_1), P_A(H_2), \dots, P_A(H_n)$ называются апостериорными вероятностями. Таким образом, формула Байеса позволяет вычислить апостериорные вероятности.

Пример 1.7.2. В специализированную больницу поступают в среднем 50% больных с заболеванием K, 30% - с заболеванием L, 20 % - с заболеванием M. Вероятность полного излечения болезни K равна 0,7; для болезней L и M эти вероятности соответственно равны 0,8 и 0,9. Больной, поступивший в больницу, был выписан здоровым. Найти вероятность того, что этот больной страдал заболеванием K?

Решение

Введем событие A – больной выписан из больницы здоровым. В данном случае имеем три гипотезы H_1 – больной поступил в больницу с заболеванием K, H_2 – больной поступил в больницу с заболеванием L, H_3 – больной поступил в больницу с заболеванием M.

Найдем вероятность того, что больной выписан из больницы здоровым:

$$P(A) = P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + P(H_3)P_{H_3}(A) =$$
$$= 0{,}5 \cdot 0{,}7 + 0{,}3 \cdot 0{,}8 + 0{,}2 \cdot 0{,}9 = 0{,}77.$$

По формуле Байеса найдем вероятность того, что больной страдал заболеванием K:

$$P_A(H_1) = \frac{P(H_1) \cdot P_{H_1}(A)}{P(A)} = \frac{0{,}5 \cdot 0{,}7}{0{,}77} = \frac{35}{77} = 0{,}45$$

Ответ 0,45.

1.8. Повторные независимые испытания

Рассмотрим n независимых повторных испытаний, которые выполняются при постоянном комплексе условий. При этом будем считать, что результат одного такого испытания – это либо успех, либо неудача. Число успехов среди n испытаний обозначим m. Например, если банк выдал $n = 100$ кредитов за один месяц на 5 лет. По истечении 5 лет только 86 кредитов полностью погашены. В этом случае можно говорить, что имеется 100 повторных независимых испытаний, а число успехов равно 86. Если, например, стрелок стреляет 20 раз в мишень и имеется ровно 18 попаданий, то можно говорить, что имеется 20 независимых испытаний, а число успехов равно 18.

Введем событие A_i – успех в i – м испытании и событие $B_{m,n}$ состоящее в том, что среди n повторных независимых испытаний было ровно m успехов. Тогда:

$$B_{m,n} = A_1 A_2 \cdots A_m \overline{A}_{m+1} \cdots \overline{A}_n + \overline{A}_1 A_2 \cdots A_m A_{m+1} \overline{A}_{m+2} \cdots \overline{A}_n + \\ + \overline{A}_1 \overline{A}_2 \cdots \overline{A}_{n-m} A_{n-m+1} \cdots A_n$$

Найдем вероятность события $B_{m,n}$:

$$P(B_{m,n}) = P_n(m) = P\big(A_1 A_2 \cdots A_m \overline{A}_{m+1} \cdots \overline{A}_n + \\ + \overline{A}_1 A_2 \cdots A_m A_{m+1} \overline{A}_{m+2} \cdots \overline{A}_n + \overline{A}_1 \overline{A}_2 \cdots \overline{A}_{n-m} A_{n-m+1} \cdots A_n\big)$$

Применим теоремы сложения и умножения вероятностей, обозначим вероятность успеха $P(A_i) = p$, а вероятность неудачи $P(\overline{A}_i) = q = 1 - p$.

$$P(B_{m,n}) = P_n(m) = \underbrace{p^m q^{n-m} + p^m q^{n-m} + \cdots + p^m q^{n-m}}_{C_n^m \text{ раз}} = \\ = C_n^m p^m q^{n-m}.$$

Последняя формула называется *формулой Бернулли*.

Таким образом, справедлива теорема.

Если вероятность p наступления события A в каждом испытании постоянна, то вероятность $P_n(m)$ того, что событие A наступит m раз в n независимых испытаниях, равна:

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

где $q = 1 - p$ – вероятность неудачи в одном испытании,

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

Пример 1.8.1 Стрелок делает 5 выстрелов по мишени. Вероятность поражения мишени при одном выстреле равна 0,8. Найти вероятность того, что число попаданий а) равно 3; б) больше 3; в) не больше 2.

Решение. В данном случае число испытаний $n = 5$, вероятность успеха в одном испытании $p = 0{,}8$, вероятность неудачи в одном испытании $q = 1 - p = 1 - 0{,}8 = 0{,}2$. Формула Бернулли примет вид:

$$P_5(m) = C_5^m 0{,}8^m 0{,}2^{5-m} = \frac{5!}{m!(5-m)!} \cdot 0{,}8^m \cdot 0{,}2^{5-m}.$$

а) $P_5(m = 3) = \frac{5!}{3!(5-3)!} \cdot 0{,}8^3 \cdot 0{,}2^2 = \frac{1\cdot2\cdot3\cdot4\cdot5}{1\cdot2\cdot3\cdot1\cdot2} \cdot 0{,}512 \cdot 0{,}04 = 0{,}2048.$

б) $P_5(m > 3) = P_5(m = 4) + P_5(m = 5) = \frac{5!}{4!(5-4)!} \cdot 0{,}8^4 \cdot 0{,}2^1 +$

$$+\frac{5!}{5!(5-5)!}\cdot 0{,}8^5\cdot 0{,}2^0 = 5\cdot 0{,}4096\cdot 0{,}2 + 0{,}32768 = 0{,}73728.$$

в) $P_5(m \le 2) = P_5(m=0) + P_5(m=1) + P_5(m=2) =$

$$= \frac{5!}{0!\,(5-0)!}\cdot 0{,}8^0\cdot 0{,}2^5 + + \frac{5!}{1!\,(5-1)!}\cdot 0{,}8^1\cdot 0{,}2^4 +$$

$$+\frac{5!}{2!\,(5-2)!}\cdot 0{,}8^2\cdot 0{,}2^3 = 0{,}00032 + 0{,}0064 + 0{,}0512 = 0{,}05792.$$

Ответ а)0,2048; б) 0,73728; в) 0,05792.

1.9. Формула Пуассона

Рассмотрим формулу Бернулли при условии, что число испытаний велико $n \to \infty$; вероятность успеха в одном испытании мала $p \to 0$, а величина $\lambda = n\cdot p$ остается постоянной:

$$P_n(m) = C_n^m p^m q^{n-m} = \frac{n!}{m!\,(n-m)!}\cdot p^m(1-p)^{n-m} =$$

$$= \frac{n(n-1)\cdots(n-m+1)}{m!}\cdot p^m(1-p)^n(1-p)^{-m}.$$

Из соотношения $\lambda = n\cdot p$ выразим $p = \lambda/n$.

$$P_n(m) = \frac{n(n-1)\cdots(n-m+1)}{m!}\cdot\left(\frac{\lambda}{n}\right)^m\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m} =$$

$$= \frac{n(n-1)\cdots(n-m+1)}{n^m}\cdot\frac{\lambda^m}{m!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m} =$$

$$= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)\cdot\frac{\lambda^m}{m!}\left(\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right)^{-\lambda}\left(1-\frac{\lambda}{n}\right)^{-m}.$$

При $n \to \infty$ будем иметь

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right) = 1;\ \lim_{n\to\infty}\left(\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right)^{-\lambda} = e^{-\lambda};$$

где $e = 2{,}73$ – основание натурального логарифма;

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-m} = 1.$$

С учетом последних предельных соотношений получим:

$$\lim_{n\to\infty} P_n(m) = P_m(\lambda) = \frac{\lambda^m}{m!}\cdot e^{-\lambda}.$$

Последняя формула называется *формулой Пуассона*.

Сформулируем теорему. Если вероятность p наступления события A в каждом испытании стремится к нулю $(p \to 0)$ при неограниченном увеличении числа n испытаний $(n \to \infty)$, причем произведение $n \cdot p$ стремиться к постоянному числу λ $(n \cdot p \to \lambda)$, то вероятность $P_n(m)$ того, что событие A появится m раз в n независимых испытаниях, удовлетворяет предельному равенству:

$$\lim_{n\to\infty} P_n(m) = P_m(\lambda) = \frac{\lambda^m}{m!} \cdot e^{-\lambda}.$$

Отметим, что более строгие оценки указывают на то, что формула Пуассона применима при условии $\lambda = n \cdot p \leq 10$ и при достаточно большом числе испытаний n.

Пример 1.9.1. Учебник издан тиражом 5000 экземпляров. Вероятность того, что учебник сброшюрован неправильно равна 0,001. Найти вероятность того, что тираж содержит бракованных книг: а) равно 2; б) менее 2; в) более 3.

Решение В нашем случае $n = 5000$, $p = 0{,}001$, $\lambda = n \cdot p = 5000 \cdot 0{,}001 = 5$.

Так как $\lambda = 5 < 10$, можем применить формулу Пуассона:

$$P_n(m) \approx P_m(\lambda) = \frac{\lambda^m}{m!} \cdot e^{-\lambda} = \frac{5^m}{m!} \cdot e^{-5}.$$

а) $P_{5000}(m = 2) = \frac{5^2}{2!} \cdot e^{-5} = 0{,}0842$;

б) $P_{5000}(m < 2) = P_{5000}(m = 0) + P_{5000}(m = 1) =$

$$= \frac{5^0}{0!} \cdot e^{-5} + \frac{5^1}{1!} \cdot e^{-5} = 0{,}0067 + 0{,}0337 = 0{,}0404;$$

в) $P_{5000}(m > 3) = 1 - P_{5000}(m \leq 3) =$

$$= 1 - \left(P_{5000}(m = 0) + P_{5000}(m = 1) + P_{5000}(m = 2) + P_{5000}(m = 3)\right) =$$

$$= 1 - \left(\frac{5^0}{0!} \cdot e^{-5} + \frac{5^1}{1!} \cdot e^{-5} + \frac{5^2}{2!} \cdot e^{-5} + \right.$$

$$\left. + \frac{5^3}{3!} \cdot e^{-5}\right) = 1 - (0{,}0067 + 0{,}0337 + 0{,}0842 + 0{,}1404) = 0{,}735.$$

Ответ а) 0,0842; б) 0,0404; в) 0,735.

Глава 2. СЛУЧАЙНЫЕ ВЕЛИЧИИНЫ

2.1. Понятие случайной величины

Под случайной величиной понимается переменная, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений. Примерами случайных величин являются:

1) число клиентов, обратившихся в банк за кредитом в данный день;

2) число выстрелов до первого попадания в мишень;

3) расход электроэнергии в квартире за месяц;

4) число кредитов, погашенных в данный день.

Случайная величина называется дискретной, если множество её значений конечное, или бесконечное, но счетное.

Под непрерывной случайной величиной будем понимать величину, бесконечное несчетное множество значений которой есть некоторый интервал (конечный или бесконечный) числовой оси.

Определение: Случайной величиной X называется функция, заданная на множестве элементарных исходов (или в пространстве элементарных событий), т.е. $X = f(\omega)$, где ω – элементарный исход (или элементарное событие, принадлежащее пространству элементарных исходов Ω).

Случайные величины будем обозначать прописными буквами $X, Y, Z, \ldots$, а их значения соответствующими строчными буквами $x, y, z, \ldots$

Определение: Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины X и соответствующими им вероятностями $P(X)$.

Для дискретной случайной величины закон распределения может быть задан в виде таблицы, аналитически (в виде формулы) и графически.

Таблица закона распределения имеет вид.

X	x_1	x_2	…	x_i	…	x_n
$P(X)$	p_1	p_2	…	p_i	…	p_n

Такая таблица называется рядом распределения дискретной случайной величины. События $X = x_1$, $X = x_2, \ldots$ $X = x_i$, $\ldots X = x_n$ являются несовместными и единственно возможными, т.е. образуют полную группу. Следовательно, вероятности p_1, p_2, …, p_i,…, p_n удовлетворяют основному свойству закона распределения:

$$p_1 + p_2 + \ldots p_i \ldots + p_n = 1.$$

Ряд распределения может быть изображен графически. Для этого нужно по оси абсцисс откладывать значения случайной величины x_i, а по оси ординат вероятности p_i. Соединяя точки прямыми, получим многоугольник распределения, представленный на рисунке 1.

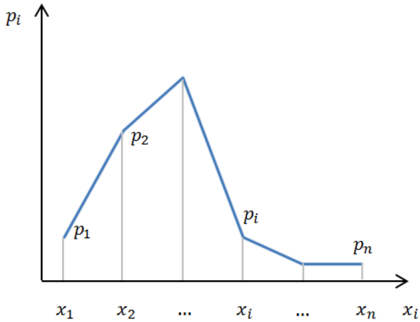

Рисунок 1. Многоугольник распределения

Пример 2.1.1. В партии из 8 деталей 5 стандартных. Наудачу отобраны 2 детали. Составить закон распределения числа стандартных деталей среди отобранных.

Решение. Число стандартных деталей принимает следующие значения $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. Вычислим вероятности этих событий.

$$P(x_1 = 0) = p_1 = \frac{C_5^0 \cdot C_3^2}{C_8^2} = \frac{1 \cdot 3}{28} = \frac{3}{28};$$

$$C_8^2 = \frac{8!}{2!\,(8-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = \frac{7 \cdot 8}{1 \cdot 2} = 28;$$

$$C_5^0 = \frac{5!}{0!\ (5-0)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 1; \quad C_3^2 = \frac{3!}{2!\,(3-2)!} = \frac{1 \cdot 2 \cdot 3}{1 \cdot 2 \cdot 1} = 3;$$

$$P(x_2 = 1) = p_2 = \frac{C_5^1 \cdot C_3^1}{C_8^2} = \frac{5 \cdot 3}{28} = \frac{15}{28};$$

$$C_5^1 = \frac{5!}{1!\ (5-1)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 1 \cdot 2 \cdot 3 \cdot 4} = 5; \quad C_3^1 = \frac{3!}{1!\,(3-1)!} = \frac{1 \cdot 2 \cdot 3}{1 \cdot 1 \cdot 2} = 3;$$

$$P(x_3 = 2) = p_3 = \frac{C_5^2 \cdot C_3^0}{C_8^2} = \frac{10 \cdot 1}{28} = \frac{10}{28};$$

$$C_5^2 = \frac{5!}{2!\ (5-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3} = 10; \quad C_3^0 = \frac{3!}{0!\,(3-0)!} = \frac{1 \cdot 2 \cdot 3}{1 \cdot 2 \cdot 3} = 1.$$

Выполним контроль:

$$p_1 + p_2 + p_3 = \frac{3}{28} + \frac{15}{28} + \frac{10}{28} = \frac{28}{28} = 1.$$

Закон распределения запишем в виде таблицы:

X	0	1	2
$P(X)$	$\frac{3}{28}$	$\frac{15}{28}$	$\frac{10}{28}$

2.2. Математическое ожидание случайной величины

Определение. *Математическим ожиданием или средним значением* $M(X)$ дискретной случайной величины X называется сумма произведений всех ее значений на соответствующие им вероятности:

$$M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n.$$

Свойства математического ожидания.

1.Математическое ожидание постоянной величины равно самой этой постоянной:

$$M(C) = C, \ \ C - \text{постоянная.}$$

2.Постоянный множитель можно выносить за знак математического ожидания:

$$M(kX) = k \cdot M(X), \quad k - \text{постоянная.}$$

3.Математическое ожидание алгебраической суммы конечного числа случайных величин равно такой же сумме их математических ожиданий:

$$M(X \pm Y) = M(X) + M(Y).$$

4.Математическое ожидание произведения конечного числа независимых случайных величин равно произведению их математических ожиданий:

$$M(X \cdot Y) = M(X) \cdot M(Y).$$

5.Если все значения случайной величины увеличить (уменьшить) на постоянную C, то на эту же постоянную C увеличится (уменьшится) математическое ожидание этой случайной величины:

$$M(X \pm C) = M(X) \pm C.$$

6.Математическое ожидание отклонения случайной величины от ее математического ожидания равно нулю:

$$M(X - M(X)) = 0.$$

Пример 2.2.1. Найти математическое ожидание случайной величины $Z = 6X - 7Y + 12$, если $M(X) = 4$, $M(Y) = 5$.

Решение. $M(Z) = M(6X - 7Y + 12) = 6M(X) - 7M(Y) + M(12) =$

$$= 6M(X) - 7M(Y) + M(12) = 6 \cdot 4 - 7 \cdot 5 + 12 = 1.$$

Ответ $M(Z) = 1$.

Пример 2.2.2. Задан ряд распределения дискретной случайной величины:

X	1	3	5	7
$P(X)$	0,1	0,2	0,5	0,2

Найти математическое ожидание.

Решение.

$$M(X) = \bar{x} = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 =$$
$$= 1 \cdot 0{,}1 + 3 \cdot 0{,}2 + 5 \cdot 0{,}5 + 7 \cdot 0{,}2 = 4{,}6.$$

Ответ $M(X) = 4{,}6$.

2.3. Дисперсия дискретной случайной величины

Определение. *Дисперсией* $D(X)$ случайной величины X называется математическое ожидание квадрата ее отклонения от математического ожидания:

$$D(X) = M[X - M(X)]^2$$

или

$$D(X) = M(X - a)^2, \quad \text{где} \quad a = M(X).$$

или

$$D(X) = (x_1 - a)^2 \cdot p_1 + (x_2 - a)^2 \cdot p_2 + \cdots + (x_n - a)^2 \cdot p_n.$$

Из определения дисперсии видно, что если величина X измеряется в каких-то единицах измерения, то дисперсия измеряется в этих же единицах измерения, возведенных в квадрат. Например, если величина X – величина вклада в рублях, то дисперсия $D(X)$ имеет размерность рубли в квадрате. Для того, чтобы образовать величину, по размерности совпадающую со случайной величиной X, вводят понятие среднего квадратического отклонения.

Определение. Средним квадратическим отклонением (стандартным отклонением или стандартом) $\sigma(X)$ случайной величины X называется арифметическое значение корня квадратного из ее дисперсии:

$$\sigma(X) = \sqrt{D(X)}.$$

Дисперсия и среднее квадратическое отклонение характеризует способность случайной величины отклоняться от своего среднего значения. Так если есть две случайные величины X_1 и X_2 и $D(X_2) > D(X_1)$, то в этом случае величина X_2 более сильно отклоняется от своего среднего значения $M(X_2)$, чем величина X_1 отклоняется от своего среднего значения $M(X_1)$.

Отметим основные свойства дисперсии

1.Дисперсия постоянной величины равна нулю:

$$D(C) = M(C - C)^2 = M(0)^2 = M(0) = 0, \text{где } C - \text{постоянная.}$$

2.Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат:

$$D(kX) = k^2 D(X),\ \ k - \text{постоянная.}$$

3.Дисперсия случайной величины равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания:

$$D(X) = M(X^2) - (M(X))^2.$$

4.Дисперсия алгебраической суммы конечного числа независимых случайных величин равна сумме их дисперсий:

$$D(X \pm Y) = D(X) + D(Y).$$

Обратим внимание, что дисперсия как суммы, так и разности независимых случайных величин X и Y равна сумме их дисперсий, т.е.:

$$D(X+Y) = D(X-Y) = D(X) + D(Y).$$

Пример 2.3.1. Найти дисперсию случайной величины $Z = 5X - 3Y + 12$, если $D(X) = 4, D(Y) = 6$.

Решение. $D(Z) = D(5X - 3Y + 12) = D(5X - 3Y + 12) =$

$$= D(5X) + D(3Y) + D(12) = 5^2 D(X) + 3^2 D(Y) + 0 =$$

$$25 \cdot 4 + 9 \cdot 6 = 154.$$

Ответ $D(Z) = 154$.

Пример 2.3.2. Задан закон распределения случайной величины X

X	2	4	6	8
$P(X)$	0,1	0,5	0,2	p_4

Найти значение p_4, а также дисперсию $D(X)$ и среднее квадратичное отклонение $\sigma(X)$.

Решение. Основное свойство закона распределения:

$$p_1 + p_2 + p_3 + p_4 = 1$$

или

$$0{,}1 + 0{,}5 + 0{,}2 + p_4 = 1$$

позволяет найти $p_4 = 1 - 0{,}8 = 0{,}2$.

Далее найдем математическое ожидание $M(X)$:

$$M(X) = a = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 =$$

$$= 2 \cdot 0{,}1 + 4 \cdot 0{,}5 + 6 \cdot 0{,}2 + 8 \cdot 0{,}2 = 5.$$

По определению дисперсии:

$$D(X) = (x_1 - a)^2 \cdot p_1 + (x_2 - a)^2 \cdot p_2 + (x_3 - a)^2 \cdot p_3 + (x_4 - a)^2 \cdot p_4 =$$

$$= (2-5)^2 \cdot 0{,}1 + (4-5)^2 \cdot 0{,}5 + (6-5)^2 \cdot 0{,}2 + (8-5)^2 \cdot 0{,}2 = 3{,}4$$

Среднее квадратическое отклонение $\sigma(X) = \sqrt{3{,}4} = 1{,}84$.

Отметим, что очень удобно находить дисперсию по свойству 3:

$$M(X^2) = a = {x_1}^2 \cdot p_1 + {x_2}^2 \cdot p_2 + {x_3}^2 \cdot p_3 + {x_4}^2 \cdot p_4 =$$

$$= 2^2 \cdot 0{,}1 + 4^2 \cdot 0{,}5 + 6^2 \cdot 0{,}2 + 8^2 \cdot 0{,}2 = 28{,}4$$

Дисперсия тогда будет:

$$D(X) = M(X^2) - (M(X))^2 = 28{,}4 - 5^2 = 3{,}4.$$

Ответ $p_4 = 0{,}2$; $D(X) = 3{,}4$; $\sigma(X) = 1{,}84$.

2.4. Функция распределения случайной величины

Определение. *Интегральной функцией распределения* случайной величины X (или просто функцией распределения) называется функция $F(x)$, выражающая для каждого x вероятность того, что случайная величина X примет значение меньшее x (рисунок 2):

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

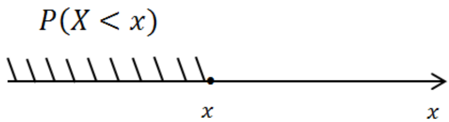

Рисунок 2. К определению функции распределения $F(x)$

Отметим свойства функции распределения:

1.Функция распределения случайной величины есть неотрицательная функция, заключенная между нулем и единицей:

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

2.Функция распределения есть не убывающая функция на всей числовой оси. Рассмотрим это свойство подробнее. Пусть есть два значения аргумента $x_1 < x_2$. Тогда: $P(X < x_2) - P(X < x_1) = P(x_1 \le X < x_2)$. Поскольку: $P(X < x_2) = F(x_2)$, $P(X < x_1) = F(x_1)$, то

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

Из этой формулы следует, что:

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

3.На минус бесконечности функция распределения равна нулю:

$$F(-\infty) = P(X < -\infty) = P(\text{невозможное событие}) = 0.$$

4.На плюс бесконечности функция распределения равна единице:

$$F(+\infty) = P(X < +\infty) = P(\text{достоверное событие}) = 1.$$

5.Вероятность попадания случайной величины в интервал $[x_1; x_2)$ (включая x_1) равна приращению ее функции распределения на этом интервале:

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

Последнее свойство 5 удобно использовать для расчета вероятностей событий.

Пример 2.4.1. Задан ряд распределения дискретной случайной величины:

X	1	3	5	9
$P(X)$	0,15	0,5	0,3	0,05

Найти функцию распределения, также вероятность попадания случайной величины в интервал $[1{,}2; 5{,}7)$.

Будем задавать различные значения x аргумента функции распределения $F(x)$ и по формуле: $F(x) = P(X < x)$ находить $F(x)$.

1. Если $x \leq 1$, то $F(x) = P(X < 1) = 0$.
2. Если $1 < x \leq 3$, то $F(x) = P(X < 3) = P(X = 1) = 0{,}15$.
3. Если $3 < x \leq 5$, то
$$F(x) = P(X < 5) = P(X = 1) + P(X = 3) = 0{,}15 + 0{,}5 = 0{,}65$$
4. Если $5 < x \leq 9$, то
$$F(x) = P(X < 9) = P(X = 1) + P(X = 3) + P(X = 5) = \\ = 0{,}15 + 0{,}5 + 0{,}3 = 0{,}95.$$
5. Если $x > 9$, то
$$F(x) = P(X = 1) + P(X = 3) + P(X = 5) + P(X = 9) = \\ = 0{,}15 + 0{,}5 + 0{,}3 + 0{,}05 = 1.$$

Функция распределения имеет вид:
$$F(x) = \begin{cases} 0 & \text{при } x \leq 1; \\ 0{,}15 & \text{при } 1 < x \leq 3; \\ 0{,}65 & \text{при } 3 < x \leq 5; \\ 0{,}95 & \text{при } 5 < x \leq 9; \\ 1 & \text{при } x > 9. \end{cases}$$

Найдем вероятность попадания случайной величины в интервал $(1{,}2; 5{,}7)$:
$$P(1{,}2 \leq X < 5{,}7) = F(5{,}7) - F(1{,}2) = 0{,}95 - 0{,}15 = 0{,}8.$$

Ответ $F(x) = \begin{cases} 0 & \text{при } x \leq 1; \\ 0{,}15 & \text{при } 1 < x \leq 3; \\ 0{,}65 & \text{при } 3 < x \leq 5; \\ 0{,}95 & \text{при } 5 < x \leq 9; \\ 1 & \text{при } x > 9. \end{cases}$ $\quad P(1{,}2 \leq X < 5{,}7) = 0{,}8.$

2.5. Непрерывные случайные величины. Функция плотности вероятности

Определение. Случайная величина X называется *непрерывной*, если ее функция распределения непрерывна в любой точке и дифференцируема всюду, кроме, быть может, отдельных точек.

Для непрерывной случайной величины справедлива теорема: Вероятность любого отдельно взятого значения случайной величины равна нулю.

Рассмотрим доказательство этого факта. Пусть x_0 – это какое-то значение случайной величины X. Представим:

$$P(X = x_0) = \lim_{x_1 \to x_0} P(x_0 \leq X < x_1).$$

Используя свойства функции распределения, запишем:

$$(X = x_0) = \lim_{x_1 \to x_0} P(x_0 \leq X < x_1) = \lim_{x_1 \to x_0} \big(F(x_1) - F(x_0)\big) =$$
$$= F(x_0) - F(x_0) = 0.$$

Что и требовалось доказать.

Данная теорема позволяет сформулировать следствие: если X – непрерывная случайная величина, то вероятность попадания случайной величины в интервал (x_1, x_2) не зависит от того, является этот интервал открытым или закрытым, т.е.:

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

Для того чтобы ввести понятие плотности вероятности непрерывной случайной величины рассмотрим некоторый интервал $[x; x + \Delta x]$. Найдем вероятность:

$$P(x \leq X \leq x + \Delta x) = F(x + \Delta x) - F(x),$$

тогда вероятность, приходящаяся на единицу длины, будет равна:

$$\frac{P(x \leq X \leq x + \Delta x)}{\Delta x} = \frac{F(x + \Delta x) - F(x)}{\Delta x},$$

Переходя к пределу при $\Delta x \to 0$, получим плотность вероятности $\varphi(x)$ в точке x:

$$\varphi(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F^{/}(x).$$

Определение. *Плотностью вероятности* $\varphi(x)$ непрерывной случайной величины X называется производная ее функции распределения $F^{/}(x)$.

Отметим основные свойства функции плотности вероятности:

1.Плотность вероятности является неотрицательной функцией $\varphi(x) \geq 0$.

2.Вероятность попадания непрерывной случайной величины в интервал $[a;b]$ равна определенному интегралу от ее плотности вероятности в пределах от a до b:

$$P(a \leq X \leq b) = \int_a^b \varphi(x)dx.$$

3.Функция распределения непрерывной случайной величины может быть выражена через плотность вероятности по формуле:

$$F(x) = \int_{-\infty}^{x} \varphi(x)dx.$$

4.Несобственный интеграл в бесконечных пределах от плотности вероятности непрерывной случайной величины равен единице:

$$F(+\infty) = \int_{-\infty}^{+\infty} \varphi(x)dx = 1.$$

Математическое ожидание $M(X)$ и дисперсию непрерывной случайной величины можно выразить через плотность вероятности по формулам:

$$a = M(X) = \int_{-\infty}^{+\infty} x\,\varphi(x)dx\,;$$

$$D(X) = \int_{-\infty}^{+\infty} (x-a)^2\,\varphi(x)dx.$$

Отметим, что на практике удобно находить дисперсию случайной величины по свойству 3:

$$D(X) = M(X^2) - (M(X))^2,$$

где:

$$M(X^2) = \int_{-\infty}^{+\infty} x^2\,\varphi(x)dx.$$

Пример 2.5.1. Функция плотности вероятности задана в виде:

$$\varphi(x) = \begin{cases} 0, \text{при } x \leq 1, \\ \dfrac{C}{x^6}, \text{при } x > 1. \end{cases}$$

Найти 1) константу C; 2) математическое ожидание $M(X)$; 3) дисперсию $D(X)$; 4) вероятность $P(1 \leq X \leq 2)$; 5) функцию распределения $F(x)$.

Решение

1) $\int\limits_{-\infty}^{+\infty} \varphi(x)dx = \int\limits_{-\infty}^{1} 0\,dx + \int\limits_{1}^{+\infty} \frac{C}{x^6}\,dx = -\frac{C}{5\,x^5}\Big|_1^{+\infty} = 0 + \frac{C}{5\cdot 1^5} = \frac{C}{5} = 1,$

$C = 5.$

2) $M(X) = \int\limits_{-\infty}^{+\infty} x\,\varphi(x)dx = \int\limits_{-\infty}^{1} x\,0\,dx + \int\limits_{1}^{+\infty} x\,\frac{5}{x^6}\,dx = \int\limits_{1}^{+\infty} \frac{5}{x^5}\,dx =$

$= -\frac{5}{4\,x^4}\Big|_1^{+\infty} = 0 + \frac{5}{4\cdot 1^4} = \frac{5}{4} = 1{,}25.$

3) $M(X^2) = \int\limits_{-\infty}^{+\infty} x^2\,\varphi(x)dx = \int\limits_{-\infty}^{1} x^2\,0\,dx + \int\limits_{1}^{+\infty} x^2\,\frac{5}{x^6}\,dx = \int\limits_{1}^{+\infty} \frac{5}{x^4}\,dx =$

$= -\frac{5}{3\,x^3}\Big|_1^{+\infty} = 0 + \frac{5}{3\cdot 1^3} = \frac{5}{3},\quad D(X) = M(X^2) - \big(M(X)\big)^2 = \frac{5}{3} - \frac{5}{4} = \frac{5}{12}.$

4) $P(1 \le X \le 2) = \int\limits_{1}^{2} \varphi(x)dx = \int\limits_{1}^{2} \frac{5}{x^6}\,dx = -\frac{5}{5\,x^5}\Big|_1^{2} = -\frac{1}{2^5} + \frac{1}{1^5} =$

$= 1 - \frac{1}{32} = \frac{31}{32}.$

5) Если $x \le 1$, то $F(x) = \int\limits_{-\infty}^{x} \varphi(x)dx = \int\limits_{-\infty}^{x} 0\,dx = 0,$

Если $x > 1$, то $F(x) = \int\limits_{-\infty}^{x} \varphi(x)dx = \int\limits_{-\infty}^{1} 0\,dx + \int\limits_{1}^{x} \frac{5}{x^6}\,dx = -\frac{5}{5\,x^5}\Big|_1^{x} =$

$= -\frac{1}{x^5}\Big|_1^{x} = -\frac{1}{x^5} + \frac{1}{1^5} = 1 - \frac{1}{x^5}.$

$$F(x) = \begin{cases} 0, & \text{при } x \le 1, \\ 1 - \frac{1}{x^5}, & \text{при } x > 1. \end{cases}$$

Ответ 1) $C = 5$; 2) $M(X) = 5/4$; 3) $D(X) = 5/12$; 4) $P(1 \le X \le 2) = 31/32$;

5) $F(x) = \begin{cases} 0, & \text{при } x \le 1, \\ 1 - \frac{1}{x^5}, & \text{при } x > 1. \end{cases}$

Глава 3. ОСНОВНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

3.1. Биномиальный закон распределения

Определение. Дискретная случайная величина X имеет *биномиальный закон распределения* с параметрами n и p, если она принимает значения 0, 1, 2, …, m, …, n с вероятностями:

$$P(X = m) = C_n^m\, p^m q^{n-m},$$

где $0 < p < 1, q = 1 - p$.

В табличном виде этот закон распределения можно представить так:

x_i	0	1	2		m		n
p_i	q^n	$C_n^1\, pq^{n-1}$	$C_n^2\, p^2 q^{n-2}$		$C_n^m\, p^m q^{n-m}$		p^n

Найдем сумму всех вероятностей данного закона:

$$\sum_{i=1}^{n} p_i = q^n + C_n^1\, pq^{n-1} + C_n^2\, p^2 q^{n-2} + \dots + C_n^m\, p^m q^{n-m} + \dots + p^n.$$

Используя формулу Бинома Ньютона, найдем:

$$\sum_{i=1}^{n} p_i = (q + p)^n = (1 - p + p)^n = 1^n = 1.$$

Поскольку основное свойство закона распределения выполняется, следовательно, формула Бернулли задает закон распределения случайной величины.

Для подсчета математического ожидания и дисперсии для данного закона распределения введем индикатор события X_k с законом распределения:

X_k	0	1
$P(X_k)$	q	p

Математическое ожидание индикатора события:

$$M(X_k) = 0 \cdot q + 1 \cdot p = p.$$

Дисперсия индикатора события:

$$D(X_k) = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p = p^2 q + q^2 p = pq(p + q) = pq.$$

Случайную величину X – распределенную по биномиальному закону можно представить как сумму n индикаторов событий:

$$X = X_1 + X_2 + \dots + X_n = \sum_{k=1}^{n} X_k.$$

Математическое ожидание этой величины тогда определится выражением:

$$M(X) = M\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} M(X_k) = \sum_{k=1}^{n} p = np.$$

Дисперсия же примет вид:

$$D(X) = D\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} D(X_k) = \sum_{k=1}^{n} pq = npq.$$

Таким образом, нами доказана теорема: Математическое ожидание и дисперсия случайной величины X, распределенной по биномиальному закону, определяются выражениями:

$$M(X) = np, \quad D(X) = npq.$$

Следствие: Математическое ожидание и дисперсия частости m/n события в n независимых испытаниях, в каждом из которых оно может наступить с одной и той же вероятностью p равны:

$$M\left(\frac{m}{n}\right) = p; \qquad D\left(\frac{m}{n}\right) = \frac{pq}{n}.$$

Рассмотрим доказательство этого следствия:

$$M\left(\frac{m}{n}\right) = M\left(\frac{X}{n}\right) = \frac{1}{n}M(X) = \frac{1}{n}np = p.$$

$$D\left(\frac{m}{n}\right) = D\left(\frac{X}{n}\right) = \frac{1}{n^2}D(X) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Что и требовалось доказать.

Пример 3.1.1. В партии 10 % нестандартных деталей. Наудачу отобраны четыре детали. Написать биномиальный закон распределения дискретной случайной величины X – числа нестандартных деталей среди четырех отобранных. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение.

Решение. Вероятность выбрать нестандартную деталь в одном испытании равна $p = 0{,}1$, а $q = 1 - p = 0{,}9$. Число испытаний $n = 4$. Число нестандартных деталей среди четырех отобранных может принимать значения $X = 0{,}1{,}2{,}3{,}4$. Вычислим вероятности этих событий:

$$P(X = 0) = C_4^0 \cdot 0{,}1^0 \cdot 0{,}9^{4-0} = \frac{4!}{0!\,(4-0)!} \cdot 1 \cdot 0{,}6561 = 0{,}6561;$$

$$P(X = 1) = C_4^1 \cdot 0{,}1^1 \cdot 0{,}9^{4-1} = \frac{4!}{1!\,(4-1)!} \cdot 0{,}1 \cdot 0{,}729 = 0{,}2916;$$

$$P(X = 2) = C_4^2 \cdot 0{,}1^2 \cdot 0{,}9^{4-2} = \frac{4!}{2!\,(4-2)!} \cdot 0{,}01 \cdot 0{,}81 = 0{,}0486;$$

$$P(X=3)=C_4^3\cdot 0{,}1^3\cdot 0{,}9^{4-3}=\frac{4!}{3!\,(4-3)!}\cdot 0{,}001\cdot 0{,}9=0{,}0036;$$

$$P(X=4)=C_4^4\cdot 0{,}1^4\cdot 0{,}9^{4-4}=\frac{4!}{4!\,(4-4)!}\cdot 0{,}0001\cdot 1=0{,}0001.$$

Закон распределения имеет вид:

X	0	1	2	3	4
$P(X)$	0,6561	0,2916	0,0486	0,0036	0,0001

Выполним контроль:

$$0{,}6561+0{,}2916+0{,}0486+0{,}0036+0{,}0001=1.$$

Математическое ожидание (непосредственный расчет):

$$M(X)=0\cdot 0{,}6561+1\cdot 0{,}2916+2\cdot 0{,}0486+3\cdot 0{,}0036+$$
$$+4\cdot 0{,}0001=0{,}4.$$

По теореме: $M(X)=np=4\cdot 0{,}1=0{,}4$.

Для вычисления дисперсии найдем:

$$M(X^2)=0^2\cdot 0{,}6561+1^2\cdot 0{,}2916+2^2\cdot 0{,}0486+$$
$$+3^2\cdot 0{,}0036+4^2\cdot 0{,}0001=0{,}52.$$

Дисперсия (непосредственный расчет):

$$D(X)=M(X^2)-\big(M(X)\big)^2=0{,}52-0{,}4^2=0{,}36.$$

По теореме $D(X)=npq=4\cdot 0{,}1\cdot 0{,}9=0{,}36$.

Среднее квадратическое отклонение $\sigma(X)=\sqrt{0{,}36}=0{,}6$.

Ответ $M(X)=0{,}4; D(X)=0{,}36;\ \sigma(X)=0{,}6$.

3.2. Закон распределения Пуассона

Определение. Дискретная случайная величина X имеет *закон распределения Пуассона* с параметром λ, если она принимает значения 0, 1, 2, …, m, … с вероятностями:

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

В табличном виде этот закон имеет вид:

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

Найдем сумму вероятностей:

$$\sum_{i=1}^{+\infty} p_i = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \cdots + \frac{\lambda^m e^{-\lambda}}{m!} + \cdots =$$

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

Поскольку сумма вероятностей равна единице, то определение закона Пуассона корректно.

Найдем математическое ожидание случайной величины X, имеющей закон распределения Пуассона:

$$M(X) = 0 \cdot e^{-\lambda} + 1 \cdot \frac{\lambda e^{-\lambda}}{1!} + 2 \cdot \frac{\lambda^2 e^{-\lambda}}{2!} + \cdots + m \cdot \frac{\lambda^m e^{-\lambda}}{m!} + \cdots =$$

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

Для того, чтобы найти дисперсию случайной величины X, вычислим:

$$M(X^2) = 0^2 \cdot e^{-\lambda} + 1^2 \cdot \frac{\lambda e^{-\lambda}}{1!} + 2^2 \cdot \frac{\lambda^2 e^{-\lambda}}{2!} + \cdots + m^2 \cdot \frac{\lambda^m e^{-\lambda}}{m!} + \cdots =$$

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

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

Следовательно, дисперсия случайной величины будет равна:

$$D(X) = M(X^2) - \big(M(X)\big)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Сформулируем теорему: Математическое ожидание и дисперсия случайной величины, распределенной по закону Пуассона, совпадают и равны параметру λ этого закона:

$$M(X) = \lambda, \quad D(X) = \lambda.$$

Пример 3.2.1. Магазин получил 1000 бутылок минеральной воды. Вероятность того, что при перевозке бутылка окажется разбитой, равна 0,003. Найти вероятность того, что при перевозке будет разбито не более 2 бутылок, а также среднее число разбитых бутылок.

Решение. По условию $n = 1000,\ p = 0{,}003$.

Среднее число разбитых бутылок:

$M(X) = \lambda = n \cdot p = 1000 \cdot 0{,}003 = 3.$

$$P_{1000}(m \le 2) = P_{1000}(m = 0) + P_{1000}(m = 1) + P_{1000}(m = 2) =$$

$$= \frac{\lambda^0\, e^{-\lambda}}{0!} + \frac{\lambda^1\, e^{-\lambda}}{1!} + \frac{\lambda^2\, e^{-\lambda}}{2!} = \left(1 + \lambda + \frac{\lambda^2}{2}\right) e^{-\lambda} =$$

$$= \left(1 + 3 + \frac{3^2}{2}\right) e^{-3} = 0{,}4232.$$

Ответ $P_{1000}(m \le 2) = 0{,}4232,\ \ M(X) = \lambda = 3.$

3.3. Равномерный закон распределения

Определение. Непрерывная случайная величина X имеет *равномерный закон распределения* на отрезке от $[a, b]$, если ее плотность вероятности $\varphi(x)$ постоянна на этом отрезке и равна нулю вне его, т.е.:

$$\varphi(x) = \begin{cases} C & \text{при } a \le x \le b \\ 0 & \text{при } \ \ x < a \text{ и } x > b. \end{cases}$$

Найдем константу C. Для этого воспользуемся свойством плотности вероятности:

$$\int_{-\infty}^{+\infty} \varphi(x) dx = 1,$$

$$\int_{-\infty}^{a} 0\, dx + \int_{a}^{b} C\, dx + \int_{b}^{+\infty} 0\, dx = C\, x|_a^b = C(b - a) = 1, \qquad C = \frac{1}{b - a}.$$

Таким образом, плотность вероятности равномерно распределенной случайной величины имеет вид:

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

Найдем функцию распределения $F(x)$ случайной величины X, распределенной по равномерному закону:

1.Если $x \le a$, то:

$$F(x) = \int_{-\infty}^{x} 0\, dx = 0.$$

2.Если $a < x \le b$, то:

$$F(x) = \int_{-\infty}^{a} 0\, dx + \int_{a}^{x} \frac{1}{b - a}\, dx = \frac{x}{b - a}\bigg|_a^x = \frac{x - a}{b - a}.$$

3.Если $x > b$, то:

$$F(x) = \int_{-\infty}^{a} 0\, dx + \int_{a}^{b} \frac{1}{b - a}\, dx + \int_{b}^{+\infty} 0\, dx = \frac{x}{b - a}\bigg|_a^b = \frac{b - a}{b - a} = 1.$$

Окончательно функция распределения будет иметь вид:

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

Найдем математическое ожидание случайной величины X, имеющей равномерный закон распределения:

$$M(X)=\int_{-\infty}^{+\infty} x\,\varphi(x)\,dx=\int_{-\infty}^{a} x\,0\,dx+\int_{a}^{b}\frac{x}{b-a}\,dx+\int_{b}^{+\infty} x\,0\,dx=$$

$$=\frac{x^2}{2(b-a)}\bigg|_a^b=\frac{b^2-a^2}{2(b-a)}=\frac{(b-a)(b+a)}{2(b-a)}=\frac{b+a}{2}.$$

Чтобы найти выражение для дисперсии, найдем математическое ожидание квадрата случайной величины:

$$M(X^2)=\int_{-\infty}^{+\infty} x^2\,\varphi(x)\,dx=\int_{-\infty}^{a} x^2\,0\,dx+\int_{a}^{b}\frac{x^2}{b-a}\,dx+\int_{b}^{+\infty} x^2\,0\,dx=$$

$$=\frac{x^3}{3(b-a)}\bigg|_a^b=\frac{b^3-a^3}{3(b-a)}=\frac{(b-a)(b^2+ab+a^2)}{3(b-a)}=\frac{b^2+ab+a^2}{3}.$$

Следовательно, дисперсия будет равна:

$$D(X)=M(X^2)-\big(M(X)\big)^2=\frac{b^2+ab+a^2}{3}-\left(\frac{b+a}{2}\right)^2=$$

$$=\frac{b^2+ab+a^2}{3}-\frac{b^2+2ab+a^2}{4}=\frac{b^2-2ab+a^2}{12}=\frac{(b-a)^2}{12}.$$

Пример 3.3.1. Автобусы некоторого маршрута идут строго по расписанию. Интервал движения 5 минут. Найти вероятность того, что пассажир, подошедший к остановке, будет ожидать очередной автобус менее 3 минут.

Решение. Пусть X – это время ожидания автобуса. Оно может быть от 0 до 5 минут. Значит $a=0,\ b=5$, функция плотности:

$$\varphi(x)=\begin{cases} \dfrac{1}{5} & \text{при } 0 \le x \le 5 \\ 0 & \text{при } \quad x > 5. \end{cases}$$

Для того, чтобы ждать автобус менее 3 минут, нужно прийти на остановку между 2 и 5 минутами. Вероятность этого события:

$$P(2 \le X \le 5) = \int_2^5 \frac{1}{5} dx = \frac{x}{5}\bigg|_2^5 = \frac{5-2}{5} = \frac{3}{5} = 0{,}6.$$

Ответ 0,6.

3.4. Показательный закон распределения

Определение. Непрерывная случайная величина X имеет *показательный закон распределения* с параметром $\lambda > 0$, если ее плотность вероятности имеет вид:

$$\varphi(x) = \begin{cases} Ce^{-\lambda x} & \text{при } x \ge 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Найдем константу C из условия нормировки:

$$\int_{-\infty}^{+\infty} \varphi(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{+\infty} Ce^{-\lambda x}\, dx = -\frac{C}{\lambda} e^{-\lambda x}\bigg|_0^{+\infty} = \frac{C}{\lambda} = 1, \ \ C = \lambda.$$

Таким образом, плотность вероятности для показательного закона распределения имеет вид:

$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} & \text{при } x \ge 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Найдем функцию распределения случайной величины X, распределенной по показательному закону:

1.Если $x < 0$, то:

$$F(x) = \int_{-\infty}^{x} 0\, dx = 0.$$

2.Если $x \ge 0$, то:

$$F(x) = \int_{-\infty}^{0} 0\, dx + \int_{0}^{x} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_0^x = 1 - e^{-\lambda x}.$$

Следовательно, функция распределения будет иметь вид:

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

Математическое ожидание случайной величины X, имеющей показательный закон распределения имеет вид:

$$M(X) = \int_{-\infty}^{+\infty} x\, \varphi(x) dx = \int_{0}^{+\infty} x\, \lambda e^{-\lambda x} dx.$$

Применим формулу вычисления интеграла по частям:

$$\int_a^b U dV = U\,V - \int_a^b V dU.$$

Обозначим части:

$$U = x, \quad dU = dx, \quad dV = \lambda e^{-\lambda x} dx, \quad V = -e^{-\lambda x},$$

Тогда:

$$M(X) = -xe^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx = -\frac{e^{-\lambda x}}{\lambda}\bigg|_0^{+\infty} = \frac{1}{\lambda}.$$

Далее найдем:

$$M(X^2) = \int_{-\infty}^{+\infty} x^2\,\varphi(x) dx = \int_0^{+\infty} x^2\,\lambda e^{-\lambda x} dx.$$

Дважды применяя формулу вычисления интеграла по частям, найдем:

$$M(X^2) = \int_0^{+\infty} x^2\,\lambda e^{-\lambda x} dx = -2xe^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} 2x\,e^{-\lambda x} dx =$$

$$= -2\frac{e^{-\lambda x}}{\lambda}\bigg|_0^{+\infty} + \int_0^{+\infty} \frac{2}{\lambda}\,e^{-\lambda x} dx = -\frac{2}{\lambda^2} e^{-\lambda x}\bigg|_0^{+\infty} = \frac{2}{\lambda^2}$$

Следовательно, дисперсия будет равна:

$$D(X) = M(X^2) - \big(M(X)\big)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Пример 3.4.1. Непрерывная случайная величина X распределена по показательному закону, заданному при $x \geq 0$ плотность распределения $f(x) = 0{,}04\,e^{-0{,}04\,x}$; при $x < 0$ функцией $f(x) = 0$. Найти вероятность того, что в результате испытания X попадает в интервал $(1, 2)$.

Решение.

$$P(1 \leq X \leq 2) = \int_1^2 0{,}04\,e^{-0{,}04\,x}\,dx = -e^{-0{,}04\,x}\big|_1^2 =$$

$$= -e^{-0{,}08} + e^{-0{,}04} = -0{,}9231 + 0{,}9608 = 0{,}0377.$$

Ответ $P(1 \leq X \leq 2) = 0{,}0377.$

3.5. Нормальный закон распределения

Определение. Непрерывная случайная величина X имеет *нормальный закон распределения* (закон Гаусса) с параметрами a и σ^2, если ее плотность вероятности имеет вид:

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

Из выражения для $\varphi(x)$ видно, что максимальное значение данная функция принимает при $x = a$, а само максимальное значение равно:

$$\varphi(a) = \varphi_{max} = \frac{1}{\sigma\sqrt{2\pi}}.$$

Таким образом, параметр a влияет на положение максимума данной функции, а параметр σ определяет высоту максимума функции (рисунок 3).

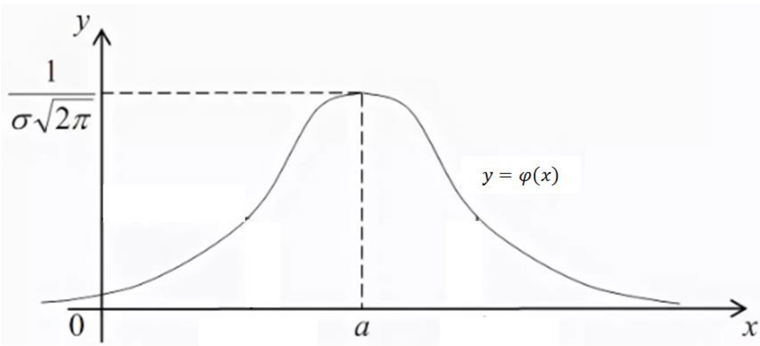

Рисунок 3. График функции плотности вероятности нормально распределенной случайной величины

Найдем функцию распределения нормально распределенной случайной величины:

$$F(x) = \int_{-\infty}^{x} \varphi(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-a)^2}{2\sigma^2}} dx.$$

Выполним замену переменной:

$$t = \frac{x-a}{\sigma}, \quad x = a + \sigma t, \;\; dx = \sigma\, dt.$$

Тогда:

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-a}{\sigma}} e^{-\frac{t^2}{2}}\,dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{t^2}{2}}\,dt+\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{x-a}{\sigma}} e^{-\frac{t^2}{2}}\,dt.$$

Первый из интегралов сведем к интегралу Эйлера-Пуассона:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{t^2}{2}}\,dt=\frac{\sqrt{2}}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\left(\frac{t}{\sqrt{2}}\right)^2}\,d\left(\frac{t}{\sqrt{2}}\right)=\frac{\sqrt{2}}{\sqrt{2\pi}}\cdot\frac{\sqrt{\pi}}{2}=\frac{1}{2}.$$

Для вычисления второго интеграла введем *функцию Лапласа* (для нее составлены таблицы):

$$\phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{t^2}{2}}\,dt.$$

Следовательно, функция распределения нормально распределенной случайной величины будет иметь вид:

$$F(x)=\frac{1}{2}+\phi\left(\frac{x-a}{\sigma}\right).$$

Отметим основные свойства функции Лапласа $\phi(x)$:

1.При $x=0$ функция Лапласа равна нулю $\phi(0)=0$.

2.Функция Лапласа является монотонно возрастающей функцией.

3.Функция Лапласа является нечетной функцией $\phi(-x)=-\phi(x)$.

4.При значениях аргумента $x>5$, функция принимает значения $\phi(x>5)=0{,}5$.

Для значений аргумента $0\le x\le 5$ значения функции Лапласа можно найти в таблице приложения.

Смысл параметров a и σ^2 определяет теорема: Математическое ожидание и дисперсия случайной величины X, распределенной по нормальному закону равны:

$$M(X)=a,\quad D(X)=\sigma^2.$$

Найдем вероятность того, что случайная величина, распределенная по нормальному закону, попадет в интервал от x_1 до x_2:

$$P(x_1\le X\le x_2)=F(x_2)-F(x_1)=\frac{1}{2}+\phi\left(\frac{x_2-a}{\sigma}\right)-\frac{1}{2}-\phi\left(\frac{x_1-a}{\sigma}\right)=$$
$$=\phi\left(\frac{x_2-a}{\sigma}\right)-\phi\left(\frac{x_1-a}{\sigma}\right).$$

Вероятность того, что отклонение случайной величины X, распределенной по нормальному закону, от математического ожидания a не превысит величину $\Delta>0$, равна:

$$P(|X-a|\le\Delta)=P(a-\Delta\le X\le a+\Delta)=$$

$$= \phi\left(\frac{a+\Delta-a}{\sigma}\right) - \phi\left(\frac{a-\Delta-a}{\sigma}\right) =$$
$$= \phi\left(\frac{\Delta}{\sigma}\right) + \phi\left(\frac{\Delta}{\sigma}\right) = 2\phi\left(\frac{\Delta}{\sigma}\right).$$

Используя последнюю формулу, вычислим вероятности:

$\Delta = \sigma \qquad P(|X-a| \le \Delta) = 2\phi\left(\frac{\sigma}{\sigma}\right) = 2\phi(1) = 2 \cdot 0{,}3413 = 0{,}6826;$

$\Delta = 2\sigma \qquad P(|X-a| \le \Delta) = 2\phi\left(\frac{2\sigma}{\sigma}\right) = 2\phi(2) = 2 \cdot 0{,}4772 = 0{,}9544;$

$\Delta = 3\sigma \qquad P(|X-a| \le \Delta) = 2\phi\left(\frac{3\sigma}{\sigma}\right) = 2\phi(3) = 2 \cdot 0{,}49865 = 0{,}9973.$

Из этих вычислений следует «*правило трех сигма*»: Если случайная величина X имеет нормальный закон распределения с параметрами a и σ^2, то практически достоверно, что ее значения заключены в интервале $(a-3\sigma^2; a+3\sigma^2)$.

Пример 3.5.1 Математическое ожидание и среднее квадратическое отклонение нормально распределенной случайной величины X соответственно равны 20 и 5. Найти вероятность того, что в результате испытания X примет значение, заключенное в интервале (15,25).

Решение. По условию $a = 20,\ \sigma = 5$.

$$P(15 \le X \le 25) = \phi\left(\frac{x_2-a}{\sigma}\right) - \phi\left(\frac{x_1-a}{\sigma}\right) =$$
$$= \phi\left(\frac{25-20}{5}\right) - \phi\left(\frac{15-20}{5}\right) =$$
$$= \phi(1) - \phi(-1) = 2 \cdot \phi(1) = 2 \cdot 0{,}3413 = 0{,}6826.$$

Ответ $P(15 \le X \le 25) = 0{,}6826$

3.6. Локальная и интегральная теоремы Муавра-Лапласа

Рассматривая биномиальный закон распределения, были доказаны формулы для математического ожидания и дисперсии:

$$M(X) = a = np, \qquad D(X) = npq.$$

При достаточно большом числе испытаний биномиальный закон распределения можно аппроксимировать нормальным распределением с параметрами:

$$a = np, \qquad \sigma^2 = npq.$$

В частности, справедливы теоремы Муавра-Лапласа.

Локальная теорема Муавра-Лапласа: если вероятность p наступления события A в каждом испытании постоянна и отлична от 0 и 1, то вероятность $P_n(m)$ того, что событие A произойдет m раз в n независимых испытаниях при достаточно большом числе n, приближенно равна:

$$P_n(m) = \frac{1}{\sqrt{npq}} f\left(\frac{m - np}{\sqrt{npq}}\right),$$

где:

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \; - \text{функция Гаусса.}$$

Интегральная теорема Муавра-Лапласа: если вероятность p наступления события A в каждом испытании постоянна и отлична от 0 и 1, то вероятность того, что число m наступления события A в n независимых испытаниях заключено в пределах от m_1 до m_2 (включительно), при достаточно большом числе n приближенно равна:

$$P_n(m_1 \leq m \leq m_2) = \phi\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \phi\left(\frac{m_1 - np}{\sqrt{npq}}\right)$$

Отметим, что данные теоремы дают приемлемые результаты при условии, что $npq \geq 20$.

Рассмотрим следствие интегральной теоремы Муавра-Лапласа.

Следствие. Если вероятность p наступления события A в каждом испытании постоянна и отлична от 0 и 1, то при достаточно большом числе n независимых испытаний вероятность того, что:

а) число m наступлений события A отличается от произведения np не более, чем на величину $\varepsilon > 0$ (по абсолютной величине), т.е.:

$$P_n(|m - np| \leq \varepsilon) = 2\phi\left(\frac{\varepsilon}{\sqrt{npq}}\right);$$

б) частость m/n события A заключена в пределах от α до β:

$$P_n\left(\alpha \leq \frac{m}{n} \leq \beta\right) = \phi\left(\frac{\beta - p}{\sqrt{pq/n}}\right) - \phi\left(\frac{\alpha - p}{\sqrt{pq/n}}\right);$$

в) частость m/n события A отличается от его вероятности p не более, чем на величину $\Delta > 0$:

$$P_n\left(\left|\frac{m}{n} - p\right| \leq \Delta\right) = 2\phi\left(\frac{\Delta\sqrt{n}}{\sqrt{pq}}\right).$$

Пример 3.6.1. Вероятность появления события в каждом из 2100 независимых испытаний равна 0,7. Найти вероятность того, что событие появится: а) не менее 1470 и не более 1500 раз; б) не менее 1470 раз; в) не более 1469 раз.

Решение.

По условию $n = 2100;\ p = 0{,}7;\quad n \cdot p = 2100 \cdot 0{,}7 = 1470;$ $\sqrt{n \cdot p \cdot q} = \sqrt{2100 \cdot 0{,}7 \cdot (1 - 0{,}7)} = 21.$

а) $$P_n(1470 \le m \le 1500) = \phi\left(\frac{1500 - 1470}{21}\right) - \phi\left(\frac{1470 - 1470}{21}\right) =$$
$$= \phi(1{,}43) - \phi(0) = 0{,}4236 - 0 = 0{,}4236.$$

б) $$P_n(1470 \le m \le 2100) = \phi\left(\frac{2100 - 1470}{21}\right) - \phi\left(\frac{1470 - 1470}{21}\right) =$$
$$= \phi(30) - \phi(0) = 0{,}5 - 0 = 0{,}5.$$

в) $$P_n(0 \le m \le 1469) = \phi\left(\frac{1469 - 1470}{21}\right) - \phi\left(\frac{0 - 1470}{21}\right) =$$
$$= \phi(-0{,}05) - \phi(-70) = -\phi(0{,}05) + \phi(70) = -0{,}0199 + 0{,}5 = 0{,}4801.$$

Ответ а) 0,4236; б) 0,5; в) 0,4801.

Вопросы для самопроверки по разделам 1-3

1. Классификация случайных событий: возможные и невозможные события, совместные и несовместные, противоположные и достоверные события.
2. Полная группа событий. Пространство элементарных исходов.
3. Классическое определение вероятности события. Свойства вероятности события.
4. Статистическое определение вероятности события. Теорема Бернулли.
5. Геометрическое определение вероятности.
6. Сумма событий и ее свойства.
7. Теорема сложения вероятностей и ее следствия.
8. Произведение событий и его свойства.
9. Условная вероятность. Зависимые и независимые события. Теорема умножения вероятностей.
10. Формулы полной вероятности и Байеса.
11. Случайная величина. Дискретная случайная величина и ее закон (ряд) распределения. Основное свойство закона распределения.
12. Функция распределения дискретной случайной величины.
13. Непрерывная случайная величина. Вероятность отдельно взятого значения непрерывной случайной величины.
14. Абсолютно непрерывная случайная величина. Плотность вероятности абсолютно непрерывной случайной величины, ее определение, свойства и график.
15. Математическое ожидание случайной величины и его свойства.

16. Дисперсия случайной величины и ее свойства. Среднее квадратическое отклонение случайной величины.
17. Закон распределения Бернулли, его определение, свойства и примеры.
18. Биномиальный закон распределения, его определение, свойства и примеры.
19. Закон распределения Пуассона, его определение, свойства и примеры.
20. Равномерный закон распределения, его определение, свойства и примеры.
21. Нормальный (гауссовский) закон распределения. Геометрический и вероятностный смысл параметров нормального закона распределения.
22. Свойства случайной величины, распределенной по нормальному закону. Правило трех сигм.
23. Показательный (экспоненциальный) закон распределения, его определение, свойства и примеры.
24. Повторные независимые испытания. Формула Бернулли.
25. Локальная и интегральная теоремы Муавра—Лапласа, условия их применимости.
26. Следствия из интегральной теоремы Муавра—Лапласа.
27. Асимптотическая формула Пуассона и условия ее применимости.

Глава 4. ОСНОВЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ ВЫБОРОЧНОГО МЕТОДА

4.1. Общие сведения о выборочном методе

При изучении свойств, качеств и характеристик отдельных объектов используют два вида наблюдений: *сплошное*, когда изучают все объекты, и не сплошное, *выборочное*, когда изучается часть объектов. Примером сплошного наблюдения является анализ качества всех объектов, выпускаемых, например, на автоматизированной линии. Если же проверке на качество будет подвергаться только часть объектов, сошедших с автоматизированной линии, тогда метод исследования качества объектов можно назвать выборочным. В математической статистике различают генеральную совокупность и выборку.

Генеральная совокупность – это вся подлежащая изучению совокупность объектов (наблюдений). Понятие генеральной совокупности следует трактовать как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий.

Та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности, называется *выборочной совокупностью*, или *выборкой*. Числа объектов (наблюдений) в генеральной и выборочной совокупности называются их *объемами*. Генеральная совокупность при этом может иметь как конечный, так и бесконечный объем. Выборку можно рассматривать как некий эмпирический аналог генеральной совокупности.

Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности (выборке) выносить суждения о ее свойствах в целом.

Выборочный метод имеет ряд преимуществ перед методом сплошного изучения свойств объектов. Отметим их:

- экономит затраты всевозможных ресурсов (временные, трудовые, денежные, материальные) при изучении свойств объектов;

- является единственно возможным методом при изучении свойств объектов, разрушающихся при исследовании.

Однако выборочный метод имеет и недостаток: ошибки репрезентативности (ошибки недостаточности объема данных). Данные ошибки возникают в результате того, что изучаются не все объекты в генеральной совокупности, а только та их часть, которая попала в выборку.

Для того, чтобы уменьшить ошибки, которые могут возникать при выборочном методе, к выборкам предъявляются определенные требования. В частности, чтобы по данным выборки иметь возможность судить о генеральной совокупности, она должна быть отобрана случайным образом. Такой отбор, например, можно осуществить с помощью датчиков или таблиц случайных чисел.

Выборка называется *репрезентативной* (представительной), если она достаточно хорошо воспроизводит генеральную совокупность.

Выборки могут быть следующих видов:

- собственно-случайная выборка, образованная случайным выбором элементов без расчленения на части или группы;

- механическая выборка, в которую элементы из генеральной совокупности отбираются через определенный интервал. Например, выбирается каждый 20 – й элемент на конвейере;

- типическая выборка, в которую случайным образом отбираются элементы из типичных групп, на которые по некоторому признаку разбивается генеральная совокупность;

- серийная (гнездовая) выборка, в которую случайным образом отбираются не элементы, а целые серии элементов. А сами серии подвергаются сплошному наблюдению.

Используют два способа образования выборок:

- повторный отбор, когда каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть отобран для изучения повторно;

- безповторный отбор, когда отобранный элемент не возвращается в общую совокупность.

4.2. Выборка и ее представление

В ходе выборочного метода фиксируются значения какой-либо случайной величины X принадлежащей генеральной совокупности. Различные значения признака (случайной величины X) называются вариантами (обозначим их через x). Если объем выборки обозначить n, то различные значения признака можно представить рядом:

$$x_1, x_2, \dots, x_n.$$

Такой ряд записан по мере фиксации значений, например, в ходе эксперимента, испытания или может быть взят из баз данных и т.д. Данный ряд не удобен для осмысления, поэтому первым шагом, который целесообразно предпринять – это упорядочить (ранжировать) данный ряд по возрастанию или убыванию значений.

Пример 4.2.1. Дан ряд:

70, 75, 100, 120, 75, 60, 100, 120, 70, 60, 65, 100, 65, 100, 70, 75, 60, 100, 100, 120, 70, 75, 70, 120, 65, 70, 75, 70, 100, 100,

тогда ряд, ранжированный по возрастанию, имеет вид:

60, 60, 60, 65, 65, 65, 70, 70, 70, 70, 70, 70, 70, 75, 75, 75, 75, 75, 100, 100, 100, 100, 100, 100, 100, 100, 120, 120, 120, 120.

Объем выборки для данного случая $n = 30$.

Из приведенного примера 4.2.1 видно, что некоторые значения в ряде встречаются несколько раз. Например, значения $x_1 = 60$ и $x_2 = 65$ встречаются каждое по 3 раза, $x_3 = 70$, встречается 7 раз и т.д.

Числа, показывающие сколько раз встречаются варианты в ряде, называются *частотами* и обозначаются n_i.

Отношение частоты варианты к объему выборки называется *относительной частотой* варианты $w_i = n_i/n$.

Вариационным рядом называется ранжированный по возрастанию или убыванию ряд вариантов с соответствующими им весами (частотами или относительными частотами).

Пример 4.2.2. Составить вариационный ряд для примера 4.2.1. Рассчитать частоты и относительные частоты вариант.

Номер группы	i	1	2	3	4	5	6
Значение варианты	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4
Относительная частота	w_i	3/30	3/30	7/30	5/30	8/30	4/30

Если объем выборки очень велик, то работа с дискретными рядами оказывается затруднительной, если мы, конечно, не используем компьютерные статистические пакеты (spss, statistica, пакет R и др.). Для удобства работы с выборками большого объема служит *интервальный ранжированный ряд*. Для построения данного ряда разобьем все варианты на отдельные интервалы. Количество интервалов m следует брать не очень большим, чтобы конечный ряд не был громоздким и не очень малым, чтобы не потерять особенности распределения признака. Обычно количество интервалов берут в диапазоне от 4 до 15 или определяют по формуле Стерджеса $m = 1 + 3{,}322 \cdot \lg n$, где n – объем выборки. Ширину одного интервала определяют по формуле $\Delta x = R/m$, где $R = x_{max} - x_{min}$ – размах вариации; x_{min} и x_{max} – наименьшее и наибольшее значения варианты.

Пример 4.2.3.

Дана выборка результатов измерения роста 105 студентов.

155, 170, 185, 180, 188, 152, 173, 178, 178, 168, 185, 173, 170, 183, 175, 173, 170, 183, 175, 180, 175, 193, 178, 183, 180, 197, 178, 181, 187, 168, 174, 179, 184, 183, 178, 180, 178, 163, 166, 178, 175, 182, 190, 167, 170, 178, 183, 170, 178, 181, 173, 168, 185, 175, 170, 155, 169, 186, 179, 189, 155, 174, 179, 179, 169, 186, 174, 171, 184, 175, 193, 178, 184, 180, 196, 175, 181, 188, 168, 179, 178, 183, 184, 178, 181, 177, 163, 166, 178, 175, 183, 190, 167, 170, 178, 183, 170, 178, 182, 173, 168, 186, 176, 171, 188.

Требуется составить интервальный вариационный ряд.

Номер интервала	Рост студентов	Частота	Относительная частота
1	$[152; 157{,}5]$	4	4/105
2	$(157{,}5; 163]$	2	2/105
3	$(163; 168{,}5]$	9	9/105
4	$(168{,}5; 174]$	20	20/105
5	$(174; 179{,}5]$	30	30/105
6	$(179{,}5; 185]$	26	26/105
7	$(185; 190{,}5]$	10	10/105
8	$(190{,}5; 196]$	4	4/105
сумма		105	1

Самое малое и самое большое значения вариант $x_{min} = 152$, $x_{max} = 196$, размах вариации $R = x_{max} - x_{min} = 196 - 152 = 44$. Вычислим количество интервалов $m = 1 + 3{,}322 \cdot \lg 105 \approx 8$. Ширина интервала тогда будет $\Delta x = R/m = 44/8 = 5{,}5$.

Для графического изображения вариационных рядов используют полигон, гистограмму, кумулятивную кривую, функцию распределения.

Полигоном частот называют ломанную, отрезки которой соединяют точки $(x_1; n_1)$, $(x_2; n_2)$, …, $(x_m; n_m)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат - частоты n_i. Каждые две точки соединяют отрезками прямых.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки $(x_1; w_1)$, $(x_2; w_2)$, …, $(x_m; w_m)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – относительные частоты w_i. Каждые две точки соединяют отрезками прямых.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длинною Δx, а высоты равны n_i. Для построения гистограммы частот на оси абсцисс откладывают интервалы Δx, а над ними проводят отрезки, параллельные оси абсцисс на высоте n_i.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длинною Δx, а высоты равны w_i. Для построения гистограммы относительных частот на оси абсцисс откладывают интервалы Δx, а над ними проводят отрезки, параллельные оси абсцисс на высоте w_i.

Кумулятивная кривая (кумулята) – кривая накопленных частот или накопленных относительных частот. Для дискретного ряда кумулята представляет ломаную, соединяющую точки $(x_1; n_1^{\text{нак}})$, $(x_2; n_2^{\text{нак}})$, …, $(x_m; n_m^{\text{нак}})$ или точки $(x_1; w_1^{\text{нак}})$, $(x_2; w_2^{\text{нак}})$, …, $(x_m; w_m^{\text{нак}})$. Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого, а ордината – накопленной частоте, равной нулю. Другие точки этой ломаной соответствуют концам интервалов.

Пример 4.2.4. Для примера 4.2.2 построить полигон частот и кумуляту.

Номер группы	i	1	2	3	4	5	6
Значение варианты	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4
Накопленные частоты	$n_i^{\text{нак}}$	3	6	13	18	26	30

Кривые построены на рисунке 4 и рисунке 5.

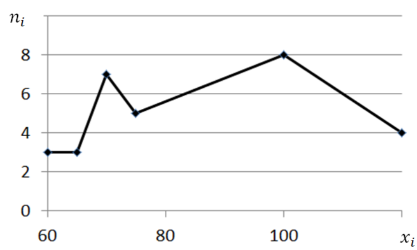

Рисунок 4. Полигон частот

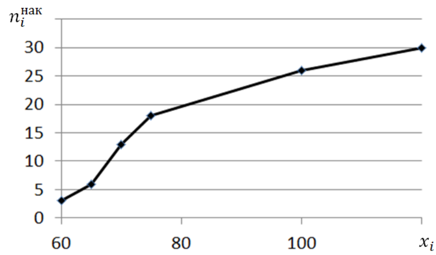

Рисунок 5. Кривая накопленных частот (кумулята)

4.3. Числовые характеристики вариационных рядов

Вариационный ряд характеризуется такими величинами как мода Mo и медиана Me. Медиана, как и мода, не зависит от крайних значений вариантов, поэтому применяется для характеристики центра в ряду распределениями с неопределенными границами.

Мода (Mo) – это значение признака (варианта), наиболее часто встречающееся в исследуемой совокупности, т.е. это одна из вариант признака, которая в ряду распределения имеет наибольшую частоту.

В дискретном ряду мода определяется визуально по максимальной частоте.

Например, в примере 4.2.4 мода $Mo = 100$, так как при $x = 100$ частота $n = 8$ принимает наибольшее значение.

В интервальном ряду по наибольшей частоте определяется модальный интервал, а конкретное значение моды в интервале вычисляется по формуле

$$Mo = x_0 + \Delta x \cdot \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})},$$

где x_0 и Δx – соответственно нижняя граница и ширина модального интервала; n_{Mo-1}, n_{Mo}, n_{Mo+1} – частоты предмодального, модального и послемодального интервала соответственно.

Пример 4.3.1. Найти моду для данных примера 4.2.3.

Модальным интервалом в данной задаче является интервал №5, т.е интервал $(174; 179{,}5]$. Значит $x_0 = 174$ и $\Delta x = 179{,}5 - 174 = 5{,}5$. Частоты интервалов $n_{Mo-1} = 20,$ $n_{Mo} = 30,$ $n_{Mo+1} = 26$. Таким образом, мода будет

$$Mo = 174 + 5{,}5 \cdot \frac{30 - 20}{(30 - 20) + (30 - 26)} = 177{,}93.$$

Ответ $Mo = 177{,}93$.

Медиана (Me) – значение признака (варианта), приходящееся на середину ранжированного вариационного ряда, т.е. это значение, которое делит ряд распределения на две равные по объему части.

В дискретном ряду медиана определяется визуально по накопленной частоте. Например, в примере 4.2.4, чтобы найти медиану нужно найти варианту, для которой накопленная частота составляет $n/2 = 30/2 = 15$. При $x_1 = 60$ накопилось $n_1^{\text{нак}} = 3$; при $x_2 = 65$ накопилось $n_2^{\text{нак}} = 6$; при $x_3 = 70$ накопилось $n_3^{\text{нак}} = 13$; при $x_4 = 70$ накопилось $n_4^{\text{нак}} = 18$. Следовательно, медиана равна $Me = 70$.

В случае интервального вариационного ряда значение медианы вычисляется по формуле:

$$Me = x_0 + \Delta x \cdot \frac{0{,}5 \cdot n - n_{Me-1}^{\text{нак}}}{n_{Me}},$$

где n – объем выборки, x_0, Δx – нижняя граница и ширина медианного интервала соответственно, $n_{Me-1}^{\text{нак}}$ – накопленная частота предмедианного интервала, n_{Me} – частота медианного интервала.

Пример 4.3.2. Для примера 4.2.3 найти медиану, построить гистограмму частот и кумуляту.

Номер интервала	Рост студентов	Частота	Накопленная частота
1	$[152; 157{,}5]$	4	4
2	$(157{,}5; 163]$	2	6
3	$(163; 168{,}5]$	9	15
4	$(168{,}5; 174]$	20	35
5	$(174; 179{,}5]$	30	65
6	$(179{,}5; 185]$	26	91
7	$(185; 190{,}5]$	10	101
8	$(190{,}5; 196]$	4	105

Объем выборки равен $n = 105$, $n/2 = 52{,}5$. Медианный интервал №5. Следовательно, $x_0 = 174$, $\Delta x = 5{,}5$, $n_{Me-1}^{\text{нак}} = 35$, $n_{Me} = 30$

$$Me = x_0 + \Delta x \cdot \frac{0{,}5 \cdot n - n_{Me-1}^{\text{нак}}}{n_{Me}} = 174 + 5{,}5 \cdot \frac{52{,}5 - 35}{30} = 177{,}208.$$

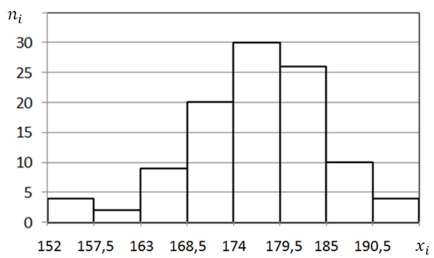

Рисунок 6. Гистограмма частот

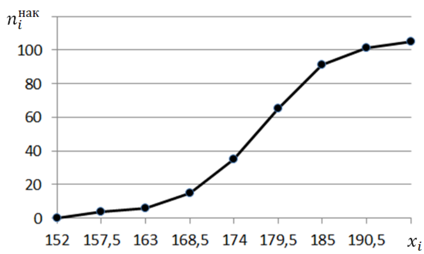

Рисунок 7. Кумулятивная кривая

Ответ $Me = 177{,}208$.

Средней арифметической вариационного ряда называется отношение суммы всех вариант к объему выборки.

Среднюю арифметическую вычисляют как среднюю арифметическую простую

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

или среднюю арифметическую взвешенную

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{m} x_i \cdot n_i\right) = \sum_{i=1}^{m} x_i \cdot w_i$$

где n – объем выборки, m – число интервалов в ранжированном ряде, n_i, w_i – частоты и относительные частоты соответственно.

Пример 4.3.3. Для примера 4.2.1 вычислить среднюю арифметическую. Используем формулу для расчета средней арифметической простой:

$$\bar{x} = (60 + 60 + 60 + 65 + 65 + 65 + 70 + 70 + 70 + 70 + 70 + 70 + \\ +70 + 75 + 75 + 75 + 75 + 75 + 100 + 100 + 100 + 100 + 100 + 100 + \\ +100 + 100 + 120 + 120 + 120 + 120)/30 = 2520/30 = 84.$$

Ответ $\bar{x} = 84$.

Пример 4.3.4. Для примера 4.2.2 вычислить среднюю арифметическую. Используем формулу для расчета средней арифметической взвешенной.

$$\bar{x} = \frac{(60 \cdot 3 + 65 \cdot 3 + 70 \cdot 7 + 75 \cdot 5 + 100 \cdot 8 + 120 \cdot 4)}{30} = \\ = \frac{2520}{30} = 84.$$

Ответ $\bar{x} = 84$.

Пример 4.3.5. Для примера 4.2.3 вычислить среднюю арифметическую. Если вычисления проводить по первоначальным данным, то нужно использовать формулу для средней арифметической простой:

$$\bar{x} = (155 + 170 + 185 + 180 + 188 + 152 + 173 + 178 + 178 + 168 + \\ +185 + 173 + 170 + 183 + 175 + 173 + 170 + 183 + 175 + 180 + \\ +175 + 193 + 178 + 183 + 180 + 197 + 178 + 181 + 187 + 168 + \\ +174 + 179 + 184 + 183 + 178 + 180 + 178 + 163 + 166 + 178 + \\ +175 + 182 + 190 + 167 + 170 + 178 + 183 + 170 + 178 + 181 + \\ +173 + 168 + 185 + 175 + 170 + 155 + 169 + 186 + 179 + 189 + \\ +155 + 174 + 179 + 179 + 169 + 186 + 174 + 171 + 184 + 175 +$$

$$+193+178+184+180+196+175+181+188+168+179+$$
$$+178+183+184+178+181+177+163+166+178+175+$$
$$+183+190+167+170+178+183+170+178+182+173+$$
$$+168+186+176+171+188)/105 = 18594/105 = 177{,}09.$$

Если вычисления проводить по интервальному вариационному ряду, то нужно использовать формулу для средней арифметической взвешенной:

$$\bar{x} = \left(\frac{152+157{,}5}{2}\cdot 4 + \frac{157{,}5+163}{2}\cdot 2 + \frac{163+168{,}5}{2}\cdot 9 + \right.$$
$$+\frac{168{,}5+174}{2}\cdot 20 + \frac{174+179{,}5}{2}\cdot 30 + \frac{179{,}5+185}{2}\cdot 26 +$$
$$\left. +\frac{185+190{,}5}{2}\cdot 10 + \frac{190{,}5+196}{2}\cdot 4\right)/105 = 18547{,}75/105 = 176{,}65.$$

Из примеров 4.3.3, 4.3.4, 4.3.5 видно, что при замене первичного ряда ранжированным интервальным рядом несколько изменяется средняя арифметическая.

Отметим основные свойства средней арифметической.

1. Средняя арифметическая постоянной равна самой постоянной.

$$\bar{c} = c, , \qquad c \text{ – постоянная.}$$

2. Если все варианты увеличить в одно и то же число раз, то средняя арифметическая увеличится во столько же раз

$$\overline{k\cdot x} = k\cdot\bar{x}, \qquad k \text{ – постоянная.}$$

3. Если все варианты увеличить на одно и то же число, то средняя арифметическая увеличится на то же число

$$\overline{x+c} = \bar{x}+c, \qquad c \text{ – постоянная.}$$

4. Средняя арифметическая отклонений вариантов от средней арифметической равна нулю

$$\overline{x-\bar{x}} = \bar{x}-\bar{x} = 0.$$

5. Средняя арифметическая алгебраической суммы нескольких признаков равна такой же сумме средних арифметических этих признаков

$$\overline{x+y} = \bar{x}+\bar{y}.$$

6. Если ряд состоит из нескольких групп, общая средняя равна средней арифметической групповых средних, причем весами являются объемы групп:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{l}\overline{x_\iota}\cdot n_i\right)$$

где $\overline{x_\iota}$ – групповая средняя i – ой группы, объем которой равен n_i, l – число групп.

Отметим так же, что среднее арифметическое относится к *аналитическим* величинам, тогда как мода и медиана относятся к *структурным* или *порядковым* величинам.

Доля вариант, удовлетворяющих какому - либо признаку – это отношение числа элементов, обладающих данным признаком, к общему числу элементов в выборке:

$$w = m/n,$$

где m – количество вариант, удовлетворяющих какому – либо признаку, n – объем выборки.

Пример 4.3.6. Для примера 4.2.2 найти долю вариант, удовлетворяющих условию $x_i \leq 70$.

$w(x_i \leq 70) = (3 + 3 + 7)/30 = 0{,}43$.

Ответ $w = 0{,}43$.

Дисперсией вариационного ряда называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической.

Для не сгруппированного ряда дисперсия вычисляется по формуле:

$$S_x^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right).$$

Для вариационного ряда формула для дисперсии имеет вид:

$$S_x^2 = \frac{1}{n}\left(\sum_{i=1}^{m}(x_i - \bar{x})^2 \cdot n_i\right) = \sum_{i=1}^{m}(x_i - \bar{x})^2 \cdot w_i.$$

Отметим основные свойства дисперсии.

1. Дисперсия постоянной равна нулю.

$$S_c^2 = 0, \qquad c \text{ – постоянная.}$$

2. Если все варианты увеличить в одно и то же число k раз, то дисперсия увеличится в k^2 раз

$$S_{kx}^2 = k^2 \cdot S_x^2, \qquad k \text{ – постоянная.}$$

3. Если все варианты увеличить на одно и то же число, то дисперсия не изменится

$$S_{x+\text{c}}^2 = S_x^2.$$

4. Дисперсия равна разности между средней арифметической квадратов вариантов и квадратом средней арифметической

$$S_x^2 = \overline{x^2} - \bar{x}^2.$$

5. Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий $\overline{S_i^2}$ (внутригрупповой дисперсии) и межгрупповой дисперсии δ^2:

$$S^2 = {S'}^2 + \delta^2,$$

где:

$$S'^2 = \frac{1}{n}\left(\sum_{i=1}^{l} S_i^2 \cdot n_i\right), \qquad \delta^2 = \frac{1}{n}\left(\sum_{i=1}^{l} (\overline{x_i} - \bar{x})^2 \cdot n_i\right),$$

где $\overline{x_i}$ – групповая средняя i – ой группы, объем которой равен n_i, l – число групп, S_i^2 – групповая дисперсия i – ой группы.

Пример 4.3.7. Для примера 4.2.1 вычислить дисперсию. Используя формулу для расчета средней арифметической простой, найдем среднее значение $\overline{x^2}$:

$$\overline{x^2} = (60^2 + 60^2 + 60^2 + 65^2 + 65^2 + 65^2 + 70^2 + 70^2 + 70^2 + \\ +70^2 + 70^2 + 70^2 + 70^2 + 75^2 + 75^2 + 75^2 + 75^2 + 75^2 + 100^2 + \\ +100^2 + 100^2 + 100^2 + 100^2 + 100^2 + 100^2 + 100^2 + 120^2 + \\ +120^2 + 120^2 + 120^2)/30 = 223500/30 = 7450.$$

Среднее значение вычислено в примере 4.3.3. Оно равно $\bar{x} = 84$.

Дисперсию вычислим по формуле:

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 7450 - 84^2 = 394.$$

Ответ $S_x^2 = 394$

Пример 4.3.8. Для примера 4.2.2 вычислить дисперсию. Используя формулу для расчета средней арифметической взвешенной, найдем среднее значение $\overline{x^2}$:

$$\overline{x^2} = (60^2 \cdot 3 + 65^2 \cdot 3 + 70^2 \cdot 7 + 75^2 \cdot 5 + 100^2 \cdot 8 + 120^2 \cdot 4)/30 = \\ = 223500/30 = 7450.$$

Среднее значение вычислено в примере 4.3.3. Оно равно $\bar{x} = 84$.

Дисперсию вычислим по формуле:

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 7450 - 84^2 = 394.$$

Ответ $S_x^2 = 394$.

Пример 4.3.9. Для примера 4.2.3 вычислить дисперсию. Используя формулу для расчета средней арифметической простой, найдем среднее значение $\overline{x^2}$ для несгруппированного ряда:

$$\overline{x^2} = (155^2 + 170^2 + 185^2 + 180^2 + 188^2 + 152^2 + 173^2 + 178^2 + \\ +178^2 + 168^2 + 185^2 + 173^2 + 170^2 + 183^2 + 175^2 + 173^2 + 170^2 + \\ +183^2 + 175^2 + 180^2 + 175^2 + 193^2 + 178^2 + 183^2 + 180^2 + 197^2 + \\ +178^2 + 181^2 + 187^2 + 168^2 + 174^2 + 179^2 + 184^2 + 183^2 + 178^2 + \\ +180^2 + 178^2 + 163^2 + 166^2 + 178^2 + 175^2 + 182^2 + 190^2 + 167^2 + \\ +170^2 + 178^2 + 183^2 + 170^2 + 178^2 + 181^2 + 173^2 + 168^2 + 185^2 + \\ +175^2 + 170^2 + 155^2 + 169^2 + 186^2 + 179^2 + 189^2 + 155^2 + 174^2 + \\ +179^2 + 179^2 + 169^2 + 186^2 + 174^2 + 171^2 + 184^2 + 175^2 + 193^2 + \\ +178^2 + 184^2 + 180^2 + 196^2 + 175^2 + 181^2 + 188^2 + 168^2 + 179^2 +$$

$$+178^2+183^2+184^2+178^2+181^2+177^2+163^2+166^2+178^2+$$
$$+175^2+183^2+190^2+167^2+170^2+178^2+183^2+170^2+178^2+$$
$$+182^2+173^2+168^2+186^2+176^2+171^2+188^2)/105=$$
$$=3300110/105=31429{,}619.$$

Среднее значение вычислено в примере 4.3.5. Оно равно $\bar{x}=177{,}09$.

Дисперсию вычислим по формуле:

$$S_x^2=\overline{x^2}-\bar{x}^2=31429{,}619-177{,}09^2=68{,}75.$$

Если вычисления проводить по интервальному вариационному ряду, то нужно использовать формулу для средней арифметической взвешенной для вычисления $\overline{x^2}$:

$$\overline{x^2}=\left(\left(\frac{152+157{,}5}{2}\right)^2\cdot 4+\left(\frac{157{,}5+163}{2}\right)^2\cdot 2+\left(\frac{163+168{,}5}{2}\right)^2\cdot 9+\right.$$
$$+\left(\frac{168{,}5+174}{2}\right)^2\cdot 20+\left(\frac{174+179{,}5}{2}\right)^2\cdot 30+\left(\frac{179{,}5+185}{2}\right)^2\cdot 26+$$
$$\left.+\left(\frac{185+190{,}5}{2}\right)^2\cdot 10+\left(\frac{190{,}5+196}{2}\right)^2\cdot 4\right)/105=$$
$$=3283630{,}563/105=31272{,}672.$$

Среднее значение вычислено в примере 4.3.5. Оно равно $\bar{x}=176{,}65$.

Дисперсию вычислим по формуле:

$$S_x^2=\overline{x^2}-\bar{x}^2=31272{,}672-176{,}65^2=67{,}45.$$

Из примера 4.3.9 видно, что при замене первичного ряда ранжированным интервальным рядом несколько изменяется дисперсия.

Среднее квадратическое отклонение – это арифметическое значение корня квадратного из дисперсии:

$$S_x=\sqrt{S_x^2}.$$

Начальный момент v_k k-го порядка вариационного ряда определяется по формуле:

$$v_k=\frac{1}{n}\left(\sum_{i=1}^{n}{x_i}^k\right)\qquad\text{или}\qquad v_k=\frac{1}{n}\left(\sum_{i=1}^{m}{x_i}^k\cdot n_i\right).$$

Центральный момент μ_k k-го порядка вариационного ряда определяется по формуле:

$$\mu_k=\frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\bar{x})^k\right)\qquad\text{или}\qquad \mu_k=\frac{1}{n}\left(\sum_{i=1}^{m}(x_i-\bar{x})^k\cdot n_i\right).$$

Очевидно, что начальные моменты v_k представляют собой средние арифметические значения различных степеней варианты:

$$v_1 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \bar{x}, \quad v_2 = \frac{1}{n}\left(\sum_{i=1}^{n} {x_i}^2\right) = \overline{x^2}, \quad v_3 = \frac{1}{n}\left(\sum_{i=1}^{n} {x_i}^3\right) = \overline{x^3} \text{ и т. д.}$$

Центральный момент первого порядка равен нулю:

$$\mu_1 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})\right) = \bar{x} - \bar{x} = 0.$$

Центральный момент второго порядка равен дисперсии:

$$\mu_2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = S_x^2.$$

Через центральные моменты 3 и 4 порядков определяются такие величины как коэффициент асимметрии и эксцесс вариационного ряда.

Коэффициент асимметрии вариационного ряда определяется по формуле:

$$A = \frac{\mu_3}{S_x^3}.$$

Если $A = 0$, то распределение имеет симметричную форму, т.е. варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту. При $A > 0$ говорят о положительной (правосторонней) асимметрии, а при $A < 0$ – отрицательной (левосторонней) асимметрии.

Эксцессом вариационного ряда называется число:

$$E = \frac{\mu_4}{S_x^4} - 3.$$

4.4. Понятие оценки параметров

Сформулируем задачу оценки параметров в общем виде. Пусть распределение признака X генеральной совокупности задается функцией вероятностей $P(X = x_i, \theta)$ или функцией плотности вероятности $f(x, \theta)$, которая содержит неизвестный параметр θ.

Для вычисления параметра θ исследовать все элементы генеральной совокупности не представляется возможным. Поэтому о параметре θ пытаются судить по выборке, состоящей из значений $x_1, x_2, \ldots, x_n$. Эти значения можно рассматривать как частные значения n независимых случайных величин $X_1, X_2, \ldots, X_n$, каждая из которых имеет тот же закон распределения, что и сама случайная величина X.

Оценкой θ_n *параметра* θ называют всякую функцию результатов наблюдений над случайной величиной X, с помощью которой судят о значении параметра θ: $\theta_n = \theta_n(X_1, X_2, \dots, X_n)$

Поскольку $X_1, X_2, \dots, X_n$ – случайные величины, то и оценка θ_n является случайной величиной, зависящей от закона распределения случайной величины X и числа n.

Рассмотрим наиболее важные свойства оценок.

1.Оценка θ_n параметра θ называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру, т.е.:

$$M(\theta_n) = \theta.$$

В противном случае оценка называется *смещенной*.

Если это равенство не выполняется, то оценка θ_n, полученная по разным выборкам, будет в среднем либо завышать значение параметра θ, либо занижать его. Таким образом, требование несмещённости гарантирует отсутствие систематических ошибок при оценивании.

2. Оценка θ_n параметра θ называется *состоятельной,* если она удовлетворяет закону больших чисел, т.е. сходится по вероятности к оцениваемому параметру:

$$\lim_{n\to\infty} P(|\theta_n - \theta| < \varepsilon) = 1\,.$$

В случае использования состоятельных оценок оправдывается увеличение объема выборки, так как при этом становятся маловероятными значительные ошибки при оценивании. Практический смысл имеют только состоятельные оценки.

Если оценка θ_n параметра θ является несмещенной, а ее дисперсия $D(\theta_n) \to 0$ при $n \to \infty$, то оценка θ_n является и состоятельной. Это вытекает из неравенства Чёбышева:

$$P(|\theta_n - \theta| < \varepsilon) \geq 1 - \frac{D(\theta_n)}{\varepsilon^2}.$$

3.Несмещенная оценка θ_n параметра θ называется *эффективной*, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, вычисленных по выборкам одного и того же объема n.

Эффективность оценки θ_n определяют соотношением:

$$e = \frac{D(\theta_n^{\text{эфективное}})}{D(\theta_n)},$$

где $D(\theta_n^{\text{эфективное}})$, $D(\theta_n)$ – соответственно дисперсии эффективной и данной оценок. Чем ближе e к единице, тем эффективнее оценка.

4.5. Метод моментов для нахождения оценок параметров

Согласно методу моментов определенное количество выборочных моментов (начальных v_k или центральных μ_k или и тех и других) приравнивается к соответствующим теоретическим моментам распределения случайной величины X.

Пример 4.5.1. На предприятии изготавливается определенный вид продукции. Ежемесячный объем выпуска этой продукции является случайной величиной, для характеристики которой принят показательный закон распределения:

$$f(x) = \lambda \cdot e^{-\lambda \cdot x} \qquad x \geq 0.$$

В течение шести месяцев проводился замер объемов выпуска продукции. Получены следующие данные:

Месяц	1	2	3	4	5	6
Объем	20	24	25	28	27	32

Найти оценку параметра λ.

Поскольку теоретический закон распределения содержит только один неизвестный параметр λ, то достаточно составить одно уравнение. Приравняем первые начальные моменты, вычисленные по выборке и по теоретическому распределению

$$v_1^{\text{выборочный}} = v_1^{\text{теоретический}}.$$

Вычислим теоретический начальный момент первого порядка

$$v_1^{\text{теоретический}} = \int_{-\infty}^{+\infty} x\, f(x) dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{+\infty} x\, \lambda \cdot e^{-\lambda \cdot x}\, dx =$$

$$= \lim_{b \to +\infty} \int_0^b x\, \lambda \cdot e^{-\lambda \cdot x}\, dx.$$

Интеграл вычислим по формуле интегрирования по частям. Примем

$$U = x, \quad dU = dx, \quad dV = \lambda \cdot e^{-\lambda \cdot x}\, dx, \quad V = -e^{-\lambda \cdot x}$$

тогда

$$v_1^{\text{теоретический}} = \lim_{b \to +\infty} \int_0^b x\, \lambda \cdot e^{-\lambda \cdot x}\, dx = \lim_{b \to +\infty} \left(-x e^{-\lambda \cdot x} |_0^b + \int_0^b e^{-\lambda \cdot x}\, dx \right) =$$

$$= \lim_{b \to +\infty} \left(-b e^{-\lambda \cdot b} - \frac{e^{-\lambda \cdot x}}{\lambda} |_0^b \right) = \lim_{b \to +\infty} \left(-b e^{-\lambda \cdot b} - \frac{e^{-\lambda \cdot b}}{\lambda} + \frac{1}{\lambda} \right) = \frac{1}{\lambda}.$$

Вычислим выборочный начальный момент первого порядка

$$v_1^{\text{выборочный}}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{6}(20+24+25+28+27+32)=26.$$

Приравнивая моменты получим:

$$\frac{1}{\lambda}=26;\qquad \lambda=\frac{1}{26}\approx 0{,}038.$$

Ответ $\lambda=0{,}038$.

Пример 4.5.2. Случайная величина X распределена по закону Пуассона с неизвестным параметром λ. Статистическое распределение выборки представлено в таблице:

x_i	0	1	2	3	4	5	6	7
n_i	199	169	87	31	9	3	1	1

Найти оценку параметра λ по методу моментов.

В случае распределения Пуассона вероятности определяются согласно выражению:

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

Приравняем первые начальные моменты

$$v_1^{\text{выборочный}}=v_1^{\text{теоретический}}.$$

Найдем первый теоретический начальный момент:

$$v_1^{\text{теоретический}}=\sum_{m=0}^{+\infty}m\cdot\frac{\lambda^m}{m!}e^{-\lambda}=e^{-\lambda}\cdot\left(\lambda+\lambda^2+\frac{\lambda^3}{2!}+\frac{\lambda^4}{3!}+\cdots\right)=$$

$$=e^{-\lambda}\cdot\lambda\left(1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\cdots\right)=e^{-\lambda}\cdot\lambda\cdot e^{\lambda}=\lambda.$$

Найдем первый выборочный начальный момент:

$$v_1^{\text{выборочный}}=\frac{1}{n}\sum_{i=1}^{m}x_i\,n_i=$$

$$=\frac{1}{500}(0\cdot 199+1\cdot 169+2\cdot 87+3\cdot 31+4\cdot 9+5\cdot 3+6\cdot 1+7\cdot 1)=$$

$$=\frac{500}{500}=1.$$

Приравнивая моменты, найдем $\lambda=1$.

Ответ $\lambda=1$.

4.6. Метод наибольшего правдоподобия для нахождения оценок параметров

Метод наибольшего правдоподобия, применяемый для определения точечной оценки, опирается на использование условий экстремума функции одной или нескольких случайных величин. В качестве такой функции принимают функцию правдоподобия.

Для дискретной случайной величины функция правдоподобия принимает вид:

$$L = P(x_1, \theta) \cdot P(x_2, \theta) \cdot \ldots \cdot P(x_n, \theta),$$

где $x_1, x_2, \ldots, x_n$ – варианты выборки, θ – оцениваемый параметр, $P(x_i, \theta)$ – вероятность события $X = x_i$, зависящая от параметра θ.

Для непрерывных случайных величин функция правдоподобия выбирается в виде:

$$L = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta),$$

где $f(x_i, \theta)$ – функция плотности вероятности в точках x_i, зависящая от параметра θ.

Так как функции L и $\ln L$ достигают максимума при одном и том же значении параметра θ, то обычно точки экстремума находят для $\ln L$.

Пример 4.6.1 Случайная величина X распределена по биномиальному закону. Статистическое распределение выборки представлено в таблице:

x_i	0	1	2	3	4	5	6	7
n_i	2	3	10	22	26	20	12	5

Найти точечную оценку вероятности успеха p указанного закона распределения случайной величины (число испытаний $r = 10$) методом наибольшего правдоподобия.

В случае биномиального распределения:

$$P_r(X = x_i, p) = C_r^{x_i} p^{x_i} (1-p)^{r-x_i}.$$

Функция правдоподобия примет вид:

$$L = P_r(x_1, p)^{n_1} \cdot P_r(x_2, p)^{n_2} \cdot \ldots \cdot P_r(x_n, p)^{n_m} =$$
$$= \left(C_r^{x_1} p^{x_1} (1-p)^{r-x_1}\right)^{n_1} \cdot \left(C_r^{x_2} p^{x_2} (1-p)^{r-x_2}\right)^{n_2} \cdot \ldots \cdot$$
$$\cdot \ldots \cdot \left(C_r^{x_m} p^{x_m} (1-p)^{r-x_m}\right)^{n_m}.$$

Найдем логарифм функции правдоподобия:

$$\ln L = n_1 ln\left(C_r^{x_1} p^{x_1} (1-p)^{r-x_1}\right) + n_2 ln\left(C_r^{x_2} p^{x_2} (1-p)^{r-x_2}\right) + \cdot \ldots \cdot$$
$$+ \cdot \ldots \cdot + n_m ln\left(C_r^{x_m} p^{x_m} (1-p)^{r-x_m}\right) =$$
$$= (n_1 x_1 + n_2 x_2 + \cdots + n_m x_m) \ln p +$$
$$+ \Big(\big(n_1(r - x_1) + n_2(r - x_2) + \cdots + n_m(r - x_m)\big) \ln(1-p)\Big) +$$

$$+n_1 \ln C_r^{x_1} + n_2 \ln C_r^{x_2} + \cdots + n_m \ln C_r^{x_m}.$$

Воспользуемся тем, что:

$$n_1 + n_2 + \cdots + n_m = n;$$

$$n_1 x_1 + n_2 x_2 + \cdots + n_m x_m = n \cdot \bar{x}.$$

$$\ln L = n \cdot \bar{x} \ \ln p + \big((n \cdot r - n \cdot \bar{x}) \ln(1-p)\big) +$$

$$+n_1 \ln C_r^{x_1} + n_2 \ln C_r^{x_2} + \cdots + n_m \ln C_r^{x_m}.$$

Найдем производную:

$$\frac{d(\ln L)}{dp} = \frac{n \cdot \bar{x}}{p} - \frac{n \cdot (r - \bar{x})}{1-p}.$$

Приравняем производную нулю:

$$\frac{\bar{x}}{p} - \frac{r - \bar{x}}{1-p} = 0, \ \ \bar{x}(1-p) - (r - \bar{x})p = 0, \quad \bar{x} - \bar{x}\,p - rp + \bar{x}\,p = 0,$$

$$\bar{x} - rp = 0, \quad p = \frac{\bar{x}}{r}.$$

Вычисляя вторую производную, можно убедиться, что данная точка будет являться точкой максимум.

Найдем среднее арифметическое вариационного ряда:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_m x_m}{n} =$$

$$= \frac{0 \cdot 2 + 1 \cdot 3 + 2 \cdot 10 + 3 \cdot 22 + 4 \cdot 26 + 5 \cdot 20 + 6 \cdot 12 + 7 \cdot 5}{100} = \frac{400}{100} = 4.$$

$$p = \frac{\bar{x}}{r} = \frac{4}{10} = 0{,}4.$$

Ответ $p = 0{,}4$.

Пример 4.6.2. Решим задачу из примера 4.5.2 методом наибольшего правдоподобия.

Вероятности событий запишем в виде:

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

Запишем функцию правдоподобия:

$$L = P(x_1, p)^{n_1} \cdot P(x_2, p)^{n_2} \cdot \ldots \cdot P(x_n, p)^{n_m} =$$

$$= \left(\frac{\lambda^{x_1}}{x_1!} e^{-\lambda}\right)^{n_1} \cdot \left(\frac{\lambda^{x_2}}{x_2!} e^{-\lambda}\right)^{n_2} \cdot \ldots \cdot \left(\frac{\lambda^{m}}{x_m!} e^{-\lambda}\right)^{n_m} =$$

$$= \frac{\lambda^{(x_1 n_1 + x_2 n_2 + \cdots + x_m n_m)}}{(x_1!)^{n_1} (x_2!)^{n_2} \ldots (x_m!)^{n_m}} e^{-\lambda(n_1 + n_2 + \cdots + n_m)}.$$

Воспользуемся тем, что:

$$n_1 + n_2 + \cdots + n_m = n;$$

$$n_1 x_1 + n_2 x_2 + \cdots + n_m x_m = n \cdot \bar{x}.$$

$$L = \frac{\lambda^{n\cdot\bar{x}}}{(x_1!)^{n_1}(x_2!)^{n_2}\dots(x_m!)^{n_m}}e^{-\lambda\, n}.$$

Найдем логарифм функции правдоподобия:

$$\ln L = n\,\bar{x}\,\ln\lambda - \lambda\, n\ln e - \ln((x_1!)^{n_1}(x_2!)^{n_2}\dots(x_m!)^{n_m}).$$

Найдем производную от логарифма функции правдоподобия:

$$\frac{d(\ln L)}{d\lambda} = \frac{n\,\bar{x}}{\lambda} - n.$$

Приравняем производную нулю и выразим λ:

$$\frac{n\,\bar{x}}{\lambda} - n = 0, \quad \lambda = \bar{x}.$$

Значение $\bar{x} = 1$ вычислено в примере 4.5.2. Таким образом, метод моментов и наибольшего правдоподобия дают один и тот же результат $\lambda = \bar{x} = 1$.

Ответ $\lambda = \bar{x} = 1$.

Пример 4.6.3. Случайная величина X распределена по показательному закону. Статистическое распределение выборки представлено в таблице:

x_i	5	15	25	35	45	55	65
n_i	365	245	150	100	70	45	25

Найти точечную оценку параметра λ.

Для показательного распределения функция плотности вероятности имеет вид:

$$f(x_i, \lambda) = \lambda\cdot e^{-\lambda\cdot x_i}.$$

Функция правдоподобия:

$$L = f(x_1,\lambda)^{n_1}\cdot f(x_2,\lambda)^{n_2}\cdot\dots\cdot f(x_m,\lambda)^{n_m} =$$
$$= \left(\lambda\cdot e^{-\lambda\cdot x_1}\right)^{n_1}\cdot\left(\lambda\cdot e^{-\lambda\cdot x_2}\right)^{n_2}\cdot\dots\cdot\left(\lambda\cdot e^{-\lambda\cdot x_m}\right)^{n_m} =$$
$$= \lambda^{n_1+n_2+\dots+n_m}\cdot e^{-\lambda\cdot(n_1x_1+n_2x_2+\dots+n_mx_m)}.$$

Воспользуемся равенствами:

$$n_1 + n_2 + \dots + n_m = n;$$
$$n_1x_1 + n_2x_2 + \dots + n_mx_m = n\cdot\bar{x}.$$
$$L = \lambda^n\cdot e^{-\lambda\cdot n\cdot\bar{x}}.$$

Найдем логарифм функции правдоподобия:

$$\ln L = n\ln\lambda - \lambda\cdot n\cdot\bar{x}\ln e.$$

Найдем производную:

$$\frac{d(\ln L)}{d\lambda} = \frac{n}{\lambda} - n\cdot\bar{x}.$$

Приравняем производную нулю и выразим λ:

$$\frac{n}{\lambda} - n \cdot \bar{x} = 0, \quad \lambda = \frac{1}{\bar{x}}.$$

Найдем:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_m x_m}{n} =$$

$$= \frac{5 \cdot 365 + 15 \cdot 245 + 25 \cdot 150 + 35 \cdot 100 + 45 \cdot 70 + 55 \cdot 45 + 65 \cdot 25}{1000} =$$

$$= \frac{20000}{1000} = 20.$$

Вычислим λ:

$$\lambda = \frac{1}{\bar{x}} = \frac{1}{20} = 0{,}05.$$

Ответ $\lambda = 0{,}05$.

4.7. Оценка параметров генеральной совокупности по собственно-случайной выборке

Оценка генеральной доли.

Пусть генеральная совокупность содержит N элементов, из которых M обладает некоторым признаком А. Генеральная доля элементов, обладающим данным свойством будет равна $p = M/N$. В качестве точечной оценки для генеральной доли p можно рассмотреть выборочную долю $w = m/n$. Где m – число элементов выборки, обладающих данным свойством А, n – объем выборки. В частности, имеет место теорема

Теорема. Выборочная доля $w = m/n$ *есть несмещенная и состоятельная оценка генеральной доли* $p = M/N$, *причем ее дисперсия*

$$\sigma_w^2 = \frac{p(1-p)}{n} \text{ – для повторной выборки;}$$

$$\sigma_w^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{p(1-p)}{n}\left(1 - \frac{n}{N}\right) \text{ – для бесповторной выборки.}$$

Оценка генеральной средней.

Пусть из генеральной совокупности объема N отобрана случайная выборка $X_1, X_2, \ldots, X_n$, где X_k – случайная величина, выражающая значение признака у k-го элемента выборки. Тогда для оценки генеральной средней $\overline{x_0}$ можно использовать выборочную среднюю:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Так имеет место теорема.

Теорема. Выборочная средняя $\bar{x}$ есть не смещенная и состоятельная оценка генеральной средней $\overline{x_0}$, причем

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \text{ – для повторной выборки;}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) \text{ – для бесповторной выборки,}$$

где σ^2 – генеральная дисперсия.

Оценка генеральной дисперсии.

Выборочная дисперсия S^2 является смещенной оценкой генеральной дисперсии σ^2. Так для повторной выборки случайные величины $X_1, X_2, \dots, X_n$ можно считать независимыми, а саму дисперсию можно записать в виде:

$$S_x^2 = S_{x-c}^2 = S_{x-\overline{x_0}}^2 = \overline{(x-\overline{x_0})^2} - (\bar{x}-\overline{x_0})^2.$$

Найдем математическое ожидание S_x^2:

$$M(S_x^2) = M\left(\overline{(x-\overline{x_0})^2} - (\bar{x}-\overline{x_0})^2\right) =$$

$$= M\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{x_0})^2\right) - M((\bar{x}-\overline{x_0})^2) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} M(X_i-\overline{x_0})^2\right) - D(\bar{x}) = \frac{1}{n}\left(\sum_{i=1}^{n} D(X_i)\right) - D(\bar{x}) =$$

$$= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Введем исправленную выборочную дисперсию:

$$\hat{S}^2 = \frac{n}{n-1}S^2.$$

Найдем математическое ожидание $\hat{S}^2$:

$$M(\hat{S}^2) = M\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}M(S^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Таким образом, исправленная выборочная дисперсия является несмещенной и состоятельной оценкой генеральной дисперсии σ^2.

Пример 4.7.1. Найти несмещенную оценку генеральной дисперсии случайной величины X на основании данного распределения выборки:

x_i	2	7	9	10
n_i	8	14	10	18

Найдем выборочную среднюю:

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \dots + n_mx_m}{n} = \frac{2\cdot 8 + 7\cdot 14 + 9\cdot 10 + 10\cdot 18}{50} = \frac{384}{50}$$
$$= 7{,}68.$$

Найдем среднее значений x^2:

$$\bar{x} = \frac{n_1 {x_1}^2 + n_2 {x_2}^2 + \cdots + n_m {x_m}^2}{n} = \frac{2^2 \cdot 8 + 7^2 \cdot 14 + 9^2 \cdot 10 + 10^2 \cdot 18}{50} =$$

$$= \frac{3328}{50} = 66{,}56.$$

Выборочная дисперсия:

$$S^2 = \overline{x^2} - \bar{x}^2 = 66{,}56 - 7{,}68^2 = 7{,}5776.$$

Исправленная выборочная дисперсия будет равна:

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{50}{50-1} \cdot 7{,}5776 = 7{,}73.$$

Ответ $\hat{S}^2 = 7{,}73$.

4.8. Интервальное оценивание параметров генеральной совокупности

Выше рассмотрены ситуации, в которых оценивание параметров генеральной совокупности осуществляется одним числом. Генеральное среднее $\overline{x_0}$ оценивается выборочным средним $\bar{x}$, генеральная доля p – выборочной долей w, генеральная дисперсия σ^2 – исправленной выборочной дисперсией $\hat{S}^2$. Такие оценки называются точечными.

Однако точечная оценка θ_n является лишь приближенным значением неизвестного параметра θ генеральной совокупности. В условиях выборки малого объема точечная оценка θ_n может значительно отличаться от оцениваемого параметра θ. Чтобы иметь возможность оценить точность и надежность оценки θ_n параметра θ, используют интервальную оценку параметра.

Интервальной оценкой параметра θ называется числовой интервал $\left(\theta_n^{(1)};\, \theta_n^{(2)}\right)$, который с заданной вероятностью γ накрывает неизвестное значение параметра θ.

Такой интервал $\left(\theta_n^{(1)};\, \theta_n^{(2)}\right)$ называется доверительным, а вероятность γ – доверительной вероятностью, уровнем доверия или надежностью оценки.

Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Величина доверительного интервала существенно зависит от объема выборки n и от значения доверительной вероятности γ.

Очень часто доверительный интервал выбирается симметричным относительно параметра θ, т.е. $(\theta - \Delta, \theta + \Delta)$.

Наибольшее отклонение Δ оценки θ_n от оцениваемого параметра θ, которое возможно с заданной доверительной вероятностью γ, называется предельной ошибкой выборки.

Ошибка Δ является ошибкой репрезентативности (представительства) выборки. Она возникает только вследствие того, что исследуется не вся совокупность, а лишь часть ее (выборка), отобранная случайно. Эту ошибку часто называют случайной ошибкой репрезентативности. Ее не следует путать с систематической ошибкой репрезентативности, появляющейся в результате нарушения принципа случайности при отборе элементов в выборку.

4.8.1. Доверительный интервал для оценки математического ожидания нормального распределения при известной генеральной дисперсии

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить генеральное математическое ожидание $\overline{x_0}$ по выборочной средней $\bar{x}$.

Будем рассматривать выборочную среднюю $\bar{x}$ как случайную величину $\bar{X}$ и выборочные значения признака $x_1, x_2, \dots, x_n$ – как одинаково распределенные независимые случайные величины $X_1, X_2, \dots, X_n$. Математическое ожидание каждой из этих величин равно $\overline{x_0}$ и среднее квадратическое отклонение σ.

Доказано, что если случайная величина X распределена нормально, то выборочная средняя $\bar{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\bar{X}$ таковы:

$$M(\bar{X}) = \overline{x_0}, \qquad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Потребуем, чтобы выполнялось соотношение $P(|\bar{X} - \overline{x_0}| < \Delta) = \gamma$.

Из теории вероятностей известно, что для нормально распределенной величины справедливо равенство:

$$P(|X - a| < \Delta) = 2\phi\left(\frac{\Delta}{\sigma_X}\right),$$

где $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$ – нормированная функция Лапласа.

Подставляя в предыдущее равенство вместо X, $\bar{X}$, а вместо a, $\overline{x_0}$ получим:

$$P(|\bar{X}-\overline{x_0}|<\Delta)=2\phi\left(\frac{\Delta}{\sigma_{\bar{X}}}\right)=2\phi(t)=\gamma,$$

где обозначено $t=\Delta/\sigma_{\bar{X}}$.

Таким образом, из соотношения $2\phi(t)=\gamma$ или $\phi(t)=\gamma/2$ по таблицам нормированной функции Лапласа находится значение t. Ошибка репрезентативности находится из соотношения $\Delta = t\sigma_{\bar{X}}$.

Доверительный интервал строится исходя из неравенства $|\bar{x}-\overline{x_0}|<\Delta$ или $\bar{x}-\Delta<\overline{x_0}<\bar{x}+\Delta$.

Пример 4.8.1.1. Найти доверительный интервал с надежностью $\gamma=0{,}95$ для оценки математического ожидания нормально распределенной случайной величины X, если известны ее среднее квадратичное отклонение $\sigma_x=4$, выборочная средняя $\bar{x}=16$ и объем выборки $n=16$.

По надежности $\gamma=0{,}95$ находим значение функции Лапласа

$$\phi(t)=\frac{\gamma}{2}=\frac{0{,}95}{2}=0{,}475.$$

По приложению 1 находим значение $t=1{,}96$.

Вычисляем $\sigma_{\bar{x}}^2=\frac{\sigma_x^2}{n}=\frac{4^2}{16}=1$. Находим $\sigma_{\bar{x}}=\sqrt{1}=1$.

Ошибка репрезентативности будет равна $\Delta = t\sigma_{\bar{x}}=1{,}96\cdot 1=1{,}96$.

Доверительный интервал

$$\bar{x}-\Delta<\overline{x_0}<\bar{x}+\Delta \text{ или } 16-1{,}96<\overline{x_0}<16+1{,}96,$$
$$14{,}04<\overline{x_0}<17{,}96.$$

Ответ $14{,}04<\overline{x_0}<17{,}96$.

4.8.2. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии

Пусть количественный признак генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ_x не известно. Оценим неизвестное генеральное среднее $\overline{x_0}$ с помощью доверительного интервала. Случайная величина:

$$T=\frac{\bar{X}-\overline{x_0}}{\sigma_{\bar{X}}},$$

где $\sigma_{\bar{X}}^2=\frac{\hat{S}^2}{n}$, $\hat{S}$ – исправленная выборочная дисперсия, имеет распределение Стьюдента с $k=n-1$ степенями свободы.

Плотность распределения Стьюдента:

$$f(t,n)=\frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma\big((n-1)/2\big)}\left(1+\frac{t^2}{n-1}\right)^{-n/2},$$

где $\Gamma(n/2)$ и $\Gamma\big((n-1)/2\big)$ - Гамма функции.

Из выражения $f(t,n)$ видно, что распределение Стьюдента зависит только от объема выборки n (или что тоже самое, от числа степеней свободы $k=n-1$) и не зависит от неизвестных параметров $\overline{x_0}$, σ_x – это является его достоинством.

Поскольку $f(t,n)$ является четной функцией, то можно записать:

$$P\left(\left|\frac{\bar{X}-\overline{x_0}}{\sigma_{\bar{X}}}\right|<t_{\gamma,n}\right)=2\int\limits_0^{t_{\gamma,n}} f(t,n)dt=\gamma.$$

Если ввести обозначение $\Delta = t_{\gamma,n}\cdot\sigma_{\bar{X}}=t_{\gamma,n}\cdot\hat{S}/\sqrt{n}$, тогда из неравенства:

$$\left|\frac{\bar{x}-\overline{x_0}}{\sigma_{\bar{x}}}\right|<t_{\gamma,n},$$

можно найти доверительный интервал в виде $\bar{x}-\Delta<\overline{x_0}<\bar{x}+\Delta$.

Отметим, что величину $t_{\gamma,n}$ можно найти в приложении 2.

Пример 4.8.2.1. Случайная величина X распределена по нормальному закону. Статистическое распределение выборки представлено в таблице

x_i	1	3	5	7	9
n_i	2	5	4	6	3

Найти с надежностью 0,95 доверительный интервал для оценки генерального среднего.

Вычислим среднее значение $\bar{x}$ и $\overline{x^2}$:

$$\bar{x}=\frac{x_1n_1+x_2n_2+\cdots+x_mn_m}{n}=\frac{1\cdot2+3\cdot5+5\cdot4+7\cdot6+9\cdot3}{20}=5{,}3,$$

$$\overline{x^2}=\frac{{x_1}^2n_1+{x_2}^2n_2+\cdots+{x_m}^2n_m}{n}=$$

$$=\frac{1^2\cdot2+3^2\cdot5+5^2\cdot4+7^2\cdot6+9^2\cdot3}{20}=34{,}2.$$

Выборочная дисперсия будет равна:

$$S^2=\overline{x^2}-\bar{x}^2=34{,}2-5{,}3^2=6{,}11.$$

Исправленная дисперсия:

$$\hat{S}^2=\frac{n}{n-1}\cdot S^2=\frac{20}{19}\cdot6{,}11=6{,}43,\quad \hat{S}=\sqrt{6{,}43}=2{,}54.$$

По надежности $\gamma = 0{,}95$ и объему выборки $n = 20$ по приложению 2 найдем $t_{\gamma,n} = 2{,}093$.

Ошибка репрезентативности:

$$\Delta = t_{\gamma,n} \cdot \hat{S}/\sqrt{n} = 2{,}093 \cdot 2{,}54/\sqrt{20} = 1{,}19.$$

Доверительный интервал будет:

$$\bar{x} - \Delta < \overline{x_0} < \bar{x} + \Delta \text{ или } 5{,}3 - 1{,}19 < \overline{x_0} < 5{,}3 + 1{,}19$$
$$\text{или } 4{,}11 < \overline{x_0} < 6{,}49.$$

Ответ $4{,}11 < \overline{x_0} < 6{,}49$.

Заметим, что при $n \to \infty$ распределение Стьюдента переходит в нормальное распределение, т.к.:

$$\lim_{n\to\infty} f(t,n) = \lim_{n\to\infty} \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma\big((n-1)/2\big)}\left(1 + \frac{t^2}{n-1}\right)^{-n/2} = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Поэтому практически при $n > 30$ можно вместо распределения Стьюдента использовать нормальное распределение. Однако, важно помнить, что для малых выборок $(n < 30)$, в особенности для малых значений n, замена распределения нормальным приводит к грубым ошибкам, а именно к неоправданному сужению доверительного интервала, т.е. к повышению точности оценки. Например, если $n = 5$ и $\gamma = 0{,}99$, то пользуясь распределением Стьюдента, найдем $t_{\gamma,n} = 4{,}6$, а используя функцию Лапласа, найдем $t = 2{,}58$, т.е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стьюдента.

То обстоятельство, что распределение Стьюдента при малой выборке дает не вполне определенные результаты (широкий доверительный интервал), вовсе не свидетельствует о слабости метода Стьюдента, а объясняется тем, что малая выборка, разумеется, содержит малую информацию об интересующем нас признаке.

Рассмотрим вопрос об истинном значении измеряемой величины.

Пусть производится n независимых равноточных измерений некоторой физической величины, истинное значение $\overline{x_0}$ которой неизвестно. Будем рассматривать результаты отдельных измерений как случайные величины $X_1, X_2, \dots, X_n$. Эти величины независимы, имеют одно и то же математическое ожидание $\overline{x_0}$ (истинное значение измеряемой величины), одинаковые дисперсии σ^2 (измерения равноточны) и распределены нормально (такое допущение подтверждается опытом). При данных предположениях истинное значение измеряемой величины можно оценить по среднему арифметическому результатов отдельных измерений при помощи доверительных интервалов. Поскольку обычно σ неизвестно, следует пользоваться формулами:

$$\bar{x} - \Delta < \overline{x_0} < \bar{x} + \Delta, \text{ где } \Delta = t_{\gamma,n} \cdot \hat{S}/\sqrt{n} \ ,$$

$t_{\gamma,n}$ – определяется по приложению 2 для распределения Стьюдента.

4.8.3. Доверительные интервалы для оценки среднего квадратического отклонения σ нормального распределения

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по исправленному выборочному среднему квадратическому отклонению $\hat{S}$. Поставим задачу найти доверительные интервалы, покрывающие параметр σ с заданной надежностью γ.

Потребуем, чтобы выполнялось условие:

$$P\left(\left|\sigma - \hat{S}\right| < \Delta\right) = \gamma \text{ или } P\left(\hat{S} - \Delta < \sigma < \hat{S} + \Delta\right) = \gamma.$$

Для того чтобы можно было пользоваться готовой таблицей, преобразуем двойное неравенство $\hat{S} - \Delta < \sigma < \hat{S} + \Delta$ в равносильное неравенство $\hat{S}\left(1 - \Delta/\hat{S}\right) < \sigma < \hat{S}\left(1 + \Delta/\hat{S}\right)$. Обозначив $q = \Delta/\hat{S}$, получим доверительный интервал для генерального среднеквадратичного отклонения:

$$\hat{S}(1 - q) < \sigma < \hat{S}(1 + q).$$

Известно, что величина $\hat{S}^2(n-1)/\sigma^2 = \chi^2$ имеет χ^2 – распределение. Извлекая квадратный корень, найдем, что величина $\hat{S}\sqrt{n-1}/\sigma = \chi$ имеет распределение χ с $n-1$ степенями свободы.

Распределение χ имеет вид:

$$R(\chi, n) = \frac{\chi^{n-2} e^{-\chi^2/2}}{2^{(n-3)/2}\Gamma\left(\frac{n-1}{2}\right)}$$

и не зависит от параметра σ.

Неравенство $\hat{S}(1 - q) < \sigma < \hat{S}(1 + q)$ преобразуем к виду:

$$\frac{1}{\hat{S}(1+q)} < \frac{1}{\sigma} < \frac{1}{\hat{S}(1-q)}; \quad \frac{\sqrt{n-1}}{(1+q)} < \frac{\hat{S}\sqrt{n-1}}{\sigma} < \frac{\sqrt{n-1}}{(1-q)}.$$

Следовательно, величина χ лежит в интервале:

$$\frac{\sqrt{n-1}}{(1+q)} < \chi < \frac{\sqrt{n-1}}{(1-q)}.$$

Значит, вероятность неравенства $P(|\sigma - \hat{S}| < \Delta) = \gamma$ можно записать в виде:

$$\int_{\frac{\sqrt{n-1}}{(1+q)}}^{\frac{\sqrt{n-1}}{(1-q)}} R(\chi, n)d\,\chi = \gamma.$$

Из последнего уравнения можно по заданным значениям n и γ найти величину q. На практике для величины q построены таблицы. Они представлены в приложении 3.

Пример 4.8.3.1. В условиях примера 4.8.2.1 найти 99 % доверительный интервал для генерального среднего квадратического отклонения.

В примере 4.8.2.1 найдено $\hat{S} = 2{,}54$, $n = 20$. По таблице приложения 3 при $n = 20$ и $\gamma = 0{,}99$ найдем $q = 0{,}58$. Доверительный интервал для генерального среднего квадратического отклонения будет

$$\hat{S}(1 - q) < \sigma < \hat{S}(1 + q)$$

или

$$2{,}54(1 - 0{,}58) < \sigma < 2{,}54(1 + 0{,}58)$$

или

$$1{,}07 < \sigma < 4{,}01.$$

Ответ $1{,}07 < \sigma < 4{,}01$.

Глава 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистической гипотезой называется любое предположение о виде или параметрах неизвестного закона распределения.

Различают простую и сложную статистические гипотезы.

Простая гипотеза – это гипотеза, которая содержит только одно предположение. Например, гипотезы «математическое ожидание нормально распределенной случайной величины равно 5» или «вероятность появления события в схеме Бернулли равна 1/3» являются простыми.

Сложная гипотеза – это гипотеза, которая состоит из конечного или бесконечного числа простых гипотез. Например, гипотезы «математическое ожидание нормально распределенной случайной величины больше 5» или «вероятность появления случайного события в схеме Бернулли заключена между 0,2 и 0,4» являются сложными.

Гипотезы о значениях параметров распределения или о сравнительной величине параметров двух распределений называются *параметрическими*, а гипотезы о виде распределения называются *непараметрическими*.

Параметрические гипотезы формулируются на основании свойств тех или иных статистических законов распределения и могут использоваться лишь в том случае, если распределение выборочных данных согласуется с этим законом распределения.

Непараметрические гипотезы применяются в том случае, если закон распределения изучаемых случайных величин неизвестен или их распределение не соответствует никакому из известных законов. В этом случае необходимо проверить, согласуются ли выборочные (эмпирические) данные с каким-либо теоретическим распределением или нет.

Проверяемую гипотезу принято называть *нулевой гипотезой* (или *основной гипотезой*) и обозначать H_0.

Наряду с нулевой гипотезой выдвигается *альтернативная гипотеза* (или *конкурирующая гипотеза)* H_1, являющаяся логическим отрицанием H_0. Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез.

Суть проверки статистической гипотезы заключается в том, что используется специально составленная выборочная характеристика (статистика) $\theta_n(x_1, x_2, \dots, x_n)$, полученная по выборке $x_1, x_2, \dots, x_n$ точное или приближенное распределение которой известно. Затем по этому выборочному распределению определяется критическое значение $\theta_{кр}$ – такое, что если гипотеза H_0 верна, то вероятность $P(\theta_n > \theta_{кр}) = \alpha$ мала. При малых значения α в соответствии с принципом практической уверенности в условиях данного исследования событие $\theta_n > \theta_{кр}$ можно считать практически невозможным. Если в данном конкретном случае обнаруживается отклонение $\theta_n > \theta_{кр}$, то гипотеза H_0 отвергается, в то время при появлении значения $\theta_n \leq \theta_{кр}$ гипотеза H_0 принимается.

Правило, по которому гипотеза H_0 отвергается или принимается, называется *статистическим критерием или статистическим тестом*.

Множество возможных значений статистики критерия θ_n разбивается на два непересекающихся подмножества: *критическую область* ω (область отклонения гипотезы H_0) и *область допустимых значений* (область принятия гипотезы H_0). Если фактически наблюдаемое значение статистики критерия θ_n попадает в критическую область, то гипотезу H_0 отвергают.

При проверке статистических гипотез можно совершить два вида ошибок.

Ошибка первого рода допускается, если в действительности гипотеза H_0 верна, а на основании выборочных данных принято решение ее отвергнуть.

Если в действительности верна альтернативная гипотеза H_1, а принято решение принять гипотезу H_0 то говорят, что допущена *ошибка второго рода.*

Все возможные случаи принятия решения сведены в таблицу:

Гипотеза H_0	Принимается	Отвергается
Верна	Правильное решение $1-\alpha = P(H_0 \mid H_0)$	**Ошибка 1-го рода** $\alpha = P(H_1 \mid H_0)$
Неверна	**Ошибка 2-го рода** $\beta = P(H_0 \mid H_1)$	Правильное решение $1-\beta = P(H_1 \mid H_1)$

Вероятность α ошибки 1-го рода, т.е. вероятность отвергнуть гипотезу H_0 при условии, что она верна, называется *уровнем значимости критерия*.

Следует устанавливать разумные значения вероятностей этих ошибок. На практике, обычно, рассматривают значения 0,005, 0,01, 0,05 и 0,1. При этом следует отметить, что нередко возникают ситуации, когда одна и та же гипотеза на одном уровне значимости принимается, а на другом отвергается.

Вероятность противоположная вероятности α, называется *уровнем доверия*, или *доверительной вероятностью*. Доверительная вероятность – вероятность принять нулевую гипотезу при условии, что она верна, т.е. $\gamma = P(H_0|H_0)$.

Вероятность ошибки 2-го рода т.е. вероятность принять гипотезу H_0 при условии, что она не верна, обозначается β.

Вероятность $1-\beta$ не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0, когда она верна, называется *мощностью критерия*.

Естественное желание состоит в том, чтобы вероятность совершить ошибку первого рода была минимальной. При этом значение мощность критерия должна быть как можно больше. Однако, это противоречивые требования. При фиксированном объеме выборки можно сделать сколь угодно малой лишь одну из вероятностей – α или β.

Уменьшение одной из них сопряжено с неизбежным увеличением другой. Одновременное уменьшение вероятностей α или β возможно только лишь при изменении объема выборки в сторону его увеличения.

Возможны три варианта расположения критической области: двусторонняя, правосторонняя, левосторонняя.

Двусторонняя критическая область $\omega = (-\infty, \theta_{\text{кр.л}}) \cup (\theta_{\text{кр.п}}, +\infty)$, где $\theta_{\text{кр.л}}$ и $\theta_{\text{кр.п}}$ определяются из условий:

$$P(\theta < \theta_{\text{кр.л}}) = \frac{\alpha}{2}, \quad P(\theta > \theta_{\text{кр.п}}) = \frac{\alpha}{2}.$$

Двусторонняя критическая область изображена на рисунке 8.

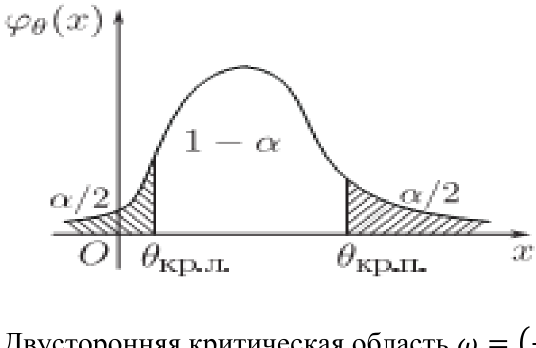

Рисунок 8. Двусторонняя критическая область $\omega = (-\infty, \theta_{\text{кр.л}}) \cup (\theta_{\text{кр.п}}, +\infty)$

Критическая область $\omega = (\theta_{\text{кр}}, +\infty)$, где $\theta_{\text{кр}}$ определяется из условия $P(\theta > \theta_{\text{кр}}) = \alpha$.

Левосторонняя критическая область $\omega = (-\infty, \theta_{\text{кр}})$, где $\theta_{\text{кр}}$ определяется из условия $P(\theta < \theta_{\text{кр}}) = \alpha$. Правосторонняя и левостороняя критические области изображены на рисунке 9.

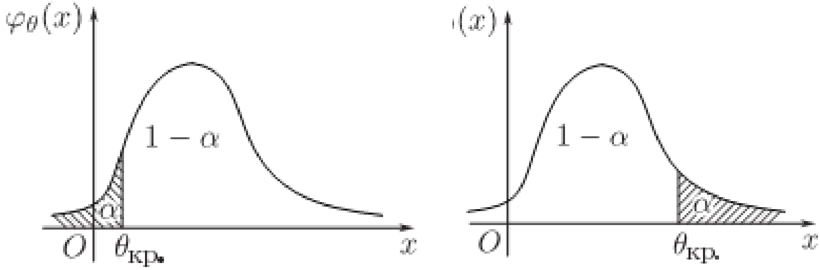

Рисунок 9. Правосторонняя (справа) и левосторонняя (слева) критические области.

5.1. Общая схема проверки гипотез

1. В зависимости от задачи исследования формулируются статистические – нулевая гипотеза H_0 и альтернативная ей гипотеза H_1.

2. Выбирается статистическая характеристика гипотезы — статистика критерия – случайная величина θ, распределение которой при справедливости нулевой гипотезы известно.

3. Задается уровень значимости α и рассчитываются границы *критической области* ω.

4. Вычисляется эмпирическое или наблюдаемое значение статистики $\theta_{\text{наб}}$.

5. Сравнивается критическое и наблюдаемое значения статистики и делается вывод. Если эмпирическое значение статистики $\theta_{\text{наб}}$ (вычисленное по конкретной выборке) попадает в критическую область ω, то нулевая гипотеза отвергается и принимается альтернативная гипотеза; в противном случае нулевая гипотеза не отвергается.

5.2. Сравнение двух дисперсий нормальных генеральных совокупностей в случае независимых выборок

Гипотезы о дисперсиях возникают довольно часто, так как дисперсия характеризует такие показатели, как точность машин, приборов, технологических процессов, степень однородности совокупностей, риск, связанный с отклонением доходности активов от ожидаемого уровня.

Пусть имеются две нормально распределенные совокупности, генеральные дисперсии которых равны σ_x^2 и σ_y^2. Сформулируем нулевую гипотезу о равенстве генеральных дисперсий $H_0\colon \sigma_x^2 = \sigma_y^2$.

Для проверки данной гипотезы из данных совокупностей взяты две независимые выборки объемов n_1 и n_2. Точечными несмещенными оценками генеральных дисперсий являются исправленные выборочные дисперсии $\hat{S}_x^2$ и $\hat{S}_y^2$. Следовательно, задача сводится к проверке равенства дисперсий $\hat{S}_x^2$ и $\hat{S}_y^2$.

При справедливости нулевой гипотезы $H_0\colon \sigma_x^2 = \sigma_y^2 = \sigma^2$ в качестве оценки σ^2 можно взять те же дисперсии $\hat{S}_x^2$ и $\hat{S}_y^2$ рассчитанные по элементам первой и второй выборок. При этом выборочные характеристики:

$$\frac{(n_1 - 1)\hat{S}_x^2}{\sigma^2} \quad \text{и} \quad \frac{(n_2 - 1)\hat{S}_y^2}{\sigma^2}$$

имеют распределения χ^2 с числом степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ соответственно. Отношение же данных величин имеет F – распределение Фишера-Снедекора.

Следовательно, случайная величина F, определяемая отношением:

$$F = \frac{\dfrac{1}{n_1 - 1} \cdot \left(\dfrac{(n_1 - 1)\hat{S}_x^2}{\sigma^2}\right)}{\dfrac{1}{n_2 - 1} \cdot \left(\dfrac{(n_2 - 1)\hat{S}_y^2}{\sigma^2}\right)} = \frac{\hat{S}_x^2}{\hat{S}_y^2}$$

имеет F – распределение Фишера-Снедекора с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы.

Отметим, что при вычислении критерия F обычно большую дисперсию делят на меньшую.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0\colon \sigma_x^2 = \sigma_y^2$. Конкурирующая гипотеза $H_1\colon \sigma_x^2 \neq \sigma_y^2$. В этом случае строят двустороннюю критическую область $\omega = (0; F_{\text{кр.л}}) \cup (F_{\text{кр.п}}; +\infty)$, исходя из требования, чтобы вероятность попадания в критическую область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α. Наибольшая мощность критерия достигается в том случае, когда вероятность попадания критерия в оба интервала одинакова и равна $\alpha/2$. Таким образом, для поиска $F_{\text{кр.л}}$ и $F_{\text{кр.п}}$ можно записать:

$$P\left(0 < F < F_{\text{кр.л}}\right) = \frac{\alpha}{2}; \qquad P\left(F > F_{\text{кр.п}}\right) = \frac{\alpha}{2}.$$

Величину $F_{\text{кр.п}} = F_{\text{кр.п}}(k_1; k_2; \alpha/2)$, которая является решение уравнения $P\left(F > F_{\text{кр.п}}\right) = \alpha/2$, находят по таблице приложения 4. Величину же $F_{\text{кр.л}} = F_{\text{кр.л}}(k_1; k_2; 1 - \alpha/2)$ можно найти из соотношения справедливого для F – критерия Фишера-Снедекора:

$$F_{\text{кр.л}}(k_1; k_2; 1 - \alpha/2) = \frac{1}{F_{\text{кр.п}}(k_2; k_1; \alpha/2)}.$$

Пример 5.2.1. По двум независимым выборкам, объемы которых $n_1 = 6$ и $n_2 = 9$, извлеченным из нормальных генеральных совокупностей X и Y, найдены выборочные дисперсии $S_x^2 = 20{,}5$ и $S_y^2 = 14{,}4$. При уровне значимости $\alpha = 0{,}1$ проверить нулевую гипотезу $H_0: \sigma_x^2 = \sigma_y^2$ и конкурирующей гипотезе $H_1: \sigma_x^2 \neq \sigma_y^2$.

Найдем исправленные дисперсии:

$$\hat{S}_x^2 = \frac{n_1}{n_1 - 1} S_x^2 = \frac{6}{5} \cdot 20{,}5 = 24{,}6; \qquad \hat{S}_y^2 = \frac{n_2}{n_2 - 1} S_y^2 = \frac{9}{8} \cdot 14{,}4 = 16{,}2.$$

Найдем наблюдаемое значение критерия Фишера-Снедекора:

$$F_{\text{наб}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} = \frac{24{,}6}{16{,}2} = 1{,}52.$$

Числа степеней свободы $k_1 = n_1 - 1 = 6 - 1 = 5$ и $k_2 = n_2 - 1 = 9 - 1 = 8$.

По таблице приложения 4 найдем:

$$F_{\text{кр.п}} = F_{\text{кр.п}}(k_1 = 5; k_2 = 8; \alpha/2 = 0{,}05) = 3{,}69.$$

Вычислим левую границу критической области:

$$F_{\text{кр.л}}(k_1 = 5; k_2 = 8; 1 - \alpha/2 = 0{,}95) = \frac{1}{F_{\text{кр.п}}(k_2 = 8; k_1 = 5; \alpha/2 = 0{,}05)};$$

$$F_{\text{кр.л}}(k_1 = 5; k_2 = 8; 1 - \alpha/2 = 0{,}95) = \frac{1}{4{,}82} = 0{,}21.$$

Критическая область имеет вид: $\omega = (0; 0{,}21) \cup (3{,}69; +\infty)$.

Поскольку $F_{\text{наб}} \notin$ ω, гипотезу H_0 принимаем.

Ответ гипотезу H_0 принимаем.

Второй случай. Нулевая гипотеза $H_0: \sigma_x^2 = \sigma_y^2$. Конкурирующая гипотеза $H_1: \sigma_x^2 > \sigma_y^2$. В этом случае строят одностороннюю критическую область $\omega = \left(F_{\text{кр.п}}; +\infty\right)$, исходя из требования, чтобы вероятность попадания в критическую область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α. Таким образом, для поиска $F_{\text{кр.п}}$ можно записать:

$$P\left(F > F_{\text{кр.п}}\right) = \alpha.$$

Величину $F_{\text{кр.п}} = F_{\text{кр.п}}(k_1; k_2; \alpha)$, которая является решение уравнения $P\left(F > F_{\text{кр.п}}\right) = \alpha$, находят по таблице приложения 4.

Пример 5.2.2. По двум независимым выборкам, объемы которых $n_1 = 9$ и $n_2 = 16$, извлеченным из нормальных генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии $\hat{S}_x^2 = 34{,}02$ и $\hat{S}_y^2 = 12{,}15$. При уровне значимости $\alpha = 0{,}01$, проверить нулевую гипотезу H_0: $\sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе H_1: $\sigma_x^2 > \sigma_y^2$.

Найдем наблюдаемое значение критерия Фишера-Снедекора:

$$F_{\text{наб}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} = \frac{34{,}02}{12{,}15} = 2{,}8,$$

$k_1 = n_1 - 1 = 9 - 1 = 8$ и $k_2 = n_2 - 1 = 16 - 1 = 15$.

По таблице приложения 4 найдем:

$$F_{\text{кр.п}} = F_{\text{кр.п}}(k_1 = 8; k_2 = 15; \alpha = 0{,}01) = 4{,}00.$$

Критическая область имеет вид $\omega = (4{,}00; +\infty)$.

Поскольку $F_{\text{наб}} \notin \omega$, гипотезу H_0 принимаем.

Ответ гипотезу H_0 принимаем.

5.3. Сравнение генеральной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

Пусть генеральная совокупность распределена нормально, причем генеральная дисперсия хотя и неизвестна, но имеются основания предполагать, что она равна гипотетическому (предполагаемому) значению σ_0^2.

Пусть из генеральной совокупности извлечена выборка объема n и по ней найдена исправленная выборочная дисперсия $\hat{S}^2$ с $k = n - 1$ степенями свободы. Требуется по исправленной дисперсии при заданном уровне значимости проверить нулевую гипотезу, состоящую в том, что генеральная дисперсия рассматриваемой совокупности равна гипотетическому значению σ_0^2.

Учитывая, что $\hat{S}^2$ является несмещенной оценкой генеральной дисперсии, нулевую гипотезу можно записать так H_0: $M\left(\hat{S}^2\right) = \sigma_0^2$. Следовательно, требуется установить, значимо или не значимо различаются исправленная выборочная и гипотетическая генеральная дисперсии.

На практике такая гипотеза проверяется, если нужно проверить точность приборов, инструментов, станков, методов исследования и устойчивость технологических процессов. Например, если известна допустимая характеристика рассеяния контролируемого размера деталей, изготовленных на станке, равна σ_0^2, а найденная по выборке окажется значимо больше σ_0^2, то станок требует наладки.

В качестве критерия проверки нулевой гипотезы выбираю величину:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma_0^2},$$

имеющую χ^2 – квадрат распределение с числом степеней свободы $k = n-1$.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза H_0: $\sigma^2 = \sigma_0^2$. Конкурирующая гипотеза H_1: $\sigma^2 > \sigma_0^2$. В данном случае строится правосторонняя критическая область $\omega = \left(\chi^2_{\text{кр.п}}; +\infty\right)$, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left(\chi^2 > \chi^2_{\text{кр.п}}\right) = \alpha.$$

Критическую точку $\chi^2_{\text{кр.п}} = \chi^2_{\text{кр.п}}(k;\alpha)$ находят по таблице приложения 5.

Пример 5.3.1. Из нормальной генеральной совокупности извлечена выборка объема $n = 17$ и по ней найдена исправленная выборочная дисперсия $\hat{S}^2 = 0{,}24$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу H_0: $\sigma^2 = \sigma_0^2 = 0{,}18$, приняв в качестве конкурирующей гипотезы H_1: $\sigma^2 > 0{,}18$.

Вычислим наблюдаемое значения критерия χ^2:

$$\chi^2_{\text{набл}} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(17-1)0{,}24}{0{,}18} = 21{,}33.$$

По таблице приложения 5 при $k = n - 1 = 17 - 1 = 16$ и $\alpha = 0{,}05$ найдем $\chi^2_{\text{кр.п}}(k = 16; \alpha = 0{,}05) = 26{,}3$. Критическая область $\omega = (26{,}3; +\infty)$. Поскольку $\chi^2_{\text{набл}} \notin \omega$, то нет оснований отвергать гипотезу H_0.

Ответ гипотезу H_0 принимаем.

Второй случай. Нулевая гипотеза H_0: $\sigma^2 = \sigma_0^2$. Конкурирующая гипотеза H_1: $\sigma^2 \neq \sigma_0^2$. В данном случае строится двусторонняя критическая область $\omega = (0; \chi^2_{\text{кр.л}}) \cup (\chi^2_{\text{кр.п}}; +\infty)$, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α. Границы критической области $\chi^2_{\text{кр.л}}$ и $\chi^2_{\text{кр.п}}$ строят исходя из того, чтобы вероятность попадания критерия в каждый из двух интервалов критической области была равна $\alpha/2$:

$$P(0 < \chi^2 < \chi^2_{\text{кр.л}}) = \alpha/2; \qquad P(\chi^2 > \chi^2_{\text{кр.п}}) = \alpha/2.$$

Правую критическую точку $\chi^2_{\text{кр.п}} = \chi^2_{\text{кр.п}}(k; \alpha/2)$ находят по таблице приложения 5. Левую критическую точку находят из следующих рассуждений. Так как:

$$P(0 < \chi^2 < \chi^2_{\text{кр.л}}) + P(\chi^2 > \chi^2_{\text{кр.л}}) = 1,$$

то:

$$P(\chi^2 > \chi^2_{\text{кр.л}}) = 1 - P(0 < \chi^2 < \chi^2_{\text{кр.л}}) = 1 - \frac{\alpha}{2}.$$

Значит левую критическую точку $\chi^2_{\text{кр.л}} = \chi^2_{\text{кр.л}}(k; 1 - \alpha/2)$ можно найти по приложению 5.

Пример 5.3.2. Из нормальной генеральной совокупности извлечена выборка объема $n = 21$ и по ней найдена исправленная выборочная дисперсия $\hat{S}^2 = 16{,}2$. Требуется при уровне значимости $\alpha = 0{,}02$ проверить нулевую гипотезу H_0: $\sigma^2 = \sigma_0^2 = 15$, приняв в качестве конкурирующей гипотезы H_1: $\sigma^2 \neq 15$.

Наблюдаемое значения критерия χ^2:

$$\chi^2_{\text{набл}} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(21-1)16{,}2}{15} = 21{,}6.$$

По таблице приложения 5 при $k = n - 1 = 21 - 1 = 20$ и $\alpha/2 = 0{,}01$ найдем: $\chi^2_{\text{кр.п}}(k = 20; \alpha/2 = 0{,}01) = 37{,}6$,

$$\chi^2_{\text{кр.л}} = \chi^2_{\text{кр.л}}(k = 20; 1 - \alpha/2 = 0{,}99) = 8{,}26.$$

Критическая область $\omega = (0; 8{,}26) \cup (37{,}6; +\infty)$. Поскольку $\chi^2_{\text{набл}} \notin \omega$, то нет оснований отвергать гипотезу H_0.

Ответ гипотезу H_0 принимаем.

Третий случай. Нулевая гипотеза H_0: $\sigma^2 = \sigma_0^2$. Конкурирующая гипотеза H_1: $\sigma^2 < \sigma_0^2$. В данном случае строится левосторонняя критическая область $\omega = (0; \chi^2_{\text{кр.л}})$, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left(0 < \chi^2 < \chi^2_{\text{кр.л}}\right) = \alpha.$$

Критическую точку $\chi^2_{\text{кр.л}} = \chi^2_{\text{кр.л}}(k; 1-\alpha)$ находят по таблице приложения 5.

Пример 5.3.3. Из нормальной генеральной совокупности извлечена выборка объема $n = 13$ и по ней найдена исправленная выборочная дисперсия $\hat{S}^2 = 10{,}3$. Требуется при уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу H_0: $\sigma^2 = \sigma_0^2 = 12$, приняв в качестве конкурирующей гипотезы H_1: $\sigma^2 < 12$.

Наблюдаемое значения критерия χ^2:

$$\chi^2_{\text{набл}} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(13-1)10{,}3}{12} = 10{,}3.$$

По таблице приложения 5 при $k = n - 1 = 13 - 1 = 12$ и $1 - \alpha = 0{,}99$ найдем $\chi^2_{\text{кр.п}}(k = 12; 1 - \alpha = 0{,}99) = 3{,}57$. Критическая область $\omega = (0; 3{,}57)$. Поскольку $\chi^2_{\text{набл}} \notin \omega$, то нет оснований отвергать гипотезу H_0.

Ответ гипотезу H_0 принимаем.

5.4. Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам одинакового объема. Критерий Кохрена

Пусть генеральные совокупности $X_1, X_2, \ldots, X_l$ распределены нормально. Из этих совокупностей извлечено l независимых выборок одинакового объема n и по ним найдены исправленные выборочные дисперсии $\hat{S}_1^2,\ \hat{S}_2^2,\ \ldots, \hat{S}_l^2$ все с одинаковым числом степеней свободы $k = n - 1$.

Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_l^2.$$

Фактически нужно проверить значимо или незначимо различаются исправленные выборочные дисперсии.

В качестве критерия проверки нулевой гипотезы примем критерий Кохрена – отношение максимальной исправленной дисперсии к сумме всех исправленных дисперсий:

$$G=\frac{\hat{S}^2_{max}}{\hat{S}^2_1+\hat{S}^2_2+\cdots+\hat{S}^2_l}.$$

Распределение этой случайной величины зависит только от числа степеней свободы $k=n-1$ и количества выборок l.

Критическую область строят правостороннюю $\omega=\left(G_{\text{кр}};\ +\infty\right)$, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости:

$$P\left(G>G_{\text{кр}}\right)=\alpha.$$

Критическое значение критерия Кохрена $G_{\text{кр}}=G_{\text{кр}}(k;l;\alpha)$ находят по таблице приложения 6.

Пример 5.4.1 По четырем независимым выборкам одинакового объема $n=37$, извлеченным из нормальных генеральных совокупностей, найдены исправленные дисперсии: 3,0; 3,2; 3,8; 4,2. Требуется при уровне значимости 0,05 проверить нулевую гипотезу об однородности генеральных дисперсий.

Найдем наблюдаемое значение критерия Кохрена:

$$G_{\text{набл}}=\frac{\hat{S}^2_{max}}{\hat{S}^2_1+\hat{S}^2_2+\cdots+\hat{S}^2_l}=\frac{4{,}2}{3{,}0+3{,}2+3{,}8+4{,}2}=0{,}2958.$$

Число степеней свободы $k=n-1=37-1=36$. По таблице приложения 6 найдем $G_{\text{кр}}(k=36;\ l=4;\ \alpha=0{,}05)\ =0{,}3720$. Критическая область $\omega=(0{,}372;+\infty)$. Поскольку наблюдаемое значение критерия $G_{\text{набл}}\notin\omega$, то гипотезу о равенстве дисперсий принимаем. Оценкой же генеральной дисперсии в данном случае может быть средняя арифметическая исправленных выборочных дисперсий $\sigma^2=(3{,}0+3{,}2+3{,}8+4{,}2)/4=3{,}55$.

Ответ генеральные дисперсии однородны.

5.5. Сравнение двух средних нормальных генеральных совокупностей с известными дисперсиями в случае независимых выборок

Пусть в генеральной совокупности случайные величины X и Y распределены нормально, причем их дисперсии известны. Дисперсии могут быть известны из предыдущего опыта или найдены теоретически. По независимым выборкам, объемы которых n_1 и n_2, извлеченных из этих совокупностей, найдены выборочные средние $\bar{x}$ и $\bar{y}$.

Требуется по выборочным средним при заданном уровне значимости α проверить нулевую гипотезу о равенстве генеральных средних $\bar{x}_0 = M(X)$ и $\bar{y}_0 = M(Y)$ т.е.:

$$H_0:\quad \bar{x}_0 = \bar{y}_0.$$

Учитывая, что выборочные средние являются несмещенными оценками генеральных средних, т.е. $M(\bar{x}) = \bar{x}_0$ и $M(\bar{y}) = \bar{y}_0$, нулевую гипотезу можно записать так:

$$H_0:\quad M(\bar{x}) = M(\bar{y}).$$

Следовательно, требуется проверить, что математические ожидания выборочных средних равны между собой. Данная задача ставится, поскольку на практике, выборочные средние оказываются численно не равны $\bar{x} \neq \bar{y}$. Встает вопрос: значимо или не значимо отличие выборочных средних?

Если окажется, что гипотеза H_0 принимается, то различие в выборочных средних связано со случайным характером величин X и Y. В данном случае его можно объяснить статистической погрешностью, т.е. случайным отбором объектов в выборки.

В качестве критерия проверки гипотезы выберем величину:

$$Z = \frac{\bar{x} - \bar{y}}{\sigma(\bar{x} - \bar{y})}.$$

Используя свойства дисперсии, найдем:

$$\sigma(\bar{x} - \bar{y}) = \sqrt{D(\bar{x} - \bar{y})} = \sqrt{D(\bar{x}) + D(\bar{y})} = \sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}.$$

Таким образом:

$$Z = \frac{\bar{x} - \bar{y}}{\sigma(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n_1 + D(Y)/n_2}}.$$

Случайная величина Z является нормированной нормальной случайной величиной и при справедливости нулевой гипотезы H_0 имеет $M(Z) = 0$ и $\sigma(Z) = 1$.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Если конкурирующая гипотеза $H_1 \colon \bar{x}_0 \neq \bar{y}_0$, то строится двухсторонняя критическая область $\omega = \left(-\infty; z_{\text{кр.л}}\right) \cup \left(z_{\text{кр.п}}; +\infty\right)$. При этом считают, что вероятность попасть в оба этих интервала равна $\alpha/2$. Т.е.:

$$P\left(Z < z_{\text{кр.л}}\right) = \frac{\alpha}{2}; \qquad P\left(Z > z_{\text{кр.п}}\right) = \frac{\alpha}{2}.$$

Так как Z нормально распределенная случайная величина с нулевым математическим ожиданием, то ее распределение симметрично относительно нуля. Значит $z_{\text{кр.л}} = -z_{\text{кр.п}}$.

Для случайной величины Z будет справедливо равенство:

$$P(Z > 0) = P\left(0 < Z < z_{\text{кр.п}}\right) + P\left(Z > z_{\text{кр.п}}\right) = \frac{1}{2}.$$

Откуда:

$$\phi\left(z_{\text{кр.п}}\right) + \frac{\alpha}{2} = \frac{1}{2}; \qquad \phi\left(z_{\text{кр.п}}\right) = \frac{1-\alpha}{2}.$$

Пример 5.5.1 По двум независимым выборкам, объемы которых $n_1 = 15$ и $n_2 = 20$, извлеченных из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 135$ и $\bar{y} = 125$. Генеральные дисперсии известны $D(x) = 80$ и $D(y) = 60$. Требуется при уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу H_0: $\bar{x}_0 = \bar{y}_0$ при конкурирующей гипотезе $H_1: \bar{x}_0 \neq \bar{y}_0$.

Наблюдаемое значение критерия:

$$Z_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n_1 + D(Y)/n_2}} = \frac{135 - 125}{\sqrt{80/15 + 60/20}} = 3{,}46.$$

Вычислим $\phi\left(z_{\text{кр.п}}\right) = (1-\alpha)/2 = 0{,}495$. По таблице из приложения 1 найдем $z_{\text{кр.п}} = 2{,}58$. Критическая область $\omega = (-\infty; -2{,}58) \cup (2{,}58; +\infty)$. Поскольку $Z_{\text{набл}} \in \omega$ гипотезу H_0 отвергаем.

Ответ нулевую гипотезу H_0 отвергаем.

Второй случай. Если конкурирующая гипотеза имеет вид $H_1: \bar{x}_0 > \bar{y}_0$, то строится односторонняя критическая область $\omega = \left(z_{\text{кр.п}}; +\infty\right)$. При этом считают, что вероятность попасть в данный интервал равна α. Т.е.:

$$P\left(Z > z_{\text{кр.п}}\right) = \alpha.$$

Следовательно, из равенства:

$$P(Z > 0) = P\left(0 < Z < z_{\text{кр.п}}\right) + P\left(Z > z_{\text{кр.п}}\right) = \frac{1}{2};$$

найдем:

$$\phi\left(z_{\text{кр.п}}\right) + \alpha = \frac{1}{2}; \qquad \phi\left(z_{\text{кр.п}}\right) = \frac{1-2\alpha}{2}.$$

Пример 5.5.2 По двум независимым выборкам, объемы которых соответственно равны $n_1 = 10$ и $n_2 = 15$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 14$ и $\bar{y} = 13$. Генеральные дисперсии известны: $D(x) = 20$ и $D(y) = 18$. При уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу вид $H_0: \bar{x}_0 = \bar{y}_0$, при конкурирующей гипотезе $H_1: \bar{x}_0 > \bar{y}_0$.

Наблюдаемое значение критерия:

$$Z_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n_1 + D(Y)/n_2}} = \frac{14 - 13}{\sqrt{20/10 + 18/15}} = 0{,}56.$$

Вычислим: $\phi\left(z_{\text{кр.п}}\right) = (1 - 2\alpha)/2 = 0{,}45.$ По таблице из приложения 1 найдем $z_{\text{кр.п}} = 1{,}65$. Критическая область $\omega = (1{,}65; +\infty)$. Поскольку $Z_{\text{набл}} \notin \omega$ гипотезу H_0 принимаем.

Ответ нулевую гипотезу H_0 принимаем.

5.6. Сравнение двух средних произвольно распределенных генеральных совокупностей при больших независимых выборках

Пусть в генеральной совокупности случайные величины X и Y имеют произвольный закон распределения, а дисперсии их неизвестны. Если обе выборки имеют большой объем $(n_1 \geq 30, n_2 \geq 30)$, тогда выборочные средние $\bar{x}$ и $\bar{y}$ распределены приближенно нормально, а их дисперсии можно заменить несмещенными точечными оценками – исправленными дисперсиями:

$$D(X) = \hat{S}_x^2 = \frac{n_1}{n_1 - 1} S_x^2; \qquad D(Y) = \hat{S}_y^2 = \frac{n_2}{n_2 - 1} S_y^2.$$

Таким образом:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n_1 + D(Y)/n_2}} = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2/(n_1 - 1) + S_y^2/(n_2 - 1)}}.$$

В остальном же проверка гипотезы проводится так же, как описано в предыдущем разделе.

Пример 5.6.1 По выборке объема $n_1 = 30$ найден средний вес $\bar{x} = 130$ г изделий, изготовленных на первом станке. По выборке объема $n_2 = 40$ найден средний вес $\bar{y} = 125$ г изделий, изготовленных на втором станке. Выборочные дисперсии равны $S_x^2 = 60\ \text{г}^2$, $S_y^2 = 80\ \text{г}^2$. Выборки независимы. Требуется при уровне значимости $\alpha = 0{,}1$, проверить нулевую гипотезу H_0: $\bar{x}_0 = \bar{y}_0$ при конкурирующей гипотезе H_1: $\bar{x}_0 \neq \bar{y}_0$.

Наблюдаемое значение критерия:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2/(n_1 - 1) + S_y^2/(n_2 - 1)}} = \frac{130 - 125}{\sqrt{60/(30 - 1) + 80/(40 - 1)}} = 2{,}46.$$

Так как конкурирующая гипотеза имеет вид H_1: $\bar{x}_0 \neq \bar{y}_0$ имеем первый случай. Вычисляем:

$$\phi(z_{\text{кр.п}}) = \frac{1-\alpha}{2} = \frac{1-0{,}1}{2} = 0{,}45.$$

По приложению 1 найдем $z_{\text{кр.п}} = 1{,}65$. Следовательно, критическая область имеет вид $\omega = (-\infty; -1{,}65) \cup (1{,}65; +\infty)$. Так как $Z_{\text{набл}} \in \omega$ нулевую гипотезу отвергаем.

Ответ нулевую гипотезу H_0 отвергаем.

5.7. Сравнение двух средних нормальных генеральных совокупностей с неизвестными, но одинаковыми дисперсиями в случае независимых выборок

Пусть генеральные совокупности X и Y распределены нормально, а дисперсии их неизвестны, но одинаковы. Данная ситуация довольно часто встречается на практике. Пусть, например, сравниваются средние размеры деталей из двух партий, выпущенных на одном и том же станке. В этом случае логично предположить, что дисперсии контролируемых размеров одинаковы, т.к. они определяются точностью станка.

Если же нет оснований считать, что дисперсии в выборках одинаковы, то следует по критерию Фишера-Снедекора проверить гипотезу о равенстве генеральных дисперсий.

Лучшей оценкой для генеральной дисперсии в данном случае будет дисперсия смешенной совокупности объема $n_1 + n_2$:

$$\hat{S}^2 = \frac{1}{n_1+n_2-2} \cdot \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right) = \frac{n_1 S_x^2 + n_2 S_y^2}{n_1+n_2-2}.$$

Оценка дисперсии разности независимых выборочных средних определится выражением:

$$\hat{S}^2_{\bar{x}-\bar{y}} = \frac{n_1 S_x^2 + n_2 S_y^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{(n_1-1)\hat{S}_x^2 + (n_2-1)\hat{S}_y^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

В качестве критерия проверки нулевой гипотезы H_0: $\bar{x}_0 = \bar{y}_0$ принимают случайную величину:

$$T = \frac{\bar{x}-\bar{y}}{\sqrt{\hat{S}^2_{\bar{x}-\bar{y}}}} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{n_1 S_x^2 + n_2 S_y^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

имеющую t-распределение Стьюдента с $n_1 + n_2 - 2 = k$ числом степеней свободы.

Критическая область зависит от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: \bar{x}_0 = \bar{y}_0$. Конкурирующая гипотеза $H_1: \bar{x}_0 \neq \bar{y}_0$. В данном случае строится двусторонняя критическая область $\omega = \left(-\infty; t_{\text{кр.л}}\right) \cup \left(t_{\text{кр.п}}; +\infty\right)$, исходя из требования, чтобы вероятность попадания критерия T в эту область была равна принятому уровню значимости α.

Наибольшая мощность критерия достигается тогда, когда левая и правая критические точки $t_{\text{кр.л}}$, $t_{\text{кр.п}}$ выбраны так, чтобы вероятность попадания критерия в каждый из двух интервалов двусторонней критической области была равна $\alpha/2$, т.е.:

$$P\left(T < t_{\text{кр.л}}\right) = \frac{\alpha}{2} \quad \text{и} \quad P\left(T > t_{\text{кр.п}}\right) = \frac{\alpha}{2}.$$

Запишем очевидное равенство:

$$P\left(T < t_{\text{кр.л}}\right) + P\left(t_{\text{кр.л}} < T < t_{\text{кр.п}}\right) + P\left(T > t_{\text{кр.п}}\right) = 1;$$

$$\frac{\alpha}{2} + P\left(t_{\text{кр.л}} < T < t_{\text{кр.п}}\right) + \frac{\alpha}{2} = 1.$$

Поскольку t-распределение Стьюдента симметрично относительно $T = 0$, то $t_{\text{кр.л}} = -t_{\text{кр.п}}$. Следовательно:

$$P\left(-t_{\text{кр.п}} < T < t_{\text{кр.п}}\right) = 1 - \alpha.$$

Значения $t_{\text{кр.п}}$ можно найти из приложения 7.

Пример 5.7.1 По двум независимым малым выборкам, объемы которых $n_1 = 12$ и $n_2 = 20$, извлеченными из нормальных генеральных совокупностей X и Y, найдены выборочные средние $\bar{x} = 31$, $\bar{y} = 29$ и исправленные дисперсии $\hat{S}_x^2 = 0{,}64$ и $\hat{S}_y^2 = 0{,}50$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу $H_0: \bar{x}_0 = \bar{y}_0$ при конкурирующей гипотезе $H_1: \bar{x}_0 \neq \bar{y}_0$.

Поскольку исправленные дисперсии различны, то проверим по критерию Фишера-Снедекора, значимо ли различаются дисперсии выборок. Проверим гипотезу $H_0: \sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе $H_1: \sigma_x^2 > \sigma_y^2$.

Найдем наблюдаемое значение критерия Фишера-Снедекора:

$$F_{\text{набл}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} = \frac{0{,}64}{0{,}5} = 1{,}28.$$

$k_1 = n_1 - 1 = 12 - 1 = 11$ и $k_2 = n_2 - 1 = 20 - 1 = 19$.

По таблице приложения 4 найдем:

$$F_{\text{кр.п}} = F_{\text{кр.п}}(k_1 = 11; k_2 = 19; \alpha = 0{,}05) = 2{,}35.$$

Критическая область имеет вид $\omega_F = (2{,}35; +\infty)$. Поскольку $F_{\text{наб}} \notin \omega$, гипотезу H_0 принимаем. Таким образом, можно считать, что на данном уровне значимости дисперсии выборок отличаются не значимо. Значит, дисперсии считаем равными.

Перейдем к проверке гипотезы $H_0: \bar{x}_0 = \bar{y}_0$ при конкурирующей гипотезе $H_1: \bar{x}_0 \neq \bar{y}_0$.

Наблюдаемое значение t – критерия Стьюдента:

$$T_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_1 - 1)\hat{S}_x^2 + (n_2 - 1)\hat{S}_y^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} =$$

$$= \frac{31 - 29}{\sqrt{\frac{(12 - 1) \cdot 0{,}64 + (20 - 1) \cdot 0{,}5}{12 + 20 - 2}\left(\frac{1}{12} + \frac{1}{20}\right)}} = 7{,}1.$$

По таблице приложения 7 найдем:

$$t_{\text{кр.п}} = t_{\text{кр.п}}(k = n_1 + n_2 - 2 = 30; \alpha = 0{,}05) = 2{,}04.$$

Критическая область имеет вид $\omega_t = (0; -2{,}04) \cup (2{,}04; +\infty)$. Поскольку $T_{\text{набл}} \in \omega$ нулевую гипотезу отвергаем.

Ответ нулевую гипотезу H_0 отвергаем.

Второй случай. Нулевая гипотеза $H_0: \bar{x}_0 = \bar{y}_0$. Конкурирующая гипотеза $H_1: \bar{x}_0 > \bar{y}_0$. В данном случае строится односторонняя критическая область $\omega = \left(t_{\text{кр.п}}; +\infty\right)$, исходя из требования, чтобы вероятность попадания критерия T в эту область была равна принятому уровню значимости α.

Запишем очевидные равенства:

$$P(-\infty < T < 0) + P\left(0 < T < t_{\text{кр.п}}\right) + P\left(T > t_{\text{кр.п}}\right) = 1;$$

$$\frac{1}{2} + P\left(0 < T < t_{\text{кр.п}}\right) + \alpha = 1; \quad P\left(0 < T < t_{\text{кр.п}}\right) = \frac{1}{2} - \alpha;$$

$$2P\left(0 < T < t_{\text{кр.п}}\right) = P\left(-t_{\text{кр.п}} < T < t_{\text{кр.п}}\right) = 1 - 2\alpha.$$

Значения $t_{\text{кр.п}}$ можно найти из приложения 7.

Пример 5.7.2 На уровне значимости $\alpha = 0{,}05$ требуется проверить гипотезу $H_0: \bar{x}_0 = \bar{y}_0$ о равенстве генеральных средних нормальных совокупностей X и Y при конкурирующей гипотезе $H_1: \bar{x}_0 > \bar{y}_0$ по малым независимым выборкам, объемы которых $n_1 = 10$ и $n_2 = 16$. Получены следующие результаты:

x_i	12,3	12,5	12,8	13,0	13,5
m_i	1	2	4	2	1

y_j	12,2	12,3	13,0
m_j	6	8	2

Вычислим средние величины и дисперсии:

$$\bar{x} = \frac{12{,}3 \cdot 1 + 12{,}5 \cdot 2 + 12{,}8 \cdot 4 + 13{,}0 \cdot 2 + 13{,}5 \cdot 1}{10} = 12{,}8;$$

$$\bar{y} = \frac{12{,}2 \cdot 6 + 12{,}3 \cdot 8 + 13{,}0 \cdot 2}{16} = 12{,}35;$$

$$\overline{x^2} = \frac{12{,}3^2 \cdot 1 + 12{,}5^2 \cdot 2 + 12{,}8^2 \cdot 4 + 13{,}0^2 \cdot 2 + 13{,}5^2 \cdot 1}{10} = 163{,}94;$$

$$\overline{y^2} = \frac{12{,}2^2 \cdot 6 + 12{,}3^2 \cdot 8 + 13{,}0^2 \cdot 2}{16} = 152{,}585.$$

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 163{,}94 - 12{,}8^2 = 0{,}1;$$

$$S_y^2 = \overline{y^2} - \bar{y}^2 = 152{,}585 - 12{,}35^2 = 0{,}0625;$$

$$\hat{S}_x^2 = \frac{n_1}{n_1 - 1} S_x^2 = \frac{10}{10 - 1} 0{,}1 = 0{,}111;$$

$$\hat{S}_y^2 = \frac{n_2}{n_2 - 1} S_x^2 = \frac{16}{16 - 1} 0{,}0625 = 0{,}067.$$

Поскольку исправленные дисперсии различны, то проверим по критерию Фишера-Снедекора, значимо ли различаются дисперсии выборок. Проверим гипотезу $H_0: \sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе $H_1: \sigma_x^2 > \sigma_y^2$.

Найдем наблюдаемое значение критерия Фишера-Снедекора:

$$F_{\text{набл}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} = \frac{0{,}111}{0{,}067} = 1{,}66.$$

$k_1 = n_1 - 1 = 10 - 1 = 9$ и $k_2 = n_2 - 1 = 16 - 1 = 15$.

По таблице приложения 4 найдем:

$$F_{\text{кр.п}} = F_{\text{кр.п}}(k_1 = 9; k_2 = 15; \alpha = 0{,}05) = 2{,}59.$$

Критическая область имеет вид $\omega_F = (2{,}59; +\infty)$. Поскольку $F_{\text{наб}} \notin \omega$, гипотезу H_0 принимаем. Таким образом, можно считать, что на данном уровне значимости дисперсии выборок отличаются не значимо. Значит, дисперсии считаем равными.

Перейдем к проверке гипотезы $H_0: \bar{x}_0 = \bar{y}_0$ при конкурирующей гипотезе $H_1: \bar{x}_0 > \bar{y}_0$.

Наблюдаемое значение t – критерия Стьюдента:

$$T_{\text{набл}} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{n_1 S_x^2 + n_2 S_y^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{12{,}8-12{,}35}{\sqrt{\frac{10\cdot 0{,}1+16\cdot 0{,}0625}{10+16-2}\left(\frac{1}{10}+\frac{1}{16}\right)}} = 3{,}9.$$

По таблице приложения 7 найдем:

$$t_{\text{кр.п}} = t_{\text{кр.п}}(k = n_1+n_2-2 = 24; \alpha = 0{,}05) = 1{,}71.$$

Критическая область имеет вид $\omega_t = (1{,}71; +\infty)$. Поскольку $T_{\text{набл}} \in \omega$ нулевую гипотезу отвергаем.

Ответ нулевую гипотезу H_0 отвергаем.

5.8. Сравнение выборочной средней с гипотетической генеральной средней нормальной совокупности

А. Дисперсия генеральной совокупности известна.

Примем, что генеральная совокупность X распределена нормально, причем генеральная средняя $\bar{x}_0$ хотя и неизвестна, но имеются основания предполагать, что она равна гипотетическому (предполагаемому) значению a.

Предположим, что дисперсия генеральной совокупности известна, например, из предшествующего опыта, или найдена теоретически, или вычислена по выборке большого объема. Заметим, что по выборке большого объема можно получить достаточно хорошую оценку дисперсии.

Пусть из нормальной генеральной совокупности извлечена выборка объема n и по ней найдена выборочная средняя $\bar{x}$, причем генеральная дисперсия σ^2 известна. Требуется по выборочной средней при заданном уровне значимости проверить нулевую гипотезу $H_0: \bar{x}_0 = a$ о равенстве генеральной средней $\bar{x}_0$ гипотетическому значению a.

В качестве критерия проверки нулевой гипотезы примем случайную величину:

$$z = \frac{\bar{x}-a}{\sigma_{\bar{x}}} = \frac{\bar{x}-a}{\sigma/\sqrt{n}},$$

Которая распределена нормально, причем при справедливости нулевой гипотезы $M(z) = 0$ и $\sigma(z) = 1$.

Критическая область зависит от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 \neq a$. В данном случае критическая область является двусторонней $\omega = \left(-\infty; z_{\text{кр.л}}\right) \cup \left(z_{\text{кр.п}}; +\infty\right)$. Где $z_{\text{кр.л}} = -z_{\text{кр.п}}$, а $z_{\text{кр.п}}$ находят из соотношения:

$$\phi\left(z_{\text{кр.п}}\right) = \frac{1-\alpha}{2},$$

по таблице приложения 1 при заданном уровне значимости α.

Второй случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 > a$. В данном случае критическая область является односторонней $\omega = \left(z_{\text{кр.п}}; +\infty\right)$. Где $z_{\text{кр.п}}$ находят из соотношения:

$$\phi\left(z_{\text{кр.п}}\right) = \frac{1-2\alpha}{2},$$

по таблице приложения 1 при заданном уровне значимости α.

Третий случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 < a$. В данном случае критическая область является односторонней $\omega = \left(-\infty; -z_{\text{кр.п}}\right)$. Где $z_{\text{кр.п}}$ находят из соотношения:

$$\phi\left(z_{\text{кр.п}}\right) = \frac{1-2\alpha}{2},$$

по таблице приложения 1 при заданном уровне значимости α.

Пример 5.8.1 Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 40$ извлечена выборка объема $n = 64$ и по ней найдена выборочная средняя $\bar{x} = 136{,}5$. Требуется при уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу $H_0: \bar{x}_0 = a = 130$ при конкурирующей гипотезе

а) $H_0: \bar{x}_0 \neq 130$; б) $H_0: \bar{x}_0 > 130$.

Наблюдаемое значение критерия:

$$z_{\text{набл}} = \frac{\bar{x}-a}{\sigma/\sqrt{n}} = \frac{136{,}5-130}{40/\sqrt{64}} = 1{,}3.$$

а) Если конкурирующая гипотеза $H_0: \bar{x}_0 \neq 130$, то критическая область двусторонняя. По приложению 1 для:

$$\phi\left(z_{\text{кр.п}}\right) = \frac{1-\alpha}{2} = \frac{1-0{,}01}{2} = 0{,}495,$$

найдем $z_{\text{кр.п}} = 2{,}58$, $z_{\text{кр.л}} = -z_{\text{кр.п}} = -2{,}58$,

$$\omega = (-\infty;\ -2{,}58) \cup (2{,}58; +\infty).$$

Поскольку $z_{\text{набл}} \notin \omega$, то нулевую гипотезу принимаем.

б) Если конкурирующая гипотеза $H_0: \bar{x}_0 > 130$, то критическая область односторонняя. По приложению 1 для:

$$\phi\left(z_{\text{кр.п}}\right) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}01}{2} = 0{,}49,$$

найдем $z_{\text{кр.п}} = 2{,}33$, $\omega = (2{,}33; +\infty)$.

Поскольку $z_{\text{набл}} \notin \omega$, то нулевую гипотезу принимаем.

Ответ а) нулевую гипотезу H_0 принимаем;

б) нулевую гипотезу H_0 принимаем.

Б. Дисперсия генеральной совокупности неизвестна.

Если дисперсия генеральной совокупности неизвестна, то в качестве критерия проверки нулевой гипотезы принимают случайную величину:

$$T = \frac{\bar{x} - a}{\hat{S}/\sqrt{n}},$$

имеющую t – распределение Стьюдента с $k = n - 1$ степенями свободы.

Критическая область строится в зависимости от конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 \neq a$. В данном случае критическая область является двусторонней $\omega = (-\infty; t_{\text{кр.л}}) \cup (t_{\text{кр.п}}; +\infty)$. Где $t_{\text{кр.л}} = -t_{\text{кр.п}}$, а $t_{\text{кр.п}}$ находят из соотношения:

$$P(-t_{\text{кр.п}} < T < t_{\text{кр.п}}) = 1 - \alpha,$$

по таблице приложения 7 при заданном уровне значимости α и числе степеней свободы k.

Второй случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 > a$. В данном случае критическая область является односторонней $\omega = (t_{\text{кр.п}}; +\infty)$. Где $t_{\text{кр.п}}$ находят из соотношения:

$$P(-t_{\text{кр.п}} < T < t_{\text{кр.п}}) = 1 - 2\alpha,$$

по таблице приложения 6 при заданном уровне значимости α и числе степеней свободы k.

Третий случай. Нулевая гипотеза $H_0: \bar{x}_0 = a$, конкурирующая гипотеза $H_1: \bar{x}_0 < a$. В данном случае критическая область является односторонней $\omega = (-\infty; -t_{\text{кр.п}})$. Где $t_{\text{кр.п}}$ находят из соотношения:

$$P(-t_{\text{кр.п}} < T < t_{\text{кр.п}}) = 1 - 2\alpha,$$

по таблице приложения 7 при заданном уровне значимости α и числе степеней свободы k.

Пример 5.8.2. По выборке объема $n = 22$, извлеченной из нормальной генеральной совокупности, найдены выборочная средняя $\bar{x} = 18$ и исправленное среднее квадратическое отклонение $\hat{S} = 5$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу $H_0: \bar{x}_0 = a = 19$, при конкурирующей гипотезе:

а) $H_1: \bar{x}_0 \neq 19$;

б) $H_1: \bar{x}_0 < 19$.

Вычислим наблюдаемое значение t – критерия Стьюдента:

$$T_{\text{набл}} = \frac{\bar{x} - a}{\hat{S}/\sqrt{n}} = \frac{18 - 19}{5/\sqrt{22}} = -0{,}94.$$

Число степеней свободы $k = 22 - 1 = 21$.

а) Если конкурирующая гипотеза $H_0: \bar{x}_0 \neq 19$, то критическая область двусторонняя. По приложению 7 найдем:

$$t_{\text{кр.п}} = t_{\text{кр.п}}(k = 21; \alpha = 0{,}05) = 2{,}08.$$

Критическая область $\omega = (-\infty;\ -2{,}08) \cup (2{,}08; +\infty)$.

Поскольку $T_{\text{набл}} \notin \omega$, то нулевую гипотезу принимаем.

б) Если конкурирующая гипотеза $H_0: \bar{x}_0 < 19$, то критическая область односторонняя. По приложению 7 найдем:

$$t_{\text{кр.п}} = t_{\text{кр.п}}(k = 21; \alpha = 0{,}05) = 1{,}72.$$

Критическая область $\omega = (-\infty;\ -1{,}72)$.

Поскольку $T_{\text{набл}} \notin \omega$, то нулевую гипотезу принимаем.

Ответ а) нулевую гипотезу H_0 принимаем;

б) нулевую гипотезу H_0 принимаем.

5.9. Проверка гипотезы о законе распределения генеральной совокупности. Критерий согласия Пирсона

Если закон распределения неизвестен, но есть основания предположить, что он имеет определенный вид (назовем его А), то проверяют нулевую гипотезу: генеральная совокупность распределена по закону А.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т.е. при помощи специально подобранной случайной величины – критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: Пирсона (χ^2 – «хи квадрат»); Колмогорова; Смирнова и др. Рассмотрим критерий Пирсона. Будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные) частоты. Эмпирические и теоретические частоты обычно различаются. Расхождение может быть случайно (незначимо) и связано с малым числом наблюдений, либо со способом их группировки, либо с другими причинами. Расхождение частот может быть не случайно (значимо) и объясняться тем, что теоретические частоты вычислены исходя из неверной гипотезы о виде закона распределения.

Пусть по выборке объема n получено эмпирическое распределение:

x_i	x_1	x_2		x_m
n_i	n_1	n_2		n_m

Допустим, что в предположении закона распределения А для генеральной совокупности вычислены теоретические частоты n_i'. При уровне значимости α требуется проверить гипотезу: генеральная совокупность имеет закон распределения А.

В качестве критерия проверки нулевой гипотезы примем случайную величину:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}.$$

Доказано, что при $n \to \infty$ данная случайная величина имеет закон распределения χ^2 независимо от того какому закону распределения подчинена генеральная совокупность.

Число степеней свободы χ^2 находят по равенству $k = m - 1 - r$, где m – число групп (частичных интервалов) выборки; r – число параметров предполагаемого распределения, которые оценены по данным выборки.

Замечание. Малочисленные интервалы следует объединять так, чтобы каждый интервал содержал не менее 5 вариант.

Если предполагаемое распределение является нормальным законом распределения, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение), поэтому $r = 2$ и число степеней свободы $k = m - 1 - 2 = m - 3$.

Если, например, генеральная совокупность распределена по закону Пуассона, то оценивают один параметр, поэтому $r = 1$ и $k = m - 2$.

Поскольку односторонний критерий более жестко отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область $\omega = \left(\chi^2_{\text{кр}}; +\infty\right)$, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left(\chi^2 > \chi^2_{\text{кр}}(k;\alpha)\right) = \alpha.$$

Величину $\chi^2_{\text{кр}} = \chi^2_{\text{кр}}(k;\alpha)$ находят по таблице приложения 5.

Пример 5.9.1 Имеются следующие данные о засоренности партии семян клевера семенами сорняков:

Число семян в одной	0	1	2	3	4	5	6	$\sum$

пробе x_i								
Число проб n_i	405	366	175	40	8	4	2	1000

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X – число семян сорняков распределена по закону Пуассона.

Найдем выборочное среднее:

$$\bar{x} = \frac{0 \cdot 405 + 1 \cdot 366 + 2 \cdot 175 + 3 \cdot 40 + 4 \cdot 8 + 5 \cdot 4 + 6 \cdot 2}{1000} = 0{,}9.$$

Используя метод моментов или метод максимального правдоподобия, можно показать, что параметр λ для распределения Пуассона равен выборочному среднему $\lambda = \bar{x} = 0{,}9$.

По формуле Пуассона:

$$P(X = x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}, \quad x_i = 0{,}1{,}2{,}3{,}4{,}5{,}6.$$

вычислим вероятности:

$$P(X = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = \frac{0{,}9^0}{0!} e^{-0{,}9} = 0{,}4066;$$

$$P(X = 1) = \frac{\lambda^1}{1!} e^{-\lambda} = \frac{0{,}9^1}{1!} e^{-0{,}9} = 0{,}3659;$$

$$P(X = 2) = \frac{\lambda^2}{2!} e^{-\lambda} = \frac{0{,}9^2}{2!} e^{-0{,}9} = 0{,}1647;$$

$$P(X = 3) = \frac{\lambda^3}{3!} e^{-\lambda} = \frac{0{,}9^3}{3!} e^{-0{,}9} = 0{,}0494;$$

$$P(X = 4) = \frac{\lambda^4}{4!} e^{-\lambda} = \frac{0{,}9^4}{4!} e^{-0{,}9} = 0{,}0111;$$

$$P(X = 5) = \frac{\lambda^5}{5!} e^{-\lambda} = \frac{0{,}9^5}{5!} e^{-0{,}9} = 0{,}0020;$$

$$P(X = 6) = \frac{\lambda^6}{6!} e^{-\lambda} = \frac{0{,}9^6}{6!} e^{-0{,}9} = 0{,}0003.$$

Вычислим теоретические частоты:

$$n'_1 = n \cdot P(X = 0) = 1000 \cdot 0{,}4066 = 406{,}6;$$

$$n'_2 = n \cdot P(X = 1) = 1000 \cdot 0{,}3659 = 365{,}9;$$

$$n'_3 = n \cdot P(X = 2) = 1000 \cdot 0{,}1647 = 164{,}7;$$

$$n'_4 = n \cdot P(X = 3) = 1000 \cdot 0{,}0494 = 49{,}4;$$

$$n'_5 = n \cdot P(X = 4) = 1000 \cdot 0{,}0111 = 11{,}1;$$

$$n'_6 = n \cdot P(X = 5) = 1000 \cdot 0{,}0020 = 2{,}0;$$

$$n'_7 = n \cdot P(X = 6) = 1000 \cdot 0{,}0003 = 0{,}3.$$

Поскольку $n'_7 < 5$, $n'_7 + n'_6 < 5$, $n'_5 + n'_6 + n'_7 > 5$ объединим последние 3 интервала. Обозначим $n'_5 = 11{,}1 + 2{,}0 + 0{,}3 = 13{,}4$. Эмпирические частоты так же нужно объединить $n_5 = 8 + 4 + 2 = 14$.

Вычислим наблюдаемое значение критерия Пирсона:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{m} \frac{(n_i - n'_i)^2}{n'_i} = \sum_{i=1}^{5} \frac{(n_i - n'_i)^2}{n'_i} =$$

$$= \frac{(405 - 406{,}6)^2}{406{,}6} + \frac{(366 - 365{,}9)^2}{365{,}9} + \frac{(175 - 164{,}7)^2}{164{,}7} +$$

$$+ \frac{(40 - 49{,}4)^2}{49{,}4} + \frac{(14 - 13{,}4)^2}{13{,}4} = 2{,}47.$$

Найдем число степеней свободы:

$$k = m - r - 1 = m - 1 - 1 = m - 2 = 5 - 2 = 3.$$

По таблице приложения 5, найдем:

$$\chi^2_{\text{кр}} = \chi^2_{\text{кр}}(k = 3; \alpha = 0{,}05) = 7{,}82.$$

Критическая область имеет вид: $\omega = (7{,}82; +\infty)$.

Поскольку наблюдаемое значение критерия Пирсона не принадлежит критической области $\chi^2_{\text{набл}} \notin \omega$, то гипотезу H_0 принимаем.

Ответ Случайная величина X – число семян сорняков распределена по закону Пуассона.

Глава 6. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

6.1. Функциональная, статистическая и корреляционная зависимости

В естественных науках часто речь идет о функциональной зависимости, когда значению одной переменной соответствует вполне определенное значение другой.

Между переменными величинами иногда может существовать такая зависимость, когда каждому значению одной переменной соответствует не какое-то определенное, а множество возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует определенное (условное) распределение другой переменной. Такая зависимость получила название статистической (или стохастической, вероятностной).

Возникновение понятия статистической связи обуславливается тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками. Примером статистической зависимости является зависимость урожайности от количества внесенных удобрений, производительности труда на предприятии от его энерговооруженности и т.п.

В силу неоднозначности статистической зависимости между Y и X для исследователя, в частности, представляет интерес усредненная по x схема зависимости, т.е. закономерность в изменении среднего значения – условного математического ожидания $M_x(Y)$ в зависимости от x.

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой, называется корреляционной. Иначе, корреляционной зависимостью между двумя переменными величинами называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Корреляционная зависимость может быть представлена в виде:

$$M_x(Y) = \varphi(x) \text{ или } M_y(X) = \psi(y).$$

Предполагается, что $\varphi(x) \neq const$ и $\psi(y) \neq const$, т.е. если при изменении x или y условные математические ожидания $M_x(Y)$ и $M_y(X)$ не изменяются, то говорят, что корреляционная зависимость между переменными X и Y отсутствует.

Сравнивая различные виды зависимости между X и Y, можно сказать, что с изменением значений переменной X при функциональной зависимости однозначно изменяется определенное значение переменной Y, при корреляционной – определенное среднее значение (условное математическое ожидание) Y, а при статистической – определенное (условное) распределение переменной Y. Таким образом, из рассмотренных зависимостей наиболее общей выступает статистическая зависимость.

Уравнения $M_x(Y) = \varphi(x)$ и $M_y(X) = \psi(y)$ называются уравнениями регрессии соответственно Y по X и X по Y. Функции $\varphi(x)$ и $\psi(y)$ называются функциями регрессии, а их графики – линиями регрессии.

На практике исследователь располагает лишь выборкой пар значений $(x_i; y_i)$ ограниченного объема. В этом случае речь может идти об оценке по выборке функции регрессии. Такой оценкой является выборочная линия регрессии Y по X:

$$\bar{y}_x = \hat{\varphi}\left(x; b_0; b_1; \dots; b_p\right)$$

где $\bar{y}_x$ – условная (групповая) средняя переменной Y при фиксированном значении переменной $X = x$, $b_0; b_1; \dots; b_p$ – параметры кривой, которые обычно находят по методу наименьших квадратов.

Аналогично определяется выборочная линия регрессии X по Y:

$$\bar{x}_y = \hat{\psi}\left(y; c_0; c_1; \dots; c_p\right)$$

где $\bar{x}_y$ – условная (групповая) средняя переменной X при фиксированном значении переменной $Y = y$, $c_0; c_1; \dots; c_p$ – параметры кривой.

6.2. Линейная парная регрессия

Пусть в результате n независимых опытов получены n пар чисел $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$. Найдем по данным наблюдений выборочное уравнение прямой линии среднеквадратической регрессии. Будем искать уравнение регрессии Y на X в виде:

$$\bar{y}_x = b_{yx} \cdot x + b_0,$$

где b_{yx} – выборочный коэффициент регрессии Y на X.

Поскольку различные значения x признака X и соответствующие им значения y признака Y наблюдались по одному разу, то группировать данные нет необходимости. Так же нет надобности использовать понятие условной средней, поэтому искомое уравнение можно записать так:

$$y = b_{yx} \cdot x + b_0.$$

Параметры b_{yx}, b_0 подберем по методу наименьших квадратов. Для этого составим сумму квадратов невязок:

$$S(b_{yx}; b_0) = \sum_{i=1}^{n} \left(b_{yx} \cdot x_i + b_0 - y_i\right)^2.$$

Вычислим производные:

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n} 2\left(b_{yx} \cdot x_i + b_0 - y_i\right); \quad \frac{\partial S}{\partial b_{yx}} = \sum_{i=1}^{n} 2\left(b_{yx} \cdot x_i + b_0 - y_i\right) \cdot x_i.$$

Приравняем данные производные к нулю и разделим на объем выборки n. Получим систему нормальных уравнений:

$$\begin{cases} b_{yx} \cdot \bar{x} + b_0 = \bar{y}; \\ b_{yx} \cdot \overline{x^2} + b_0 \cdot \bar{x} = \overline{xy}. \end{cases}$$

Из первого уравнения найдем $b_0 = \bar{y} - b_{yx} \cdot \bar{x}$ и подставим во второе уравнение:

$$b_{yx} \cdot \overline{x^2} + \left(\bar{y} - b_{yx} \cdot \bar{x}\right) \cdot \bar{x} = \overline{xy}.$$

Из данного уравнения найдем коэффициент регрессии Y на X:

$$b_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x^2} = \frac{\mu_{yx}}{S_x^2},$$

где μ_{yx} – выборочный коэффициент ковариации.

Подставляя $b_0 = \bar{y} - b_{yx} \cdot \bar{x}$ в уравнение прямой регрессии Y на X, найдем окончательный вид этого уравнения:

$$y - \bar{y} = b_{yx} \cdot (x - \bar{x}).$$

Из данного уравнения видно, что *коэффициент b_{yx} регрессии Y на X показывает, на сколько единиц в среднем изменяется переменная Y при увеличении переменной X на одну единицу*.

Производя формальные замены $x \to y$ и $y \to x$, найдем коэффициент регрессии X на Y:

$$b_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{y^2} - \bar{y}^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_y^2} = \frac{\mu_{xy}}{S_y^2}$$

и уравнение прямой регрессии X на Y:

$$x - \bar{x} = b_{xy} \cdot (y - \bar{y}).$$

Следовательно, *коэффициент b_{xy} регрессии X на Y показывает, на сколько единиц в среднем изменяется переменная X при увеличении переменной Y на одну единицу*.

Пример 6.2.1. Найти выборочные уравнения прямых регрессии Y на X и X на Y по данным $n = 5$ наблюдений:

x_i	1	2	4	5	6
y_i	2	3	4	5	7

Для удобства решения составим таблицу:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	1	2	2	1	4
	2	3	6	4	9
	4	4	16	16	16
	5	5	25	25	25
	6	7	42	36	49
$\sum$	18	21	91	82	103

Вычислим средние значения величин:

$$\bar{x} = \frac{18}{5} = 3{,}6;\ \bar{y} = \frac{21}{5} = 4{,}2;\ \overline{xy} = \frac{91}{5} = 18{,}2;$$

$$\overline{x^2} = \frac{82}{5} = 16{,}4;\ ;\ \overline{y^2} = \frac{103}{5} = 20{,}6.$$

Вычислим коэффициенты регрессии:

$$b_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{18{,}2 - 3{,}6 \cdot 4{,}2}{16{,}4 - 3{,}6^2} = 0{,}895;$$

$$b_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{y^2} - \bar{y}^2} = \frac{18{,}2 - 3{,}6 \cdot 4{,}2}{20{,}6 - 4{,}2^2} = 1{,}041.$$

Уравнение прямой регрессии Y на X примет вид:

$$y - 4{,}2 = 0{,}895 \cdot (x - 3{,}6);$$

$$y = 0{,}978 + 0{,}895 \cdot x.$$

Уравнение прямой регрессии X на Y примет вид:

$$x - 3{,}6 = 1{,}041 \cdot (y - 4{,}2);$$

$$x = -0{,}772 + 1{,}041 \cdot y.$$

Ответ $y = 0{,}978 + 0{,}895 \cdot x$; $x = -0{,}772 + 1{,}041 \cdot y$.

6.3. Коэффициент корреляции

Перейдем к оценке тесноты корреляционной зависимости. Рассмотрим наиболее важный для практики и теории случай линейной зависимости.

На первый взгляд подходящим измерителем тесноты связи Y от X является коэффициент регрессии b_{yx} так как, он показывает, на сколько единиц в среднем изменяется Y, когда X увеличивается на одну единицу. Однако b_{yx} зависит от единиц измерения переменных.

Для исправления b_{yx} как показателя тесноты связи нужна такая стандартная система единиц измерения, в которой данные по различным характеристикам оказались бы сравнимы между собой. Статистика знает такую систему единиц. Эта система использует в качестве единиц измерения переменной ее среднее квадратическое отклонение S.

Представим уравнение регрессии Y на X в следующем виде:

$$\frac{\bar{y}_x - \bar{y}}{S_y} = \frac{b_{yx} \cdot S_x}{S_y} \cdot \frac{(x - \bar{x})}{S_x}.$$

Видно, что величина:

$$r_{yx} = \frac{b_{yx} \cdot S_x}{S_y} = \frac{\mu_{yx} \cdot S_x}{S_x^2 \cdot S_y} = \frac{\mu_{yx}}{S_x \cdot S_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}$$

показывает, *на сколько величин S_y изменится в среднем величина Y, когда X увеличится на одно S_x.*

Величина r_{yx} является показателем тесноты линейной связи и называется выборочным коэффициентом корреляции.

Из выражения для r_{yx}, видно, что данный коэффициент симметричен относительно замены $x \to y$ и $y \to x$, т.е. $r_{yx} = r_{xy} = r$. Таким образом:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_y \cdot S_x} = \frac{\mu_{yx} \cdot S_y}{S_y^2 \cdot S_x} = \frac{b_{xy} \cdot S_y}{S_x}.$$

Найдем произведение:

$$r_{yx} \cdot r_{xy} = \frac{b_{yx} \cdot S_x}{S_y} \cdot \frac{b_{xy} \cdot S_y}{S_x} = b_{yx} \cdot b_{xy} = r^2.$$

Выражая из последнего уравнения коэффициент корреляции, найдем:

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}\,,$$

где знак $+$ выбирают, если $b_{yx} > 0$ и $b_{xy} > 0$ и знак $-$, если $b_{yx} < 0$ и $b_{xy} < 0$. Следовательно, *коэффициент корреляции r переменных Y и X есть средняя геометрическая коэффициентов регрессии, имеющая их знак.*

Пример 6.3.1. Найти коэффициент корреляции для примера 6.2.1.

Коэффициенты регрессии в примере 6.2.1 равны $b_{yx} = 0{,}895$ и $b_{xy} = 1{,}041$. Следовательно, $r = +\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{0{,}895 \cdot 1{,}041} = 0{,}965$. Знак $+$ выбран так как коэффициенты регрессии b_{yx} и b_{xy} положительны.

Ответ $r = 0{,}965$.

Отметим основные свойства коэффициента корреляции.

1.Коэффициент корреляции принимает значения на отрезке $[-1; 1]$, т.е. $-1 \leq r \leq 1$.

В зависимости от того, насколько $|r|$ приближается к 1, различают связь слабую, умеренную, заметную, достаточно тесную, тесную и весьма тесную, т.е. чем ближе $|r|$ к 1, тем теснее связь.

2.Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

3.При $r = \pm 1$ корреляционная связь представляет линейную функциональную зависимость. При этом линии регрессии Y по X и X по Y совпадают и все наблюдаемые значения располагаются на общей прямой.

Найдем $tg\,\varphi$ между двумя прямыми регрессии с угловыми коэффициентами $k_1 = b_{yx}$ и $k_2 = 1/b_{xy}$:

$$tg\,\varphi = \left|\frac{k_2 - k_1}{1 + k_1 k_2}\right| = \left|\frac{1 - b_{yx} b_{xy}}{b_{xy} + b_{yx}}\right| = \left|\frac{1 - r^2}{r} \frac{S_x S_y}{S_x^2 + S_y^2}\right|.$$

Из полученной формулы видно, что чем теснее связь и чем ближе $|r|$ к 1, тем меньше угол φ между прямыми регрессии, а при $r = \pm 1$ $tg\,\varphi = 0$ и линии регрессии сливаются.

4.При $r = 0$ линейная корреляционная связь отсутствует. При этом групповые средние переменных совпадают с их общими средними, а линии регрессии Y по X и X по Y параллельны осям координат.

На самом деле уравнения регрессии:

$$\bar{y}_x - \bar{y} = \frac{r \cdot S_y}{S_x} \cdot (x - \bar{x});$$

$$\bar{x}_y - \bar{x} = \frac{r \cdot S_x}{S_y} \cdot (y - \bar{y});$$

при $r = 0$ примут вид $\bar{y}_x = \bar{y}$ и $\bar{x}_y = \bar{x}$, т.е. линии регрессии параллельны осям координат.

Отметим, что равенство $r = 0$ указывает только на отсутствие линейной корреляционной зависимости, но не вообще об отсутствии корреляционной, а тем более статистической зависимости.

6.4. Проверка гипотезы о значимости коэффициента корреляции

Пусть двумерная генеральная совокупность (X, Y) распределена нормально. Из этой совокупности извлечена выборка объема n и по ней найден выборочный коэффициент корреляции r, который оказался отличным от нуля. Так как выборка отобрана случайно, то ещё нельзя заключить, что коэффициент корреляции генеральной совокупности ρ также отличен от нуля. В конечном счете нас интересует именно этот коэффициент, поэтому возникает необходимость при заданном уровне значимости α проверить нулевую гипотезу $H_0: \rho = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: \rho \neq 0$.

Если нулевая гипотеза отвергается, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля, а X и Y коррелированы, т.е. связаны линейной зависимостью.

Если нулевая гипотеза будет принята, то выборочный коэффициент корреляции незначим, а X и Y некоррелированные, т.е. не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы примем случайную величину:

$$T = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}.$$

При справедливости нулевой гипотезы данная величина имеет распределение Стьюдента с $k = n - 2$ степенями свободы.

Поскольку конкурирующая гипотеза $H_1: \rho \neq 0$, критическая область строится двусторонней $\omega = \left(-\infty; t_{\text{кр.л}}\right) \cup \left(t_{\text{кр.п}}; +\infty\right)$. Левая критическая точка определяется из равенства $t_{\text{кр.л}} = -t_{\text{кр.п}}$. Правая критическая точка выбирается по приложению 7 исходя из равенства

$$P\left(-t_{\text{кр.п}} < T < t_{\text{кр.п}}\right) = 1 - \alpha.$$

Пример 6.4.1 По выборке объема $n = 62$, извлеченной из нормальной двумерной совокупности, найден выборочный коэффициент корреляции $r = 0{,}5$. При уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции.

Найдем наблюдаемое значение критерия:

$$T_{\text{набл}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}5 \cdot \sqrt{62-2}}{\sqrt{1-0{,}5^2}} = 4{,}47.$$

Число степеней свободы $k = n - 2 = 62 - 2 = 60$. По таблице приложения 6 найдем $t_{\text{кр.п}} = t_{\text{кр.п}}(k = 60, \alpha = 0{,}05) = 2{,}00$. Критическая область примет вид $\omega = (-\infty; -2) \cup (2; +\infty)$. Поскольку $T_{\text{набл}} \in \omega$ нулевую гипотезу отвергаем. Следовательно, генеральный коэффициент корреляции ρ значимо отличается от нуля, а величины X и Y коррелированы.

Ответ коэффициент корреляции ρ значимо отличается от нуля.

6.5. Методика построения корреляционной таблицы

При исследовании корреляционной зависимости между X и Y основой является достаточно большое количество эмпирических точек $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$. Если количество точек n очень велико, то при исследовании зависимости между X и Y используют корреляционную таблицу. Для ее построения нужно проделать следующие действия:

1.Найти наименьшие x_{min}, y_{min} и наибольшие x_{max}, y_{max} значения переменных.

2.Вычислить размахи вариаций $R_x = x_{max} - x_{min}$, $R_y = y_{max} - y_{min}$ для каждой переменной.

3.Выбрать количество интервалов по каждой переменной m_x, m_y.

4.Вычислить ширину интервалов для каждой переменной $\Delta x = R_x/m_x$, $\Delta y = R_y/m_y$.

5.Сформировать интервалы для каждой переменной:

$$[x_{min}; x_{min} + \Delta x); \quad [x_{min} + \Delta x; x_{min} + 2\Delta x); \quad [x_{max} - \Delta x; x_{max}];$$
$$[y_{min}; y_{min} + \Delta y); \quad [y_{min} + \Delta y; y_{min} + 2\Delta y); \quad [y_{max} - \Delta y; y_{max}].$$

6.Найти количество n_{ij} эмпирических точек, попадающих в интервалы.

В результате можно сформировать корреляционную таблицу.

Y \ X	$[x_{min}; x_{min} + \Delta x)$	…	$[x_{max} - \Delta x; x_{max}]$
$[y_{min}; y_{min} + \Delta y)$	n_{11}	…	n_{1m_x}
…	…	…	…
$[y_{max} - \Delta y; y_{max}]$	$n_{m_y 1}$	…	$n_{m_y m_x}$

Числа n_{ij} называются частотами. Очевидно, что сумма всех частот в таблице равна объему выборки.

Пример 6.5.1. Даны 50 эмпирических точек.

$(5,7; 12,8), (16,9; 19,3), (13,4; 18,4), (8,6; 15,8), (2,5; 11,2), (19,6; 21,4), (14,2; 17), (2,6; 11,9), (5,6; 12,9), (0,3; 10,1)$ $(16,1; 17,3), (19; 24,4), (13,9; 18,9), (18,3; 23,5), (3,2; 12,1), (16,2; 18), (2,2; 11,4), (4; 13,1), (18,9; 21,8), (1,2; 11), (14,9; 16), (10,2; 15,4), (10,1; 16,1), (5,9; 12,8), (13,5; 16,4), (5,6; 13,4), (8,1; 13,5), (11,9; 18,5), (18,4; 20,4), (15,4; 21,2), (9,5; 15,2), (14,6; 18,8), (13,8; 19), (15,6; 18,1), (5,7; 14), (3,8; 12,1), (5,7; 13,6), (7; 13,5), (8,1; 13,6), (11,5; 18,1), (6,2; 12,7), (15,7; 18,3), (1,4; 11,1), (18,3; 19,7), (14,4; 17,2), (16,1; 19,2), (6,6; 13,6), (6,6; 14,1),$

$$(1{,}5;\,10{,}9),\,(9{,}3;\,15{,}2).$$

Сформировать корреляционную таблицу.

Найдем минимальные и максимальные значения x, y:

$$x_{min} = 0{,}3, \quad x_{max} = 19{,}6, \quad y_{min} = 10{,}1, \quad y_{max} = 24{,}4.$$

Для удобства дальнейших расчетов примем:

$$x_{min} = 0, \quad x_{max} = 20, \quad y_{min} = 10, \quad y_{max} = 25, \quad m_x = m_y = 5.$$

Вычислим:

$$R_x = x_{max} - x_{min} = 20 - 0 = 20,\ R_y = y_{max} - y_{min} = 25 - 10 = 15,$$

$$\Delta x = \frac{R_x}{m_x} = \frac{20}{5} = 4, \quad \Delta y = \frac{R_y}{m_y} = \frac{15}{5} = 3.$$

Сформируем корреляционную таблицу:

Y \ X	$[0;4)$	$[4;8)$	$[8;12)$	$[12;16)$	$[16;20]$
$[10;13)$	9	4			
$[13;16)$		7	6		
$[16;19)$			3	9	2
$[19;22)$				2	6
$[22;25]$					2

Для проверки найдем сумму частот:

$$9 + 4 + 7 + 6 + 3 + 9 + 2 + 2 + 6 + 2 = 50.$$

Пример 6.5.2. Используя корреляционную таблицу примера 6.5.1, найти уравнения прямых регрессии Y по X и X по Y, коэффициент корреляции. Проверить гипотезу о значимости коэффициента корреляции при уровне значимости $\alpha = 0{,}05$.

В корреляционной таблице примера 6.5.1 заменим интервалы средними значениями:

Y \ X	2	6	10	14	18
11,5	9	4			
14,5		7	6		
17,5			3	9	2
20,5				2	6
23,5					2

По корреляционной таблице вычислим среднее значение:

$$\overline{xy} = \frac{1}{50} \cdot (11{,}5 \cdot (2 \cdot 9 + 6 \cdot 4) + 14{,}5 \cdot (6 \cdot 7 + 10 \cdot 6) +$$
$$+17{,}5 \cdot (10 \cdot 3 + 14 \cdot 9 + 18 \cdot 2) + 20{,}5 \cdot (14 \cdot 2 + 18 \cdot 6) +$$
$$+23{,}5 \cdot 18 \cdot 2) = 179{,}12.$$

Составим закон распределения случайной величины X:

X	2	6	10	14	18
n	9	11	9	11	10

По данному распределению вычислим средние характеристики:

$$\bar{x} = \frac{2 \cdot 9 + 6 \cdot 11 + 10 \cdot 9 + 14 \cdot 11 + 18 \cdot 10}{50} = 10{,}16;$$

$$\overline{x^2} = \frac{2^2 \cdot 9 + 6^2 \cdot 11 + 10^2 \cdot 9 + 14^2 \cdot 11 + 18^2 \cdot 10}{50} = 134{,}56;$$

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 134{,}56 - 10{,}16^2 = 31{,}334.$$

Составим закон распределения случайной величины Y:

Y	11,5	14,5	17,5	20,5	23,5
n	13	13	14	8	2

По данному распределению вычислим средние характеристики:

$$\bar{y} = \frac{11{,}5 \cdot 13 + 14{,}5 \cdot 13 + 17{,}5 \cdot 14 + 20{,}5 \cdot 8 + 23{,}5 \cdot 2}{50} = 15{,}88;$$

$$\overline{y^2} = \frac{11{,}5^2 \cdot 13 + 14{,}5^2 \cdot 13 + 17{,}5^2 \cdot 14 + 20{,}5^2 \cdot 8 + 23{,}5^2 \cdot 2}{50} = 264{,}13;$$

$$S_y^2 = \overline{y^2} - \bar{y}^2 = 264{,}13 - 15{,}88^2 = 11{,}956.$$

Найдем выборочный коэффициент ковариации:

$$\mu_{yx} = \mu_{xy} = \mu = \overline{xy} - \bar{x} \cdot \bar{y} = 179{,}12 - 10{,}16 \cdot 15{,}88 = 17{,}779.$$

Коэффициенты регрессии будут равны

$$b_{yx} = \frac{\mu}{S_x^2} = \frac{17{,}779}{31{,}334} = 0{,}567; \;\; b_{xy} = \frac{\mu}{S_y^2} = \frac{17{,}779}{11{,}956} = 1{,}487.$$

Уравнения регрессии:

$$\bar{y}_x - \bar{y} = b_{yx} \cdot (x - \bar{x});$$

$$\bar{y}_x - 15{,}88 = 0{,}567 \cdot (x - 10{,}16); \qquad \bar{y}_x = 10{,}119 + 0{,}567 \cdot x;$$

$$\bar{x}_y - \bar{x} = b_{xy} \cdot (y - \bar{y});$$

$$\bar{x}_y - 10{,}16 = 1{,}487 \cdot (y - 15{,}88); \qquad \bar{x}_y = -13{,}454 + 1{,}487 \cdot y.$$

Коэффициент корреляции найдем по формуле:

$$r = \frac{\mu}{S_x \cdot S_y} = \frac{17{,}779}{\sqrt{31{,}334} \cdot \sqrt{11{,}956}} = 0{,}92.$$

Проверим гипотезу о равенстве генерального коэффициента корреляции нулю $H_0: \rho = 0$ при конкурирующей гипотезе $H_1: \rho \neq 0$.

Наблюдаемое значение критерия Стьюдента:

$$T_{\text{набл}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1 - r^2}} = \frac{0{,}92 \cdot \sqrt{50 - 2}}{\sqrt{1 - 0{,}92^2}} = 16{,}26.$$

Число степеней свободы $k = n - 2 = 50 - 2 = 48$. По таблице приложения 7 найдем $t_{\text{кр.п}} = t_{\text{кр.п}}(k = 48, \alpha = 0{,}05) = 2{,}02$. Критическая область примет вид $\omega = (-\infty; -2{,}02) \cup (2{,}02; +\infty)$. Поскольку $T_{\text{набл}} \in \omega$ нулевую гипотезу отвергаем. Следовательно, генеральный коэффициент корреляции ρ значимо отличается от нуля.

Ответ $\bar{y}_x = 10{,}119 + 0{,}567 \cdot x$; $\bar{x}_y = -13{,}454 + 1{,}487 \cdot y$; $r = 0{,}92$; коэффициент корреляции ρ значимо отличается от нуля.

6.6. Выборочное корреляционное отношение и индекс корреляции

Введенный выше коэффициент корреляции, как уже отмечалось, является полноценным показателем тесноты связи лишь в случае линейной зависимости между переменными. Однако часто возникает необходимость в достоверном показателе интенсивности связи при любой форме зависимости.

Для получения такого показателя вспомним правило сложения дисперсий:

$$S_y^2 = {S'_y}^2 + \delta_y^2,$$

где S_y^2 – общая дисперсия переменной, ${S'_y}^2$ – внутригрупповая дисперсия, δ_y^2 – межгрупповая дисперсия.

Внутригрупповая дисперсия характеризует ту часть вариации Y, которая возникает из-за изменчивости неучтенных факторов, не зависящих от X. Межгрупповая дисперсия выражает ту часть вариации Y, которая обусловлена изменчивостью X.

Выборочным корреляционным отношением Y к X называют отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y:

$$\eta_{yx} = \frac{\delta_y}{S_y}$$

где:

$$\delta_y = \sqrt{\frac{1}{n}\sum_{i=1}^{m_x} n_{xi} \cdot (\bar{y}_{xi} - \bar{y})^2}\,; \qquad S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{m_y} n_{yj} \cdot \left(y_j - \bar{y}\right)^2}.$$

n – объем выборки, n_{xi} – частота значения признака X, n_{yj} – частота значения признака Y, $\bar{y}$ – общая средняя признака Y, $\bar{y}_{xi}$ – групповая средняя признака Y.

Аналогично определяется выборочное корреляционное отношение X к Y:

$$\eta_{xy} = \frac{\delta_x}{S_x}.$$

Отметим основные свойства корреляционного отношения. Поскольку η_{xy} обладает теми же свойства, что и η_{yx}, перечислим свойства только выборочного корреляционного отношения η_{yx}.

1.Корреляционное отношение удовлетворяет двойному неравенству:

$$0 \leq \eta_{yx} \leq 1.$$

2.Если $\eta_{yx} = 0$, то переменная y с переменной x корреляционной зависимостью не связана. Отметим, что справедливо и обратное предположение: если переменная y не связана с переменной x корреляционной зависимостью, то $\eta_{yx} = 0$.

3.Если $\eta_{yx} = 1$, то переменная y с переменной x связана функциональной зависимостью. Отметим, что справедливо и обратное предположение: если переменная y связана с переменной x функциональной зависимостью, то $\eta_{yx} = 1$.

4.Выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции: $\eta_{yx} \geq |r|$.

5.Если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.

Эмпирическое корреляционное отношение η_{yx} является показателем рассеяния точек корреляционного поля относительно эмпирической линии регрессии, выражаемой ломаной, соединяющей значения $\bar{y}_x$. Однако в связи с тем, что закономерное изменение $\bar{y}_x$ нарушается случайными зигзагами ломаной, возникающими вследствие остаточного действия неучтенных факторов, η_{yx} преувеличивает тесноту связи. Поэтому наряду с η_{yx} рассматривается показатель тесноты связи R_{yx}, характеризующий рассеяние точек корреляционного поля относительно теоретической линии регрессии $\bar{y}_x$. Показатель R_{yx} получил название теоретического корреляционного отношения или индекса корреляции Y по X. Рассчитывается R_{yx} таким же образом, как и η_{yx}:

$$R_{yx} = \frac{\delta_y}{S_y}$$

но при подсчете межгруппового среднего квадратического отклонения δ_y, вместо групповой средней $\bar{y}_{xi}$ используется условная средняя $\bar{y}_x$, вычисленная по уравнению регрессии Y на X.

Аналогично вводится индекс корреляции X на Y:

$$R_{xy} = \frac{\delta_x}{S_x}.$$

Достоинством рассмотренных показателей η_{yx} и R_{xy} является то, что они могут быть вычислены при любой форме связи между переменными. Хотя η_{yx} и завышает тесноту связи по сравнению с R_{xy}, но для его вычисления не нужно знать уравнение регрессии. Корреляционные отношения η_{yx} и R_{xy} связаны с коэффициентом корреляции r следующим неравенством:

$$0 \leq |r| \leq R_{yx} \leq \eta_{yx} \leq 1.$$

Так же отметим, что в случае линейной модели регрессии, т.е. зависимости $\bar{y}_x - \bar{y} = b_{yx} \cdot (x - \bar{x})$, индекс корреляции R_{yx} равен коэффициенту корреляции r по абсолютной величине $R_{yx} = |r|$ (или $R_{xy} = |r|$).

Коэффициент детерминации R^2_{yx}, равный квадрату индекса корреляции, показывает долю общей вариации зависимой переменной, обусловленной регрессией или изменчивостью объясняющей переменной.

Чем ближе R^2_{yx} к единице, тем лучше регрессия аппроксимирует эмпирические данные, тем теснее эмпирические точки примыкают к линии регрессии. Если $R^2_{yx} = 1$, то эмпирические точки $(x; y)$ лежат на линии регрессии и между переменными Y и X существует линейная функциональная зависимость. Если $R^2_{yx} = 0$, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных, и линия регрессии параллельна оси абсцисс.

Расхождение между η^2_{yx} и R^2_{yx} может быть использовано для проверки линейности корреляционной зависимости.

Рассмотрим проверку гипотезы о значимости корреляционного отношения η_{yx}. Сформулируем нулевую гипотезу: генеральное корреляционное отношение равно нулю $H_0\colon \eta_{yx} = 0$. Альтернативную гипотезу сформулируем в виде: генеральное корреляционное отношение значимо отличается от нуля $H_1\colon \eta_{yx} \neq 0$.

Критерием для проверки нулевой гипотезы случит величина:

$$F=\frac{\eta_{yx}^2(n-m)}{\left(1-\eta_{yx}^2\right)(m-1)},$$

которая имеет распределение Фишера-Снедекора с $k_1=m-1$ и $k_2=n-m$ степенями свободы, где m – число интервалов по группировочному признаку.

При данной конкурирующей гипотезе строят двустороннюю критическую область $\omega=\left(0;F_{\text{кр.л}}\right)\cup\left(F_{\text{кр.л}};+\infty\right)$. Критические точки области находят из условий:

$$P\left(0<F<F_{\text{кр.л}}\right)=\frac{\alpha}{2};\qquad P\left(F_{\text{кр.п}}<F<+\infty\right)=\frac{\alpha}{2},$$

по таблице приложения 4.

Гипотеза о значимости индекса корреляции R_{yx} проверяется аналогично гипотезе о значимости корреляционного отношения η_{yx} за тем исключением, что наблюдаемое значение критерия вычисляется по формуле:

$$F=\frac{R_{yx}^2(n-2)}{\left(1-R_{yx}^2\right)},$$

а числа степеней свободы определяются выражениями $k_1=1$ и $k_2=n-2$.

Пример 6.6.1. Дана корреляционная таблица:

Y \ X	0	2	4	6	8
10	9	4			
14	5	10	10		
18	2	5	5	7	5
22			5	10	15
26				3	5

Найти:

1) групповые средние $\bar{y}_x$ и $\bar{x}_y$ и построить их графики;

2) выборочные корреляционные отношения η_{yx} и η_{xy};

3) уравнения прямых регрессии Y на X и X на Y и построить их графики;

4) коэффициент корреляции r;

5) индексы корреляции R_{yx} и R_{xy};

6) коэффициент детерминации R^2.

1. Групповые средние $\bar{y}_x$:

$$X=0: \qquad \bar{y}_x = \frac{10\cdot 9 + 14\cdot 5 + 18\cdot 2}{9+5+2} = 12{,}25;$$
$$X=2: \qquad \bar{y}_x = \frac{10\cdot 4 + 14\cdot 10 + 18\cdot 5}{4+10+5} = 14{,}21;$$
$$X=4: \qquad \bar{y}_x = \frac{14\cdot 10 + 18\cdot 5 + 22\cdot 5}{10+5+5} = 17{,}00;$$
$$X=6: \qquad \bar{y}_x = \frac{18\cdot 7 + 22\cdot 10 + 26\cdot 3}{7+10+3} = 21{,}20;$$
$$X=8: \qquad \bar{y}_x = \frac{18\cdot 5 + 22\cdot 15 + 26\cdot 5}{5+15+5} = 22{,}00.$$

Групповые средние $\bar{x}_y$:

$$Y=10: \qquad \bar{x}_y = \frac{0\cdot 9 + 2\cdot 4}{9+4} = 0{,}62;$$
$$Y=14: \qquad \bar{x}_y = \frac{0\cdot 5 + 2\cdot 10 + 4\cdot 10}{5+10+10} = 2{,}40;$$
$$Y=18: \qquad \bar{x}_y = \frac{0\cdot 2 + 2\cdot 5 + 4\cdot 5 + 6\cdot 7 + 8\cdot 5}{2+5+5+7+5} = 4{,}67;$$
$$Y=22: \qquad \bar{x}_y = \frac{4\cdot 5 + 6\cdot 10 + 8\cdot 15}{5+10+15} = 6{,}67;$$
$$Y=26: \qquad \bar{x}_y = \frac{6\cdot 3 + 8\cdot 5}{3+5} = 7{,}25.$$

Средние значения $\bar{x}$ и $\bar{y}$:

$$\bar{y} = \frac{10\cdot 13 + 14\cdot 25 + 18\cdot 24 + 22\cdot 30 + 26\cdot 8}{100} = 17{,}80,$$
$$\bar{x} = \frac{0\cdot 16 + 2\cdot 19 + 4\cdot 20 + 6\cdot 20 + 8\cdot 25}{100} = 4{,}38.$$

Графики групповых средних (эмпирических линий регрессии) представлены на рисунке 10 и на рисунке 11.

2. Для удобства дальнейших вычислений дополним корреляционную таблицу:

$Y \backslash X$	0	2	4	6	8	n_y	$\bar{x}_y$
10	9	4				13	0,62
14	5	10	10			25	2,40
18	2	5	5	7	5	24	4,67
22			5	10	15	30	6,67
26				3	5	8	7,25
n_x	16	19	20	20	25	$n=100$	
$\bar{y}_x$	12,25	14,21	17,00	21,20	22,00		

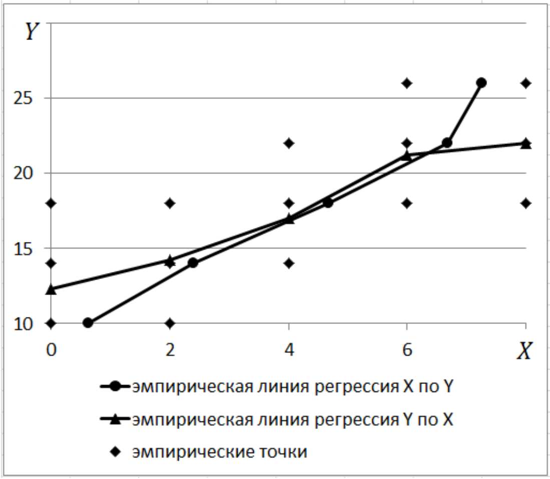

Рисунок 10. График групповых средних (эмпирических линий регрессии)

Средние значения $\overline{x^2}$ и $\overline{y^2}$:

$$\overline{y^2} = \frac{10^2 \cdot 13 + 14^2 \cdot 25 + 18^2 \cdot 24 + 22^2 \cdot 30 + 26^2 \cdot 8}{100} = 339{,}04,$$

$$\overline{x^2} = \frac{0^2 \cdot 16 + 2^2 \cdot 19 + 4^2 \cdot 20 + 6^2 \cdot 20 + 8^2 \cdot 25}{100} = 27{,}16.$$

Общие дисперсии S_y^2 и S_x^2:

$$S_y^2 = \overline{y^2} - \bar{y}^2 = 339{,}04 - 17{,}80^2 = 22{,}2;$$

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 27{,}16 - 4{,}38^2 = 7{,}98;$$

Межгрупповые дисперсии δ_y^2 и δ_x^2:

$$\delta_y^2 = \{(12{,}25 - 17{,}8)^2 \cdot 16 + (14{,}21 - 17{,}8)^2 \cdot 19 + (17 - 17{,}8)^2 \cdot 20 + \\ +(21{,}2 - 17{,}8)^2 \cdot 20 + (22 - 17{,}8)^2 \cdot 25\}/100 = 14{,}23;$$

$$\delta_x^2 = \{(0{,}62 - 4{,}38)^2 \cdot 13 + (2{,}4 - 4{,}38)^2 \cdot 25 + (4{,}67 - 4{,}38)^2 \cdot 24 + \\ +(6{,}67 - 4{,}38)^2 \cdot 30 + (7{,}25 - 4{,}38)^2 \cdot 8\}/100 = 5{,}07.$$

Средние квадратические отклонения:

$$S_y = \sqrt{22{,}2} = 4{,}71; \quad S_x = \sqrt{7{,}98} = 2{,}82;$$

$$\delta_y = \sqrt{14{,}23} = 3{,}77; \quad \delta_x = \sqrt{5{,}07} = 2{,}25.$$

Корреляционные отношения:

$$\eta_{yx} = \frac{\delta_y}{S_y} = \frac{3{,}77}{4{,}71} = 0{,}80; \quad \eta_{xy} = \frac{\delta_x}{S_x} = \frac{2{,}25}{2{,}82} = 0{,}80.$$

3. Среднее значение $\overline{xy}$:

$$\overline{xy} = \{10 \cdot (0 \cdot 9 + 2 \cdot 4) + 14 \cdot (0 \cdot 5 + 2 \cdot 10 + 4 \cdot 10) +$$

$$+18 \cdot (0 \cdot 2 + 2 \cdot 5 + 4 \cdot 5 + 6 \cdot 7 + 8 \cdot 5) + 22 \cdot (4 \cdot 5 + 6 \cdot 10 + 8 \cdot 15) + $$
$$+26 \cdot (6 \cdot 3 + 8 \cdot 5)\}/100 = 88{,}44.$$

Выборочный коэффициент ковариации:

$$\mu = \overline{xy} - \bar{x} \cdot \bar{y} = 88{,}44 - 4{,}38 \cdot 17{,}80 = 10{,}48.$$

Коэффициенты регрессии:

$$b_{yx} = \frac{\mu}{S_x^2} = \frac{10{,}48}{7{,}98} = 1{,}31; \qquad b_{xy} = \frac{\mu}{S_y^2} = \frac{10{,}48}{22{,}2} = 0{,}47.$$

Уравнения регрессий:

$$\bar{y}_x - \bar{y} = b_{yx} \cdot (x - \bar{x});$$
$$\bar{y}_x - 17{,}80 = 1{,}31 \cdot (x - 4{,}38); \qquad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x;$$
$$\bar{x}_y - \bar{x} = b_{xy} \cdot (y - \bar{y});$$
$$\bar{x}_y - 4{,}38 = 0{,}47 \cdot (y - 17{,}80); \qquad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y.$$

Графики прямых регрессии представлены на рисунке 8.

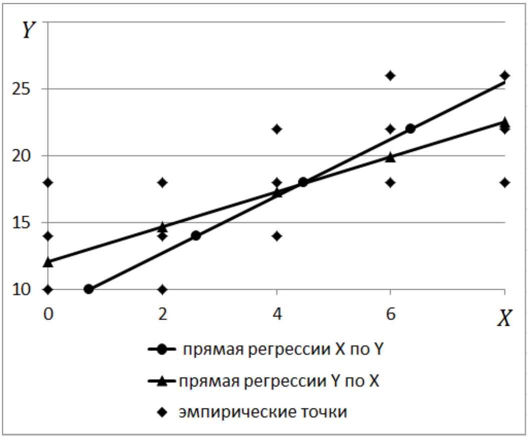

Рисунок 11. График групповых средних (эмпирических линий регрессии)

4. Коэффициент корреляции:

$$r = +\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{1{,}31 \cdot 0{,}47} = 0{,}78.$$

5. Используя, уравнения регрессий найдем условные средние $\bar{y}_x$ и $\bar{x}_y$:

$X = 0$: $\quad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x = 12{,}06 + 1{,}31 \cdot 0 = 12{,}06;$

$X = 2$: $\quad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x = 12{,}06 + 1{,}31 \cdot 2 = 14{,}68;$

$X = 4$: $\quad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x = 12{,}06 + 1{,}31 \cdot 4 = 17{,}30;$

$X = 6$: $\quad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x = 12{,}06 + 1{,}31 \cdot 6 = 19{,}92;$

$X = 8$: $\quad \bar{y}_x = 12{,}06 + 1{,}31 \cdot x = 12{,}06 + 1{,}31 \cdot 8 = 22{,}54;$

$$Y=10: \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y = -3{,}99 + 0{,}47 \cdot 10 = 0{,}71;$$
$$Y=14: \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y = -3{,}99 + 0{,}47 \cdot 14 = 2{,}59;$$
$$Y=18: \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y = -3{,}99 + 0{,}47 \cdot 18 = 4{,}47;$$
$$Y=22: \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y = -3{,}99 + 0{,}47 \cdot 22 = 6{,}35;$$
$$Y=26: \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y = -3{,}99 + 0{,}47 \cdot 26 = 8{,}23.$$

Вычислим межгрупповые дисперсии δ_y^2 и δ_x^2, используя условные средние $\bar{y}_x$ и $\bar{x}_y$, рассчитанные по уравнениям регрессии:

$$\delta_y^2 = \{(12{,}06-17{,}8)^2 \cdot 16 + (14{,}68-17{,}8)^2 \cdot 19 + (17{,}30-17{,}8)^2 \cdot 20 + \\ +(19{,}92-17{,}8)^2 \cdot 20 + (22{,}54-17{,}8)^2 \cdot 25\}/100 = 13{,}69;$$

$$\delta_x^2 = \{(0{,}71-4{,}38)^2 \cdot 13 + (2{,}59-4{,}38)^2 \cdot 25 + (4{,}47-4{,}38)^2 \cdot 24 + \\ +(6{,}35-4{,}38)^2 \cdot 30 + (8{,}23-4{,}38)^2 \cdot 8\}/100 = 4{,}90.$$

Средние квадратические отклонения:

$$\delta_y = \sqrt{13{,}69} = 3{,}7; \quad \delta_x = \sqrt{4{,}90} = 2{,}21.$$

Индексы корреляции:

$$R_{yx} = \frac{\delta_y}{S_y} = \frac{3{,}7}{4{,}71} = 0{,}78; \quad R_{xy} = \frac{\delta_x}{S_x} = \frac{2{,}21}{2{,}82} = 0{,}78.$$

Поскольку индексы корреляции 0,78 близок к корреляционному отношению 0,8, это указывает на то, что зависимость между переменными является линейной.

Отметим так же, что данный пример позволил проверить, что в случае линейной корреляционной зависимости выполняется равенство $R = r$, т.е. индекс корреляции равен коэффициенту корреляции.

6. Коэффициент детерминации $R^2 = 0{,}78^2 = 0{,}61$. Эта величина показывает, что вариация зависимой переменной y на 61% объясняется вариацией независимой переменной x.

Ответ 1)

X	0	2	4	6	8
$\bar{y}_x$	12,25	14,21	17,00	21,20	22,00

$\bar{x}_y$	0,62	2,40	4,67	6,67	7,25
Y	10	14	18	22	26

2) $\eta_{yx} = 0{,}80; \quad \eta_{xy} = 0{,}80;$

3) $\bar{y}_x = 12{,}06 + 1{,}31 \cdot x; \quad \bar{x}_y = -3{,}99 + 0{,}47 \cdot y;$

4) $r = 0{,}78;$

5) $R_{yx} = 0{,}78; \quad R_{xy} = 0{,}78;$

6) $R^2 = 0{,}61.$

6.7. Ранговая корреляция

На практике часто встречаются с необходимостью изучения связи между порядковыми переменными, измеренными в так называемой порядковой шкале. В этой шкале можно установить лишь порядок, в котором объекты выстраиваются по степени проявления признака (например, качество жилищных условий, тестовые баллы, экзаменационные оценки и т.п.). Например, если по некоторой дисциплине два студента имеют оценки «отлично» и «удовлетворительно», то можно лишь утверждать, что уровень подготовки по этой дисциплине первого студента выше, чем второго, но нельзя сказать, на сколько или во сколько раз больше.

В таких случаях проблема оценки тесноты связи разрешима, если упорядочить, или ранжировать объекты анализа по степени выраженности измеряемых признаков. При этом каждому объекту присваивается определенный номер, называемый рангом. Объекту с наименьшим проявлением признака присваивается ранг 1, следующему за ним присваивается ранг 2 и т.д. Объекты можно располагать и в порядке убывания проявления признака. Если объекты ранжированы по двум признакам, то имеется возможность оценить тесноту связи между признаками, основываясь на рангах, т.е. тесноту ранговой корреляции.

Пусть проведена выборка из n – объектов. Каждый объект характеризуется двумя признаками A и B. Расположим сначала объекты выборки в порядке ухудшения (улучшения) качества по признаку A. Припишем объекту, стоящему на -м месте, число – ранг x_i (например, x_i может быть равно порядковому номеру). Проделаем такую же операцию по признаку B. Присвоим каждому объекту число – ранг y_i по признаку B. В результате сформируем таблицу:

Номер объекта	1	2	….	n
Ранг по признаку A	x_1	x_2	….	x_n
Ранг по признаку B	y_1	y_2	….	y_n

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$\rho = 1 - \frac{6}{n^3 - n} \cdot \sum_{i=1}^{n} (x_i - y_i)^2.$$

Отметим свойства коэффициента ранговой корреляции Спирмена:

1. Если между качественными признаками A и B имеется полная прямая зависимость в том смысле, что ранги объектов совпадают при всех значениях i, то выборочный коэффициент ранговой корреляции Спирмена равен единице.

2. Если между качественными признаками A и B имеется противоположная зависимость в том смысле, что рангу $x_1 = 1$ соответствует ранг $y_1 = n$; рангу $x_2 = 2$ соответствует ранг $y_2 = n - 1$; рангу $x_n = n$ соответствует ранг $y_n = 1$ то выборочный коэффициент ранговой корреляции Спирмена равен минус единице.

3. Если между качественными признаками A и B нет ни полной прямой, ни противоположной зависимостей, то коэффициент ρ заключен между минус единицей и единицей $-1 < \rho < 1$, причем чем ближе к нулю его абсолютная величина, тем зависимость меньше.

При ранжировании иногда сталкиваются со случаями, когда невозможно найти существенные различия между объектами по величине проявления рассматриваемого признака. Объекты, как говорят, оказываются связанными. Связанным объектам приписывают одинаковые средние ранги, такие, чтобы сумма всех рангов оставалась такой же, как и при отсутствии связанных рангов. Например, если четыре объекта оказались равнозначными в отношении рассматриваемого признака и невозможно определить, какие из четырех рангов 4, 5, 6, 7 приписать этим объектам, то каждому объекту приписывается средний ранг, равный (4+5+6+7)/4=5,5.

При наличии связанных рангов ранговый коэффициент корреляции Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{1}{\frac{1}{6}(n^3 - n) - \left(T_x - T_y\right)} \cdot \sum_{i=1}^{n} (x_i - y_i)^2 \,;$$

где:

$$T_x = \frac{1}{12} \sum_{i=1}^{m_x} (t_x^3 - t_x)\,; \qquad T_y = \frac{1}{12} \sum_{i=1}^{m_y} \left(t_y^3 - t_y\right);$$

m_x, m_y – число групп неразличимых рангов по признакам A и B, t_x, t_y – число рангов, входящих в группу неразличимых рангов по признакам A и B.

При проверке значимости коэффициента корреляции Спирмена ρ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными при $n > 10$ статистика:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

имеет t – распределение Стьюдента с $k = n - 2$ степенями свободы.

Пример 6.7.1 По результатам тестирования 10 студентов по двум дисциплинам A и B на основе набранных баллов получены следующие ранги:

Номер студента	1	2	3	4	5	6	7	8	9	10
Ранг по дисциплине A	2	4	5	1	6	8,5	7	8,5	3	10
Ранг по дисциплине B	2	5,5	2	2	5,5	8	5,5	5,5	9	10

Вычислить ранговый коэффициент Спирмена и проверить его значимость при уровне значимости $\alpha = 0{,}05$.

По дисциплине A имеем одну группу неразличимых рангов $m_x = 1$ с числом рангов, входящих в группу $t_x = 2$. Для дисциплины B имеем $m_y = 2$, $t_y = 4$ (для рангов 5,5) и $t_y = 3$ (для рангов 2). Следовательно:

$$T_x = \frac{1}{12}(2^3 - 2) = 0{,}5; \quad T_y = \frac{1}{12}(4^3 - 4) + \frac{1}{12}(3^3 - 3) = 5 + 2 = 7.$$

$$\sum_{i=1}^{n}(x_i - y_i)^2 = (2-2)^2 + (4-5{,}5)^2 + (5-2)^2 + (1-2)^2 +$$
$$+(6-5{,}5)^2 + (8{,}5-8)^2 + (7-5{,}5)^2 + (8{,}5-5{,}5)^2 +$$
$$+(3-9)^2 + (10-10)^2 = 60.$$

Коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{60}{\frac{1}{6}(10^3 - 10) - (0{,}5 - 7)} = 0{,}65.$$

Проверим значимость коэффициента Спирмена. Нулевую гипотезу сформулируем $H_0: \rho = 0$, конкурирующую $H_0: \rho \neq 0$. Наблюдаемое значение статистики:

$$t_{\text{набл}} = \frac{0{,}65 \cdot \sqrt{10-2}}{\sqrt{1 - 0{,}65^2}} = 2{,}42.$$

По таблице приложения 7 при числе степеней свободы $k = n - 2 = 8$ и уровне значимости $\alpha = 0{,}05$ для двусторонней критической области найдем $t_{\text{кр}} = 2{,}31$. Критическая область $\omega = (-\infty; -2{,}31) \cup (2{,}31; +\infty)$. Поскольку $t_{\text{набл}} \in \omega$, нулевую гипотезу отклоняем. Значит коэффициент корреляции значимо отличается от нуля, а между успеваемостью по дисциплинам A и B имеется взаимосвязь.

Ответ $\rho = 0{,}65$, ρ значимо отличается от нуля.

Можно оценивать связь между двумя качественными признаками, используя коэффициент ранговой корреляции Кенделла. Для этого нужно ранжировать объекты по одной переменной x в порядке возрастания рангов. Далее для каждого ранга y_i нужно определить количество рангов y_i превосходящих y_i и располагающихся правее от y_i. Т.е. R_1 число рангов, стоящих справа от y_1 и превосходящих y_1; R_2 число рангов, стоящих справа от y_2 и превосходящих y_2; …; R_{n-1} число рангов, стоящих справа от y_{n-1} и превосходящих y_{n-1}.

Выборочный коэффициент ранговой корреляции Кендалла определяется формулой:

$$\tau = \frac{4R}{n \cdot (n-1)} - 1,$$

где:

$$R = \sum_{i=1}^{n-1} R_i.$$

Ранговый коэффициент Кенделла имеет те же свойства, что и ранговый коэффициент Спирмена.

Сравнивая коэффициенты ранговой корреляции ρ Спирмена и τ Кенделла, можно отметить, что хотя вычисление τ более трудоемко, коэффициент τ обладает некоторыми преимуществами перед ρ при исследовании его статистических свойств и большим удобством его пересчета при добавлении к n статистически обследованным объектам новых, т.е. при удлинении анализируемых ранжировок.

Заметим, что при умеренно больших значениях n $(n > 10)$ и условии, что абсолютные величины значений этих коэффициентов не слишком близки к единице, их связывает простое приближенное соотношение $\rho = 3\tau/2$.

Для проверки значимости коэффициента корреляции Кенделла τ проверяют нулевую гипотезу $H_0\colon \tau = 0$ при конкурирующей гипотезе $H_1\colon \tau \neq 0$. Критическая область имеет вид $\omega = \left(-\infty;\ -T_{\text{кр}}\right) \cup \left(T_{\text{кр}};+\infty\right)$. Где $T_{\text{кр}}$ вычисляют по формуле:

$$T_{\text{кр}} = z_{\text{кр}} \cdot \sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

где n – объем выборки, а $z_{\text{кр}}$ – критическая точка двусторонней критической области, которую находят по равенству

$$\phi(z_{\text{кр}}) = \frac{1-\alpha}{2}.$$

Пример 6.7.2 Найти ранговые коэффициенты корреляции Спирмена и Кенделла по данным рангам выборки объема $n = 10$. Проверить значимость коэффициентов корреляции при уровне значимости $\alpha = 0{,}05$.

Ранг по признаку A	1	2	3	4	5	6	7	8	9	10
Ранг по признаку B	8	9	10	6	7	5	4	3	1	2

Для решения задачи составим вспомогательную таблицу:

Ранг по признаку A	1	2	3	4	5	6	7	8	9	10
Ранг по признаку B	8	9	10	6	7	5	4	3	1	2
$x_i - y_i$	7	7	7	2	2	1	3	5	8	8
R_i	2	1	0	1	0	0	0	0	1	

Найдем:

$$\sum_{i=1}^{n}(x_i - y_i)^2 = (-7)^2 + (-7)^2 + (-7)^2 + (-2)^2 + (-2)^2 +$$

$$+(1)^2 + (3)^2 + (5)^2 + (8)^2 + (8)^2 = 318.$$

Коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6}{10^3 - 10} \cdot 318 = -0{,}93.$$

Проверим гипотезу о значимости коэффициента корреляции Спирмена. Нулевая гипотеза H_0: $\rho = 0$. Конкурирующая гипотеза H_0: $\rho \neq 0$. Наблюдаемое значение статистики:

$$t_{\text{набл}} = \frac{-0{,}93 \cdot \sqrt{10-2}}{\sqrt{1-(-0{,}93)^2}} = -7{,}16.$$

По таблице приложения 7 при числе степеней свободы $k = n - 2 = 8$ и уровне значимости $\alpha = 0{,}05$ для двусторонней критической области найдем $t_{\text{кр}} = 2{,}31$. Критическая область $\omega = (-\infty; -2{,}31) \cup (2{,}31; +\infty)$. Поскольку $t_{\text{набл}} \in \omega$, нулевую гипотезу отклоняем. Значит коэффициент корреляции Спирмена значимо отличается от нуля.

Найдем:

$$R = \sum_{i=1}^{n-1} R_i = 2 + 1 + 0 + 1 + 0 + 0 + 0 + 0 + 1 = 5.$$

Коэффициент корреляции Кендалла:

$$\tau = \frac{4R}{n \cdot (n-1)} - 1 = \frac{4 \cdot 5}{10 \cdot (10-1)} - 1 = -0{,}78.$$

Проверим гипотезу о значимости коэффициента корреляции Кендалла. Нулевая гипотеза $H_0: \tau = 0$. Конкурирующая гипотеза $H_0: \tau \neq 0$. Из соотношения:

$$\phi(z_{\text{кр}}) = \frac{1-\alpha}{2} = \frac{1-0{,}05}{2} = 0{,}475$$

по таблице приложения 1 найдем $z_{\text{кр}} = 1{,}96$. Найдем критическую точку:

$$T_{\text{кр}} = z_{\text{кр}} \cdot \sqrt{\frac{2(2n+5)}{9n(n-1)}} = 1{,}96 \cdot \sqrt{\frac{2(2 \cdot 10 + 5)}{9 \cdot 10 \cdot (10-1)}} = 0{,}49.$$

Критическая область $\omega = (-\infty; -0{,}49) \cup (0{,}49; +\infty)$. Поскольку $\tau \in \omega$, нулевую гипотезу отклоняем. Значит коэффициент корреляции Кендалла значимо отличается от нуля.

Ответ $\rho = -0{,}93, \tau = -0{,}78$, оба коэффициента значимо отличаются от нуля.

В практике статистических исследований встречаются случаи, когда совокупность объектов характеризуется не двумя, а несколькими последовательностями рангов и необходимо установить статистическую связь между несколькими переменными. Такие задачи возникают при анализе экспертных оценок, когда необходимо установить меру их согласованности.

В качестве такого измерителя используют коэффициент конкордации рангов Кендалла:

$$W = \frac{12}{m^2(n^3 - n)} \cdot \sum_{i=1}^{n} D_i^2,$$

где n – число объектов, m – число анализируемых порядковых переменных,

$$D_i = \sum_{j=1}^{m} r_{ij} - \frac{m(n+1)}{2}.$$

Коэффициент конкордации лежит на промежутке $0 \leq W \leq 1$. При совпадении всех ранжировок $W = 1$.

Проверка значимости коэффициента конкордации W основана на том, что в случае справедливости нулевой гипотезы $H_0\colon W = 0$ об отсутствии корреляционной связи при $n > 7$ статистика:

$$\chi^2 = m(n-1)W,$$

имеет приближенно χ^2 – распределение с $k = n - 1$ степенями свободы. Конкурирующую гипотезу сформулируем $H_1\colon W > 0$. Критическая область будет односторонней $\omega = \left(\chi^2_{\text{кр}}(k;\alpha);+\infty\right)$, где $\chi^2_{\text{кр}}(k;\alpha)$ находится по таблице приложения 5.

Пример 6.7.3 Группа из 4 экспертов оценивает качество изделий, изготовленных на 6 предприятиях. Их предпочтения представлены в таблице:

	Номер предприятия							
	1	2	3	4	5	6	7	8
1-й эксперт	1	2	4	3	6	5	7	8
2-й эксперт	1	3	6	2	4	5	8	7
3-й эксперт	1	2	5	3	4	6	7	8
4-й эксперт	2	1	6	5	4	3	8	7

Вычислить коэффициент конкордации рангов и оценить его значимость на уровне значимости $\alpha = 0{,}05$.

Для данной задачи $n = 8$, $m = 4$. Вычислим величины D_i:

$$D_1 = \sum_{j=1}^{4} r_{1j} - \frac{4(8+1)}{2} = 1 + 1 + 1 + 2 - 18 = -13;$$

$$D_2 = \sum_{j=1}^{4} r_{2j} - \frac{4(8+1)}{2} = 2 + 3 + 2 + 1 - 18 = -10;$$

$$D_3 = \sum_{j=1}^{4} r_{3j} - \frac{4(8+1)}{2} = 4 + 6 + 5 + 6 - 18 = 3;$$

$$D_4 = \sum_{j=1}^{4} r_{4j} - \frac{4(8+1)}{2} = 3 + 2 + 3 + 5 - 18 = -5;$$

$$D_5 = \sum_{j=1}^{4} r_{5j} - \frac{4(8+1)}{2} = 6 + 4 + 4 + 4 - 18 = 0;$$

$$D_6 = \sum_{j=1}^{4} r_{6j} - \frac{4(8+1)}{2} = 5 + 5 + 6 + 3 - 18 = 1;$$

$$D_7=\sum_{j=1}^{4} r_{7j}-\frac{4(8+1)}{2}=7+8+7+8-18=12;$$

$$D_8=\sum_{j=1}^{4} r_{8j}-\frac{4(8+1)}{2}=8+7+8+7-18=12.$$

Коэффициент конкордации рангов Кендалла:

$$W=\frac{12}{4^2(8^3-8)}\cdot((-13)^2+(-10)^2+3^2+(-5)^2+0^2+$$
$$+1^2+12^2+12^2)=0{,}88.$$

Наблюдаемое значение статистики:

$$\chi^2_{\text{набл}}=4(8-1)\cdot 0{,}88=24{,}64.$$

При числе степеней свободы $k=n-1=8-1=7$ и уровне значимости $\alpha=0{,}05$ по приложению 5 найдем критическое значение критерия Хи-квадрат $\chi^2_{\text{кр}}(k;\alpha)=14{,}1$. Критическая область имеет вид $\omega=(14{,}1;+\infty)$. Поскольку $\chi^2_{\text{набл}}\in\omega$ нулевую гипотезу отклоняем. Следовательно, коэффициент конкордации Кенделла значимо отличается от нуля.

Ответ $W=0{,}88$, мнения экспертов согласованы.

Глава 7. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Пусть генеральные совокупности $X_1, X_2, \dots, X_p$ распределены нормально и имеют одинаковую, хотя и неизвестную, дисперсию. Математические ожидания также неизвестны, но могут быть различными. Требуется при заданном уровне значимости по выборочным средним проверить нулевую гипотезу $H_0: M(X_1) = M(X_2) = \dots = M(X_p)$ о равенстве всех математических ожиданий. Другими словами, требуется установить, значимо или незначимо различаются выборочные средние. Казалось бы, для сравнения нескольких средних $(p > 2)$ можно сравнить их попарно. Однако с возрастанием числа средних возрастает и наибольшее различие между ними: среднее новой выборки может оказаться больше наибольшего или меньше наименьшего из средних, полученных до нового опыта. По этой причине для сравнения нескольких средних пользуются другим методом, который основан на сравнении дисперсий и поэтому назван дисперсионным анализом.

На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет p уровней $F_1, F_2, \dots, F_p$ на изучаемую величину X. Например, если требуется выяснить, какой вид удобрений наиболее эффективен для получения наибольшего урожая, то фактор F – удобрение, а его уровни – виды удобрений.

Основная идея дисперсионного анализа состоит в сравнении факторной дисперсии, порождаемой воздействием фактора, и остаточной дисперсии, обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на X. В этом случае среднее наблюдаемых значений на каждом уровне различаются также значимо.

Если уже установлено, что фактор существенно влияет на X, а требуется выяснить, какой из уровней оказывает наибольшее воздействие, то дополнительно производят попарное сравнение средних.

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей. Если дисперсии этих совокупностей одинаковы по предположению, а дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны. Однородные же совокупности можно объединить в одну и тем самым получить о ней более полную информацию, следовательно, и более надежные выводы.

7.1. Общая, факторная и остаточная суммы квадратов отклонений

Пусть на количественный нормально распределенный признак X воздействует фактор F, который имеет p постоянных уровней. Будем предполагать, что число наблюдений (испытаний) на каждом уровне одинаково и равно q.

Пусть наблюдалось $n = p \cdot q$ значений x_{ij} признака X, где i – номер испытания $(i = 1, 2, \dots, q)$, j – номер уровня фактора $(j = 1, 2, \dots, p)$. Результаты наблюдений удобно располагать в таблице:

Номер испытания	Уровень фактора F_j			
	F_1	F_2	$\dots$	F_p
1	x_{11}	x_{12}	$\dots$	x_{1p}
2	x_{21}	x_{22}	$\dots$	x_{2p}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
q	x_{q1}	x_{q2}	$\dots$	x_{qp}
Групповая средняя	$\bar{x}_{\text{гр}1}$	$\bar{x}_{\text{гр}2}$	$\dots$	$\bar{x}_{\text{гр}p}$

Введем величины: общая сумма квадратов отклонений наблюдаемых значений от общей средней $\bar{x}$:

$$s_{\text{общ}} = \sum_{j=1}^{p} \sum_{i=1}^{q} \left(x_{ij} - \bar{x}\right)^2;$$

факторная сумма квадратов отклонений групповых средних от общей средней, которая характеризует рассеяние между группами:

$$s_{\text{факт}} = q \cdot \sum_{j=1}^{p} \left(\bar{x}_{\text{гр}j} - \bar{x}\right)^2;$$

Остаточная сумма квадратов отклонений наблюдаемых значений группы от своей групповой средней, которая характеризует рассеяние внутри групп:

$$s_{\text{ост}} = \sum_{i=1}^{q} \left(x_{i1} - \bar{x}_{\text{гр}1}\right)^2 + \sum_{i=1}^{q} \left(x_{i2} - \bar{x}_{\text{гр}2}\right)^2 + \dots + \sum_{i=1}^{q} \left(x_{ip} - \bar{x}_{\text{гр}p}\right)^2.$$

Эти суммы удовлетворяют равенству:

$$s_{\text{общ}} = s_{\text{факт}} + s_{\text{ост}}.$$

Отметим, что для практических расчетов удобнее использовать следующие формулы:

$$P_j = \sum_{i=1}^{q} x_{ij}^2, \quad R_j = \sum_{i=1}^{q} x_{ij},$$

$$s_{\text{общ}} = \sum_{j=1}^{p} P_j - \frac{1}{p \cdot q}\left(\sum_{j=1}^{p} R_j\right)^2, \quad s_{\text{факт}} = \frac{1}{q}\sum_{j=1}^{p} R_j^2 - \frac{1}{p \cdot q}\left(\sum_{j=1}^{p} R_j\right)^2,$$

$$s_{\text{ост}} = s_{\text{общ}} - s_{\text{факт}}.$$

7.2. Общая, факторная и остаточная дисперсии

Чтобы найти несмещенные оценки дисперсий, найдем числа степеней свободы дисперсий. Число степеней общей дисперсии – это объем выборки минус один $k_{\text{общ}} = n - 1 = p \cdot q - 1$. Число степеней свободы факторной дисперсии $k_{\text{факт}} = p - 1$. Число степеней свободы остаточной дисперсии найдем как разность:

$$k_{\text{ост}} = k_{\text{общ}} - k_{\text{факт}} = p \cdot q - 1 - (p - 1) = p \cdot q - p = p \cdot (q - 1).$$

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$S_{\text{общ}}^2 = \frac{s_{\text{общ}}}{p \cdot q - 1}, \qquad S_{\text{факт}}^2 = \frac{s_{\text{факт}}}{p - 1}, \qquad S_{\text{ост}}^2 = \frac{s_{\text{ост}}}{p \cdot (q - 1)},$$

где p – число уровней фактора, q – число наблюдений на каждом уровне фактора.

7.3. Сравнение нескольких средних методом дисперсионного анализа

Вернемся к задаче сравнения нескольких $(p > 2)$ средних нормальных совокупностей с неизвестными, но одинаковыми дисперсиями. Решение этой задачи сводится к сравнению факторной и остаточной дисперсий по критерию Фишера-Снедекора:

1.Пусть нулевая гипотеза $H_0: M(X_1) = M(X_2) = \cdots = M(X_p)$ о равенстве нескольких средних правильна. В этом случае факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Если сравнить эти оценки по критерию F, то очевидно, критерий укажет, что нулевую гипотезу о равенстве факторной и остаточной дисперсий следует принять.

Таким образом, если гипотеза о равенстве групповых средних правильна, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2.Пусть нулевая гипотеза о равенстве групповых средних ложна. В этом случае с возрастанием расхождения между групповыми средними увеличивается факторная дисперсия, а вместе с ней и отношение:

$$F_{\text{набл}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}.$$

В итоге $F_{\text{набл}}$ окажется больше $F_{\text{кр}}$ и, следовательно, гипотеза о равенстве дисперсий будет отвергнута.

Таким образом, если гипотеза о равенстве групповых средних ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Итак, для того чтобы проверить нулевую гипотезу о равенстве групповых средних нормальных совокупностей с одинаковыми дисперсиями, достаточно проверить по критерию F нулевую гипотезу о равенстве факторной и остаточной дисперсий. В этом и состоит метод дисперсионного анализа.

Отметим, что если факторная дисперсия окажется меньше остаточной, то уже отсюда следует справедливость гипотезы о равенстве групповых средних и, значит, нет надобности прибегать к критерию F.

Пример 7.3.1 Имеются четыре партии сырья для текстильной промышленности. Из каждой партии отобрано по пять образцов и проведены испытания на определение величины разрывной нагрузки. Результаты испытаний приведены в таблице:

Номер испытания	Номер партии (уровни фактора)			
	№1 (F_1)	№2 (F_2)	№3 (F_3)	№4 (F_4)
1	200	190	230	150
2	140	150	190	170
3	170	210	200	150
4	145	150	190	170
5	165	150	200	180

При уровне значимости $\alpha = 0{,}05$ выяснить существенно ли влияние различных партий сырья на величину разрывной нагрузки.

В данном случае $p = 4$, $q = 5$. Вычислим суммы:

$$P_1 = \sum_{i=1}^{5} x_{i1}^2 = 200^2 + 140^2 + 170^2 + 145^2 + 165^2 = 136750;$$

$$R_1 = \sum_{i=1}^{5} x_{i1} = 200 + 140 + 170 + 145 + 165 = 820;$$

$$P_2 = \sum_{i=1}^{5} x_{i2}^2 = 190^2 + 150^2 + 210^2 + 150^2 + 150^2 = 147700;$$

$$R_2 = \sum_{i=1}^{5} x_{i2} = 190 + 150 + 210 + 150 + 150 = 850;$$

$$P_3 = \sum_{i=1}^{5} x_{i3}^2 = 230^2 + 190^2 + 200^2 + 190^2 + 200^2 = 205100;$$

$$R_3 = \sum_{i=1}^{5} x_{i3} = 230 + 190 + 200 + 190 + 200 = 1010;$$

$$P_4 = \sum_{i=1}^{5} x_{i4}^2 = 150^2 + 170^2 + 150^2 + 170^2 + 180^2 = 135200;$$

$$R_4 = \sum_{i=1}^{5} x_{i4} = 150 + 170 + 150 + 170 + 180 = 820;$$

$$s_{\text{общ}} = \sum_{j=1}^{4} P_j - \frac{1}{4 \cdot 5}\left(\sum_{j=1}^{4} R_j\right)^2$$

$$= 136750 + 147700 + 205100 + 135200 -$$

$$-\frac{1}{4 \cdot 5}(820 + 850 + 1010 + 820)^2 = 624750 - \frac{1}{20}(3500)^2 = 12250;$$

$$s_{\text{факт}} = \frac{1}{5}\sum_{j=1}^{4} R_j^2 - \frac{1}{4 \cdot 5}\left(\sum_{j=1}^{4} R_j\right)^2 = \frac{1}{5}(820^2 + 850^2 + 1010^2 + 820^2) -$$

$$-\frac{1}{4 \cdot 5}(820 + 850 + 1010 + 820)^2 = \frac{3087400}{5} - \frac{1}{20}(3500)^2 = 4980;$$

$$s_{\text{ост}} = s_{\text{общ}} - s_{\text{факт}} = 12250 - 4980 = 7270.$$

Вычислим дисперсии:

$$S_{\text{факт}}^2 = \frac{s_{\text{факт}}}{p-1} = \frac{4980}{3} = 1660; \qquad S_{\text{ост}}^2 = \frac{s_{\text{ост}}}{p \cdot (q-1)} = \frac{7270}{16} = 454{,}375.$$

Наблюдаемое значение критерия Фишера-Снедекора:

$$F_{\text{набл}} = \frac{S_{\text{факт}}^2}{S_{\text{ост}}^2} = \frac{1660}{454{,}375} = 3{,}65.$$

Число степеней свободы факторной дисперсии $k_1 = k_{\text{факт}} = p - 1 = 3$. Число степеней свободы остаточной дисперсии $k_2 = k_{\text{ост}} = p \cdot (q - 1) = 16$. При уровне значимости $\alpha = 0{,}05$ и числах степеней свободы $k_1 = 3$, $k_2 = 16$ по таблице приложения 4 найдем критическое значение критерия Фишера-Снедекора $F_{\text{кр}} = 3{,}24$. Критическая область $\omega = (3{,}24; +\infty)$.

Наблюдаемое значение критерия принадлежит критической области $F_{\text{набл}} \epsilon\, \omega$, следовательно, нулевая гипотеза отвергается. Таким образом, различие между партиями сырья оказывает существенное влияние на величину разрывной нагрузки.

Ответ Величина разрывной нагрузки зависит от номера партии.

Вопросы для самопроверки по разделам 4-7

1. Вариационный ряд, его разновидности. Средняя арифметическая и дисперсия ряда. Гистограмма.
2. Генеральная совокупность. Выборки и способы их получения. Репрезентативная выборка.
3. Точечные оценки неизвестных параметров генеральной совокупности и их свойства: несмещенность, состоятельность, эффективность.
4. Выборочная доля как точечная оценка генеральной доли, ее несмещенность и состоятельность.
5. Выборочная средняя как точечная оценка генеральной средней, ее несмещенность и состоятельность.
6. Выборочная дисперсия как точечная оценка генеральной дисперсии, ее смещенность и состоятельность. Несмещенная оценка генеральной дисперсии.
7. Интервальные оценки неизвестных параметров генеральной совокупности. Доверительная вероятность. Предельная ошибка выборки. Средние квадратические ошибки выборок.
8. Построение доверительного интервала для генеральной доли признака.
9. Построение доверительного интервала для генеральной средней.
10. Определение необходимого объема повторной и бесповторной выборок при оценке генеральной средней и доли.
11. Основные принципы проверки статистических гипотез.
12. Критерий – χ^2 Пирсона для проверки гипотезы о виде закона распределения.

13. t – критерий Стьюдента для проверки значимости коэффициента корреляции.
14. Функциональная, статистическая и корреляционная зависимости. Различия между ними. Основные задачи теории корреляции.
15. Линейная парная регрессия. Система нормальных уравнений для определения параметров прямых регрессии. Выборочная ковариация.
16. Оценка тесноты связи. Выборочный коэффициент корреляции и его свойства.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТРЫ

1. Балдин К.В. Теория вероятностей и математическая статистика: учебник / К.В. Балдин, В.Н. Башлыков, А.В Рукосуев. - М.: Дашков и К, 2016. - 472 с.

2. Балдин К.В. Теория вероятностей и математическая статистика: учебник / К.В. Балдин, В.Н. Башлыков. - М.: Дашков и К, 2016. - 472 с.

3. Бирюкова Л.Г. Теория вероятностей и математическая статистика: учебное пособие / Л.Г. Бирюкова, Г.И. Бобрик, В.И. Матвеев. - М.: Инфра-М, 2019. - 160 с.

4. Блягоз З.У. Теория вероятностей и математическая статистика. Курс лекций: учебное пособие / З.У. Блягоз. - СПб.: Лань, 2018. - 224 с.

5. Бондаренко П.С. Теория вероятностей и математическая статистика (для бакалавров) / П.С. Бондаренко, Г.В. Горелова, И.А. Кацко. - М.: КноРус, 2018. - 384 с.

6. Борзых Д.А. Теория вероятностей и математическая статистика в задачах: более 360 задач и упражнений / Д.А. Борзых. - М.: Ленанд, 2018. - 240 с.

7. Боровков А.А. Теория вероятностей / А.А. Боровков. - М.: КД Либроком, 2016. - 656 с.

8. Боровков А.А. Теория вероятностей / А.А. Боровков. - М.: КД Либроком, 2018. - 656 с.

9. Ватутин В.А. Теория вероятностей и математическая статистика в задачах / В.А. Ватутин, Г.И. Ивченко, Ю.И. Медведев. - М.: Ленанд, 2015. - 384 с.

10. Ватутин В.А. Теория вероятностей и математическая статистика в задачах / В.А. Ватутин, Г.И. Ивченко, Ю.И. Медведев, В.П. Чистяков. - М.: Ленанд, 2015. - 384 с.

11. Высоцкий В.С. ЕГЭ 2019. Математика. Теория вероятностей. Задача 4. (проф. уровень). Задача 10 (баз)Рабочая тетрадь / В.С. Высоцкий. - М.: МЦНМО, 2019. - 64 с.

12. Высоцкий И.Р. ЕГЭ 2017. Математика. Теория вероятностей. Задача 4(проф уровень). Задачи 10(базов уровень) Рабочая / И.Р. Высоцкий. - М.: МЦНМО, 2017. - 64 с.

13. Высоцкий И.Р. Теория вероятностей. Задачи и контрольные работы. 10 класс / И.Р. Высоцкий. - М.: МЦНМО, 2019. - 101 с.

14. Высоцкий И.Р. ЕГЭ 2016. Математика. Теория вероятностей. Задача 4 (профильный уровень). Задача 10 (базовый уровень / И.Р. Высоцкий. - М.: МЦНМО, 2016. - 64 с.

15. Ганичева А.В. Теория вероятностей: Учебное пособие / А.В. Ганичева. - СПб.: Лань, 2017. - 140 с.

16. Геворкян П.С. Теория вероятностей и математическая статистика / П.С. Геворкян, А.В. Потемкин, И.М. Эйсымонт. - М.: Физматлит, 2016. - 176 с.

17. Геворкян П.С. Теория вероятностей и математическая статистика / П.С. Геворкян. - М.: Физматлит, 2016. - 176 с.

18. Гливенко В.И. Теория вероятностей: учебник для высших педагогических учебных заведений / В.И. Гливенко. - М.: Ленанд, 2019. - 138 с.

19. Гмурман В.Е. Теория вероятностей и математическая статистика: учебник для прикладного бакалавриата / В.Е. Гмурман. - Люберцы: Юрайт, 2016. - 479 с.

20. Гмурман В.Е. Теория вероятностей и математическая статистика: учебник для СПО / В.Е. Гмурман. - Люберцы: Юрайт, 2016. - 479 с.

21. Горобец Б.С. Теория вероятностей, математическая статистика и элементы случайных процессов: упрощенный курс / Б.С. Горобец. - М.: КД Либроком, 2016. - 232 с.

22. Зеленцов Б.П. Теория вероятностей в познавательных и забавных задачах / Б.П. Зеленцов, О.И. Тутынина. - М.: КД Либроком, 2015. - 128 с.

23. Зеленцов Б.П. Теория вероятностей в познавательных и забавных задачах / Б.П. Зеленцов, О.И. Тутынина. - М.: Ленанд, 2019. - 128 с.

24. Золотаревская Д.И. Теория вероятностей: задачи с решениями / Д.И. Золотаревская. - М.: КД Либроком, 2016. - 168 с.

25. Золотаревская Д.И. Теория вероятностей: задачи с решениями / Д.И. Золотаревская. - М.: КД Либроком, 2018. - 168 с.

26. Ивашев-Мусатов О.С. Теория вероятностей и математическая статистика: учебник и практикум для СПО / О.С. Ивашев-Мусатов. - Люберцы: Юрайт, 2016. - 224 с.

27. Калинина В.Н. Теория вероятностей и математическая статистика: учебник для академического бакалавриата / В.Н. Калинина. - Люберцы: Юрайт, 2015. - 472 с.

28. Кацман Ю.Я. Теория вероятностей и математическая статистика. примеры с решениями: учебное пособие для прикладного бакалавриата / Ю.Я. Кацман. - Люберцы: Юрайт, 2016. - 130 с.

29. Кацман Ю.Я. Теория вероятностей и математическая статистика. примеры с решениями: учебник для СПО / Ю.Я. Кацман. - Люберцы: Юрайт, 2016. - 130 с.

30. Ковалев Е.А. Теория вероятностей и математическая статистика для экономистов: учебник и практикум для бакалавриата и магистратуры / Е.А. Ковалев, Г.А. Медведев. - Люберцы: Юрайт, 2016. - 284 с.

31. Козырь И.Е. Теория вероятностей в инженерных приложениях: учебное пособие / И.Е. Козырь, И.Ф. Пикалова, Н.В. Ханов. - СПб.: Лань, 2015. - 368 с.

32. Коледов Л.А. Теория вероятностей и математическая статистика: учебное пособие КПТ / Л.А. Коледов. - СПб.: Лань КПТ, 2016. - 224 с.

33. Колемаев В.А. Теория вероятностей и математическая статистика. учебник / В.А. Колемаев, В.Н. Калинина. - М.: КноРус, 2017. - 304 с.

34. Кочетков Е.С. Теория вероятностей и матем. статистика: уч. / Е.С. Кочетков, С.О. Смерчинская, В.В. Соколов. - М.: Форум, 2018. - 352 с.

35. Кочетков Е.С. Теория вероятностей в задачах и упражнениях: учебное пособие / Е.С. Кочетков, С.О. Смерчинская. - М.: Форум, 2018. - 559 с.

36. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник / Н.Ш. Кремер. - М.: Юнити, 2016. - 240 с.

37. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник и практикум для академического бакалавриата / Н.Ш. Кремер. - Люберцы: Юрайт, 2016. - 514 с.

38. Кристалинский В.Р. Теория вероятностей в системе Mathematica: учебное пособие / В.Р. Кристалинский. - СПб.: Лань, 2018. - 136 с.

39. Максимов Ю.Д. Теория вероятностей: опорный конспект / Ю.Д. Максимов. - М.: Проспект, 2016. - 88 с.

40. Мятлев В.Д. Теория вероятностей и математическая статистика. Математические модели: учебное пособие / В.Д. Мятлев. - М.: Академия, 2018. - 240 с.

41. Мятле, В.Д. Теория вероятностей и математическая статистика. Математические модели / В.Д. Мятлев. - М.: Academia, 2017. - 32 с.

42. Опойцев В.И. Школа Опойцева: Теория вероятностей / В.И. Опойцев. - М.: Ленанд, 2018. - 280 с.

43. Палий И.А. Теория вероятностей: учебное пособие / И.А. Палий. - М.: Инфра-М, 2018. - 352 с.

44. Пашкевич А.В. Теория вероятностей и математическая статистика для социологов и менеджеров: учебник / А.В. Пашкевич; Под ред. Макарова А.А. - М.: Academia, 2018. - 40 с.

45. Пашкевич А.В. Теория вероятностей и математическая статистика для социологов и менеджеров: учебник / А.В. Пашкевич. - М.: Академия, 2018. - 256 с.

46. Рыбников К.А. История математики: подисциплинарное изложение. Геометрия. Алгебра и теория чисел. Математический анализ. Теория вероятностей и математическая статистика. Дискретная математика / К.А. Рыбников. - М.: Ленанд, 2018. - 536 с.

47. Сапожников П.Н. Теория вероятностей, математическая статистика в примерах, задачах и тестах: учебное пособие / П.Н. Сапожников, А. Макаров, М.В. Радионова. - М.: Инфра-М, 2017. - 192 с.

48. Семенов В.А. Теория вероятностей и математическая статистика: учебное пособие / В.А. Семенов. - СПб.: Питер, 2015. - 195 с.

49. Серовайский С.Я. История математики: Эволюция математических идей: Вычислительная математика. Теория вероятностей. Информатика. Математическая логика / С.Я. Серовайский. - М.: Ленанд, 2019. - 240 с.

50. Сидняев Н.И. Теория вероятностей и математическая статистика: учебник / Н.И. Сидняев. - Люберцы: Юрайт, 2015. - 219 с.

51. Сидняев Н.И. Теория вероятностей и математическая статистика: учебник для СПО / Н.И. Сидняев. - Люберцы: Юрайт, 2016. - 219 с.

52. Спирина М.С. Теория вероятностей и математическая статистика: учебник / М.С. Спирина. - М.: Academia, 2019. - 144 с.

53. Спирина М.С. Теория вероятностей и математическая статистика: учебник / М.С. Спирина. - М.: Академия, 2015. - 176 с.

54. Спирина М.С. Теория вероятностей и математическая статистика стер / М.С. Спирина. - М.: Academia, 2016. - 163 с.

55. Спирина М.С. Теория вероятностей и математическая статистика: Сборник задач: учебное пособие / М.С. Спирина. - М.: Academia, 2016. - 144 с.

56. Спирина М.С. Теория вероятностей и математическая статистика. Сборник задач: учебное пособие / М.С. Спирина. - М.: Academia, 2018. - 176 с.

57. Спирина М.С. Теория вероятностей и математическая статистика: Сборник задач: учебное пособие / М.С. Спирина. - М.: Academia, 2017. - 144 с.

58. Татарников О.В. Теория вероятностей и математическая статистика для экономистов (для бакалавров) / О.В. Татарников, Е.В. Швед. - М.: КноРус, 2018. - 352 с.

59. Трухан А.А. Теория вероятностей в инженерных приложениях: учебное пособие / А.А. Трухан, Г.С. Кудряшев. - СПб: Лань, 2015. - 368 с.

60. Тутубалин В.Н. Теория вероятностей / В.Н. Тутубалин. - М.: Academia, 2018. - 210 с.

61. Хуснутдинов Р.Ш. Теория вероятностей: уч. / Р.Ш. Хуснутдинов. - М.: Инфра-М, 2018. - 384 с.

62. Шведов А.С. Теория вероятностей и математическая статистика: промежуточный уровень: учебное пособие / А.С. Шведов. - М.: ИД ВШЭ, 2016. - 280 с.

ПРИЛОЖЕНИЯ

Приложение 1

Нормированная функция Лапласа $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$. (4 знака после запятой)

x	0	1	2	3	4	5	6	7	8	9
0,0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2703	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3883	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4858
2,2	4861	4865	4868	4872	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4951
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3,0	4987	4987	4988	4988	44988	4989	4989	4990	4990	4990
3,1	4991	4991	4991	4992	4992	4992	4992	4993	4993	4993
3,2	4993	4994	4994	4994	4994	4995	4995	4995	4995	4995
3,3	4995	4996	4996	4996	4996	4996	4996	4996	4997	4997
3,4	4997	4997	4997	4997	4997	4997	4998	4998	4998	4998
3,5	4998	4998	4998	4998	4998	4998	4998	4998	4999	4999
3,6	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999

3,7	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999
3,8	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999
3,9	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999
4,0	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999

Нормированная функция Лапласа имеет свойства

$$\phi(-x) = -\phi(x),$$

$\phi(x > 4) = 0{,}5.$

Приложение 2

Таблица значений $t_{\gamma,n}$ найденных по распределению Стьюдента, из соотношения

$$\int_0^{t_{\gamma,n}} \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma\big((n-1)/2\big)}\left(1+\frac{t^2}{n-1}\right)^{-n/2} dt = \frac{\gamma}{2}.$$

Объем выборки n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,001	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Приложение 3

Таблица значений $q = q(\gamma, n)$ найденных по χ – распределению, из соотношения

$$\int_{\frac{\sqrt{n-1}}{(1+q)}}^{\frac{\sqrt{n-1}}{(1-q)}} \frac{\chi^{n-2} e^{-\chi^2/2}}{2^{(n-3)/2}\Gamma\left(\frac{n-1}{2}\right)} d\,\chi = \gamma.$$

Объем выборки n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение 4

Критические точки F- распределения Фишера-Снедекора, найденные из соотношения

$$\int_{F(k_1;k_2;\delta)}^{+\infty} \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} \frac{x^{\frac{k_1}{2}-1}}{\left(1+\frac{k_1}{k_2}x\right)^{\frac{k_1+k_2}{2}}}dx = \delta$$

$\delta = 0,05$																			
k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	240	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	6,61	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,83	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

$\delta = 0{,}01$																			
k_1 / k_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Приложение 5

Значения $\chi^2(k;\delta)$ критерия Пирсона, найденное из соотношения

$$\int_{\chi^2(k;\delta)}^{+\infty} \frac{x^{\frac{k}{2}-1}e^{-x/2}}{2^{\frac{k}{2}}\ \Gamma\left(\frac{k}{2}\right)}dx = \delta.$$

Число степеней свободы k	Вероятность δ												
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,00	0,00	0,00	0,02	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58	1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,3
4	0,30	0,43	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,7	13,3
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,1	13,4	15,1
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,6	12,6	15,0	16,8
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,0	14,1	16,6	18,5
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,0	13,4	15,5	18,2	20,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2
11	3,05	3,61	4,58	5,58	6,99	8,15	10,3	12,9	14,6	17,3	19,7	22,6	24,7
12	3,57	4,18	5,23	6,30	7,81	9,03	11,3	14,0	15,8	18,5	21,0	24,1	26,2
13	4,11	4,76	5,89	7,04	8,63	9,93	12,3	15,1	17,0	19,8	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79	9,47	10,8	13,3	16,2	18,1	21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6
16	5,81	6,61	7,96	9,31	11,1	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8
19	7,63	8,57	10,1	11,6	13,7	15,3	18,3	21,7	23,9	27,2	30,1	33,7	36,2
20	8,26	9,24	10,8	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6
21	8,90	9,92	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9
22	9,54	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,7	44,3
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6
27	12,9	14,1	16,1	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6
30	14,9	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9

Приложение 6

Таблица критических точек $G_{кр}(k, l, \alpha)$ критерия Кохрена

Уровень значимости $\alpha = 0{,}01$							
l	k						
	1	2	3	4	5	6	7
2	0,9999	0,9950	0,9794	0,9586	0,9373	0,9172	0,8988
3	0,9933	0,9423	0,8831	0,8335	0,7933	0,7606	0,7335
4	0,9676	0,8643	0,7814	0,7212	0,6761	0,6410	0,6129
5	0,9279	0,7885	0,6957	0,6329	0,5875	0,5531	0,5259
6	0,8828	0,7218	0,6258	0,5635	0,5195	0,4866	0,4608
7	0,8376	0,6644	0,5685	0,5080	0,4659	0,4347	0,4105
8	0,7945	0,6152	0,5209	0,4627	0,4226	0,3932	0,3704
9	0,7544	0,5727	0,4810	0,4251	0,3870	0,3592	0,3378
10	0,7175	0,5358	0,4469	0,3934	0,3572	0,3308	0,3106
12	0,6528	0,4751	0,3919	0,3428	0,3099	0,2861	0,2680
15	0,5747	0,4069	0,3317	0,2882	0,2593	0,2386	0,2228
20	0,4799	0,3297	0,2654	0,2288	0,2048	0,1877	0,1748
24	0,4247	0,2871	0,2295	0,1970	0,1759	0,1608	0,1495
30	0,3632	0,2412	0,1913	0,1635	0,1454	0,1327	0,1232
40	0,2940	0,1915	0,1508	0,1281	0,1135	0,1033	0,0957
60	0,2151	0,1371	0,1069	0,0902	0,0796	0,0722	0,0668
120	0,1225	0,0759	0,0585	0,0489	0,0429	0,0387	0,0357
∞	0,0000	0,0000	0,0000	0,000	0,0000	0,0000	0,0000
Уровень значимости $\alpha = 0{,}01$							
l	k						
	8	9	10	16	36	144	∞
2	0,8823	0,8674	0,8539	0,7949	0,7067	0,6062	0,5000
3	0,7107	0,6912	0,6743	0,6059	0,5153	0,4230	0,3333
4	0,5897	0,5702	0,5536	0,4884	0,4057	0,3251	0,2500
5	0,5037	0,4854	0,4697	0,4094	0,3351	0,2644	0,2000
6	0,4401	0,4229	0,4084	0,3529	0,2858	0,2229	0,1667
7	0,3911	0,3751	0,3616	0,3105	0,2494	0,1929	0,1429
8	0,3522	0,3373	0,3248	0,2779	0,2214	0,1700	0,1250
9	0,3207	0,3067	0,2950	0,2514	0,1992	0,1521	0,1111
10	0,2945	0,2813	0,2704	0,2297	0,1811	0,1376	0,1000
12	0,2535	0,2419	0,2320	0,1961	0,1535	0,1157	0,0833
15	0,2104	0,2002	0,1918	0,1612	0,1251	0,0934	0,0667
20	0,1646	0,1567	0,1501	0,1248	0,0960	0,0709	0,0500
24	0,1406	0,1338	0,1283	0,1060	0,0810	0,0595	0,0417
30	0,1157	0,1100	0,1054	0,0867	0,0658	0,0480	0,0333
40	0,0898	0,0853	0,0816	0,0668	0,0503	0,0363	0,0250
60	0,0625	0,0594	0,0567	0,0461	0,0344	0,0245	0,0167

120	0,0334	0,0316	0,0302	0,0242	0,0178	0,0125	0,0083
∞	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Уровень значимости $\alpha = 0,05$							
l	k						
	1	2	3	4	5	6	7
2	0,9985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332
3	0,9669	0,8709	0,7977	0,7457	0,7071	0,6771	0,6530
4	0,9065	0,7679	0,6841	0,6287	0,5895	0,5598	0,5365
5	0,8412	0,6338	0,5981	0,5440	0,5063	0,4783	0,4564
6	0,7808	0,6161	0,5321	0,4803	0,4447	0,4184	0,3980
7	0,7271	0,5612	0,4800	0,4307	0,3974	0,3726	0,3535
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185
9	0,6385	0,4775	0,4027	0,3584	0,3286	0,3067	0,2901
10	0,6020	0,4450	0,3733	0,3311	0,3029	0,2823	0,2666
12	0,5410	0,3924	0,3624	0,2880	0,2624	0,2439	0,2299
15	0,4709	0,3346	0,2758	0,2419	0,2195	0,2034	0,1911
20	0,3894	0,2705	0,2205	0,1921	0,1735	0,1602	0,1501
24	0,3434	0,2354	0,1907	0,1656	0,1493	0,1374	0,1286
30	0,2929	0,1980	0,1593	0,1377	0,1237	0,1137	0,1061
40	0,2370	0,1576	0,1259	0,1082	0,0968	0,0887	0,0827
60	0,1737	0,1131	0,0895	0,0765	0,0682	0,0623	0,0583
120	0,0998	0,0632	0,0495	0,0419	0,0371	0,0337	0,0312
∞	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Уровень значимости $\alpha = 0,05$							
l	k						
	8	9	10	16	36	144	∞
2	0,8159	0,8010	0,7880	0,7341	0,6602	0,5813	0,5000
3	0,6333	0,6167	0,6025	0,5466	0,4748	0,4031	0,3333
4	0,5175	0,5017	0,4884	0,4366	0,3720	0,3093	0,2500
5	0,4387	0,4241	0,4118	0,3645	0,3066	0,2013	0,2000
6	0,3817	0,3682	0,3568	0,3135	0,2612	0,2119	0,1667
7	0,3384	0,3259	0,3154	0,2756	0,2278	0,1833	0,1429
8	0,3043	0,2926	0,2829	0,2462	0,2022	0,1616	0,1250
9	0,2768	0,2659	0,2568	0,2226	0,1820	0,1446	0,1111
10	0,2541	0,2439	0,2353	0,2032	0,1655	0,1308	0,1000
12	0,2187	0,2098	0,2020	0,1737	0,1403	0,1100	0,0833
15	0,1815	0,1736	0,1671	0,1429	0,1144	0,0889	0,0667
20	0,1422	0,1357	0,1303	0,1108	0,0879	0,0675	0,0500
24	0,1216	0,1160	0,1113	0,0942	0,0743	0,0567	0,0417
30	0,1002	0,0958	0,0921	0,0771	0,0604	0,0457	0,0333
40	0,0780	0,0745	0,0713	0,0595	0,0462	0,0347	0,0250
60	0,0552	0,0520	0,0497	0,0411	0,0316	0,0234	0,0167
120	0,0292	0,0279	0,0266	0,0218	0,0165	0,0120	0,0083
∞	0,0000	0,000	0,0000	0,0000	0,0000	0,0000	0,0000

Приложение 7

Таблица значений $t_{\gamma,k}$ найденных по распределению Стьюдента, из соотношений:

для двусторонней критической области

$$\int_{-t_{\gamma,k}}^{t_{\gamma,k}} \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \left(1+\frac{t^2}{k}\right)^{-(k+1)/2} dt = 1-\alpha;$$

для односторонней критической области

$$\int_{-t_{\gamma,k}}^{t_{\gamma,k}} \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \left(1+\frac{t^2}{k}\right)^{-(k+1)/2} dt = 1-2\alpha.$$

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71

27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Федеральное государственное образовательное бюджетное
учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»
Ярославский филиал

Жаров А.Н.
Минеичева И.Г.

АНАЛИЗ ДАННЫХ

Учебное пособие

ISBN 978-5-6044447-4-0

9 785604 444740

Подписано в печать 28.08.2020. Формат 60х90/16.
Усл. печ. л. 9,25. Тираж 30 экз. Заказ № 4096.

Отпечатано в ООО «ПКФ «СОЮЗ-ПРЕСС»
150062, г. Ярославль, пр-д Доброхотова, д. 16, кв. 158.
Тел.: (4852) 58-76-33, 58-76-37.
E-mail: kancler2007@yandex.ru.