Министерство образования Российской Федерации

Московский государственный университет экономики, статистики и информатики

Евразийский открытый институт

В.П. Лисьев

Теория вероятностей и математическая статистика

Учебное пособие

Москва 2006

УДК 519.2
ББК 22.171
Л 638

Лисьев В.П. **ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА:** Учебное пособие/ Московский государственный университет экономики, статистики и информатики. – М., 2006. – 199 с.

ISBN 5-374- 00005-5

© В.П. Лисьев, 2006
© Московский государственный университет экономики, статистики и информатики, 2006

Содержание

Сведения об авторе **5**

Общие сведения о дисциплине **5**

Цель и задачи дисциплины **6**

Рекомендации по изучению дисциплины **7**

1. Случайные события **9**
1.1. События. Пространство элементарных событий 10
1.2. Элементы комбинаторного анализа 11
1.3. Отношения между событиями 13
1.4. Вероятность события 14
1.5. Простейшие свойства вероятности 16
1.6. Условная вероятность. Формула умножения вероятностей. Зависимые и независимые события 18
1.7. Формула сложения вероятностей 19
1.8. Формула полной вероятности и формула Байеса 20
1.9. Последовательность независимых испытаний. Формула Бернулли 21
1.10. Асимптотические приближения формулы Бернулли 23
Контрольные вопросы и вопросы для обсуждения 25

2. Случайные величины **27**
2.1. Определение, классификация, способы задания случайных величин 28
2.2. Функция распределения вероятностей и её свойства 29
2.3. Плотность распределения вероятностей и её свойства 31
2.4. Функция случайной величины. Математическое ожидание 33
2.5. Числовые характеристики случайных величин. 35
2.6. Квантили, квартили и вероятное отклонение. 40
2.7. Моменты случайных величин. Асимметрия и эксцесс 41
2.8. Производящие функции 43
2.9. Примеры дискретных законов распределения 45
2.10. Примеры непрерывных распределений 46
Контрольные вопросы и вопросы для обсуждения 53

3. Многомерные случайные величины **55**
3.1. Определение многомерных случайных величин 56
3.2. Функция распределения вероятностей двухмерной случайной величины 57
3.3. Плотность распределения вероятностей двухмерной случайной величины 60
3.4. Условные законы распределения. Статистическая зависимость. 62
3.5. Числовые характеристики многомерных случайных величин. Ковариационный момент и коэффициент корреляции 64
3.6. Условные числовые характеристики. Линии регрессии. Корреляционное отношение 68
3.7. Двухмерное нормальное распределение 71
Контрольные вопросы и вопросы для обсуждения 75

4. Функциональные преобразования случайных величин **77**
4.1. Функция одной случайной величины 78
4.2. Функция нескольких случайных величин 79
4.3. Теоремы о математических ожиданиях и дисперсиях 80

4.4. Некоторые специальные законы распределения 81
Контрольные вопросы и вопросы для обсуждения 83

5. Закон больших чисел. Центральная предельная теорема 85
5.1. Предварительные замечания 86
5.2. Неравенство Чебышева 86
5.3. Теорема Чебышева 88
5.4. Теорема Бернулли 89
5.5. Центральная предельная теорема 90
Контрольные вопросы и вопросы для обсуждения 91

6. Статистическая обработка экспериментальных данных. Оценка параметров 93
6.1. Задачи математической статистики 94
6.2. Выборка. Вариационный ряд. Эмпирические законы распределения 95
6.3. Эмпирические числовые характеристики 98
6.4. Точечные оценки параметров. Свойства эмпирических характеристик 101
6.5. Доверительные интервалы. Общие определения 105
6.6. Доверительные интервалы параметров нормального распределения 106
6.7. Построение доверительного интервала для вероятности события 111
Контрольные вопросы и вопросы для обсуждения 114

7. Проверка статистических гипотез. 115
7.1. Общие положения 116
7.2. Проверка гипотез о параметрах распределений 117
7.3. Критерий квантилей 119
7.4. Проверка гипотез о распределениях 120
Контрольные вопросы и вопросы для обсуждения 124

8. Дисперсионный анализ 125
8.1. Постановка задачи дисперсионного анализа 126
8.2. Однофакторный дисперсионный анализ 127
8.3. Двухфакторный дисперсионный анализ 129
8.4. Трёхфакторный дисперсионный анализ. План «латинский квадрат» 131
Контрольные вопросы и вопросы для обсуждения 133

9. Регрессионный анализ 135
9.1. Постановка и схема решения задачи регрессионного анализа 136
9.2. Одномерный линейный регрессионный анализ 139
9.3. Многомерный линейный регрессионный анализ 142
9.4. Одномерный нелинейный регрессионный анализ 144
Контрольные вопросы и вопросы для обсуждения 146

10. Применение ЭВМ. 147
10.1. Общие замечания 148
10.2. Средства решения статистических задач в пакете MathCAD. 148
10.3. Решение статистических задач в среде Microsoft Excel. 149

Практикум 152

Список используемой литературы 199

Сведения об авторе

Лисьев Владимир Парфёнович, 1941 года рождения.

Кандидат физико-математических наук, доцент. Стаж работы в системе образования около 40 лет. Работал в Томском государственном университете, в Алтайском политехническом институте (старое название вуза), в Восточно-Казахстанском государственном техническом университете, три учебных года работал в Браззавильском университете в Народной Республике Конго, заведовал кафедрой высшей математики и кафедрой прикладной математики и вычислительной техники. В момент создания УМК по данной дисциплине – доцент Усть-Каменогорского филиала МЭСИ.

Автор является специалистом в области прикладной математики: исследование операций, теория массового обслуживания, теория восстановлений, теория вероятностей и математическая статистика. Автор читает курсы лекций по высшей математике, по линейной алгебре, по математическим методам исследования операций, по теории вероятностей и математической статистике.

Общие сведения о дисциплине

Весь курс разбит на 10 тем. Лекционные занятия рекомендуется проводить по 9 темам, причём темы 4 и 5 можно объединить для того, чтобы получилось 8 обзорных лекций при семнадцати недельном семестре. Каждой обзорной лекции соответствуют два практических занятия, индивидуальные задания по теме, одна аудиторная консультация или консультация в режиме «On-line». Желательно также по каждой теме предусмотреть тест типа «Самопроверка». По 10-й теме «Применение ЭВМ» проводится только практическое занятие непосредственно в компьютерных классах. Рекомендуется проводить один промежуточный экзамен для проведения текущей аттестации и один итоговый экзамен. Экзамены рекомендуется проводить по тестам, в которые можно включить часть вопросов из тестов, предназначенных для самопроверки. В любом случае в экзаменационных тестах должны быть представлены все темы курса. Перед экзаменами рекомендуется проводить как аудиторную консультацию, так и консультацию по компьютерной сети, например в режиме «Чат», если имеется соответствующее техническое обеспечение.

Последовательность мероприятий по изучению дисциплины с привязкой по времени и с информацией о порядке зачёта пройденного материала должна быть представлена календарным планом, в котором указываются сроки мероприятий, а также непосредственные источники и инструменты их выполнения. Весь процесс ориентирован на 51 час аудиторных занятий. Если число аудиторных часов увеличено до 68, то в календарном плане следует добавить по одной лекции на каждую тему. Рекомендуется вести непрерывный контроль над степенью усвоения материалов студентами. Самостоятельная работа студентов включает изучение материала по темам (текст пособия), чтение дополнительных источников, выполнение самостоятельных заданий (практикум), ответы на вопросы для самопроверки и прохождение тестов по самопроверке. Вся эта работа должна выполняться в привязке к изучаемой теме в точном соответствии с календарным планом. Рекомендуется пользоваться некоторой системой зачётных баллов для оценки своевременности и качества выполняемой студентами работы. Итоговую оценку по курсу следует выставлять студенту с учётом накопленных баллов в течение семестра. Для контроля самостоятельной работы студентов рекомендуется использовать компьютерную сеть в любых доступных вариантах.

Цель и задачи дисциплины

Курс «Теория вероятностей и математическая статистика» является составной частью цикла математических дисциплин, составляющих фундамент математического образования специалиста. В любой области человеческой деятельности имеют место случайные явления, которые не позволяют осуществить точный прогноз результатов этой деятельности. Теория вероятностей и математическая статистика изучают закономерности случайных явлений. Знание этих закономерностей помогает принимать решения в условиях неопределённости, направленные на достижение поставленных целей.

Курс «Теория вероятностей и математическая статистика» является основой для изучения последующих дисциплин, таких как «Эконометрика», «Статистические методы прогнозирования», «Исследование операций», «Методы оптимизации», «Теория массового обслуживания», «Теория восстановлений», «Основы актуарных расчётов» и т.д.

Преподавание теории вероятностей и математической статистики имеет целью: ознакомить студентов с основами теории, необходимыми для решения прикладных задач, развить способности студентов к логическому и алгоритмическому мышлению, привить навыки самостоятельного изучения литературы по данной дисциплине и ее приложениям. На практических занятиях по дисциплине необходимо развить навыки составления и анализа математических моделей несложных задач прикладного характера, связанных со случайными явлениями, научить способам вычисления вероятностей простых и сложных событий, методам оценки неизвестных параметров на основе экспериментальных данных, методам проверки гипотез и правилам принятия решений, методам аппроксимации статистических связей между величинами или факторами.

Приступая к изучению курса теории вероятностей и математической статистики, студент должен знать:

- элементарную математику в объёме средней школы;
- дифференцирование и интегрирование функций из курса высшей математики;
- основные сведения по сбору и первичной обработке данных из общего курса статистики.

В результате изучения дисциплины студент должен знать:

- правила вычисления вероятностей случайных событий;
- способы определения и построения законов распределения вероятностей случайных величин и вычисления их числовых характеристик;
- основные понятия, связанные со статистической зависимостью между случайными величинами;
- способы оценки неизвестных параметров по экспериментальным данным;
- способы проверки гипотез по экспериментальным данным;
- методы анализа статистической связи между величинами и аппроксимации её функциональной связью.

В результате изучения дисциплины студент должен уметь:

- самостоятельно разбираться в материалах, содержащихся в специальной литературе по вопросам, связанным с данной дисциплиной;
- выбирать метод исследования и доводить решение задач до практически приемлемого результата;
- вычислять вероятности простых и сложных событий;
- находить необходимые характеристики случайных величин по известным законам распределения вероятностей или оценки этих характеристик по экспериментальным данным;
- практически применять известные критерии проверки статистических гипотез;
- проводить дисперсионный анализ с целью установления факта влияния или невлияния на исследуемую величину некоторых факторов;
- проводить регрессионный анализ с целью получения аппроксимирующей функции, по которой можно осуществлять прогнозирование.

Рекомендации по изучению дисциплины

Прежде всего, заметим, что теорию вероятностей условно можно разделить на три части: случайные события, случайные величины и случайные процессы. В такой последовательности излагается и изучается теория вероятностей. Нарушение этой последовательности освоения теории невозможно, поскольку материал каждой последующей части опирается на понятия, определяемые в предыдущих частях. Математическая статистика, излагаемая после теории вероятностей, является, по сути, прикладной частью теории вероятностей. Она направлена на обработку экспериментальных данных, полученных в условиях частичной или полной неопределённости. В математической статистике используются все основные законы и понятия теории вероятностей. В данном пособии рассматриваются случайные события, случайные величины и элементы математической статистики. Ключевой темой, открывающей возможность правильно воспринимать весь последующий материал, является тема «Случайные события». Поэтому при изучении курса рекомендуется уделить этой теме максимальное внимание. Особенно важно хорошо освоить операции над случайными событиями. Нужно научиться выражать сложные события, связанные с данным опытом, через простые (элементарные) события, не испытывая при этом особых затруднений. Не следует забывать также об относительности понятия случайного события. Это понятие связано с некоторым опытом, т.е. жёсткой фиксацией некоторого комплекса условий, малейшее изменение которых может изменить соотношения между событиями. Это замечание относится и к случайным величинам.

В начале изучения курса следует просмотреть учебную программу курса, затем внимательно просмотреть оглавление данного учебного пособия, с тем, чтобы понять общую направленность изложения материала – от теории к практике. После этого можно приступать к последовательному изучению тем. Порядок изучения темы: прослушать обзорную лекцию ведущего преподавателя, изучить материал по теме, представленный в пособии, разобраться с непонятными вопросами во время практических занятий, ответить на вопросы для самопроверки, приведённые в конце каждой темы, и поразмышлять над вопросом для обсуждения. После этого следует выполнить индивидуальные практи-

ческие задания по теме из раздела пособия «Практикум», посмотреть рекомендуемые дополнительные источники, получить ответы на непонятные вопросы темы на консультации, сдать тест по данной теме в режиме «Самопроверка». Последнее можно делать многократно, пока не будет достигнут необходимый уровень (не менее 70% правильных ответов). Порядок представления отчётов по самостоятельным заданиям определяется ведущим преподавателем (тьютором).

ТЕМА 1.

Случайные события

Студент должен
освоить:

- понятия случайного события, операции над событиями, вероятность события, правила вычисления вероятностей;

приобрести навыки:

- решения задач на вычисление вероятностей простых и сложных событий.

Краткое содержание

Введение. Случайные события, пространство элементарных событий, алгебра событий. Основные формулы комбинаторики. Вероятность события и ее свойства. Условная вероятность. Формула умножения вероятностей. Статистическая зависимость между событиями. Формула сложения вероятностей. Формула полной вероятности и формула Байеса. Последовательность независимых испытаний. Формула Бернулли и следствия из нее. Предельные теоремы Муавра – Лапласа и Пуассона.

1.1. События. Пространство элементарных событий

Под опытом (или испытанием) будем понимать создание некоторого комплекса условий S, предполагая при этом, что опыт (т. е. комплекс условий S) может быть повторён многократно, а его результаты поддаются точному описанию. Результаты опыта будем называть событиями. События, которые происходят обязательно, если только реализован комплекс условий S, будем называть *достоверными* событиями и обозначать буквой U. События, которые никогда не осуществляются при реализации S, будем называть *невозможными* событиями, и обозначать буквой V. Наибольший интерес представляет третья группа, в которую входят те события, которые при реализации S могут либо произойти, либо не произойти и заранее неизвестно, произойдёт событие или нет. Такие события называют *случайными* событиями и обозначают большими латинскими буквами A, B, C, ...

По своему логическому содержанию события могут быть составными, т. е. разложимыми на более простые события, и *элементарными*, которые разложить на более простые события уже невозможно. Пусть, например, опыт заключается в том, что из десяти карточек, пронумерованных числами от 1 до 6, случайным образом выбирается одна. Обозначим событие, состоящее в том, что выбранная карточка имеет номер k, через W_k. Таких событий шесть. Они полностью описывают возможные результаты опыта и являются элементарными событиями. На этих событиях строятся более сложные события. Пусть B – событие, состоящее в том, что номер выбранной карточки чётный; C – событие, состоящее в том, что номер карточки делится на 3 без остатка. Очевидно, что B осуществляется тогда, когда осуществляется одно из следующих событий: W_2, W_4, W_6. Это можно условно отобразить записью $B = \{W_2, W_4, W_6\}$. Событие C выполняется только тогда, когда осуществляется одно из событий W_3, W_6. Таким образом, $C = \{W_3, W_6\}$. На основе элементарных событий W_k, $k = 1, 2, ..., 6$ можно построить и другие составные события.

Неразложимость элементарных событий означает, что при осуществлении опыта обязательно произойдёт одно и только одно из них. Это эквивалентно записи: $U = \{W_1, W_2, ..., W_n\}$, что можно прочитать так: «достоверно, что произойдёт одно из перечисленных событий». Совокупность всех элементарных событий называется *пространством элементарных событий*. Пространство элементарных событий принято обозначать буквой Ω, а число элементарных событий, составляющих Ω, принято называть $\mathrm{Card}(\Omega)$. В терминологии теории множеств $\mathrm{Card}(\Omega)$ – мощность дискретного конечного множества.

Любое случайное событие представляет собой подмножество множества Ω. Совокупность всех таких подмножеств (событий) является, в свою очередь, множеством случайных событий. Отметим, что событие U совпадает с Ω, а событие V является пустым подмножеством $\varnothing$. В дальнейшем будем пользоваться обозначениями U и V тогда, когда события рассматриваются как конкретные физические явления. На уровне множеств следует пользоваться обозначениями Ω и $\varnothing$. Определение случайного события как подмножества пространства элементарных событий является основой аксиоматического построения теории вероятностей. Построение пространства элементарных событий являет-

ся ответственным этапом. Неверное построение этого пространства может привести не только к количественным, но и к логическим ошибкам, вплоть до мнимых парадоксов.

1.2. Элементы комбинаторного анализа

Число элементов множеств вычисляют по формулам комбинаторного анализа. Пусть имеется n различимых объектов. Последовательность k объектов из n будем называть комбинацией из n объектов по k. Комбинации называют упорядоченными в случае, когда две комбинации, содержащие одинаковые объекты, рассматриваются как различные, если только эти объекты расположены в различном порядке. Комбинации называются неупорядоченными, если комбинации, содержащие одни и те же объекты, считаются одинаковыми независимо от порядка расположения этих объектов.

Пример 1.1. Из трёх объектов (⊗, □, ∇) можно составить 6 упорядоченных комбинаций по 2 объекта: (⊗, □), (⊗, ∇), (□, ∇), (□, ⊗), (∇, ⊗), (∇, □). Комбинации (⊗, □) и (□, ⊗) считаются различными. Из этих объектов можно составить 3 неупорядоченные комбинации по 2: (⊗, □), (⊗, ∇), (□, ∇). Комбинации (⊗, □) и (□, ⊗) считаются одинаковыми.

При определении числа различных комбинаций применяют *основное правило комбинаторики*, заключающееся в следующем. Пусть требуется последовательно выполнить i действий, причём 1-е действие можно выполнить n_1 способами, 2-е действие – n_2 способами и т.д., наконец, i-е действие – n_i способами. Число N возможных способов выполнения всей последовательности действий определяется произведением: $N = n_1 n_2 \dots n_i$.

Пример 1.2. Сколько существует трёхзначных чётных чисел, составленных только из цифр 2, 4, 5, 6, 8, 9? Процесс составления трёхзначного числа – это три действия: первое действие – выбор первой цифры, второе – выбор второй цифры, третье – выбор третьей цифры. Первая цифра может быть любой, поэтому первое действие можно выполнить шестью способами. Второе – также шестью способами. Третья цифра должна быть чётной, следовательно, третье действие можно выполнить 4-мя способами. Таким образом, существует всего $6 \cdot 6 \cdot 4 = 144$ числа, которые удовлетворяют требованиям задачи.

1. *Размещения с повторениями*. Пусть имеется n различимых объектов. Будем размещать эти объекты по k местам следующим образом. Извлекаем объект из общего числа n, отводим ему очередное не занятое место и возвращаем этот объект обратно. Комбинации, полученные таким образом, называются размещениями с повторением из n объектов по k, а процедура их построения – выбором с возвращением. Число вариантов выбора при каждом извлечении равно n. Следовательно, общее число таких размещений равно

$$N = n^k. \qquad (1.1)$$

Пример 1.3. Число размещений с повторениями из 3 объектов по 2 равно $3^2 = 9$. Возьмём объекты из примера 1.1: (⊗, □, ∇). Тогда размещениями с повторением будут следующие комбинации: (⊗, ⊗), (⊗, □), (⊗, ∇), (□, □), (□, ∇), (□, ⊗), (∇, ∇), (∇, □), (∇, ⊗).

2. *Размещения без повторений*. Будем снова размещать n объектов по k местам, не возвращая обратно извлечённые объекты. Полученные комбинации называют размещениями без повторений или просто размещениями. Число размещений из n объектов по k будем обозначать A_n^k. Очевидно, что при первом извлечении имеется n вариантов выбора, при втором – $(n-1)$ вариант, при третьем – $(n-2)$ и т.д. Число вариантов выбора для k-го извлечения равно $(n-k+1)$. Применяя основное правило комбинаторики, получим:

$$A_n^k = n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}. \qquad (1.2)$$

Здесь и далее выражение типа $m!$, где m – целое положительное число, означает произведение целых чисел от 1 до m включительно и читается как «m-факториал» (0! = 1).

Заметим, что в формуле (1.2) обязательно выполнение неравенства $0 \le k \le n$, а в формуле (1.1) k может быть любым целым положительным числом.

Пример 1.4. В бригаде из 20 рабочих выбирается бригадир, хозяйственник и общественный кассир. Сколько вариантов выбора имеется в распоряжении рабочих?

Здесь $n = 20$, $k = 3$. Число вариантов выбора будет равно $N = 20 \cdot 19 \cdot 18 = 6840$.

3. *Перестановки.* Перестановками называются комбинации, которые отличаются друг от друга только порядком расположения объектов. Перестановки являются частным случаем размещений без повторений, когда число мест совпадает с числом объектов, т. е. когда $k = n$. Таким образом, число перестановок из n объектов определяется формулой:

$$N = A_n^n = n! \qquad (1.3)$$

Пример 1.5. Число перестановок из трёх объектов равно 3! = 6. Перечислим все перестановки из объектов примера 1.1: (⊗, □, ∇), (□, ∇, ⊗), (∇, ⊗, □), (⊗, ∇, □), (∇, □, ⊗), (□, ⊗, ∇).

4. *Сочетания.* Сочетаниями из n объектов по k называются такие комбинации по k объектов из данных n, которые отличаются друг от друга хотя бы одним объектом. Число сочетаний обозначают C_n^k. Все сочетания можно получить из размещений путём исключения лишних перестановок из k объектов. Следовательно,

$$N = C_n^k = \frac{A_n^k}{A_k^k} = \frac{n!}{k!(n-k)!}. \qquad (1.4)$$

Пример 1.6. В бригаде из 20 рабочих требуется выбрать троих для направления в командировку. Сколько существует вариантов выбора?

Последовательность выбора в данном случае не имеет значения. Комбинации «Иван, Пётр, Николай» и «Пётр, Николай, Иван» по условию задачи являются не различимыми. Значит число вариантов выбора равно числу сочетаний из 20 объектов (рабочих) по 3:

$$\mathrm{N} = \mathrm{C}_{20}^3 = \frac{20!}{3!(20-3)!} = \frac{20!}{3!\,17!} = \frac{20 \cdot 19 \cdot 18}{6} = 1140.$$

5. *Перестановки с повторениями.* Пусть имеется группа из n объектов, состоящая из r подгрупп. Первая подгруппа содержит n_1 объектов, вторая – n_2 объектов и т.д., наконец, последняя группа содержит n_r объектов, причём внутри каждой подгруппы объекты не различимы. В таком случае число перестановок из n объектов определяется формулой:

$$N = \frac{n!}{n_1! n_2! \ldots n_r!}. \qquad (1.5)$$

Здесь $n_1 + n_2 + \ldots + n_r = n$. Заметим, что при $r = 2$ формула (1.5) совпадает с формулой (1.4).

Пример 1.7. На 6 карточках написаны цифры 2, 2, 3, 4, 4, 4. Сколько различных шестизначных чисел можно получить, располагая эти карточки последовательно?

Здесь $n = 6$, $r = 3$, при этом $n_1 = 2$, $n_2 = 1$, $n_3 = 3$. Следовательно, можно составить указанным в задаче способом $N = \frac{6!}{2! \cdot 1! \cdot 3!} = \frac{4 \cdot 5 \cdot 6}{2} = 60$ различных шестизначных чисел.

Для более сложных задач при подсчёте числа комбинаций следует действовать в соответствии с основным правилом комбинаторики.

1.3. Отношения между событиями

Пусть $\Omega = \{W_1, W_2, ..., W_n\}$ – пространство элементарных событий. Число событий, построенных из k элементарных событий, где $0 \le k \le n$, будет равно числу сочетаний из n по k. Исходя из этого, можно показать, что число N всех случайных событий, которые можно построить на Ω, равно $N = 2^n$. Рассмотрим операции над событиями.

1. Любое событие A является подмножеством пространства элементарных событий. Пусть в результате осуществления опыта произошло элементарное событие W. Если $W \in A$, то говорят, что событие A произошло, если, напротив, $W \notin A$, то произошло *противоположное* событие. Событие, противоположное по отношению к A, обозначается $\overline{A}$ (читается «не A»). Это событие называют ещё *отрицанием* события A.

2. Говорят, что событие A *влечёт за собой событие* B, если при наступлении события A событие B обязательно происходит. Записывают: $A \subset B$. Если $A \subset B$ и $B \subset A$, то события A и B называют *эквивалентными* (*равносильными*, *равными*). Записывают: $A = B$. Заметим, что по физическому содержанию события A и B могут быть различными.

3. Суммой или *объединением* двух событий A и B называется событие C, которое происходит тогда и только тогда, когда происходит или A, или B, или оба события одновременно. Это отражается записью: $C = A \cup B$ или $C = A + B$.

4. Произведением или *пересечением* двух событий A и B называется событие C, которое происходит тогда и только тогда, когда происходят и A и B одновременно. Записывают: $C = A \cap B$ или $C = AB$.

5. Разность событий A и B – это событие C, которое происходит тогда и только тогда, когда событие A происходит, а B не происходит. Записывают так: $C = A \setminus B$ или $C = A - B$.

6. События $A_1, A_2, \ldots, A_n$ представляют собой *полную группу событий*, если в результате опыта хотя бы одно из них обязательно произойдёт. Это эквивалентно тому, что сумма указанных событий равна достоверному событию:

$$A_1 + A_2 + \ldots + A_n = U. \tag{1.6}$$

7. События A и B называют *несовместными*, если они не могут произойти одновременно, т. е. $AB = V$. В противном случае события называют *совместными*. События $A_1, A_2, \ldots, A_n$ представляют *полную группу попарно несовместных событий*, если выполняются соотношения:

$$A_1 + A_2 + \ldots + A_n = U \text{ и } A_i A_k = V,\ i \ne k. \tag{1.7}$$

Приведём некоторые очевидные тождества:

$$\begin{aligned} &A + V = A, \quad A + U = U, \quad A + A = A, \quad A + \overline{A} = U, \\ &AV = V, \quad AU = A, \quad AA = A, \quad A\overline{A} = V \end{aligned} \tag{1.8}$$

Операции над событиями обладают свойствами коммутативности, ассоциативности и дистрибутивности:

$$\begin{gathered} A + B = B + A,\ AB = BA; \\ (A + B) + C = A + (B + C) = A + B + C,\ (AB)C = A(BC) = ABC; \\ A + (BC) = (A + B)(A + C),\ A(B + C) = AB + AC. \end{gathered} \tag{1.9}$$

Пример 1.8. Опыт: два стрелка одновременно стреляют по мишени. В этом случае пространство элементарных событий содержит 4 события: W_1 – мишень не поражена; W_2 – мишень поражена только 1-м стрелком; W_3 – мишень поражена только 2-м стрелком; W_4 – мишень поражена обеими стрелками. Таким образом, $\Omega = \{W_1, W_2, W_3, W_4\}$. На пространстве Ω можно построить $2^4 = 16$ событий:

$A_1 = \{W_1\}$, $A_2 = \{W_2\}$, $A_3 = \{W_3\}$, $A_4 = \{W_4\}$, $A_5 = \{W_1, W_2\}$,
$A_6 = \{W_1, W_3\}$, $A_7 = \{W_1, W_4\}$, $A_8 = \{W_2, W_3\}$, $A_9 = \{W_2, W_4\}$,
$A_{10} = \{W_3, W_4\}$, $A_{11} = \{W_1, W_2, W_3\}$, $A_{12} = \{W_2, W_3, W_4\}$, $A_{13} = \{W_1, W_3, W_4\}$,
$A_{14} = \{W_1, W_2, W_4\}$, $A_{15} = \{W_1, W_2, W_3, W_4\} = \Omega$, $A_{16} = \varnothing$.

Каждое событие из множества $\{A_1, A_2, \ldots, A_{16}\}$ имеет конкретный физический смысл. Например, событие A_7 означает, что мишень не была поражена или она была поражена двумя пулями. События A_8 – мишень поражена только одной пулей. Событие A_{12} означает, что мишень поражена, т.е. было или одно или два попадания.

Приведём примеры соотношений между событиями.

1. Противоположное событие: $A_7 = \overline{A}_8$; $A_1 = \overline{A}_{12}$; $A_5 = \overline{A}_{10}$; $A_9 = \overline{A}_6$; и т.д.
2. Сумма событий: $A_{14} = A_4 + A_5$; $A_{10} = A_3 + A_4$; $A_{14} = A_1 + A_9$; $A_{11} = A_5 + A_6$ и т.д.
3. Произведение событий: $A_1 = A_6 A_7$; $A_6 = A_{11} A_{13}$; $A_{10} = A_{10} A_{13}$; $A_7 = A_{13} A_{14}$ и т.д.
4. Разность событий: $A_2 = A_8 - A_3$; $A_8 = A_{12} - A_4$; $A_5 = A_{15} - A_{10}$; $A_4 = A_7 - A_1$ и т.д.

События A_3 и A_9 являются несовместными, $A_3 A_9 = V$, события A_2, A_5 и A_{10} составляют полную группу, так как $A_2 + A_5 + A_{10} = U$. События A_1, A_4 и A_8 представляют полную группу попарно несовместных событий: $A_1 + A_4 + A_8 = U$ и, кроме того, $A_1 A_4 = V$, $A_1 A_8 = V$, $A_4 A_8 = V$.

Пример 1.9. Событие D связано с событиями A, B и C следующей формулой:

$$D = AB + C + (A + B)(A + C).$$

Можно ли выразить эту связь более простым соотношением?

Воспользуемся свойствами операций над событиями и основными тождествами:

$$(A + B)(A + C) = AA + BA + AC + BC = A + AB + AC + BC =$$
$$= A(U + B + C) + BC = AU + BC = A + BC.$$

Это равенство является свойством дистрибутивности из (1.9). Таким образом,

$$D = AB + C + A + BC = A(B + U) + C(U + B) = AU + CU = A + C,$$

т.е. D зависит только от A и C.

1.4. Вероятность события

Вероятность события – это основное понятие теории вероятностей. Вообще *вероятность события* есть объективная мера возможности осуществления данного события.

Математическое определение вероятности. Каждому событию A ставится в соответствие некоторая мера $P(A)$, которая называется вероятностью этого события и которая удовлетворяет следующим аксиомам: для любого события $0 \le P(A) \le 1$; вероятность невозможного события равна нулю, $P(V) = 0$; вероятность достоверного события равна единице, $P(U) = 1$; если $AB = V$, то $P(A + B) = P(A) + P(B)$, т.е. если события A и B являются несовместными, то вероятность суммы событий равна сумме вероятностей событий.

Все рассматриваемые ниже определения являются, по существу, следствием математического определения вероятностей.

Классическое определение. Вероятностью $P(A)$ события A называют отношение числа исходов опыта N_A, приводящих к осуществлению события A, к общему числу исходов опыта N в предположении, что все исходы опыта являются равновозможными:

$$P(A) = \frac{N_A}{N}. \qquad (1.10)$$

Под исходами опыта здесь понимаются элементарные события, так что в терминах теории множеств N_A = Card(A), N = Card(Ω). Предположение о равной возможности исходов опыта вносит долю субъективизма в это определение, что и является его недостатком.

Пример 1.10. Бросается игральная кость. Какова вероятность того, что число выпавших очков будет не менее 5 (событие A)?

Игральная кость имеет 6 граней. Следовательно, общее число исходов опыта равно $N = 6$. К осуществлению события А приводят только 2 исхода, когда выпадает или 5 или 6 очков, т.е. $N_A = 2$. Таким образом, искомая вероятность равна $P(A) = 2/6 = 1/3$.

Пример 1.11. Из 10 теннисных мячей, среди которых 4 мяча новые, для очередной игры случайным образом берут три. Какова вероятность того, что среди взятых мячей два мяча будут новыми (событие A)?

Можно считать, что все результаты выбора являются равновозможными. Порядок выбора мячей не имеет значения. Это значит, что возможные исходы опыта следует рассматривать как сочетания. Так как взято 3 мяча из 10, то общее число исходов опыта будет равно C_{10}^3. К событию A приводят такие варианты, когда в число взятых мячей попали 2 мяча из четырёх новых (C_4^2) и один мяч из шести старых (C_6^1). Следовательно,

$$N = C_{10}^3 = \frac{10!}{3!7!} = \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} = 120, \;\; N_A = C_4^2 \cdot C_6^1 = \frac{4!}{2!2!} \cdot 6 = 36.$$

Тогда $P(A) = N_A/N = 36/120 = 0{,}3$.

Статистическое определение. Пусть опыт, с которым связано событие A, повторяется n раз, а событие A осуществляется при этом k раз ($0 \le k \le n$). Отношение k к n принято называть частотой появления события. Эта частота будет принимать различные значения в разных сериях из n опытов. При увеличении числа опытов n эти значения будут всё плотнее и плотнее группироваться вокруг некоторого (неизвестного) числа. Это число, очевидно, является некоторой объективной характеристикой, связывающей условия данного опыта с возможностью осуществления рассматриваемого события. Частота осуществления события k/n с увеличением числа повторений опыта стремится в определённом смысле к некоторому постоянному числу, которое и принимается за вероятность события A. Такое определение вероятностей называют *статистическим определением*. В данном виде оно не даёт точного правила вычисления вероятности события, так как n всегда конечно. При достаточно больших значениях n можно принять, что $P(A) \approx k/n$. Заметим, что это есть лишь оценка истинной вероятности.

Рассмотрим пример. Некоторый экспериментатор для определения вероятности появления «орла» при бросании монеты осуществил подбрасывание монеты 1000 раз. При этом оказалось, что «орёл» выпал $k_1 = 481$ раз. Оценка вероятности появления «орла» будет равна $P(A) \approx p_1^* = 0{,}481$. Второй экспериментатор также провёл серию из $n = 1000$ таких опытов и получил значение $k_2 = 512$. В этом случае $P(A) \approx p_2^* = 0{,}512$. Третий экспериментатор получил бы третью оценку. Наблюдается разброс оценок вероятности в разных сериях опытов. Разница в оценках уменьшается при увеличении числа n, но полное совпадение оценок при конечных n практически не возможно. Теоретически факт приближения частоты появления события к его вероятности обосновывается законом больших чисел.

Пример 1.12. Стрелок сделал 20 выстрелов по мишени, и было зафиксировано 12 попаданий. Оценить вероятность попадания в мишень (событие A) этим стрелком.

Здесь $n = 20$, а $k = 12$. Следовательно, $P(A) \approx 12/20 = 0{,}6$.

Геометрическое определение. Если пространство элементарных событий не является конечным множеством, то классическое определение вероятности оказывается не приемлемым. Пусть, например, некоторая точка случайным образом «бросается» в круг радиуса R. Требуется определить вероятность того, что эта точка окажется в круге меньшего радиуса r ($r < R$), который целиком находится в круге радиуса R. Исходом данного опыта являются координаты местоположения «вброшенной» точки. Предположим, что центр большого круга находится в начале координат. Тогда координаты (x, y) «вброшенной» точки удовлетворяют неравенству $x^2 + y^2 \le R^2$. Таким образом, данный круг представляет собой непрерывное множество исходов рассматриваемого опыта. Будем далее считать, что вероятность попадания точки в любую область не зависит от местоположения области в круге, а зависит только от её размера. Это предположение оправдано тем, что точка «бросается» в круг случайным образом. Таким образом, вероятность попадания точки в заданную область можно определить как отношение размера этой области к размеру всего круга. В рассмотренном примере искомая вероятность будет равна отношению площади маленького круга к площади большого круга, $\pi r^2/\pi R^2 = r^2/R^2$.

Решение задач с привлечением подобной геометрической интерпретации возможно не только на плоскости, но и на линии или в пространстве. Если исходы опыта образуют непрерывное множество и являются равновозможными в указанном выше смысле, то

$$P(A) = \frac{mes\{A\}}{mes\{\Omega\}}, \qquad (1.11)$$

где *mes* – геометрическая мера множества. Мерой может быть длина, площадь, объём. Равенство (1.11) представляет собой *геометрическое определение* вероятности. Оно охватывает и классическое определение, так как в случае конечных множеств в качестве меры множества можно взять число его элементов.

Пример 1.13. В круг радиуса r случайным образом ставится точка. Какова вероятность того, что эта точка окажется внутри вписанного в круг квадрата (событие A), если любое место расположения точки является равновозможным?

Здесь мера множеств – площадь. Множество всех элементарных событий представляет собой круг, а множество событий, приводящих к осуществлению события, образует квадрат. Таким образом, $mes\{\Omega\} = \pi r^2$, а $mes\{A\} = 2r^2$ (площадь вписанного квадрата). Следовательно, $P(A) = 2/\pi$.

В любом случае смысл вероятности состоит в следующем. Если событие имеет вероятность осуществления 0,4, то это означает, что при многократном повторении опыта событие произойдёт в среднем в 40% случаев.

1.5. Простейшие свойства вероятности

Свойство 1. Для вероятности любого события A выполняется неравенство:

$$0 \le P(A) \le 1. \qquad (1.12)$$

Действительно, если N_A есть число исходов опыта, приводящих к осуществлению A, а N – общее число исходов, то $0 \le N_A \le N$. Поделив это неравенство на N, получим (1.12).

Свойство 2. Вероятность невозможного события равна нулю:

$$P(V) = 0. \qquad (1.13)$$

Для невозможного события $N_V = 0$ и, следовательно, $P(V) = 0$.

Обратное утверждение не всегда верно. Если $P(A) = 0$, то нельзя утверждать, что A является невозможным событием. Предположим, что на отрезок [0; 2] случайным образом

ставится точка. Вероятность того, что точка попадёт в интервал $[x, x + \Delta x)$, будет равна $|\Delta x|/2$. При $\Delta x \to 0$ эта вероятность будет стремиться к нулю. Значит вероятность того, что точка будет иметь координату, совпадающую с x, равна нулю для любого значения x. Но в результате опыта точка будет иметь конкретную координату, т. е. произойдёт событие с нулевой вероятностью. Равенство $P(A) = 0$ при бесконечном числе исходов опыта нужно понимать в том смысле, что событие A имеет бесконечно малую вероятность и для его реализации опыт придётся повторять неограниченное число раз.

Свойство 3. Вероятность достоверного события равна единице, т.е.

$$P(U) = 1. \tag{1.14}$$

Для достоверного события $N_U = N$ и, следовательно, $P(U) = 1$.

Здесь также обратное утверждение не всегда справедливо. Действительно, если $P(A) = 1$, то $P(\overline{A}) = 0$. Событие $\overline{A}$, имея нулевую вероятность, может иногда произойти (см. примечание к свойству 2). Но в таком случае не произойдёт событие A, хотя $P(A) = 1$.

Свойство 4. Для несовместных событий A и B

$$P(A + B) = P(A) + P(B), \text{ если } AB = V. \tag{1.15}$$

Так как события A и B не пересекаются, то число исходов опыта N_{A+B}, приводящих к осуществлению событий A или B, равно сумме исходов N_A и N_B, приводящих соответственно к осуществлению событий A и B, т.е. $N_{A+B} = N_A + N_B$. Поделив это равенство на N, получим равенство (1.15). Это свойство называют свойством аддитивности вероятностей.

Вообще, если $A_1, A_2, \ldots, A_n$ группа попарно несовместных событий, т.е. $A_iA_k = V$, когда $i \neq k$, то вероятность суммы событий равна сумме вероятностей этих событий:

$$P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) \tag{1.16}$$

Если эта группа событий является полной группой, т.е. $A_1 + A_2 + \ldots + A_n = U$, то $P(A_1 + A_2 + \ldots + A_n) = P(U) = 1$, из чего следует, что

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1. \tag{1.17}$$

События A и $\overline{A}$ несовместны и составляют полную группу. Следовательно,

$$P(A) + P(\overline{A}) = 1 \text{ или } P(A) = 1 - P(\overline{A}). \tag{1.18}$$

Свойство 5. Если $A \subset B$, то $P(A) \leq P(B)$.

Соотношение $A \subset B$ означает, что множество A составляет часть множества B. Поэтому всегда $N_A \leq N_B$. Поделив это неравенство на N, получим подтверждение свойства 5.

Пример 1.14. Из 10 деталей, среди которых имеется 4 нестандартных, случайным образом взято 3. Найти вероятность того, что в числе взятых деталей бракованных окажется: 1) не менее 2 (событие B); 2) хотя бы одна (событие C).

Пусть A_i – событие, состоящее в том, что среди взятых деталей оказалось i нестандартных ($i = 0, 1, 2, 3$). Тогда $B = A_2 + A_3$ («не менее 2» означает или 2, или 3). События A_i, $i = 0, 1, 2, 3$, являются попарно несовместными. Поэтому $P(B) = P(A_2 + A_3) = P(A_2) + P(A_3)$. Найдём вероятности $P(A_2)$ и $P(A_3)$, пользуясь классической схемой. Обозначим через N_i число исходов опыта, приводящих к осуществлению события A_i ($i = 0, 1, 2, 3$). Определим N_2, N_3, а также общее число исходов N: $N_2 = C_4^2 C_6^1 = 36$; $N_3 = C_4^3 C_6^0 = 4$; $N = C_{10}^3 = 120$. Тогда, $P(A_2) = N_2/N = 0,3$; $P(A_3) = N_3/N = 1/30$. Следовательно, $P(A_3) = 0,3 + 1/30 = 1/3$.

Очевидно, что события A_0 и C являются противоположными событиями, т.е. $C = \overline{A}_0$.

По формуле (1.18) $P(C) = P(\overline{A}_0) = 1 - P(A_0) = 1 - \dfrac{N_0}{N} = 1 - \dfrac{C_4^0 C_6^3}{120} = 1 - \dfrac{5 \cdot 4}{120} = 1 - \dfrac{1}{6} = \dfrac{5}{6}$.

1.6. Условная вероятность. Формула умножения вероятностей. Зависимые и независимые события

Условной вероятностью события A по отношению к событию B называется вероятность события A, вычисленная при условии, что событие B произошло. Обозначается: $P(A \mid B)$ (кратко читается так: "P от A при условии B"). Условная вероятность обладает всеми свойствами безусловной вероятности. В частности, $P(A \mid B) + P(\overline{A} \mid B) = 1$. Отметим также, что если $B \subset A$, то $P(A \mid B) = 1$; если $AB = V$, то $P(A \mid B) = 0$.

Рассмотрим задачу вычисления условной вероятности в общем случае. Преобразуем вероятность $P(AB)$ следующим образом: $\mathrm{P(AB)} = \mathrm{N_{AB}} / \mathrm{N} = (\mathrm{N_B} / \mathrm{N}) \cdot (\mathrm{N_{AB}} / \mathrm{N})$. Вторая дробь в этом выражении представляет собой условную вероятность $P(A|B)$. Действительно, поскольку событие B произошло, то имел место один из исходов, приводящих к событию B, из общего числа N_B. Первая дробь есть не что иное, как $P(B)$. Таким образом,

$$P(AB) = P(B) \cdot P(A|B). \tag{1.19}$$

Эта формула называется *формулой умножения вероятностей*. Из этой формулы получаем правило вычисления условной вероятности:

$$P(A|B) = \frac{P(AB)}{P(B)}. \tag{1.20}$$

Аналогично можно записать формулы, содержащие условную вероятность $P(B \mid A)$:

$$P(AB) = P(A) \cdot P(B|A), \; P(B|A) = \frac{P(AB)}{P(A)}. \tag{1.21}$$

Если условная вероятность совпадает с безусловной вероятностью, т.е. $P(A \mid B) = P(A)$, то события A и B называются *статистически независимыми событиями*. Это значит, что вероятность осуществления события A не зависит от того, произошло или нет событие B. Если $P(A \mid B) \neq P(A)$, то события A и B называются *статистически зависимыми событиями*. Заметим, что статистическая зависимость или независимость является взаимной. Если $P(A \mid B) = P(A)$, то обязательно выполняется и равенство $P(B \mid A) = P(B)$.

Формула умножения вероятностей обобщается для произвольного числа событий:

$$P(A_1A_2 \ldots A_n) = P(A_1 \ldots A_{n-1})P(A_n \mid A_1 \ldots A_{n-1}), \tag{1.22}$$

где $P(A_n \mid A_1 \ldots A_{n-1})$ – условная вероятность события A_n по отношению к произведению всех остальных событий, т. е. вероятность события A_n, вычисленная при условии, что события A_1, A_2, …, A_{n-1} уже произошли. Безусловную вероятность в (1.22) можно представить формулой умножения вероятностей. Повторяя это действие n – 1 раз, получим:

$$P(A_1A_2 \ldots A_n) = P(A_1)P(A_2 \mid A_1) \ldots P(A_{n-1} \mid A_1 \ldots A_{n-2})\, P(A_n \mid A_1 \ldots A_{n-1}).$$

В частности, при $n = 3$,

$$P(A_1A_2A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2). \tag{1.23}$$

Будем называть события A_1, A_2, …, A_n *статистически независимыми*, если условная вероятность любого события по отношению к любому произведению из остальных событий, совпадает с безусловной вероятностью данного события. Другими словами, вероятность осуществления любого события не зависит от того, произошли или не произошли другие события. Для произведения независимых событий справедливо равенство:

$$P(A_1A_2 \ldots A_n) = P(A_1)\, P(A_2) \ldots P(A_n). \tag{1.24}$$

В частности, если A и B – статистически независимые события, то

$$P(AB) = P(A) \cdot P(B). \tag{1.25}$$

Пример 1.15. Из коробки, содержащей 3 белых и 5 чёрных шаров, наугад взяли 3 шара. Какова вероятность того, что все 3 шара оказались чёрного цвета?

Пусть A_i есть событие, состоящее в том, что i-й извлечённый шар оказался чёрным (i = 1, 2, 3). Тогда искомая вероятность есть вероятность произведения событий $P(A_1A_2A_3)$.

По формуле умножения вероятностей $P(A_1A_2A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)$.

Когда берётся первый шар, то в коробке находится 8 шаров, из которых 5 шаров имеют чёрный цвет. Следовательно, $P(A_1) = 5/8$. Когда берётся второй шар при условии, что событие A_1 уже произошло, то в коробке находится 7 шаров, причём чёрных шаров осталось 4 (один уже был взят). Значит $P(A_2 \mid A_1) = 4/7$. После извлечения второго шара в коробке осталось 6 шаров, в том числе 3 чёрных. Таким образом, $P(A_3 \mid A_1A_2) = 3/6$. Искомая вероятность равна: $P(A_1A_2A_3) = (3/6) \cdot (4/7) \cdot (5/8) = 5/28$.

Эту вероятность можно получить и непосредственно по формуле для числа сочетаний.

Пример 1.16. Два стрелка одновременно, независимо друг от друга стреляют по мишени. Вероятность попадания в мишень для первого стрелка равна 0,6, а для второго – 0,8. Какова вероятность того, что в мишени окажется две пробоины (событие C)?

Пусть событие A – попадание в мишень первым стрелком, событие B – попадание в мишень вторым стрелком. Тогда $C = AB$, причём события A и B являются (по условию задачи) независимыми. Следовательно, $P(AB) = P(A) \cdot P(B) = 0{,}6 \cdot 0{,}8 = 0{,}48$.

1.7. Формула сложения вероятностей

Для любых событий A и B выполняется равенство:

$$P(A + B) = P(A) + P(B) - P(AB). \qquad (1.26)$$

Это *формула сложения вероятностей*. Она справедлива для любых событий. Действительно, если A и B являются несовместными, то $A \cdot B = V$ и $P(AB) = P(V) = 0$. При этом получается формула сложения вероятностей несовместных событий, записанная ранее:

$$P(A + B) = P(A) + P(B). \qquad (1.27)$$

Формула (1.26) легко доказывается. Если N_A – число исходов опыта, приводящих к осуществлению события A, а N_B – к осуществлению события B, то в сумме $N_A + N_B$ будут дважды учтены те исходы, которые одновременно приводят к осуществлению и A и B. Число исходов, приводящих к осуществлению $A + B$, равно $N_A + N_B - N_{AB}$, где N_{AB} – число исходов, приводящих к осуществлению AB. Эти рассуждения приводят к формуле (1.26).

Применяя дважды формулу (1.26), для трёх событий получим:

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC). \qquad (1.28)$$

Вероятность суммы большого числа событий вычисляют обычно через противоположное событие. Сумма событий $A_1 + A_2 + \ldots + A_n$ есть событие, состоящее в том, что произойдёт хотя бы одно из суммируемых событий. Противоположным по отношению к нему является событие $\overline{A}_1\overline{A}_2 \ldots \overline{A}_n$. Это значит, что $\overline{A}_1\overline{A}_2 \ldots \overline{A}_n + A_1 + A_2 + \ldots + A_n = U$. События $A_1 + A_2 + \ldots + A_n$ и $\overline{A}_1\overline{A}_2 \ldots \overline{A}_n$ являются несовместными, а $P(U) = 1$. Поэтому

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A}_1\overline{A}_2 \ldots \overline{A}_n). \qquad (1.29)$$

Пример 1.17. Два стрелка одновременно, независимо друг от друга стреляют по мишени. Вероятность попадания в мишень для первого стрелка равна 0,6, а для второго стрелка – 0,8. Какова вероятность того, что мишень будет поражена (событие C)?

Мишень поражена, если будет хотя бы одно попадание. Это значит, что $C = A + B$. По условию задачи события A и B статистически независимы. Следовательно, $P(C) = P(A + B) = P(A) + P(B) - P(AB) = P(A) + P(B) - P(A)P(B) = 0{,}6 + 0{,}8 - 0{,}6 \cdot 0{,}8 = 0{,}92$. С другой сторо-

ны, вероятности противоположных по отношению к A и B событий соответственно равны 0, 4 и 0, 2. По формуле (1.29) получим: $P(C) = P(A + B) = 1 - 0{,}4 \cdot 0{,}2 = 1 - 0{,}08 = 0{,}92$.

1.8. Формула полной вероятности и формула Байеса

В некоторых задачах имеется дополнительная неопределённость, которая не позволяет непосредственно определить искомую вероятность. Формула полной вероятности и формула Байеса применяются при решении одного класса таких задач.

Рассмотрим следующую задачу. Пусть события H_1, H_2, …, H_n составляют полную группу попарно несовместных событий, т. е.

$$H_1 + H_2 + \ldots + H_n = U, \tag{1.30}$$

причём, $H_iH_k = V$, если только $i \neq k$. Пусть, далее, имеется некоторое событие A, которое может произойти с одним из событий H_i, $i = 1, 2, \ldots, n$. Вероятности $P(H_i)$ и $P(A \mid H_i)$ являются известными. Требуется вычислить вероятность $P(A)$.

Умножим равенство (1.30) на событие A, имея в виду, что $AU = A$:

$$A = AU = A(H_1 + H_2 + \ldots + H_n) = AH_1 + AH_2 + \ldots + AH_n.$$

Имеем сумму попарно несовместных событий: если $i \neq k$, то $AH_iAH_k = AH_iH_k = AV = V$. Поэтому, $P(A) = P(AH_1 + AH_2 + \ldots + AH_n) = P(AH_1) + P(AH_2) + \ldots + P(AH_n)$. Кроме того, $P(AH_1) = P(A \mid H_i)\, P(H_i)$. Получили формулу, которую называют *формулой полной вероятности*:

$$P(A) = \sum_{i=1}^{n} P(A|H_i) \cdot P(H_i)\,. \tag{1.31}$$

Заметим, что всегда $\sum_{i=1}^{n} P(H_i) = 1$, так как $H_1 + H_2 + \ldots + H_n = U$.

Пример 1.18. Три автоматические линии изготовляют однотипные изделия и работают на общий конвейер. Производительности 1-ой, 2-ой и 3-ей линий находятся в соотношении 2 : 3 : 5. Вероятность изготовления дефектного изделия на 1-ой линии равна 0,05, для 2-ой линии эта вероятность равна 0, 08, для 3-ей – 0,1. С общего конвейера наугад берётся изделие. Какова вероятность $P(A)$ того, что это изделие не имеет дефектов?

Возможны три гипотезы: H_1 – изделие было изготовлено на 1-ой линии, H_2 – на 2-ой, H_3 – на 3-ей. Из соотношения производительностей линий находим, что $P(H_1) = 0{,}2$; $P(H_2) = 0{,}3$; $P(H_3) = 0{,}5$. Вероятности изготовления бездефектного изделия на линиях равны 1 – 0,05 = 0,95, т.е. $P(A \mid H_1) = 0{,}95$, $P(A \mid H_2) = 0{,}92$ и $P(A \mid H_3) = 0{,}9$. Тогда $P(A)$ равна:

$$P(A) = P(A \mid H_1)P(H_1) + P(A \mid H_2)P(H_2) + P(A \mid H_3)P(H_3) =$$
$$= 0{,}2 \cdot 0{,}95 + 0{,}3 \cdot 0{,}92 + 0{,}5 \cdot 0{,}9 = 0{,}916.$$

Этот результат означает, что в среднем 916 изделий из 1000 будут без дефектов.

Пусть в результате опыта событие A произошло. Какое из событий H_1, H_2, …, H_n стало причиной осуществления события A? Точно на этот вопрос ответить невозможно, но можно определить вероятность того, что причиной стало то или иное событие.

По формуле умножения вероятностей $P(AH_k) = P(A \mid H_k)P(H_k) = P(H_k \mid A)\, P(A)$. Отсюда:

$$P(H_k|A) = \frac{P(A|H_k)P(H_k)}{P(A)} \quad \text{или} \quad P(H_k|A) = \frac{P(A|H_k) \cdot P(H_k)}{\sum_{i=1}^{n} P(A|H_i) \cdot P(H_i)}\,. \tag{1.32}$$

Формулу (1.32) называют *формулой Байеса*. По формуле Байеса находится вероятность того, что имела место гипотеза H_k при условии, что событие A произошло.

Пример 1.19. Группа из 12 стрелков включает в себя трёх мастеров спорта, четырёх кандидатов в мастера и пятерых перворазрядников. Мастер спорта может попасть в мишень с вероятностью 0,9, кандидат в мастера – с вероятностью 0,85, перворазрядник – с вероятностью 0,75. Наудачу выбранный стрелок сделал выстрел, мишень была поражена (событие A). Какова вероятность того, что этот стрелок является мастером спорта?

Неизвестно, какой стрелок сделал выстрел. Имеются гипотезы: H_1 – стрелял мастер спорта, H_2 – стрелял кандидат в мастера, H_3 – стрелял перворазрядник. По условию, $P(H_1) = 3/12$, $P(H_2) = 4/12$, $P(H_3) = 5/12$. Кроме того, даны вероятности попадания для стрелков каждой группы: $P(A \mid H_1) = 0{,}9$, $P(A \mid H_2) = 0{,}85$, $P(A \mid H_3) = 0{,}75$. Таким образом,

$$P(A) = P(A \mid H_1)P(H_1) + P(A \mid H_2)P(H_2) + P(A \mid H_3)P(H_3) =$$
$$= (0,9 \cdot 3 + 0,85 \cdot 4 + 0,75 \cdot 5)/12 \approx 0,821.$$

Тогда, по формуле (1.32), $P(H_1 \mid A) = \dfrac{P(A \mid H_1)P(H_1)}{P(A)} = \dfrac{0,9 \cdot 3}{12 \cdot 0,821} \approx 0,274$.

Вероятности $P(H_i)$, $i = 1, 2, \ldots, n$, называют *априорными вероятностями* (вычисленными до опыта), а вероятности $P(H_i \mid A)$ – *апостериорными* (вычисленными после опыта).

1.9. Последовательность независимых испытаний. Формула Бернулли

Пусть некоторый опыт (испытание) повторяется n раз. Будем считать, что вероятность осуществления события A, связанного с данным опытом, при каждом повторении опыта остаётся неизменной и равна p. Тогда вероятность того, что событие A не осуществится, также будет неизменной и равной $q = 1 - p$. Такая последовательность проведения одного и того же опыта называется последовательностью независимых испытаний. Независимость понимается в том смысле, что вероятность осуществления события A в любом по номеру повторении опыта не зависит от результатов опыта при всех других повторениях. Найдём вероятность $P_n(k)$ того, что событие A в n испытаниях произойдёт k раз.

Пронумеруем повторяемые опыты целыми числами от 1 до n и обозначим через A_i событие, состоящее в том, что в i-ом опыте событие A произошло. Тогда $P(A_i) = p$ и $P(\overline{A}_i) = q$) для любого i. Если принять всю последовательность испытаний за один опыт, то результатами этого опыта будут события вида $C_j = A_1 A_2 \overline{A}_3 A_4 \overline{A}_5 \ldots \overline{A}_{n-2} A_{n-1} \overline{A}_n$.

Так как на любом i-ом месте этого произведения может стоять либо A_i, либо $\overline{A}_i$, то всего таких событий будет 2^n (по общему правилу комбинаторики), т.е. $j = 1, 2, 3, \ldots, 2^n$. Рассмотрим только те из них, которые содержат k множителей типа A_i и n-k множителей типа $\overline{A}_i$, причём взаимное расположение различных типов множителей не имеет значения. Анализ показывает, что число таких событий равно числу сочетаний из n объектов по k. Все события C_j являются попарно несовместными. Так как испытания независимы, то вероятность любого события C_j равна произведению вероятностей всех множителей. Таким образом, вероятность любого события C_j, содержащего k множителей типа A_i, будет равна $P(C_j) = p^k q^{n-k}$, следовательно,

$$\mathrm{P_n(k)} = \sum_j P(C_j) = C_n^k \cdot p^k q^{n-k}.$$

Здесь суммирование проводится только по тем значениям j, для которых число множителей типа A_i в событии C_j равно k.

Итак, вероятность появления события A в n независимых испытаниях k раз равна

$$P_n(k) = C_n^k \cdot p^k q^{n-k}, \qquad (1.33)$$

где $q = 1 - p$, а $C_n^k = \dfrac{n!}{k!\cdot(n-k)!}$ – число сочетаний из n элементов по k. Эта формула называется *формулой Бернулли*. Рассмотрим следствия, вытекающие из формулы Бернулли.

1. Вероятность того, что в n независимых испытаниях событие A произойдёт не менее чем i раз, равна

$$P\{k \ge i\} = \sum_{k=i}^{n} C_n^k \cdot p^k q^{n-k} \tag{1.34}$$

или

$$P\{k \ge i\} = 1 - \sum_{k=0}^{i-1} C_n^k \cdot p^k q^{n-k} \ . \tag{1.35}$$

Действительно, пусть B_k – событие, состоящее в том, что событие A произойдёт k раз. Тогда событие $B_i + B_{i+1} + \ldots + B_n$ означает, что A произойдёт не менее чем i раз. Противоположным этому событию является событие $B_0 + B_1 + \ldots + B_{i-1}$. Слагаемые в суммах – попарно несовместные события. Отсюда,

$$P\{k \ge i\} = P(B_i + B_{i+1} + \ldots + B_n) = \sum_{k=i}^{n} P(B_k)$$

или, в соответствии с формулой (1.18), $P\{k \ge i\} = 1 - P(B_0 + B_1 + \ldots + B_{i-1}) = 1 - \sum_{k=0}^{i-1} P(B_k)$.

Эти соотношения и доказывают справедливость формул (1.34) и (1.35).

2. Вероятность того, что событие A произойдёт хотя бы один раз в n независимых испытаниях, определяется формулой (частный случай формулы (1.35) при $i = 1$):

$$P\{\, k \ge 1 \,\} = 1 - q^n. \tag{1.36}$$

3. Если выбрать такое n, что $n \ge \dfrac{|\ln(1 - P)|}{|\ln(1 - p)|}$, (1.37)

то с вероятностью, не меньшей P, событие A произойдёт в испытаниях хотя бы один раз.

Действительно, положим $P\{k \ge 1\} = 1 - q^n \ge P$. Тогда $1 - P \ge q^n$. Прологарифмируем это неравенство: $\ln(1 - P) \ge n\cdot\ln q$ или $\ln(1 - P) \ge n\cdot\ln(1 - p)$. Поделив последнее неравенство на $\ln(1 - p)$ и учитывая, что оба логарифма отрицательны, получим формулу (1.37). В качестве n берётся ближайшее целое число, превосходящее значение правой части неравенства.

Пример 1.20. Проводятся стендовые испытания шести приборов, каждый из которых, независимо от других, с вероятностью $p = 0{,}4$ может быть признан годным без дополнительных регулировок. Событие B – только два прибора не потребуют регулировок, событие C – хотя бы два прибора не потребуют регулировок. Найти $P(B)$ и $P(C)$.

Очевидно, имеет место схема независимых последовательных испытаний при $n = 6$, $p = 0{,}4$, $q = 0{,}6$. Тогда вероятность $P(B)$ определяется непосредственно формулой Бернулли:

$$P(B) = P_6(2) = C_6^2 p^2 q^4 = \frac{6!}{2!\cdot 4!}\cdot 0{,}4^2 \cdot 0{,}6^4 = 15 \cdot 0{,}16 \cdot 0{,}1296 \approx 0{,}311 .$$

Вероятность события C вычислим через вероятность противоположного события:

$$P(C) = 1 - P(\overline{C}) = 1 - P\{k < 2\} = 1 - P_6(0) - P_6(1) = 1 - C_6^0 p^0 q^6 -$$
$$- C_6^1 p^1 q^5 = 1 - 0{,}6^6 - 6 \cdot 0{,}4 \cdot 0{,}6^5 \approx 1 - 0{,}047 - 0{,}187 = 0{,}766.$$

Пример 1.21. Вероятность попадания в мишень при одном выстреле равна 0,3. Сколько нужно сделать выстрелов, чтобы с вероятностью, не меньшей чем 0, 95, мишень имела хотя бы одну пробоину?

Здесь достаточно воспользоваться формулой (1.37) при $p = 0{,}3$ и $P = 0{,}95$:

$$n \geq \frac{|\ln(1-P)|}{|\ln(1-p)|} = \frac{|\ln(1-0{,}95)|}{|\ln(1-0{,}3)|} = \frac{|\ln(0{,}05)|}{|\ln(0{,}7)|} \approx \frac{2{,}9957}{0{,}3567} \approx 8{,}398\,.$$

Число n должно быть целым, поэтому $n = 9$. Таким образом, если сделать 9 выстрелов, то с вероятностью, не меньшей чем $P = 0{,}95$, мишень получит хотя бы одну пробоину.

Отметим, что вероятности $P_n(k)$ совпадают с соответствующими членами разложения бинома $(p + q)^n$ по степеням p: $(p+q)^n = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = \sum_{k=0}^{n} P_n(k)$. Поэтому распределение вероятностей (1.33) (по различным значениям k) называют биномиальным законом распределения вероятностей. Из этого соотношения следует, что сумма всех вероятностей равна единице (условие нормировки вероятностей), так как $(p + q)^n = 1^n = 1$.

Каждое из чисел k ($k = 0, 1, 2, \ldots, n$) имеет свою вероятность при проведении последовательности независимых испытаний. Число s, которому соответствует наибольшая вероятность $P_n(k)$, называется *наивероятнейшим числом* появлений события в последовательности n независимых испытаний. По определению, $P_n(s) \geq P_n(s-1)$ и $\mathrm{P_n(s)} \geq \mathrm{P_n(s+1)}$.

Раскрывая эти неравенства и упрощая их, получим:

$$np - q \leq s \leq np + p. \qquad (1.38)$$

Если границы промежутка определяются дробными числами, то в промежутке имеется одно целое число, которое и является наивероятнейшим числом. Если границы определяются целыми числами, то существует два наивероятнейших числа: $s_1 = np - q$ и $s_2 = np + p$.

Пример 1.22. Вероятность появления события A в каждом испытании одна и та же и равна 0, 7. Определить наивероятнейшее число появлений события в n испытаниях и вероятность этого числа при $n = 8$ и при $n = 9$.

Положим $n = 8$, $p = 0{,}7$, $q = 0{,}3$. Подставляя в неравенство (1.38) эти данные, получим $5{,}3 \leq s \leq 6{,}3$. Следовательно, $s = 6$ и $P_8(6) = C_8^6 \cdot 0{,}7^6 \cdot 0{,}3^2 \approx 28 \cdot 0{,}1176 \cdot 0{,}09 \approx 0{,}296$.

Запишем неравенство для $n = 9$: $9 \cdot 0{,}7 - 0{,}3 \leq s \leq 9 \cdot 0{,}7 + 0{,}7$ или $6 \leq s \leq 7$. Следовательно, существует два наивероятнейших числа: $s_1 = 6$ и $s_2 = 7$. Вероятности этих чисел одинаковы: $P_9(7) = P_9(6) = C_9^6 \cdot 0{,}7^6 0{,}3^3 \approx 84 \cdot 0{,}1176 \cdot 0{,}0027 \approx 0{,}267$.

1.10. Асимптотические приближения формулы Бернулли

Формулой Бернулли пользоваться неудобно, если число испытаний достаточно велико. Рассмотрим приближённые формулы, к которым асимптотически приближается формула Бернулли при неограниченном увеличении числа испытаний ($n \to \infty$).

Локальная предельная теорема Муавра – Лапласа утверждает следующее. При неограниченном возрастании n вероятность того, что в n независимых испытаниях событие наступит k раз, определяется приближённым равенством:

$$P_n(k) \approx \frac{1}{\sqrt{npq}} \cdot f_0(x), \text{ где } f_0(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\{-\frac{x^2}{2}\}, \text{ а } x = \frac{k - np}{\sqrt{npq}} \qquad (1.39)$$

Точность этого приближённого равенства тем выше, чем больше число испытаний n. Наибольшая точность достигается при значениях p, близких к 0,5.

Интегральная предельная теорема Муавра – Лапласа является, в определённом смысле, продолжением локальной. Она утверждает, что при достаточно больших n, вероят-

ность того, что число k появлений события в n независимых испытаниях окажется в интервале $(k_1; k_2)$, определяется приближенно равенством:

$$P\{k_1 \le k < k_2\} \approx \Phi_0(x_2) - \Phi_0(x_1), \quad x_1 = \frac{k_1 - np}{\sqrt{npq}}, \quad x_2 = \frac{k_2 - np}{\sqrt{npq}}, \tag{1.40}$$

где $\Phi_0(x)$ – функция Лапласа: $\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\{-\frac{t^2}{2}\} dt$. (1.41)

Формула (1.40) даёт наибольшую точность также при $p = q = 0,5$.

Для функций $f_0(x)$ и $\Phi_0(x)$ имеются подробные таблицы, что существенно облегчает процесс вычислений. Функция $f_0(x)$ является чётной, а $\Phi_0(x)$ – нечётной функцией: $f_0(-x) = f_0(x)$, $\Phi_0(-x) = -\Phi_0(x)$. Поэтому таблицы приводятся только для $x \ge 0$.

Пример 1.23. Вероятность рождения мальчика равна 0,51. Найти вероятность того, что среди 80 новорождённых окажется 45 мальчиков.

Здесь $p = 0,51$, $q = 0,49$, $n = 80$ и $k = 45$. Применим локальную предельную теорему Муавра – Лапласа: $np = 40,8$, $npq = 19,992$, $\sqrt{npq} \approx 4,4712$, $x = \frac{k - np}{\sqrt{npq}} \approx \frac{4,2}{4,4712} \approx 0,9393$.

По таблице значений функции $f_0(x)$ для $x = 0,9393$ находим $f_0(x) \approx 0,2567$. Получим:

$$P_n(k) \approx \frac{1}{\sqrt{npq}} \cdot f_0(x) \approx \frac{1}{4,4712} \cdot 0,2567 \approx 0,0574.$$

Пример 1.24. Завод выпускает 80% изделий первого сорта и 20% изделий второго сорта. Наугад взято 100 штук из общего потока изделий. Какова вероятность того, что число изделий первого сорта среди выбранных будет находиться в пределах от 75 до 90?

Применим интегральную предельную теорему, полагая $n = 100$, $p = 0,8$ и $q = 0,2$:

$$np = 80,\ npq = 16,\ \sqrt{npq} = 4,\ k_1 = 75,\ k_2 = 90,\ x_1 = -1,25,\ x_2 = 2,5.$$

Следовательно, $P\{k_1 \le k < k\} \approx \Phi_0(x_2) - \Phi_0(x_1) = \Phi_0(2,5) - \Phi_0(-1,25) = \Phi_0(2,5) + \Phi_0(1,25)$. По таблице значений функции $\Phi_0(x)$ для $x_1 = 1,25$ и $x_2 = 2,5$ находим $\Phi_0(1,25) = 0,3944$ и $\Phi_0(2,5) = 0,4938$. Таким образом, искомая вероятность равна $P\{75 \le k < 90\} = 0,4938 + 0,3944 = 0,8882$.

При больших значениях n и достаточно малых p рекомендуется пользоваться *теоремой Пуассона*, согласно которой

$$P_n(k) \approx \frac{\lambda^k}{k!} \cdot \exp\{-\lambda\}, \text{ где } \lambda = np. \tag{1.42}$$

Пример 1.25. Вероятность повреждения изделия при транспортировке равна 0,001. Какова вероятность того, что при транспортировке 2000 изделий будет повреждено не более двух из них (событие A)?

Так как n велико, а p мало, то можно воспользоваться теоремой Пуассона: $n = 2000$, $p = 0,001$, $np = \lambda = 2$. Тогда

$$P(A) = P\{k = 0\} + P\{k = 1\} + P\{k = 2\} =$$

$$= \frac{2^0}{0!} \cdot \exp\{-2\} + \frac{2^1}{1!} \cdot \exp\{-2\} + \frac{2^2}{2!} \cdot \exp\{-2\} = 5 \cdot \exp\{-2\} \approx 0,6765.$$

Термины для запоминания

- Размещения, перестановки, сочетания.
- Случайное событие, достоверное событие, невозможное событие, элементарное событие, пространство элементарных событий, противоположное событие, сумма событий, произведение событий, совместные и несовместные события.
- Вероятность события, классическое определение вероятности, статистическое определение вероятности, геометрическое определение вероятности, аксиоматическое определение вероятности, формула сложения вероятностей, формула умножения вероятностей, формула полной вероятности, формула Байеса.
- Последовательность независимых испытаний, формула Бернулли, наивероятнейшее число, асимптотические приближения формулы Бернулли, локальная предельная теорема Муавра – Лапласа, интегральная предельная теорема Муавра – Лапласа, теорема Пуассона.

Список контрольных вопросов

1. Чем отличаются комбинации размещения и комбинации сочетания?
2. Дайте определение суммы и произведения случайных событий.
3. Сформулируйте классическое определение вероятности.
4. Сформулируйте геометрическое определение вероятности.
5. Запишите формулы сложения вероятностей для совместных и несовместных событий. Чем они отличаются и почему?
6. Чему равна вероятность суммы событий, составляющих полную группу?
7. Что такое условная вероятность события?
8. Какие события называются статистически независимыми?
9. Запишите формулы умножения вероятностей для зависимых и независимых событий. Чем они отличаются и почему?
10. Запишите формулу полной вероятности и формулу Байеса.
11. Каким условиям должны удовлетворять гипотезы в формуле полной вероятности и в формуле Байеса?
12. Что называется последовательностью независимых испытаний? В каком смысле понимается здесь независимость?
13. Запишите формулу Бернулли и объясните, как она получается.
14. При каких условиях применяются предельные теоремы Муавра – Лапласа для приближённой замены формулы Бернулли?
15. При каких условиях применяются предельные теоремы Пуассона для приближённой замены формулы Бернулли?
16. Что такое наивероятнейшее число появлений события в последовательности независимых испытаний и как оно определяется?

Вопросы для обсуждений

1. Попробуйте найти в окружающем нас мире неслучайные (на первый взгляд) события.
2. Попробуйте затем аргументированно опровергнуть это мнение о данном событии.

ТЕМА 2.

Случайные величины

Студент должен
освоить:

- понятие случайной величины, законы распределения вероятностей, числовые характеристики случайных величин;

приобрести навыки:

- решения задач на вычисление вероятностей попадания случайной величины в заданный интервал, задач на вычисление числовых характеристик.

Краткое содержание

Случайные величины: определение, классификация, способы задания. Функция распределения вероятностей случайной величины и ее свойства. Плотность распределения вероятностей случайной величины и ее свойства. Функция случайной величины. Математическое ожидание функции случайной величины. Числовые характеристики случайных величин. Математическое ожидание случайной величины и его свойства. Мода и медиана. Дисперсия и среднее квадратическое отклонение случайной величины и их свойства. Коэффициент вариации. Квантили, квартили и вероятное отклонение. Моменты случайных величин. Скошенность и эксцесс. Производящие функции. Примеры дискретных и непрерывных распределений (биномиальное, пуассоновское, равномерное, показательное, нормальное, Эрланга, Вейбулла).

2.1. Определение, классификация, способы задания случайных величин

Случайной величиной называется величина, принимающая в результате опыта одно из своих возможных значений, причём заранее не известно, какое именно значение она примет. Случайные величины обозначают большими латинскими буквами (X, Y, Z_1, …), а их возможные значения – соответствующими малыми буквами (x, y, z, …). Событие, состоящее в том, что случайная величина X примет значение x, будем отражать записью $X = x$. С левой стороны этого равенства стоит имя случайной величины, а справа – принимаемое ею значение. Вероятность этого события будем обозначать $P\{X = x\}$. Аналогично, $P\{X < x\}$ – вероятность того, что случайная величина X примет значение, меньшее, чем x.

Случайные величины делятся на дискретные, непрерывные и на величины смешанного типа. Возможные значения *дискретной* случайной величины составляют конечное или бесконечное, но счётное, множество. Возможные значения *непрерывной* случайнойвеличины составляют непрерывное множество. Все возможные значения дискретной случайной величины можно пронумеровать (хотя бы в принципе, если они составляют бесконечное множество), для непрерывной случайной величины это сделать невозможно. *Случайная величина смешанного типа* кроме непрерывного множества возможных значений имеет ещё возможные значения, изолированные от этого множества.

Совокупность всех возможных значений случайной величины с соответствующими им вероятностями называется *законом распределения вероятностей* этой случайной величины. Перечислить возможные значения и указать их вероятности можно таблично, графически или аналитически. Если $p_k = P\{X = x_k\}$ – есть вероятность того, что случайная величина X примет значение x_k (k = 1, 2, ..., n), то таблица из двух строк представляет *табличное задание* величины X:

x_k	x_1	x_2	……	x_n
p_k	p_1	p_2	……	p_k

При *аналитическом задании* связь между возможными значениями случайной величины и их вероятностями задаётся с помощью математической формулы, например:

$$P\{X = k\} = p_k = p(1-p)^k,\ k = 0,1,2,\ldots \qquad (2.1)$$

здесь $x_k = k$, $0 < p < 1$. Если эта связь задана с помощью графика, то говорят, что случайная величина задана *графически* (рис. 2.1). В любом случае, если общее число возможных значений случайной величины равно n, то $\sum_{k=1}^{n} p_k = 1$ – *условие нормировки вероятностей*.

Пример 2.1. Выпущено 1000 лотерейных билетов: на 5 из них выпадает выигрыш в сумме 500 рублей, на 10 – выигрыш в 100 рублей, на 20 – 50 рублей, на 50 – 10 рублей. Определить закон распределения вероятностей случайной величины X – выигрыша на один билет.

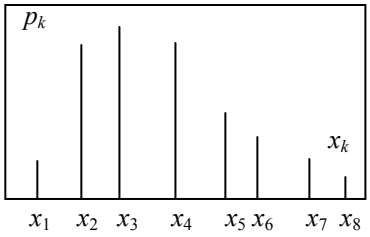

Рис. 2.1. **Графическое задание дискретной случайной величины**

Величина X может принять одно из пяти возможных значений: 0, 10, 50, 100 и 500. Заметим, что число билетов без выигрыша равно 1000 – 5 – 10 – 20 – – 50 = 915. Следовательно, $P\{X = 0\} = 915/1000 = 0{,}915$. Аналогично находим все другие вероятности: $P\{X = 10\} = 0{,}05$, $P\{X = 50\} = 0{,}02$, $P\{X = 100\} = 0{,}01$, $P\{X = 500\} = 0{,}005$. Найденный закон представим таблично:

x_k	0	10	50	100	500
p_k	0,915	0,05	0,02	0,01	0,005

Для непрерывных величин эти способы задания являются неприемлемыми, так как невозможно перечислить все возможные значения непрерывной случайной величины.

2.2. Функция распределения вероятностей и её свойства

Функцией распределения вероятностей $F(x)$ случайной величины X в точке x называется вероятность того, что в результате опыта случайная величина примет значение, меньшее, чем x, т.е. $F(x) = P\{X < x\}$. Рассмотрим свойства функции $F(x)$.

1. $F(-\infty) = \lim\limits_{x\to-\infty} F(x) = 0$. Действительно, по определению, $F(-\infty) = P\{X < -\infty\}$. Событие $(X < -\infty)$ является невозможным событием: $F(-\infty) = P\{X < -\infty\} = P\{V\} = 0$.
2. $F(\infty) = \lim\limits_{x\to\infty} F(x) = 1$, так как по определению, $F(\infty) = P\{X < \infty\}$. Событие $X < \infty$ является достоверным событием. Следовательно, $F(\infty) = P\{X < \infty\} = P\{U\} = 1$.
3. Вероятность того, что случайная величина примет значение из интервала $[\alpha, \beta)$, равна приращению функции распределения вероятностей на этом интервале.
$$P\{\alpha \le X < \beta\} = F(\beta) - F(\alpha). \qquad (2.2)$$
4. $F(x_2) \ge F(x_1)$, если $x_2 > x_1$, т.е. функция распределения вероятностей является неубывающей функцией.
5. Функция распределения вероятностей непрерывна слева.

Докажем 3-е свойство. Обозначим событие $X < \alpha$ через A, событие $X < \beta$ – через B и событие $\alpha \le X < \beta$ – через C. Очевидно, что $B = A + C$ (рис. 2.2), причём события A и C являются несовместными. Следовательно, $P(B) = P(A) + P(C)$ или $P(C) = P(B) - P(A)$.

Четвёртое свойство вытекает из третьего. Положим, в (2.2) $\alpha = x_1$, а $\beta = x_2$. Получим: $P\{x_1 \le X < x_2\} = F(x_2) - F(x_1)$. Но $P\{x_1 \le X < x_2\} \ge 0$ всегда, значит $F(x_2) \ge F(x_1)$.

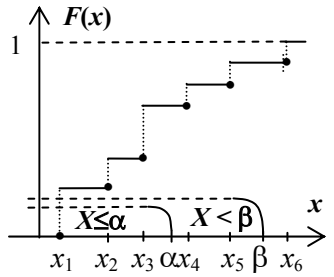

Рис. 2.2. **Функции распределения дискретной случайной величины**

Различия между функциями распределения вероятностей дискретной и непрерывной случайных величин хорошо иллюстрировать графиками. Пусть, например, дискретная случайная величина имеет n возможных значений, вероятности которых равны

$P\{X = x_k\} = p_k$, $k = 1, 2, \ldots, n$. Если $x \le x_1$, то $F(x) = 0$, так как левее x нет возможных значений случайной величины. Если $x_1 < x \le x_2$, то левее x находится всего одно возможное значение, а именно, значение x_1. Значит, $F(x) = P\{X = x_1\} = p_1$. При $x_2 < x \le x_3$ слева от x находится уже два возможных значения, поэтому $F(x) = P\{X = x_1\} + P\{X = x_2\} = p_1 + p_2$. Рассуждая аналогично, приходим к выводу, что если $x_k < x \le x_{k+1}$, то $F(x) = p_1 + p_2 + \ldots + p_k$, $1 \le k < n$. Наконец, если $x > x_n$, то $F(x) = 1$, так как функция будет равна сумме вероятностей всех возможных значений, которая, по условию нормировки, равна единице. Таким образом, график функции распределения дискретной случайной величины является ступенчатым (рис. 2.2). Возможные значения непрерывной величины располагаются плотно на интервале задания этой величины, что обеспечивает плавное возрастание функции распределения $F(x)$, т.е. её непрерывность (рис. 2.3).

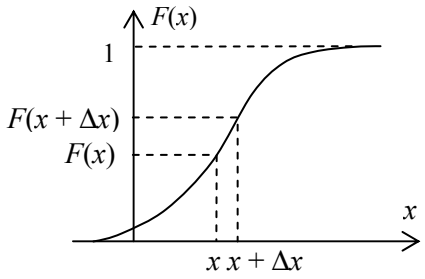

Рис. 2.3. **Функции распределения непрерывной случайной величины**

Рассмотрим вероятность попадания случайной величины в интервал $[x, x + \Delta x)$, $\Delta x > 0$:

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

Перейдём к пределу при $\Delta x \to 0$:

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

Предел равен вероятности того, что случайная величина примет значение, равное x. Если функция $F(x)$ непрерывна в точке x, то $\lim\limits_{\Delta x \to 0} F(x + \Delta x) = F(x)$, т.е. $P\{X = x\} = 0$. Если $F(x)$ имеет разрыв в точке x, то вероятность $P\{X = x\}$ будет равна скачку функции в этой точке.

Таким образом, вероятность появления любого возможного значения для непрерывной величины равна нулю. Выражение $P\{X = x\} = 0$ следует понимать как предел вероятности попадания случайной величины в бесконечно малую окрестность точки x при неограниченном уменьшении этой окрестности. Вероятности $P\{\alpha < X < \beta\}$, $P\{\alpha \le X < \beta\}$, $P\{\alpha < X \le \beta\}$ и $P\{\alpha \le X \le \beta\}$ равны, если X – непрерывная случайная величина. Для дискретных величин эти вероятности не одинаковы в том случае, когда границы интервала α и (или) β совпадают с возможными значениями случайной величины. Для дискретной случайной величины необходимо строго учитывать тип неравенства в формуле (2.2).

Пример 2.2. Задана функция распределения вероятностей случайной величины X:

$$F(x) = \begin{cases} 0, & \text{если } x \le 0, \\ \alpha \cdot \sin x, & \text{если } 0 < x \le \pi/2, \\ 1, & \text{если } x > \pi/2. \end{cases}$$

Определить постоянную α и вероятность того, что случайная величина примет значение, лежащее в интервале $(\pi/6, \pi/3)$. Построить график функции распределения вероятностей.

Так как $F(x) = 0$ при $x \le 0$ и $F(x) = 1$ при $x > \pi/2$, то возможных значений нет левее точки $x = 0$ и правее точки $x = \pi/2$. Функция распределения вероятностей непрерывной случайной величины является непрерывной функцией, поэтому $F(\pi/2) = 1$. Или $\alpha \cdot \sin(\pi/2) = 1$, отсюда получаем $\alpha = 1$.

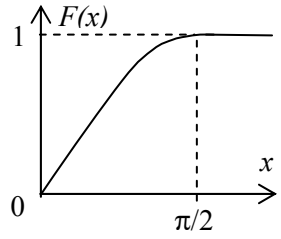

Рис. 2.4. **Иллюстрация к примеру 2.2**

Чтобы вычислить искомую вероятность, воспользуемся свойством 3 функции распределения вероятностей:

$$P\{\frac{\pi}{6} < x < \frac{\pi}{3}\} = F(\frac{\pi}{3}) - F(\frac{\pi}{6}) = \sin(\frac{\pi}{3}) - \sin(\frac{\pi}{6}) = \frac{\sqrt{3}}{2} - \frac{1}{2} \approx 0{,}366\,.$$

График функции распределения представлен на рис. 2.4.

2.3. Плотность распределения вероятностей и её свойства

Из формулы (2.2) следует, что вероятность попадания случайной величины в заданный интервал определяется скоростью изменения функции распределения вероятностей на этом интервале. Скорость изменения непрерывной функции равна её производной. Это позволяет ввести новую функцию для задания случайной величины.

Рассмотрим снова вероятность попадания случайной величины в интервал $[x, x + \Delta x)$:

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

Пусть X – непрерывная случайная величина. Тогда для малых значений Δx эта вероятность будет также достаточно малой. Поделим её на Δx и перейдём к пределу при $\Delta x \to 0$:

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

Если этот предел существует, то он равен производной от функции распределения $F(x)$:

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x) = f(x)\,. \qquad (2.3)$$

Функция $f(x)$ называется *плотностью распределения вероятностей* случайной величины X. Из определения следует, что при малых значениях Δx справедливо равенство:

$$P\{x \le X < x + \Delta x\} \approx f(x) \cdot \Delta x. \qquad (2.4)$$

Рассмотрим свойства плотности распределения $f(x)$.

1. Всегда $f(x) \ge 0$, так как функция $F(x)$ является неубывающей функцией.

2. Для функции распределения $F(x)$ справедливо равенство:

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

Действительно, так как по определению $f(x) = F'(x)$, то $F(x)$ является первообразной функцией по отношению к плотности распределения $f(x)$. Следовательно,

$$\int_{-\infty}^{x} f(t)dt = F(t)\Big|_{-\infty}^{x} = F(x) - F(-\infty) = F(x) - 0 = F(x)\,.$$

3. Вероятность попадания случайной величины в заданный интервал $[\alpha, \beta)$ равна:

$$P\{\alpha \le X < \beta\} = \int_{\alpha}^{\beta} f(t)dt\,. \qquad (2.6)$$

Действительно, в соответствии с формулой Ньютона – Лейбница этот определённый интеграл равен $F(\beta) - F(\alpha)$. По свойству 3 функции распределения вероятностей эта разность и представляет собой вероятность $P\{\alpha \le X < \beta\}$.

4. Интеграл от плотности распределения вероятности по всей области задания случайной величины равен единице:

$$\int_{-\infty}^{\infty} f(t)dt = 1. \qquad (2.7)$$

Равенство (2.7) представляет условие нормировки вероятностей для непрерывных случайных величин. По смыслу интеграл в (2.7) есть не что иное, как $F(\infty)$, а по свойству 2 функции распределения вероятностей $F(\infty) = 1$. Если областью задания случайной величины является отрезок $[a; b]$, то $\int_a^b f(t)dt = 1$. Условие нормировки вероятностей часто используется для определения неизвестного параметра закона распределения.

Для иллюстрации геометрического смысла перечисленных свойств приведём пример графика плотности распределения вероятностей. Для большей наглядности на рис. 2.5 представлен также график соответствующей функции распределения вероятностей.

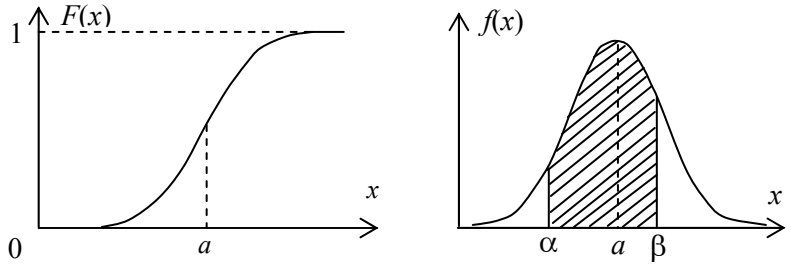

Рис. 2.5. **Функция и плотность распределения вероятностей непрерывной случайной величины**

Вся кривая плотности распределения вероятностей располагается выше оси OX (свойство 1), причём максимум плотности достигается в точке $x = a$, в которой функция распределения вероятностей имеет наибольшую крутизну. Вероятность попадания случайной величины в интервал $[\alpha; \beta)$ численно равна площади криволинейной трапеции, построенной на этом интервале как на основании и ограниченной сверху графиком плотности распределения (заштрихованная область на рис 2.5). Площадь всей криволинейной трапеции, заключённой между осью OX и графиком плотности распределения, всегда равна единице. Любая функция, удовлетворяющая перечисленным выше свойствам, может быть плотностью распределения некоторой непрерывной случайной величины.

Пример 2.3. Задана плотность распределения вероятностей случайной величины X:

$$f(x) = \begin{cases} 0, & \text{если } x \le 0 \\ m \cdot e^{-2x}, & \text{если } x > 0. \end{cases}$$

Определить m. Найти вероятность попадания случайной величины в интервал $[1; 2)$.

Для определения неизвестного параметра m воспользуемся условием нормировки:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad \text{или} \qquad \int_0^{\infty} m \cdot e^{-3x}dx = -\frac{m}{3} \cdot e^{-3x}\Big|_0^{\infty} = -\frac{m}{3}(0-1) = \frac{m}{3} = 1.$$

При интегрировании было учтено, что $f(x) = 0$, если $x \in (-\infty; 0)$. Из полученного равенства находим, что $m = 3$ и, следовательно, $f(x) = 3 \cdot e^{-3x}$, если $x > 0$, и $f(x) = 0$, если $x \le 0$.

Вероятность попадания случайной величины в интервал [1; 2) можно вычислить, опираясь на свойство 3 плотности распределения:

$$P\{1 \le X < 2\} = \int_1^2 f(x)dx = 3\int_1^2 e^{-3x}dx = 3 \cdot \frac{1}{-3} \cdot e^{-3x}\Big|_1^2 = -(e^{-6} - e^{-3}) \approx 0{,}047$$

Пример 2.4. Задана плотность распределения случайной величины X: $f(x) = 4x \cdot e^{-2x}$, если $x > 0$ и $f(x) = 0$, если $x \le 0$. Определить функцию распределения вероятностей $F(x)$.

Если $x \le 0$, то $F(x) = 0$, так как $f(x) = 0$ при $x \le 0$. Пусть теперь $x \in (0; \infty)$. Тогда

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_0^x f(t)dt = \int_{-\infty}^{0} 0 \cdot dt + 4\int_0^x t \cdot e^{-2t}dt\,.$$

Проводя интегрирование по частям, получим: $F(x) = 1 - (1 + 2x) \cdot e^{-2x}$. Заметим, что $F(x) \to 1$ при $x \to \infty$. Полная запись ответа: $F(x) = \begin{cases} 0, & \text{если } x \le 0, \\ 1 - (1 + 2x) \cdot e^{-2x}, & \text{если } x > 0. \end{cases}$

2.4. Функция случайной величины. Математическое ожидание

Пусть X – дискретная случайная величина, имеющая n возможных значений $x_1, x_2, \ldots, x_n$, вероятности которых равны соответственно $p_1, p_2, \ldots, p_n$. Рассмотрим некоторую функцию $y = \varphi(x)$. Возможным значениям $x_1, x_2, \ldots, x_n$ случайной величины X будут соответствовать значения $y_1, y_2, \ldots, y_n$, причём вероятностями этих значений будут также $p_1, p_2, \ldots, p_n$. Значения $y_1, y_2, \ldots, y_n$ – это возможные значения случайной величины Y, являющейся функцией случайной величины X: $Y = \varphi(X)$.

Сделаем N измерений случайной величины X, вычислим соответствующие значения функции $\varphi(X)$ и найдём её среднее арифметическое значение на основе полученных данных. Пусть $N > n$. Тогда в последовательности измерений величины X каждое из её возможных значений встретится многократно. Предположим, что значение x_1 повторяется N_1 раз, x_2 – N_2 раза, …, x_n – N_n раз, причём $N_1 + N_2 + \ldots + N_n = N$. Тогда среднее арифметическое значение m_φ^* функции $\varphi(X)$ определяется следующей суммой:

$$m_\varphi^* = \sum_{i=1}^{n} \varphi(x_i) \cdot \frac{N_i}{N} = \sum_{i=1}^{n} \varphi(x_i) \cdot p_i^*\,.$$

Будем увеличивать число измерений случайной величины X. При $N \to \infty$ частота $p_i^* = N_i / N$ появлений значения x_i среди измерений будет в некотором смысле приближаться к истинной вероятности p_i этого значения. При $N \to \infty$ величина m_φ^* будет приближаться к величине m_φ, определяемой следующей формулой: $m_\varphi = \sum_{i=1}^{n} \varphi(x_i) \cdot p_i$. Эта величина называется *математическим ожиданием функции* $Y = \varphi(X)$.

Пусть теперь X является непрерывной случайной величиной с плотностью распределения вероятностей $f(x)$. Будем считать, что эта величина определена на отрезке $[a; b]$, т. е. $f(x) \ne 0$ при $x \in [a; b]$ и $f(x) = 0$ при $x \notin [a; b]$. Разделим отрезок $[a; b]$ на n элементарных частей точками $x_0 = a, x_1, x_2, \ldots, x_n = b$ так, что $x_0 < x_1 < x_2 < \ldots < x_n$. В каждом частичном интервале выберем произвольно точку z_k: $x_{k-1} < z_k < x_k$. Вероятность p_k попадания случайной величины в k-й частичный интервал будет равна $p_k = \int_{x_{k-1}}^{x_k} f(x)dx \approx f(z_k) \cdot \Delta x_k$, где

$\Delta x_k = x_k - x_{k-1}$, причём точность приближения последнего равенства будет тем выше, чем меньше Δx_k. Будем рассматривать новую дискретную случайную величину Z, полагая, что эта величина принимает значение z_k, если только X попадает в k-й частичный интервал. Тогда Z будет иметь n возможных значений $z_1, z_2, \ldots, z_n$ с соответствующими вероятностями $p_1, p_2, \ldots, p_n$. Следовательно, можно вычислить математическое ожидание функции $\varphi(Z)$:

$$m_\varphi = \sum_{k=1}^{n} \varphi(z_k) \cdot p_k \approx \sum_{k=1}^{n} \varphi(z_k) \cdot f(z_k) \cdot \Delta x_k .$$

Получилась интегральная сумма, которая стремится к определённому интегралу при стремлении к нулю всех Δx_k (при этом $n \to \infty$, а z_k стремится к текущему значению x):

$$m_\varphi = \int_a^b \varphi(x) f(x) dx .$$

Формула будет более общей, если интегрировать в пределах от $-\infty$ до ∞, так как интегрирование на участках числовой оси, где $f(x) = 0$, не изменяет значения интеграла.

Математическим ожиданием функции $\varphi(X)$ случайной величины X называется величина m_φ, которая имеет символическое обозначение $M\{\varphi(X)\}$ и определяется по формулам:

$$m_\varphi = M\{\varphi(X)\} = \begin{cases} \sum_{i=1}^{n} \varphi(x_i) p_i, & \text{для дискретных сл. вел.,} \\ \int_{-\infty}^{\infty} \varphi(x) f(x) dx, & \text{для непрерывных сл. вел.,} \end{cases} \qquad (2.8)$$

где x_i ($i = 1, 2, \ldots, n$) – возможные значения дискретной случайной величины; p_i – вероятности этих значений; n – общее число возможных значений; $f(x)$ – плотность распределения вероятностей непрерывной случайной величины.

Пример 2.5. Случайная величина X задана таблично:

x_i	0	1	2	3	4	5
p_i	0,05	0,1	0,2	0,3	0,2	0,15

Вычислить математическое ожидание функции $Y = X^2$.

Применим формулу (2.8) для дискретной случайной величины ($n = 6$):

$$M\{X^2\} = \sum_{i=1}^{6} x_i^2 \cdot p_i = 0^2 \cdot 0{,}05 + 0^2 \cdot 0{,}05 + 1^2 \cdot 0{,}1 + 2^2 \cdot 0{,}2 + 3^2 \cdot 0{,}3 + 4^2 \cdot 0{,}2 + 5^2 \cdot 0{,}15 = 8{,}65$$

Пример 2.6. Задана плотность распределения вероятностей случайной величины X: $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0$, $x \geq 0$, для $x < 0$ плотность $f(x) = 0$. Найти математическое ожидание функции $\varphi(X) = X^k$, где k – целое число ($k \geq 0$). Опираясь на полученный результат, вычислить $M\{X^2\} - [M\{X\}]^2$.

Воспользуемся формулой (2.8) для непрерывной случайной величины:

$$M\{\varphi(X)\} = \int_{-\infty}^{\infty} \varphi(x) f(x) dx = \int_0^{\infty} x^k \cdot \lambda e^{-\lambda x} dx = \left| t = \lambda x,\ dt = \lambda dx \right| = \frac{1}{\lambda^k} \cdot \int_0^{\infty} t^k e^{-t} dt = \frac{k!}{\lambda^k}.$$

Последний интеграл вычислен интегрированием по частям k раз. Таким образом,

$$M\{X^2\} - [M\{X\}]^2 = \frac{2!}{\lambda^2} - \left[\frac{1!}{\lambda}\right]^2 = \frac{1}{\lambda^2}.$$

Рассмотренный закон распределения вероятностей случайной величины называется показательным (или экспоненциальным) законом. Полученный здесь результат

$$M\{X^k\} = \frac{k!}{\lambda^k} \tag{2.9}$$

будет использован в дальнейшем изложении материала.

Из формул (2.8) следует, что математическое ожидание функции не всегда существует. Если интеграл в (2.8) оказывается несобственным интегралом, то он может сходиться или не сходиться. Аналогично, в дискретном случае, когда множество возможных значений случайной величины является неограниченным ($n = \infty$), математическое ожидание функции будет представлять собой числовой ряд, который также может сходиться или расходиться. Математическое ожидание функции существует, если соответствующий числовой ряд (для дискретной величины) или соответствующий интеграл (для непрерывной величины) сходятся абсолютно. Если нет абсолютной сходимости, то говорят, что математическое ожидание функции не существует.

2.5. Числовые характеристики случайных величин

Числовые характеристики случайных величин можно условно разделить на основные и вспомогательные. К основным характеристикам относятся характеристики положения случайной величины и характеристики рассеяния. Характеристики положения указывают некоторую точку на числовой оси, вокруг которой группируются возможные значения случайной величины. К ним относятся математическое ожидание, мода и медиана случайной величины. Характеристики рассеяния являются некоторой мерой разброса возможных значений случайной величины около своего центра рассеяния, например, математического ожидания. Характеристиками рассеяния являются дисперсия и вероятное отклонение случайной величины. К этим характеристикам можно отнести и среднее квадратическое отклонение случайной величины, хотя эта характеристика однозначно определяется дисперсией.

Дополнительные характеристики применяются для дальнейшего уточнения свойств случайной величины. К ним, прежде всего, относятся асимметрия (или скошенность) и эксцесс (или островершинность) закона распределения случайной величины. К дополнительным характеристикам относится и коэффициент вариации случайной величины, который характеризует относительный разброс возможных значений случайной величины.

Математическое ожидание случайной величины является основной характеристикой положения случайной величины на числовой оси. Если в формуле для математического ожидания функции от случайной величины положить $\varphi(X) = X$, то получим математическое ожидание m_x самой случайной величины:

$$m_x = M\{X\} = \begin{cases} \sum_{i=1}^{n} x_i p_i, & \text{для дискретной сл. величины,} \\ \int_{-\infty}^{\infty} x f(x) dx, & \text{для непрерывной сл. величины.} \end{cases} \tag{2.10}$$

Математическое ожидание m_x указывает точку на числовой оси, вокруг которой группируются возможные значения случайной величины. В определённом смысле эта точка является центром распределения вероятностей случайной величины. Заметим, что математическое ожидание случайной величины может не существовать.

Отметим простейшие свойства математического ожидания.

1. Математическое ожидание постоянной величины C равно этой постоянной величине, т.е. $M\{C\} = C$. Постоянную величину можно считать частным случаем дискретной величины с одним возможным значением, вероятность которого равна единице.

1. Постоянный множитель можно выносить за знак математического ожидания, т.е. $M\{C \cdot X\} = C \cdot M\{X\}$, где C – постоянный множитель. Это следует из того, что в (2.10) постоянный множитель можно выносить как за знак суммы, так и за знак интеграла.

2. Математическое ожидание суммы случайных величин равно сумме математических ожиданий этих случайных величин, т.е. $M\{X + Y\} = M\{X\} + M\{Y\}$.

Это свойство обобщается на любое конечное число случайных величин. Далее свойства математического ожидания будут рассмотрены более подробно.

Пример 2.7. Найти математическое ожидание дискретной случайной величины, заданной таблично:

x_k	0,5	1,5	2	3	3,5	4
p_k	0,1	0,2	0,3	0,2	0,1	0,1

По формуле (2.10) для дискретных случайных величин получаем: $m_x = 0{,}5 \cdot 0{,}1 + 1{,}5 \cdot 0{,}2 + 2 \cdot 0{,}3 + 3 \cdot 0{,}2 + 3{,}5 \cdot 0{,}1 + 4 \cdot 0{,}1 = 2{,}3$. Полученное значение математического ожидания $m_x = 2{,}3$ не совпадает ни с одним из возможных значений случайной величины. Сама величина может принимать только свои возможные значения, но при многократных измерениях величины среднее арифметическое измеренных значений будет неограниченно приближаться к значению 2,3 при неограниченном увеличении числа измерений.

Пример 2.8. Задана плотность распределения вероятностей случайной величины X:

$$f(x) = \begin{cases} \dfrac{x}{2}, & \text{если } 0 \le x \le 2, \\ 0, & \text{если } x < 0 \quad \text{или} \quad x > 2. \end{cases}$$

Требуется найти математическое ожидание этой случайной величины.

Воспользуемся формулой (2.10) для непрерывной случайной величины:

$$m_x = M\{X\} = \int_{-\infty}^{\infty} x f(x)dx = \int_0^2 x \cdot \frac{x}{2} dx = \left. \frac{x^3}{2 \cdot 3} \right|_0^2 = \frac{8}{2 \cdot 3} = \frac{4}{3} \approx 1{,}33 .$$

Если математическое ожидание случайной величины не существует, то в качестве характеристики положения случайной величины применяют моду или медиану.

Модой случайной величины X называется такое значение x, при котором плотность распределения вероятностей случайной величины $f(x)$ принимает максимальное значение. Из определения следует, что мода определяется только для непрерывных случайных величин. Аналогом моды для дискретной случайной величины является её наивероятнейшее значение. Мода обозначается через Mod_x.

Для нахождения моды нужно исследовать плотность распределения вероятностей на максимум. Для этого нужно найти стационарные точки, как корни уравнения $f'(x) = 0$, затем применить к найденным точкам один из достаточных признаков максимума. Если максимумов вообще нет, то говорят, что моды не существует. Если максимум один, то закон распределения называют одномодальным, если максимумов несколько, то – многомодальным. На рис. 2.6. представлены примеры графиков одномодального и двухмодального законов распределения.

Пример 2.9. Задана плотность распределения вероятностей случайной величины X:

$$f(x) = \begin{cases} \lambda^2 x \cdot e^{-\lambda x}, & \lambda > 0, \quad \textit{если } x \geq 0, \\ 0, & \textit{если } x < 0. \end{cases}$$

Требуется найти моду этой случайной величины.

Найдём производную от плотности распределения: $f'(x) = \lambda^2(e^{-\lambda x} - \lambda x e^{-\lambda x}) = \lambda^2 e^{-\lambda x}(1 - \lambda x)$.

Уравнение $f'(x) = 0$ имеет один корень: $1 - \lambda x = 0$ или $x = 1/\lambda$. Таким образом, $\mathrm{Mod}_x = 1/\lambda$.

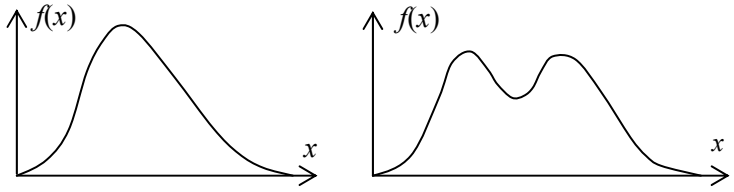

Рис. 2.6. **Примеры одномодального и двухмодального законов распределения вероятностей**

Медианой случайной величины X называется такое значение x, которое разбивает всю область возможных значений случайной величины на две равновероятные части, т. е.

$$P\{X < x\} = P\{X \geq x\} = 0{,}5.$$

Из определения следует, что медиана, как и мода, точно может быть определена только для непрерывных случайных величин. Медиана обычно обозначается через Med_x. По определению, медиана находится как решение уравнения $F(x) = 0{,}5$. Очевидно, что для дискретных величин это уравнение либо не имеет решения, либо имеет бесчисленное множество решений в силу ступенчатого характера функции распределения вероятностей. В некоторых задачах медиана используется и для дискретных величин. За медиану принимается некоторое значение x, достаточно близкое к решению указанного уравнения.

В случае симметричных распределений медиана совпадает с математическим ожиданием случайной величины. Если, к тому же, распределение является одномодальным, то совпадают все три характеристики – математическое ожидание, мода и медиана.

Пример 2.10. Задана функция распределения вероятностей случайной величины X: $F(x) = 1 - \exp\{-\dfrac{x^2}{2\sigma^2}\}$, если $x > 0$ и $F(x) = 0$, если $x \leq 0$ (*распределение Релея*).

Найти медиану случайной величины.

Медиану находим как решение уравнения $F(x) = 0{,}5$:

$$F(x) = 1 - \exp\{-\frac{x^2}{2\sigma^2}\} = 0{,}5 \qquad \text{или} \qquad \exp\{-\frac{x^2}{2\sigma^2}\} = 0{,}5.$$

Прологарифмируем последнее равенство: $-\dfrac{x^2}{2\sigma^2} = \ln\dfrac{1}{2} = -\ln 2$ или $x = \pm\sqrt{2\ln 2}\cdot\sigma$.

Отрицательной медиана быть не может, так как величина X определена на положительной полуоси. Следовательно, $\mathrm{Med}_x = \sqrt{2\ln 2}\cdot\sigma$. Для сравнения приводим здесь другие характеристики положения для данного распределения: $Mod_x = \sigma, \quad m_x = \sigma\sqrt{\pi/2}$.

Дисперсия D_x, или $D\{X\}$, случайной величины X является основной мерой рассеяния возможных значений случайной величины вокруг математического ожидания. Дис-

персия определяется как математическое ожидание квадрата отклонения случайной величины от своего математического ожидания. В соответствии с этим определением,

$$D_x = D\{X\} = M\{(X - m_x)^2\}. \quad (2.11)$$

Применяя формулу, определяющую математическое ожидание произвольной функции от случайной величины, получим:

$$D_x = M\{(X - m_x)^2\} = \begin{cases} \sum_{i=1}^{n} (x_i - m_x)^2 p_i, \\ \int_{-\infty}^{\infty} (x - m_x)^2 f(x)dx. \end{cases} \quad (2.12)$$

В (2.12) первая строка соответствует дискретной случайной величине, вторая – непрерывной. Для существования дисперсии необходима сходимость числового ряда (если n = ∞) в (2.12) для дискретной величины, или сходимость интеграла в (2.12) для непрерывной величины.

Дисперсия обладает следующими свойствами:

1. Дисперсия постоянной величины равна нулю, т.е. $D\{C\} = 0$, если C = const. Постоянная величина не имеет рассеяния, так как обладает единственным возможным значением.
2. Постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат, т.е. $D\{C \cdot X\} = C^2 \cdot D\{X\}$, где C – постоянный множитель. Действительно,
$$D[CX\} = M\{(CX - M\{CX\})^2\} = M\{(CX - C \cdot m_x)^2\} =$$
$$= M\{C^2 \cdot (X - m_x)^2\} = C^2 \cdot M\{(X - m_x)^2\} = C^2 \cdot D\{X\}.$$
3. Для любой постоянной величины C справедливо следующее соотношение:
$$D\{X\} = M\{(X - C)^2\} - (C - m_x)^2.$$

Это равенство легко доказывается непосредственным применением свойств математического ожидания. Из этого следует, что для любого C выполняется неравенство:

$$D\{X\} \le M\{(X - C)^2\},$$

т. е. рассеяние случайной величины около произвольной точки C числовой оси всегда не меньше, чем дисперсия. Равенство достигается тогда, когда $C = m_x$. Это свойство часто используется для упрощения вычислений при нахождении дисперсии.

Дисперсия не всегда является удобной мерой рассеяния случайной величины. Если, например, случайная величина X измеряется в метрах, то дисперсия D_x будет измеряться в квадратных метрах. Это обстоятельство вносит определённые неудобства. Поэтому часто применяют *среднее квадратическое отклонение* σ_x:

$$\sigma_x = \sqrt{D_x} \text{ или } D_x = \sigma_x^2 . \quad (2.13)$$

Иногда вместо термина «среднее квадратическое отклонение» применяют термин «*стандартное отклонение*».

Дисперсию можно вычислить иным способом. Раскрывая (2.11), получим: $D_x = M\{(X - m_x)^2\} = M\{X^2 - 2Xm_x + m_x^2\} = M\{X^2\} - 2m_x M\{X\} + m_x^2$. Поскольку $M\{X\} = m_x$, то

$$D_x = M\{X^2\} - m_x^2. \quad (2.14)$$

Пример 2.11. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины X, заданной таблично:

x_k	–1	0	1	2	2,5	3
p_k	0,1	0,1	0,2	0,3	0,2	0,1

Сначала найдём математическое ожидание: m_x = –1 · 0,1 + 0 · 0,1 + 1 · 0,2 + 2·0,3 + + 2,5 · 0,2 + 3·0,1 = 1,5. Дисперсию вычислим непосредственно по формуле (2.12):

$$D_x = \sum_{k=1}^{6} (x_k - m_x)^2 p_k = (-1-1{,}5)^2 \cdot 0{,}1 + (0-1{,}5)^2 \cdot 0{,}1 + (1-1{,}5)^2 \cdot 0{,}2 + (2-1{,}5)^2 \cdot 0{,}3 +$$

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

Таким образом, математическое ожидание m_x = 1,5, дисперсия D_x = 1,4, следовательно, $\sigma_x = \sqrt{D_x} = \sqrt{1{,}4} = 1{,}183$.

Пример 2.12. Задана плотность распределения вероятностей случайной величины *X*:

$$f(x) = \begin{cases} \frac{1}{2}\sin(x), & \textit{если} \quad 0 \le x \le \pi, \\ 0, & \textit{если} \quad x < 0 \quad \textit{или} \quad x > \pi. \end{cases}$$

Определить дисперсию этой случайной величины.

Воспользуемся формулой (2.14). Вычислим математическое ожидание:

$$m_x = \int_{-\infty}^{\infty} x \cdot f(x)dx = \frac{1}{2}\int_0^{\pi} x \cdot \sin(x)dx = \left| \begin{matrix} U = x & dU = dx \\ dV = \sin(x)dx & V = -\cos(x) \end{matrix} \right| = \frac{1}{2}[-x \cdot \cos(x)\Big|_0^{\pi} +$$

$$= \int_0^{\pi} \cos(x)dx] = \frac{\pi}{2} + \frac{1}{2}\sin(x)\Big|_0^{\pi} = \frac{\pi}{2}.$$

Найдём теперь $M\{X^2\}$: $M\{X^2\} = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{2}\int_0^{\pi} x^2 \sin(x)dx$. Интегрируя дважды по частям, получим: $M\{X^2\} = \frac{1}{2}\left[-x^2\cos(x) + 2x \cdot \sin(x) + 2\cos(x)\right] = \frac{1}{2}\left(\pi^2 - 4\right)$. Таким образом,

$$D_x = M\{X^2\} - m_x^2 = \frac{1}{2}(\pi^2 - 4) - \left(\frac{\pi}{2}\right)^2 = \frac{\pi^2}{4} - 2\,.$$

В качестве меры относительного отклонения случайной величины от своего математического ожидания применяется коэффициент вариации. *Коэффициентом вариации* K_v случайной величины называется отношение среднего квадратического отклонения этой величины к её математическому ожиданию, т.е.

$$K_v = \frac{\sigma_x}{m_x}. \qquad (2.15)$$

Коэффициент вариации широко применяется в теории надёжности. Пусть, например, случайная величина T представляет собой время безотказной работы некоторой технической системы. Тогда, чем меньше коэффициент вариации этой величины, тем выше качество данной системы. Таким образом, для повышения качества системы нужно повышать среднее время безотказной работы и понижать среднее квадратическое отклонение времени безотказной работы от своего среднего значения. Заметим, что не для всех законов распределения коэффициент вариации можно изменять как угодно. Например, для экспоненциального закона распределения коэффициент вариации всегда равен единице, а для некоторых распределений он может изменяться только ступенчатым образом.

2.6. Квантили, квартили и вероятное отклонение

Квантилью порядка p случайной величины X называют такое число x_p, для которого выполняется следующее равенство:

$$P\{X < x_p\} = p \text{ или } F(x_p) = p. \qquad (2.16)$$

Квантили удобны для сравнения различных законов распределения вероятностей. В некоторых случаях пользуются *децилями*: $x_{0,1}$, $x_{0,2}$, $x_{0,3}$, …, $x_{0,9}$. Однако наибольшее распространение получили квартили. *Квартилями* называют квантили порядков 0,25, 0,5 и 0,75. Будем их обозначать соответственно как k_1, k_2, k_3. Квартили k_1 и k_3 называют обычно нижней и верхней квартилями. Вторая квартиль k_2 совпадает, очевидно, с рассмотренной ранее медианой распределения. На рис. 2.7 показаны квартили на числовой оси.

С помощью квантилей удобно сравнивать распределения по величине рассеяния возможных значений случайных величин. В частности, если дисперсия случайной величины не существует, то в качестве меры рассеяния можно применить величину $E = 0{,}5(k_3 - k_1)$, которую называют *вероятным отклонением* или *срединным отклонением*. Для симметричных распределений вероятное отклонение можно определить из следующего уравнения: $P\{|X - m_x| \le E\} = 0{,}5$.

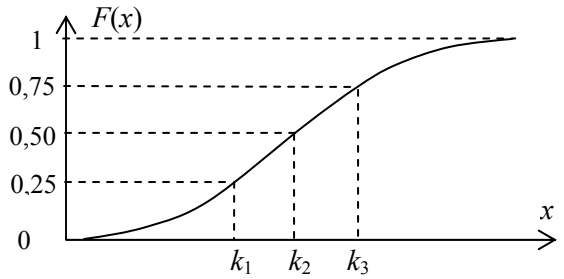

Рис. 2.7. **Иллюстрация к понятию квартилей**

Чем больше известно квантилей, тем более полное представление можно составить о характере закона распределения случайной величины. Однако, описание случайных величин с помощью квантилей применимо лишь к непрерывным случайным величинам. Для дискретных случайных величин функция распределения вероятностей имеет ступенчатый характер. Поэтому уравнение (2.16) будет иметь решения только для некоторых значений p.

Пример 2.13. Дана функция распределения вероятностей случайной величины X (распределение Релея):

$$F(x) = 1 - \exp\{-\frac{x^2}{2\sigma^2}\},\ x \ge 0.$$

Определить квартили и вероятное отклонение этой случайной величины.

Квантили любого порядка p определяются уравнением (2.16), т.е. $F(x) = 1 - \exp\{-\frac{x^2}{2\sigma^2}\} = p$ или $\exp\{-\frac{x^2}{2\sigma^2}\} = 1 - p$. Разрешая это уравнение относительно x, получим квантиль порядка p: $x_p = \sqrt{2} \cdot \sigma\sqrt{-\ln(1-p)}$. Второй корень уравнения не попадает в область определения случайной величины. Отметим также, что подкоренное выражение в полученной формуле не может быть отрицательным, так как $1 - p \le 1$ и, следовательно, всегда $\ln(1 - p) \le 0$. Подставляя в полученную формулу p = 0, 25, 0, 5 и 0, 75 и проводя преобразования, получим три квартили: $k_1 = \sigma \cdot \sqrt{2(\ln 4 - \ln 3)} \approx 0{,}76\sigma$, $k_2 = \sigma \cdot \sqrt{2\ln 2} \approx 1{,}18\sigma$, $k_3 = 2\sigma \cdot \sqrt{\ln 2} \approx 1{,}67\sigma$. Следовательно, $E = 0{,}5(k_3 - k_1) = 0{,}455\sigma$.

2.7. Моменты случайных величин. Асимметрия и эксцесс

Иногда знание математического ожидания и дисперсии оказывается недостаточным для характеристики случайной величины. В этом случае привлекают моменты случайной величины, частным случаем которых являются математическое ожидание и дисперсия.

Начальным моментом ν_k порядка k случайной величины X называется математическое ожидание величины X^k:

$$\nu_k = M\{X^k\} = \begin{cases} \sum_{i=1}^{n} x_i^k p_i, & \text{если } X \text{ дискретная сл. вел.,} \\ \int_{-\infty}^{\infty} x^k f(x)dx, & \text{если } X \text{ непрерывная сл. вел.} \end{cases} \tag{2.17}$$

Здесь $k = 0, 1, 2, 3, \ldots$. Очевидно, что всегда $\nu_0 = 1$, а $\nu_1 = m_x$.

Центральным моментом μ_k порядка k случайной величины X называется математическое ожидание величины $(X - \nu_1)^k$:

$$\mu_k = M\{(X-\nu_1)^k\} = \begin{cases} \sum_{i=1}^{n} (x_i - \nu_1)^k p_i, \\ \int_{-\infty}^{\infty} (x-\nu_1)^k f(x)dx, \end{cases} \tag{2.18}$$

Здесь первая строка соответствует дискретной случайной величине, вторая – непрерывной, $k = 0, 1, 2, 3, \ldots$. Очевидно, что всегда $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = D_x$. Таким образом, дисперсия случайной величины – это её центральный момент второго порядка.

Применяя простейшие преобразования и пользуясь свойствами математического ожидания, получим формулы связи между начальными и центральными моментами:

$$\mu_2 = \nu_2 - \nu_1^2; \quad \mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3; \quad \mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4. \tag{2.19}$$

Рассмотрим, например, последовательность получения формулы для μ_3:

$$\mu_3 = M\{(X-\nu_1)^3\} = M\{X^3 - 3\cdot X^2\nu_1 + 3\cdot X\cdot\nu_1^2 - \nu_1^3\} = M\{X^3\} - \\ -3\nu_1\cdot M\{X^2\} + 3\nu_1^2\cdot M\{X\} - \nu_1^3.$$

Учитывая, что $M\{X\} = \nu_1$, приходим к формуле для μ_3 в (2.18). Центральные моменты рекомендуется вычислять через начальные моменты, пользуясь формулами (2.19).

Центральный момент третьего порядка μ_3 характеризует асимметрию закона распределения вероятностей случайной величины. За меру асимметрии принимают величину

$$Sk_x = \frac{\mu_3}{\sigma_x^3}, \tag{2.20}$$

которую называют *асимметрией* (или *скошенностью)* распределения случайной величины. Очевидно, что асимметрия является безразмерной характеристикой. Если $Sk_x = 0$, то закон распределения вероятностей является симметричным. На рис. 2.8 приведены графики плотностей распределения вероятностей с отрицательной, нулевой и положительной асимметрией.

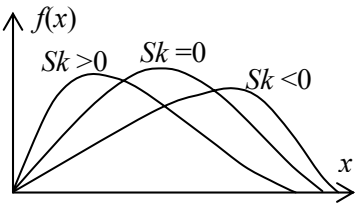

Рис. 2.8. **Законы распределения с разной величиной асимметрии**

Центральный момент четвёртого порядка μ_4 характеризует островершинность закона распределения вероятностей случайной величины. За меру островершинности закона распределения принимается величина Ex_x, определяемая формулой

$$Ex_x = \frac{\mu_4}{\sigma_x^4} - 3. \qquad (2.21)$$

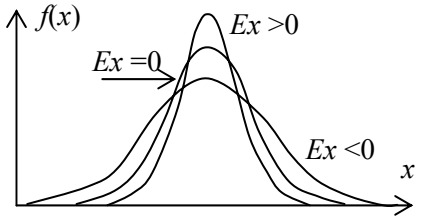

Рис. 2.9. **Законы распределения с разной величиной эксцесса**

Эта величина называется *эксцессом* закона распределения. На рис. 2.9 приведены графики плотностей распределения вероятностей с разными эксцессами. В (2.21) деление на σ_x^4 делается для того, чтобы эксцесс был безразмерной характеристикой. Число 3 вычитается с той целью, чтобы эксцесс нормального распределения, используемого в качестве эталонного, был равен нулю. Плотность нормального распределения вероятностей имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \cdot \exp\{-\frac{(x - m_x)^2}{2\sigma_x^2}\}, \quad -\infty < x < \infty. \qquad (2.22)$$

Математическое ожидание и дисперсия соответственно равны константам m_x и σ_x^2, а отношение μ_4/σ_x^4 равно 3. Следовательно, $Ex_x = 0$. Таким образом, если $Ex_x < 0$, то вершина распределения менее острая, чем у нормального распределения, если $Ex_x > 0$, то, напротив, вершина более острая.

Пример 2.14. Определить скошенность и эксцесс для экспоненциального закона распределения: $f(x) = \lambda \cdot e^{-\lambda x}, \quad \lambda > 0, \quad x > 0.$

Согласно (2.9), $\nu_k = M\{X^k\} = \frac{k!}{\lambda^k}, \quad k = 1, 2, 3, \ldots$. Найдём центральные моменты:

$$\mu_2 = \nu_2 - \nu_1^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}; \quad \mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 = \frac{6}{\lambda^3} - 3 \cdot \frac{2}{\lambda^2} \cdot \frac{1}{\lambda} + 2 \cdot \left(\frac{1}{\lambda}\right)^3 = \frac{2}{\lambda^3};$$

$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 = \frac{24}{\lambda^4} - 4 \cdot \frac{6}{\lambda^3} \cdot \frac{1}{\lambda} + 6 \cdot \frac{2}{\lambda^2} \cdot \left(\frac{1}{\lambda}\right)^2 - 3 \cdot \left(\frac{1}{\lambda}\right)^4 = \frac{9}{\lambda^4}.$$

Таким образом, для экспоненциального закона $m_x = 1/\lambda$, $\sigma_x^2 = 1/\lambda^2$. Следовательно,

$$Sk_x = \frac{\mu_3}{\sigma_x^3} = \frac{2}{\lambda^3} \cdot \lambda^3 = 2, \qquad Ex_x = \frac{\mu_4}{\sigma_x^4} - 3 = \frac{9}{\lambda^4} \cdot \lambda^4 - 3 = 6.$$

Заметим, что скошенность и эксцесс не зависят в данном случае от параметра λ.

Пример 2.15. Требуется найти все основные характеристики случайной величины X (математическое ожидание, дисперсию, скошенность и эксцесс), заданной таблично:

x_k	1	2	3	4	5	6	7
p_k	0,1	0,2	0,35	0,15	0,1	0,05	0,05

Перемножая возможные значения случайной величины на соответствующие вероятности и складывая все произведения, получим математическое ожидание: $m_x = 3{,}3$. Воспользуемся формулой (2.18). Все данные для этого представлены в следующей таблице:

x_k	1	2	3	4	5	6	7	Σ
p_k	0,1	0,2	0,35	0,15	0,1	0,05	0,05	1
$x_k - m_x$	–2,3	–1,3	–0,3	0,7	1,7	2,7	3,7	– –
$(x_k - m_x)p_k$	–0,23	–0,26	–0,105	0,105	0,17	0,135	0,185	0
$(x_k - m_x)^2 p_k$	0,529	0,338	0,032	0,074	0,289	0,365	0,685	2,312
$(x_k - m_x)^3 p_k$	–1,217	–0,439	–0,01	0,051	0,491	0,984	2,533	2,393
$(x_k - m_x)^4 p_k$	2,798	0,571	0,003	0,036	0,835	2,657	9,371	16,271

В последнем столбце таблицы представлены центральные моменты: $\mu_1 = 0$, $\mu_2 = 2{,}312$, $\mu_3 = 2{,}393$, $\mu_4 = 16{,}271$. Таким образом, дисперсия $D_x = 2{,}312$, среднее квадратическое отклонение $\sigma_x = 1{,}521$. Следовательно,

$$Sk_x = \frac{\mu_3}{\sigma_x^3} = \frac{2.393}{1{,}521^3} \approx 0{,}68, \quad Ex_x = \frac{\mu_4}{\sigma_x^4} - 3 = \frac{16.271}{1{,}521^4} - 3 \approx 0{,}044.$$

2.8. Производящие функции

Все моменты случайной величины можно "запомнить" с помощью производящей функции. *Производящей функцией начальных моментов* случайной величины X называют математическое ожидание функции $\exp\{uX\}$, где u – некоторая вспомогательная переменная:

$$\varphi_\nu(u) = M\{\exp\{uX\}\} = \begin{cases} \sum\limits_{k=0}^{\infty} \exp\{u \cdot x_k\} p_k, \\ \int\limits_{-\infty}^{\infty} \exp\{ux\} f(x)dx. \end{cases} \tag{2.23}$$

Первая строка соответствует дискретной случайной величине, вторая – непрерывной.

Чтобы понять смысл производящей функции, рассмотрим разложение экспоненты $\exp\{uX\}$ в ряд Маклорена:

$$\exp\{uX\} = 1 + uX + \frac{(uX)^2}{2!} + \frac{(uX)^3}{3!} + \ldots + \frac{(uX)^n}{n!} + \ldots .$$

Отсюда получаем: $\varphi_\nu(u) = M\{\exp\{uX\}\} = 1 + \nu_1 u + \frac{\nu_2}{2!}u^2 + \frac{\nu_3}{3!}u^3 + \ldots + \frac{\nu_n}{n!}u^n + \ldots$,

т.е. коэффициентами разложения функции $\varphi_\nu(u)$ в ряд Маклорена являются начальные моменты, поделённые на соответствующие факториалы. Эта формула также непосредственно вытекает из самого определения производящей функции (2.23).

Производящей функцией центральных моментов случайной величины X называют математическое ожидание функции $\exp\{u(X - m_x)\}$:

$$\varphi_\mu(u) = M\{\exp\{u(X - m_x)\}\} = \begin{cases} \sum\limits_{k=0}^{\infty} \exp\{u \cdot (x_k - m_x)\} p_k, \\ \int\limits_{-\infty}^{\infty} \exp\{u(x - m_x)\} f(x)dx. \end{cases} \tag{2.24}$$

Коэффициентами разложения этой функции в ряд Маклорена являются центральные моменты, поделённые на соответствующие факториалы:

$$\varphi_\mu(u) = M\{\exp\{u(X - m_x)\}\} = 1 + \mu_1 u + \frac{\mu_2}{2!}u^2 + \frac{\mu_3}{3!}u^3 + \ldots + \frac{\mu_n}{n!}u^n + \ldots .$$

Таким образом, зная производящие функции, легко определить любой начальный и любой центральный момент по формулам

$$\nu_k = \left.\frac{d^k \varphi_\nu(u)}{du^k}\right|_{u=0}, \qquad \mu_k = \left.\frac{d^k \varphi_\mu(u)}{du^k}\right|_{u=0}. \tag{2.25}$$

Вообще в математике производящей функцией последовательности $a_0, a_1, \ldots, a_n, \ldots$ называют функцию

$$\varphi(u) = a_0 + a_1 u + a_2 u^2 + \ldots + a_n u^n + \ldots,$$

если только эта функция существует. Функция $\varphi_\nu(u)$ является производящей функцией последовательности $\nu_n/n!$ (n = 0, 1, 2, …), а функция $\varphi_\mu(u)$ – производящей функцией последовательности $\mu_n/n!$ (n = 0, 1, 2, …). Возьмём в качестве последовательности вероятности $p_0, p_1, \ldots, p_n\ldots$ возможных значений $x_0, x_1, \ldots, x_n, \ldots$ дискретной случайной величины X. Тогда получим функцию

$$\varphi_p(u) = p_0 + p_1 u + p_2 u^2 + \ldots + p_n u^n + \ldots, \tag{2.26}$$

которая называется *производящей функцией вероятностей* случайной величины X. Очевидно, что $\varphi_p(1) = 1$ – условие нормировки вероятностей, а $\varphi_p(0) = p_0$. Кроме того, зная $\varphi_p(u)$ можно определить все вероятности p_k (k = 0, 1, 2, …), т.е. полностью восстановить закон распределения вероятностей случайной величины:

$$p_k = \left.\frac{1}{k!}\frac{d^k \varphi_p(u)}{du^k}\right|_{u=0}, \quad k = 1,2,3,\ldots. \tag{2.27}$$

Пример 2.16. Требуется найти производящую функцию начальных моментов случайной величины X, распределённой по закону Бернулли:

$$P\{X = k\} = p_k = C_n^k p^k q^{n-k},\ k = 0, 1, 2, \ldots, 0 \le p \le 1,\ q = 1 - p.$$

Используя эту производящую функцию, определить математическое ожидание и дисперсию случайной величины X.

Применим формулу (2.23) для дискретных величин:

$$\varphi_\nu(u) = \sum_{k=0}^{n} e^{uk} p_k = \sum_{k=0}^{n} e^{uk} C_n^k p^k q^{n-k} = \sum_{k=0}^{n} C_n^k (pe^u)^k q^{n-k}.$$

Пользуясь формулой бинома Ньютона $(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}$, получим:

$$\varphi_\nu(u) = (pe^u + q)^n = (pe^u + 1 - p)^n = (1 + p(e^u - 1))^n.$$

По формуле (2.25) определим первые два начальных момента:

$$\nu_1 = \left.\frac{\partial \varphi_\nu(u)}{\partial u}\right|_{u=0} = n(1 + p(e^u - 1))^{n-1} pe^u\Big|_{u=0} = np,\ \nu_2 = \left.\frac{\partial^2 \varphi_\nu(u)}{\partial u^2}\right|_{u=0} =$$

$$= \left[n(n-1)(1 + p(e^u - 1))^{n-2} p^2 e^{2u} + n(1 + p(e^u - 1))^{n-1} pe^u\right]\Big|_{u=0} = n(n-1)p^2 + np.$$

Математическое ожидание m_x совпадает с ν_1, т.е. $m_x = np$. Дисперсию D_x вычислим через начальные моменты:

$$D_x = \nu_2 - \nu_1^2 = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq.$$

Пример 2.17. Задана производящая функция вероятностей случайной величины X, принимающей значения k = 1, 2, 3, …:

$$\varphi_p(u) = \frac{pu}{1-qu}, \quad q = 1-p.$$

Требуется восстановить закон распределения вероятностей случайной величины.

Воспользуемся формулой (2.27):

$$P\{X=k\} = p_k = \frac{1}{k!}\cdot\frac{\partial^k \varphi_p(u)}{\partial u^k}\bigg|_{u=0} = \frac{1}{k!}\cdot\frac{k!\,pq^{k-1}}{(1-qu)^k}\bigg|_{u=0} = pq^{k-1},$$

где k = 1, 2, 3, … . Этот закон распределения вероятностей называется геометрическим законом, или распределением Фарри.

2.9. Примеры дискретных законов распределения

Некоторые законы распределения вероятностей достаточно часто встречаются в практических задачах. Рассмотрим несколько таких законов.

Равномерное распределение. Говорят, что случайная величина X распределена по равномерному закону, если она может принимать только целые неотрицательные значения от 1 до n, а вероятности всех возможных значений одинаковы. Таким образом,

$$P\{X=k\} = p_k = \frac{1}{n},\ k = 1,2,...,n. \tag{2.28}$$

Основные числовые характеристики:

математическое ожидание $m_x = \frac{n+1}{2}$; *дисперсия* $D_x = \frac{n^2-1}{12}$; *скошенность* $Sk_x = 0$;

эксцесс $Ex_x = -1.2 + \frac{4}{n^2-1}$.

Биномиальное распределение (распределение Бернулли). Говорят, что случайная величина X распределена по биномиальному закону (или по закону Бернулли), если она может принимать только целые неотрицательные значения от 0 до n, а вероятности этих значений определяются формулой:

$$P\{X=k\} = P_n(k) = C_n^k p^k q^{n-k},\ k = 0, 1, 2,...n;\ 0 \le p \le 1,\ q = 1-p. \tag{2.29}$$

По смыслу случайная величина X является числом появления некоторого события в n независимых испытаниях, если вероятность появления этого события в каждом испытании постоянна p.

Основные числовые характеристики:

математическое ожидание $m_x = np$; *дисперсия* $D_x = npq$; *наивероятнейшее значение* s,

$np-q \le s \le np+p$; *скошенность* $Sk_x = \frac{1-2p}{\sqrt{npq}}$; *эксцесс* $Ex_x = \frac{1-6pq}{npq}$.

Таким образом, скошенность распределения может быть и положительной и отрицательной. Распределение становится симметричным при p=0,5. Величина эксцесса всегда отрицательна и приближается к нулю при возрастании n. Отметим, что при $n \to \infty$ биномиальный закон распределения вероятностей неограниченно приближается к нормальному закону.

Биномиальный закон распределения вероятностей применяется в теории стрельбы, в теории и практике статистического контроля качества продукции, в теории массового обслуживания, в теории надежности и т.д. Этот закон может применяться во всех случаях, когда имеет место последовательность независимых испытаний.

Распределение Пуассона. Если число независимых испытаний n устремить к ∞, а вероятность p устремить к нулю так, чтобы np стремилось к некоторой постоянной λ, то биномиальное распределение будет неограниченно приближаться к распределению Пуассона. В общем случае говорят, что случайная величина X распределена по закону Пуассона, если она может принимать только целые неотрицательные значения, а вероятности этих значений определяются формулой:

$$P\{X = k\} = P_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots; \ \lambda \geq 0. \tag{2.30}$$

Основные числовые характеристики:

математическое ожидание $m_x = \lambda$; *дисперсия* $D_x = \lambda$; *наиверoятнейшее значение* s , $\lambda - 1 \leq s \leq \lambda$;

скошенность $Sk_x = \frac{1}{\sqrt{\lambda}}$; *эксцесс* $Ex_x = \frac{1}{\lambda}$.

Геометрическое распределение (Фарри). Этот закон распределения вероятностей также связан с последовательностью независимых испытаний. Пусть случайной величиной является число проведённых испытаний до первого осуществления наблюдаемого события. Например, число выстрелов до первого попадания в цель, число проверенных изделий до первого появления бракованного изделия и т.п. Очевидно, что возможными значениями такой случайной величины будут целые числа 1, 2, 3,…, причём эти значения не ограничены (теоретически) сверху. Если p есть вероятность данного события в одном испытании, а q=1-p, то

$$P\{X = k\} = p_k = pq^{k-1}, \quad k = 1, 2, 3, \ldots \tag{2.31}$$

Эта формула и представляет геометрический закон распределения вероятностей. Заметим, что при любом значении p, не равном нулю или единице, наиверoятнейшим значением является единица. С ростом k вероятности монотонно убывают.

Основные числовые характеристики:

математическое ожидание $m_x = 1/p$; *дисперсия* $D_x = q/p^2$.

2.10. Примеры непрерывных распределений

Равномерное распределение. Говорят, что случайная величина X распределена равномерно в интервале [a, b], если её плотность распределения вероятностей определяется формулой:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{если } a \leq x \leq b, \\ 0, & \text{если } x < a \text{ или } x > b. \end{cases} \tag{2.32}$$

Соответствующая функция распределения вероятностей имеет следующий вид:

$$F(x) = \begin{cases} 0, & \text{если } x < a, \\ \frac{x-a}{b-a}, & \text{если } a \leq x \leq b, \\ 1, & \text{если } x > b. \end{cases} \tag{2.33}$$

Основные числовые характеристики:

$$\text{математическое ожидание } m_x = \frac{a+b}{2};$$

$$\text{дисперсия } D_x = \frac{(b-a)^2}{12};$$

$$\text{скошенность } Sk_x = 0;$$

$$\text{эксцесс } Ex_x = -1.2;$$

$$\textit{коэффициент вариации } K_V = \frac{1}{\sqrt{3}} \cdot \frac{b-a}{a+b}.$$

Графики плотности и функции распределения вероятностей приведены на рис. 2.10.

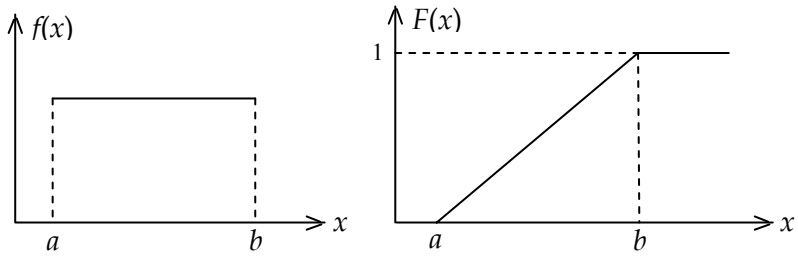

Рис. 2.10. **Графики функции и плотности равномерного распределения вероятностей.**

Равномерно распределены ошибки округления при выполнении числовых расчётов, время ожидания автобуса, (трамвая и т.п.) при условии, что интервал движения постоянен и строго выполняется, а пассажир прибывает на остановку в случайный момент времени. Равномерно распределено и время ожидания начала обслуживания заявки, поступающей в случайный момент в систему массового обслуживания, если обслуживающее устройство включается через равные промежутки времени. При этом предполагается, что все заявки, поступившие к моменту включения обслуживающего устройства, немедленно принимаются к обслуживанию.

Равномерно распределённые величины применяются для получения (генерации) реализаций случайных величин, имеющих другие законы распределения вероятностей. Равномерное распределение является самым простым из непрерывных законов распределения. Поэтому путём определённых операций достаточно легко получить последовательность чисел, которые можно приближённо считать реализациями случайной величины, распределённой равномерно в заданном интервале. Такие числа, а также их функциональные преобразования, называют *псевдослучайными числами*. Существует много алгоритмов получения равномерно распределённых псевдослучайных чисел. Все они работают так, что через некоторый период начинают повторяться те числа, которые уже встречались. В этом и заключается смысл псевдослучайности полученных чисел. С позиций теории вероятностей повторение одной и той же реализации непрерывной случайной величины является событием с нулевой вероятностью. Если же само это событие повторяется многократно, то это очевидный признак того, что имеет место отклонение от равномерного закона распределения. Заметим, что генератор псевдослучайных чисел тем лучше, чем больший период повторения он имеет. В то же время следует заметить, что при практическом использовании полученных чисел они всегда округляются. Следовательно, даже при самом лучшем генераторе псевдослучайных чисел отклонение от заданного закона распределения вероятностей неизбежно.

В программном обеспечении современных ЭВМ имеется программа генерации равномерно распределённых псевдослучайных чисел с параметрами a=0 и b=1. Можно приближенно считать, что числа, полученные с помощью такой программы, являются реализациями случайной величины X, распределённой равномерно в интервале (0; 1) и имеющей математическое ожидание m_x=0,5 и дисперсию D_x=1/12.

Экспоненциальное (показательное) распределение. Говорят, что случайная величина X распределена по экспоненциальному закону, если её плотность распределения вероятностей определяется формулой:

$$f(x) = \lambda \cdot e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0. \tag{2.34}$$

Соответствующая функция распределения вероятностей имеет следующий вид:

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0. \tag{2.35}$$

Графики плотности и функции распределения вероятностей приведены на рис. 2.11.

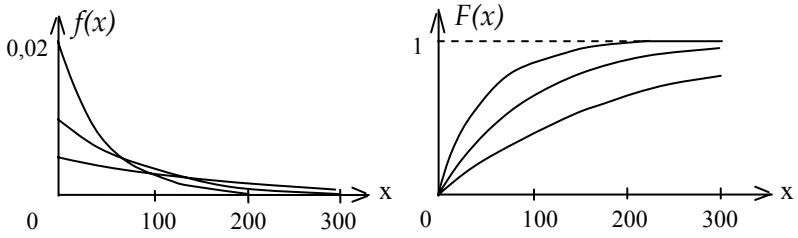

Рис. 2.11. **Функция и плотность распределения вероятностей случайной величины, распределённой по экспоненциальному закону**

Основные числовые характеристики:

математическое ожидание $m_x = \frac{1}{\lambda}$; *дисперсия* $D_x = \frac{1}{\lambda^2}$; *скошенность* $Sk_x = 2$; *эксцесс* $Ex_x = 6$; *коэффициент вариации* $K_V = 1$.

Экспоненциальное распределение зависит от единственного параметра λ. Этим фактом определяются некоторые характерные свойства данного распределения.
В частности, среднее квадратическое отклонение σ_x всегда совпадает с математическим ожиданием, т.е. всегда коэффициент вариации равен единице.

Последовательность реализаций случайной величины, распределённой по экспоненциальному закону, легко получить экспериментально. Пусть T есть случайная величина, распределённая равномерно в интервале (0; 1), а t – одно из значений этой величины. Тогда число x, определяемое формулой

$$x = -\frac{1}{\lambda} \cdot \ln(1 - t), \tag{2.36}$$

будет одним из значений случайной величины X, распределённой по экспоненциальному закону с параметром λ. Если t получать с помощью генератора случайных чисел, то x будет реализацией псевдослучайной величины X, распределённой по этому закону. Заметим, что эта формула вытекает непосредственно из формулы (2.35), если предполагать, что сама функция $F(x)$ является случайной величиной равномерно распределённой в интервале (0; 1).

Экспоненциальное (или, иначе, показательное) распределение широко применяется в теории надёжности и в теории массового обслуживания. Оказывается, что случайное время безотказной работы многих технических систем обладает именно экспоненциальным распределением. Кроме того, время восстановления (ремонта) отказавшей технической системы часто также оказывается распределённым по этому закону. В связи с этими фактами экспоненциальное распределение получило относительно широкое применение.

Экспоненциальное распределение имеет определённую связь с распределением Пуассона. Предположим, что мы наблюдаем некоторую последовательность событий, развивающуюся с течением времени. Если интервалы времени между моментами осуще-

ствления последовательных событий распределены по экспоненциальному закону с параметром λ, то число событий, осуществившихся на интервале времени (0; *t*), будет распределено по закону Пуассона с параметром λ*t*. В связи с этим закон распределения Пуассона часто называют законом редких событий.

Нормальное распределение. Говорят, что случайная величина *X* распределена по нормальному закону, если её плотность распределения вероятностей определяется следующей формулой:

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}},\quad -\infty<x<0. \tag{2.37}$$

Соответствующая функция распределения вероятностей имеет следующий вид:

$$F(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}}dx. \tag{2.38}$$

Графики плотности и функции распределения вероятностей приведены на рис. 2.12. На рисунке представлены графики двух нормальных законов распределения вероятностей с разными числовыми характеристиками.

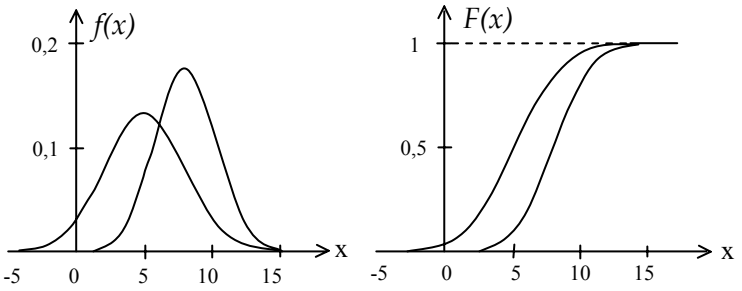

Рис. 2.12. **Графики плотности и функции распределения вероятностей нормально распределённой случайной величины.**

Основные числовые характеристики:
математическое ожидание $m_x = m$; *дисперсия* $D_x = \sigma^2$; *скошенность* $Sk_x = 0$;
эксцесс $Ex_x = 0$; *коэффициент вариации* $K_V = \sigma/m$.

Отметим, что нормальное распределение симметрично относительно своего математического ожидания. В связи с этим, мода, медиана и математическое ожидание нормально распределенной случайной величины равны между собой.

Нормальное распределение удобно использовать в качестве эталонного распределения при сравнении случайных величин. Однако для сравнения используется не произвольная нормально распределённая случайная величина, а величина с нулевым математическим ожиданием и единичной дисперсией: m_x=0, D_x=1. Обозначим плотность и функцию распределения вероятностей такой величины соответственно через $f_0(x)$ и $F_0(x)$. Тогда

$$F_0(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt,\quad f_0(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}},\quad -\infty<x<\infty. \tag{2.39}$$

Делая замену переменной интегрирования в (2.38), легко получить связь между функциями $F(x)$ и $F_0(x)$:

$$F(x)=F_0\left(\frac{x-m}{\sigma}\right). \tag{2.40}$$

Таким образом, вероятность попадания нормально распределённой случайной величины в интервал [α;β) определяется формулой:

$$P\{\alpha \le X < \beta\} = F_0(\frac{\beta - m}{\sigma}) - F_0(\frac{\alpha - m}{\sigma}). \quad (2.41)$$

Отметим, что преобразование случайной величины по формуле $T=(X-m)/\sigma$ называют операцией центрирования (вычитание математического ожидания) и нормирования (деление на среднее квадратическое отклонение) случайной величины. Полученную таким образом случайную величину T называют центрированной (с нулевым математическим ожиданием) и нормированной (с единичной дисперсией). Очевидно, что центрирование и нормирование можно проводить раздельно.

При решении задач, связанных с нормальным распределением, часто вместо функции распределения вероятностей $F_0(x)$ используется функция Лапласа $\Phi_0(x)$:

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt.$$

Очевидно, что $F_0(x)=F_0(0)+\Phi_0(x)$. Но $F_0(0)=0{,}5$. Поэтому

$$F_0(x)=0.5+\Phi_0(x). \quad (2.42)$$

Таким образом, вероятность попадания нормально распределённой случайной величины в интервал [α;β) можно находить и по следующей формуле:

$$P\{\alpha \le X < \beta\} = \Phi_0(\frac{\beta - m}{\sigma}) - \Phi_0(\frac{\alpha - m}{\sigma}). \quad (2.43)$$

Выбор формул (2.41) или (2.43) зависит только от того, какие таблицы имеются в распоряжении. В справочной литературе (а также в учебниках, в сборниках задач и т.п.) можно найти таблицы значений как функции распределения вероятностей $F_0(x)$, так и функции Лапласа $\Phi_0(x)$. Заметим, что функция $\Phi_0(x)$ является нечётной функцией, т.е. $\Phi_0(-x) = -\Phi_0(x)$, а $f_0(x)$ является чётной функцией, т.е. $f_0(-x) = f_0(x)$. Поэтому обычно приводятся таблицы значений этих функций только для положительных значений аргументов.

Нормально распределённые случайные величины встречаются очень часто в практических задачах. Это связано с тем, что нормально распределённая случайная величина является следствием аддитивного (суммарного) взаимодействия большого числа других величин (или факторов), каждая из которых вносит малую долю в общий результат. Вообще при определённых условиях закон распределения суммы случайных величин при увеличении числа слагаемых неограниченно приближается к нормальному закону. Об этом говорит центральная предельная теорема, которая формулируется в теме 5. Частным случаем этой теоремы являются локальная и интегральная теоремы Муавра-Лапласа, рассмотренные в первой теме. Искусственное получение реализаций нормально распределённой случайной величины также опирается на центральную предельную теорему. Оказывается, что уже двенадцать одинаково равномерно распределённых случайных величин в сумме представляют величину, закон распределения которой достаточно близок к нормальному закону. Если t_i, i=1, 2,…,12, есть реализации случайной величины, распределённой равномерно в интервале (0; 1), то величина

$$x = \sum_{i=1}^{12} (t_i - 0{,}5) \quad (2.44)$$

будет реализацией центрированной и нормированной случайной величины, распределённой приближенно по нормальному закону.

Гамма-распределение и *распределение Эрланга.* Более сложным законом распределения вероятностей непрерывных случайных величин является гамма-распределение. Плотность распределения вероятностей величины *T*, распределённой по этому закону имеет вид:

$$f(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp\{-\lambda t\},\ t \ge 0,,\quad \mathrm{t} \ge 0, \tag{2.45}$$

где $\lambda>0$, $\alpha>0$, а $\Gamma(\alpha)$ - гамма-функция Эйлера:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp\{-t\} dt \tag{2.46}$$

При целых значениях α это распределение совпадает с *распределением Эрланга*. Если $\alpha=n$ – целое, то $\Gamma(n)= (n-1)!$. Тогда плотность распределения запишется следующим образом:

$$f(t) = \frac{\lambda \cdot (\lambda t)^{n-1}}{(n-1)!} \exp\{-\lambda t\} \tag{2.47}$$

Это распределение называют распределением Эрланга *n*-го порядка. В частности, распределение Эрланга 1-го (при $n=1$) совпадает с экспоненциальным законом распределения (2.34).

Для гамма-распределения можно записать функцию распределения вероятностей следующим образом:

$$F(t) = \begin{cases} 0, & \text{если} \quad t \le 0, \\ \dfrac{\Gamma(\alpha, \lambda t)}{\Gamma(\alpha)}, & \text{если} \quad t > 0, \end{cases} \tag{2.48}$$

где $\Gamma(\alpha,\lambda t)$ – неполная гамма-функция:

$$\Gamma(\alpha, \lambda t) = \int_0^{\lambda t} z^{\alpha-1} \exp\{-z\} dz\,. \tag{2.49}$$

В случае распределения Эрланга, когда $\alpha=n$, интеграл (2.49) можно вычислить. Тогда

$$F(t) = \begin{cases} 0, & \text{если} \quad t \le 0, \\ 1 - \exp\{-\lambda t\} \cdot \sum\limits_{k=0}^{n-1} \dfrac{(\lambda t)^k}{k!}, & \text{если} \quad t > 0. \end{cases} \tag{2.50}$$

Гамма-распределение имеет следующие основные числовые характеристики:

математическое ожидание $m_x = \frac{\alpha}{\lambda}$; *дисперсия* $D_x = \frac{\alpha}{\lambda^2}$; *скошенность* $Sk_x = \frac{2}{\sqrt{\alpha}}$;

эксцесс $Ex_x = \frac{6}{\alpha}$; *коэффициент вариации* $K_V = \frac{1}{\sqrt{\alpha}}$.

Чтобы получить основные числовые характеристики распределения Эрланга, достаточно вместо α подставить порядок распределения *n*.

Примеры графиков плотности распределения вероятностей гамма-распределения

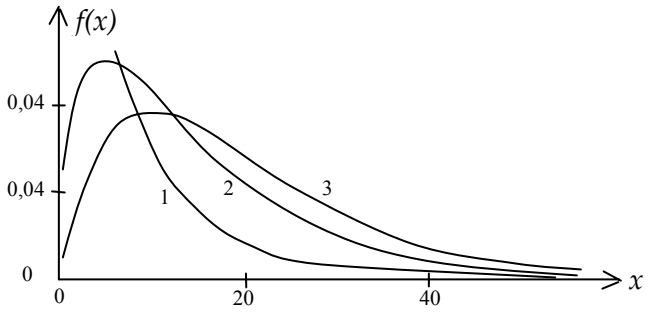

Рис. 2.13. **Графики плотности распределения вероятностей гамма-распределения при разных значениях параметра α.**

представлены на рис. 2.13. Все графики получены при значении λ=0,1. Значения второго параметра равны соответственно: α=0,7 для графика 1, α=1,5 для графика 2 и α=2 для графика 3.

Гамма-распределение можно использовать при описании процессов массового обслуживания или процессов восстановления. Легко убедиться, что интервал времени, распределённый по закону Эрланга порядка n с параметром λ, представляет собой сумму n интервалов, распределённых по экспоненциальному закону с тем же параметром α. Поэтому с помощью распределения Эрланга часто описываются процессы многоэтапного обслуживания. Гамма-распределение является более гибким законом, поскольку для величины, распределённой по этому закону, коэффициент вариации можно изменять плавно, изменяя значения параметров λ и α.

Пример 2.18. Некоторые приборы последовательно проходят профилактическое обслуживание. Время обслуживания одного прибора является случайной величиной, распределённой по экспоненциальному закону с параметром λ=2 час$^{-1}$ (это означает, что среднее время обслуживания одного прибора равно $1/\lambda$=0,5 часа). Какова вероятность того, что время обслуживания четырёх приборов будет не более чем 1,5 часа.

Пусть t_i – случайный интервал времени обслуживания i-го (по очереди) прибора. Тогда время обслуживания четырёх приборов будет равно величине $T=t_1+t_2+t_3+t_4$. Случайная величина T будет распределена по закону Эрланга четвёртого порядка, так как она является суммой четырёх случайных величин, распределённых по одному и тому же экспоненциальному закону. Следовательно,

$$F(t) = 1 - \exp\{-2t\} \cdot \sum_{k=0}^{3} \frac{(2t)^k}{k!}, \quad \textit{если} \quad t > 0 \quad \textit{и} \quad F(t) = 0, \quad \textit{если} \quad t \le 0.$$

Таким образом, $P\{T<1,5\}=F(1,5)=1-\exp\{-3\}[9/2+9/2+3+1]=0,353$.

Распределение Вейбулла. Это распределение широко применяется в задачах надёжности и в теории восстановление. Этому закону распределения вероятностей подчиняется случайный интервал времени безотказной работы некоторых устройств или их элементов. Плотность распределения вероятностей величины T, распределённой по закону Вейбулла, имеет вид:

$$f(t) = \lambda\alpha \cdot t^{\alpha-1} \exp\{-\lambda t^{\alpha}\}, \; t \ge 0, \tag{2.51}$$

где λ>0, α>0. Функция распределения вероятностей записывается следующим образом:

$$F(t) = 1 - \exp\{-\lambda t^{\alpha}\}, \quad t \ge 0. \tag{2.52}$$

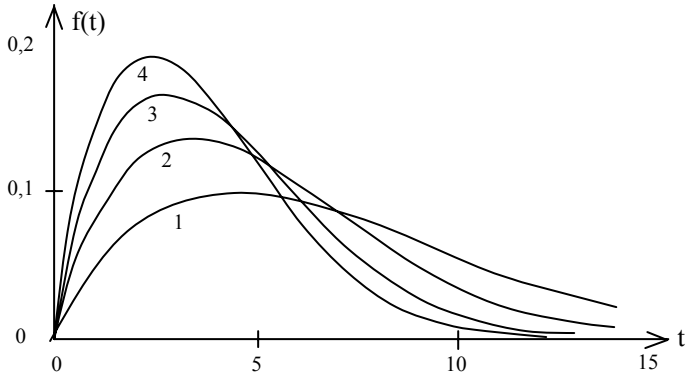

Рис. 2.14. **Графики плотности распределения вероятностей для закона распределения Вейбулла.**

Распределение Вейбулла имеет два параметра, которые могут принимать любые положительные значения. Поэтому оно является достаточно гибким и удобным для аппроксимации законов распределения, полученных опытным путём. При α=1 распределение Вейбулла становится экспоненциальным распределением. Основные числовые характеристики распределения Вейбулла: *математиче-*

ское ожидание $m_x = \lambda^{-\frac{1}{\alpha}}\Gamma(1+\frac{1}{\alpha})$; *дисперсия* $D_x = \lambda^{-\frac{2}{\alpha}}[\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{1}{\alpha})]$. Гамма-функция Эйлера определена в (2.46).

Графики плотности закона распределения Вейбулла представлены на рис. 2.14. Графики построены при значении параметра α=1,7. Кривая 1 соответствует значению λ=0,03, кривая 2 – значению 0,005, кривая 3 – значению 0,07 и кривая 4 – значению 0,09.

Термины для запоминания

- Случайная величина, дискретная случайная величина, непрерывная случайная величина, закон распределения вероятностей, функция распределения вероятностей, условие нормировки вероятностей.
- Математическое ожидание, дисперсия, среднее квадратическое отклонение, коэффициент вариации, мода, медиана, квантили, квартили, вероятное отклонение, начальные моменты случайных величин, центральные моменты, скошенность, асимметрия, эксцесс.

Список контрольных вопросов

1. Дайте определение случайной величины.
2. В чём различие дискретной и непрерывной случайных величин?
3. Что называется законом распределения вероятностей случайной величины?
4. Какими способами можно задать дискретную случайную величину?
5. Дайте определение функции распределения вероятностей случайной величины.
6. Перечислите основные свойства функции распределения вероятностей случайной величины и объясните их смысл.
7. Дайте определение плотности распределения вероятностей случайной величины.
8. Перечислите основные свойства плотности распределения вероятностей случайной величины и объясните их смысл.
9. Чему равен интеграл от плотности распределения вероятностей по всей области задания случайной величины?
10. Перечислите характеристики положения случайной величины на числовой оси.
11. Перечислите характеристики рассеяния случайной величины.
12. Как определяются математическое ожидание и дисперсия дискретной случайной величины?
13. Как определяются математическое ожидание и дисперсия непрерывной случайной величины?
14. Что такое мода и медиана и как они определяются?
15. Дайте определение начальных моментов случайной величины.
16. Дайте определение центральных моментов случайной величины.
17. Что такое асимметрия и эксцесс закона распределения и как они определяются?
18. Что такое квантили и квартили закона распределения и как они определяются?

Вопросы для обсуждений

1. Есть ли в окружающем нас мире «независимые» случайные величины?
2. Попробуйте найти примеры таких величин.
3. Обоснуйте относительность понятия «независимость».

ТЕМА 3.

Многомерные случайные величины

Студент должен
освоить:

- понятия многомерной случайной величины, условные и безусловные законы распределения, статистической зависимости, числовые характеристики случайных величин, линии регрессии;

приобрести навыки:

- решения задач на вычисление характеристик связи между величинами (ковариационный момент, коэффициент корреляции).

Краткое содержание

Определение, классификация, способы задания многомерных случайных величин. Функция распределения вероятностей двухмерной случайной величины и ее свойства. Плотность распределения вероятностей двухмерной случайной величины и ее свойства. Условные законы распределения вероятностей. Статистическая зависимость между величинами. Числовые характеристики многомерных случайных величин. Ковариационный момент и коэффициент корреляции. Условные числовые характеристики. Линии регрессии. Корреляционное отношение. Двухмерное нормальное распределение.

3.1. Определение многомерных случайных величин

Совокупность n случайных величин принято называть n-мерной случайной величиной или системой n случайных величин или n-мерным вектором. В дальнейшем n-мерную случайную величину будем обозначать $(X_1, X_2, \ldots, X_n)$, где X_i, $i = 1, 2, \ldots, n$ – отдельные случайные величины. В частности (X, Y) – двухмерная случайная величина, (X, Y, Z) – трёхмерная случайная величина и т.д. Очевидно, что n-мерную случайную величину можно трактовать как вектор, компоненты которого являются случайными величинами. Двухмерная случайная величина представляет собой вектор на координатной плоскости XOY, а трёхмерная случайная величина – вектор в трёхмерном пространстве.

Для определения типа n-мерной случайной величины достаточно указать типы составляющих её компонент. Здесь рассматриваются только полностью дискретные величины (достаточно кратко) и полностью непрерывные величины (более подробно). Слово «полностью» означает в данном случае тот факт, что все компоненты случайной величины относятся к одному и тому же типу. Почти все вопросы, касающиеся изучения n-мерных случайных величин, можно рассмотреть на примере двухмерной случайной величины.

Пусть в случайной величине (X, Y) возможными значениями компоненты X являются значения $x_1, x_2, \ldots, x_s$, а возможными значениями Y являются $y_1, y_2, \ldots, y_q$. Перечисление всех пар возможных значений (x_i, y_j), $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, q$, с указанием вероятностей этих пар p_{ij} называется законом распределения вероятностей двухмерной случайной величины (X, Y). Для дискретных величин часто применяют табличный способ задания:

x_i \ y_i	y_1	y_2	….	y_q	p_{i*}
x_1	p_{11}	p_{12}	…	p_{1q}	p_{1*}
x_2	p_{21}	p_{22}	…	p_{2q}	p_{2*}
…	…	…	…	…	…
x_s	p_{s1}	p_{s2}	…	p_{sq}	p_{s*}
p_{*j}	p_{*1}	p_{*2}	…	p_{*q}	1

В первом столбце перечислены возможные значения случайной величины X, а в первой строке – возможные значения случайной величины Y. На пересечении i-й строки и j-го столбца указана вероятность того, что случайная величина X примет значение x_i и, одновременно, случайная величина Y примет значение y_j, т.е. $p_{ij} = P\{X = x_i,\ Y = y_j\}$. Суммируя все вероятности p_{ij} по одному из индексов, сохраняя постоянным другой индекс, можно получить закон распределения вероятностей для одномерной величины:

$$P\{X = x_i\} = \sum_{j=1}^{q} p_{ij},\ \ i = 1, 2, ..., s, \qquad P\{Y = y_j\} = \sum_{i=1}^{s} p_{ij},\ \ j = 1, 2, ..., q.$$

Эти вероятности будем обозначать соответственно p_{i*} и p_{*j}. Символ «звёздочка» означает, что по данному индексу проведено суммирование вероятностей двухмерной случайной величины. Вероятности p_{i*} и p_{*j} записаны в последнем столбце и в последней строке таблицы. Очевидно, что обычное условие нормировки вероятностей для p_{i*} и p_{*j} сохраняется, т.е. $\sum_{i=1}^{s} p_{i*} = 1$ и $\sum_{j=1}^{q} p_{*j} = 1$. Из этого следует, что сумма всех вероятностей p_{ij} всегда будет равна единице (*условие нормировки двухмерных вероятностей*):

$$\sum_{i=1}^{s}\sum_{j=1}^{q} p_{ij} = 1. \qquad (3.1)$$

Аналогично можно задавать и случайные величины большей размерности. С ростом размерности процесс задания случайной величины становится громоздким. Так, для трёхмерной случайной величины потребуется трёхмерная таблица, т. е. несколько таблиц представленного выше типа.

Пример 3.1. Двухмерная случайная величина задана табличным способом.

x_i \ y_j	–1	0	1	2	3
3	0,03	0,04	0,03	0,03	0,02
4	0,04	0,06	0,07	0,05	0,03
6	0,05	0,08	0,09	0,08	0,05
8	0,04	0,08	0,06	0,04	0,03

Требуется определить одномерные законы распределения случайных величин X и Y. Вычислить вероятность P того, что случайная величина X примет значение из интервала (3; 7) и, одновременно, Y примет значение из интервала [0; 3).

Для определения вероятностей p_{i*} и p_{*j} нужно просуммировать заданные в таблице вероятности соответственно по строкам и по столбцам. Суммируя по строкам, получим закон распределения вероятностей случайной величины X в табличной форме, суммируя по столбцам, получим закон распределения вероятностей случайной величины Y.

x_i	3	4	5	6
p_{i*}	0,15	0,25	0,35	0,25

y_j	–1	0	1	2	3
p_{*j}	0,16	0,26	0,25	0,20	0,13

В интервал (3; 7) попадает всего два возможных значений величины X: $x = 4$ и $x = 6$. В интервал [0; 3) попадает три возможных значения величины Y: $y = 0$, $y = 1$ и $y = 2$. Таким образом, в указанную в условии задачи область попадает всего шесть точек: (4; 0), (4; 1), (4; 2), (6; 0), (6; 1) и (6; 2). Искомая вероятность равна сумме вероятностей этих точек: P = = 0,06 + 0,07 + 0,05 + 0,08 + 0,09 + 0,08 = 0,43.

3.2. Функция распределения вероятностей двухмерной случайной величины

Универсальным способом задания двумерной случайной величины является задание её функции распределения вероятностей. *Функцией распределения вероятностей двухмерной случайной величины* (X, Y) в точке (x, y) называется вероятность того, что случайная

величина X примет значение меньшее, чем x, а случайная величина Y одновременно примет значение меньшее, чем y, т.е.

$$F(x, y) = P\{X < x, Y < y\}. \quad (3.2)$$

Функция распределения вероятностей $F(x, y)$ равна вероятности попадания точки $(X; Y)$ в область D, заштрихованную часть координатной плоскости на рис. 3.1.

Из определения вытекают свойства двухмерной функции распределения.

1. $F(-\infty, -\infty) = F(-\infty, y) = F(x, -\infty) = 0$. Действительно, события $X < -\infty$ и $Y < -\infty$ представляют невозможные события, следовательно, их вероятность равна нулю. Поэтому одновременное осуществление любого другого события с данными событиями является невозможным. Здесь и далее замена любого аргумента функции $F(x, y)$ символом ∞ (или $-\infty$) означает предельный переход при стремлении к ∞ (или $-\infty$) данного аргумента.

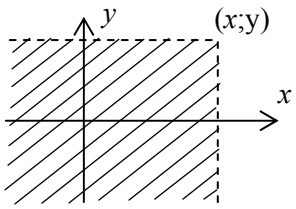

Рис. 3.1. **Область *D***

2. $F(\infty, \infty) = 1$. Действительно, $F(\infty, \infty) = P\{X < \infty, Y < \infty\}$. Но события $X < \infty$ и $Y < \infty$ представляют собой достоверные события. Следовательно, $F(\infty, \infty) = P\{X < \infty, Y < \infty\} = P\{U\} = 1$.

3. $F(x, \infty) = F_X(x)$ и $F(\infty, y) = F_Y(y)$, где $F_X(x)$ и $F_Y(y)$ есть одномерные функции распределения вероятностей соответственно для величин X и Y. Докажем, например, первое равенство. По определению, $F(x, \infty) = P\{X < x, Y < \infty\}$, но событие $Y < \infty$ является достоверным, следовательно, $P\{X < x, Y < \infty\} = P\{X < x\}$. Таким образом, $F(x, \infty) = P\{X < x\}$, а эта вероятность является функцией распределения $F_X(x)$.

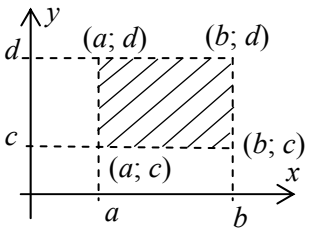

Рис. 3.2. **К определению вероятности попадания точки $(X; Y)$ в прямоугольник**

4. Вероятность того, что случайная величина X примет значение из интервала $[a; b)$, а величина Y примет значение из интервала $[c; d)$, определяется формулой:

$$P\{a \le X < b, c \le Y < d\} = F(b, d) - F(a, d) - F(b, c) + F(a, c). \quad (3.3)$$

Эта вероятность представляет собой вероятность попадания случайной точки $(X; Y)$ в заштрихованный прямоугольник, изображённый на рис. 3.2.

Доказательство равенства (3.3) вытекает из свойств вероятности. При определении вероятности попадания точки в данный прямоугольник возьмём вероятность попадания точки в квадрант с вершиной $(b; d)$. Отнимем от неё вероятности попадания точки в квадранты с вершинами соответственно $(a; d)$ и $(b; c)$. Тогда вероятность попадания точки в квадрант с вершиной $(a; c)$ будет отнята дважды. Поэтому в (3.3) указанная вероятность один раз прибавляется.

5. Функция распределения вероятностей F(x, y) является неубывающей.

Эти свойства справедливы и для многомерных случайных величин более высокой размерности. Необходимо лишь делать соответствующие поправки на размерность. Так, для трёхмерной случайной величины (X, Y, Z) функция распределения в точке (x, y, z) равна $F_3(x, y, z) = P\{X < x, Y < y, Z < z\}$. При этом $F_3(-\infty, -\infty, -\infty) = F_3(-\infty, y, z) = F_3(x, -\infty, z) = F_3(x, y, -\infty) = 0$, а $F_3(\infty, \infty, \infty) = 1$. Далее, $F(x, y, \infty) = F_2(x, y)$, где $F_2(x, y)$ есть функция распределения вероятностей двухмерной случайной величины (X, Y). Аналогично можно получить и остальные двухмерные функции распределения вероятностей. Нижние индексы у функций указывают на их размерность и позволяют различать эти функции. Индекс у двухмерной функции распределения вероятностей будем указывать только в случае необходимости. У одномерных функций в качестве индексов будем всегда указывать имя величины, к которой она относится.

Свойство 3 является правилом перехода к законам распределения вероятностей более низкой размерности. Чтобы из многомерной функции распределения вероятностей получить функцию распределения вероятностей более низкой размерности нужно перейти к пределу при стремлении к бесконечности части аргументов исходной функции. Например, из четырёхмерной функции распределения $F_4(x, y, z, t)$ можно получить двухмерную: $F_2(x, t) = F_4(x, \infty, \infty, t)$. Эту формальную запись следует рассматривать как предельный переход от функции $F_4(x, y, z, t)$ при $y \to \infty$ и $z \to \infty$.

Стремление аргументов к ∞ относится к тому случаю, когда возможные значения (x, y) величины (X, Y) располагаются на всей координатной плоскости. Если область определения случайной величины ограничена, то вместо бесконечности нужно использовать граничные значения переменных x и y, соответствующие этой области. Пусть область возможных значений (x, y) такова, что $\alpha \le x < \beta$ при любом значении y и $\delta \le y < \gamma$ при любом значении x (область не обязательно прямоугольная). Тогда свойства функции распределения можно записать так: $F(\alpha, y) = F(x, \delta) = F(\alpha, \delta) = 0$; $F(\beta, \gamma) = 1$; $F(x, \gamma) = F_X(x)$, $F(\beta, y) = F_Y(y)$.

Графиком двухмерной функции распределения является поверхность. Для дискретных величин эта поверхность имеет ступенчатую форму. Для непрерывных величин график не имеет разрывов в области определения функции распределения.

Пример 3.2. Задана функция распределения вероятностей двухмерной случайной величины (X, Y): $F(x,y) = a \cdot \left(arctg\ x + \frac{\pi}{2} \right)\left(arctg\ y + \frac{\pi}{2} \right), \quad -\infty < x < \infty, \quad -\infty < y < \infty.$

Определить коэффициент a. Найти одномерные функции распределения величин X и Y. Вычислить вероятность попадания точки (X, Y) в квадрат: $-1 \le x \le 1$; $-1 \le y \le 1$.

По второму свойству $F(\infty, \infty) = 1$, следовательно,

$F(\infty,\infty) = a \cdot \left(arctg\ \infty + \frac{\pi}{2} \right)\left(arctg\ \infty + \frac{\pi}{2} \right) = a \cdot \pi^2 = 1.$ Отсюда находим: $a = 1/\pi^2$.

Воспользуемся свойством 3 функции распределения вероятностей:

$$F_X(x) = \lim_{y \to \infty} F(x, y) = \lim_{y \to \infty} \frac{1}{\pi^2}\left(arctg\ x + \frac{\pi}{2} \right)\left(arctg\ y + \frac{\pi}{2} \right) =$$

$$= \frac{1}{\pi^2}\left(arctg\ x + \frac{\pi}{2} \right)\left(arctg\ \infty + \frac{\pi}{2} \right) = \frac{1}{\pi}\left(arctg\ x + \frac{\pi}{2} \right).$$

Поскольку переменные x и y входят в функцию $F(x, y)$ симметрично, то, очевидно, функция $F(y)$ имеет аналогичную структуру: $F_X(y) = \frac{1}{\pi}\left(arctg\ y + \frac{\pi}{2} \right).$

Вершинами квадрата, заданного в условии задачи, являются точки $A(-1; -1)$, $B(-1; 1)$, $C(1; 1)$ и $D(1; -1)$. По свойству 4 двухмерной функции распределения получим:

$$P\{-1 \le X \le 1, -1 \le Y \le 1\} = F(1,1) - F(-1,1) - F(1,-1) + F(-1,-1\}$$

$$= \frac{1}{\pi^2}(arctg\,1 + \frac{\pi}{2})(arctg\,1 + \frac{\pi}{2}) - \frac{1}{\pi^2}(arctg\,(-1) + \frac{\pi}{2})(arctg\,1 + \frac{\pi}{2}) -$$

$$- \frac{1}{\pi^2}(arctg\,1 + \frac{\pi}{2})(arctg\,(-1) + \frac{\pi}{2}) + \frac{1}{\pi^2}(arctg\,(-1) + \frac{\pi}{2})(arctg\,(-1) + \frac{\pi}{2}).$$

Так как arctg1 = π/4 и arctg(–1) = –π/4, то P{–1 ≤ X ≤ 1, –1 ≤ Y ≤ 1} = 1/4 = 0,25.

3.3. Плотность распределения вероятностей двухмерной случайной величины

Найдём вероятность попадания двухмерной случайной величины (X, Y) в прямоугольник со сторонами Δx и Δy, полагая, что функция распределения вероятностей $F(x, y)$ является непрерывной и дифференцируемой. Пусть вершинами прямоугольника являются точки $A(x; y)$, $B(x; y + \Delta y)$, $C(x + \Delta x; y + \Delta y)$ и $D(x + \Delta x; y)$. Тогда $P\{x \le X < x + \Delta x, y \le Y < y + \Delta y\} = F(x + \Delta x, y + \Delta y) - F(x, y + \Delta y) - F(x + \Delta x, y) + F(x, y)$.

Поделим эту вероятность на $\Delta x \Delta y$ и перейдём к пределу при $\Delta x \to 0$ и $\Delta y \to 0$:

$$\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}} \frac{P\{x \le X < x + \Delta x, y \le Y < y + \Delta y\}}{\Delta x \cdot \Delta y} = \lim_{\Delta x\to 0} \frac{1}{\Delta x} \{\lim_{\Delta y\to 0} \frac{F(x + \Delta x, y + \Delta y) - F(x + \Delta x, y)}{\Delta y} -$$

$$- \lim_{\Delta y\to 0} \frac{F(x, y + \Delta y) - F(x, y)}{\Delta y}\} = \lim_{\Delta x\to 0} \frac{1}{\Delta x} \{\frac{\partial F(x + \Delta x, y)}{\partial y} - \frac{\partial F(x, y)}{\partial y}\}.$$

Последний предел равен, очевидно, второй смешанной частной производной. Этот предел называют *плотностью распределения вероятностей $f(x, y)$ двухмерной случайной величины* (X, Y), или *двухмерной плотностью распределения вероятностей*:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}. \tag{3.4}$$

Плотность распределения вероятностей является удобной формой задания закона распределения вероятностей случайных величин, но применение её ограничено только непрерывными случайными величинами. Графиком двумерной плотности распределения вероятностей является поверхность, расположенная над координатной плоскостью XOY. Сформулируем свойства двумерной плотности распределения вероятностей.

1. Всегда $f(x, y) \ge 0$, так как $F(x, y)$ является неубывающей функцией.
2. Непосредственно из (3.4) вытекает следующее равенство:

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) dv du. \tag{3.5}$$

3. Используя свойство 2 и, учитывая, что $F(\infty, \infty) = 1$, получим:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \tag{3.6}$$

4. Вероятность попадания случайной точки (X, Y) в произвольную область D равна интегралу от плотности распределения вероятностей по этой области:

$$P\{(X, Y) \in D\} = \iint_D f(x, y) dx dy. \tag{3.7}$$

Если D прямоугольник с вершинами в точках $A(a; c)$, $B(a; d)$, $C(b; d)$ и $E(b; c)$, то

$$P\{(X, Y) \in D\} = \int_a^b dx \int_c^d f(x, y) dy. \tag{3.8}$$

5. Из (3.5) следует, что

$$\int_{-\infty}^{\infty} f(x, y) dy = f_X(x) \qquad \text{и} \qquad \int_{-\infty}^{\infty} f(x, y) dx = f_Y(y). \tag{3.9}$$

Действительно, $F(x, \infty) = F_X(x)$ и $f_X(x) = F_X'(x)$. С другой стороны, по свойству 2,

$$F(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v) dv du.$$

Дифференцируя это равенство по x, получим первое равенство из (3.9).

Свойство 3 представляет *условие нормировки вероятностей*. Если D охватывает все возможные значения компонент X и Y, то (3.6) запишется так:

$$\iint_D f(x,y)dxdy = 1. \qquad (3.10)$$

Это означает, что объём тела, заключённого между поверхностью $z = f(x, y)$ и координатной плоскостью XOY, всегда равен единице. На рис. 3.3 изображён график плотности распределения вероятностей для величины (X, Y), все возможные значения которой (x, y) находятся в области D. На рис. 3.4 приводится график плотности распределения вероятностей случайной величины (X, Y), равномерно распределённой в прямоугольной области: $f(x, y) = 1/S_D$, если $(x, y)\in D$ и $f(x, y) = 0$, если $(x, y)\notin D$, $S_D = (b - a)(d - c)$.

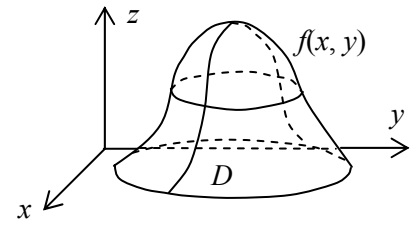

Рис. 3.3. **График плотности распределения вероятностей двухмерной величины (X, Y) в области D**

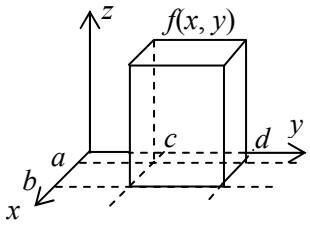

Рис. 3.4. **График плотности равномерного распределения величины (X, Y) в прямоугольнике**

Свойство 5 является общим правилом перехода от плотности распределения вероятностей размерности n к плотностям распределения вероятностей размерности n – 1. Для такого перехода достаточно n-мерную плотность распределения проинтегрировать по одному из аргументов в пределах от –∞ до ∞.

Пример 3.3. Задана плотность распределения вероятностей величины (X, Y):

$$f(x,y) = a\cdot\cos(x-y),\quad 0\le x\le\frac{\pi}{2},\quad 0\le y\le\frac{\pi}{2}.$$

Определить коэффициент a. Найти одномерные плотности $f_X(x)$ и $f_Y(y)$.

Воспользуемся условием нормировки вероятностей (3.6):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = \int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}} a\cdot\cos(x-y)dydx = a\cdot\int_0^{\frac{\pi}{2}}[\sin(\frac{\pi}{2}-y)-\sin(-y)]dy =$$

$$= a\cdot[\cos(\frac{\pi}{2}-\frac{\pi}{2})-\cos(\frac{\pi}{2}-0)-\cos(\frac{\pi}{2})+\cos(0)] = 2a = 1.$$ Отсюда находим $a = 1/2$.

Одномерные плотности определяем по свойству 5: $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{2}\int_0^{\frac{\pi}{2}}\cos(x-y)dx =$

$$= \frac{1}{2}[\sin(\frac{\pi}{2}-y)+\sin y],\quad f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{2}\int_0^{\frac{\pi}{2}}\cos(x-y)dy = -\frac{1}{2}[\sin(x-\frac{\pi}{2})-\sin x].$$ Применяя формулы приведения для синуса, получим:

$$f_X(x) = \frac{1}{2}(\cos x+\sin x),\ 0\le x\le\frac{\pi}{2};\quad f_Y(y) = \frac{1}{2}(\cos y+\sin y),\ 0\le y\le\frac{\pi}{2}.$$

3.4. Условные законы распределения. Статистическая зависимость

Зная плотность распределения вероятностей $f(x, y)$ двухмерной случайной величины (X, Y), можно определить одномерные плотности распределения $f_X(x)$ и $f_Y(y)$, пользуясь свойством 5. При таком переходе вся информация о характере связи между величинами X и Y будет потеряна. Обратный переход от одномерных плотностей $f_X(x)$ и $f_Y(y)$ к двухмерной плотности $f(x, y)$ оказывается невозможным.

Для установления связи между двухмерной и одномерными плотностями распределения вероятностей вводится понятие условной плотности распределения вероятностей. *Условной плотностью распределения вероятностей* случайной величины Y по отношению к величине X называют плотность распределения вероятностей величины Y, вычисленную при условии, что величина X приняла значение x. Обозначают $f(y \mid x)$. Можно показать, что для двух непрерывных величин

$$f(x, y) = f(y \mid x) \cdot f_X(x) \text{ и } f(x, y) = f(x \mid y) \cdot f_Y(y). \qquad (3.11)$$

Двухмерная плотность распределения вероятностей равна произведению условной плотности распределения вероятности одной из этих величин по отношению к другой на безусловную плотность распределения вероятности второй величины. Для восстановления двухмерной плотности необходимо знать одну из условных плотностей и соответствующую одномерную плотность. Формулы (3.11) называют формулами умножения плотностей распределения вероятностей. Их структура совпадает со структурой формулы умножения вероятностей событий.

Условные плотности распределения вероятностей содержат всю информацию о зависимости между случайными величинами. Может оказаться, что условная плотность полностью совпадает с безусловной плотностью. В таком случае говорят, что данные величины являются *статистически независимыми*. Заметим, что если величина Y не зависит от величины X, то и X не зависит от Y. Действительно, пусть $f(y \mid x) = f_Y(y)$. Левые части равенств (3.11) равны, поэтому равны и правые части, т.е.

$$f(x \mid y) \cdot f_Y(y) = f(y \mid x) \cdot f_X(x) \qquad \text{или} \qquad f(x \mid y) = \frac{f(y \mid x) \cdot f_X(x)}{f_Y(y)}.$$

Отсюда следует, что если $f(y \mid x) = f_Y(y)$, то $f(x \mid y) = f_X(x)$.

Если условные плотности распределения вероятностей не равны безусловным плотностям, то случайные величины X и Y являются *статистически зависимыми*. Формулы умножения плотностей распределения (3.11) справедливы в любом случае. Для статистически независимых величин формулы сводятся к одной:

$$f(x, y) = f_X(x) \cdot f_y(y). \qquad (3.12)$$

Это равенство является признаком независимости случайных величин.

Из записанных соотношений вытекает следующая формула для определения условных плотностей распределения вероятностей по двухмерной плотности:

$$f(y \mid x) = \frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y)dy} \qquad \text{и} \qquad f(x \mid y) = \frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y)dx}. \qquad (3.13)$$

Аналогично вводится понятие условной плотности распределения вероятностей для многомерных величин более высокой размерности.

Вообще *условным законом распределения вероятностей* одной случайной величины по отношению к другой называется её закон распределения, полученный при условии, что вторая величина приняла конкретное значение. Пусть (X, Y) дискретная случайная величина. Возможными значениями компоненты X являются значения $x_1, x_2, \ldots, x_s$, а возможными значениями Y – $y_1, y_2, \ldots, y_q$. Заданы вероятности $p_{ij} = P\{X = x_i, Y = y_j\}$. Условный закон распределения вероятностей величины X при условии, что величина Y приняла значение y_j, будет иметь следующий вид:

$$P\{X = x_i \mid Y = y_j\} = \frac{p_{ij}}{\sum_{i=1}^{s} p_{ij}} = \frac{p_{ij}}{p_{*j}}, \quad i = 1, 2, ..., s. \tag{3.14}$$

Здесь изменяется индекс i, но не изменяется индекс j. Заметим, что по своему смыслу эта формула полностью эквивалентна второй формуле из (3.13). Другими словами, любая строка (любой столбец) в таблице задания вероятностей p_{ij} (см. п. 3.1), поделённая на сумму вероятностей этой строки (этого столбца), представляет собой условный закон распределения вероятностей одной из величин по отношению к фиксированному значению другой величины. Если, например, все вероятности второй строки таблицы поделить на p_{2*}, то получим условный закон распределения случайной величины Y при условии, что величина X приняла значение x_2.

Для дискретных величин статистическая зависимость определяется аналогично. Если условное распределение вероятностей одной величины по отношению к другой совпадает с безусловным законом, т.е. $P\{X = x_i \mid Y = y_j\} = p_{i*}$, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, q$, то величины X и Y являются статистически независимыми. Если это условие не выполняется, то величины являются статистически зависимыми.

Пример 3.4. Задана плотность распределения вероятностей двухмерной случайной величины (X, Y): $f(x,y) = \begin{cases} \dfrac{1}{\pi R^2} & \text{при} \quad x^2 + y^2 < R, \\ 0 & \text{при} \quad x^2 + y^2 \geq R, \quad R > 0. \end{cases}$

Доказать, что случайные величины X и Y являются зависимыми.

Случайные величины X и Y будут зависимыми, если $f(x \mid y) \neq f_X(x)$. Найдём $f_X(x)$:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\pi R^2} \cdot \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy = \frac{2\sqrt{R^2 - x^2}}{\pi R^2}, \quad -R < x < R.$$

Здесь при расстановке пределов интегрирования было учтено, что плотность $f(x, y)$ равна нулю вне круга радиуса R. В силу симметрии плотности $f(x, y)$, плотность $f_Y(y)$ имеет такую же структуру: $f_Y(y) = \dfrac{2\sqrt{R^2 - y^2}}{\pi R^2}$, $-R < y < R$. Следовательно, $f(x \mid y) = \dfrac{f(x,y)}{f_Y(y)} = \dfrac{1}{\pi R^2} \cdot \dfrac{\pi R^2}{2\sqrt{R^2 - y^2}} = \dfrac{1}{2\sqrt{R^2 - y^2}} \neq f_X(x) = \dfrac{2\sqrt{R^2 - x^2}}{\pi R^2}$. Это доказывает, что X и Y являются зависимыми величинами. Отметим, что условные плотности $f(x \mid y)$ и $f(y \mid x)$ также имеют одинаковые структуры. Следует иметь в виду, что они определены только внутри круга: $x^2 + y^2 < R^2$. Вне этого круга они равны нулю.

Для доказательства существования зависимости между величинами необязательно находить условные плотности распределения вероятностей. Достаточно убедиться, что $f(x, y) \neq f_X(x) f_Y(y)$. В данном случае это неравенство очевидно.

3.5. Числовые характеристики многомерных случайных величин. Ковариационный момент и коэффициент корреляции

При изучении свойств двухмерной случайной величины (X, Y) можно говорить о математических ожиданиях отдельных компонент m_x и m_y, о их дисперсиях D_x и D_y и т. п. Однако знание характеристик изолированных компонент не позволяет делать выводы о существовании статистической связи между этими компонентами и о характере этой связи. При изучении многомерных величин дополнительно привлекают такие характеристики, которые отражают статистическую связь. К таким характеристикам относятся смешанные моменты случайных величин.

Пусть для двухмерной случайной величины (X, Y) задана плотность распределения вероятностей $f(x, y)$. *Начальным моментом* ν_{gh} порядка $k = g + h$ двухмерной величины (X, Y) называется математическое ожидание произведения $X^g \cdot Y^h$:

$$\nu_{gh} = M\{X^g \cdot Y^h\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^g y^h f(x,y)dxdy\,. \qquad (3.15)$$

Если (X, Y) двухмерная дискретная случайная величина с возможными значениями (x_i, y_j), $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, q$, и с вероятностями p_{ij}, то

$$\nu_{gh} = M\{X^g \cdot Y^h\} = \sum_{i=1}^{s}\sum_{j=1}^{q} x_i^g y_j^h p_{ij}\ . \qquad (3.16)$$

В формулах (3.15) и (3.16) g и h – целые числа, $g \geq 0$, $h \geq 0$, $g + h = k$, $k = 0, 1, 2, \ldots$.

Число моментов порядка k будет равно $k + 1$. При $k = 0$ имеется всего один момент ν_{00}, который всегда равен единице (условие нормировки вероятностей). При $k = 1$ имеется два момента ν_{01} и ν_{10}, которые совпадают с математическими ожиданиями величин X и Y: $\nu_{01} = m_y$ и $\nu_{10} = m_x$. Докажем, например, первое равенство:

$$\nu_{10} = M\{X^1 \cdot Y^0\} = M\{X\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f(x,y)dxdy = \int_{-\infty}^{\infty} x \cdot [\int_{-\infty}^{\infty} f(x,y)dy]dx = \int_{-\infty}^{\infty} x \cdot f_X(x)dx = m_x\,.$$

Перечислим моменты порядка k: ν_{0k}, $\nu_{1(k-1)}$, $\nu_{2(k-2)}$, …, $\nu_{(k-1)1}$, ν_{k0}. Моменты ν_{0k} и ν_{k0} являются моментами k-ого порядка соответственно величин Y и X. Остальные моменты (из перечисленных) при $k > 1$ представляют собой смешанные моменты, которые и характеризуют статистическую связь между компонентами X и Y.

Для анализа статистической связи удобнее использовать центральные моменты. *Центральный момент* μ_{gh} порядка k двухмерной величины (X, Y) это математическое ожидание произведения $(X - m_x)^g \cdot (Y - m_y)^h$, так что $g + h = k$, т.е.

$$\mu_{gh} = M\{(X - m_x)^g \cdot (Y - m_y)^h\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - m_x)^g (y - m_y)^h f(x,y)dxdy..\qquad (3.17)$$

Если (X, Y) двухмерная дискретная случайная величина с возможными значениями (x_i, y_j), $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, q$, и с вероятностями p_{ij}, то

$$\mu_{gh} = M\{(X - m_x)^g \cdot (Y - m_y)^h\} = \sum_{i=1}^{s}\sum_{j=1}^{q} (x_i - m_x)^g (y_j - m_y)^h p_{ij}\ . \qquad (3.18)$$

В этих формулах g и h – целые числа, $g \geq 0$, $h \geq 0$, $g + h = k$, $k = 0, 1, 2, \ldots$.

Рассмотрим моменты 2-го порядка. Момент μ_{20} является, очевидно, дисперсией D_x случайной величины X, а момент μ_{02} – дисперсией D_y величины Y. Смешанный момент

μ_{11} имеет первостепенное значение при изучении зависимости между случайными величинами. Этот момент принято называть *ковариационным моментом*, *моментом связи* или просто *ковариацией* и обозначать через K_{xy}. Ковариационный момент характеризует линейную связь между рассматриваемыми величинами. Легко показать, что если X и Y являются статистически независимыми величинами, то корреляционный момент равен нулю. Действительно, если X и Y независимы, то $f(x, y) = f_X(x)f_Y(y)$ и, следовательно,

$$K_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_x)(y - m_y) f(x, y)dxdy =$$

$$= \int_{-\infty}^{\infty}(x - m_x) f_X(x)dx \int_{-\infty}^{\infty}(y - m_y) f_Y(y)dy = \mu_{x1} \cdot \mu_{y1} = 0.$$

Здесь μ_{x1} и μ_{y1} есть центральные моменты первого порядка величин соответственно X и Y. Эти моменты всегда равны нулю.

Таким образом, если величины X и Y являются статистически независимыми, то $K_{xy} = 0$. Однако обратное утверждение не является верным, т.е. если $K_{xy} = 0$, то это ещё не значит, что данные величины являются статистически независимыми. Равенство нулю ковариационного момента означает факт, что между величинами отсутствует линейная связь. Однако может существовать нелинейная связь, которая может выступать в разных формах. Форма связи при $K_{xy} = 0$ определяется моментами более высокого порядка. Например, момент μ_{21} определяет параболическую связь между величиной Y и величиной X, момент μ_{31} – кубическую связь и т.д.

Рассмотрим подробнее вопрос о линейной связи между двумя величинами. Предположим, что между случайными величинами X и Y существует статистическая зависимость, причём точно известно, что эта зависимость линейная. Тогда справедливо равенство $Y = aX + b + \xi$. В этом равенстве слагаемое $aX + b$ представляет собой компоненту величины Y, линейную относительно величины X, причём a и b являются постоянными коэффициентами. Слагаемое ξ представляет собой некоторую случайную величину, не зависящую от X. Пусть m_x и m_ξ есть математические ожидания величин X и ξ, а D_x – дисперсия величины X. Тогда математическое ожидание величины Y будет равно $m_y = M\{aX + b + \xi\} = am_x + b + m_\xi$. Следовательно, $K_{xy} = M\{(Y - m_y)(X - m_x)\} = M\{(aX + b + \xi - am_x - b - m_\xi)(X - m_x)\} = M\{[a(X - m_x) + (\xi - m_\xi)](X - m_x)\} = aM\{(X - m_x)^2\} + M\{(\xi - m_\xi)(X - m_x)\} = aD_x + K_{\xi x}$. Но $K_{\xi x} = 0$, так как, по предположению, связь между величинами X и ξ отсутствует. Таким образом, $K_{xy} = aD_x$. Это означает, что ковариационный момент пропорционален коэффициенту, обеспечивающему линейную связь между величинами. Если линейная связь отсутствует, то $a = 0$ и ковариационный момент также обращается в нуль.

Изучение линейной связи между величинами является наиболее простой и, в то же время, неизбежной задачей при исследовании статистической зависимости между величинами. Именно поэтому придаётся большое значение определению ковариационного момента. При изучении свойств совокупности n случайных величин X_1, X_2, …, X_n используется понятие *ковариационной матрицы* – матрицы, состоящей из ковариационных моментов попарно рассматриваемых величин из данной совокупности:

$$K = \begin{pmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{pmatrix}. \qquad (3.19)$$

Здесь K_{ij} – ковариационный момент между величинами X_i и X_j, $1 \le i \le n$, $1 \le j \le n$. Для унификации записи введено обозначение: $\sigma_i^2 = K_{ii}$. Последнее означает, что диспер-

сия σ_i^2 формально рассматривается как ковариационный момент между X_i и X_i. Ковариационная матрица даёт полное представление о линейной связи между величинами рассматриваемой совокупности. Анализ матрицы позволяет выявить те компоненты, между которыми линейная зависимость наиболее существенна.

Величина ковариационного момента K_{xy} зависит от дисперсий σ_x^2 и σ_y^2 величин X и Y. Поэтому затруднительно судить о силе линейной связи между величинами, опираясь на значение коэффициента K_{xy}. Если, например, $K_{xy} > K_{uv}$, то ещё нельзя утверждать, что линейная связь между Y и X сильнее, чем между величинами V и U. Это связано с тем, что у пары величин Y и X дисперсии могут значительно отличаться от дисперсий пары V и U. С целью устранения этого неудобства для измерения силы линейной связи был введён *коэффициент корреляции*:

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}. \tag{3.20}$$

Коэффициент корреляции является безразмерной величиной, не зависящей от дисперсий σ_x^2 и σ_y^2, причём всегда выполняется неравенство:

$$|r_{xy}| \le 1 \text{ или } -1 \le r_{xy} \le 1. \tag{3.21}$$

Если $|r_{xy}| = 1$, то имеется функциональная линейная связь между величинами Y и X, т.е. $Y = aX + b$, где a и b – некоторые постоянные коэффициенты. Если $r_{xy} = -1$, то $a < 0$, если $r_{xy} = 1$, то $a > 0$. Другими словами, при $r_{xy} = -1$ имеем монотонно убывающую функцию связи между Y и X, а при $r_{xy} = 1$ – монотонно возрастающую функцию.

Если $r_{xy} = 0$, то линейная связь между величинами Y и X отсутствует. Последнее означает, что или величины Y и X являются статистически независимыми, или между ними существуют только нелинейные виды связей.

При рассмотрении совокупности n случайных величин ($n > 2$) используется *корреляционная матрица* – матрица, составленная из коэффициентов корреляции:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}. \tag{3.22}$$

Корреляционная матрица более наглядно представляет линейные связи между рассматриваемыми величинами. Из определения ковариационного момента и коэффициента корреляции следует, что всегда имеют место равенства $K_{xy} = K_{yx}$ и $r_{xy} = r_{yx}$. Таким образом, ковариационная (3.19) и корреляционная (3.22) матрицы являются симметричными относительно главных диагоналей.

Если коэффициент r_{xy} между величинами X и Y равен нулю, то говорят, что эти случайные величины являются *некоррелированными*. Если $r_{xy} \neq 0$, величины называют *коррелированными*. Статистически независимые величины являются некоррелированными. Обратное утверждение в общем случае не является верным.

Вычисление начальных моментов почти всегда менее трудоёмко, чем вычисление центральных моментов. Поэтому для определения центральных моментов часто применяют формулы связи между начальными и центральными моментами. В частности, для ковариационного момента $K_{xy} = M\{(X - m_x)\cdot(Y - m_y)\} = M\{XY - m_xY - m_yX + m_xm_y\}$. Пользуясь свойствами математического ожидания, получим:

$$K_{xy} = M\{XY\} - m_xm_y \text{ или } K_{xy} = \nu_{11} - m_xm_y. \tag{3.23}$$

Из формулы (3.23) вытекает одно очень важное свойство математических ожиданий. Если ковариационный момент (или коэффициент корреляции) равен нулю, то математическое ожидание произведения случайных величин равно произведению их математических ожиданий. Справедливо и обратное утверждение. Однако из равенства $\nu_{11} = m_x m_y$ ещё не следует, что величины X и Y являются статистически независимыми.

Пример 3.5. Вычислить числовые характеристики m_x, m_y, D_x, D_y, K_{xy} и r_{xy} для двухмерной дискретной случайной величины, заданной в примере 3.1.

Для определения характеристик m_x, m_y, D_x и D_y воспользуемся одномерными законами распределения вероятностей, уже определёнными в примере 3.1:

$$m_x = \sum_{i=1}^{s} x_i p_{i*} = 3\cdot 0{,}15 + 4\cdot 0{,}25 + 6\cdot 0{,}35 + 8\cdot 0{,}25 = 5{,}55; \quad m_y = \sum_{j=1}^{q} y_j p_{*j} = -1\cdot 0{,}16 + 0\cdot 0{,}26 +$$

$$+ 1\cdot 0{,}25 + 2\cdot 0{,}20 + 3\cdot 0{,}13 = 0{,}88.$$ Найдём 2-е начальные моменты для величин X и Y:

$$M\{X^2\} = \sum_{i=1}^{s} x_i^2 p_{i*} = 9\cdot 0{,}15 + 16\cdot 0{,}25 + 36\cdot 0{,}35 + 64\cdot 0{,}25 = 33{,}95; \; M\{Y^2\} =$$

$$= \sum_{j=1}^{q} y_j^2 p_{*j} = +1\cdot 0{,}25 + 2\cdot 0{,}20 + 3\cdot 0{,}13 = 0{,}88.$$

Следовательно, $D_x = 33, 95 - (5, 55)^2 \approx 3, 15$ и $D_y = 2, 38 - (0, 88)^2 \approx 1, 61$. Тогда средние квадратические отклонения будут равны:

$$\sigma_x = \sqrt{D_x} = \sqrt{3{,}15} \approx 1{,}77; \qquad \sigma_y = \sqrt{D_y} = \sqrt{1{,}61} \approx 1{,}27.$$

Ковариационный момент вычисляется непосредственно по исходной таблице вероятностей. Применим формулу (3.23)

$$K_{xy} = \nu_{11} - m_x m_y = M\{XY\} - m_x m_y = \sum_{i=1}^{s}\sum_{j=1}^{q} x_i y_j - m_x m_y =$$

$$= 3\cdot((-1)\cdot 0{,}03 + 0\cdot 0{,}04 + 1\cdot 0{,}03 + 2\cdot 0{,}03 + 3\cdot 0{,}02) +$$

$$+ 4\cdot((-1)\cdot 0{,}04 + 0\cdot 0{,}06 + 1\cdot 0{,}07 + 2\cdot 0{,}05 + 3\cdot 0{,}03) +$$

$$+ 6\cdot((-1)\cdot 0{,}05 + 0\cdot 0{,}08 + 1\cdot 0{,}09 + 2\cdot 0{,}08 + 3\cdot 0{,}05) +$$

$$+ 8\cdot((-1)\cdot 0{,}04 + 0\cdot 0{,}08 + 1\cdot 0{,}06 + 2\cdot 0{,}04 + 3\cdot 0{,}03) - 5{,}55\cdot 0{,}88 = -0{,}024.$$

Следовательно, коэффициент корреляции будет равен $r_{xy} = \dfrac{K_{xy}}{\sigma_x \sigma_y} = \dfrac{-0{,}024}{1{,}77\cdot 1{,}27} \approx -0{,}011.$

Существует очень слабая линейная связь между X и Y. Однако коэффициент корреляции настолько мал, что этой линейной связью можно пренебречь. Для выявления нелинейной связи необходимо вычислять смешанные моменты более высоких порядков.

Пример 3.6. Вычислить числовые характеристики m_x, m_y, D_x, D_y, K_{xy} и r_{xy} для двухмерной непрерывной случайной величины, заданной в примере 3.3.

Одномерные плотности распределения вероятностей, найденные в примере 3.3, имеют одинаковую структуру. Поэтому математические ожидания и дисперсии величин X и Y будут равными. Одномерная плотность $f_X(x)$ имеет следующий вид: $f_X(x) = \dfrac{1}{2}(\cos x + \sin x)$. Найдём m_x и D_x, применяя интегрирование по частям:

$$m_x = \int_{-\infty}^{\infty} x \cdot f_X(x)dx = \frac{1}{2}\int_0^{\frac{\pi}{2}} x(\cos x + \sin x)dx = \frac{1}{2}[(x\sin x + \cos x)\Big|_0^{\frac{\pi}{2}} - (x\cos x - \sin x)\Big|_0^{\frac{\pi}{2}}] =$$

$$= \frac{1}{2}[\frac{\pi}{2} - 1 + 1] = \frac{\pi}{4}. \quad D_x = M\{X^2\} - m_x^2 = \frac{1}{2}\int_0^{\frac{\pi}{2}} x^2[\cos x + \sin x]dx - m_x^2. \quad \text{Интегрируя, получим:}$$

$D_x = \frac{\pi^2}{8} + \frac{\pi}{2} - 2 - \left(\frac{\pi}{4}\right)^2 = \frac{\pi^2}{16} + \frac{\pi}{2} - 2.$ Итак, $m_x = m_y = \frac{\pi}{4}$ è $D_x = D_y = \frac{\pi^2}{16} + \frac{\pi}{2} - 2.$

Ковариационный момент вычислим по формуле (3.23):

$$\nu_{11} = M\{XY\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \frac{1}{2}\int_0^{\frac{\pi}{2}}\int_0^{\frac{\pi}{2}} xy\cos(x-y)dxdy = \frac{1}{2}\int_0^{\frac{\pi}{2}} y\int_0^{\frac{\pi}{2}} x\cos(x-y)dxdy =$$

$$= \frac{1}{2}\int_0^{\frac{\pi}{2}} y[\frac{\pi}{2}\sin(\frac{\pi}{2} - y) + \cos(\frac{\pi}{2} - y) - \cos y]dy = \frac{1}{2}\int_0^{\frac{\pi}{2}} y[(\frac{\pi}{2} - 1)\cos y + \sin y]dy = \frac{\pi^2}{8} - \frac{\pi}{2} + 1.$$

Тогда $K_{xy} = \nu_{11} - m_x^2 = \frac{\pi^2}{8} - \frac{\pi}{2} + 1 - \frac{\pi^2}{16} = \frac{\pi^2}{16} - \frac{\pi}{2} + 1$. Следовательно, коэффициент корреляции r_{xy} будет равен: $r_{xy} = \frac{K_{xy}}{\sigma_x\sigma_y} = (\frac{\pi^2}{16} - \frac{\pi}{2} + 1)\Big/(\frac{\pi^2}{16} + \frac{\pi}{2} - 2) = \frac{(\pi-4)^2}{\pi^2 + 8\pi - 32}$.

Численные расчёты показывают, что $m_x = m_y \approx 0{,}785$, $D_x = D_y \approx 0{,}188$, $K_{xy} \approx 0{,}046$ и $r_{xy} \approx$ $\approx 0{,}246$. Последнее означает, что между величинами X и Y существует линейная связь. Если ставится задача по прогнозированию ожидаемых значений величины Y по известным значениям величины X, то можно пользоваться линейной функцией $y = ax + b$ (вопрос определения коэффициентов a и b этой функции будет рассмотрен позднее). При этом точность прогноза будет достаточно низкой, так как коэффициент корреляции относительно мал. Можно предположить, что имеется ещё и нелинейная связь. Если учесть её, то можно повысить точность прогнозирования.

3.6 Условные числовые характеристики. Линии регрессии. Корреляционное отношение

Числовые характеристики, найденные на основе условных законов распределения вероятностей, называются *условными числовыми характеристиками* случайных величин. При анализе зависимости между величинами часто ограничиваются рассмотрением *условного математического ожидания и условной дисперсии*.

Для двухмерной случайной величины (X, Y) можно определить два условных закона распределения вероятностей: условный закон для величины Y по отношению к X и условный закон распределения величины X по отношению к Y. Таким образом, можно выделить две группы условных числовых характеристик: условные числовые характеристики случайной величины Y по отношению к X и условные числовые характеристики случайной величины X по отношению к Y. Условное математическое ожидание величины Y по отношению к X будем обозначать через $m_y(x)$, а условное математическое ожидание величины X по отношению к Y – через $m_x(y)$. Условные дисперсии будем обозначать через $D_y(x)$ и $D_x(y)$.

Для непрерывных случайных величин (X, Y)

$$m_y(x) = \int_{-\infty}^{\infty} y \cdot f(y \mid x)dy \qquad \text{и} \qquad m_x(y) = \int_{-\infty}^{\infty} x \cdot f(x \mid y)dx. \tag{3.24}$$

Для дискретных величин (X, Y) с вероятностями $p_{ij} = P\{X < x_i, Y < y_j\}$, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, q$, условные математические ожидания определяются формулами (см. (3.14)):

$$m_y(x_i) = \frac{1}{p_{i*}} \cdot \sum_{j=1}^{q} y_j p_{ij} \qquad \text{и} \qquad m_x(y_j) = \frac{1}{p_{*j}} \cdot \sum_{i=1}^{s} x_i p_{ij}. \tag{3.25}$$

Условное математическое ожидание имеет большое значение при исследовании статистической зависимости между величинами. Если нужно предсказывать ожидаемые значения величины Y по известным значениям величины X, то следует использовать условное математическое ожидание. Функция

$$y = m_y(x) \tag{3.26}$$

называется *регрессией* величины Y на X, а график этой функции – *линией регрессии* Y на X. Оказывается, что применение этой функции для прогнозирования ожидаемых значений Y обеспечивает минимальную среднеквадратическую ошибку прогнозирования. Пусть для прогнозирования мы применяем функцию $y = \varphi(x)$. Тогда средний квадрат ошибки прогнозирования будет равен $M\{[Y - \varphi(X)]^2\}$. При любых законах распределения минимум этой величины достигается в случае $\varphi(x) = m_y(x)$.

Аналогично определяется регрессия величины X на Y: $x = m_x(y)$. Линии регрессии $y = m_y(x)$ и $x = m_x(y)$ расположены на координатной плоскости XOY симметрично относительно биссектрисы первого и третьего координатных углов. В связи с этим регрессии $y = m_y(x)$ и $x = m_x(y)$ можно считать взаимно обратными функциями.

Регрессия не охватывает всю глубину статистической связи между величинами. Может оказаться, что $m_y(x) = m_y$. В таком случае регрессия оказывается бесполезной с точки зрения возможного прогнозирования. Однако нельзя делать вывод, что величина X не влияет на Y. Действительно, величина X, не оказывая влияния на математическое ожидание величины Y, может существенно влиять на её дисперсию или на её моменты более высоких порядков. Для уточнения характера влияния одной величины на другую используются условные моменты более высоких порядков, в частности условная дисперсия.

Для непрерывных случайных величин (X, Y)

$$D_y(x) = \int_{-\infty}^{\infty} [y - m_y(x)]^2 \cdot f(y \mid x)dy; \quad D_x(y) = \int_{-\infty}^{\infty} [x - m_x(y)]^2 \cdot f(x \mid y)dx. \tag{3.27}$$

Аналогичные формулы для дискретных величин (X, Y) с вероятностями $p_{ij} = P\{X < x_i, Y < y_j\}$, $i = 1, 2, \ldots, s$; $j = 1, 2, \ldots, q$, имеют следующий вид:

$$D_y(x_i) = \frac{1}{p_{i*}} \cdot \sum_{j=1}^{q} [y_j - m_y(x_i)]^2 p_{ij}; \quad D_x(y_j) = \frac{1}{p_{*j}} \cdot \sum_{i=1}^{s} [x_i - m_x(y_j)]^2 p_{ij}. \tag{3.28}$$

Подчеркнём ещё раз, что тождественное совпадение условных и безусловных дисперсий не позволяет делать вывод о независимости случайных величин. Это замечание относится к любым условным характеристикам. Если хотя бы одна из условных характеристик не совпадает с соответствующей безусловной характеристикой, то рассматриваемые величины являются статистически зависимыми.

Вернёмся снова к упомянутой выше задаче прогнозирования ожидаемых значений одной величины по известным значениям другой величины. Пусть прогнозируются значения величины Y по известным значениям величины X с использованием известной

регрессии $y = m_y(x)$. На самом деле в результате опыта будут появляться значения величины Y, отличающиеся от прогнозируемых значений: $y = m_y(x) + \xi$. Случайная величина ξ является ошибкой прогнозирования. Тогда средний квадрат $D_{ост}$ ошибки ξ будет равен:

$$D_{ост} = M\{\xi^2\} = M\{[Y - m_y(X)]^2\}. \quad (3.29)$$

Величина $D_{ост}$ называется *остаточной дисперсией*. Смысл термина «остаточная дисперсия» заключается в следующем. Неопределённость в предсказании ожидаемых значений Y полностью определяется дисперсией D_y. Сама дисперсия D_y обусловлена, во-первых, влиянием на Y величины X, во-вторых, всеми другими (неучтёнными в данном опыте) факторами. Если проводится прогнозирование Y с использованием регрессии $y = m_y(x)$, то неопределённость, связанная с влиянием X, снимается. Остаётся неопределённость, обусловленная неучтёнными факторами, которая характеризуется остаточной дисперсией. Не следует путать остаточную дисперсию с условной дисперсией. Первая из них есть число, а вторая – функция.

Представим дисперсию D_y в виде суммы двух компонент:

$$D_y = M\{[Y - m_y]^2\} = M\{[Y - m_y(X) + m_y(X) - m_y]^2\} = M\{[Y -$$
$$- m_y(X)]^2\} + 2M\{[Y - m_y(X)][m_y(X) - m_y]\} + M\{[m_y(X) - m_y]^2\}.$$

Можно показать, что второе слагаемое в этом выражении равно нулю. Тогда

$$D_y = M\{[Y - m_y(X)]^2\} + M\{[m_y(X) - m_y]^2\}. \quad (3.30)$$

Поделим равенство (3.30) на D_y: $\dfrac{M\{[Y - m_y(X)]^2\}}{D_y} + \dfrac{M\{[m_y(X) - m_y]^2\}}{D_y} = 1.$

Первое слагаемое в левой части этого соотношения представляет долю дисперсии D_y, обусловленную неконтролируемой компонентой ξ, второе слагаемое – долю дисперсии D_y, обусловленную контролируемой компонентой $m_y(X)$. Это второе отношение используется в качестве меры статистической связи между величинами Y и X. Оно называется *корреляционным отношением* и обозначается $\eta^2_{y/x}$ (читается: «эта квадрат игрек на икс»). Таким образом,

$$\eta^2_{y/x} = \frac{M\{[m_y(X) - m_y]^2\}}{D_y} = 1 - \frac{M\{[Y - m_y(X)]^2\}}{D_y} = 1 - \frac{D_{ост}}{D_y}. \quad (3.31)$$

Для корреляционного отношения $\eta^2_{y/x}$ всегда справедливо неравенство:

$$0 \le \eta^2_{y/x} \le 1. \quad (3.32)$$

Равенство $\eta^2_{y/x} = 0$ означает, что $m_y(X) \equiv m_y$. При этом оказывается, что остаточная дисперсия $D_{ост}$ совпадает с дисперсией D_y, т.е. прогнозирование вообще невозможно. Из этого факта нельзя сделать вывод о полной независимости рассматриваемых случайных величин, поскольку может иметь место влияние X на другие числовые характеристики величины Y. Если $\eta^2_{y/x} = 1$, то существует функциональная связь между Y и X, которая точно описывается регрессией $y = m_y(x)$. Остаточная дисперсия $D_{ост}$ будет равна нулю, и прогнозирование ожидаемых значений величины Y по известным значениям величины X с помощью регрессии будет точным.

Аналогично определяется и второе корреляционное отношение $\eta^2_{x/y}$. Корреляционные отношения могут совпадать или не совпадать. Поэтому при изучении зависимости между величинами необходимо вычислять оба отношения.

Корреляционные отношения характеризуют и линейную и нелинейную связь в совокупности. Поэтому всегда справедливы два неравенства:

$$\eta^2_{y/x} \geq r^2_{xy} \qquad u \qquad \eta^2_{x/y} \geq r^2_{xy}, \tag{3.33}$$

где r_{xy} – коэффициент корреляции между величинами X и Y. Сравнивая коэффициент корреляции с корреляционными отношениями, можно сделать определённые выводы о характере статистической зависимости. Так, если корреляционные отношения велики по сравнению с квадратом коэффициента корреляции, то преобладает нелинейная связь между величинами. Если, напротив, $\eta^2_{y/x} \approx r^2_{xy}$ и, одновременно, $\eta^2_{x/y} \approx r^2_{xy}$, то преобладает линейная связь. Если хотя бы одно из корреляционных отношений равно нулю, то линейная связь полностью отсутствует.

3.7. Двухмерное нормальное распределение

Двухмерная случайная величина (X,Y) с нормальным законом распределения вероятностей встречается относительно часто при рассмотрении конкретных практических задач. Проведём краткий анализ такой двухмерной случайной величины. Говорят, что двухмерная случайная величина (X,Y) распределена по нормальному закону, если её плотность распределения вероятностей имеет следующий вид:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2_{xy}}} \cdot \varphi(x,y), \quad \text{где} \quad \varphi(x,y) = \tag{3.34}$$

$$= \exp\left\{ -\frac{1}{2(1-r^2_{xy})} \left[\frac{(x-m_x)^2}{\sigma^2_x} - 2r_{xy}\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma^2_y} \right] \right\}.$$

Найдём одномерную плотность распределения вероятностей $f_X(x)$. Выражение, стоящее в квадратных скобках, можно преобразовать, выделяя полный квадрат, получим:

$$\left[\frac{y-m_y}{\sigma_y} - r_{xy}\frac{x-m_x}{\sigma_x} \right]^2 + (1-r^2_{xy})\frac{(x-m_x)^2}{\sigma^2_x}.$$

Тогда $f_X(x) = \int\limits_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\{-\frac{(x-m_x)^2}{2\sigma^2_x}\} \cdot \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r^2_{xy}}} \times$

$$\times \int\limits_{-\infty}^{\infty} \exp\{-\frac{1}{2(1-r^2_{xy})}[\frac{y-m_y}{\sigma_y} - r_{xy}\frac{x-m_x}{\sigma_x}]^2 dy\,.$$

Сделаем замену переменной, положив $\frac{1}{\sqrt{1-r^2_{xy}}}[\frac{y-m_y}{\sigma_y} - r_{xy}\frac{x-m_x}{\sigma_x}] = t, \quad \frac{dy}{\sigma_y\sqrt{1-r^2_{xy}}} = dt.$

Получим: $f_X(x) = \cdot\frac{1}{\sqrt{2\pi}\sigma_x}\exp\{-\frac{(x-m_x)^2}{2\sigma^2_x}\} \cdot \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \exp\{-\frac{t^2}{2}\}dt\,.$

Последний интеграл вместе с коэффициентом, стоящим перед ним, представляет собой интеграл от плотности распределения вероятностей некоторой случайной величины T. Этот интеграл называют ещё интегралом Пуассона, он равен единице:

$$\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \exp\{-\frac{t^2}{2}\}dt = 1.$$

Окончательно,

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\{-\frac{(x-m_x)^2}{2\sigma_x^2}\}. \quad (3.35)$$

Очевидно, что плотность $f_Y(y)$будет иметь такую же структуру, однако в ней вместо параметров m_x и σ_x^2 будут стоять m_y и σ_y^2

Таким образом, $f_X(x)$ представляет собой обычную плотность нормального распределения одномерной случайной величины X. Это означает, что математическое ожидание и дисперсия величины X равны соответственно m_x и σ_x^2. Аналогично можно убедиться, что m_y и σ_y^2 являются соответственно математическим ожиданием и дисперсией случайной величины Y. Достаточно громоздкие, но простые вычисления приводят к выводу, что параметр r_{xy} является не чем иным, как коэффициентом корреляции между величинами X и Y. Таким образом, все параметры двухмерной плотности по своему смыслу соответствуют их исходным обозначениям.

Поскольку известны $f(x,y)$ и $f_X(x)$, то можно определить условную плотность распределения вероятностей $f(y \mid x)$, пользуясь тем, что $f(x,y)=f(y \mid x)f_X(x)$:

$$f(y \mid x) = \frac{1}{\sigma_y\sqrt{2\pi(1-r_{xy}^2)}}\exp\{-\frac{1}{2\sigma_y^2(1-r_{xy}^2)}[y-m_y-r_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x)]^2\}.$$

Структура условной плотности распределения вероятностей полностью соответствует структуре нормального распределения. Положим

$$m_y(x) = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x) \quad и \quad D_y(x) = \sigma_y^2(x) = \sigma_y^2\left(1-r_{xy}^2\right). \quad (3.36)$$

Тогда можно записать:

$$f(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma_y(x)}\exp\{-\frac{1}{2\sigma_y^2(x)}[y-m_y(x)]^2\}. \quad (3.37)$$

Из этого следует, что $m_y(x)$ и $D_y(x)$ из (3.36) есть соответственно условное математическое ожидание и условная дисперсия случайной величины Y при фиксированном значении величины X, что не противоречит введённым обозначениям.

Если коэффициент корреляции r_{xy} будет равен нулю, то окажется, что $m_y(x)=m_y$ и $D_y(x)=D_y$. В таком случае условная плотность $f(y \mid x)$ совпадёт с безусловной плотностью $f(y)$ и, следовательно, случайные величины X и Y окажутся статистически независимыми. При этом, естественно, выполнится равенство $f(x,y)=f_X(x)f_Y(y)$, т.е.

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{1}{2}\left[\frac{(x-m_x)^2}{\sigma_x^2}+\frac{(y-m_y)^2}{\sigma_y^2}\right]\right\}. \quad (3.38)$$

Таким образом, вопрос о том, являются ли нормально распределённые величины зависимыми или независимыми, разрешается значением всего одного параметра. Если $r_{xy}=0$, то величины X и Y являются статистически независимыми. Если, напротив, $r_{xy}\neq 0$, то эти случайные величины статистически зависимы.

Всё выше изложенное приводит к выводу, что если X и Y распределены нормально, то регрессия одной величины на другую всегда будет линейной:

$$m_y(x) = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x) \;\; и \;\; m_x(y) = m_x + r_{xy}\frac{\sigma_x}{\sigma_y}(y-m_y). \quad (3.39)$$

Линии регрессии являются прямыми линиями, пересекающимися в точке $(m_x; m_y)$, графики линий симметричны относительно биссектрисы первого и третьего координатных углов. Другой связи между нормально распределёнными величинами быть не может.

Для прогнозирования ожидаемых значений величины Y по заданным значениям X следует воспользоваться регрессией (3.39):

$$y = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x)\,. \tag{3.40}$$

При этом остаточная дисперсия $D_{ост}$ совпадает с условной дисперсией $D_y(x)$:

$$D_{ост} = D_y(x) = \sigma_y^2(x) = \sigma_y^2(1 - r_{xy}^2) \tag{3.41}$$

Действительно, $D_{ост} = M\{[Y - m_y(X)]^2 = M\{[(Y - m_y) - r_{xy}\frac{\sigma_y}{\sigma_x}(X - m_x)]^2\} =$

$$= M\{[Y - m_y]^2\} - 2r_{xy}\frac{\sigma_y}{\sigma_x}M\{(X - m_x)(Y - m_y)\} +$$

$$+ r_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}M\{[X - m_x]^2\} = \sigma_y^2 - 2r_{xy}\frac{\sigma_y}{\sigma_x}K_{xy} + r_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2.$$

Подставляя сюда $K_{xy}=r_{xy}\sigma_x\sigma_y$, в соответствии с (3.20), получим равенство (3.41). Таким образом, корреляционное отношение в данном случае совпадает с квадратом коэффициента корреляции: $\eta_{y/x}^2 = 1 - \frac{D_{ост}}{D_y} = 1 - \frac{\sigma_y^2(1 - r_{xy}^2)}{\sigma_y^2} = 1 - (1 - r_{xy}^2) = r_{xy}^2$. Это ещё раз подтверждает, что между величинами X и Y может существовать только линейная связь.

Пусть коэффициент корреляции r_{xy} равен нулю. Тогда двухмерная плотность распределения вероятностей представляется выражением (3.38). Уравнение

$$\frac{(x - m_x)^2}{\sigma_x^2} + \frac{(y - m_y)^2}{\sigma_y^2} = k^2, \tag{3.42}$$

определяет эллипс на координатной плоскости XOY с центром в точке $(m_x; m_y)$ и с осями, параллельными осям координат. Поделим это уравнение на k^2 (k>0):

$$\frac{(x - m_x)^2}{(k\sigma_x)^2} + \frac{(y - m_y)^2}{(k\sigma_y)^2} = 1.$$

Таким образом, полуоси эллипса пропорциональны средним квадратическим отклонениям и равны соответственно $k\sigma_x$ и $k\sigma_y$. Придавая константе k различные значения, получаем эллипсы различного размера. Такие эллипсы называются *эллипсами рассеяния*. На линии эллипса рассеяния плотность распределения вероятностей $f(x,y)$ является постоянной, т.е. эллипсы рассеяния представляют собой линии равного уровня функции $f(x,y)$. Если $\sigma_x=\sigma_y$, то рассеяние называют круговым. Эллипс, полученный при k=1, называют *единичным эллипсом рассеяния*.

Найдём вероятность попадания точки $(X;Y)$ внутрь эллипса рассеяния. Обозначим эту область через D. Тогда

$$P\{(X;Y)\in D\} = \iint_D f(x,y)dxdy = \frac{1}{2\pi\sigma_x\sigma_y}\iint_D \exp\{-\frac{(x - m_x)^2}{2\sigma_x^2} - \frac{(y - m_y)^2}{2\sigma_y^2}\}dxdy$$

Перейдём к центрированным и нормированным случайным величинам, делая замену

$$\frac{x - m_x}{\sigma_x} = u, \quad \frac{y - m_y}{\sigma_y} = v, \quad \text{тогда} \quad \frac{dx}{\sigma_x} = du, \quad \frac{dy}{\sigma_y} = dv.$$

Уравнение (3.42) на координатной плоскости UOV окажется уравнением окружности: $u^2+v^2=k^2$. Тогда $P\{(X;Y)\in D\} = \frac{1}{2\pi}\iint_D \exp\{-\frac{u^2 + v^2}{2}\}dudv$, причём область D является

кругом радиуса k. Интегрирование удобно провести в полярной системе координат. Пусть $u = \rho\cos\theta$, $v = \rho\sin\theta$. Тогда элемент площади $dxdy$ изменится на $\rho d\rho d\theta$. Получим:

$$P\{(X;Y)\in D\} = \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{k} \exp\{-\frac{\rho^2}{2}\}\rho d\rho = -\frac{1}{2\pi}\cdot 2\pi\cdot\exp\{-\frac{\rho^2}{2}\}\Big|_0^k .$$

Окончательно,

$$\mathrm{P}\{(\mathrm{X};\mathrm{Y})\in \mathrm{D}\} = 1-\exp\{-\frac{\mathrm{k}^2}{2}\}. \qquad (3.43)$$

Таким образом, вероятность попадания точки $(X;Y)$ в эллипс рассеяния не зависит от дисперсий. Вероятность попадания точки $(X;Y)$ в единичный эллипс равна приблизительно 0,39. При k=3 вероятность будет равна примерно 0,99. Это означает, что почти всегда точка $(X;Y)$ попадёт в эллипс рассеяния с полуосями $3\sigma_x$ и $3\sigma_y$.

Также легко определяется вероятность попадания двухмерной случайной величины $(X;Y)$ в прямоугольную область со сторонами, параллельными осям координат, если только X и Y статистически независимы. Если область D определяется неравенствами: $a\le x<b$, $c\le y<d$, то

$$P_D = P\{(X;Y)\in D\} = \iint_D f(x,y)dxdy = \frac{1}{2\pi\sigma_x\sigma_y}\int_a^b dx\int_c^d \exp\{-\frac{(x-m_x)^2}{2\sigma_x^2}-\frac{(y-m_y)^2}{2\sigma_y^2}\}dy =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_x}\cdot\int_a^b \exp\{-\frac{(x-m_x)^2}{2\sigma_x^2}\}dx\cdot\frac{1}{\sqrt{2\pi}\sigma_y}\cdot\int_c^d \exp\{-\frac{(y-m_y)^2}{2\sigma_y^2}\}dy.$$

Применяя формулу (2.41), получим:

$$P_D = [F_0(\frac{b-m_x}{\sigma_x}) - F_0(\frac{a-m_x}{\sigma_x}][F_0(\frac{d-m_y}{\sigma_y}) - F_0(\frac{c-m_y}{\sigma_y})], \qquad (3.44)$$

или, в соответствии с (2.43),

$$P_D = [\Phi_0(\frac{b-m_x}{\sigma_x}) - \Phi_0(\frac{a-m_x}{\sigma_x}][\Phi_0(\frac{d-m_y}{\sigma_y}) - \Phi_0(\frac{c-m_y}{\sigma_y})]. \qquad (3.45)$$

Функция $\Phi_0(z)$ – функция Лапласа (1.43), а $F_0(z)$ – функция распределения вероятностей нормированной и центрированной нормально распределённой случайной величины.

Если область D отличается от рассмотренных, то задача по аналитическому определению вероятности попадания в эту область может оказаться достаточно сложной. То же самое можно сказать и о том случае, когда величины X и Y оказываются статистически зависимыми. Заметим, однако, что простым поворотом системы координат можно перейти от зависимых нормально распределённых случайных величин к независимым. Кроме того, современная вычислительная техника позволяет вычислять практически любые интегралы. В таком случае можно пользоваться непосредственно формулой (3.7).

Пример 3.7. Двухмерная случайная величина (X,Y) имеет нормальное распределение с параметрами m_x=2, σ_x=4, m_y=3, σ_y=2, r_{xy}=-0,8. Записать уравнение линии регрессии Y на X и вычислить остаточную дисперсию.

Воспользуемся готовой формулой (3.40): $y = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x)$. Подставим сюда заданные параметры закона распределения: $y = 3-0{,}8\cdot\frac{2}{4}(x-2)$, или $y = -0{,}4x+3{,}8$.

Если пользоваться этим уравнением для прогнозирования ожидаемых значений величины Y по известным значениям величины X, то средний квадрат ошибки прогнозирования будет равен остаточной дисперсии: $D_{ост} = \sigma_y^2\left(1 - r_{xy}^2\right) = 2^2(1 - 0{,}8^2) = 1{,}44$.

Пример 3.8. Определить вероятность попадания нормально распределённой двухмерной величины (X,Y) в эллипс рассеяния с полуосями $2\sigma_x$, $2\sigma_y$, если известно, что величины X и Y являются статистически независимыми.

Так как случайные величины статистически независимы, то коэффициент корреляции равен нулю. Таким образом, достаточно применить формулу (3.43) при k=2. Искомая вероятность P будет равна: $P = 1 - \exp\{-\frac{k^2}{2}\}\Big|_{k=2} = 1 - \exp\{-\frac{2^2}{2}\} = 1 - e^{-2} \approx 0{,}865$.

Термины для запоминания

- Условный закон распределения вероятностей, статистическая зависимость, ковариационный момент, ковариация, корреляция, коэффициент корреляции, ковариационная матрица, условное математическое ожидание, условная дисперсия, регрессия, линии регрессии, остаточная дисперсия, корреляционное отношение.

Список контрольных вопросов

1. Дайте определение многомерной случайной величины.
2. Что называется законом распределения вероятностей двухмерной случайной величины?
3. Как можно задать двухмерную дискретную случайную величину?
4. Дайте определение функции распределения вероятностей двухмерной величины.
5. Перечислите основные свойства двухмерной функции распределения вероятностей и объясните их смысл.
6. Как получить двухмерную плотность распределения вероятностей, зная функцию распределения вероятностей двухмерной величины?
7. Перечислите основные свойства двухмерной плотности распределения вероятностей и объясните их смысл.
8. Как определить вероятность попадания двухмерной случайной величины в заданную область?
9. Дайте определение условной плотности распределения вероятностей.
10. В каком случае две случайные величины называются статистически независимыми?
11. Дайте определение начальных моментов двухмерной случайной величины.
12. Дайте определение центральных моментов двухмерной случайной величины.

13. Что характеризуют смешанные моменты двухмерной случайной величины?
14. Что такое корреляционный момент?
15. Что такое коэффициент корреляции и как он определяется?
16. Дайте определение условного математического ожидания.
17. Что такое регрессия и линия регрессии?

Вопрос для обсуждения

В каких задачах желательно знать смешанные моменты случайных величин?

ТЕМА 4.

Функциональные преобразования случайных величин

Студент должен
освоить:

- правила функциональных преобразований случайных величин и специальных законов распределения;

приобрести навыки:

- определения закона распределения функции случайных величин и её числовых характеристик.

Краткое содержание

Функция одной случайной величины. Функция нескольких случайных величин. Теоремы о математических ожиданиях и дисперсиях функций от случайных величин. Некоторые специальные законы распределения, применяемые в математической статистике.

4.1. Функция одной случайной величины

Пусть задано некоторое однозначное функциональное преобразование случайной величины X: $Y = \varphi(X)$. Величина Y также будет случайной величиной, а её возможные значения полностью определяются возможными значениями величины X.

Закон распределения вероятностей величины Y также полностью определяется законом распределения вероятностей величины X. Если X дискретная случайная величина с возможными значениями $x_1, x_2, \ldots, x_n$ и вероятностями этих значений, соответственно, $p_1, p_2, \ldots, p_n$, то возможными значениями величины Y будут $y_1, y_2, \ldots, y_n$, определяемые преобразованием: $y_k = \varphi(x_k)$, $k = 1, 2, \ldots, n$. Таким образом, $P\{Y = y_k\} = P\{X = x_k\} = p_k$ для всех значений k. Это и есть закон распределения вероятностей для величины Y. Это относится к случаю, когда функция $Y = \varphi(X)$ является взаимно однозначной.

Рассмотрим непрерывную случайную величину X с плотностью распределения вероятностей $f_X(x)$. Если функция $Y = \varphi(X)$ является монотонной функцией, то

$$f_Y(y) = f_X(x(y)) \,|\, x'(y) \,|\,. \tag{4.1}$$

Если функции $Y = \varphi(X)$ не является монотонной, то вся область определения этой функции разбивается на интервалы монотонности. Пусть n – число таких интервалов. Выражение обратной функции по отношению к данной на k–м интервале обозначим через $x_k(y)$. Тогда

$$f_Y(y) = \sum_{k=1}^{n} f_X(x_k(y)) \,|\, x_k'(y) \,|\, \tag{4.2}$$

Пример 4.1. Требуется найти распределение случайной величины $Y = X^2$, если величина X имеет плотность распределения вероятностей $f_X(x)$. Рассмотреть случай, когда величина X распределена нормально с математическим ожиданием $m_x = 0$ и дисперсией σ^2.

Функция $y = x^2$ имеет два интервала монотонности: $(-\infty; 0)$ и $(0, \infty)$. На интервале монотонного убывания функции $x = x_1(y) = -\sqrt{y}$, а на интервале монотонного возрастания функции $x = x_2(y) = \sqrt{y}$. Следовательно, $x_1'(y) = -\dfrac{1}{2\sqrt{y}}$, а $x_2'(y) = \dfrac{1}{2\sqrt{y}}$. Применим формулу (4.2): $f_Y(y) = f_X(x_2(y)) \cdot |\, x_2'(y) \,| + f_X(x_1(y)) \cdot |\, x_1'(y) \,|$. Подставляя сюда функции $x_1(y)$ и $x_2(y)$, получим окончательно: $f_Y(y) = \dfrac{1}{2\sqrt{y}}[f_X(-\sqrt{y}) + f_X(\sqrt{y})], \quad y > 0$. Если X имеет нормальное распределение, то $f_X(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \cdot \exp\{-\dfrac{x^2}{2\sigma^2}\}$, причём $f_X(-\sqrt{y}) = f_X(\sqrt{y})$. Следовательно, $f_Y(y) = \dfrac{1}{\sqrt{2\pi y}\sigma} \cdot \exp\{-\dfrac{y}{2\sigma^2}\}, \quad y > 0$. Если $y < 0$, то $f_Y(y) = 0$.

Рассмотрим более подробно линейное преобразование случайной величины X. Пусть $Y = \alpha X + \beta$, $\alpha \neq 0$. Тогда $x(y) = \frac{1}{\alpha}(y - \beta)$, $x'(y) = \frac{1}{\alpha}$ и $|x'(y)| = \frac{1}{|\alpha|}$. Подставляя это в формулу (4.3), получим: $f_Y(y) = \frac{1}{|\alpha|} \cdot f_X(\frac{1}{\alpha}(y - \beta))$.

Линейное преобразование не изменяет типа закона распределения вероятностей. Оно приводит лишь к изменению масштаба графика плотности распределения и к переносу начала координат в точку $y_0 = \beta$. Это свойство часто используется при решении задач.

4.2. Функция нескольких случайных величин

Рассмотрим случайную величину Z, которая является функцией двух случайных величин X и Y: $Z = \varphi(X, Y)$. Будем считать, что любой паре значений (x, y) двухмерной величины (X, Y) соответствует единственное значение z величины Z. Последнее означает, что функция $z = \varphi(x, y)$ является однозначной. По определению, $F_Z(z) = P\{Z < z\}$, так что $F_Z(z)$ представляет собой вероятность попадания точки $(X; Y)$ в ту область значений (x, y), для которой выполняется неравенство $\varphi(x, y) < z$. Обозначим эту область через D_z. Тогда

$$F_Z(z) = \iint_{D_Z} f(x, y)dxdy. \qquad (4.3)$$

Здесь D_z зависит от z, а $f(x, y)$ – плотность распределения вероятностей двухмерной случайной величины (X, Y).

Закон распределения суммы случайных величин. Пусть $Z = X + Y$. Тогда область интегрирования в (4.3) представляет собой полуплоскость $x + y < z$ (рис. 4.1):

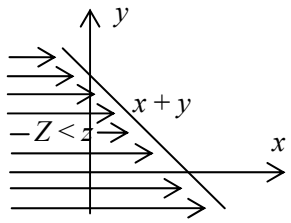

Рис. 4.1. **Иллюстрация к выводу формулы свёртки**

$$F_Z(z) = \iint_{x+y<z} f(x, y)dxdy = \int_{-\infty}^{\infty} dy \int_{-\infty}^{z-y} f(x, y)dx.$$

Порядок интегрирования показан стрелками. Заменим переменную: $x = t - y$, $dx = dt$. Тогда

$$F_Z(z) = \int_{-\infty}^{\infty} dy \int_{-\infty}^{z} f(t - y, y)dt = \int_{-\infty}^{z} [\int_{-\infty}^{\infty} f(t - y, y)dy]dt.$$

Найдём плотность распределения вероятностей $f_Y(y)$. Производная от интеграла по верхнему пределу интегрирования равна подынтегральной функции, вычисленной при значении переменной интегрирования, равной её верхнему пределу, т. е.

$$f_Z(z) = \int_{-\infty}^{\infty} f(z - y, y)dy. \qquad (4.4)$$

Формула (4.4) справедлива для любых непрерывных величин X и Y. Если они статистически независимы, то двухмерная плотность распределения равна произведению одномерных плотностей. В этом случае

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)dy. \qquad (4.5)$$

Формулы типа (4.5) называются *формулами свёртки* двух функций. Плотность распределения вероятностей суммы двух случайных величин равна свёртке плотностей рас-

пределения. Закон распределения вероятностей суммы двух случайных величин называется композицией распределений этих величин. Применяя многократно формулу свёртки, можно получить закон распределения вероятностей суммы большего числа статистически независимых случайных величин.

4.3. Теоремы о математических ожиданиях и дисперсиях

После рассмотрения многомерных случайных величин и функций от случайных величин появилась возможность рассмотреть свойства математических ожиданий и дисперсий более подробно. Сформулируем эти свойства в виде теорем.

Теорема 1. *Математическое ожидание суммы случайных величин равно сумме их математических ожиданий:*

$$M\{X + Y\} = M\{X\} + M\{Y\}. \tag{4.6}$$

Если X, Y – непрерывные случайные величины с плотностью распределения $f(x, y)$, то $M\{X+Y\}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy=\int_{-\infty}^{\infty}x[\int_{-\infty}^{\infty}f(x,y)dy]dx+\int_{-\infty}^{\infty}y[\int_{-\infty}^{\infty}f(x,y)dx]dy$. В первом слагаемом внутренний интеграл представляет собой одномерную плотность распределения $f_X(x)$, а во втором слагаемом – одномерную плотность распределения $f_Y(y)$. Таким образом,

$$M\{X+Y\} = \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy = M\{X\} + M\{Y\}.$$

Аналогично доказывается теорема для дискретных случайных величин. Теорема легко обобщается на случай произвольного числа слагаемых:

$$M\left\{\sum_{i=1}^{n} X_i\right\} = \sum_{i=1}^{n} M\{X_i\}. \tag{4.7}$$

Следствие: математическое ожидание линейной функции от случайных величин равно линейной функции от их математических ожиданий, т.е.

$$M\left\{\sum_{i=1}^{n} a_i X_i + b\right\} = \sum_{i=1}^{n} a_i M\{X_i\} + b. \tag{4.8}$$

Теорема 2. *Математическое ожидание произведения двух случайных величин равно произведению их математических ожиданий плюс их ковариационный момент, т.е.*

$$M\{XY\} = M\{X\}\cdot M\{Y\} + K_{xy}. \tag{4.9}$$

В п. 3.5 было показано (3.23), что $K_{xy} = M\{XY\} - m_x m_y$. Это и доказывает (4.9).

Следствие: если две величины являются некоррелированными, то математическое ожидание произведения этих величин равно произведению их математических ожиданий. Это справедливо и для несколько случайных величин, если они статистически независимы.

Пример 4.2. Вычислить математическое ожидание функции $Z = 2X^2 + 3X - 1$, если математическое ожидание случайной величины X равно $m_x = 2$, а её дисперсия равна $D_x = 5$.

По свойствам математического ожидания $M\{Z\} = M\{2X^2 + 3X - 1\} = 2M\{X^2\} + 3M\{X\} - 1$.

Но $M\{X^2\} = D_x + m_x^2$, следовательно, $M\{Z\} = 2D_x + 2m_x^2 + 3m_x - 1 = 2 \cdot 5 + 2 \cdot 4 + 3 \cdot 2 - 1 = 23$.

Теорема 3. *Дисперсия суммы или разности случайных величин определяется формулой:*

$$D\{X \pm Y\} = D\{X\} + D\{Y\} \pm 2K_{xy}, \tag{4.10}$$

где K_{xy} – ковариационный момент величин X и Y.

Действительно, рассмотрим, например, сумму случайных величин X + Y:

$$D\{X + Y\} = M\{[X + Y - M\{X + Y\}]^2\} = M\{[X - M\{X\} + Y - M\{Y\}]^2\} =$$
$$= M\{[X - M\{X\}]^2\} + M\{[Y - M\{Y\}]^2\} + 2M\{(X - M\{X\})(Y - M\{Y\})\}.$$

Аналогично равенство (4.10) доказывается и для разности случайных величин.

Теорема 3 обобщается на случай n случайных величин. Если $Z = X_1 + X_2 + \ldots + X_n$, то

$$D\{Z\} = D\{\sum_{k=1}^{n} X_k\} = \sum_{k=1}^{n} D\{X_k\} + 2\sum_{k=1}^{n}\sum_{i\neq k} K_{ik}\ , \qquad (4.11)$$

где K_{ik} – ковариационный момент между i-й и k-й случайными величинами.

Следствие: если случайные величины являются некоррелированными, то дисперсия суммы случайных величин равна сумме их дисперсий:

$$\mathrm{D\{Z\} = D\{\sum_{k=1}^{n} X_k\} = \sum_{k=1}^{n} D\{X_k\}}\ . \qquad (4.12)$$

Теорема 4. *Дисперсия произведения двух статистически независимых случайных величин определяется следующей формулой*:

$$D\{XY\} = D\{X\}D\{Y\} + m_x^2 D\{Y\} + m_y^2 D\{X\}\ . \qquad (4.13)$$

Действительно, для независимых случайных величин справедливо равенство $M\{XY) = M\{X\}M\{Y\} = m_x m_y$. Далее, $M\{X^2Y^2\} = M\{X^2\}M\{Y^2\} = [D\{x\} + m_x^2]\cdot[D\{Y\} + m_y^2]$. Таким образом,

$$D\{XY\} = M\{[XY - M\{XY\}]^2\} = M\{[XY - m_x m_y]^2\} =$$
$$= M\{[XY]^2\} - 2m_x m_y M\{XY\} + m_x^2 m_y^2 = M\{X^2\}M\{Y^2\} - m_x^2 m_y^2 =$$
$$= [D\{X\} + m_x^2]\cdot[D\{Y\} + m_y^2] - m_x^2 m_y^2 = D\{X\}D\{Y\} + m_x^2 D\{Y\} + m_y^2 D\{X\}.$$

Дисперсия произведения двух случайных величин равна произведению их дисперсий только для статистически независимых и центрированных ($m_x = m_y = 0$) случайных величин.

Пример 4.3. Производятся заготовки прямоугольной формы. Пусть X – ширина заготовки, а Y – её длина. Технология производства обеспечивает следующие параметры: $m_x = 100$, $m_y = 200$, $D_x = 4$ и $D_y = 9$. Вычислить математическое ожидание и дисперсию площади заготовки при условии, что X и Y являются статистически независимыми величинами.

Площадь заготовки $S = XY$. Так как величины X и Y являются статистически независимыми, то $M\{S\} = M\{XY\} = M\{X\}M\{Y\} = m_x m_y = 100\cdot200 = 20000$. По теореме 4

$$D\{Z\} = D\{XY\} = D_x D_y + m_x^2 D_y + m_y^2 D_x = 4\cdot9 + 100^2\cdot9 + +200^2\cdot4 = 250036.$$

Среднее квадратическое отклонение σ_s будет равно: $\sigma_s = \sqrt{D_s} = \sqrt{250036} \approx 500{,}04$.

Применение теорем о математических ожиданиях и дисперсиях позволяет во многих случаях определять числовые характеристики функций случайных величин, не определяя предварительно закон распределения вероятностей для этих функций.

4.4. Некоторые специальные законы распределения

Рассмотрим некоторые законы распределения вероятностей, которые применяются в математической статистике и которые получены путём функциональных преобразований нормально распределённых случайных величин.

Распределение χ^2. Пусть статистически независимые случайные величины X_1, X_2, …, X_n распределены по нормальному закону с математическим ожиданием m_x и дисперсией σ_x^2. Пусть U сумма квадратов центрированных и нормированных случайных величин:

$$U = \sum_{i=1}^{n} \frac{(X_i - m_x)^2}{\sigma_x^2}. \tag{4.14}$$

Пользуясь правилами преобразования случайных величин, можно получить плотность распределения вероятностей величины U в следующем виде:

$$f(u) = \frac{1}{2^{n/2}\Gamma(n/2)} \cdot u^{n/2-1} \cdot \exp\{-\frac{u}{2}\}, \quad u > 0, \tag{4.15}$$

где $\Gamma(z)$ – гамма-функция. Этот закон распределения вероятностей называется распределением χ^2 с n степенями свободы. Распределение зависит только от одного параметра – от числа степеней свободы n. Это число представляет собой число статистически независимых величин, представленных в сумме (4.14). При неограниченном увеличении n распределение (4.15) неограниченно приближается к нормальному распределению. Математическое ожидание случайной величины U равно числу степеней свободы n. Отметим одно важное свойство. Сумма двух величин, распределённых по закону χ^2 с n_1 и с n_2 степенями свободы, также распределена по закону χ^2 с $n = n_1 + n_2$ степенями свободы.

Распределение Стьюдента. Пусть случайная величина X распределена нормально с математическим ожиданием m_x и дисперсией σ_x^2, а величина U распределена по закону χ^2 с n степенями свободы. Тогда величина

$$T = \frac{X - m_x}{\sigma_x \sqrt{U/n}} \tag{4.16}$$

будет распределена по закону Стьюдента с n степенями свободы:

$$f(t) = \frac{1}{\sqrt{\pi n}} \cdot \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} (1 + \frac{t^2}{n})^{-\frac{n+1}{2}}, \quad -\infty < t < \infty. \tag{4.17}$$

Распределение Стьюдента зависит от одного параметра и также приближается к нормальному распределению при неограниченном увеличении n. Этот закон является симметричным относительно начала координат. График плотности распределения (4.17) напоминает график плотности нормального распределения.

Распределение Фишера – Снедекора (F-распределение). Пусть случайная величина U распределена по закону χ^2 с k степенями свободы, а величина V – по закону χ^2 с n степенями свободы. Тогда величина $Z = \frac{U/k}{V/n} = \frac{n}{k} \cdot \frac{U}{V}$ будет распределена по закону Фишера – Снедекора с k и n степенями свободы:

$$f(z) = \left(\frac{k}{n}\right)^{\frac{k}{2}} \frac{\Gamma((k+n)/2)}{\Gamma(k/2)\Gamma(n/2)} \cdot z^{\frac{k}{2}-1} (1 + \frac{k}{n} z)^{-\frac{k+n}{2}}, \quad z > 0. \tag{4.18}$$

Этот закон распределения вероятностей зависит от двух степеней свободы k и n.

Термины для запоминания

- Функциональное преобразование; композиция распределений; распределение χ^2 (распределение «хи-квадрат»); распределение Стьюдента; распределение Фишера – Снедекора; F-распределение.

Список контрольных вопросов

1. Как определить закон распределения функция дискретной случайной величины?
2. Каким будет закон распределения функции от случайной величины при линейном преобразовании.
3. Если нормально распределённая случайная величина преобразуется линейно, то какой закон распределения будет иметь функция?
4. Что называется композицией распределений двух случайных величин?
5. Сформулируйте первую теорему о математических ожиданиях случайных величин.
6. Сформулируйте вторую теорему о математических ожиданиях случайных величин.
7. В каком случае математическое ожидание произведения двух случайных величин равно произведению их математических ожиданий?
8. Сформулируйте первую теорему о дисперсиях случайных величин.
9. В каком случае дисперсия суммы двух случайных величин равна сумме их дисперсий?
10. Сформулируйте вторую теорему о дисперсиях случайных величин.
11. В каком случае дисперсия произведения двух случайных величин равна произведению их дисперсий?
12. Перечислите специальные распределения, полученные путём функциональных преобразований нормально распределённых случайных величин.

Вопрос для обуждения

Почему нормальное распределение имеет такое широкое применение? Объясните это, используя материал, рассмотренный в теме «Функциональные преобразования случайных величин».

ТЕМА 5.

Закон больших чисел. Центральная предельная теорема

Студент должен
освоить:

- законы больших чисел в форме Чебышева и центральной предельной теоремы;

приобрести навыки:

- применения закона больших чисел для оценки вероятности попадания случайной величины в симметричные (относительно математического ожидания) интервалы.

Краткое содержание

Закон больших чисел в форме Чебышева. Неравенство Чебышева. Теорема Чебышева. Теорема Бернулли. Центральная предельная теорема.

5.1. Предварительные замечания

Закон больших чисел говорит о свойствах сумм случайных величин при неограниченном увеличении числа слагаемых. Пусть Z – среднее арифметическое n случайных величин

$$Z = \frac{1}{n}\sum_{k=1}^{n} X_k .$$

Закон больших чисел в форме Чебышева – это совокупность теорем, которые позволяют оценить вероятность попадания величины Z в интервал, симметричный относительно математического ожидания этой суммы. Такие суммы появляются при многократных измерениях случайных величин.

Пусть требуется оценить математическое ожидание случайной величины X, не зная закона распределения вероятностей. Сделаем n измерений этой величины и в качестве оценки m_x^* математического ожидания возьмём среднее арифметическое полученных измерений: $m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$. Здесь x_i, i = 1, 2, …, n, – измеренные значения X. Насколько близко значение m_x^* к истинному значению математического ожидания m_x? Сколько нужно сделать измерений, чтобы можно было оценить m_x с удовлетворительной точностью? Можно ли, хотя бы приближённо, судить о законе распределения вероятностей такой суммы? На некоторые из таких вопросов можно ответить, применяя выводы закона больших чисел и центральной предельной теоремы.

Для формулировки теорем закона больших чисел введём понятие сходимости по вероятности. Пусть Z_1, Z_2, …, Z_n – последовательность случайных величин. Говорят, что эта последовательность сходится по вероятности к числу m_z, если для любого как угодно малого положительного числа ε выполняется равенство:

$$\lim_{n\to\infty} P\{|\, Z_n - m_z \,| < \varepsilon\} = 1 . \qquad (5.1)$$

Особенность сходимости по вероятности состоит в том, что нельзя полностью отвергать событие $|Z_n - m_z| \geq \varepsilon$ при $n\to\infty$. Здесь уместно вспомнить, что вероятность невозможного события равна 0, но обратное утверждение не всегда является верным, т.е. событие, имеющее нулевую вероятность, может, тем не менее, произойти.

Свойства сумм случайных величин рассматривает также *центральная предельная теорема*. При определённых условиях закон распределения суммы случайных величин при неограниченном увеличении числа слагаемых неограниченно приближается к нормальному закону. Центральная предельная теорема указывает на эти условия.

5.2. Неравенство Чебышева

Пусть Z – произвольная случайная величина, имеющая математическое ожидание m_z и дисперсию D_z. *Неравенство Чебышева* утверждает, что вероятность отклонения слу-

чайной величины от своего математического ожидания на величину, не меньшую чем ε, ограничена сверху величиной D_z/ε^2, т.е.

$$P\{|\,Z - m_z\,| \geq \varepsilon\} \leq \frac{D_z}{\varepsilon^2} \tag{5.2}$$

Докажем это неравенство для случая непрерывной случайной величины. Вероятность попадания случайной величины Z в интервал, указанный в (5.2), равна интегралу от плотности распределения вероятностей по области значений z, в которой $|\,z - m_z\,| \geq \varepsilon$:

$$P\{|\,Z - m_z\,| \geq \varepsilon\} = \int\limits_{|z-m_z|\geq\varepsilon} f_z(z)dz \leq \int\limits_{|z-m_z|\geq\varepsilon} \frac{(z-m_z)^2}{\varepsilon^2} f_z(z)dz\,.$$

Последнее неравенство записано на том основании, что в области интегрирования отношение $(z - m_z)^2/\varepsilon^2 \geq 1$ для всех значений z. Так как подынтегральная функция не отрицательна, то расширение области интегрирования может привести только к увеличению этого интеграла. Следовательно, $P\{|Z - m_z| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}\int\limits_{-\infty}^{\infty}(z - m_z)^2 f_z(z)dz = \frac{D_z}{\varepsilon^2}$.

Неравенство Чебышева можно записать в другой форме:

$$P\{|\,Z - m_z\,| < \varepsilon\} > 1 - \frac{D_z}{\varepsilon^2}. \tag{5.3}$$

Неравенство Чебышева даёт достаточно грубую оценку вероятности. Иногда такие оценки становятся просто бесполезными. Например, пользуясь неравенством (5.3), можно иногда прийти лишь к выводу, что вероятность попадания в данный интервал не отрицательна, а неравенство (5.2) может привести к выводу, что вероятность не больше единицы. Такие оценки не несут никакой новой информации. Однако, в некоторых случаях оценки вероятности попадания случайной величины в интервал вполне приемлемы. Неравенством Чебышева пользуются тогда, когда другого подходящего способа оценки вероятности нет. Наибольшее значение неравенство Чебышева имеет в теоретическом плане, поскольку теоремы закона больших чисел базируются на неравенстве Чебышева.

Пример 5.1. С помощью неравенства Чебышева оценить вероятность того, что случайная величина X отклонится от своего математического ожидания m_x менее чем на $3\sigma_x$, где σ_x – её среднее квадратическое отклонение.

В данном случае $\varepsilon = 3\sigma_x$. Воспользуемся неравенством (5.3):

$$P\{|\,X - m_x\,| < 3\sigma_x\} > 1 - \frac{\sigma_x^2}{(3\sigma_x)^2} = 1 - \frac{1}{9} = \frac{8}{9} \approx 0{,}89.$$

Пример (5.1) показывает, что вероятность попадания любой случайной величины в интервал $(m_x - 3\sigma_x;\ m_x + 3\sigma_x)$ является величиной, большей, чем 8/9. Этот нижний предел достигается в самом худшем случае. Для конкретных случайных величин эта вероятность значительно выше. Так, например, для нормально распределенной случайной величиины эта вероятность равна приблизительно 0,997. В связи с вышесказанным, интервал $(m_x - 3\sigma_x;\ m_x + 3\sigma_x)$ принимают часто за интервал практически возможных значений случайной величины. Сам факт, состоящий в том, что для всех величин вероятность попадания в такой интервал достаточно велика, называют «*законом трёх сигм*».

Пример 5.2. Пусть $m_x = 2$, $D_x = 4$. С помощью неравенства Чебышева оценить вероятность того, что случайная величина примет значение из интервалов (0; 4) и (–1; 5).

Для 1-го интервала $\varepsilon = 2$. Тогда $P\{|X - 2| < 2\} > 1 - 4/2^2 = 0$. Таким образом, вероятность попадания случайной величины X в интервал (0; 4) больше нуля. Этот результат не несёт новой информации, поскольку любая вероятность не отрицательна. Для 2-го интервала $\varepsilon = 3$. Следовательно, $P\{|X - 2| < 3\} > 1 - 4/3^2 \approx 0{,}56$. Этот результат уже представляет интерес.

5.3. Теорема Чебышева

Теорема Чебышева. *Пусть случайные величины X_1, X_2, …, X_n попарно независимы и имеют дисперсии, ограниченные одной и той же постоянной, т.е. $D\{X_1\} \le C$, $D\{X_2\} \le C$, …, $D\{X_n\} \le C$. Тогда среднее арифметическое этих величин при $n \to \infty$ сходится по вероятности к среднему арифметическому их математических ожиданий*:

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} M\{X_k\}\right| < \varepsilon\right\} = 1. \qquad (5.4)$$

Рассмотрим случайную величину $Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Здесь слагаемые независимы. Поэтому $m_{zn} = M\{Z_n\} = \frac{1}{n}\sum_{k=1}^{n} M\{X_k\}$, $D_{zn} = D\{Z_n\} = \frac{1}{n^2}\sum_{k=1}^{n} D\{X_k\}$. В последней сумме каждая из дисперсий $D\{X_k\}$ меньше или равна C. Следовательно, всегда выполняется неравенство $D_{zn} \le C/n$. Таким образом, $P\{|Z_n - m_{zn}| < \varepsilon\} \ge 1 - \frac{D_{zn}}{\varepsilon^2} \ge 1 - \frac{C}{n\varepsilon^2}$. Переходя к пределу при $n \to \infty$, получим: $\lim_{n\to\infty} P\{|Z_n - m_{zn}| < \varepsilon\} \ge 1$. Из этого следует (5.4).

Для практики представляет интерес частный случай теоремы Чебышева, когда случайные величины X_k, $k = 1, 2, \ldots, n$, имеют одинаковые математические ожидания и одинаковые дисперсии, т.е. $M\{X_k\} = m_x$ и $D\{X_k\} = D_x$ для всех значений k. В этом случае

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - m_x\right| < \varepsilon\right\} = 1. \qquad (5.5)$$

Рассмотрим практическую интерпретацию равенства (5.5). Пусть величина X с математическим ожиданием m_x и дисперсией D_x измерена n раз. Измерения x_1, x_2, …, x_n можно также считать независимыми случайными величинами с теми же характеристиками m_x и D_x. Тогда случайная величина $m_x^* = \frac{1}{n}\sum_{k=1}^{n} x_k$ будет иметь характеристики:

$$M\{m_x^*\} = \frac{1}{n}\sum_{k=1}^{n} M\{x_k\} = \frac{1}{n}\sum_{k=1}^{n} m_x = m_x, \qquad D\{m_x^*\} = D\{\frac{1}{n}\sum_{k=1}^{n} x_k\} = \frac{1}{n^2}\sum_{k=1}^{n} D\{x_k\} = \frac{1}{n^2}\sum_{k=1}^{n} D_x = \frac{D_x}{n}.$$

Для величины m_x^* получим: $\lim_{n\to\infty} P\left\{\left|m_x^* - m_x\right| < \varepsilon\right\} = 1$, т.е. оценка m_x^* с характеристиками

$$M\{m_x^*\} = m_x \quad \text{и} \quad D\{m_x^*\} = \frac{D_x}{n} \qquad (5.6)$$

при $n \to \infty$ сходится по вероятности к истинному математическому ожиданию m_x случайной величины X. Таким образом, m_x^* можно использовать в качестве оценки неизвестного математического ожидания. Дисперсия оценки стремится к нулю с возрастанием n.

В практических задачах используются соответствующие неравенства Чебышева. Например, равенству (5.5) соответствует следующее неравенство Чебышева:

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k - m_x\right| < \varepsilon\right\} \geq 1 - \frac{D_x}{n\varepsilon^2}. \qquad (5.7)$$

Пример 5.4. Технологическое оборудование обеспечивает среднее квадратическое отклонение длины изготовляемой детали от математического ожидания этой длины не более чем 0,05 см. Были измерены 50 деталей. Какова вероятность того, что среднее арифметическое этих измерений отклонится от истинного математического ожидания не более чем на 0,02?

В данном случае ε = 0,02, дисперсия измеряемой величины $D_x = 0{,}05^2$, число измерений $n = 50$. По формуле (5.7) $P\left\{\left|\frac{1}{50}\sum_{k=1}^{20}X_k - m_x\right| < 0{,}02\right\} = 1 - \frac{0{,}05^2}{50 \cdot 0{,}02^2} = 1 - \frac{5}{40} = 0{,}875$.

5.4. Теорема Бернулли

Теорема Бернулли. *Относительная частота появления некоторого события в последовательности независимых испытаний стремится по вероятности к истинной вероятности этого события при неограниченном увеличении числа испытаний.*

Пусть проводится последовательно n независимых испытаний, в каждом из которых событие A может осуществиться с вероятностью p. Пусть в n испытаниях событие A осуществилось k раз. Тогда величина $p^* = k/n$ является частотой появления события в n испытаниях. Теорема Бернулли утверждает, что истинную неизвестную вероятность появления события A в одном испытании можно оценить с помощью p^*.

Пусть X_i, $i = 1, 2, \ldots, n$, – двоичная случайная величина:

$$X_i = \begin{cases} 1, & \text{если событие } A \text{ произошло в } i-\text{ом испытании,} \\ 0, & \text{если событие } A \text{ не произошло в } i-\text{ом испытании.} \end{cases}$$

Тогда $p^* = \frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n}X_i$ причём, $M\{p^*\} = p$ и $D\{p^*\} = \frac{pq}{n}$. (5.8)

Теорема Бернулли утверждает, что величина p^* при $n \to \infty$ стремится по вероятности к истинной вероятности p осуществления события A в одном испытании, т.е.

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i - p\right| < \varepsilon\right\} = 1 \quad \text{или} \quad \lim_{n\to\infty} P\left\{\left|\frac{k}{n} - p\right| < \varepsilon\right\} = 1. \qquad (5.9)$$

Таким образом, величина p^* может быть использована в качестве оценки истинной неизвестной вероятности p осуществления события A в одном испытании.

При конечных n теорема Бернулли применяется в форме неравенства Чебышева:

$$P\left\{\left|\frac{k}{n} - p\right| < \varepsilon\right\} \geq 1 - \frac{pq}{n\varepsilon^2}. \qquad (5.10)$$

Это неравенство зависит от вероятности p. Вспомним, что величина $pq = p(1 - p)$ имеет наибольшее значение при $p = 1/2$ (на интервале $0 < p < 1$). При этом $pq = 1/4$. Если вместо величины pq подставить её наибольшее значение, то неравенство может только усилиться:

$$P\left\{\left|\frac{k}{n}-p\right|<\varepsilon\right\}\geq 1-\frac{1}{4n\varepsilon^2}. \qquad (5.11)$$

Пример 5.5. Сколько нужно сделать выстрелов по мишени, чтобы с вероятностью не меньшей чем 0,9 можно было утверждать, что относительная частота попаданий отклонится от истинной вероятности попадания в мишень менее чем на 0,05?

Здесь ε = 0,05, задан нижний предел вероятности осуществления неравенства $|p^* - p| < \varepsilon$, а число испытаний n неизвестно. Воспользуемся формулой (5.11):

$$P\left\{\left|k/n-p\right|<\varepsilon\right\}\geq 1-1/(4n\varepsilon^2)=1-1/(4n\cdot 0{,}05^2)=1-100/n=0{,}9.$$

Отсюда, n = 1000.

5.5. Центральная предельная теорема

В *центральной предельной теореме* говорится об условиях, при которых закон распределения суммы случайных величин $Z_n=\sum_{k=1}^{n}X_k$ неограниченно приближается к нормальному закону при $n\to\infty$. Ограничимся формулировкой теоремы для статистически независимых, одинаково распределённых, случайных величин.

Теорема. Если $X_1, X_2, \ldots, X_n, \ldots$ – независимые случайные величины, имеющие один и тот же закон распределения с математическим ожиданием m_x и дисперсией σ_x^2, то при неограниченном увеличении n закон распределения суммы Z_n неограниченно приближается к нормальному закону.

Другие формулировки центральной предельной теоремы расширяют область её действия. Смысл условий, предъявляемых слагаемым, сводится к тому, чтобы каждая из суммируемых величин вносила относительно малую долю в общую сумму.

Практически центральная предельная теорема применяется при конечных значениях числа слагаемых величин. Если для величины Z_n выполняются условия центральной предельной теоремы, то говорят, что эта величина имеет *асимптотически нормальное распределение*. При выполнении условий теоремы можно считать, что

$$f_z(z)\approx\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\{-\frac{(z-m_z)^2}{2\sigma_z^2}\},\quad -\infty<z<\infty, \qquad (5.12)$$

где $m_z = nm_x$ и $\sigma_z^2 = n\sigma_x^2$. Равенство будет тем точнее, чем большее n. С помощью центральной предельной теоремы можно обосновать выбор нормального распределения для аппроксимации закона распределения исследуемой случайной величины. Достаточно показать, что эта величина образуется в результате аддитивного воздействия многих мелких факторов. Рассмотренные ранее предельные теоремы Муавра – Лапласа являются одним из вариантов центральной предельной теоремы.

Термины для запоминания

- Закон больших чисел в форме Чебышева. Неравенство Чебышева. Теорема Чебышева. Теорема Бернулли. Центральная предельная теорема.

Список контрольных вопросов

1. О чём говорит закон больших чисел?
2. Что называют законом больших чисел в форме Чебышева?
3. Сформулируйте неравенство Чебышева в двух его формах.
4. К каким случайным величинам можно применить неравенство Чебышева?
5. Сформулируйте теорему Чебышева. В каких условиях она применима?
6. Какой практический вывод делается на основе теоремы Чебышева по поводу оценки математического ожидания случайной величины?
7. Сформулируйте теорему Бернулли.
8. Какой практический вывод делается на основе теоремы Бернулли по поводу оценки неизвестной вероятности события?
9. Сформулируйте центральную предельную теорему.
10. Как практически используется центральная предельная теорема?

Вопрос для обсуждения

Почему нормальное распределенпе имеет такое широкое применение в окружающем нас мире? Объясните это, опираясь на законы больших чисел.

ТЕМА 6.

Статистическая обработка экспериментальных данных. Оценка параметров

Студент должен
освоить:

- понятия выборки, вариационного ряда, эмпирических законов распределения вероятностей, эмпирических числовых характеристик, точечных и интервальных оценок неизвестных параметров;

приобрести навыки:

- построения гистограммы, полигона, эмпирической функции распределения вероятностей, вычисления эмпирических числовых характеристик, построения доверительных инвервалов.

Краткое содержание

Выборка. Генеральная совокупность. Вариационный ряд. Эмпирические законы распределения вероятностей. Полигон и гистограмма. Точечные оценки параметров распределения. Свойства оценок. Понятие несмещенности, эффективности, состоятельности оценок. Методы нахождения оценок. Доверительные интервалы. Построение доверительных интервалов для параметров нормального распределения. Построение доверительного интервала для неизвестной вероятности события.

6.1. Задачи математической статистики

Любое понятие теории вероятностей есть отражение определённого физического процесса. Оно появляется в связи с требованием практики и представляет собой следствие некоторых экспериментов. Задачей математической статистики является обработка экспериментальных данных с целью изучения свойств случайных явлений, получения некоторой математической модели изучаемого объекта. Математическая модель объекта – это система математических соотношений, которые описывают с определённым уровнем приближения свойства величин, характеризующих объект, а также их возможные связи. Математическая модель не полностью описывает реальный объект, однако по мере изучения самой модели (вместо объекта) она может уточняться, всё полнее отражая свойства этого объекта. Если в объекте имеют место случайные явления, то его математическая модель обязательно будет включать в себя некоторые понятия теории вероятностей. Изучить определённые закономерности случайных явлений можно только с помощью многократных повторений опыта с объектом с последующей обработкой результатов.

Кроме основной задачи – обработки экспериментальных данных – математическая статистика рассматривает и смежные задачи, такие, как методы регистрации, способы описания экспериментальных данных, а также само планирование эксперимента. Прежде чем проводить эксперимент, необходимо чётко сформулировать и осознать цель этого эксперимента. Характер эксперимента должен максимально соответствовать его цели в том смысле, что результаты его будут достаточными для оценки тех или иных характеристик объекта.

Задачу по изучению свойств случайной величины с помощью экспериментальных данных можно поставить в разных вариантах. Во-первых, можно ограничиться лишь оценкой числовых характеристик величины X. В этом случае задача относится к группе задач *оценки неизвестных параметров*. В эту группу входят задачи выбора структуры оценок и задачи анализа качества этих оценок. Оценить можно не только параметры, относящиеся к одной величине (математическое ожидание, дисперсию и т.п.), но и параметры, связанные с двумя или более величинами (коэффициент корреляции, корреляционное отношение и т.п.).

Более общая задача – *установление закона распределения вероятностей* величины X. Пусть x_1, x_2, ..., x_n – n измерений этой величины. При отсутствии иной информации остаётся предполагать, что эти значения являются единственно возможными значениями данной случайной величины, а вероятность каждого из них равна $1/n$. Исходя из этого, можно построить функцию распределения вероятностей величины X как дискретной случайной величины: $F^*(x) = k/n$, где k – число измерений, лежащих на числовой оси левее x. Возникает вопрос, нельзя ли при конечном значении n провести сглаживание функции $F^*(x)$ так, чтобы получить функцию, близкую к истинной функции распределения? Этот вопрос рассматривается в группе задач, связанных с установлением закона распределения вероятностей по экспериментальным данным.

Следующая группа задач – это задачи *проверки статистических гипотез*. В качестве гипотез могут быть гипотезы о виде закона распределения вероятностей, гипотезы о предположительном значении тех или иных числовых характеристик, гипотезы о существовании или отсутствии статистической связи между величинами и т.д. Разработано достаточно много методов проверки статистических гипотез. Часть из них предполагает обработку данных постоянного объёма, другая часть – данных переменного объёма.

Если рассматриваются несколько случайных величин в совокупности, то одной из задач математической статистики является изучение и аппроксимация статистических связей между ними. Рассмотрим, например, две случайные величины X и Y. Можно поставить следующий вопрос: зависит ли случайная величина Y от случайной величины X? Ответом на этот вопрос будет либо «да», либо «нет». Задачу в такой постановке можно отнести к группе задач *дисперсионного анализа*. В задачах дисперсионного анализа рассматривают обычно некоторые факторы, которые могут влиять на исследуемую величину. Факторы могут быть как количественными, так и качественными. Например, зависит ли количество дорожно-транспортных происшествий (фактор B – количественный) от цвета автомобилей (фактор A – качественный). Факторов может быть и более двух. Задачи такого типа называют ещё задачами факторного анализа. Предположим теперь, что X и Y являются статистически зависимыми величинами. Тогда можно поставить следующую задачу: найти такую функцию $Y = \varphi(X)$, которая бы максимально отражала статистическую зависимость между величинами. С помощью такой функции можно было бы наиболее точно по заданным значениям X прогнозировать значения, которые будет принимать величина Y. Эта задача относится к группе задач *регрессионного анализа*. Такая задача может быть одномерной, многомерной, линейной или нелинейной.

6.2. Выборка. Вариационный ряд. Эмпирические законы распределения

Пусть проводятся n независимых испытаний над случайной величиной X при неизменном комплексе условий, от которых зависят конкретные реализации этой величины. В результате испытаний получены n измерений величины X: $X_n = \{x_1, x_2, \ldots, x_n\}$. Совокупность X_n таких измерений называют в*ыборкой* случайной величины X, сами измерения называют элементами выборки, а число измерений n – объёмом выборки. Элементы выборки можно считать независимыми одинаково распределёнными случайными величинами, поскольку они являются результатом проведения последовательности независимых испытаний с одной и той же случайной величиной X.

Предположим, что в выборке X_n представлены все возможные значения случайной величины X с частотами, пропорциональными истинным (неизвестным) вероятностям этих значений. Тогда выборку называют *генеральной совокупностью*. Генеральная совокупность полностью представляет закон распределения вероятностей случайной величины. Если X – дискретная случайная величина с конечным числом возможных значений, то, в принципе, можно получить генеральную совокупность при неограниченном увеличении числа измерений. По теореме Бернулли (см. п. 5.4) относительные частоты появления тех или иных возможных значений будут стремиться по вероятности к истинным вероятностям этих значений. Если X – непрерывная случайная величина, то ситуация усложняется, поскольку возможные значения случайной величины в совокупности представляют непрерывное множество. Из сказанного следует, что, во-первых, для любого конечного n нельзя утверждать, что X_n является генеральной совокупностью, во-вторых, чем больше n, тем ближе, в определённом смысле, X_n к генеральной совокупности.

Та выборка X_n, которая достаточно хорошо сохраняет пропорции генеральной совокупности, называется *представительной* (или *репрезентативной*). Это определение представительности выборки не позволяет делать конкретные выводы, так как не указана общая мера соответствия между представительной выборкой и генеральной совокупностью. Вопрос о представительности выборки нужно решать в конкретных частных задачах, опираясь на конкретные критерии соответствия.

Выборка X_n, элементы которой расположены в порядке возрастания, называется *простым вариационным рядом*. Разность R между наибольшим и наименьшим значениями измерений называют *широтой распределения* или *размахом варьирования*. Если одинаковые по значению элементы объединить в группы, то получается *сгруппированный вариационный ряд*, который представляется в виде таблицы:

x_i	x_1	x_2	…	x_r
n_i	n_1	n_2	…	n_r

В таблице x_i (i = 1, 2, ..., r) – различающиеся значения элементов; n_i – число элементов, имеющих значение x_i; r – число различных значений в выборке, причём $n_1 + n_2 + ... + n_r = n$.

По вариационному ряду строится *эмпирическая функция распределения вероятностей* исследуемой случайной величины. Эмпирическая функция распределения вероятностей $F^*(x)$ определяется как отношение числа $\alpha(x)$ элементов выборки, меньших, чем x, к общему числу элементов n: $F^*(x) = \alpha(x)/n$. Эта функция будет иметь ступенчатый график. Если все элементы выборки различны, то величина ступенек будет равна $1/n$. С ростом объёма выборки n величина ступенек уменьшается и стремится к нулю при $n\to\infty$. Для непрерывной величины при $n\to\infty$ эмпирическая функция $F^*(x)$ будет неограниченно приближаться к некоторой непрерывной функции $F(x)$. Эту сходимость следует понимать как сходимость по вероятности. Если выборка имеет повторяющиеся по величине элементы, что характерно для дискретных величин, то удобнее пользоваться сгруппированным вариационным рядом. В случае дискретной величины ступенчатый характер функции $F^*(x)$ с возрастанием n сохраняется. С ростом n могут появляться дополнительные ступеньки до тех пор, пока не будут зарегистрированы все возможные значения дискретной случайной величины (если число их конечно).

При большом объёме выборки и большом числе различных по величине элементов выборки пользоваться простым и сгруппированным вариационными рядами неудобно. В таком случае пользуются *интервальным вариационным рядом*, который строится следующим образом. Вся широта распределения разбивается на r частичных интервалов и подсчитывается число элементов n_i, попавших в i-й интервал (i = 1, 2, ..., r). Для каждого интервала указываются его правая α_{i-1} и левая α_i границы и его середина $\bar{x}_i$. Вся эта информация представляется в виде таблицы произвольной формы. Приведём один из вариантов такой таблицы:

Номер интервала	1	2	3	…	r
Границы интервала	$\alpha_0 \div \alpha_1$	$\alpha_1 \div \alpha_2$	$\alpha_2 \div \alpha_3$	…	$\alpha_{r-1} \div \alpha_r$
Середина интервала	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	…	$\bar{x}_r$
Число точек в интервале	n_1	n_2	n_3	…	n_r
Относительная частота	n_1/n	n_2/n	n_3/n	…	n_r/n

Для наглядного представления о форме плотности распределения случайной величины X используются понятия *полигона и гистограммы* распределения, которые строятся по интервальному вариационному ряду. Для построения полигона нужно из середины каждого частичного интервала восстановить перпендикуляр длиной $p_i^* = n_i/n$ и соединить отрезками прямых вершины этих перпендикуляров. Вершины крайних перпендикуляров соединяются с концами крайних частичных интервалов. Относительные частоты p_i^* представлены в таблице последней строкой. Чтобы построить гистограмму, нужно на каждом частичном интервале построить прямоугольник высотой p_i^* (рис. 6.1).

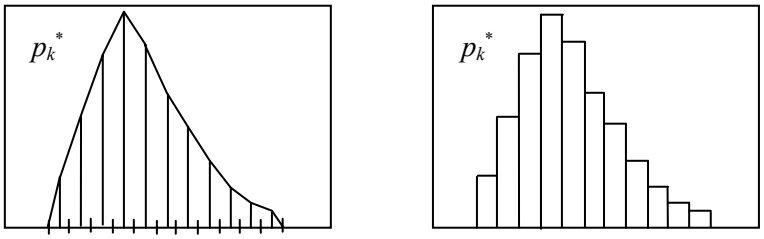

Рис. 6.1. **Полигон и гистограмма эмпирического распределения**

Относительные частоты p_i^* есть не что иное, как эмпирические вероятности попадания случайной величины в соответствующие интервалы (здесь и далее символ * означает, что величина определена по экспериментальным данным). Если по оси OY откладывать не p_i^*, а отношения p_i^*/Δ_i, где Δ_i – длины частичных интервалов, то полигон и гистограмма будут различными формами представления *эмпирической плотности распределения* вероятностей.

Любая группировка исходных данных, подобная той, которая применяется при построении интервального вариационного ряда, приводит к частичной потере информации. Интервальный вариационный ряд не содержит точных значений элементов выборки, так как все элементы, попавшие в i-й интервал (i = 1, 2, ..., r), фактически приравниваются к значению $\bar{x}_i$, находящемуся в середине интервала. Современная вычислительная техника позволяет проводить обработку данных, исходя непосредственно из простого вариационного ряда при любом объёме выборки. Использовать интервальный вариационный ряд рационально тогда, когда этого требует сам метод обработки экспериментальных данных.

Пример 6.1. Для определения среднего размера мужской обуви, на который нужно ориентироваться при оптовых закупках, был проделан следующий эксперимент. В течение определённого времени в обувном магазине фиксировался размер обуви, которую приобретали покупатели. В результате была получена следующая выборка:

39, 43, 42, 40, 44, 39, 42, 41, 41, 40, 42, 41, 42, 45, 43, 44, 40, 43, 41, 42,
41, 43, 38, 41, 42, 40, 43, 40, 44, 41, 43, 41, 39, 45, 43, 46, 42, 43, 42, 40,
43, 42, 41, 43, 39, 44, 40, 43, 41, 42, 41, 43, 42, 45, 44, 42, 41, 42, 40, 44.

Построить сгруппированный вариационный ряд и график эмпирической функции распределения вероятностей $F^*(x)$ случайной величины X – размера обуви, которую носит мужское население данного города.

Объединяя одинаковые значения размера обуви, получим следующий сгруппированный вариационный ряд:

x_i	38	39	40	41	42	43	44	45	46
n_i	1	4	8	12	13	12	6	3	1

Функция распределения вероятностей в точке x определяется следующим образом. Число элементов вариационного ряда, меньших, чем x, делится на объём выборки n. Например, для $x \in (x_1; x_2]$ слева находится всего один элемент, т.е $F^*(x) = 1/60$. Для $x \in (x_2; x_3]$ слева уже находится пять элементов $(n_1 + n_2)$, т. е. $F^*(x) = 5/60 = 1/12$, и т.д.

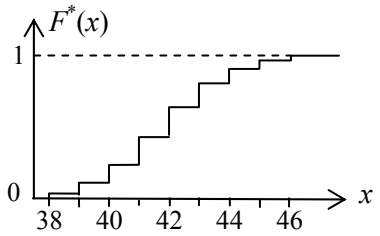

Рис. 6.2. ***F*(x)* для дискретной случайной величины**

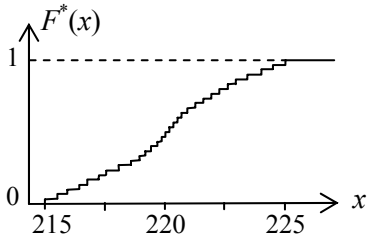

Рис. 6.3. ***F*(x)* для непрерывной случайной величины**

График эмпирической функции распределения вероятностей приведён на рис. 6.2. Для сравнения на рис. 6.3 приводится график эмпирической функции распределения вероятностей непрерывной случайной величины, рассмотренной в примере 6.2. Для непрерывных случайных величин вероятность совпадения по величине двух или нескольких измерений случайной величины равна нулю. Все измерения практически всегда оказываются различными, и скачки функции $F^*(x)$ будут равны величине $1/n$.

Пример 6.2. Напряжение (X) в электрической сети зависит от подключаемой к ней нагрузки, которая имеет случайный характер. Для изучения колебаний напряжения в сети в течение определённого периода времени было сделано 30 измерений с точностью до десятых долей вольта. Был получен следующий вариационный ряд:

215,0 215,5 215,9 216,4 216,8 217,3 217,5 218,1 218,6 218,9
219,2 219,4 219,7 219,8 220,0 220,2 220,3 220,5 220,7 220,9
221,3 221,6 221,9 222,3 222,6 222,9 223,4 224,0 224,5 225,0.

Построить график эмпирической функции распределения вероятностей случайной величины X, найти вероятность того, что напряжение в сети будет находиться в интервале от 219 до 221 вольт.

Эмпирическая функция распределения вероятностей $F^*(x)$ случайной величины X строится так же, как в примере 6.1. Величина всех ступенек будет одинаковой, равной 1/30 (рис. 6.3). Вероятность попадания случайной величины в заданный интервал определяется по обычным правилам: $P^*\{219 < X < 221\} = F^*(221) - F^*(219) = 20/30 - 10/30 = 1/3$.

Таким образом, примерно в 33% случаев будет наблюдаться напряжение в сети, величина которого находится между 219 и 221 вольтами.

6.3. Эмпирические числовые характеристики

Числовые характеристики случайных величин, найденные на основе экспериментальных данных, называются *точечными оценками* этих характеристик или *эмпирическими характеристиками*. Чтобы понять структуру формул, определяющих эмпирические моменты случайной величины, рассмотрим простой вариационный ряд $X_n = \{x_1, x_2, \ldots, x_n\}$. Можно формально считать, что рассматривается дискретная случайная величина, имеющая n возможных значений с вероятностями $1/n$. Математическое ожидание этой случайной величины и дисперсия определяются по общему правилу:

$$m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad D_x^* = M\{(X - m_x^*)^2\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x^*)^2. \qquad (6.1)$$

Формулы (6.1) соответствуют простому вариационному ряду. Для сгруппированного вариационного ряда число слагаемых в (6.1) уменьшится до *r*, где *r* – число различных по величине элементов выборки, за счёт группирования одинаковых слагаемых. Для интервального вариационного ряда формулы будут иметь такую же структуру, однако вместо непосредственных измерений, в ней фигурируют середины частичных интервалов $\bar{x}_i$. Учитывая эти особенности, можно записать общие формулы для вычисления начальных ν_k^* и центральных μ_k^* *эмпирических моментов* случайной величины:

$$\nu_k^* = \begin{cases} \dfrac{1}{n}\sum_{i=1}^{n} x_i^k, \\ \dfrac{1}{n}\sum_{i=1}^{r} x_i^k n_i, \\ \dfrac{1}{n}\sum_{i=1}^{r} \bar{x}_i^k n_i, \end{cases} \qquad \mu_k^* = \begin{cases} \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \nu_1^*)^k, \\ \dfrac{1}{n}\sum_{i=1}^{r} (x_i - \nu_1^*)^k n_i, \\ \dfrac{1}{n}\sum_{i=1}^{r} (\bar{x}_i - \nu_1^*)^k n_i, \ k = 1, 2, ... \end{cases} \tag{6.2}$$

В этих формулах первая строка соответствует простому вариационному ряду, вторая – сгруппированному, третья – интервальному вариационному ряду. Формулы связи между центральными и начальными моментами не изменяется, т.е.

$$\mu_2^* = \nu_2^* - \nu_1^{*2}, \quad \mu_3^* = \nu_3^* - 3\nu_2^*\nu_1^* + 2\nu_1^{*3}, \quad \mu_4^* = \nu_4^* - 4\nu_3^*\nu_1^* + 6\nu_2^*\nu_1^{*2} - 3\nu_1^{*4}. \tag{6.3}$$

Эмпирическое математическое ожидание случайной величины совпадает с первым начальным моментом ν_1^*, а её эмпирическая дисперсия совпадает со вторым центральным моментом μ_2^*. Формулы, определяющие основные характеристики случайной величины, также сохраняют свою структуру. В них достаточно заменить теоретические моменты ν_k и μ_k на эмпирические ν_k^* и μ_k^*. Таким образом, эмпирические характеристики асимметрия (скошенность) и эксцесс определяются по обычным формулам:

$$Sk_x^* = \frac{\mu_3^*}{\sigma_x^{*3}}, \qquad Ex_x^* = \frac{\mu_4^*}{\sigma_x^{*4}} - 3, \tag{6.4}$$

где $\sigma_x^* = \sqrt{D_x^*}$ – эмпирическое среднее квадратическое отклонение величины *X*.

При вычислении эмпирических характеристик можно делать некоторые предварительные преобразования выборки, которые приводят к упрощению вычислений. При этом опираются на соответствующие свойства математического ожидания, дисперсии и т.п. Например, математическое ожидание можно вычислять по формуле:

$$m_x^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - C) + C. \tag{6.5}$$

Постоянная величина *C* выбирается так, чтобы суммирование оказалось наиболее простым. Преобразование типа $X - C$ означает сдвиг всей выборки по числовой оси на величину *C*. Дисперсия не изменяется, т.е. $D^*\{X\} = D^*\{X - C\}$. Можно вводить масштабный коэффициент, т.е. рассматривать величину αX вместо величины *X*, где α – масштабирующий множитель. При вычислениях следует учитывать, что $M\{\alpha X\} = \alpha M\{X\}$, а $D\{\alpha X\} = \alpha^2 D\{X\}$. Такие преобразования часто приводят к упрощению вычислений. Если вычисления проводятся на ЭВМ, то эти преобразования не целесообразны.

Пример 6.3. По выборке из примера 6.1 вычислить эмпирические математическое ожидание и дисперсию размера обуви, который пользуется спросом у населения.

Воспользуемся формулой (6.5). Выберем $C = 42$ и рассмотрим величину $Y = X - 42$:

$x_i - 42$	–4	–3	–2	–1	0	1	2	3	4
n_i	1	4	8	12	13	12	6	3	1

Эмпирические математическое ожидание $m_y^* = \nu_1^*$ и второй начальный момент будем вычислять по второй формуле из (6.2) для ν_k^*:

$$m_y^* = \frac{1}{n}\sum_{i=1}^{n} y_i n_i =$$

$$= \frac{1}{60}[(-4)\cdot 1 + (-3)\cdot 4 + (-2)\cdot 8 + (-1)\cdot 12 + 0\cdot 13 + 1\cdot 12 + 2\cdot 6 + 3\cdot 3 + 4\cdot 1] =$$

$$= -7/60 \approx -0{,}117$$

Следовательно, $m_x^* = m_y^* + 42 = 41{,}883$. Это средний (эмпирический) размер обуви, которую носит мужское население данного города. Вычислим дисперсию, пользуясь тем, что $D_x^* = D_y^*$:

$$D_x^* = D_y^* = \nu_2^* - \nu_1^{*2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 n_i - \nu_1^{*2} =$$

$$= \frac{1}{60}[(-4)^2\cdot 1 + (-3)^2\cdot 4 + (-2)^2\cdot 8 + (-1)^2\cdot 12 + 0^2\cdot 13 + 1^2\cdot 12 + 2^2\cdot 6 + 3^2\cdot 3 + 4^2\cdot 1] - 0{,}117^2 \approx$$

$$\approx 2{,}89.$$

Пример 6.4. По данной выборке случайной величины X вычислить все основные эмпирические характеристики: математическое ожидание m_x^*, дисперсию D_x*, среднее квадратическое отклонение σ_x^*, асимметрию Sk_x^* и эксцесс Ex_x^*.

1,8	1,3	2,3	2,7	4,7	3,4	1,0	0,1	0,2	2,7	0,3	2,1	0,7	3,3	8,0
0,8	4,0	0,2	0,3	0,6	3,0	3,5	4,6	0,5	0,6	4,1	2,7	0,3	0,4	1,2
4,5	1,6	1,5	9,6	4,0	0,3	0,7	7,3	2,5	2,0	3,7	0,1	0,9	4,9	0,1
1,2	0,5	0,0	1,3	2,8	0,6	1,4	0,8	1,1	0,9	0,4	1,2	0,2	0,1	0,7.

Наибольший элемент выборки равен 9,6, наименьший – 0, размах выборки равен 9,6. Учитывая, что элементы выборки распределены не равномерно на этом интервале, ширину первых пяти частичных интервалов выберем равной 0,6, двух следующих – 1,2, а последнего интервала – 4,2. Составим интервальный вариационный ряд, подсчитав число элементов выборки, попавших в каждый частичный интервал. Если значение элемента совпадает с левой границей частичного интервала, то его следует относить к данному интервалу. Значение, совпадающее с правой границей, не включается в данный интервал. В последний интервал включается и то значение, которое совпадает с его правой границей. Вычисления проводятся в таблице, содержащей вариационный ряд и строки, необходимые для вычисления начальных моментов.

По результатам вычислений, получим: $m_x^* = \nu_1^* = \frac{1}{n}\cdot\sum_{i=1}^{r}\bar{x}_i n_i = \frac{1}{60}\cdot 120 = 2,$

$$\nu_2^* = \frac{1}{n}\cdot\sum_{i=1}^{r}\bar{x}_i^2 n_i = \frac{1}{60}\cdot 447{,}84 = 7{,}46, \qquad \nu_3^* = \frac{1}{n}\sum_{i=1}^{r}\bar{x}_i^3 n_i = \frac{1}{60}\cdot 2252{,}88 = 37{,}55,$$

$$\nu_4^* = \frac{1}{n}\sum_{i=1}^{r}\bar{x}_i^4 n_i = \frac{1}{60}\cdot 13351{,}26 = 222{,}52, \quad D_x^* = \nu_2^* - \nu_1^{*2} = 7{,}46 - 2^2 = 3{,}46, \quad \sigma = \sqrt{D_x^*} = 1{,}86.$$

i	$z_{i-1} \div z_i$	$\bar{x}_i$	n_i	$\bar{x}_i n_i$	$\bar{x}_i^2 n_i$	$\bar{x}_i^3 n_i$	$\bar{x}_i^4 n_i$
1	0 ÷ 0,6	0,3	16	4,8	1,44	0,43	0,13
2	0,6 ÷ 1,2	0,9	12	10,8	9,72	8,75	7,87
3	1,2 ÷ 1,8	1,5	8	12	18	27	40,5
4	1,8 ÷ 2,4	2,1	4	8,4	17,64	37,04	77,79
5	2,4 ÷ 3,0	2,7	5	13,5	36,45	92,42	265,72
6	3,0 ÷ 4,2	3,6	8	28,8	103,68	373,25	1343,69
7	4, 2 ÷ 5,4	4,8	4	19,2	92,16	442,37	2123,37
8	5,4 ÷ 9,6	7,5	3	22,5	168,75	1265,63	9492,19
∑			60	120	447,84	2252,88	13351,26

Вычислим центральные моменты по формулам (6.3):

$$\mu_3^* = 37{,}55 - 3 \cdot 7{,}46 \cdot 2 + 2 \cdot 2^3 = 8{,}79,$$

$$\mu_4^* = 222{,}52 - 4 \cdot 37{,}55 \cdot 2 + 6 \cdot 7{,}46 \cdot 2^2 - 3 \cdot 2^4 = 53{,}16$$

Таким образом, асимметрия и эксцесс будут равны:

$$Sk_x^* = \frac{\mu_3^*}{\sigma_x^{*3}} = \frac{8{,}76}{1{,}86^3} = 1{,}36, \qquad Ex_x^* = \frac{\mu_4^*}{\sigma_x^{*4}} - 3 = \frac{53{,}27}{3{,}46^2} - 3 = 1{,}44.$$

6.4. Точечные оценки параметров. Свойства эмпирических характеристик

Требуется оценить некоторый параметр θ, связанный со случайной величиной X, используя выборку $X_n = \{x_1, x_2, \ldots, x_n\}$. Пусть в качестве такой оценки выбрана однозначная функция от элементов выборки $\theta^* = \theta^*(x_1, x_2, \ldots, x_n)$. Для конкретных значений элементов выборки эта оценка представляет собой одно число. Такие оценки называются *точечными оценками параметров*, так как на числовой оси они изображаются одной точкой. Задача состоит в том, чтобы найти такую оценку θ*, которая была бы в определённом смысле наиболее близкой к оцениваемому параметру θ.

Как функция элементов выборки, оценка θ* является случайной величиной. Определим её математическое ожидание. Оно, очевидно, будет зависеть от истинных числовых характеристик изучаемой величины X и от объёма выборки n. Пусть получено равенство:

$$M\{\theta^*\} = \theta + \varphi(\theta, n), \qquad (6.6)$$

где $\varphi(\theta, n)$ – некоторая функция истинного значения параметра θ. Желательно, чтобы функция $\varphi(\theta, n)$ равнялась нулю. Это бы означало, что математическое ожидание оценки параметра равно истинному значению этого параметра. Оценка θ*, обладающая таким свойством, называется *несмещённой оценкой* параметра θ. Если $\varphi(\theta, n) \neq 0$, то θ* называется *смещённой оценкой* параметра θ, а сама функция $\varphi(\theta, n)$ называется *смещением*.

Если при $n \to \infty$ оценка параметра сходится по вероятности к истинному значению параметра, то оценка θ* называется *состоятельной оценкой* параметра θ. Для дальнейшего изучения свойств оценки θ* можно определить её дисперсию, которая также окажется функцией от истинных числовых характеристик изучаемой случайной величины X и от объёма выборки n, т.е. $D\{\theta^*\} = D(\theta, n)$. Если оценка состоятельная, то $D(\theta, n)$ стремится к нулю при $n \to \infty$. Различные оценки одного и того же параметра будут иметь разные дис-

персии. Та из них, которая имеет наименьшую дисперсию, называется *эффективной оценкой* данного параметра.

Проведём краткий анализ эмпирических числовых характеристик. Найдём математическое ожидание и дисперсию оценки m_x^* истинного математического ожидания m_x случайной величины X: $M\{m_x^*\} = M\{\frac{1}{n}\sum_{i=1}^{n} x_i\} = \frac{1}{n}\sum_{i=1}^{n} M\{x_i\} = \frac{1}{n}\sum_{i=1}^{n} m_x = \frac{1}{n}\cdot nm_x = m_x$.

$$D\{m_x^*\} = D\{\frac{1}{n}\sum_{i=1}^{n} x_i\} = \frac{1}{n^2}\sum_{i=1}^{n} D\{x_i\} = \frac{1}{n^2}\sum_{i=1}^{n} D_x = \frac{1}{n^2}\cdot nD_x = \frac{D_x}{n}.$$

Здесь учтено, что элементы выборки x_i, являясь независимыми реализациями случайной величины X, имеют те же самые характеристики, что и сама величина X. Таким образом,

$$M\{m_x^*\} = m_x, \quad D\{m_x^*\} = \frac{D_x}{n}. \tag{6.7}$$

Из этого следует, что m_x^* является несмещённой и состоятельной оценкой истинного математического ожидания m_x случайной величины X ($D\{m_x^*\}\to 0$ при $n\to\infty$).

Аналогичный анализ для эмпирической дисперсии D_x^* показывает, что

$$M\{D_x^*\} = D_x - \frac{D_x}{n}.$$

Таким образом, эмпирическая дисперсия является смещённой оценкой дисперсии. Смещение равно $(-D_x/n)$ и стремится к нулю при $n\to\infty$. Однако при малом объёме выборки это смещение оказывается существенным. Для его устранения вводится поправочный коэффициент, при умножении которого на D_x^* получается другая оценка дисперсии, не имеющая смещения. Эта оценка обозначается обычно через S^2 (читается: «S – квадрат»):

$$S^2 = \frac{n}{n-1} D_x^* \quad \text{или} \quad S^2 = \frac{1}{n-1}\sum (x_i - m_x^*)^2 . \tag{6.8}$$

Очевидно, что $M\{S^2\} = D_x$.

Вычисление дисперсии величины S^2 не представляет принципиальных трудностей, но оказывается достаточно громоздким. Вычисления показывают, что эта дисперсия пропорциональна величине $1/n$ и, следовательно, стремится к нулю при $n\to\infty$. Таким образом, величина S^2 является несмещённой и состоятельной оценкой истинной дисперсии D_x. Её рекомендуется использовать вместо оценки D_x^*, особенно при малых значениях n.

Свойством несмещённости обладают только первые два эмпирических момента. Моменты более высоких порядков ни при каких весовых коэффициентах суммирования таким свойством не обладают, т. е. они всегда имеют неустранимое смещение.

Рассмотрим кратко методы нахождения оценок. Один из методов предполагает задание структуры оценки с точностью до неизвестных параметров, которые определяются из условия минимума дисперсии оценки. Примером применения этого метода является определение оценки математического ожидания случайной величины в случае неравноточных измерений. Пусть по выборке $X_n = \{x_1, x_2, \ldots, x_n\}$ требуется оценить параметры m_x и D_x, причём измерения x_i были произведены с разной точностью, т.е.

$$M\{x_i\} = m_x,\ M\{(x_i - m_x)^2\} = \sigma_i^2,\ i = 1, 2, \ldots, n.$$

Чем меньше дисперсия измерения, тем больше доверия этому измерению, т.е. измерения должны учитываться оценкой с разными весовыми коэффициентами. Исходя из этого, выберем следующую структуру оценки:

$$m_x^* = \sum_{i=1}^{n} b_i x_i \,. \qquad (6.9)$$

Чтобы оценка была несмещённой, необходимо выполнение условия:

$$M\{m_x^*\} = M\{\sum_{i=1}^{n} b_i x_i\} = \sum_{i=1}^{n} b_i M\{x_i\} = m_x \sum_{i=1}^{n} b_i = m_x,$$

т.е. $\sum_{i=1}^{n} b_i = 1$. Коэффициенты b_i нужно выбрать так, чтобы они минимизировали бы дисперсию оценки. Так как измерения статистически независимы, то $D\{m_x^*\} = D\{\sum_{i=1}^{n} b_i x_i\} = \sum_{i=1}^{n} b_i^2 \sigma_i^2$. Коэффициенты b_i должны определяться из условия: $\sum_{i=1}^{n} b_i^2 \sigma_i^2 = \min_{b_i}$, $i = 1, 2, ..., n$, при ограничении $\sum_{i=1}^{n} b_i = 1$.

Проводя минимизацию, получим:

$$b_i = \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j}, \text{ где } \alpha_i = \frac{\sigma_n^2}{\sigma_i^2}. \qquad (6.10)$$

Итак, при неравноточных измерениях для оценки математического ожидания следует пользоваться формулой (6.9), подставляя в неё коэффициенты b_i из (6.10).

Второй метод нахождения оценок – *метод моментов*. В этом методе используются теоретические формулы, которые связывают оцениваемый параметр с моментами случайной величины. Для получения оценки неизвестного параметра нужно в соответствующую формулу подставить вместо теоретических моментов эмпирические моменты. Предположим, например, что случайная величина X распределена по экспоненциальному закону: $f(x) = \lambda \exp\{-\lambda x\}$, где $\lambda > 0$, $x > 0$, причём параметр λ неизвестен. Требуется оценить этот параметр. Пусть по выборке X_n получена оценка математического ожидания m_x^* исследуемой случайной величины. С другой стороны, известна формула, связывающая параметр λ экспоненциального распределения с математическим ожиданием m_x: $\lambda = 1/m_x$. Подставляя в эту формулу вместо m_x оценку m_x^*, получим оценку параметра λ: $\lambda^* = 1/m_x^*$. В некоторых случаях оценки, полученные этим простым способом, совпадают с оценками, полученными с помощью других, более сложных методов.

Третий метод – *метод наибольшего правдоподобия*. Этот метод требует знания закона распределения случайной величины с точностью до неизвестных параметров. Предположим, что плотность распределения вероятностей величины X равна $f_X(x, \theta)$, где θ – неизвестный параметр, который требуется оценить. Тогда каждое измерение x_i из выборки $X_n = \{x_1, x_2, \ldots, x_n\}$ будет иметь плотность распределения $f_X(x_i, \theta)$. Элементы выборки x_i являются статистически независимыми, поэтому n – мерная плотность распределения вероятностей выборки равна произведению одномерных плотностей, т.е.

$$f(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} f_X(x_i, \theta). \qquad (6.11)$$

Эту плотность называют *функцией правдоподобия*. Можно предполагать, что в выборке чаще встречаются те возможные значения величины X, для которых плотность распределения имеет относительно большие значения. Из этого следует, что в качестве оценки параметра θ логично взять такое значение, которое максимизирует функцию

правдоподобия. Однако с целью упрощения вычислений используют не функцию правдоподобия непосредственно, а её натуральный логарифм:

$$L(x_1, x_2, ..., x_n, \theta) = \sum_{i=1}^{n} \ln[f_X(x_i, \theta)]. \qquad (6.12)$$

Доказано, что эта функция имеет максимум, причём значение $\theta = \theta^*(x_1, x_2, \ldots, x_n)$, при котором достигается этот максимум, является оценкой параметра θ, обладающей наименьшей дисперсией. Таким образом, для определения оценки нужно решить уравнение

$$\frac{\partial L(x_1, x_2, ..., x_n, \theta)}{\partial \theta} = 0. \qquad (6.13)$$

Решение этого уравнения $\theta^*(x_1, x_2, \ldots, x_n)$ и будет оценкой параметра θ. Оценки, найденные таким способом, называются *оценками максимального (или наибольшего) правдоподобия*. Оценки максимального правдоподобия всегда являются эффективными оценками.

Пример 6.5. По выборке объёма $n = 20$ найдена оценка дисперсии D_x^* случайной величины X: $D_x^* = 4{,}8$. Определить несмещённую оценку S^2 неизвестной дисперсии.

Воспользуемся формулой (6.9): $S^2 = \frac{n}{n-1} D_x^* = \frac{20}{19} \cdot 4{,}8 \approx 5{,}05$.

Пример 6.6. Некоторая постоянная величина C была измерена тремя измерительными приборами по 4 раза каждым. Измерения, полученные с помощью первого прибора, равны 4,75, 4,9, 5,05, 4,85, с помощью второго прибора – 4,9, 5,1, 4,7, 5,15, с помощью третьего прибора – 4,7, 4,95, 5,05, 5,25. Класс точности измерительных приборов такой, что первый из них обеспечивает среднее квадратическое отклонение измерений $\sigma_1 = 0{,}1$, второй – $\sigma_2 = 0{,}15$, третий – $\sigma_3 = 0{,}2$. Требуется оценить истинное значение величины C.

Измерения величины C являются неравноточными. Поэтому применим для оценки формулу (6.9). Вычислим коэффициенты b_i по формулам (6.10): $\alpha_{1\div4} = \frac{\sigma_{12}^2}{\sigma_1^2} = \frac{0{,}2^2}{0{,}1^2} = 4$, $\alpha_{5\div8} = \frac{\sigma_{12}^2}{\sigma_2^2} = \frac{0{,}2^2}{0{,}15^2} = \frac{16}{9}$, $\alpha_{9\div12} = 1$. Поскольку каждым прибором было сделано по 4 измерения, то $n = 12$, и, следовательно, $\sum_{i=1}^{12} \alpha_i = 4 \cdot 4 + 4 \cdot \frac{16}{9} + 4 \cdot 1 = 4 \cdot \frac{61}{9} \approx 27{,}11$.

Таким образом, $b_{1\div4} = \frac{4}{27{,}11} \approx 0{,}148$, $b_{5\div8} = \frac{16}{27{,}11 \cdot 9} \approx 0{,}066$, $b_{9\div12} = \frac{1}{27{,}11} \approx 0{,}037$.

Подставим найденные значения и соответствующие измерения в формулу (6.10):

$$m_x^* = C^* = 0{,}148(4{,}75 + 4{,}9 + 5{,}05 + 4{,}85) + 0{,}066(4{,}9 + 5{,}1 + 4{,}7 + 5{,}15) +$$
$$+ 0{,}037(4{,}7 + 4{,}95 + 5{,}05 + 5{,}25) = 2{,}8934 + 1{,}3101 + 0{,}7382 \approx 4{,}94.$$

Итак, в качестве оценки истинного значения величины C следует взять $C^* = 4{,}94$.

Пример 6.7. Построить оценку максимального правдоподобия для параметра λ случайной величины X с плотностью распределения вероятностей $f_X(x) = \lambda \exp\{-\lambda x\}$, $x > 0$.

Пусть имеется выборка: $X_n = \{x_1, x_2, \ldots, x_n\}$. Тогда функция правдоподобия будет равна

$$f(x_1, x_2, ..., x_n, \lambda) = \prod_{i=1}^{n} f_X(x_i, \lambda) = \prod_{i=1}^{n} [\lambda \exp(-\lambda x_i)] = \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\}.$$

Найдём логарифм функции правдоподобия:

$$L(x_1, x_2, ..., x_n, \lambda) = \ln[f(x_1, x_2, ..., x_n, \lambda)] = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i .$$

Приравнивая производную этой функции по параметру λ к нулю, получим уравнение, определяющее оценку этого параметра: $\frac{\partial L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$ откуда следует, что $\lambda^* = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{m_x^*}$. Полученная максимально правдоподобная оценка полностью совпадает с оценкой этого параметра с помощью метода моментов.

6.5. Доверительные интервалы. Общие определения

Точечные оценки оценивают неизвестное значение параметра одним числом. Недостатком точечных оценок является то, что в них не указывается точность оценки параметра при выборках конечного объёма. Можно лишь сказать, что при $n \to \infty$ оценки параметров сходятся по вероятности к истинным значениям этих параметров. Иногда удобнее оценивать значение параметра с помощью интервала, в который это значение попадает с определённой вероятностью. Пусть θ – оцениваемый параметр, а θ_1 и θ_2 – две функции элементов выборки $x_1, x_2, \ldots, x_n$, такие, что $\theta_1 < \theta_2$. Если выполняется соотношение

$$P\{\theta_1 < \theta < \theta_2\} = \gamma, \qquad (6.14)$$

то интервал (θ_1, θ_2) называется 100 γ-процентным *доверительным интервалом* параметра θ. Другими словами, доверительный интервал – это интервал, в котором с заданной вероятностью находится значение неизвестного параметра. Значения θ_1 и θ_2 называют соответственно нижней и верхней границами доверительного интервала, а γ – *доверительной вероятностью* или *коэффициентом доверия*. Неважно, каким образом были получены границы интервала θ_1 и θ_2, важен сам факт выполнения соотношения (6.14). Доверительный интервал даёт определённую информацию о точности оценки данного параметра.

Для построения доверительного интервала необходимо знать тот или иной закон распределения вероятностей. Предположим, например, что неизвестный параметр θ можно интерпретировать как некоторую случайную величину с известной плотностью распределения вероятностей $f(\theta)$. Пусть $\theta^* = \theta^*(x_1, x_2, \ldots, x_n)$ – точечная оценка параметра θ. Тогда в некоторых случаях можно определить условную плотность распределения вероятностей $f(\theta \mid \theta^*) = \frac{f(\theta^* \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(\theta^* \mid \theta) f(\theta) d\theta}$. Следовательно, из соотношения $P\{\theta_1 < \theta < \theta_2\} = \int_{\theta_1}^{\theta_2} f(\theta \mid \theta^*) d\theta = \gamma$ можно определить границы θ_1 и θ_2 доверительного интервала с доверительной вероятностью γ.

Однако не всегда можно задать неизвестный параметр плотностью распределения вероятностей. Обычно неизвестный параметр является некоторой постоянной величиной. Поэтому при построении доверительного интервала пользуются не условной плотностью распределения $f(\theta \mid \theta^*)$, а условной плотностью $f(\theta^* \mid \theta)$. Рассмотрим один из возможных способов построения доверительного интервала с использованием этой плотно-

сти. Зададим некоторую доверительную вероятность γ и рассмотрим соотношение $P\{|\ \theta^* - \theta\ | < \delta\} = \gamma$.

Это соотношение определяет симметричный относительно θ доверительный интервал. Рассматривая это соотношение как уравнение относительно δ, можно определить δ, используя известную плотность $f(\theta^*|\theta)$. Тем самым доверительный интервал будет найден. Действительно, $P\{|\ \theta^* - \theta\ | < \delta\} = P\{-\delta < \theta^* - \theta < \delta\} = P\{\theta - \delta < \theta^* < \theta + \delta\} = P\{\theta^* - \delta < \theta < \theta^* + \delta\} = \gamma$. Последнее равенство означает, что $\theta \in (\theta^* - \delta, \theta^* + \delta)$ с вероятностью γ. Данные соотношения показывают, что при построении доверительного интервала можно исходить как из плотности $f(\theta^*\ |\ \theta)$, так и из плотности $f(\theta\ |\ \theta^*)$.

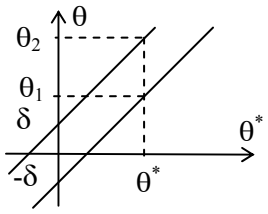

Рис. 6.4. **Доверительная область с прямолинейными границами**

Величина δ определяет ширину доверительного интервала. Для фиксированного значения доверительной вероятности γ и для неизменной плотности $f(\theta^*\ |\ \theta)$ эта величина является постоянной. Границы доверительного интервала определяются равенствами $\theta = \theta^* - \delta$ и $\theta = \theta^* + \delta$. Если считать θ и θ^* переменными, то эти два равенства являются уравнениями прямых линий. Вся область, заключённая между этими прямыми, называется *доверительной областью* (рис. 6.4). Располагая доверительной областью можно определить доверительный интервал для любого значения оценки θ^*. На рисунке указаны границы доверительного интервала θ_1 и θ_2 для произвольного значения оценки θ^*.

Границы доверительной области не обязательно являются прямыми линиями. Форма доверительной области определяется структурой самой оценки и, прежде всего, видом закона распределения $f(\theta^*\ |\ \theta)$. Здесь рассмотрен самый простой случай построения доверительной области. На рис. 6.5 изображена доверительная область для неизвестной вероятности события. Границами этой области является линия эллипса (см. п. 6.7).

Для дискретных случайных величин не всегда можно найти доверительный интервал, имеющий коэффициент доверия, в точности равный γ, если γ задано произвольно. Это связано с тем, что закон распределения дискретной случайной величины имеет ступенчатый характер.

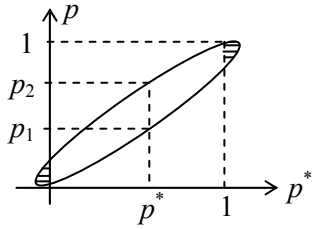

Рис. 6.5. **Доверительная область в форме эллипса**

Установление доверительного интервала не означает того факта, что неизвестный параметр принадлежит этому интервалу. Можно лишь утверждать, что с вероятностью γ этот параметр находится внутри интервала. При этом, разумеется, с вероятностью 1 – γ данный параметр находится вне этого интервала. Доверительную вероятность γ выбирают достаточно большой (γ = 0,9 ÷ 0,99). Следует иметь в виду, что при увеличении доверительной вероятности увеличивается длина доверительного интервала. Таким образом, при выборе значения доверительной вероятности следует придерживаться разумного компромисса. Если есть необходимость повысить доверительную вероятность при сохранении длины доверительного интервала, то нужно увеличить объём выборки.

6.6. Доверительные интервалы параметров нормального распределения

Пусть X – случайная величина, распределённая по нормальному закону с математическим ожиданием m_x и дисперсией σ_x^2. Требуется оценить неизвестное математиче-

ское ожидание m_x. По выборке $X_n = \{x_1, x_2, \ldots, x_n\}$ вычислим эмпирическое математическое ожидание m_x^*, которое является точечной оценкой параметра m_x:

$$m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad M\{m_x^*\} = m_x, \quad D\{m_x^*\} = \frac{D_x}{n} = \frac{\sigma_x^2}{n}.$$

Поскольку оценка m_x^* является линейной относительно измерений нормально распределённой случайной величины, то сама она также будет иметь нормальное распределение с математическим ожиданием m_x и дисперсией σ_x^2/n.

Пусть σ_x^2 является известной величиной. Преобразуем оценку m_x^* в центрированную и нормированную случайную величину T:

$$T = \frac{m_x^* - m_x}{\sigma_x/\sqrt{n}} = \frac{\sqrt{n}(m_x^* - m_x)}{\sigma_x}, \quad M\{T\} = 0, \quad D\{T\} = 1. \tag{6.15}$$

Величина T распределена нормально с нулевым математическим ожиданием и единичной дисперсией. Пользуясь таблицами нормального распределения и исходя из заданного значения вероятности γ, можно решить следующее уравнение относительно ε_γ:

$$P\{|T| < \varepsilon_\gamma\} = \gamma \quad \text{или} \quad P\{-\varepsilon_\gamma < T < \varepsilon_\gamma\} = \gamma. \tag{6.16}$$

Для нормального распределения вероятность попадания случайной величины в заданный интервал выражается через функцию Лапласа: $P\{-\varepsilon_\gamma < T < \varepsilon_\gamma\} = P\{-\varepsilon_\gamma < T < \varepsilon_\gamma\} = \Phi(\varepsilon_\gamma) - \Phi(-\varepsilon_\gamma) = 2\Phi(\varepsilon_\gamma)$. Здесь учтена нечётность функции Лапласа, т. е. $\Phi(-z) = -\Phi(z)$. Для определения ε_γ нужно, пользуясь таблицей функции Лапласа, решить уравнение

$$\Phi(\varepsilon_\gamma) = \gamma/2. \tag{6.17}$$

Вернёмся вновь к оценке m_x^*. Подставим в (6.16) выражение для T из (6.15):

$$P\{-\varepsilon_\gamma < T < \varepsilon_\gamma\} = P\{-\varepsilon_\gamma < \frac{\sqrt{n}(m_x^* - m_x)}{\sigma_x} < \varepsilon_\gamma\}.$$

Разрешая внутреннее неравенство относительно m_x, приходим к выражению:

$$P\{m_x^* - \varepsilon_\gamma \frac{\sigma_x}{\sqrt{n}} < m_x < m_x^* + \varepsilon_\gamma \frac{\sigma_x}{\sqrt{n}}\} = \gamma.$$

Итак, если m_x^* – эмпирическое математическое ожидание нормально распределённой величины с истинным математическим ожиданием m_x и известной дисперсией σ^2, то

$$m_x \in \left(m_x^* - \varepsilon_\gamma \frac{\sigma_x}{\sqrt{n}}; m_x^* + \varepsilon_\gamma \frac{\sigma_x}{\sqrt{n}} \right) \tag{6.18}$$

с вероятностью γ. При этом величина ε_γ находится по таблицам нормального распределения, как решение уравнения (6.17).

Пример 6.8. Класс точности измерительного прибора обеспечивает среднюю квадратическую погрешность измерений $\sigma_x = 0{,}05$, причём ошибка измерений распределена по нормальному закону с нулевым средним значением. При измерении некоторой постоянной величины были получены следующие значения: 5,25; 5,23; 5,29; 5,31; 5,22; 5,26; 5,23; 5,26; 5,26; 5,24; 5,25; 5,21; 5,27; 5,24; 5,28; 5,25. Построить доверительный интервал для истинного значения измеряемой величины с доверительной вероятностью 0,95.

Измерение X является случайной величиной с неизвестным математическим ожиданием m_x и известным средним квадратическим отклонением σ_x. Так как систематическая ошибка прибора равна нулю, то значение измеряемой величины совпадает с m_x. Суммируя все измерения и деля полученную сумму на число измерений $n = 16$, найдём оценку m_x^*: $m_x^* \approx 5{,}253$. Решим уравнение (6.17) при $\gamma = 0{,}95$: $\Phi(\varepsilon_\gamma) = 0{,}95/2$ или $\Phi(\varepsilon_\gamma) = 0{,}475$. По значению функции Лапласа, равному 0,475, по таблице определяем соответствующее значение аргумента этой функции: $\varepsilon_\gamma = 1{,}96$. Тогда с вероятностью 0,95 справедливо выражение: $m_x \in \left(5{,}253 - 1{,}96 \cdot \frac{0{,}05}{\sqrt{16}};\ 5{,}253 + 1{,}96 \cdot \frac{0{,}05}{\sqrt{16}}\right)$. Вычисляя, получим доверительный интервал: $m_x \in (5{,}228;\ 5{,}278)$ с вероятностью $\gamma = 0{,}95$.

Пример 6.9. Случайная величина X распределена нормально и имеет дисперсию, равную 16. Сколько раз нужно измерить случайную величину, чтобы с вероятностью не меньшей чем 0,9 можно было утверждать, что эмпирическое математическое ожидание отклонится от истинного математического ожидания на величину не большую чем 0,2?

Здесь $\gamma = 0{,}9$, $\sigma_x^2 = 16$, $\sigma_x = 4$. Неизвестной величиной является объём выборки n. Решим уравнение $\Phi(\varepsilon_\gamma) = 0{,}9/2 = 0{,}45$. Используя таблицу, получим: $\varepsilon_\gamma = 1{,}645$. Отклонение оценки m_x^* от m_x будет не более чем $\varepsilon_\gamma \cdot \frac{\sigma_x}{\sqrt{n}} = 1{,}645 \cdot \frac{4}{\sqrt{n}} = \frac{6{,}580}{\sqrt{n}}$. По условию это отклонение должно быть не более чем 0,2: $\frac{6{,}580}{\sqrt{n}} \le 0{,}2$, или $\sqrt{n} \ge \frac{6{,}58}{0{,}2} = 32{,}9$, или $n \ge 1082{,}41$. Объём выборки n – целое число. Поэтому берём ближайшее целое число: $n = 1083$.

Пусть теперь дисперсия σ_x^2 случайной величины X неизвестна. Пусть S^2 – несмещённая оценка дисперсии этой величины, т.е. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - m_x^*]^2$. Доказано, что величина $[(n-1)S^2]/\sigma_x^2$ распределена по закону χ^2 с $(n-1)$-й степенью свободы (см. п. 4.4). Число степеней свободы равно числу независимых слагаемых в S^2. Всего слагаемых n, однако, измерения x_i, входящие в S^2, входят также и в m_x^*, что и приводит к уменьшению числа независимых слагаемых на единицу. Доказано также, что величина $T = \frac{\sqrt{n}(m_x^* - m_x)}{S}$ распределена по закону Стьюдента с $(n-1)$-й степенью свободы.

Следовательно, пользуясь таблицей распределения Стьюдента, можно решить уравнение (6.16) относительно ε_γ при заданном значении доверительной вероятности γ. Дальнейшее построение доверительного интервала аналогично построению доверительного интервала в случае известной дисперсии σ_x:

$$m_x \in \left(m_x^* - \varepsilon_\gamma \frac{S}{\sqrt{n}};\ m_x^* + \varepsilon_\gamma \frac{S}{\sqrt{n}}\right) \quad \text{с вероятностью} \quad \gamma. \qquad (6.19)$$

По своей структуре формула (6.19) совпадает с формулой (6.18). Различие заключается в том, что в (6.19) используется величина S, а не σ_x. Кроме того, в данном случае при решении уравнения (6.16) применяется таблица распределения Стьюдента.

Пример 6.10. Пусть условия примера 6.9 сохраняются, но средняя квадратическая ошибка измерительного прибора является неизвестной. Требуется построить доверительный интервал для измеряемой величины.

В примере 6.8 уже было найдено значение $m_x^* = 5{,}253$. Для построения доверительного интервала нужно вычислить оценку S^2 дисперсии σ_x^2. Найдём второй начальный момент. Для этого просуммируем квадраты всех элементов выборки и поделим полученную сумму на n:

$$\nu_2^* = \frac{1}{n}\sum_{i=1}^{n} x_i^2 =$$

$$= \frac{1}{16}(5{,}21^2 + 5{,}22^2 + 2\cdot 5{,}23^2 + 2\cdot 5{,}24^2 + 3\cdot 5{,}25^2 + 3\cdot 5{,}26^2 + 5{,}27^2 + 5{,}28^2 + 5{,}29^2 + 5{,}31^2) \approx$$

$$\approx 27{,}596.$$

Таким образом, $D_x^* = \nu_2^* - m_x^{*2} = 27{,}596 - 5{,}253^2 = 0{,}002$, $S^2 = \frac{n}{n-1} D_x^* = \frac{16}{15} 0{,}002 \approx 0{,}0021$.

Извлекая корень, находим $S \approx 0{,}046$.

Решим уравнение (6.16), имея в виду, что доверительная вероятность γ равна 0,95, а число степеней свободы равно $r = n - 1 = 15$. По таблице находим $\varepsilon_\gamma = 2{,}13$. Подставляя вычисленные значения в (6.19), получим: $m_x \in \left(5{,}253 - 2{,}13\cdot\frac{0{,}046}{\sqrt{16}};\ 5{,}253 + 2{,}13\cdot\frac{0{,}046}{\sqrt{16}}\right)$.

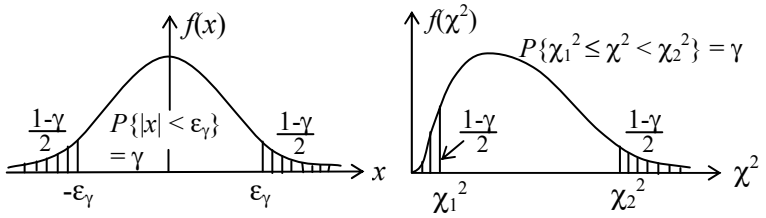

Рис. 6.6. **Построение доверительных интервалов в случае симметричного и несимметричного закона распределения**

Проводя вычисления, найдём окончательно доверительный интервал

$$m_x \in (5{,}229;\ 5{,}277) \quad \text{с вероятностью} \quad \gamma = 0{,}95.$$

Полученный доверительный интервал оказался более узким, чем интервал, найденный в примере 6.8. Это связано с тем, что данная конкретная выборка указывает на более высокую реальную точность прибора, чем та, которая определена изготовителем прибора. Однако для подтверждения этого нужно значительно увеличить объём выборки.

Рассмотрим процесс построения доверительного интервала для дисперсии нормально распределённой случайной величины. Пусть

$$\chi^2 = \frac{(n-1)S^2}{\sigma_x^2}. \tag{6.20}$$

Эта величина имеет распределение χ^2 с $(n-1)$-й степенью свободы. Распределение χ^2 не является симметричным. Поэтому невозможно использовать уравнение типа (6.16) для определения границ интервала. Построение доверительного интервала в случаях симметричных и несимметричных распределений иллюстрируется на рис. 6.6.

Будем исходить из следующего уравнения:

$$P\{\chi_1^2 \le \chi^2 < \chi_2^2\} = \gamma. \tag{6.21}$$

Это уравнение содержит две неизвестные, поэтому имеет бесчисленное множество решений. Чтобы получить одно решение, нужно иметь одно дополнительное условие. Вероятность попадания в интервал $(\chi_1^2; \chi_2^2)$ равна γ. Выберем интервал так, чтобы оставшаяся вероятность $1 - \gamma$ была распределена поровну между интервалами $(0; \chi_1^2)$ и $[\chi_2^2; \infty)$ (на рис. 6.6 они заштрихованы), т.е. положим $P\{\chi^2 < \chi_1^2\} = (1 - \gamma)/2$ и $P\{\chi^2 \geq \chi_2^2\} = (1 - \gamma)/2$. Тогда $P\{\chi^2 \geq \chi_1^2\} = 1 - (1 - \gamma)/2 = (1 + \gamma)/2$. Границы χ_1^2 и χ_2^2 определяем из двух уравнений:

$$P\{\chi^2 \geq \chi_1^2\} = \frac{1+\gamma}{2} \quad \text{и} \quad P\{\chi^2 \geq \chi_2^2\} = \frac{1-\gamma}{2} \tag{6.22}$$

Используется таблица значений χ_α^2, для которых $P\{\chi^2 \geq \chi_\alpha^2\} = \alpha$. По таблицам решаются уравнения (6.22) относительно χ_1^2 и χ_2^2 при заданном значении γ. Учитывая (6.21), получим: $P\{\chi_1^2 \leq \frac{(n-1)S^2}{\sigma_x^2} < \chi_2^2\} = \gamma$ или $P\{\frac{(n-1)S^2}{\chi_2^2} \leq \sigma_x^2 < \frac{(n-1)S^2}{\chi_1^2}\} = \gamma$. Таким образом, доверительный интервал для дисперсии будет иметь следующий вид:

$$\sigma_x^2 \in [\frac{(n-1)S^2}{\chi_2^2}; \frac{(n-1)S^2}{\chi_1^2}) \quad \text{с вероятностью} \quad \gamma. \tag{6.23}$$

Отсюда получим доверительный интервал для среднего квадратического отклонения:

$$\sigma_x \in [\frac{\sqrt{n-1} \cdot S}{\chi_2}; \frac{\sqrt{n-1} \cdot S}{\chi_1}) \quad \text{с вероятностью} \quad \gamma. \tag{6.24}$$

Пример 6.11. Пусть в условиях примера 6.8 средняя квадратическая ошибка измерительного прибора является неизвестной. На основании выборки требуется построить доверительный интервал для σ_x с доверительной вероятностью $\gamma = 0{,}95$.

Несмещённая оценка дисперсии S^2 была уже найдена по данной выборке в примере 6.10. Получено $S^2 = 0{,}0021$ и $S = 0{,}046$. Применяя таблицы, решаем два уравнения:

$$P\{\chi^2 \geq \chi_1^2\} = \frac{1+\gamma}{2} = \frac{1+0{,}95}{2} = 0{,}975, \quad P\{\chi^2 \geq \chi_2^2\} = \frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025.$$

По числу степеней свободы $r = n - 1 = 15$ и по вероятности, равной 0,975, находим $\chi_1^2 = 6{,}26$. По вероятности, равной 0,025, определяем $\chi_2^2 = 27{,}5$. Следовательно, с вероятностью 0,95

$$\sigma_x^2 \in [\frac{15 \cdot 0{,}0021}{27{,}5}; \frac{15 \cdot 0{,}0021}{6{,}26}) \quad \text{и} \quad \sigma_x \in [\frac{\sqrt{15} \cdot 0{,}046}{\sqrt{27{,}5}}; \frac{\sqrt{15} \cdot 0{,}046}{\sqrt{6{,}26}}).$$

Проводя соответствующие вычисления, окончательно получаем:

$\sigma_x^2 \in [0{,}0012; 0{,}0050)$ и $\sigma_x \in [0{,}034; 0{,}071)$ с вероятностью $\gamma = 0{,}95$.

Отметим, что если известно математическое ожидание m_x случайной величины, то вместо оценки дисперсии S^2 следует использовать оценку

$$D^* = \frac{1}{n}\sum_{i=1}^{n}[x_i - m_x]^2. \tag{6.25}$$

Эта оценка является несмещённой оценкой истинной дисперсии D_x и распределена по закону χ^2 с n степенями свободы.

Рассмотренные способы построения доверительных интервалов можно применять и для асимптотически нормальных оценок. Центральная предельная теорема подтвер-

ждает стремление закона распределения суммы случайных величин к нормальному закону при увеличении числа слагаемых. Это оправдывает применение рассмотренных способов построения доверительных интервалов при больших n. Таблицы распределений Стьюдента и χ^2 рассчитаны на малые объёмы выборок. При больших объёмах выборок вместо распределения Стьюдента используется непосредственно нормальное распределение. Что касается распределения χ^2, то граничные значения χ_1^2 и χ_2^2 вычисляются по формулам:

$$\chi_1^2 = n-1-\varepsilon_\gamma \cdot \sqrt{2(n-1)}, \quad \chi_2^2 = n-1+\varepsilon_\gamma \cdot \sqrt{2(n-1)}, \tag{6.26}$$

где ε_γ является решением уравнения (6.17) при заданной доверительной вероятности γ.

6.7. Построение доверительного интервала для вероятности события

Пусть в n последовательных независимых испытаниях событие A осуществилось k раз. Тогда вероятность события A можно оценить с помощью относительной частоты его осуществления, т.е.

$$p^* = \frac{k}{n}. \tag{6.27}$$

Величина k распределена по закону Бернулли, так что $M\{p^*\} = \frac{1}{n} M\{k\} = \frac{1}{n} \cdot np = p$ и $D\{p^*\} = \frac{1}{n^2} D\{k\} = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$, где $q = 1 - p$. Величина p^* является несмещённой и состоятельной оценкой вероятности p осуществления события A в одном испытании.

Преобразуем оценку p^* в центрированную и нормированную случайную величину T:

$$T = \frac{p^* - p}{\sqrt{pq/n}} = \sqrt{\frac{n}{pq}} \cdot (p^* - p), \quad M\{T\} = 0, \quad D\{T\} = 1.$$

Будем считать, что n достаточно велико. В соответствии с предельной теоремой Муавра–Лапласа, можно предполагать, что величина T распределена приближённо по нормальному закону с нулевым математическим ожиданием и единичной дисперсией. По таблицам нормального распределения, исходя из заданного значения доверительной вероятности γ, можно решить уравнение (6.17) относительно ε_γ. Тогда

$$P\{-\varepsilon_\gamma < T < \varepsilon_\gamma\} = P\{-\varepsilon_\gamma < \sqrt{\frac{n}{pq}} \cdot (p^* - p) < \varepsilon_\gamma\} = \gamma. \tag{6.28}$$

Разрешая внутреннее неравенство относительно p, получим окончательно:

$$P\{p^* - \varepsilon_\gamma \sqrt{\frac{pq}{n}} < p < p^* + \varepsilon_\gamma \sqrt{\frac{pq}{n}}\} = \gamma.$$

Таким образом, с вероятностью γ $p \in \left(p^* - \varepsilon_\gamma \sqrt{\frac{pq}{n}};\ p^* + \varepsilon_\gamma \sqrt{\frac{pq}{n}} \right)$. (6.29)

Границы этого интервала зависят от оцениваемого параметра p. Для избавления от этого можно использовать один из следующих способов. В первом способе в выражениях для границ интервала p заменяют на p^*. Тогда с вероятностью, приближённо равной γ,

$$p \in \left(p^* - \varepsilon_\gamma \sqrt{\frac{p^* q^*}{n}};\ p^* + \varepsilon_\gamma \sqrt{\frac{p^* q^*}{n}} \right),\ q^* = 1 - p^*.$$

Второй способ предполагает расширение интервала до максимально возможного. Для этого величина pq заменяется её наибольшим значением. Величина pq принимает максимальное значение 0,25 при $p = 0{,}5$. Тогда с вероятностью, не меньшей чем, γ

$$p \in \left(p^* - \varepsilon_\gamma \cdot \frac{0{,}5}{\sqrt{n}};\ p^* + \varepsilon_\gamma \cdot \frac{0{,}5}{\sqrt{n}} \right). \qquad (6.30)$$

Третий способ предполагает точное определение границ интервала. Рассмотрим основное неравенство, которое используется для получения доверительного интервала:

$$|T| < \varepsilon_\gamma \ \text{ или } \ |p^* - p| < \varepsilon_\gamma \sqrt{\frac{pq}{n}}, \quad \text{или} \quad n(p^* - p)^2 < \varepsilon_\gamma^2 p(1-p).$$

Это неравенство определяет доверительную область для неизвестной вероятности p. Приведём это неравенство к более удобному виду:

$$(n + \varepsilon_\gamma^2)p^2 - (2np^* + \varepsilon_\gamma^2)p + np^{*2} < 0. \qquad (6.31)$$

Доверительная область представляет собой множество пар значений (p, p*), попадающих внутрь эллипса, уравнение которого имеет вид:

$$(n + \varepsilon_\gamma^2)p^2 - (2np^* + \varepsilon_\gamma^2)p + np^{*2} = 0. \qquad (6.32)$$

На рис. 6.5. приведёно изображение области (6.31). Эллипс выходит за рамки квадрата с длиной стороны, равной единице, в направлении оси абсцисс. Заштрихованные части не входят в доверительную область, поскольку всегда должно выполняться неравенство $0 \le p^* \le 1$. Эффект выхода эллипса за рамки квадрата обусловлен тем, что распределения Бернулли для величины k заменено нормальным распределением. С ростом n эллипс становится всё более узким, «лишние» части уменьшаются. При $n \to \infty$ эллипс стягивается к диагонали квадрата с уравнением $p = p^*$, т. е. оценка p^* будет совпадать (по вероятности) с истинным значением вероятности p (см. теорему Бернулли из закона больших чисел).

Для определения доверительного интервала нужно решить квадратное уравнение (6.32) при известном значении оценки p^*. Получим два корня:

$$p_1 = \alpha(p^*, \varepsilon_\gamma, n) - \beta(p^*, \varepsilon_\gamma, n), \quad p_2 = \alpha(p^*, \varepsilon_\gamma, n) + \beta(p^*, \varepsilon_\gamma, n), \qquad (6.33)$$

$$\alpha(p^*, \varepsilon_\gamma, n) = \frac{2np^* + \varepsilon_\gamma^2}{2(n + \varepsilon_\gamma^2)},\ \beta(p^*, \varepsilon_\gamma, n) = \frac{\varepsilon_\gamma \sqrt{4np^*(1-p^*) + \varepsilon_\gamma^2}}{2(n + \varepsilon_\gamma^2)}.$$

Найденные значения p_1 и p_2 ($p_1 < p_2$) являются границами доверительного интервала (рис. 6.5), так что с вероятностью γ,

$$p \in (p_1; p_2). \qquad (6.34)$$

При больших значениях n все три способа дают практически одинаковые результаты. Аппроксимация распределения Бернулли нормальным распределением считается допустимой, если выполняется неравенства: $np \ge 4$ и $nq \ge 4$. Отсюда следует, что наименьшее значение n достигается при $p = 0{,}5$. Тогда должно быть $n \ge 8$. Если $p = 0{,}05$ или $p = 0{,}95$, то должно быть $n \ge 80$. Исходя из этих оценок, можно приближённо судить о правомерности применения рассмотренных здесь способов. Так как вероятность p неизвестна, то для оценки этой правомерности вычисляют np^* и nq^*. Можно построить доверительный интервал, опираясь на закон распределения Бернулли, однако эта процедура достаточно громоздкая. Целесообразность применения этой процедуры очень сомнительна, так как при малых значениях n точность оценки вероятности p настолько низка,

что выбор способа построения доверительного интервала уже не имеет практического значения. Отметим, что доверительная область для вероятности события, построенная с использованием закона распределения Бернулли, ограничена ступенчатой линией, напоминающей линию эллипса. Но она не будет выходить за рамки квадрата: $0 \le p \le 1$, $0 \le p^* \le 1$.

Пример 6.12. Было проведено 36 независимых испытаний, при этом событие A произошло в 9-ти испытаниях. Оценить истинную вероятность p осуществления события A в одном испытании с доверительной вероятностью $\gamma = 0{,}9$.

В данном случае $p^* = 9/36 = 0{,}25$. Поскольку $np^* = 9 > 4$ и $nq^* = 27 > 4$ ($q^* = 0{,}75$), то допустима аппроксимация распределения Бернулли нормальным распределением. Решая уравнение (6.17) относительно ε_γ при $\gamma/2 = 0{,}9/2 = 0{,}45$, по таблице значений функции Лапласа находим $\varepsilon_\gamma = 1{,}645$. Воспользуемся формулой (6.30):

$$p \in \left(0{,}25 - 1{,}645 \cdot \sqrt{\frac{0{,}25 \cdot 0{,}75}{36}};\ 0{,}25 + 1{,}645 \cdot \sqrt{\frac{0{,}25 \cdot 0{,}75}{36}} \right).$$

Производя соответствующие вычисления, получим: $p \in (0{,}131;\ 0{,}369)$ с вероятностью, приближённо равной 0,9.

Найдём теперь доверительный интервал по формуле (6.31):

$$p \in \left(0{,}25 - 1{,}645 \cdot \frac{0{,}5}{\sqrt{36}};\ 0{,}25 + 1{,}645 \cdot \frac{0{,}5}{\sqrt{36}} \right),$$

или $p \in (0{,}113;\ 0{,}387)$ с вероятностью, не меньшей чем 0,9.

Вычислим точные значения границ доверительного интервала. Для этого найдём корни уравнения (6.32) по формулам (6.33):

$$\alpha(p^*, \varepsilon_\gamma, n) = \frac{2 \cdot 36 \cdot 0{,}25 + 1{,}645^2}{2(36 + 1{,}645^2)} = 0{,}2675,$$

$$\beta(p^*, \varepsilon_\gamma, n) = \frac{1{,}645\sqrt{4 \cdot 36 \cdot 0{,}25 \cdot 0{,}75 + 1.{,}45^2}}{2(36 + 1{,}645^2)} = 0{,}1158\,.$$

Следовательно, $p_1 = 0{,}2675 - 0{,}1158 \approx 0{,}152$, $p_2 \approx 0{,}383$, и $p \in (0{,}152;\ 0{,}383)$ с вероятностью, равной 0,9.

Три интервала, полученные разными способами, мало отличаются друг от друга. Длины интервалов равны соответственно 0,238, 0,274 и 0,231. Заметим, что выражение «точные границы интервала» является условным, поскольку распределение Бернулли было приближённо заменено нормальным распределением. В особо ответственных случаях из этих трёх интервалов предпочтительно выбирать самый большой интервал.

Термины для запоминания

- Выборка; вариационный ряд; простой вариационный ряд; сгруппированный вариационный ряд; интервальный вариационный ряд; эмпирическая функция распределения; полигон; гистограмма; эмпирическая плотность распределения.
- Эмпирические числовые характеристики; точечные оценки параметров; смещённые оценки; несмещённые оценки; состоятельные оценки; эффективные оценки; метод

моментов; метод наибольшего правдоподобия; функция правдоподобия; оценка наибольшего правдоподобия.

- Доверительные интервалы; доверительная вероятность; коэффициент доверия.

Список контрольных вопросов

1. Перечислите основные задачи математической статистики.
2. Что называют выборкой случайной величины?
3. Перечислите виды вариационных рядов и поясните, чем они отличаются друг от друга.
4. Что такое размах выборки?
5. Как построить эмпирическую функцию распределения вероятностей?
6. Что такое полигон и что такое гистограмма эмпирического распределения?
7. Как построить эмпирическую плотность распределения вероятностей?
8. Как вычисляются эмпирические числовые характеристики случайных величин?
9. Какие оценки называют точечными оценками параметров?
10. Какие оценки называются несмещёнными?
11. Какие оценки называются состоятельными?
12. Какие оценки называются эффективными?
13. Перечислите методы получения оценок и поясните их смысл.
14. Как построить оценку математического ожидания случайной величины по её измерениям?
15. Как построить оценку дисперсии случайной величины по её измерениям?
16. Как построить несмещённую оценку дисперсии случайной величины?
17. Что такое доверительный интервал и что такое доверительная вероятность?
18. Какие законы распределения применяются при построении доверительных интервалов для параметров нормально распределённой случайной величины?

Вопросы для обсуждения

1. Влияют ли схема и условия измерения случайной величины для статистического анализа на результаты это анализа?
2. Могут ли здесь возникать чисто логические ошибки?

ТЕМА 7.

Проверка статистических гипотез

Студент должен
освоить:

- понятие статистической гипотезы, статистического критерия, ошибок первого и второго рода;

приобрести навыки:

- использования наиболее распространённых критериев проверки гипотез, в частности, критерия χ^2-Пирсона.

Краткое содержание

Статистическая проверка гипотез. Проверка гипотез о параметрах нормального распределения. Критерий квантилей. Критерий согласия Пирсона и его применение.

7.1. Общие положения

Любое предположение относительно исследуемой случайной величины, истинность которого требуется проверить на основе выборки этой случайной величины, называют *статистической гипотезой*. Гипотеза может касаться неизвестных параметров закона распределения вероятностей, самого закона распределения, смешанных статистических моментов, описывающих связь между величинами, и т.п. Существует большое разнообразие статистических гипотез и, соответственно, большое количество разнообразных методов проверки этих гипотез.

Общая идея проверки статистических гипотез заключается в следующем. Пусть имеется выборка $X_n = \{x_1, x_2, \ldots, x_n\}$ случайной величины X и выдвигается некоторая гипотеза относительно этой величины. Эту гипотезу называют *нулевой гипотезой* (H_0), а противоположную ей – *альтернативной гипотезой* (H_1). Предположим, что каким-то образом определена такая n-мерная область D, что при $(x_1, x_2, \ldots, x_n) \notin D$ нулевая гипотеза принимается, а при $(x_1, x_2, \ldots, x_n) \in D$ эта гипотеза отвергается. Тогда область D называется *критической областью*. При принятии решения на основе конкретной выборки возможны следующие варианты: гипотеза H_0 является верной, и она принимается в соответствии с критерием; H_0 является ложной гипотезой и она отвергается в соответствии с критерием. Эти решения являются верными. Возможны варианты, когда принятое решение является ошибочным: H_0 отвергается, являясь истинной гипотезой, или H_0 принимается, хотя на самом деле она является ложной. В первом случае ошибку в решении называют *ошибкой первого рода*, во втором случае – *ошибкой второго рода*.

Алгоритм построения критической области и правило принятия решения называют *статистическим критерием* данной гипотезы или просто критерием гипотезы. Наилучшим критерием для данной гипотезы является тот, который обеспечивает наименьшую величину вероятностей ошибок первого и второго рода. Обозначим вероятность ошибки первого рода через α, а вероятность ошибки второго рода – через β. Величину α называют ещё *уровнем значимости*. Доказано, что одновременное уменьшение этих ошибок невозможно: уменьшение одной из них неизбежно влечёт увеличение другой (при постоянном объёме выборки n). Поэтому при построении критической области проводится минимизация по одной из вероятностей ошибок, в то время как другая вероятность фиксируется на определённом значении. С другой стороны, стоимость ошибок или потери от них, не всегда являются одинаковыми. В связи с этим выбор значений величин α и β определяется в зависимости от конкретного содержания решаемой задачи.

Решение об истинности гипотезы H_0 принимается на основе сравнения некоторой вычисленной величины R с одним или двумя критическими значениями, являющимися границей или границами критической области. Если для критической области используется неравенство $R \geq R_в$, то говорят о *правосторонней* критической области, если неравенство $R \leq R_н$, то говорят о *левосторонней* критической области. Если, наконец, применяются оба неравенства, то критическую область называют *двухсторонней*. Отметим ещё, что схе-

ма построения критической области прямым образом связана со схемой построения доверительного интервала.

7.2. Проверка гипотез о параметрах распределений

Рассмотрим некоторые простые критерии, применяемые к гипотезам о параметрах случайных величин, таких, как математическое ожидание и дисперсия.

Сравнение математических ожиданий. Пусть имеется две выборки некоторой одной случайной величины, полученные в разных условиях, или двух разных случайных величин. Требуется проверить, одинаковы ли истинные математические ожидания, соответствующие этим выборкам. Такую ситуацию можно просмотреть на следующем примере. Имеется две производственные линии, выпускающие одинаковые изделия. Качество изготовленного изделия характеризуется случайной величиной X. Был осуществлён контроль n изделий, изготовленных на первой линии, и получена выборка $X_n = (x_1, x_2, \ldots, x_n)$. По второй линии осуществлён контроль над k изделиями и получена выборка $X'_k = (x'_1, x'_2, \ldots, x'_k)$. По этим выборкам нужно принять решение о том, одинаковые ли истинные математические ожидания величины X для этих линий.

Будем предполагать, что X имеет нормальное распределение: $N(m_1, \sigma_1^2)$ для первой линии и $N(m_2, \sigma_2^2)$ для второй линии. Рассмотрим сначала вариант, когда σ_1 и σ_2 известны. Выдвигаем гипотезу H_0: $m_1 = m_2$, т.е. математические ожидания одинаковы. Альтернативной гипотезой будет H_1: $m_1 \neq m_2$.

Оценка m_1^* будет иметь нормальное распределение с параметрами m_1 и σ_1^2/n, а оценка m_2^* – нормальное распределение с параметрами m_2 и σ_2^2/k. Если гипотеза H_0 верна, то разность $m_1^* - m_2^*$ будет распределена по нормальному закону с параметрами 0 и $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{k} = \frac{k\sigma_1^2 + n\sigma_2^2}{nk}$. Следовательно, величина

$$t = \frac{(m_1^* - m_2^*) \cdot \sqrt{nk}}{\sqrt{k\sigma_1^2 + n\sigma_2^2}} \qquad (7.1)$$

является центрированной и нормированной нормально распределённой случайной величиной. Найдём двухстороннюю критическую область, пользуясь этим фактом. Положим $P\{|t| \geq t_\alpha\} = \alpha$, или $1 - P\{|t| < t_\alpha\} = \alpha$, или $2\Phi(t_\alpha) = 1 - \alpha$, , где $\Phi(t_\alpha)$ – функция Лапласа. Таким образом, чтобы найти критическое значение t_α, нужно решить уравнение

$$\Phi(t_\alpha) = \frac{1-\alpha}{2}, \qquad (7.2)$$

при заданном уровне значимости α, пользуясь таблицей нормального распределения.

Критическая область определяется неравенством $|t| \geq t_\alpha$. Отсюда вытекает правило принятия решения: если вычисленное значение t удовлетворяет неравенству $-t_\alpha < t < t_\alpha$, то гипотеза H_0 принимается, если, напротив, $|t| \geq t_\alpha$, то гипотеза H_0 отвергается. Смысл этого правила состоит в том, что t имеет нормальное эталонное распределение только при равенстве истинных математических ожиданий. Для неравенства $|t| \geq t_\alpha$ задана достаточно малая вероятность α. Если оно реализуется, то это означает, что произошло маловероятное событие. Мы предполагаем, что где-то в наших рассуждениях допущена ошибка. Ошибка может быть только в одном, а именно, в предположении о том, что гипотеза H_0 истинна. Следовательно, гипотезу нужно отвергнуть. Заметим, что в таком случае мы совершаем ошибку с вероятностью α.

Исследуемая величина не обязательно должна иметь нормальное распределение. При больших n и k (порядка десятков) можно применять нормальное распределение, что оправдывается законом больших чисел. Если дисперсии одинаковы $\sigma_1^2 = \sigma_2^2 = \sigma^2$ и одинаковы объёмы выборок $n = k$, то вычисление t упрощается:

$$t = \frac{(m_1^* - m_2^*) \cdot \sqrt{n}}{\sqrt{2} \cdot \sigma}. \qquad (7.3)$$

Рассмотрим теперь случай, когда σ_1 и σ_2 неизвестны. Будем считать, что $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Вычислим оценки этих дисперсии по первой и второй выборкам S_1^2 и S_2^2:

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_1^*)^2, \quad S_2^2 = \frac{1}{k-1}\sum_{i=1}^{k}(x_i' - m_2^*)^2.$$

Если гипотеза H_0 верна, то разность $m_1^* - m_2^*$ будет распределена по нормальному закону с нулевым математическим ожиданием, а величина

$$t = \frac{(m_1^* - m_2^*) \cdot \sqrt{nk}}{S\sqrt{k+n}}, \qquad S = \sqrt{\frac{S_1^2(n-1) + S_2^2(k-1)}{n+k-2}}, \qquad (7.4)$$

будет распределена по закону Стьюдента с $n + k - 2$ степенями свободы. Таким образом, уравнение $P\{|t| \geq t_\alpha\} = \alpha$ следует решать по таблицам распределения Стьюдента.

Пример 7.1. На двух заводах выпускаются автомобильные шины одной и той же марки. Были исследованы данные по времени нормальной эксплуатации шин. По 20 шинам первого завода были вычислены $m_1^* = 36200$ (км. пробега) и $S_1^2 = 252400$. По 40 шинам второго завода были вычислены $m_2^* = 37800$ (км. пробега) и $S_2^2 = 326200$. С уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о том, что средний пробег шин, изготовленных на этих заводах, одинаков.

Вычислим S:

$$S = \sqrt{\frac{S_1^2(n-1) + S_2^2(k-1)}{n+k-2}} = \sqrt{\frac{252400 \cdot 19 + 326200 \cdot 39}{58}} \approx 550.$$

Далее вычислим t:

$$t = \frac{(m_1^* - m_2^*) \cdot \sqrt{nk}}{S\sqrt{k+n}} = \frac{(36200 - 37800) \cdot \sqrt{800}}{550\sqrt{60}} \approx -10{,}6.$$

Решим уравнение $P\{|t| \geq t_\alpha\} = 0{,}05$. По таблицам распределения Стьюдента с числом степеней свободы 58 находим $t_\alpha = 2$. Так как $|t| = 10, 6 > 2 = t_\alpha$, то гипотезу о равенстве среднего пробега шин, изготовленных на разных заводах, следует отвергнуть.

Сравнение дисперсий. Рассмотрим снова две выборки. Первая выборка $X_n = (x_1, x_2, \ldots, x_n)$ представляет измерения случайной величины X, вторая выборка $Y_k = (y_1, y_2, \ldots, y_k)$ представляет измерения случайной величины Y. Пусть вычислены оценки дисперсий:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2, \quad S_y^2 = \frac{1}{k-1}\sum_{i=1}^{k}(y_i - m_y^*)^2.$$

Требуется проверить гипотезу о том, что истинные дисперсии этих величин одинаковы.

Выдвигаем гипотезу H_0: $\sigma_x^2 = \sigma_y^2$, т.е. дисперсии одинаковы. Альтернативной гипотезой будет H_1: $\sigma_x^2 \neq \sigma_y^2$. Если величины X и Y имеют нормальное распределение, то отношение $F = S_x^2 / S_y^2$ будет иметь F-распределение (Фишера) со степенями свободы $n - 1$ и $k - 1$. Так как F-распределение не является симметричным, то можно построить двухсто-

роннюю критическую область, полагая $P\{F \le F_1\} = \alpha/2$ и $P\{F \ge F_2\} = \alpha/2$. Решая эти уравнения по таблице F-распределения, находим критические значения F_1 и F_2. Правило принятия решения: если $F_1 < F < F_2$, то гипотеза о равенстве дисперсий принимается, в противном случае эта гипотеза отвергается.

Удобнее строить одностороннюю критическую область. Для этого в отношении F в знаменателе всегда нужно ставить меньшую оценку из S_x^2 и S_y^2. Тогда F будет всегда больше единицы и нижняя граница не потребуется. Полагают

$$F = \frac{\max(S_x^2, S_y^2)}{\min(S_x^2, S_y^2)} \tag{7.5}$$

и решают уравнение

$$P\{F \ge F_\alpha) = \alpha, \tag{7.6}$$

используя таблицы F-распределения с учётом чисел степеней свободы. Если $F < F_\alpha$, то гипотеза о равенстве дисперсий принимается, если $F \ge F_\alpha$, то эта гипотеза отвергается.

Пример 7.2. По данным примера 7.1. с уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о равенстве истинных дисперсий пробега шин, изготовленных на первом и втором заводах. В примере 7.1. мы предполагали, что они одинаковы. Так ли это?

Так как оценки дисперсий уже вычислены, то строим отношение F по формуле (7.5):

$$F = \frac{326300}{252400} \approx 1{,}3.$$

Решаем по таблицам уравнение $P\{F \ge F_\alpha\} = 0{,}05$, учитывая, что число степеней свободы меньшей дисперсии равно 19, а большей дисперсии – 39. Находим критическое значение: $F_\alpha = 2{,}02$. Так как $F = 1{,}3 < 2{,}02 = F_\alpha$, то гипотезу о равенстве истинных дисперсий можно принять.

Данный критерий можно применить и для сравнения двух или нескольких математических ожиданий. В частности, этот критерий в таком варианте применяется в дисперсионном анализе.

7.3. Критерий квантилей

Пусть выборка $X_n = (x_1, x_2, \ldots, x_n)$ представляет измерения случайной величины X. Критерий квантилей используется для проверки гипотезы о том, что истинное распределение величины X имеет квантиль x_p порядка p, т.е. $F(x_p) = p$, где $F(x)$ – истинная функция распределения этой величины.

Здесь гипотеза H_0: число x_p удовлетворяет равенству $F(x_p) = p$, где p – заданная вероятность; H_1: это равенство не выполняется. Критическая область строится следующим образом. Находится число k элементов выборки, попавших в интервал $(-\infty, x_p)$. Далее рассматривается уравнение $P\{\left|\frac{k}{n} - p\right| \ge \varepsilon_\alpha\} = \alpha$, которое можно разделить на два уравнения: $P\{k \le np - \varepsilon_\alpha\} = \alpha/2$ и $P\{k \ge np + \varepsilon_\alpha\} = \alpha/2$. Так как k распределено по закону Бернулли, то определяются такие целые числа k_1 и k_2, для которых выполняются приближённые равенства $P\{k \le k_1\} \approx \alpha/2$ и $P\{k \ge k_2\} \approx \alpha/2$. Правило принятия решения: если $k_1 < k < k_2$, то гипотеза H_0 принимается, в противном случае гипотеза отвергается.

Рассмотренным критерием можно пользоваться только при малых объёмах выборки. При большом объёме выборки пользоваться распределением Бернулли неудобно. Поэтому пользуются тем фактом, что величина

$$t = \frac{k - np}{\sqrt{npq}},\ q = 1 - p, \tag{7.8}$$

при достаточно больших значениях n распределена приближённо по нормальному закону с нулевым математическим ожиданием и единичной дисперсией. Следовательно, по таблицам нормального распределения можно решить уравнение (7.2) и найти критическое значение t_α. Правило принятия решения обычное: если вычисленное значение t удовлетворяет неравенству $-t_\alpha < t < t_\alpha$, то гипотеза H_0 принимается, если, напротив, $|t| \geq t_\alpha$, то гипотеза H_0 отвергается.

Если в критерии квантилей положить $p = 0{,}5$, то получим критерий проверки гипотезы о медиане истинного распределения. Критерий квантилей удобен тем, что сами элементы выборки в их числовом выражении не используются. Используется лишь сам факт попадания элемента выборки в определённый интервал. Критерий квантилей применяется также при проверке некоторых других гипотез.

Пример 7.3. Сделано 30 измерений случайной величины X: 2,8; 2,6; 3,6; 3,2; 3,9; 3,4; 4,1; 3,7; 2,7; 3,9; 2,8; 3,8; 3,5; 3,1; 3,3; 3,5; 3,4; 3,7; 3,0; 3,3; 3,5; 3,7; 4,0; 3,6; 3,5; 3,1; 3,8; 3,5; 3,9; 3,6. С уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о том, что медианой истинного закона распределения является число 3,5.

В данном случае $n = 30$, $p = 0{,}5$, $q = 0,5$, $np = 15$. Число $k = 12$ – число элементов выборки, которые меньше, чем 3,5. Тогда $t = \frac{12 - 15}{\sqrt{30 \cdot 0{,}5 \cdot 0{,}5}} = -\frac{3}{2{,}74} \approx -1{,}1$. Решим уравнение $\Phi(t_\alpha) = (1 - 0{,}05)/2 = 0{,}475$. По таблицам находим: $t_\alpha = 1{,}96$. Так как $|t| < t_\alpha$, то гипотезу о том, что медиана может быть равной 3,5, принимаем.

7.4. Проверка гипотез о распределениях

Рассмотрим задачу *проверки гипотез о законе распределения*. Эмпирические законы распределения вероятностей имеют дискретный характер независимо от того, является ли эта величина дискретной или непрерывной. Использование такого закона в различных расчётах оказывается не всегда удобным. Поэтому возникает задача замены его некоторым теоретическим законом распределения, который был бы в определённом смысле близким к эмпирическому закону. Эта задача решается следующим образом. По виду гистограммы, полигона или графика эмпирической функции распределения по справочникам выбирается подходящий теоретический закон распределения и выдвигается гипотеза (H_0) о том, что именно этот выбранный закон является истинным законом распределения изучаемой величины. Затем по определённому критерию близости теоретического и эмпирического законов распределений выдвинутая гипотеза принимается или отвергается. Сами критерии близости могут быть различными, поэтому истинность выдвинутой гипотезы можно проверить различным образом.

Критерий χ^2-Пирсона. Одним из наиболее распространённых критериев является критерий χ^2-Пирсона. Вся широта эмпирического распределения разбивается на r частичных интервалов, и сравниваются вероятности попадания измеренных значений случайной величины в эти интервалы для случая эмпирического распределения и для случая теоретического распределения. В качестве меры отклонения берётся величина

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \qquad (7.9)$$

где n – объём выборки, n_i – число элементов, попавших в i-й интервал, p_i – вероятность попадания в i-тый интервал, вычисленная на основе теоретического распределения. Смысл величины χ^2 заключается в следующем. В качестве меры отклонения эмпирического распределения от теоретического в i-м интервале можно взять квадрат разности между эмпирической и теоретической вероятностью попадания изучаемой величины в этот интервал: $(p_i^* - p_i)^2 = \left(\frac{n_i}{n} - p_i\right)^2 = \frac{1}{n^2}(n_i - np_i)^2$. Если теперь суммировать квадраты отклонений по всем интервалам, то все отклонения будут учтены с одинаковым весом независимо от величины p_i. При одном и том же абсолютном отклонении относительная величина отклонения будет тем больше, чем меньше вероятность p_i. Целесообразно ввести весовые коэффициенты, обратно пропорциональные вероятностям p_i. Выбор коэффициентов в виде n/p_i приводит к формуле (7.9).

Вероятности p_i вычисляются по теоретической (гипотетической) функции распределения $F(x)$ или по соответствующей плотности $f(x)$:

$$p_i = F(z_i) - F(z_{i-1}) = \int_{z_{i-1}}^{z_i} f(t)dt,$$

где z_{i-1} и z_i – левая и правая границы i-ого интервала.

Оказывается, что если гипотеза верна, то величина χ^2 распределена по закону χ^2 с $k = r - 1$ степенями свободы. Этим фактом объясняется выбор структуры весового коэффициента. По таблицам распределения χ^2 решается уравнение относительно значения χ^2_α:

$$P\{\chi^2 \geq \chi^2_\alpha\} = \alpha, \qquad (7.10)$$

где α – малая величина, выбираемая в пределах 0,01 ÷ 0,1. Эта величина (*уровень значимости*) по смыслу является вероятностью отвергнуть истинную гипотезу, т.е. вероятностью ошибки первого рода. Величина χ^2_α – критическое значение величины χ^2. Она является нижней границей односторонней критической области $[\chi^2_\alpha; \infty)$ данного критерия. Правило принятия решения основывается на сравнении вычисленного значения χ^2 с критическим значением χ^2_α. Если $\chi^2 < \chi^2_\alpha$, то гипотеза о соответствии эмпирического и теоретического законов принимается, в противном случае, когда вычисленное χ^2 попадает в критическую область, эта гипотеза отвергается.

Идея этого критерия заключается в следующем. Величина χ^2 распределена по закону χ^2 только в том случае, когда гипотеза H_0 является истинной. Именно при этом условии было определено критическое значение χ^2_α таким образом, что вероятность выполнения неравенства $\chi^2 \geq \chi^2_\alpha$ очень мала (α). Предположим теперь, что это неравенство, тем не менее, выполнилось, т. е. вычисленное значение χ^2 попало в критическую область. Это значит, что произошло маловероятное событие. Если все наши рассуждения верны, то такой результат является достаточно редким. Поэтому мы подвергаем сомнению наши рассуждения. Мы могли сделать ошибку лишь тогда, когда предположили истинность проверяемой гипотезы H_0. Следовательно, эту гипотезу нужно отвергнуть. Принимая такое решение, мы можем совершить ошибку с вероятностью α, поскольку именно с этой вероятностью указанное маловероятное событие может произойти.

Если теоретическое распределение содержит неизвестные параметры, то перед вычислением вероятностей p_i, эти параметры (например, математическое ожидание, дисперсия и т.п.) оцениваются по той же самой выборке. В таком случае число степеней свободы k распределения χ^2 уменьшается: $k = r - s - 1$, где s – число оценённых параметров. Число частичных интервалов следует брать не менее 8, однако, если для некоторых интервалов при вычислениях оказывается $np_i < 5$, то их нужно объединять с соседними интервалами, чтобы было $np_i \geq 5$ (в ответственных случаях не допускается $np_i < 10$). Общее число интервалов при этом, естественно, сокращается.

Критерий ω^2. Недостатком критерия χ^2-Пирсона является то, что для его применения необходим большой объём выборки. Это связано с тем, что в критерии используется интервальный вариационный ряд, при получении которого имела место частичная потеря информации (за счёт группирования данных). В случае малых объёмов выборок целесообразно использовать критерии, непосредственно опирающиеся на измерения исследуемой величины. Наиболее распространённым критерием такого типа является критерий ω^2 (и его модификации). В этом критерии в качестве меры отклонения эмпирического распределения от теоретического применяется величина

$$\varpi^2 = \int_{-\infty}^{\infty} [F^*(x) - F(x)]^2 \, dF(x),$$

где $F^*(x)$ – эмпирическая функция распределения вероятностей; $F(x)$ – теоретическая (гипотетическая) функция. Интегрирование приводит к следующему выражению:

$$\varpi^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left[F(x_i) - \frac{2i-1}{2n}\right]^2, \qquad (7.11)$$

где x_i – элементы выборки. Доказано, что распределение величины $n\omega^2$ при возрастании n достаточно быстро сходится к некоторому предельному распределению, для которого составлена таблица.

Практически критерием можно пользоваться уже при $n > 40$, что выгодно отличает этот критерий от критерия χ^2. Схема принятия решения аналогична схеме, рассмотренной выше. Решается уравнение

$$P\{n\omega^2 \geq z_\alpha\} = \alpha, \qquad (7.12)$$

т.е. по заданному уровню значимости α по таблицам находится критическое значение z_α. Если оказывается, что $n\omega^2 < z_\alpha$, то гипотеза о соответствии эмпирического и теоретического распределений принимается, в противном случае – отвергается.

Пример 7.4. По выборке, заданной в примере 6.4, построить полигон и гистограмму. Подобрать подходящий теоретический закон распределения вероятностей и проверить гипотезу о соответствии эмпирического закона распределения выбранному теоретическому закону при уровне значимости $\alpha = 0{,}05$.

По данным примера 6.4 составим следующую таблицу:

i	1	2	3	4	5	6	7	8	$\sum$
$z_{i-1} \div z_i$	0÷0,6	0,6÷1,2	1,2÷1,8	1,8÷2,4	2,4÷3,0	3,0÷4,2	4,2÷5,4	5,4÷9,6	
Δ_i	0,6	0,6	0,6	0,6	0,6	1,2	1,2	4,2	
$\overline{x}_i$	0,3	0,9	1,5	2,1	2,7	3,6	4,8	7,5	
n_i	15	12	8	4	5	8	4	3	60
p_i^*	0,267	0,2	0,133	0,067	0,083	0,133	0,067	0,05	1
p_i^*/Δ_i	0,445	0,333	0,222	0,112	0,138	0,111	0,056	0,012	

В таблице используются следующие обозначения: Δ_i – ширина i–го интервала; p_i^* – эмпирическая вероятность попадания в i-й интервал. В последней строке представлены значения эмпирической плотности распределения вероятностей для частичных интервалов. По данным этой строки строим гистограмму и полигон распределения, которые в данном случае являются двумя различными графическими представлениями эмпирической плотности распределения вероятностей изучаемой величины (рис. 7.1).

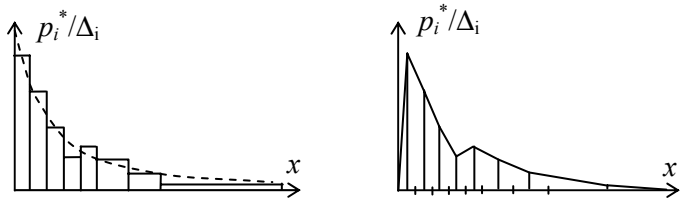

Рис. 7.1. **Гистограмма и полигон распределения, построенные по интервальному вариационному ряду**

Из графиков видно, что эмпирическое распределение очень похоже на теоретическое экспоненциальное распределение. В пользу выбора экспоненциального распределения в качестве гипотетического говорит и тот факт, что оценки математического ожидания и среднего квадратического отклонения достаточно близки. Известно, что в случае экспоненциального распределения эти параметры равны.

Для оценки параметра λ экспоненциального распределения $F(x) = 1 - e^{-\lambda x}$, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$, $\lambda > 0$, используем соотношение $m_x = 1/\lambda$, заменив математическое ожидание m_x его оценкой m_x^*, вычисленной в примере 6.4: $\lambda^* = 1/m_x^* = 1/2{,}033 = 0{,}492 \approx 0{,}5$.

Таким образом, гипотетическое распределение имеет вид:

$$F(x) = 1 - e^{-0,5x}, \quad f(x) = 0{,}5 \cdot e^{-0,5x}, \quad x \geq 0.$$

Теоретические вероятности попадания случайной величины в частичные интервалы вычисляются по формуле: $p_i = F(z_i) - F(z_{i-1})$, $i = 1, 2, \ldots, 8$. Например, $p_1 = F(0, 6) - F(0) = (1 - e^{-0,5 \cdot 0,6}) - (1 - e^{-0,5 \cdot 0}) = 1 - 0{,}741 = 0{,}259$. Тогда $np_1 = 60 \cdot 0{,}259 = 15{,}54$. Выполняя аналогичные вычисления для всех интервалов, внесём полученные результаты в таблицу. Правая граница последнего интервала принята равной ∞, так как в случае гипотетического распределения случайная величина может принимать значения в интервале (0; ∞).

В соответствии с теорией значения np_i не должны быть меньше пяти. В таблице это было учтено путём объединения 5-го интервала с 6-м, а 7-го – с 8-м. В двух последних столбцах проводится вычисление величины χ^2 непосредственно по формуле (7.9). Полученное значение равно $\chi^2 = 1{,}4306$.

i	$z_{i-1} \div z_i$	n_i	$F(z_i)$	$F(z_{i-1})$	p_i	np_i	$n_i - np_i$	$(n_i - np_i)^2/(np_i)$
1	0÷0, 6	16	0,259	0	0,259	15,54	0,46	0,0136
2	0,6÷1,2	12	0,451	0,259	0,192	11,52	0,48	0,02
3	0,6÷1,2	8	0,593	0,451	0,142	8,52	–0,52	0,0317
4	0,6÷1,2	4	0,699	0,593	0,106	6,36	–2,36	0,8757
5	0,6÷1,2	5	0,777	0,699	0,078	4,68	2,26	0,4756
6	0,6÷1,2	8	0,878	0,777	0,101	6,06		
7	0,6÷1,2	4	0,933	0,878	0,055	3,3	–0,32	0,014
8	0,6÷1,2	3	1,000	0,933	0,067	4,02		
Σ		60			1,000	60	0	1,4306

Так как окончательный расчёт величины χ^2 проводился по 6-ти интервалам и один параметр был оценён (λ) по этой же выборке, то число степеней свободы распределения χ^2 будет равно $k = 6 - 2 = 4$. По таблицам распределения χ^2 с 4 степенями свободы и с уровнем значимости $\alpha = 0{,}05$ находим критическое значение $\chi_\alpha^2 = 9{,}5$. Вычисленное значение χ^2 меньше критического. Следовательно, гипотеза о соответствии эмпирического и теоретического распределения принимается. Это означает, что можно рассматривать исследуемую величину как величину, распределённую по экспоненциальному закону с параметром $\lambda = 0{,}5$.

В заключение заметим, что принятие любой гипотезы вовсе не означает, что она действительно истинна. Факт принятия гипотезы означает лишь то, что данная выборка не противоречит этой гипотезе. Не следует также забывать, что при принятии или отклонении гипотезы мы можем совершить одну из ошибок: либо ошибку первого рода, либо ошибку второго рода. Поэтому при пополнении статистических данных (увеличении объёма выборок) гипотезу снова рекомендуется проверять.

Термины для запоминания

- Статистическая гипотеза; нулевая гипотеза; альтернативная гипотеза; критическая область; правосторонняя критическая область; левосторонняя критическая область; двухсторонняя критическая область; ошибка первого рода; ошибка второго рода; статистический критерий; критерий квантилей; критерий χ^2-Пирсона; критерий ω^2.

Список контрольных вопросов

1. Что называется статистической гипотезой?
2. Что такое альтернативная или конкурирующая гипотеза?
3. Что называется статистическим критерием?
4. Дайте определение ошибок первого и второго рода.
5. Что такое критическая область?
6. Поясните, чем отличается односторонняя и двухсторонняя критические области?
7. Какие законы распределения можно применить для построения критической области в случае проверки гипотезы о математических ожиданиях?
8. Какой закон распределения применяется для построения критической области в случае проверки гипотезы о дисперсиях?
9. Поясните, какая гипотеза проверяется с помощью критерия квантилей?
10. Какой критерий можно применить при проверке гипотезы о медиане?
11. Какова основная идея критерия χ^2-Пирсона проверки гипотез о законах распределения?
12. В чём принципиальная разница между критерием χ^2-Пирсона и критерием ω^2?

Вопрос для обсуждения

В каких ситуациях возможна проверка конкретной гипотезы статистическими методами?

ТЕМА 8.

Дисперсионный анализ

Студент должен
освоить:

- понятие фактора, уровней фактора, плана эксперимента, освоение схем однофакторного, двухфакторного анализа, трёхфакторного анализа (план «латинский квадрат»);

приобрести навыки:

- проведения дисперсионного анализа по имеющимся исходным данным.

Краткое содержание

Постановка задачи дисперсионного анализа. Однофакторный дисперсионный анализ. Двухфакторный дисперсионный анализ. Трёхфакторный дисперсионный анализ. План «латинский квадрат».

8.1. Постановка задачи дисперсионного анализа

В *дисперсионном анализе* устанавливается факт зависимости или независимости исследуемой случайной величины от одного или нескольких факторов. Сами факторы могут быть как количественными, так и качественными. Можно, например, поставить вопрос, зависит ли количество дорожно-транспортных происшествий (ДТП) от цвета автомобилей? Здесь фактор – цвет автомобиля – является качественным. Можно поставить другой вопрос: зависит ли количество ДТП от возраста водителя? Здесь фактор – возраст водителя – является количественным. В результате дисперсионного анализа может быть принято одно из двух решений: или «да», или «нет». Анализ можно проводить как по каждому фактору отдельно, так и по двум или нескольким факторам одновременно. В первом случае анализ называют *однофакторным дисперсионным анализом*, во втором – *двухфакторным* или *многофакторным*.

Общую постановку задачи дисперсионного анализа рассмотрим для случая, когда имеется три фактора. Пусть X – некоторая случайная величина, на которую могут влиять факторы A, B и C. Требуется выяснить, какие из этих факторов являются существенными в смысле влияния на величину X и какие – несущественными. Для ответа на этот вопрос все факторы квантуются в определённом смысле или, как принято говорить, разбиваются на уровни. Например, для фактора A – «цвет» – уровнями являются различные цвета: уровень A_1 – «красный», A_2 – «синий», A_3 – «зелёный» и т.д. Уровнями фактора B – «возраст» – могут быть различные непересекающиеся интервалы, например, уровень B_1 – возраст от 20 до 30 лет, уровень B_2 – от 30 до 40 и т.д.

Пусть фактор A имеет уровни A_1, A_2 A_3, …, A_n, фактор B – уровни B_1, B_2, B_3, …, B_n, фактор C – уровни C_1, C_2, C_3, …, C_n. Делаются измерения величины X при различных комбинациях уровней факторов, например, при комбинациях (A_1, B_1, C_1), (A_2, B_1, C_1), (A_1, B_1, C_2), (A_3, B_2, C_1) и т.д. Схема выбора комбинаций уровней факторов, при которых следует делать измерения, называется *планом эксперимента*. Если перебираются все возможные комбинации уровней факторов, то план эксперимента называют *полным факторным планом*. Предположим, что каждый фактор имеет по s уровней. Тогда число различных комбинаций уровней факторов будет равно s^3. Это число может оказаться большим, что приводит к большому объёму выборки. Чтобы сократить требуемый объём выборки, применяют другие планы эксперимента. Классическим примером такого плана является *план «латинский квадрат»*, который применяется для трёхфакторного анализа. Смысл этого плана состоит в том, что для измерений отбираются только такие комбинации уровней факторов, среди которых комбинации с одинаковыми парами уровней встречаются только один раз. Такой план позволяет сократить минимальное число измерений с s^3 до s^2.

Измерения, сделанные при комбинации уровней факторов (A_i, B_j, C_k), будем обозначать через x_{ijkr}, где r – номер измерения при данной комбинации. В дисперсионном анализе предполагается, что все измерения являются статистически независимыми и распределёнными по нормальному закону с одной и той же дисперсией σ^2. Предполагается также, что

$$x_{ijkr} = m_{ijk} + \xi_{ijkr},$$

где m_{ijk} – математическое ожидание величины X при комбинации уровней (A_i, B_j, C_k); ξ_{ijkr} – нормально распределённая случайная величина с нулевым математическим ожиданием и дисперсией σ^2. Требуется проверить гипотезу H_0 о том, что все математические ожидания одинаковы, т. е. $m_{ijk} = m$ для всех значений i, j, k. Эта гипотеза эквивалентна предположению, что ни один из факторов не влияет на величину X.

Альтернативной или конкурирующей гипотезой H_1 является предположение, что хотя бы один фактор влияет на величину X. Принимая решение H_0, мы отвергаем влияние на X одновременно всех факторов. Принимая решение H_1, мы признаём, что на величину X оказывает существенное влияние хотя бы один фактор. Сделать вывод более детально можно только после дополнительного анализа на отсутствие взаимодействия между факторами. Если такое взаимодействие имеет место, то может оказаться, что каждый фактор индивидуально не влияет на величину X, но за счёт их взаимодействия влияние факторов на X в совокупности будет существенным. Если взаимодействия между факторами нет, то решение о влиянии на величину X можно принимать по каждому фактору независимо.

8.2. Однофакторный дисперсионный анализ

Пусть проверяется гипотеза H_0 о том, что фактор A не влияет на величину X. Предположим, что фактор A имеет k уровней: A_1, A_2, …, A_k. На каждом уровне фактора сделаем по n измерений. Получим совокупность измерений x_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$. Обозначим через m_i математическое ожидание величины X при i-м уровне фактора A. Общее математическое ожидание величины X обозначим через m. На основе экспериментальных данных требуется проверить гипотезу H_0 о равенстве всех математических ожиданий:

$$H_0: m_1 = m_2 = \ldots = m_k = m. \tag{8.1}$$

Для проверки гипотезы вычисляются оценки математических ожиданий на всех уровнях фактора и оценка общего математического ожидания:

$$m_i^* = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad i = 1, 2, \ldots, k; \qquad m^* = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} = \frac{1}{k}\sum_{i=1}^{k} m_i^*. \tag{8.2}$$

Зависимость исследуемой величины от фактора A может обнаружиться при сравнении оценок математических ожиданий m_i^* (для всех i) с оценкой общего математического ожидания m^*. Однако такое сравнение в дисперсионном анализе делается не непосредственно, а с помощью эмпирических дисперсий. Именно поэтому такой анализ называется дисперсионным.

Опуская коэффициент $1/(kn)$, который фигурирует в оценке дисперсии величины X, рассмотрим сумму квадратов отклонений измеренных значений этой величины от общего эмпирического математического ожидания:

$$kn \cdot D_x^* = kn \cdot \frac{1}{kn} \cdot \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*]^2 = Q = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*]^2.$$

Идея дисперсионного анализа состоит в том, что эта сумма разбивается на две компоненты, одна из которых обусловлена влиянием фактора A, другая – всеми другими, неконтролируемыми, факторами. Проведём такое разбиение:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*]^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}[(x_{ij} - m_i^*) + (m_i^* - m^*)]^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m_i^*]^2 + \sum_{i=1}^{k}\sum_{j=1}^{n} 2 \cdot (x_{ij} - m_i^*) \cdot (m_i^* - m^*) + \sum_{i=1}^{k}\sum_{j=1}^{n}[m_i^* - m^*]^2.$$

Учитывая, что второе слагаемое равно нулю (в силу независимости измерений величины X), получим:

$$Q = Q_1 + Q_2, \tag{8.3}$$

где $Q = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*]^2$, $Q_1 = n\sum_{i=1}^{k}[m_i^* - m^*]^2$, $Q_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m_i^*]^2$. Здесь в сумме Q_1 проведено суммирование по индексу j.

Компонента Q_1 характеризует отклонение средних значений величины X на разных уровнях фактора от общего среднего значения, т.е. характеризует влияние фактора A на рассматриваемую величину. Величина Q_1 называется *рассеиванием по фактору*. Компонента Q_2 указывает на рассеяние случайной величины внутри уровней фактора, т. е. эта компонента связана с влиянием на величину X некоторых других факторов. Величина Q_2 называется *остаточным рассеиванием*. Сравнивая эти две компоненты, можно сделать определённые выводы относительно того, существенно ли влияние фактора A.

Очевидно, что сравнение указанных компонент должно проводиться на основе некоторого критерия. Так как величина X по условию имеет нормальное распределение с дисперсией σ^2, то величина Q/σ^2 будет иметь распределение χ^2 с $(nk - 1)$ степенью свободы. С другой стороны, все величины m_i^* ($i = 1, 2, \ldots, k$) также будут распределены по нормальному закону с одинаковой дисперсией σ^2/n. Предположим теперь, что гипотеза H_0 верна. Тогда величина Q_1/σ^2 будет распределена по закону χ^2 с $(k - 1)$ степенью свободы. Опираясь на свойства распределения χ^2, можно сделать вывод, что отношение Q_2/σ^2 также распределено по закону χ^2 с числом степеней свободы, равным разности $(nk - 1) - (k - 1) = k(n - 1)$. Следовательно, отношение F, определяемое формулой

$$F = \frac{Q_1/(k-1)}{Q_2/[k(n-1)]}, \tag{8. 4}$$

будет иметь F-распределение (Фишера – Снедекора) со степенями свободы $k - 1$ и $k(n - 1)$.

Далее проверка гипотезы проводится обычным образом. Задаётся уровень значимости α (вероятность ошибки первого рода) и по таблицам F-распределения решается уравнение

$$P\{F \geq F_\alpha\} = \alpha. \tag{8.5}$$

Из этого уравнения находится критическое значение F_α, которое является нижней границей критической области значений F. Правило принятия решения следующее: если F – вычисленное значение по формуле (8.4), то при $F < F_\alpha$ гипотеза H_0 принимается, а при $F \geq F_\alpha$ гипотеза отвергается. Смысл этого правила очевиден. Если фактор A не оказывает существенного влияния на X, т.е. гипотеза H_0 истинна, то величина Q_1 должна быть относительно малой по отношению к величине Q_2. Если же отношение F оказывается большим, то влияние фактора A следует признать существенным.

При «ручных» вычислениях удобнее пользоваться следующими формулами:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - n \cdot k \cdot m^{*2}, \; Q_1 = n\left(\sum_{i=1}^{k} m_i^{*2} - k \cdot m^{*2}\right), \; Q_2 = Q - Q_1 . \tag{8.6}$$

Сами вычисления обычно проводят в таблицах. Один из вариантов такой таблицы приводится в примере 8.1.

Пример 8.1. *Исследуется* зависимость процентного содержания брака (величина X) среди изделий, изготовленных за единицу времени, от температуры окружающей среды (фактор A). Был произведён подсчёт количества бракованных изделий для пяти интервалов времени при трёх различных температурах окружающей среды. Результаты измерений представлены в следующей таблице.

Процент брака при повышенной температуре	2,5	3,3	2,4	3,0	2,6
Процент брака при нормативной температуре	2,4	3,2	2,2	2,7	2,3
Процент брака при пониженной температуре	2,6	3,4	3,0	3, 1	2,8

Методом дисперсионного анализа проверить гипотезу о влиянии температуры среды на процентное содержание брака среди изготовленных изделий.

Пусть уровень фактора A_1 соответствует повышенной температуре, A_2 – нормативной температуре, A_3 – пониженной температуре. Все вычисления проведём по формулам (8.3) в одной таблице.

A_i	x_{ij}					m_i^*	x_{ij}^2					m_i^{*2}
A_1	2,5	3,3	2,4	3,0	2,6	2,76	6,25	10,89	5,76	9,0	6,76	7,6176
A_2	2,4	3,2	2,2	2,7	2,3	2,56	5,76	10,24	4,84	7,29	5,29	6,5536
A_3	2,6	3,4	3,0	3,1	2,8	2,98	6,76	11,56	9,0	9,61	7,84	8,8804
Σ						8,3	18,77	32,69	19,6	25,9	19,89	23,052

В последней строке таблицы записаны суммы элементов столбцов. Таким образом,

$$m^* = \frac{8,3}{3} \approx 2,7667,\ m^{*2} \approx 7,6546,\ Q = 18,77 + 32,69 + 19,6 + 25,9 + 19,89 - 3 \cdot 5 \cdot 7,6546 \approx 2,03,$$

$$Q_1 = 5(23,052 - 3 \cdot 7,6546) \approx 0,439,\quad Q_2 = 2,03 - 0,439 = 1,591,\quad F = \frac{0,439/2}{1,591/12} \approx 1,65.$$

По таблице F-распределения при степенях свободы $k_1 = 2$ и $k_2 = 12$ и уровне значимости $\alpha = 0,05$ находим критическое значение $F_\alpha = 3,88$. Так как вычисленное значение $F = 1,65$ меньше критического значения F_α, то принимается гипотеза H_0. Результаты анализа показывают, что влияние температурного режима на процентное содержание брака в готовой продукции не является существенным.

8.3. Двухфакторный дисперсионный анализ

Двухфакторный анализ проводится аналогично. По каждому фактору вычисляется своё отношение F и находится своё критическое значение F_α. Отличием схемы двухфакторного анализа от однофакторного является то, что в двухфакторном анализе появляется необходимость проверки гипотезы об отсутствии взаимодействия между факторами.

Пусть проверяется гипотеза о влиянии факторов A и B на величину X. Предположим, что фактор A имеет k уровней, а фактор B – n уровней. Измерение величины X на i-ом уровне фактора A и на j-ом уровне фактора B будем обозначать x_{ijr}, где r – номер измерения на данной комбинации уровней. Отбросим временно третий индекс, считая, что на каждой комбинации уровней факторов сделано по одному измерению. Введём обозначения:

$$m_{i+}^* = \frac{1}{n}\sum_{j=1}^{n} x_{ij},\ i = 1,2,...,k;\quad m_{+j}^* = \frac{1}{k}\sum_{i=1}^{k} x_{ij},$$
$$j = 1,2,...,n;\ m^* = \frac{1}{nk}\sum_{j=1}^{n}\sum_{i=1}^{k} x_{ij},\ i = 1,2,...,k. \qquad (8.7)$$

Здесь m^*_{i+} есть эмпирическое математическое ожидание величины X на i-ом уровне фактора A, m^*_{+j} – на j-ом уровне фактора, m* – общее эмпирическое математическое ожидание величины. Пусть m_{i+}, m_{+j}, m – соответствующие истинные (неизвестные) математические ожидания величины X. Тогда основная гипотеза H_0 заключается в равенстве всех истинных математических ожиданий:

$$H_0: \; m_{1+} = m_{2+} = ... = m_{k+} = m_{+1} = m_{+2} = ... = m_{+n} = m.$$

Для проверки этой гипотезы сумма квадратов отклонений измерений от общего эмпирического математического ожидания разбивается на три компоненты:

$$\begin{gathered} Q = Q_1 + Q_2 + Q_3, \text{ где } \; Q = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*]^2, \; Q_1 = n \cdot \sum_{i=1}^{k}[m^*_{i+} - m^*]^2, \\ Q_2 = k \cdot \sum_{i=1}^{n}[m^*_{+j} - m^*]^2, \quad Q_3 = \sum_{i=1}^{k}\sum_{j=1}^{n}[x_{ij} - m^*_{i+} - m^*_{+j} + m^*]^2. \end{gathered} \tag{8.8}$$

Компонента Q_1 характеризует влияние фактора A на рассматриваемую величину. Величина Q_1 называется *рассеиванием по фактору A*. Компонента Q_2 характеризует влияние фактора B на рассматриваемую величину и называется *рассеиванием по фактору B*. Компонента Q_3 указывает на рассеяние случайной величины внутри комбинаций уровней факторов, т. е. эта компонента связана с влиянием на величину X некоторых других факторов. Величина Q_3 называется *остаточным рассеиванием*.

Так как величина X по условию имеет нормальное распределение с дисперсией σ^2, то величина Q/σ^2 будет иметь распределение χ^2 с $(nk - 1)$ степенью свободы. Все величины m^*_{i+} (i = 1, 2, …, k) также будут распределены по нормальному закону с одинаковой дисперсией σ^2/n, а величины m^*_{+j} (j = 1, 2, …, n) – по нормальному закону с дисперсией σ^2/k. Предположим теперь, что гипотеза H_0 верна. Тогда величина Q_1/σ^2 будет распределена по закону с χ^2 $(k - 1)$ степенью свободы, а величина Q_2/σ^2 будет распределена по закону с χ^2 $(n - 1)$ степенью свободы. Опираясь на свойства распределения χ^2, можно сделать вывод, что отношение Q_3/σ^2 также распределено по закону χ^2 с числом степеней свободы, равным разности $(nk - 1) - (k - 1) - (n - 1) = nk - k - n + 1 = (k - 1)(n - 1)$. Следовательно, отношение F_1, определяемое формулой

$$F_1 = \frac{Q_1/(k-1)}{Q_3/[(k-1)(n-1)]} = \frac{Q_1 \cdot (n-1)}{Q_3}, \tag{8.9}$$

будет иметь F-распределение (Фишера – Снедекора) со степенями свободы $k - 1$ и $(k - 1)$–$(n - 1)$, а отношение F_2, определяемое формулой

$$F_2 = \frac{Q_2/(n-1)}{Q_3/[(k-1)(n-1)]} = \frac{Q_2 \cdot (k-1)}{Q_3}, \tag{8.10}$$

будет иметь F-распределение со степенями свободы $n - 1$ и $(k - 1)(n - 1)$.

Далее задаётся уровень значимости α и по таблицам F-распределения дважды решается уравнение $P\{F \geq F_\alpha\} = \alpha$. По степеням свободы $k - 1$ и $(k - 1)(n - 1)$ определяется критическое значение $F_{\alpha 1}$, а по степеням свободы $n - 1$ и $(k - 1)(n - 1)$ – критическое значение $F_{\alpha 2}$. Правило принятия решения следующее: если $F_1 < F_{\alpha 1}$ и $F_2 < F_{\alpha 2}$ то гипотеза H_0 принимается. Это означает, что оба фактора не оказывают существенного влияния (при заданном уровне значимости) на исследуемую величину. В случае, когда взаимодействие между факторами отсутствует, решение можно принимать по каждому фактору индивидуально. Если $F_1 \geq F_{\alpha 1}$ и $F_2 < F_{\alpha 2}$, то фактор A влияет, а фактор B не влияет на величину X. При $F_1 < F_{\alpha 1}$ и $F_2 \geq F_{\alpha 2}$, напротив, фактор A не влияет, а фактор B влияет на величину X. Наконец, если

$F_1 \geq F_{\alpha 1}$ и $F_2 \geq F_{\alpha 2}$, то нужно признать влияние на X обоих факторов. Не следует забывать, что эти выводы имеют вероятностный характер, так как имеют место вероятности ошибок первого и второго рода.

При «ручных» вычислениях удобнее пользоваться следующими формулами:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - n \cdot k \cdot m^{*2}, \quad Q_1 = n(\sum_{i=1}^{k} m_{i+}^{*2} - k \cdot m^{*2}),$$
$$Q_2 = k(\sum_{i=1}^{n} m_{+j}^{*2} - n \cdot m^{*2}), \quad Q_3 = Q - Q_1 - Q_2. \tag{8.11}$$

Для проверки гипотезы об отсутствии взаимодействия между факторами необходимо иметь по нескольку измерений на каждой комбинации уровней факторов. Предположим, что на каждой комбинации уровней факторов сделано по t измерений. Тогда рассматривается обычная задача двухфакторного анализа, в которой вместо x_{ij} берутся величины x_{ij+}:

$$x_{ij+} = \frac{1}{t}\sum_{r=1}^{t} x_{ijr}, \; i = 1,2,...,k; \; j = 1,2,...,n.$$

Гипотеза об отсутствии взаимодействия проверяется следующим образом. Вычисляется дополнительное отношение F_3:

$$F_3 = \frac{Q_3 /[(k-1)(n-1)]}{Q_4 /[kn \cdot (t-1)]},$$

где $Q_4 = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{r=1}^{t}(x_{ijr} - x_{ij+})^2$.

Отношение F_3 будет иметь F-распределение со степенями свободы $(k - 1)(n - 1)$ и $kn(t - 1)$. Как обычно, по уровню значимости α и данным степеням свободы определяется критическое значение $F_{\alpha 3}$. Если $F_3 < F_{\alpha 3}$, то взаимодействие между факторами не является существенным. Это означает, что решение о влиянии факторов можно принимать индивидуально по каждому фактору. Если $F_3 \geq F_{\alpha 3}$, то взаимодействие между факторами признаётся существенным. В таком случае нужно провести однофакторный анализ по каждому фактору отдельно, затем учесть, что в совокупности эти факторы также оказывают влияние на исследуемую величину.

8.4. Трёхфакторный дисперсионный анализ. План «латинский квадрат»

Рассмотренную схему проведения дисперсионного анализа можно применять и для большего числа факторов. Однако с ростом числа факторов резко повышается требуемый объём измерений случайной величины. В связи с этим иногда отказываются от полного факторного плана и применяют тот план эксперимента, который позволяет сократить число измерений. Рассмотрим одним из таких планов – план «латинский квадрат», применяемый в трёхфакторном анализе.

Имеется случайная величина X, на которую, возможно, влияют три фактора A, B и C. Пусть каждый фактор имеет n уровней. Тогда для реализации полного факторного плана потребуется не менее чем n^3 измерений. Идея плана «латинский квадрат» состоит в следующем. Измерения делаются по такой схеме, в которой каждая пара уровней раз-

личных факторов встречается только один раз. Представим таблицей такую схему для $n = 3$:

схема измерений $\begin{pmatrix} A_1B_1C_1 & A_1B_2C_2 & A_1B_3C_3 \\ A_2B_1C_2 & A_2B_2C_3 & A_2B_3C_1 \\ A_3B_1C_3 & A_3B_2C_1 & A_3B_3C_2 \end{pmatrix}$, матрица $CK = \begin{pmatrix} C_1 & C_2 & C_3 \\ C_2 & C_3 & C_1 \\ C_3 & C_1 & C_2 \end{pmatrix}$.

Здесь любая пара уровней встречается только один раз. Строки таблицы соответствуют уровням фактора *A*, столбцы – уровням фактора *B*. Для удобства чтения таблицы измерений можно рядом записать ключевую матрицу СК для уровней фактора *C*. Такой план эксперимента позволяет сократить число измерений с n^3 до n^2 (с 27 до 9 при $n = 3$).

Измерения величины *X* индексируются двумя индексами: по факторам *A* и *B*. Принадлежность измерения одному из уровней фактора *C* определяется по матрице СК:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}.$$

Эмпирические математические ожидания находятся по формулам:

$$m^*_{i+} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad m^*_{+j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad m^*_t = \frac{1}{n}\sum_{t} x_{ij}, \quad m^* = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}. \tag{8.12}$$

Здесь m^*_{I+} – эмпирическое математическое ожидание на *i*-ом уровне фактора *A*, m^*_{+j} – то же на *j*-ом уровне фактора *B*, m^*_t – на *t*-ом уровне фактора *C*, m^* – общее эмпирическое математическое ожидание. Все индексы изменяются от 1 до *n*.

Далее процедура анализа аналогична схеме двухфакторного анализа. Приведём соответствующие формулы:

$$Q = Q_1 + Q_2 + Q_3 + Q_4, \text{ где } Q = \sum_{i=1}^{n}\sum_{j=1}^{n}[x_{ij} - m^*]^2, \quad Q_1 = n \cdot \sum_{i=1}^{n}[m^*_{i+} - m^*]^2,$$
$$Q_2 = n \cdot \sum_{j=1}^{n}[m^*_{+j} - m^*]^2, \quad Q_3 = n \cdot \sum_{t=1}^{n}[m^*_t - m^*]^2, \quad Q_4 = \sum_{i=1}^{n}\sum_{j=1}^{n}[x_{ij} - m^*_{i+} - m^*_{+j} - m^*_t + 2m^*]^2. \tag{8.13}$$

После вычисления этих сумм находятся три отношения:

$$F_1 = \frac{Q_1 \cdot (n-2)}{Q_4}, \quad F_2 = \frac{Q_2 \cdot (n-2)}{Q_4}, \quad F_3 = \frac{Q_3 \cdot (n-2)}{Q_4}. \tag{8.14}$$

Число степеней свободы одинаково для Q_1, Q_2, Q_3, и равно $(n - 1)$. У суммы Q_4 имеется $(n - 1)(n - 2)$ степеней свободы. Таким образом, критическое значение F_α находится по таблицам F-распределения для $(n - 1)$ и $(n - 1)(n - 2)$ степеней свободы при заданном уровне значимости α, как решение уравнения (8.5).

Решение принимается точно так же, как в двухфакторном анализе. Если выполняются неравенства $F_1 < F_\alpha$, $F_2 < F_\alpha$, $F_3 < F_\alpha$, то гипотеза H_0 принимается. Это значит, что влияние всех трёх факторов на исследуемую величину признаётся не существенным.

Если хотя бы одно из указанных неравенств не выполняется, то для принятия решения по каждому фактору индивидуально необходима полная уверенность в том, что взаимодействия между факторами нет. Если такой уверенности нет, то придётся дополнительно проверять гипотезу об отсутствии взаимодействия между факторами. Эта процедура аналогична той, которая была рассмотрена в схеме двухфакторного анализа.

Термины для запоминания

- Дисперсионный анализ; однофакторный дисперсионный анализ; двухфакторный и многофакторный дисперсионный анализ; план эксперимента; полный факторный план; план «латинский квадрат»; рассеивание по фактору; остаточное рассеивание.

Список контрольных вопросов

1. Какую задачу решает дисперсионный анализ?
2. Как понимать термины «однофакторный анализ» и «многофакторный анализ»?
3. Что такое «план эксперимента» в дисперсионном анализе?
4. В чём заключается идея полного факторного плана и плана «латинский квадрат»?
5. Какой критерий проверки гипотезы используется в дисперсионном анализе?
6. Как сравниваются математические ожидания исследуемой величины на разных уровнях факторов?
7. Что называется рассеиванием по фактору?
8. Что называется остаточным рассеиванием?
9. Объясните правило принятия решения в однофакторном дисперсионном анализе.
10. Объясните правило принятия решения в двухфакторном дисперсионном анализе.
11. Как проверить гипотезу об отсутствии взаимодействия между факторами?
12. На что указывает ключевая матрица в плане «латинский квадрат» дисперсионного анализа?

Вопрос для обсуждения

Предположим, что вас попросили выяснить с помощью статистических методов, влияет ли цвет автомобилей на количество автомобильных аварий. Как бы вы решали эту задачу?

ТЕМА 9.

Регрессионный анализ

Студент должен
освоить:

- методику построения функций, аппроксимирующих статистические зависимости между величинами, и схем проведения регрессионного анализа;

приобрести навыки:

- проведения регрессионного анализа: линейного, нелинейного, одномерного, многомерного.

Краткое содержание

Постановка задачи регрессионного анализа. Линейный и нелинейный одномерный регрессионный анализ. Многомерный линейный регрессионный анализ.

9.1. Постановка и схема решения задачи регрессионного анализа

Задача регрессионного анализа – это задача об *аппроксимации*, т.е. приближённой замене, статистической связи между величинами функциональной связью. В регрессионный анализ входят и сопутствующие задачи, связанные с оценкой точности аппроксимации, её достоверности, а также задачи по выявлению тех переменных, которые существенным образом влияют на исследуемую случайную величину.

Рассмотрим общую постановку задачи регрессионного анализа. Пусть на величину Z могут (предположительно) влиять величины X, Y, …, T. Имеются и другие величины (или факторы), влияющие на величину Z, которые не являются контролируемыми. В связи с этим точное вычисление (прогнозирование) значений величины Z по известным значениям величин X, Y, …, T оказывается невозможным. В регрессионном анализе предполагается, что для возможных значений указанных величин выполняется соотношение

$$z = \varphi(x, y, \ldots, t) + \xi, \qquad (9.1)$$

где $\varphi(x, y, \ldots, t)$ – неизвестная функция, а ξ – случайная компонента, обусловленная неучтёнными факторами и независящая от переменных x, y, …, t. Предполагается, что случайная компонента ξ имеет нормальное распределение с нулевым математическим ожиданием и дисперсией σ_ξ^2. Приближённое определение функции $\varphi(x, y, \ldots, t)$ по экспериментальным данным и является основной задачей регрессионного анализа.

Заметим, что аргументы искомой функции могут быть либо случайными, либо не случайными величинами. Здесь существенно то, что неконтролируемая компонента является случайной величиной. Эта особенность отличает задачу регрессионного анализа от других задач подобного типа, например, от задачи интерполяции.

В п. 3.6 уже говорилось о том, что лучшей функцией для прогнозирования значений одной случайной величины по значениям другой величины является регрессия. Это положение справедливо и для случая, когда у зависимой переменной есть не один, а несколько аргументов. Действительно, если бы была известна условная плотность распределения $f(z \mid x, y, \ldots, t)$, то можно было бы определить условное математическое ожидание величины Z по отношению к величинам X, Y, …, T:

$$m_z(x, y, ..., t) = M\{Z \mid X, Y, ..., T\} = \int_{-\infty}^{\infty} z \cdot f(z \mid x, y, ..., t)dz .$$

Эта функция была бы наилучшей в смысле минимума среднего квадрата отклонения истинных значений величины z от прогнозируемых. На практике получить условный закон распределения вероятностей невозможно, поскольку исследователь располагает только измерениями случайных величин, не имея никакой дополнительной информации. В связи с этим, функция $\varphi(x, y, \ldots, t)$ может быть установлена только приближённо. Эту функцию называют *аппроксимирующей функцией* или *прогнозирующей функцией*.

Задача регрессионного анализа условно разбивается на два этапа. На первом этапе выбирается структура аппроксимирующей функции с точностью до неизвестных параметров. Этот выбор делается на основе некоторых известных физических, химических, экономических (и т.п.) свойств рассматриваемого процесса или на основе иных соображе-

ний. В частности, если изучается зависимость между двумя величинами, т.е. ищется аппроксимирующая функция $z = \varphi(x)$, то можно нанести экспериментальные точки (x_i, y_i), $i = 1, 2, \ldots, n$, на координатную плоскость и по характеру расположения этих точек сделать предположения о структуре аппроксимирующей функции. При полном отсутствии информации о структуре аппроксимирующей функции в качестве такой функции выбирают многочлен от её аргументов. Степень многочлена может быть уточнена при дальнейшем исследовании.

На втором этапе определяют неизвестные коэффициенты аппроксимирующей функции из условия минимума среднего квадрата ошибки прогнозирования. Делается это следующим образом. Пусть имеется n независимых измерений соответствующих значений величин $Z, X, Y, \ldots, T$:

$$\begin{matrix} z_1 & x_1 & y_1 & \ldots & t_1 \\ z_2 & x_2 & y_2 & \ldots & t_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ z_n & x_n & y_n & \ldots & t_n \end{matrix}$$

Тогда для каждого z_i, $i = 1, 2, \ldots, n$, можно записать:

$$z_i = \varphi(x_i, y_i, \ldots, t_i, a, b, \ldots, h) + \xi_i, \tag{9.2}$$

где $\varphi(x_i, y_i, \ldots, t_i, a, b, \ldots, h)$ – функция, структура которой была выбрана на первом этапе с точностью до неизвестных коэффициентов $a, b, \ldots, h$. Неконтролируемая компонента ξ_i представляет собой *ошибку аппроксимации*. Таким образом, средний квадрат ошибки аппроксимации будет равен:

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 = \frac{1}{n}\sum_{i=1}^{n}[z_i - \varphi(x_i, y_i, \ldots, t_i)]^2 = \frac{1}{n}\cdot S. \tag{9.3}$$

Коэффициент $1/n$ не влияет на процесс минимизации. Поэтому обычно минимизируют сумму S. Необходимым условием экстремума функции нескольких переменных является равенство нулю частных производных по этим переменным. Легко видеть, что величина S имеет минимум и не имеет максимума. Поэтому для определения точки экстремума (минимума) достаточно решить систему уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum\limits_{i=1}^{n}[z_i - \varphi(x_i, y_i, \ldots, t_i, a, b, \ldots, h)]\cdot\dfrac{\partial\varphi}{\partial a} = 0, \\ \dfrac{\partial S}{\partial b} = -2\sum\limits_{i=1}^{n}[z_i - \varphi(x_i, y_i, \ldots, t_i, a, b, \ldots, h)]\cdot\dfrac{\partial\varphi}{\partial b} = 0, \\ ---------- \\ \dfrac{\partial S}{\partial h} = -2\sum\limits_{i=1}^{n}[z_i - \varphi(x_i, y_i, \ldots, t_i, a, b, \ldots, h)]\cdot\dfrac{\partial\varphi}{\partial h} = 0. \end{cases} \tag{9.4}$$

Пусть значения $a^*, b^*, \ldots, h^*$ есть решение данной системы уравнений. Тогда функция

$$\tilde{z} = \varphi(x, y, \ldots, t, a^*, b^*, \ldots, h^*) \tag{9.5}$$

будет наилучшей аппроксимирующей функцией в смысле минимума среднего квадрата ошибки прогнозирования (наилучшей для выбранной структуры).

Точность аппроксимации определяется *остаточной дисперсией*. Смысл остаточной дисперсии рассмотрен в п. 3.6. Если в качестве функции $\varphi(x, y, \ldots, t)$ выбрать условное математическое ожидание $m_z(x, y, \ldots, t)$, то дисперсия величины Z будет равна

$$D_z = D_\varphi + D_\xi, \tag{9.6}$$

где D_φ – дисперсия контролируемой компоненты $\varphi(x, y, \ldots, t)$ из (9.1); D_ξ – дисперсия неконтролируемой компоненты ξ. При прогнозировании значений z по формуле (9.5) неопределённость, связанная со случайным характером первой компоненты в (9.1), снимается, а неоп-

ределённость, обусловленная случайным характером неконтролируемой компоненты ξ, остаётся. Поэтому дисперсию D_ξ называют остаточной дисперсией. Точно определить эту дисперсию невозможно, поскольку неизвестно условное математическое ожидание $m_z(x, y, \ldots, t)$. Для аппроксимирующей функции (9.5) остаточная дисперсия определяется формулой:

$$D^*_{\text{ост}} = \frac{1}{n-k}\sum_{i=1}^{n}[z_i - \tilde{z}_i]^2 . \qquad (9.7)$$

В этой формуле число k является числом оценённых параметров аппроксимирующей функции. Эти параметры были оценены на основе одной и той же выборки измерений. Поэтому число независимых слагаемых в сумме (9.7) будет равно $n - k$, а не n, как в (9.3).

Предположим, что получена некоторая аппроксимирующая функция типа (9.5). Тогда для этой функции можно записать равенство: $D_z^* = D_\varphi^* + D^*_{\text{ост}}$. При этом $D^*_{\text{ост}}$ не будет минимальной, так как аппроксимирующая функция является всего лишь приближённым выражением регрессии. Тем не менее, можно ввести некоторый коэффициент связи, аналогичный корреляционному отношению, который характеризует силу связи между величиной Z и всеми аргументами аппроксимирующей функции в совокупности:

$$R^2 = \frac{D^*_\varphi}{D^*_z} = 1 - \frac{D^*_{\text{ост}}}{D^*_z} . \qquad (9.8)$$

Если аппроксимирующая функция имеет всего один аргумент, например X, то величина R^2 будет являться оценкой корреляционного отношения $\eta^2_{z/x}$. Если, кроме того, имеет место только линейная связь, то величина R^2 будет оценкой квадрата коэффициента корреляции r^2_{zx}. Если аргументов несколько, а статистическая связь только линейная, то $|R|$ называют *обобщённым коэффициентом корреляции*.

Рассмотрим подробнее случай, когда аппроксимируется зависимость величины Z всего от одной случайной величины X. Выберем структуру аппроксимирующей функции в виде многочлена степени q:

$$z = a_q x^q + a_{q-1}x^{q-1} + \ldots + a_1 x + a_0. \qquad (9.9)$$

Тогда система уравнений (9.4) запишется следующим образом:

$$\begin{cases} \dfrac{\partial S}{\partial a_q} = -2\sum\limits_{i=1}^{n}[z_i - (a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_q x_i^q)]\cdot x_i^q = 0, \\ \dfrac{\partial S}{\partial a_{q-1}} = -2\sum\limits_{i=1}^{n}[z_i - (a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_q x_i^q)]\cdot x_i^{q-1} = 0, \\ \text{- -} \\ \dfrac{\partial S}{\partial a_0} = -2\sum\limits_{i=1}^{n}[z_i - (a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_q x_i^q)] = 0. \end{cases}$$

Преобразуя эту систему, получим:

$$\begin{cases} a_q\sum\limits_{i=1}^{n}x_i^{2q} + a_{q-1}\sum\limits_{i=1}^{n}x_i^{2q-1} + \ldots + a_1\sum\limits_{i=1}^{n}x_i^{q+1} + a_0\sum\limits_{i=1}^{n}x_i^q = \sum\limits_{i=1}^{n}z_i x_i^q , \\ a_q\sum\limits_{i=1}^{n}x_i^{2q-1} + a_{q-1}\sum\limits_{i=1}^{n}x_i^{2q-2} + \ldots + a_1\sum\limits_{i=1}^{n}x_i^{q} + a_0\sum\limits_{i=1}^{n}x_i^{q-1} = \sum\limits_{i=1}^{n}z_i x_i^{q-1} , \\ \text{- -} \\ a_q\sum\limits_{i=1}^{n}x_i^{q} + a_{q-1}\sum\limits_{i=1}^{n}x_i^{q-1} + \ldots + a_1\sum\limits_{i=1}^{n}x_i + a_0 n = \sum\limits_{i=1}^{n}z_i . \end{cases} \qquad (9.10)$$

Особенностью данной системы является то, что её матрица коэффициентов симметрична относительно своей главной диагонали. Кроме того, элементы матрицы могут значительно различаться между собой. Это приводит к неустойчивости процесса реше-

ния системы, т.е. малые отклонения в значениях коэффициентов могут привести к большим отклонениям в самом решении системы. В целях избежания этого эффекта рекомендуется вычислять коэффициенты системы с достаточно большой точностью.

Если поделить все уравнения системы на объём выборки n, то окажется, что матрица системы полностью будет состоять из эмпирических моментов случайной величины X, а правые части системы будут представлять эмпирические смешанные моменты случайных величин X и Z. Таким образом, коэффициенты аппроксимирующей функции определяются эмпирическими моментами рассматриваемых случайных величин.

9.2. Одномерный линейный регрессионный анализ

При $q = 1$ получаем линейную аппроксимирующую функцию $z = a_1x + a_0$. Для удобства будем использовать другое обозначение: $z = ax + b$. В таком случае

$$S = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} [z_i - ax_i - b]^2.$$

Коэффициенты a и b определяются системой уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i z_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} z_i. \end{cases} \tag{9.11}$$

Решая систему по правилу Крамера, получим следующие оценки для неизвестных параметров аппроксимирующей функции:

$$a^* = \frac{n\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad b^* = \frac{1}{n}\left(\sum_{i=1}^{n} z_i - a^* \sum_{i=1}^{n} x_i\right). \tag{9.12}$$

Таким образом, аппроксимирующая функция и остаточная дисперсия определяются формулами:

$$\tilde{z} = a^* x + b^*, \qquad D_{ост} = \frac{1}{n-2}\sum_{i=1}^{n} [z_i - \tilde{z}_i]^2 .. \tag{9.13}$$

В соответствии с (9.12), $b^* = m_z^* - a^* m_x^*$, где m_z^* и m_x^* есть эмпирические математические ожидания величин Z и X. Учитывая это, можно сразу задать структуру аппроксимирующей функции в виде: $z = a \cdot (x - m_x^*) + m_z^*$. Проводя оптимизацию по параметру a, получим:

$$a^* = \frac{K_{zx}^*}{D_x^*} = r_{zx}^* \cdot \frac{\sigma_z^*}{\sigma_x^*}, \qquad b^* = m_z^* - a_1^* m_x^*. \tag{9.14}$$

Следовательно,

$$\tilde{z} = m_z^* + r_{zx}^* \cdot \frac{\sigma_z^*}{\sigma_x^*}(x - m_x^*). \tag{9.15}$$

В формулах (9.14) и (9.15) K_{zx}^* и r_{zx}^* есть соответственно *эмпирические ковариационный момент и коэффициент корреляции* для величин X и Z:

$$K_{zx}^* = \frac{1}{n}\sum_{i=1}^{n} \{x_i - m_x^*)\{z_i - m_z^*), \qquad r_{zx}^* = \frac{K_{zx}^*}{\sigma_z^* \sigma_x^*}, \tag{9.16}$$

где $\sigma_я^* = \sqrt{D_z^*}$ и $\sigma_x^* = \sqrt{D_x^*}$ – эмпирические средние квадратические отклонения. Для эмпирического ковариационного момента справедливо следующее соотношение:

$$K_{zx}^* = \frac{1}{n}\sum_{i=1}^{n} x_i z_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right). \qquad (9.17)$$

Эта формула эквивалентна формуле (3.23) для теоретических моментов.

Можно найти коэффициенты аппроксимирующей функцию $z = ax + b$ исходя непосредственно из теории случайных величин. Для этого нужно минимизировать математическое ожидание квадрата отклонения случайной величины Z от функции $aX + b$:

$$M\{[Z - aX - b]^2\} = M\{[(Z - m_z) - a(X - m_x) - (b - m_z - am_x]^2\} \to \min_{a,b}.$$

Возводя в квадрат и пользуясь свойствами математического ожидания, получим:

$\sigma_z^2 - 2aK_{xz} + a^2\sigma_x^2 + (b - m_z - am_x)^2 \to \min_{a,b}$. Отсюда следует, что

$$a = r_{xz}\frac{\sigma_z}{\sigma_x},\ b = m_z - r_{xz}\frac{\sigma_z}{\sigma_x}m_x,\ \text{т.е. } z = m_z + r_{xz}\frac{\sigma_z}{\sigma_x}(x - m_x).$$

Формула совпадёт с формулой (9.15), если в ней заменить теоретические моменты эмпирическими.

Если известно, что величины X и Z имеют нормальное распределение, то найденная аппроксимирующая функция по структуре будет точно совпадать с теоретической регрессией. Теоретическая регрессия повторяет структуру формулы (9.15), но вместо эмпирических характеристик содержит теоретические характеристики. Это совпадение связано с тем, что, согласно законам преобразования случайных величин, две нормально распределённые величины могут иметь только линейную связь. Никаких других связей между ними быть не может. В окружающем мире широко распространено нормальное распределение (согласно центральной предельной теореме). Поэтому и линейный регрессионный анализ имеет широкое применение.

Построение доверительного интервала для коэффициентов связи r^* и a^* в общем случае является трудной задачей. Если величины X и Z имеют нормальное распределение или достаточно близкое к нормальному распределению, то r^* будет распределено нормально с математическим ожиданием r и дисперсией $\sigma_r^2 \approx (1 - r^{*2})^2/n$. Величина $T = \dfrac{r^* - r}{\sigma_r} = \dfrac{(r^* - r)\sqrt{n}}{1 - r^{*2}}$ при больших n будет центрированной и нормированной нормально распределённой случайной величиной. Рассмотрим уравнение $P\{|T| < \varepsilon_\gamma\} = \gamma$. Очевидно, что $P\{|T| < \varepsilon_\gamma\} = 2\Phi(\varepsilon_\gamma\}$. Решая уравнение $\Phi(\varepsilon_\gamma\} = \gamma/2$ по таблицам нормального распределения при заданной доверительной вероятности γ, найдём ε_γ. Получим доверительный интервал:

$$r^* - \varepsilon_\gamma\frac{1 - r^{*2}}{\sigma_r} < r < r^* + \varepsilon_\gamma\frac{1 - r^{*2}}{\sigma_r} \qquad (9.18)$$

с вероятностью γ. Из этого интервала легко получается доверительный интервал для a^*:

$$a^* - \varepsilon_\gamma\frac{1 - r^{*2}}{\sigma_r}\cdot\frac{\sigma_z^*}{\sigma_x^*} < a < a^* + \varepsilon_\gamma\frac{1 - r^{*2}}{\sigma_r}\cdot\frac{\sigma_z^*}{\sigma_x^*}. \qquad (9.19)$$

При малых значениях n нужно пользоваться распределением Стьюдента.

Пример 9.1. При исследовании зависимости между случайными величинами X и Y была получена следующая таблица измерений соответствующих значений этих величин:

x_i: 0,3 0,4 0,5 0,5 0,6 0,7 0,8 0,9 0,9 1,0 1,1 1,4

y_i: 0,2 0,8 1,2 1,1 1,8 2,5 3,1 3,4 3,8 4,1 4,4 5,9

Вычислить ковариационный момент и коэффициент корреляции. Аппроксимировать статистическую зависимость между этими величинами линейной функцией $y = ax + b$.

Все вычисления удобно проводить в специальной таблице. В левой части таблицы вычисляются все необходимые суммы (последняя строка таблицы) для определения коэффициентов аппроксимирующей функции. Найдём сначала эмпирические математические ожидания: $m_x^* = \frac{9,1}{12} \approx 0,758, \qquad m_y^* = \frac{32,3}{12} \approx 2,691.$

Пользуясь формулами (9.12), вычислим коэффициенты аппроксимирующей функции:

$$a^* = \frac{12 \cdot 30,52 - 9,1 \cdot 32,3}{12 \cdot 8,03 - 9,1^2} \approx 5,34, \quad b^* = 2,691 - 5,34 \cdot 0,758 \approx -1,36.$$

Таким образом, аппроксимирующая функция имеет вид: $\tilde{y} = 5,34x - 1,36$.

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\tilde{y}_i$	$y_i - \tilde{y}_i$	$(y_i - \tilde{y}_i)^2$
1	0,3	0,2	0,06	0,09	0,04	0,242	–0,042	0,0018
2	0,4	0,8	0,32	0,16	0,64	0,776	0,024	0,0006
3	0,5	1,2	0,60	0,25	1,44	1,310	–0,110	0,0121
4	0,5	1,1	0,55	0,25	1,21	1,310	–0,210	0,0441
5	0,6	1,8	1,08	0,36	3,24	1,844	–0,044	0,0019
6	0,7	2,5	1,75	0,49	6,25	2,378	0,122	0,0149
7	0,8	3,1	2,48	0,64	9,61	2,912	0,188	0,0353
8	0,9	3,4	3,06	0,81	11,56	3,446	–0,046	0,0021
9	0,9	3,8	3,42	0,81	14,44	3,446	0,354	0,1253
10	1,0	4,1	4,10	1, 00	16,81	3,980	0,120	0,0144
11	1,1	4,4	4,84	1,21	19,36	4,514	–0,114	0,0130
12	1,4	5,9	8,26	1,96	34,81	6,116	–0,216	0,0467
Σ	9,1	32,3	30,52	8,03	119,41	32,274	–0, 026	0,3122

Во второй части таблицы вычисляется остаточная дисперсия по формуле (9.13) с учётом того, что были оценены два параметра (a и b):

$$D_{ост} = \frac{1}{12 - 2} \cdot 0,3122 = 0,03122.$$

Таким образом, средняя квадратическая ошибка аппроксимации составит величину $\sigma_{ост} = \sqrt{D_{ост}} = \sqrt{0,03122} = 0,177$. Вычислим ковариационный момент по формуле (9.17):

$$K_{xy}^* = \frac{1}{12} \cdot 30,52 - 0,758 \cdot 2,691 \approx 0,494.$$

Для определения коэффициента корреляции необходимо вычислить оценки дисперсий величин X и Y:

$$D_x^* = \frac{1}{12} \cdot 8,03 - 0,758^2 \approx 0,094, \qquad \sigma_x^* = \sqrt{0,094} \approx 0,307.$$

$$D_y^* = \frac{1}{12} \cdot 119,41 - 2,691^2 = 2,71, \quad \sigma_y^* = \sqrt{2,71} \approx 1,646.$$

Следовательно, $r_{xy}^* = \dfrac{K_{xy}^*}{\sigma_x^* \sigma_x^*} = \dfrac{0{,}494}{0{,}307 \cdot 1{,}646} \approx 0{,}978$.

Полученный коэффициент корреляции очень высок. Этим объясняется малая величина средней квадратической ошибки аппроксимации.

9.3. Многомерный линейный регрессионный анализ

Рассмотрим случай, когда случайная величина Z зависит от нескольких случайных величин X_i, $i = 1, 2, \ldots, k$, причём зависимость предполагается линейной. Тогда будем иметь следующую структуру аппроксимирующей функции:

$$y = b + \sum_{i=1}^{k} a_i x_i. \tag{9.20}$$

Коэффициенты a_i и b этой функции нужно оценить. Пусть имеется n групп соответствующих измерений y_j и x_{ij}. Здесь i – номер переменной, j – номер измерения, $j = 1, 2, \ldots, n$. Коэффициенты определяются из условия минимума суммы квадратов отклонений истинных значений величины Y от прогнозируемых:

$$\sum_{j=1}^{n} (y_j - b - \sum_{i=1}^{k} a_i x_{ij})^2 = \min_{\vec{a}, b}, \ \vec{a} = (a_1, a_2, \ldots, a_k). \tag{9.21}$$

Проводя минимизацию по b, получим: $b = m_y^* - \sum_{i=1}^{k} a_i m_{xi}^*$, где $m_y^* = \dfrac{1}{n} \sum_{j=1}^{n} y_j$, $m_{xi}^* = \dfrac{1}{n} \sum_{j=1}^{n} x_{ij}$. Подставим найденное b в выражение (9.21):

$$\sum_{j=1}^{n} [(y_j - m_y^*) - \sum_{i=1}^{k} a_i (x_{ij} - m_{xi}^*)]^2 = \min_{\vec{a}}, \ \vec{a} = (a_1, a_2, \ldots, a_k).$$

Приравнивая частные производные по коэффициентам к нулю, получим систему уравнений:

$$\begin{cases} \sum_{j=1}^{n} [(y_j - m_y^*) - \sum_{i=1}^{k} a_i (x_{ij} - m_{xi}^*)](x_{1j} - m_{x1}^*) = 0, \\ \sum_{j=1}^{n} [(y_j - m_y^*) - \sum_{i=1}^{k} a_i (x_{ij} - m_{xi}^*)](x_{2j} - m_{x2}^*) = 0, \\ ---------- \\ \sum_{j=1}^{n} [(y_j - m_y^*) - \sum_{i=1}^{k} a_i (x_{ij} - m_{xi}^*)](x_{kj} - m_{xk}^*) = 0. \end{cases}$$

Минимизация сводится к решению этой системы k уравнений с k неизвестными. Введём для краткости записи эмпирические коэффициенты ковариации:

$$K_{si} = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - m_{xi}^*)(x_{sj} - m_{xs}^*), \ i, s = 1, 2, \ldots, k;$$

$$K_{ys} = \frac{1}{n} \sum_{j=1}^{n} (y_j - m_y^*)(x_{sj} - m_{xs}^*), \ s = 1, 2, \ldots, k.$$

Получим следующую запись системы:

$$\begin{cases} K_{11}a_1 + K_{12}a_2 + K_{13}a_3 + ... + K_{1k}a_k = K_{y1}, \\ K_{21}a_1 + K_{22}a_2 + K_{23}a_3 + ... + K_{2k}a_k = K_{y2}, \\ - \\ K_{k1}a_1 + K_{k2}a_2 + K_{k3}a_3 + ... + K_{kk}a_k = K_{yk}. \end{cases} \quad (9.22)$$

Решая её, найдём коэффициенты a_i^*. Аппроксимирующая функция запишется:

$$\tilde{y} = m_y^* + \sum_{i=1}^{k} a_i^* (x_i - m_{xi}^*). \quad (9.23)$$

Можно теперь оценить остаточную дисперсию:

$$D_{\text{ост}}^* = \frac{1}{n-\kappa-1} \sum_{j=1}^{n} [(y_j - m_y^* - \sum_{i=1}^{k} a_i^* (x_{ij} - m_{xi}^*)]^2. \quad (9.24)$$

Систему уравнений (9.22) в принципе можно решить любым способом (матричным способом, по формулам Крамера, методом Гаусса или Жордана – Гаусса). Однако нужно принять все способы сохранения точности решения. Матрица коэффициентов системы является симметрической (так как $K_{is} = K_{si}$). Кроме того, коэффициенты могут значительно различаться по абсолютной величине. Эти обстоятельства часто приводят к неустойчивости решений системы.

Для упрощения структуры аппроксимирующей функции рекомендуется провести подробный анализ корреляционной матрицы с целью выявления тех аргументов, которые наиболее существенно влияют на величину функции:

$$\begin{pmatrix} 1 & r_{12}^* & r_{13}^* & ... & r_{1k}^* & r_{1y}^* \\ r_{21}^* & 1 & r_{23}^* & ... & r_{2k}^* & r_{2y}^* \\ r_{31}^* & r_{32}^* & 1 & ... & r_{3k}^* & r_{3y}^* \\ ... & ... & ... & ... & ... & ... \\ r_{k1}^* & r_{k2}^* & r_{k3}^* & ... & 1 & r_{ky}^* \\ r_{y1}^* & r_{y2}^* & r_{y3}^* & ... & r_{yk}^* & 1 \end{pmatrix}, \; r_{is} = \frac{K_{is}^*}{\sigma_i^* \sigma_s^*}, \; r_{iy} = \frac{K_{ys}^*}{\sigma_y^* \sigma_s^*}.$$

Если найдутся две переменные, для которых $|r_{is}^*| \approx 1$, то одну из этих переменных можно не учитывать. Если найдётся переменная, для которой $|r_{iy}^*| \approx 0$, то эту переменную также можно не учитывать.

Рассмотрим подробнее случай двух аргументов. Пусть аппроксимирующая функция имеет вид: $y = b + a_x x + a_z z$. Имеется n групп измерений (x_j, z_j, y_j), $j = 1, 2, \ldots, n$. Тогда

$$b = m_y^* - a_x m_x^* - a_z m_z^*,$$

где $m_y^* = \frac{1}{n}\sum_{j=1}^{n} y_j$, $m_x^* = \frac{1}{n}\sum_{j=1}^{n} x_j$, $m_z^* = \frac{1}{n}\sum_{j=1}^{n} x_j$. Имеем систему:

$$\begin{cases} K_{11}^* a_x + K_{12}^* a_z = K_{1y}^* \\ K_{21}^* a_x + K_{22}^* a_z = K_{2y}^* \end{cases}.$$

Решая систему по правилу Крамера, находим:

$$a_x^* = \frac{K_{1y}^* K_{22}^* - K_{2y}^* K_{12}^*}{K_{11}^* K_{22}^* - K_{21}^* K_{12}^*}, \quad a_z^* = \frac{K_{2y}^* K_{11}^* - K_{1y}^* K_{21}^*}{K_{11}^* K_{22}^* - K_{21}^* K_{12}^*}. \quad (9.25)$$

Здесь коэффициенты выражены через центральные моменты второго порядка, т.е. коэффициенты ковариаций и дисперсии ($K_{ii} = \sigma_i^2$ – дисперсия i-ой величины). Можно выразить коэффициенты через коэффициенты корреляции, пользуясь тем, что $K_{is} = r_{is}\sigma_i\sigma_s$:

$$a_x^* = \frac{\sigma_y^*}{\sigma_y^*} \cdot \frac{r_{xz}^{*2} - r_{yz}^* r_{xz}^*}{1 - r_{xz}^{*2}}, \quad a_z^* = \frac{\sigma_y^*}{\sigma_z^*} \cdot \frac{r_{xz}^{*2} - r_{yx}^* r_{xz}^*}{1 - r_{xz}^{*2}}. \tag{9.26}$$

Заметим, что если отсутствует зависимость между аргументами функции, т.е. $r_{xz}^* \approx 0$, то каждый коэффициент по структуре совпадает с соответствующим коэффициентом одномерной линейной регрессии.

Остаточная дисперсия оценивается формулой

$$D_{\text{ост}}^* = \frac{1}{n-3} \sum_{j=1}^{n} [(y_j - m_y^* - a_x^*(x_j - m_x^*) - a_z^*(z_j - m_z^*)]^2. \tag{9.27}$$

Отметим, что в случае двух аргументов обобщённый коэффициент корреляции можно вычислить по формуле

$$R^2 = \frac{r_{xy}^{*2} + r_{zy}^{*2} - 2r_{xy}^* r_{zy}^* r_{xz}^*}{1 - r_{xz}^{*2}}. \tag{9.28}$$

9.4. Одномерный нелинейный регрессионный анализ

В соответствии с общими положениями, при нелинейном анализе можно применять любые нелинейные функции, имея в виду структуру (9.2). Если для выбора аппроксимирующей функции нет никакой дополнительной информации, то в качестве такой функции берут многочлены. Полагая в (9.9) $q = 2$, $q = 3$ и т.д., можно рассмотреть случай квадратичной аппроксимирующей функции, кубической и т.д. Например, для аппроксимирующей функции вида $y = ax^2 + bx + c$ ($q = 2$) неизвестные коэффициенты определяются следующей системой уравнений:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^4 + b\sum\limits_{i=1}^{n} x_i^3 + c\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i x_i^2 \\ a\sum\limits_{i=1}^{n} x_i^3 + b\sum\limits_{i=1}^{n} x_i^2 + c\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i x_i \\ a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i + cn = \sum\limits_{i=1}^{n} y_i \end{cases}. \tag{9.29}$$

После определения оценок неизвестных коэффициентов (a^*, b^*, c^*) остаточная дисперсия находится по общей формуле

$$D_{ост} = \frac{1}{n-3} \sum_{i=1}^{n} [y_i - \tilde{y}_i]^2, \quad \text{где} \quad \tilde{y}_i = a^* x_i^2 + b^* x_i + c^*. \tag{9.30}$$

Оценка корреляционного отношения определяется по формуле

$$\eta_{y/x}^{*2} = 1 - \frac{D_{ост}}{D_y^*}. \tag{9.31}$$

Пример 9.2. При исследовании зависимости между случайными величинами X и Y была получена следующая таблица измерений соответствующих значений этих величин:

x_i:	1,3	1,5	1,5	1,8	1,9	2,0	2,1	2,4	2,4	2,5
y_i:	0,2	0,8	1,2	1,1	1,8	3,2	2,5	3,1	3,4	3,8

Аппроксимировать статистическую связь между Y и X многочленом второго порядка.

Вносим данные в таблицу, в которой вычисляем все необходимые суммы, которые фигурируют в системе уравнений (9.29):

x	y	x^2	x^3	x^4	xy	yx^2	y^2	$(y-y^*)^2$
1,3	4,1	1,69	2,197	2,8561	5,33	6,929	16,81	0,2134
1,5	6,2	2,25	3,375	5,0625	9,3	13,95	38,44	0,3249
1,5	8,1	2,25	3,375	5,0625	12,15	18,225	65,61	1,7689
1,8	9,3	3,24	5,832	10,4976	16,74	30,132	86,49	0,5213
1,9	11,4	3,61	6,859	13,0321	21,66	41,154	129,96	0,0961
2	12,2	4	8	16	24,4	48,8	148,84	0,0025
2,1	13,7	4,41	9,261	19,4481	28,77	60,417	187,69	0,2480
2,4	14,8	5,76	13,824	33,1776	35,52	85,248	219,04	2,2801
2 4	16,2	5,76	13,824	33,1776	38,88	93,312	262,44	0,0121
2,5	18,4	6,25	15,625	39,0625	46	115	338,56	1,1449
19,40	114,4	39,22	82,172	177,3766	238,75	513,167	1493,88	6,6122

В последней строке таблицы приводятся суммы элементов столбцов. Таким образом, получаем систему:

$$\begin{cases} 177,38\,a + 82,17\,b + 39,22\,c = 513,17 \\ 82,17\,a + 39,22\,b + 19,4\,c = 238,75 \\ 39,22\,a + 19,4\,b + 10\,c = 114,4 \end{cases}.$$

Решая систему, находим: a = –0,4; b = 12,16; c = –10,57; $\tilde{y} = -0,4x^2 + 12,16x - 10,57$. В последнем столбце таблицы вычислена сумма квадратов отклонений прогнозируемых значений от истинных значений Y. Остаточная дисперсия равна $D_{ост} = \frac{1}{n-3}\sum_{i=1}^{n}[y_i - \tilde{y}_i]^2 = \frac{1}{10-3}\cdot 6,612 \approx 0,945$. Оценка дисперсии Y: $D_y^* = \frac{1493,9}{10} - (\frac{114,4}{10})^2 \approx 18,5$. Оценка корреляционного отношения: $\eta^{*2}_{y/x} = 1 - \frac{D_{ост}}{D_y^*} = 1 - \frac{0,945}{18,5} \approx 0,949$. Такое высокое корреляционного отношения говорит о сильной зависимости между величинами.

Если выбирается структура аппроксимирующей функции в виде $y = a\varphi(x) + b$, где $\varphi(x)$ – любая известная нелинейная функция, то рекомендуется сделать замену $z = \varphi(x)$ и рассматривать задачу линейного одномерного регрессионного анализа с аппроксимирующей функцией $y = az + b$. Приведём примеры таких замен.

1. $y = ax^2 + b$ – параболическая аппроксимация. Замена $z = x^2$.

2. $y = \frac{a}{x} + b$ – гиперболическая аппроксимация. Замена $z = x^{-1}$.

3. $y = ae^x + b$ – аппроксимация показательной функцией. Замена $z = e^x$.

Можно привести примеры и более сложных аппроксимирующих функций, которые путём одной или нескольких замен переменных сводятся к линейным функциям. Такие замены обычно рекомендуются, хотя общий объём вычислений при таких заменах практически не уменьшается.

Заметим, что здесь в краткой форме изложены только основные положения регрессионного анализа. Для более подробного знакомства с регрессионным анализом нужно смотреть специальную литературу, посвящённую этому вопросу.

Термины для запоминания

- Регрессионный анализ; аппроксимация; аппроксимирующая функция; прогнозирующая функция; ошибка аппроксимации; остаточная дисперсия; обобщённый коэффициент корреляции.

Список контрольных вопросов

1. Какую задачу решает дисперсионный анализ?
2. Что называется аппроксимирующей или прогнозирующей функцией?
3. Какая функция теоретически является наилучшей для прогнозирования в смысле минимума среднего квадрата ошибки прогнозирования?
4. Дайте словесную формулировку задачи регрессионного анализа (постановка задачи).
5. Как классифицируются задачи регрессионного анализа?
6. Что минимизируется в регрессионном анализе и по каким параметрам?
7. Как выбирается структура аппроксимирующей функции в регрессионном анализе?
8. Что такое остаточная дисперсия, что она определяет?
9. Что называют «обобщённым коэффициентом корреляции»?
10. Как определяются оценки корреляционного момента и коэффициента корреляции по экспериментальным данным?
11. В каком случае можно ограничиться получением линейной аппроксимирующей функции при одномерном регрессионном анализе?
12. От каких параметров зависит линейная аппроксимирующая функция в одномерном регрессионном анализе?
13. Как сократить число аргументов аппроксимирующей функции в многомерном линейном регрессионном анализе?
14. Какими особенностями обладает матрица коэффициентов системы уравнений, определяющей параметры аппроксимирующей функции?
15. В каких случаях задачу нелинейного одномерного регрессионного анализа можно свести к задаче линейного одномерного анализа?

Вопрос для обсуждения

Прибыль торгового предприятия зависит от многих факторов. Попробуйте перечислить основные факторы и проанализируйте, какие из них имеют взаимные статистические связи.

ТЕМА 10.

Применение ЭВМ

Студент должен
освоить:

- ознакомление с пакетами прикладных программ;

приобрести навыки:

- применения пакетов прикладных программ для решения задач статистического анализа.

Краткое содержание

Решение вероятностных задач в среде Microsoft Excel. Средства решения статистических задач в пакете MathCAD. Специализированные пакеты программ.

10.1. Общие замечания

Для решения задач теории вероятностей и математической статистики можно использовать любые пакеты программ, предназначенные для решения математических задач широкого класса. В настоящее время существует много пакетов программ, в которых предусмотрено решение вероятностных задач и задач математической статистики. Имеется много и специальных пакетов программ, ориентированных именно на решение статистических задач. Выбор пакета должен определяться квалификацией специалиста в данной области. Специалисту, слабо знакомому с методами теории вероятностей и математической статистики, нецелесообразно выбирать сложный, универсального плана, пакет программ. Освоение сложного пакета является не всегда простым делом. Кроме того, недостаточно компетентный человек в вопросах статистики не сможет адекватно реагировать на запросы программ, не понимая сути этих вопросов. В таком случае применение компьютера даст только отрицательный результат. В связи с этим изучающим данный курс рекомендуется начинать с простых (в освоении) пакетов универсального характера. Одним из доступных и достаточно простых для освоения пакетов является MathCAD. В этом пакете имеется широкий набор операторов и функций, которые позволяют решать многие задачи теории вероятностей и математической статистики. В среде Excel Microsoft Office также имеются аналогичные функции. Возможности этих пакетов примерно равные, различие заключается, прежде всего, в форме представления информации. В первом пакете программ информация представляется в более естественной форме, привычной по курсу высшей математики. Во втором пакете информация размещается в электронных таблицах. Рассмотрим методику решения задачи линейного программирования в программной среде MathCAD (версии 8 и выше) и в среде Excel Microsoft Office. Эти пакеты программ являются наиболее доступными для рядового пользователя ЭВМ.

10.2. Средства решения статистических задач в пакете MathCAD

Пакет MathCAD предназначен специально для решения математических задач. Он имеет три встроенных редактора: текстовой редактор, графический редактор и редактор математических формул. Это дополняется достаточно мощными средствами реализации процессов вычислений. В связи с этим пакет удобен как для решения задач, так и для оформления документов, содержащих математические формулы и процессы вычислений. Изменение значений параметров вначале документа приводит к пересчёту всех формул, содержащих эти параметры, по всему документу. Это свойство пакета удобно при анализе влияния параметров формулы на результат решения поставленной задачи.

Перечислим кратко некоторые из основных функций, касающиеся обработки экспериментальных данных. Все эти функции следует искать в меню «Функции» в разделах «Статистика» и «Регрессия».

1. mean(X) – среднее арифметическое измерений, содержащихся в векторе X (оценка математического ожидания).

2. var(X) – смещённая оценка дисперсии (D^*), Var(X) – несмещённая оценка дисперсии (S^2).

3. line(X, Y) – линейная регрессия вида $y = b + ax$. Векторы X и Y представляют соответствующие измерения величин X и Y. Результат выдаётся в виде вектора, первая компонента которого содержит оценку b, а вторая – оценку a.

4. cvar(X, Y) – ковариационный момент.

5. korr(X, Y) – коэффициент корреляции.

6. stderr(X, Y) – остаточная дисперсия после применения линейной регрессии.

Рассмотрим пример обработки двух массивов измерений случайных величин X и Y, поставив задачу оценить математические ожидания и дисперсии этих величин и найти линейную аппроксимацию статистической связи между величинами в виде $y = b + ax$. Ниже представлена полная распечатка решения данной задами.

Исходные данные задаём в виде векторов-столбцов. Вычислим сначала оценки математических ожиданий и оценки дисперсий.

$$X := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \quad \begin{matrix} \text{mean}(X) = 3 \\ \text{var}(X) = 2 \\ \text{Var}(X) = 2{,}5 \end{matrix} \qquad Y := \begin{pmatrix} -1 \\ 1 \\ 4 \\ 7 \\ 9 \end{pmatrix} \quad \begin{matrix} \text{mean}(Y) = 4 \\ \text{var}(Y) = 13{,}6 \\ \text{Var}(Y) = 17 \end{matrix} .$$

Определим коэффициент линейной регрессии, ковариационный момент, коэффициент корреляции и остаточную дисперсию.

$$\text{line}(X, Y) = \begin{pmatrix} -3{,}8 \\ 2{,}6 \end{pmatrix} \qquad c\,\text{var}(X, Y) = 5{,}2 \qquad \text{corr}(X, Y) = 0{,}997$$
$$\text{stderr}(X, Y) = 0{,}365$$

Таким образом, $m_x^* = 3$, $m_y^* = 4$, $D_x^* = 2$, $D_y^* = 13{,}6$, $S_x^2 = 2{,}5$, $S_y^2 = 17$. Линия линейной регрессии: $y = 2{,}6x - 3{,}8$. Остаточная дисперсия при такой аппроксимации равна $D_{ост} = 0{,}365$.

Заметим, что этот пример носит чисто иллюстративный характер. Поэтому объём выборки взят очень малым. Из этих же соображений измерения взяты с заведомо большим коэффициентом корреляции. В пакете имеются и много других функции для статистических вычислений. В пакете представлен большой набор законов распределения вероятностей, позволяющий решать и чисто вероятностные задачи. Читателю предлагается разобраться с ними самостоятельно. Отметим также, что размерность векторов в данном пакете программ ограничена.

10.3. Решение статистических задач в среде Microsoft Excel

Процесс решения задач в среде Excel менее привлекателен, поскольку он не позволяет записывать эти задачи в их обычной форме. Перечислим основные функции, которые можно использовать при решении задач теории вероятностей и математической статистики.

1. СРЗНАЧ(X) – среднее арифметическое измерений, содержащихся в массиве X (оценка математического ожидания)

2. ДИСПР(X) – смещённая оценка дисперсии (D^*), ДИСП(X) – несмещённая оценка дисперсии (S^2).

3. СТАНДОТКЛОН(X) – оценка среднего квадратического отклонения (*S*).

4. СТЬЮДРАСПОБР(1 – α, r) – критическое значение, найденное по распределению Стьюдента для построения доверительного интервала. Здесь α – доверительная вероятность, r – число степеней свободы.

5. МОДА(X) – оценка моды случайной величины.

6. МЕДИАНА(X) – оценка медианы случайной величины.

7. ЭКСЦЕСС(X) – оценка эксцесса случайной величины.

8. СКОС(X) – оценка асимметрии распределения случайной величины.

9. ЛИНЕЙН(Y, X) – линейная регрессия вида $y = b + ax$. Массивы X и Y представляют соответствующие измерения величин X и Y. Результат выдаётся в виде вектора, первая компонента которого содержит оценку b, а вторая – оценку a.

10. КОВАР(X, Y) – ковариационный момент.

11. КОРРЕЛ (X, Y) – коэффициент корреляции.

12. СТОШYX(X, Y) – корень квадратный из остаточной дисперсии после применения линейной регрессии.

Массивы можно задавать простым перечислением чисел, или с помощью адресов ячеек, в которых находятся числа, или с помощью задания диапазонов адресов.

Рассмотрим пример обработки двух массивов объёма $n = 30$ измерений случайных величин X и Y, поставив задачу оценить математические ожидания, дисперсии, моду, медиану, асимметрию и эксцесс этих величин и найти линейную аппроксимацию статистической связи между величинами в виде $y = b + ax$. Построим также доверительный интервал для математических ожиданий с коэффициентом доверия 0,95, пользуясь распределением Стьюдента. Ниже представлена полная распечатка решения данной задами. Массив X представлен диапазоном адресов ячеек B2 : G6, а массив Y – диапазоном адресов B21 : G25.

Исходный массив данных						
X:	2,3	3,9	5,4	2,8	5,1	3,3
	3,8	2,7	2,9	3,4	2,9	3,6
	4,8	3,5	2,8	5,3	3,5	3,4
	2,4	2,7	3,9	4,4	3,6	4,1
	2,1	4,3	3,6	4,5	3,7	3,2
Оценка математического ожидания						3,5967
Оценка дисперсии						0,7283
Оценка несмещённой дисперсии:						0,7534
Среднее квадратическое отклонение						0,868
Критическое значение						2,0452
Левая граница доверительного интервала						3,2725
Правая граница доверительного интервала						3,9208
Мода распределения						3,6
Медиана распределения						3,55
Асимметрия						0,4102
Эксцесс						–0,34

Вторая таблица

Исходный массив данных						
Y:	4,1	6,2	8,4	3,8	7,1	3,9
	5,8	3,7	4,7	5,4	3,9	6,6
	7,2	5,5	4,2	6,3	4,5	5,4
	4,6	4,6	5,1	5,9	4,6	6,1
	3,8	6,1	5,6	6,5	7,1	4,2
Оценка математического ожидания:						5,3633
Оценка дисперсии:						1,4637
Оценка несмещённой дисперсии:						1,5284
Среднее квадратическое отклонение						1,2305
Критическое значение						2,0452
Левая граница доверительного интервала						4,9039
Правая граница доверительного интервала						5,8228
Мода распределения						4,6
Медиана распределения						5,4
Асимметрия						0,4864
Эксцесс						–0,4261
Ковариационный момент						0,8785
Коэффициент корреляции						0,8509
Параметры линейной регрессии					1,20626	1,02483

Все вычисления проведены с помощью перечисленных выше функций, кроме вычисления границ доверительного интервала. Для вычисления границ в соответствующие ячейки вводились формулы: $m^* - t_\alpha \cdot S/\sqrt{n}$ – для левой границы интервала и $m^* + t_\alpha \cdot S/\sqrt{n}$ – для правой границы интервала. Значения m^*, t_α, S взяты из предыдущих вычислений.

Таким образом, получена следующая аппроксимирующая функция: y = 1,206x + + 1,025. Получены также доверительные интервалы: $m_x \in$ (3,273; 3,921) с вероятностью 0,95; $m_y \in$ (4,904; 5,823) с вероятностью 0,95.

Сделаем одно очень важное замечание. Если применяется функция, результатом которой является матрица (в рассмотренном примере это функция ЛИНЕЙН, которая выдаёт два параметра), то формулу функции необходимо ввести в ячейки как формулу массива. Для этого нужно выполнить следующие действия. Нужно ввести формулу в ячейку, в которой будет размещаться первый элемент матрицы – результата. Затем следует выделить все ячейки, в которых будет размещена матрица – результат, начиная с той, в которой уже введена формула. Далее следует нажать клавишу F2, а затем – клавиши CTRL + SHIFT + ENTER. Если формула не будет введена как формула массива, то будет выведено лишь одно значение в ту ячейку, в которую была введена формула.

Специальные пакеты программ, созданные для решения больших статистических задач, следует изучать и применять после освоения тех простых средств решения статистических задач, которые имеются в рассмотренных пакетах.

Практикум

1. Методика выполнения задач

Все предлагаемые задачи разделены на две группы. Задачи первой группы предназначены для решения во время аудиторных практических занятий. Задачи этой группы могут использоваться студентами для самостоятельной тренировки. Однако предлагать их к обязательному выполнению не рекомендуется. Задачи второй группы предназначены для выполнения студентами в качестве самостоятельной работы. Эти задачи предлагаются студентам в виде некоторого типового расчёта, рассчитанного на весь семестр. Вариант следует выбирать по номеру студента в списке студентов группы. Последовательность и время выполнения заданий должны точно соответствовать календарному плану. Выполнение типового расчёта обязательно для всех студентов, в том числе студентов заочного обучения.

Работа студентов над типовым расчётом должна начинаться с первых занятий по дисциплине «Теория вероятностей и математическая статистика». Прежде чем выполнять конкретное задание, необходимо тщательно разобраться с теоретическим материалом, акцентируя внимание на постановку задачи в общем виде. Не следует запоминать формулы, если не понятен их смысл. В первую очередь нужно осмыслить те условия, при которых эти формулы возникают. После освоения теоретического материала необходимо разобрать примеры, решённые на лекционных и практических занятиях, и сравнить их с тем заданием, которое предстоит выполнить. Само задание выполняется с подробным описанием всех этапов решения. Начинать решение нужно с описания обозначений тех событий, которые фигурируют в задаче. Например, «Пусть А – событие, состоящее в том, что будет одно попадание в цель при одном выстреле» и т.п. В конце решения следует записать ответы к задаче. Следует помнить, что решение заданий типового расчёта не является самоцелью. Это лишь средство освоения материала изучаемой дисциплины. Подробное описание этапов решения задачи, мотивация применения тех или иных формул с познавательной точки зрения имеет большее значение, чем сами вычисления и ответ задачи. Таким образом, описание этапов решения задачи является определённым показателем уровня освоения материала студентом.

Рассмотренный в конце практикума пример выполнения типового расчёта может, в определённой мере, помочь выполнению студентом своего задания. Однако этот пример нельзя интерпретировать как некий шаблон, который справедлив в любом случае. Большинство заданий являются индивидуальными не только по содержанию, но и по логической схеме. Таким образом, в разных вариантах одного (по номеру) задания могут встретиться разные формулы. Некоторые задачи допускают различные подходы к решению, приводящие к правильному результату. В таком случае можно выбрать любой из них, если только в формулировке задания нет специальных оговорок по этому вопросу.

Часть представленных здесь задач составлена непосредственно автором. Другая часть задач взята из различных источников, список которых прилагается в конце практикума.

2. Задачи для аудиторной и самостоятельной работы

Тема 1. Случайные события

1.1. Сколько трёхзначных чисел можно составить из цифр 1, 2, 3, 4, 5, 6?

1.2. Сколько трёхзначных чисел можно составить из цифр 1, 2, 3, 4, 5, 6, если эти числа не должны иметь повторяющихся цифр?

1.3. Сколько существует двузначных чисел, у которых обе цифры чётные?

1.4. Сколькими способами можно выбрать трёх студентов из 8?

1.5. Сколькими способами можно выбрать председателя, секретаря и казначея жилищного кооператива, если кооператив состоит из 12 членов?

1.6. Сколькими способами можно разместить 5 покупателей в очереди в кассу?

1.7. Сколько слов можно составить путём перестановки букв в слове «кроссворд» (под «словом» понимается любая комбинация букв)?

1.8. Группа туристов из 12 человек разбивается на три равные по численности подгруппы. Сколько существует вариантов такого разделения (без учёта перестановок подгрупп)?

1.9. Чему равно выражение $(A+B)(A+C)(A+\overline{C})$?

1.10. Упростите выражение: $C(\overline{B}+C)(A+B)(B+\overline{C})$.

1.11. Три стрелка одновременно делают по одному выстрелу по мишени. Событие A_k – попадание в мишень k-м стрелком (k = 1, 2 , 3), событие B – имело место только одно попадание, событие C – попаданий было не менее одного. Выразить события B и C через события A_k.

1.12. Монета бросается 3 раза. A_k – событие, состоящее в том, что герб появится при k-м броске (k = 1, 2, 3), B – событие, состоящее в том, что герб появится 2 раза. Выразить событие B через события A_k.

1.13. В коробке находится 5 деталей, из которых 2 детали имеют скрытые дефекты. Наугад берётся деталь. Какова вероятность того, что эта деталь имеет скрытый дефект?

1.14. В условиях предыдущей задачи было наугад взято из коробки две детали. Какова вероятность того, что только одна из них имеет скрытый дефект?

1.15. На столе лежит 15 экзаменационных билетов с номерами 1, 2, ..., 15. Преподаватель наугад берёт 2 билета. Какова вероятность того, что они из первых четырёх?

1.16. Из отрезка [–2; 2] наудачу взяты два числа. Какова вероятность того, что их сумма больше единицы, а модуль разности меньше единицы?

1.17. Исходя из условий задачи 1.8 найти вероятность того, что туристы К и М попадут в одну подгруппу.

1.18. Бросаются три игральные кости. Определить вероятность того, что сумма выпавших очков будет равна 14.

1.19. Вероятность преодолеть планку на данной высоте для прыгуна в высоту равна 0,8. Какова вероятность того, что спортсмен возьмёт эту высоту, если он имеет право на три попытки, причём очередная попытка делается только тогда, когда предыдущие попытки были неудачными?

1.20. Исходя из условий задачи 1.11 найти вероятности событий B и C, если известно, что $P(A_1) = 0{,}5$, $P(A_2) = 0{,}7$, $P(A_3) = 0{,}6$.

1.21. Двое поочерёдно бросают монету. Выигрывает тот, кто первым в результате получит орла. Какова вероятность того, что игра закончится не позднее чем после второго бросания монеты вторым игроком?

1.22. В тире имеется 5 ружей, вероятность попадания из которых равна 0,5, три ружья с вероятностью попадания 0,7 и два ружья с вероятностью попадания 0,8. Определить

вероятность попадания в мишень при одном выстреле, если стрелок берёт одно из ружей наудачу.

1.23. В продажу поступают телевизоры трех заводов. Продукция первого завода содержит 10% телевизоров со скрытым дефектом, второго – 8% и третьего – 6%. Приобретённый телевизор оказался без дефектов. Какова вероятность того, что этот телевизор был изготовлен на первом заводе, если в магазин поступило 40% телевизоров с первого завода, 25% – со второго и 35% – с третьего?

1.24. Известно, что 95% выпускаемой продукции удовлетворяет стандарту. Упрощенная схема контроля признаёт пригодной стандартную продукцию с вероятностью 0,9 и нестандартную – с вероятностью 0,08. Определить вероятность того, что изделие, прошедшее упрощённый контроль, удовлетворяет стандарту.

1.25. Батарея дала 6 выстрелов по военному объекту, вероятность попадания в который равна 0,3. Найти:

а) наивероятнейшее число попаданий и его вероятность;

б) вероятность разрушения объекта, если для его разрушения требуется не менее 2 попаданий.

1.26. Круг разделён на три равные части. Какова вероятность того, что из 6 точек, брошенных наудачу в круг, три попадут в одну часть круга, две – во вторую, одна – в третью?

1.27. Станок-автомат штампует детали. Вероятность того, что изготовленная деталь окажется бракованной, равна 0,02. Пользуясь формулой Пуассона, найти вероятность того, что среди 100 деталей бракованных окажется не менее 2 и не более 4.

1.28. Тестовый билет содержит 5 вопросов. На каждый вопрос предлагается два варианта ответа, из которых один является правильным. Для получения зачёта нужно правильно ответить не менее чем на три вопроса. Какова вероятность того, что некто, выбирая ответы наудачу, может получить зачёт?

1.29. Доля изделий первого сорта в продукции завода составляет 70%. Какова вероятность того, что из отобранных 400 изделий окажется от 270 до 300 изделий первого сорта?

1.30. Вероятность отказа прибора при испытании равна 0,2. Сколько таких приборов нужно испытать, чтобы с вероятностью, не меньшей чем 0,99, получить хотя бы один отказ?

Тема 2. Случайные величины

2.1. Игральная кость бросается два раза. Пусть X – суммарное число выпавших очков. Составить закон распределения вероятностей случайной величины X.

2.2. Среди десяти деталей имеется четыре бракованных. Из всей партии случайным образом взято три детали. Составить закон распределения числа бракованных деталей среди взятых.

2.3. Вероятность попадания в цель при каждом выстреле равна 0,6. Стрельба ведётся до первого попадания в цель. Составить закон распределения случайной величины X – числа сделанных выстрелов.

2.4. В дополнение к условиям задачи 2.3 предполагается, что имеется всего три патрона. Определить и представить графически функцию распределения вероятностей случайной величины X числа сделанных выстрелов до первого попадания.

2.5. Функция распределения вероятностей непрерывной случайной величины X определяется формулой:

$$F(x)=\begin{cases}0, & \text{если} \quad x\le 0,\\ ax^3, & \text{если} \quad 0<x\le 3,\\ 1, & \text{если} \quad x>3.\end{cases}$$

Определить коэффициент a и плотность распределения вероятностей $f(x)$. Найти вероятность того, что величина X примет значение из интервала (1; 2).

2.6. Задана плотность распределения вероятностей непрерывной случайной величины X:

$$f(x) = \begin{cases} a \cdot \sin x, & \text{если} \quad 0 < x \le \pi, \\ 0, & \text{если} \quad x \le 0 \quad \text{или} \quad x > \pi. \end{cases}$$

Определить коэффициент a и функцию распределения вероятностей величины X. Найти вероятность того, что величина X примет значение из интервала $(\pi/3; \pi/2)$.

2.7. Плотность распределения случайной величины X задана формулой

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty$$

(закон Коши). Определить вероятность того, что случайная величина примет значение из интервала (–1; 1).

2.8. Показать, что математическое ожидание дискретной случайной величины, распределённой по закону Пуассона с параметром λ, равно самому параметру λ.

2.9. Вычислить математическое ожидание и дисперсию дискретной случайной величины X, заданной таблично:

x_k	–1	0	1	2	4	6	8
p_k	0,1	0,15	0,35	0,2	0,1	0,05	0,05

2.10. Вычислить математическое ожидание и дисперсию непрерывной случайной величины X, если

$$f(x) = \begin{cases} 4x \cdot e^{-2x}, & \text{если } x \ge 0, \\ 0, & \text{если } x < 0. \end{cases}$$

2.11. Вычислить коэффициент вариации для случайной величины, определённой в задаче 2.6.

2.12. Определить все квартили случайной величины X, распределённой по экспоненциальному закону: $f(x) = 3e^{-3x}$, $x > 0$. Найти вероятное отклонение случайной величины.

2.13. Вычислить математическое ожидание, дисперсию, скошенность и эксцесс непрерывной случайной величины X, распределённой равномерно в интервале [1; 5].

2.14. Вычислить первые три центральных момента случайной величины, определённой в задаче 2.8.

2.15. Непрерывная случайная величина X подчинена закону распределения с плотностью $f(x) = 0{,}5 \cdot e^{-|x|}$, $-\infty < x < \infty$. Определить математическое ожидание, дисперсию, скошенность и эксцесс этой случайной величины.

2.16. Определить производящую функцию начальных моментов дискретной случайной величины X, распределённой по геометрическому закону: $P\{X = k\} = pq^{k-1}$, $0 < p < 1$, $q = 1 - p$, $k = 1, 2, 3, \ldots$. С помощью производящей функции найти математическое ожидание и дисперсию этой величины.

2.17. Определить производящую функцию начальных моментов случайной величины X, распределённой по закону Пуассона: $P\{X = k\} = \dfrac{\lambda^k}{k!} \cdot e^{-\lambda}$, $\lambda > 0$, $k = 0, 1, 2, \ldots$.

С помощью производящей функции найти математическое ожидание и дисперсию X.

2.18. Интервал времени безотказной работы технической системы распределён по экспоненциальному закону с параметром $\lambda = 0{,}1$ (час$^{-1}$). Определить вероятность того, что

отказ произойдет не ранее чем через 5 часов безотказной работы. Вычислить вероятность того, что отказ произойдет между 5-ю и 6-ю часами после начала работы системы.

2.19. Нормально распределённая случайная величина X имеет математическое ожидание $m_x = 2$ и дисперсию $D_x = 4$. Определить вероятность того, что эта случайная величина примет значение из интервала (1; 3).

2.20. По цели, имеющей вид полосы (автострада), ширина которой равна 16 метров, ведётся стрельба в направлении, перпендикулярном автостраде. Прицеливание ведётся по средней линии автострады. Среднее квадратическое отклонение в направлении стрельбы равно σ_x = 6м. Имеется систематическая ошибка в направлении стрельбы – недолёт 3 метра, которая при наводке не учитывается. Определить вероятность того, что при трёх выстрелах произойдёт только два попадания в автостраду.

Тема 3. Многомерные случайные величины

3.1. Двухмерная случайная величина (X,Y) с одинаковой вероятностью может принимать значения: (2; 1), (2; 2), (3; 2), (3; 3), (3; 4), (4; 5). Найти вероятность того, что точка $(X; Y)$ окажется в области D, если область D определена неравенствами $x \le 3$ и $y > 2$.

3.2. Закон распределения дискретной двухмерной случайной величины (X,Y) представлен следующей таблицей:

x_i \ y_j	–2	–1	0	1	2
1	0,04	0,04	0,03	0,03	0,01
3	0,04	0,07	0,06	0,05	0,03
5	0,05	0,08	0,09	0,08	0,05
7	0,03	0,04	0,04	0,06	0,08

Определить одномерные законы распределения случайных величин X и Y.

3.3. Функция распределения двухмерной случайной величины (X, Y) имеет следующий вид:

$$F(x,y) = 1 - e^{-ax} - e^{-by} + e^{-ax-by}, \quad x \ge 0,\ y \ge 0, \quad (a > 0, b > 0).$$

Определить одномерные функции распределения и одномерные плотности распределения вероятностей величин X, Y.

3.4. Двухмерная случайная величина (X, Y) имеет плотность распределения вероятностей

$$f(x,y) = \frac{A}{\pi^2 (16 + x^2)(25 + y^2)}.$$

Требуется определить величину A и найти функцию распределения вероятностей $F(x, y)$.

3.5. Определить вероятность попадания точки с координатами $(X; Y)$ в область, определяемую неравенствами $1 \le x \le 2$, $1 \le y \le 2$, если двухмерная функция распределения вероятностей имеет вид:

$$F(x,y) = \begin{cases} 1 - a^{-x^2} - a^{-2y^2} + a^{-x^2-2y^2}, & \text{если} \quad x \ge 0,\ y \ge 0, \quad a > 1, \\ 0, & \text{если} \quad x < 0 \ \text{ или } \ y < 0. \end{cases}$$

3.6. По двухмерной плотности распределения $f(x, y)$, определённой в задаче 3.4, найти одномерную плотность $f_X(x)$.

3.7. Исходя из условий задачи 3.2 определить условный закон распределения $P\{Y = yj \mid X = 3\}$. Являются ли величины X и Y статистически зависимыми?

3.8. Двухмерная случайная величина (X, Y) распределена равномерно в треугольнике, ограниченном прямыми $x = 0$, $y = 0$, $x + y = 1$. Вычислить математические ожидания m_x и m_y, дисперсии σ_x и σ_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}.

3.9. Для двухмерной случайной величины, заданной в примере 3.2, вычислить математические ожидания m_x и m_y, дисперсии σ_x и σ_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}.

3.10. Для двухмерной случайной величины, определённой в примере 3.8, вычислить условное математическое ожидание $m_y(x)$ и условную дисперсию $\sigma_y^2(x)$.

3.11. Определить в точке (2, 2) плотность распределения двухмерной нормально распределённой случайной величины (X, Y), если известно, что $m_x = m_y = 0$ и $K = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

3.12. Координаты $(X; Y)$ случайной точки A на плоскости подчинены нормальному закону

$$f(x, y) = \frac{1}{2\pi ab} \exp\left\{-\frac{1}{2}\left[\frac{x^2}{a^2} + \frac{y^2}{b^2}\right]\right\}.$$

Определить вероятность того, что точка A окажется внутри эллипса с центром в начале координат с полуосями $3a/2$ и $3b/2$, совпадающими с осями координат.

3.13. Плотность распределения вероятностей двухмерной случайной величины (X, Y) задана в виде

$$f(x, y) = k \cdot \exp\{-4x^2 - 6xy - 9y^2\}.$$

Определить постоянную k, коэффициент корреляции r_{xy} и условные законы распределения вероятностей $f(x \mid y)$ и $f(y \mid x)$.

Тема 4. Функциональные преобразования случайных величин

4.1. Дискретная случайная величина X, заданной таблично:

x_k	–2	–1	0	1	2	6	8
p_k	0,1	0,15	0,3	0,2	0,1	0,1	0,05

Найти распределение вероятностей величины $Y = X^2$.

4.2. Случайная величина X распределена равномерно в интервале $[a; b]$. Определить закон распределения величины $Y = e^X$.

4.3. Случайная величина X распределена нормально с математическим ожиданием m_x и дисперсией σ_x^2. Определить закон распределения величины $Y = e^X$.

4.4. Определить закон распределения случайной величины $Y = \ln X$, если величина X распределена по экспоненциальному закону с параметром λ.

4.5. Найти плотность распределения вероятностей для площади S квадрата, если его сторона X является случайной величиной, распределённой равномерно в интервале $[0; 1]$.

4.6. Найти закон распределения суммы Z двух независимых случайных величин X и Y, каждая из которых имеет равномерное распределение на интервале $(0; 1)$.

4.7. Случайные независимые величины X и Y распределены по одному и тому же показательному закону с параметром λ. Найти закон распределения величин $U = 2X$ и $V = X + Y$ и их математические ожидания.

4.8. Найти закон распределения произведения двух независимых случайных величин X и Y, каждая из которых имеет равномерное распределение на интервале (0; 1).

4.9. Случайные величины X и Y распределены по показательному закону с параметром λ. Найти закон распределения величины $Z = Y/X$.

4.10. Решить задачу 4.9 при условии, что параметры распределений являются различными и равны соответственно λ_x и λ_y.

4.11. Известны математическое ожидание и дисперсия случайной величины X: $m_x = 2$, $D_x = 3$. Найти математическое ожидание и дисперсию случайной величины $Y = 3x - 5$.

4.12. Случайная величина X распределена равномерно в интервале [0; 2]. Найти математическое ожидание величины $Y = -X^2 + 3X - 2$.

4.13. Случайная величина X распределена равномерно в интервале [0; 1]. Найти дисперсию случайной величины $Y = 2X^2$.

4.14. Двухмерная случайная величина (X, Y) имеет следующие характеристики: $m_x = 2$, $m_y = 0$, $D_x = 1$, $D_y = 2$, $K_{xy} = -1$. Определить математическое ожидание и дисперсию случайной величины $Z = X - 2Y$.

4.15. Случайные величины X и Y имеют математические ожидания $m_x = -1$, $m_y = 1$ и дисперсии $D_x = 4$ и $D_y = 9$. Найти математическое ожидание случайной величины $Z = 3XY + 5$, если известен коэффициент корреляции $r_{xy} = 0{,}5$.

Тема 5. Закон больших чисел

5.1. С помощью неравенства Чебышева оценить вероятность того, что случайная величина X примет значение из интервала (3; 5), симметричного относительно математического ожидания величины X. Известно, что дисперсия X равна 0,36.

5.2. Оценить длину d интервала, в который может попасть измеренное значение случайной величины X с вероятностью не меньшей чем 0,9, если среднее квадратическое значение случайной величины равно 0,5. Считать, что указанный интервал является симметричным относительно математического ожидания исследуемой величины.

5.3. В дополнение к условию задачи 5.2 известно, что $m_x = 2$. Указать интервал, в который попадёт измеренное значение величины X с вероятностью не меньшей чем 0,9.

5.4. Среднее квадратическое отклонение каждой из 2500 независимых случайных величин не превосходит 3. Оценить вероятность того, что абсолютная величина отклонения среднего арифметического случайных величин от среднего арифметического их математических ожиданий не превзойдёт 0,3.

5.5. Вероятность появления события A в одном опыте равна 0,4. Можно ли с вероятностью большей 0,97 утверждать, что число k появлений события A в 1000 независимых испытаний будет в пределах от 300 до 500?

5.6. Сколько нужно провести последовательных независимых испытаний, чтобы с вероятностью не меньшей чем 0,95 можно было утверждать, что относительная частота появления события A отклонится от истинной вероятности события менее чем на 0,08?

Тема 6. Статистическая обработка экспериментальных данных. Оценка параметров

6.1. Задана выборка случайной величины X: 0,5; 0,8; 1,2; 1,4; 1,6; 1,7; 1,8; 2,0; 2,2; 2,5. Построить график эмпирической функции распределения вероятностей $F^*(x)$.

6.2. Задан интервальный вариационный ряд (в первой строке – границы интервалов, во второй строке – число точек, попавших в данный интервал):

0,4÷0,8	0,8÷1,2	1,2÷1,6	1,6÷2,0	2,0÷2,4	2,4÷2,8	2,8÷3,2	3,2÷3,6
2	5	8	14	16	10	8	7

Построить полигон и гистограмму распределения.

6.3. Задан сгруппированный вариационный ряд (в первой строке – возможные значения случайной величины, появившиеся в выборке, во второй строке – число таких значений в выборке):

0,2	0,4	0,8	0,9	1,0	1,1	1,2	1,3	1,5	1,8
5	7	12	15	20	18	13	5	3	2

Вычислить эмпирические характеристики: математическое ожидание, дисперсию и несмещённую дисперсию.

6.4. По интервальному вариационному ряду, заданному в задаче 6.2, определить эмпирические числовые характеристики: математическое ожидание, смещённую и несмещённую дисперсии.

6.5. При исследовании эффективности работы системы массового обслуживания были зафиксированы интервалы времени обслуживания 60 заявок:

0,5 0,6 1,4 0,8 1,0 1,8 0,2 0,4 0,1 0,3 1,1 0,9 0,7 0,2 1,2
0,1 0,6 0,4 0,8 1,2 0,3 1,7 0,2 1,6 0,5 0,2 0,1 1,5 1,0 0,9
1,3 0,4 1,6 0,3 0,1 0,6 1,5 0,1 0,5 0,8 1,1 0,7 1,1 0,6 0,5
0,7 0,4 1,4 0,6 0,5 1,3 0,3 1,2 0,2 1,0 0,1 0,8 0,4 0,6 0,1

Вычислить эмпирические характеристики времени обслуживания заявки: математическое ожидание, дисперсию, асимметрию и эксцесс.

6.6. Известно, что средние квадратические отклонения измеренного значения от истинного для четырёх дальномеров равны соответственно 0,5; 0,8; 1,0 и 1,2. Измерения до цели этими дальномерами показали соответственно следующие результаты: 12,6; 13,2; 11,8 и 12,2. Оценить среднее расстояние до цели по этим данным.

6.7. Найти оценку наибольшего правдоподобия математического ожидания случайной величины, распределённой по закону Пуассона.

6.8. Построить доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$ для неизвестного математического ожидания случайной величины X, если известно, что она имеет нормальное распределение с дисперсией $\sigma^2 = 9$, а оценка математического ожидания по выборке объёма $n = 100$ равна $m^* = 4{,}6$.

6.9. Построить доверительный интервал с доверительной вероятностью $\gamma = 0{,}9$ для неизвестного математического ожидания случайной величины X по вариационному ряду, представленному в задаче 6.2.

6.10. По выборке, представленной в задаче 6.3, построить доверительный интервал для дисперсии с доверительной вероятностью $\gamma = 0{,}95$.

6.11. Для оценки вероятности осуществления некоторого события A было произведено 50 последовательных независимых испытаний. В этих испытаниях событие A осуществилось 40 раз. Построить доверительный интервал для неизвестной вероятности события с доверительной вероятностью $\gamma = 0{,}9$. Воспользоваться тремя способами построения доверительного интервала.

Тема 7. Проверка статистических гипотез

7.1. Даны две выборки:

X: 0,5 0,6 1,4 0,8 1,0 1,8 0,2 0,4 0,1 0,3 1,1 0,9 0,7 0,2 1,2;

Y: 0,7 0,4 1,4 0,6 0,5 1,3 0,3 1,2 0,2 1,0.

По этим выборкам проверить гипотезу о том, что величины *X* и *Y* имеют одинаковые математические ожидания, если известны дисперсии величин: $\sigma_x^2 = 0{,}2$, $\sigma_y^2 = 0{,}25$. Уровень значимости принять равным $\alpha = 0{,}05$.

7.2. Даны две выборки:

X: 2,5 3,6 2,4 2,8 4,0 3,9 3,2 2,4 3,1 2,3 4,1 2,9;

Y: 2,7 3,1 2,1 2,5 4,2 3,9 3,3 2,2 3,6 2,4.

По этим выборкам проверить гипотезу о том, что величины *X* и *Y* имеют одинаковые математические ожидания, если дисперсии величин одинаковы: $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Уровень значимости принять равным $\alpha = 0{,}05$.

7.3. По данным примера 7.2 проверить гипотезу о равенстве дисперсий величин *X* и *Y* . Построить одностороннюю критическую область с уровнем значимости $\alpha = 0{,}05$.

7.4. Дана выборка случайной величины *X*:

2,5 2,8 3,6 3,7 2,4 4,5 2,8 4,3 4,0 3,9 3,2 2,4 3,1 2,3 4,1 2,9.

С помощью критерия квантилей проверить гипотезу о том, что медианой этого распределения является $Med_x = 3$. Уровень значимости принять равным $\alpha = 0{,}05$.

7.5. По вариационному ряду, представленному ниже, проверить гипотезу о том, что исследуемая случайная величина имеет экспоненциальное распределение. Уровень значимости принять равным $\alpha = 0{,}05$.

0÷4	4÷8	8÷12	12÷16	16÷20	20÷24	24÷28	28÷32
28	20	16	13	8	5	6	4

В первой строке представлены границы интервалов, во второй – число точек, попавших в данный интервал.

7.6. По вариационному ряду, представленному ниже, проверить гипотезу о том, что исследуемая случайная величина имеет нормальное распределение. Уровень значимости принять равным $\alpha = 0{,}05$.

0÷2	2÷4	4÷6	6÷8	8÷10	10÷12	12÷14	14÷16
5	7	10	18	16	12	8	4

В первой строке представлены границы интервалов, во второй строке – число точек, попавших в данный интервал.

Тема 8. Дисперсионный анализ

8.1. Сделано по 5 измерений случайной величины *X* на каждом из четырёх уровней фактора *A*. Полученные результаты представлены в следующей таблице:

Уровни фактора	Номер измерения				
	1	2	3	4	5
A_1	28	32	36	34	32
A_2	34	36	38	32	35
A_3	30	29	31	30	33
A_4	36	35	34	36	38

Методом дисперсионного анализа проверить гипотезу о том, что фактор A не влияет на математическое ожидание величины X. Уровень значимости принять равным $\alpha = 0{,}05$.

8.2. Фактор A имеет 4 уровня, фактор B – 5 уровней. Сделано по одному измерению случайной величины X на каждой комбинации уровней факторов. Полученные результаты представлены в следующей таблице:

Уровни фактора A	Уровни фактора B				
	B_1	B_2	B_3	B_4	B_5
A_1	38	32	46	44	35
A_2	44	41	45	42	33
A_3	32	33	40	37	33
A_4	31	36	38	36	34

Методом дисперсионного анализа проверить гипотезу о том, что факторы A и B не влияют на математическое ожидание величины X. Предполагать, что взаимодействия между факторами нет. Уровень значимости принять равным $\alpha = 0{,}05$.

8.3. Проверить гипотезу о влиянии факторов A, B, и C на исследуемую величину X, если все факторы имеют по четыре уровня, а измерения сделаны по плану эксперимента «латинский квадрат». Результаты измерений представлены в следующей таблице:

Уровни фактора A	Уровни фактора B			
	B_1	B_2	B_3	B_4
A_1	5,5	4,4	6,1	5,8
A_2	5,1	5,1	5,3	6,3
A_3	5,8	4,3	5,8	5,9
A_4	5,0	4,8	5,4	6,6

Предполагать, что взаимодействия между факторами нет. Уровень значимости принять равным $\alpha = 0{,}05$.

Тема 9. Регрессионный анализ

9.1. В результате эксперимента получена следующая таблица соответствующих значений случайных величин X и Y:

X: 1,5 1,7 1,4 2,2 2,7 2,1 2,0 1,8 2,8 3,0 1,2 1,3 1,8 1,9 2,9;

Y: 2,2 2,5 2,1 2,4 4,9 3,2 3,0 2,6 5,2 6,0 2,0 2,1 2,7 2,8 5,6.

Аппроксимировать статистическую зависимость величины Y от X функцией $Y = ax + b$. Вычислить остаточную дисперсию, найти оценку коэффициента корреляции.

9.2. По данным, представленным в примере 9.1, аппроксимировать статистическую зависимость величины Y от X функцией $Y = ax^2 + bx + c$. Вычислить остаточную дисперсию, найти оценку корреляционного отношения. Сравнить полученные результаты с результатами примера 9.1.

9.3. По данным, представленным ниже, аппроксимировать статистическую зависимость величины Y от X и Z функцией $y = ax + bz + c$. Вычислить остаточную дисперсию, найти оценку обобщённого коэффициента корреляции.

X: 1,1 1,1 1,2 1,2 1,3 1,4 1,6 1,8 2,1 2,2 2,2 2,4 2,5 2,6 2,8;

Z: 1,3 1,4 1,5 1,9 1,4 1,3 2,1 1,5 1,8 1,6 2,1 2,2 2,1 2,5 2,9;

Y: 0,1 0,5 0,8 1,8 1,2 1,7 3,6 2,1 3,6 3,5 4,7 5,4 5,3 6,7 8,8.

3. Задания для типовых расчётов

Тема 1. Случайные события

Задание 1. Выполнить действия, указанные в задаче, пользуясь операциями над событиями и их свойствами.

1. Доказать тождество: $(B+C)(B+\overline{C})(\overline{B}+C)=BC$.
2. Монета бросается до первого появления герба. A_k – событие, состоящее в том, что герб появится при k-ом броске, B – событие, состоящее в том, что до первого появления герба придётся сделать не менее 3 бросков. Выразить B через события A_k.
3. Сделано 3 выстрела по мишени. Событие A_k – попадание при k-том выстреле, событие B – две пули попали в мишень. Выразить событие B через события A_k.
4. Упростить выражение: $A(\overline{B}+C)(A+B)(A+\overline{C})$.
5. Брошены две игральные кости. Событие A_i – на 1-ой кости выпало i очков, событие B_k – на 2-ой кости выпало k очков (i, k = 1, 2, …, 6), событие C – сумма выпавших очков равна 10. Выразить событие C через события A_i и B_k.
6. Четверо студентов сдают экзамен по математике. Событие A_k – k-й студент успешно сдаёт экзамен, событие B – только три студента смогли успешно сдать экзамен. Выразить событие B через события A_k.
7. Известно, что события A и B несовместны. Чему равно в таком случае выражение $(A+B)(A+C)(A+\overline{C})$?
8. Три детали случайным образом размещаются по трём ящикам. Событие A_{ik} – i-я деталь попадает в k-й ящик, событие B – 3-ий ящик после размещения деталей оказывается пустым. Выразить событие B через события A_{ik}.
9. Известно, что события A, B и C составляют полную группу событий. Чему в таком случае равно выражение $A+AB+AC+BC+B\overline{A}+C\overline{A}$?
10. Упростить выражение: $(\overline{A}+B)(A+B)(A+\overline{B})$.
11. В коробке находятся красные, синие и жёлтые шары. Из ящика наудачу извлекается 3 шара. Пусть A_k, B_k и C_k – события, состоящие в том, что k-й извлечённый шар имеет соответсвенно красный, синий и жёлтый цвет, а событие D – в числе извлечённых шаров только один красный. Выразить событие D через события A_k, B_k и C_k (k = 1, 2, 3).
12. Известно, что $A \subset B$. Чему в таком случае равно выражение $(AC + A)(BC + A)(C + B)$?
13. Шахматист играет в шахматы со своим партнёром 4 партии. Событие A_k – он выигрывает k-ую партию, событие B – всего он выиграет 2 партии. Выразить событие B через события A_k (k = 1, 2, 3, 4).
14. Доказать тождество: $(B+A)(B+\overline{A})(\overline{B}+A)=AB$.
15. В составе установки 2 блока одного типа и 3-го блока второго типа. Событие A_i – исправен i-й блок 1-го типа (i = 1, 2), B_k – исправен k-тый блок второго типа (k = 1, 2, 3), событие D – установка работоспособьна. Выразить событие D через A_i и B_k, если установка работоспособна в случае исправности хотя бы одного блока первого типа и хотя бы двух – второго типа.
16. Упростить выражение: $C(\overline{B}+C)(A+B)(B+\overline{C})$.
17. Монета бросается 4 раза. A_k – событие, состоящее в том, что герб появится при k-том броске, B – событие, состоящее в том, что герб появится 3 раза. Выразить событие B через события A_k.

18. Известно, что события A и B несовместны. Чему равно в таком случае выражение $(A+\overline{B})(A+C)(A+C)$?
19. Брошены две игральные кости. Событие A_i – на 1-й кости выпало i очков, событие B_k – на 2-й кости выпало k очков (i, k = 1, 2, …, 6), событие C – произведение выпавших очков равно 12. Выразить событие C через события A_i и B_k.
20. В коробке находятся красные, синие и жёлтые шары. Из ящика наудачу извлекается 3 шара. Пусть A_k, B_k и C_k – события, состоящие в том, что k-й извлечённый шар имеет соответсвенно красный, синий и жёлтый цвет, а событие D – в числе извлечённых шаров только один синий. Выразить событие D через события A_k, B_k и C_k (k = 1, 2, 3).
21. Известно, что события A, B и C составляют полную группу событий. Чему в таком случае равно выражение $B+AB+AC+BC+A\overline{B}+C\overline{B}$?
22. Шахматист играет в шахматы со своим партнёром 3 партии. Событие A_k – он выигрывает k-ю партию, событие B – всего он выиграет одну партию. Выразить событие B через события A_k (k = 1, 2, 3, 4).
23. Известно, что $A \subset B$. Чему в таком случае равно выражение $(AB + B)(AC + B)(A + C)$?
24. Трое студентов сдают экзамен по математике. Событие A_k – k-й студент успешно сдаёт экзамен, событие B – только один студент смог успешно сдать экзамен. Выразить событие B через события A_k.
25. Чему равно выражение $(B+C)(B+\overline{C})(\overline{B}+C)$.

Задание 2. Вычислить вероятности событий, указанных в тексте.

1. За круглым столом сидят 5 мужчин и 5 женщин. Какова вероятность того, что два лица одинакового пола не сидят рядом, если места занимались случайно?
2. На столе лежат 20 экзаменационных билетов с номерами 1, 2, ..., 20. Преподаватель берёт 3 любых билета. Какова вероятность того, что они из первых четырёх?
3. Имеется 6 отрезков, длины которых равны соответственно 2, 4, 6, 8, 10, 12 единицам. Найти вероятность того, что с помощью взятых наугад трёх отрезков можно построить треугольник.
4. Пять студентов из группы изучают английский язык, шесть студентов – немецкий и семь студентов – французский язык. Случайным образом выбрано четыре студента. Какова вероятность того, что двое из них изучают английский язык, один изучает французский и один – немецкий?
5. На семи карточках написаны цифры от 1 до 7. Наудачу извлекаются две карточки. Какова вероятность того, что сумма цифр на этих карточках будет чётной?
6. В мастерскую для ремонта поступило 10 телевизоров, из которых 3 нуждаются в общем ремонте. Мастер наугад берёт первые 5 штук. Какова вероятность того, что два из них нуждаются в общем ремонте?
7. Бросаются две игральные кости. Найти вероятность того, что выпадет одинаковое число очков на обеих костях, и вероятность того, что на обеих костях выпадет чётное число очков.
8. Из полной колоды карт (52 карты) вынимается наугад три карты. Найти вероятность того, что этими картами будут тройка, семёрка и туз.
9. Телефонный номер состоит из пяти цифр. Найти вероятность того, что номер телефона случайно выбранного абонента не содержит одинаковых цифр.
10. В шахматном турнире участвуют 20 человек, которые разбиваются на две группы по 10 человек. Определить вероятность того, что четыре наиболее сильных игрока разделятся между группами поровну.

11. Четырёхтомное сочинение расположено на полке в случайном порядке. Найти вероятность того, что тома стоят в должном порядке справа налево или слева направо.
12. Какова вероятность того, что четырехзначный номер случайно взятого автомобиля в большом городе имеет все цифры разные и вероятность того, что он имеет все цифры одинаковые?
13. Полная колода карт (52 карты) делится пополам. Найти вероятность того, что число чёрных и красных карт в обеих пачках будет одинаковым.
14. На полке лежат 15 учебников, из них 7 – по математике. Студент наудачу берёт 3 учебника. Какова вероятность того, что взятые учебники – учебники по математике?
15. Игральная кость бросается два раза. Какова вероятность того, что сумма выпавших очков будет не менее 7 и не более 10?
16. В урне 2 белых и 4 чёрных шара. Из урны парами последовательно извлекают все шары. Какова вероятность того, что в последней паре оба шара будут чёрными?
17. Студент знает 15 из 20 вопросов учебной программы. На экзамене предлагается ответить на 3 вопроса, которые выбираются случайным образом. Какова вероятность того, что студент сможет ответить на предложенные вопросы?
18. Отрезок прямой, длина которого равна 2, делится случайным образом на 3 части. Найти вероятность того, что из полученных частей можно построить треугольник.
19. Спортивная команда состоит из 20 спортсменов, из которых 5 боксёров, 7 штангистов и 8 борцов. Для беседы с журналистом было выбрано случайным образом 3 спортсмена. Определить вероятность того, что выбранные спортсмены представляют различные дисциплины спорта.
20. На восьми карточках написаны цифры от 1 до 8. Наудачу извлекаются две карточки. Какова вероятность того, что сумма цифр, написанных на этих карточках, будет не менее 12?
21. Бросаются две игральные кости. Найти вероятность того, что сумма выпавших очков будет не более чем 10.
22. Каждая из цифр 1, 3, 5, 6 и 8 написана на одной из пяти карточек. Карточки перемешивают и раскладывают в ряд. Какова вероятность того, что полученное пятизначное число будет делиться на 4?
23. Наугад выбирается двухзначное число. Определить вероятность того, что сумма цифр этого числа является простым числом.
24. Из полного набора домино (28 штук) наудачу выбирают 7 костей. Какова вероятность того, что среди них окажется три кости с шестью очками?
25. Группа, состоящая из 8 человек, занимает места за круглым столом в случайном порядке. Какова вероятность того, что при этом два определенных лица окажутся сидящими рядом?

Задание 3. Вычислить вероятности событий, указанных в тексте.

1. В первом ящике находятся шары с номерами от 1 до 5, а во втором – с номерами от 6 до 10. Из каждого ящика вынули по одному шару. Какова вероятность того, что сумма номеров вынутых шаров равна 11?
2. В случайный момент времени $X \in [0, T]$ появляется радиосигнал длительностью t_1. В случайный момент времени $Y \in [0, T]$ включается приемник на время $t_2 < t_1$. Найти вероятность обнаружения сигнала, если приемник настраивается мгновенно.
3. Две игральные кости бросаются один раз. Найти вероятность того, что сумма выпавших очков является простым числом.

4. Время прихода двух пароходов независимо равновозможное в течение данных суток. Определить вероятность того, что одному из пароходов придется ожидать освобождения причала, если время стоянки первого парохода – 4 часа, а второго – 2 часа.
5. Даны две концентрические окружности радиусов $r_1 = 2$ и $r_2 = 1$. На большей окружности наудачу ставятся две точки A и B. Какова вероятность того, что отрезок AB не пересечет малую окружность?
6. В первом ящике находятся шары с номерами от 1 до 5, а во втором – с номерами от 6 до 10. Из каждого ящика вынули по одному шару. Какова вероятность того, что произведение номеров вынутых шаров будет не меньше 20?
7. Найти вероятность того, что при случайном размещении 3 шаров по 3 ящикам один ящик окажется пустым.
8. Из отрезка [–1, 2] наудачу взяты два числа. Какова вероятность того, что их сумма больше единицы, а модуль разности меньше единицы?
9. Два лица условились встретиться в определённом месте между 12 и 13 часами. Пришедший первым ждёт второго в течении 15 минут, после чего уходит. Определить вероятность того, что встреча состоится, если каждый может прибыть на место встречи в любой момент в течении указанного часа, а моменты времени их прихода независимы.
10. Бросаются три игральные кости. Определить вероятность того, что сумма выпавших очков будет равна 12.
11. На отрезке длиной d наудачу выбраны две точки. Какова вероятность того, что расстояние между ними будет меньше $0{,}5d$?
12. Два игрока по очереди бросают игральную кость, каждый по одному разу. Выигравшим считается тот, кто получит большее число очков. Найти вероятность выигрыша первого игрока.
13. Из отрезка [0, 2] наудачу взяты два числа. Какова вероятность того, что их сумма меньше 3, а модуль разности больше единицы?
14. На отрезке длиной $2d$ наудачу выбраны две точки. Какова вероятность того, что расстояние между ними будет больше d?
15. Зенитная батарея, состоящая из 4 орудий, производит залп по группе, состоящей из 4 самолётов. Каждое орудие выбирает себе цель наудачу независимо от остальных. Какова вероятность того, что все орудия выстрелят по разным самолётам?
16. Код кодового замка состоит из трёх цифр. Некто, не зная кода, стал наудачу пробовать различные комбинации из трёх цифр. На каждую попытку он тратит по 20 секунд. Какова вероятность того, что замок будет открыт за один час?
17. На плоскости начерчены параллельные прямые, находящиеся на расстоянии $2d$ друг от друга. На плоскость наудачу бросается монета радиуса $r < d$. Найти вероятность того, что монета не пересечёт ни одной прямой.
18. В группе из 20 студентов 5 являются отличниками, 9 – хорошистами и 6 студентов учатся удовлетворительно. Комиссия по проверке качества подготовки студентов наудачу выбирает четырёх студентов из этой группы. Какова вероятность того, что среди выбранных окажется один отличник, два хорошиста и один студент, обучающийся на оценку удовлетворительно?
19. Точка взята наудачу внутри круга радиуса R. Найти вероятность того, что эта точка окажется внутри вписанного в круг правильного треугольника.
20. На отрезке AB длиной d поставлены наудачу две точки L и M . Найти вероятность того, что точка L окажется ближе к точке A, чем к точке M.

21. Колода карт (36 карт) делится пополам. Найти вероятность того, что в одной пачке чёрных карт будет в 2 раза больше, чем красных.
22. Точка взята наудачу внутри круга радиуса R. Найти вероятность того, что эта точка окажется внутри вписанного в круг квадрата.
23. Из колоды карт (36 карт) вынимается наудачу три карты. Определить вероятность того, что вынутыми картами будут дама, семёрка и туз.
24. На бесконечную шахматную доску со стороной квадрата d бросается наудачу монета радиуса r, $2r < d$. Найти вероятность того, что монета целиком попадёт внутрь одного квадрата.
25. Найти вероятность того, что дни рождения 12 человек приходятся на разные месяцы года (предполагается, что день рождения каждого человека с равными вероятностями может приходиться на любой месяц года).

Задание 4. Вычислить вероятности событий, пользуясь формулами сложения и (или) умножения вероятностей.

1. Три стрелка одновременно делают по одному выстрелу по мишени. Какова вероятность того, что мишень будет поражена только одной пулей, если вероятность попадания в мишень для первого стрелка равна 0,8, для второго – 0,7, для третьего – 0,6?
2. В классе учатся 10 мальчиков и 8 девочек. По жребию выбирают 5 учеников этого класса. Какова вероятность того, что среди них окажется не менее трёх девочек?
3. Из сосуда, содержащего 2 белых и 4 чёрных шара, двое поочерёдно извлекают шар (без возвращения). Найти вероятность того, что каждый из участников вынет первым белый шар.
4. В урне 4 белых и 6 чёрных шаров. Из урны наугад извлекают 4 шара. Какова вероятность того, что среди них будет хотя бы два чёрных шара?
5. Двое поочерёдно бросают монету. Выигрывает тот, который первым в результате получит орла. Какова вероятность того, что игра закончится не позднее чем после второго бросания монеты вторым игроком?
6. Вероятности попадания в цель при стрельбе из трёх орудий соответственно равны: 0,8; 0,7; 0,9. Вычислить вероятность двух попаданий при одном залпе из всех орудий.
7. Среди 100 лотерейных билетов есть 5 выигрышных. Найти вероятность того, что среди трёх наудачу выбранных билетов будет не менее двух выигрышных.
8. Чтобы получить положительную оценку на экзамене студент должен ответить по меньшей мере на три вопроса из предложенных пяти. Какова вероятность того, что студент успешно выдержит экзамен, если он знает 20 из 30 вопросов программы?
9. В ящике 12 деталей, среди которых 8 окрашенных. Сборщик наудачу извлекает 4 детали. Найти вероятность того, что хотя бы две детали из взятых будут окрашенными.
10. Автобусный маршрут обслуживается тремя автобусами. Вероятности возникновения неисправностей автобусов на маршруте в течении смены равны соответственно: 0,2; 0,1; 0,08. Определить вероятность того, что в течении смены неисправность возникнет только у одного автобуса.
11. Три баскетболиста производят по одному броску мяча. Вероятность попадания мяча в корзину для первого, второго и третьего баскетболиста равны соответственно 0,9, 0,8 и 0,7. Найти вероятность того, что удачными будут только два броска.
12. Студент знает 20 из 25 вопросов программы. Зачет считается сданным, если студент ответит не менее чем на три из четырех поставленных в билете вопросов. Взглянув на первый вопрос билета, студент обнаружил, что он его знает. Какова вероятность того, что студент сдаст зачет?

13. Радист трижды вызывает корреспондента. Вероятность того, что будет принят первый вызов, равна 0,2, второй – 0,3, третий вызов – 0,4. По условиям приёма, события, состоящие в том, что данный вызов будет услышан, независимы. Найти вероятность того, что корреспондент вообще услышит вызов.
14. В ящике 12 деталей, среди которых 8 высшего качества. Сборщик последовательно наудачу извлекает из ящика по одной детали до первого появления детали высшего качества. Найти вероятность того, что будет произведено не более трёх извлечений.
15. В ящике имеются 10 монет по 20 копеек, 5 монет по 15 коп. И 2 монеты по 10 коп. Наугад берётся 6 монет. Какова вероятность того, что в сумме они составят не более одного рубля?
16. В двух урнах находятся шары, причём в первой урне 5 белых шаров, 11 чёрных и 8 красных, а во второй соответственно 10, 8 и 6. Из обеих урн наудачу извлекается по одному шару. Какова вероятность того, что оба шара одного цвета?
17. Вероятность улучшить свой предыдущий результат с одной попытки для данного спортсмена равна p. Определить вероятность того, что на соревнованиях спортсмен улучшит свой результат, если разрешается делать две попытки.
18. Два стрелка поочерёдно стреляют по мишени до первого попадания. Вероятность попадания для первого стрелка равна 0,2, а для второго – 0,3. Найти вероятность того, что всего будет сделано 2 или 3 выстрела.
19. В лотерее выпущено 100 билетов, из которых 20 – выигрышные. Куплено 5 билетов. Какова вероятность того, что среди купленных билетов не менее двух являются выигрышными?
20. Жюри состоит из трех судей. Первый и второй судьи принимают правильное решение независимо друг от друга с вероятностью 0,8, а третий судья для решения бросает монету. Окончательное решение жюри принимает по большинству голосов. Какова вероятность того, что жюри примет правильное решение?
21. Пять студентов из группы изучают английский язык, шесть студентов – немецкий и семь студентов – французский язык. Случайным образом выбрано три студента. Какова вероятность того, что два из них изучают один и тот же иностранный язык?
22. Вероятность попадания в цель для первого орудия равна 0,3, для второго – 0,4, для третьего – 0,6. Для разрушения цели требуется хотя бы 2 попадания. Какова вероятность того, что при одном залпе из всех орудий цель будет разрушена?
23. Три стрелка одновременно делают по одному выстрелу по мишени. Какова вероятность того, что мишень будет поражена не менее чем двумя пулями, если вероятность попадания в мишень для первого стрелка равна 0,5, для второго – 0,6, для третьего – 0,8?
24. Два стрелка поочерёдно стреляют по мишени до первого попадания. Вероятность попадания для первого стрелка равна 0,3, а для второго – 0,4. Найти вероятность того, мишень будет поражена первым стрелком.
25. Двое поочерёдно бросают монету. Выигрывает тот, который первым в результате получит орла. Какова вероятность того, что выиграет игрок, бросающий монету первым?

Задание 5. Вычислить вероятности событий, пользуясь формулой полной вероятности и (или) формулой Байеса.

1. На заводе 30% деталей производится цехом №1, 45% – цехом №2 и 25% – цехом №3. Вероятность изготовления бракованной детали для 1-ого цеха равна 0,05, для 2-го – 0,01, для 3-го – 0,04. Наугад выбранная из общего потока деталь оказалась бракованной. Определить вероятность того, что эта деталь была изготовлена 1-м цехом.

2. Трое охотников одновременно выстрелили по медведю, который был убит одной пулей. Определить вероятность того, что медведь был убит первым охотником, если вероятности попадания для охотников равны соответственно 0,4, 0,35 и 0,3.
3. В урне находится два белых и четыре чёрных шара. Из урны извлекают два шара, цвет которых остаётся неизвестным, и откладывают их в сторону, после чего вынимают третий шар. Определить вероятность того, что этот шар белый.
4. Известно, что 5% всех мужчин и 0,25% всех женщин – дальтоники. На обследование прибыло одинаковое число мужчин и женщин. Наудачу выбранное лицо оказалось дальтоником. Какова вероятность того, что это мужчина?
5. В продажу поступают телевизоры трех заводов. Продукция первого завода содержит 20% телевизоров со скрытым дефектом, второго – 10% и третьего – 5%. Какова вероятность приобрести исправный телевизор, если в магазин поступило 30% телевизоров с первого завода, 20% – со второго и 50% – с третьего?
6. Два цеха штампуют однотипные детали. Первый цех дает α% брака, а второй – β%. Для контроля отобрано n_1 деталей из первого цеха и n_2 из второго. Эти $n_1 + n_2$ деталей смешаны в одну партию, и из нее наудачу извлекают одну деталь. Какова вероятность того, что она бракованная?
7. В ящике лежит 20 теннисных мячей, в том числе 15 новых. Для игры наудачу выбираются два мяча и после игры возвращаются обратно. Затем для второй игры также наудачу извлекаются еще два мяча. Какова вероятность того, что вторая игра будет проводиться новыми мячами?
8. В тире имеется 5 ружей, вероятность попадания из которых равна 0,5, три ружья с вероятностью попадания 0,7 и два ружья с вероятностью попадания 0,8. Определить вероятность попадания в мишень при одном выстреле, если стреляющий берёт одно из ружей наудачу.
9. Для контроля продукции из трёх партий деталей взята для испытания одна деталь. Как велика вероятность обнаружения бракованной продукции, если в первой партии 2/3 деталей бракованные, во второй – 1/3, а в третьей – все детали доброкачественные?
10. Из полного набора костей домино наугад берутся две кости. Определить вероятность того, что вторую кость можно приставить к первой.
11. На двух станках обрабатываются однотипные детали. Вероятность брака для 1-го станка составляет 0,03, для второго – 0,02. Обработанные детали поступают на общий конвейер. Наудачу взятая с конвейера деталь оказалась без брака. Определить вероятность того, что эта деталь была обработана на 1-м станке.
12. В первой урне находится 6 белых и 4 чёрных шаров, а во второй – 5 белых и 4 чёрных. Из первой урны во вторую переложили один шар, после чего из второй урны извлекли один шар, оказавшийся белым. Какова вероятность того, что из первой урны во вторую был переложен белый шар?
13. Группа из 15 спортсменов стрелков включает 3 мастера спорта, 6 кандидатов в мастера и 6 перворазрядников. Вероятности поражения мишени для спортсменов равны соответственно 0,9, 0,8 и 0,7. Наудачу выбранный стрелок сделал выстрел и поразил мишень. Какова вероятность того, этот стрелок является мастером спорта?
14. Три стрелка произвели залп, причём две пули поразили мишень. Найти вероятность того, что третий стрелок попал в цель, если вероятности попадания в мишень первым, вторым и третьим стрелками равны соответственно 0,6, 0,5, 0,4.
15. Вероятность попадания в цель при одном выстреле равна 0,4. Вероятность поражения цели при одном попадании равна 0,1, при двух попаданиях – 0,3, при трёх попада-

ниях – 0,6. Произведено три выстрела, в результате которых цель была поражена. Каково наиболее вероятное число попаданий?

16. В партии из 5 изделий количество бракованных изделий может быть любым с одинаковой вероятностью. Из партии взято одно изделие, оказавшееся бракованным. Какое предположение о количестве бракованных изделий наиболее вероятно, если принять во внимание результат этого опыта?
17. В первой урне находится 6 белых и 4 чёрных шаров, а во второй – 4 белых и 6 чёрных. Из каждой урны наудачу извлекли по одному шару, а затем из этих двух шаров наудачу взят один шар. Найти вероятность того, что взят белый шар.
18. На сборку поступают детали с двух автоматов. Первый даёт в среднем 0,2% брака, второй – 0,1%. Найти вероятность попадания на сборку бракованной детали, если с первого автомата поступило 2000 деталей, а со второго – 3000.
19. В трёх одинаковых коробках находится по 10 деталей, включая бракованные. В первой коробке – 2 детали бракованные, во второй – 3, в третьей – 1. Некто выбирает наугад одну из коробок и наудачу берёт из нее деталь. Какова вероятность того, что эта деталь окажется бракованной?
20. В сосуд, содержащий 5 шаров, опущен белый шар. Какова вероятность, действуя наудачу, вынуть белый шар из этого сосуда? Все предположения о первоначальном составе шаров считаются одинаково возможными.
21. Радиолампа может принадлежать к одной из трёх партий с вероятностями 0,2, 0,3 и 0,5. Вероятности того, что лампа проработает заданное число часов для этих партий, равны соответственно 0,7, 0,8 и 0,9. Определить вероятность того, что лампа проработает заданное число часов.
22. Определить вероятность того, что 100 лампочек, взятых наудачу из 1000, окажутся исправными, если известно, что число испорченных лампочек на 1000 штук может быть равным от 0 до 5 с одинаковой вероятностью.
23. В продажу поступают телевизоры трех заводов. Продукция первого завода содержит 10% телевизоров со скрытым дефектом, второго – 8% и третьего – 5%. Приобретённый телевизор оказался без дефектов. Какова вероятность того, что этот телевизор был изготовлен на первом заводе, если в магазин поступило 30% телевизоров с первого завода, 20% – со второго и 50% – с третьего?
24. В тире имеется 5 ружей, вероятность попадания из которых равна 0,5, три ружья с вероятностью попадания 0,7 и два ружья с вероятностью попадания 0,8. Сделан выстрел из наудачу взятого ружья, при этом зафиксировано попадание в цель. Определить вероятность того, что это ружьё принадлежало первой группе.
25. Известно, что 95% выпускаемой продукции удовлетворяет стандарту. Упрощенная схема контроля признаёт пригодной стандартную продукцию с вероятностью 0,9 и нестандартную – с вероятностью 0,08. Определить вероятность того, что изделие, прошедшее упрощённый контроль, удовлетворяет стандарту.

Задание 6. Вычислить вероятности событий, пользуясь формулой Бернулли, следствиями из неё, или её асимптотическими приближениями.

1. Вероятность отказа прибора при испытании равна 0,2. Сколько таких приборов нужно испытать, чтобы с вероятностью 0,99 получить хотя бы один отказ?
2. Вероятность изготовления детали отличного качества равна 0,8. Какова вероятность того, что среди 10 деталей не менее 9 отличного качества?

3. Если в среднем левши составляют 1%, какова вероятность того, что среди 200 человек окажется четверо левшей? Какова вероятность среди 200 человек обнаружить не менее 4 левшей?
4. В некотором семействе 8 детей. Вероятность рождения мальчика или девочки равна 0,5. Найти вероятность того, что а) имеется 4 мальчика и 4 девочки; б) число мальчиков заключено между 2 и 6 (включительно).
5. Вероятность хотя бы одного появления события при четырёх независимых испытаниях равна 0,59. Какова вероятность появления события при одном опыте, если при каждом опыте эта вероятность одинакова?
6. Батарея дала 14 выстрелов по военному объекту, вероятность попадания в который равна 0,2. Найти а) наивероятнейшее число попаданий и его вероятность: б) вероятность разрушения объекта, если для его разрушения требуется не менее 4 попаданий.
7. В некотором обществе имеется 1% дальтоников. Каков должен быть объём случайной выборки (с возвращением), чтобы вероятность встретить в ней хотя бы одного дальтоника была не менее 0,95?
8. Игральная кость бросается 6 раз. Найти вероятность того, что хотя бы два раза появится число очков, кратное трём.
9. Вероятность попадания в цель равна 0,2. Сбрасывается одиночно 7 бомб. Найти вероятность того, что будет а) не менее 6 попаданий; б) не менее 2 попаданий.
10. На отрезок [0, 10] наудачу брошено 5 точек. Найти вероятность того, что две точки попадут в отрезок [0, 2], одна – в [2, 3], две – в [3, 10].
11. Вероятность попадания в цель при каждом выстреле равна 0,001. Найти вероятность попадания в цель двумя и более выстрелами, если сделано 5000 выстрелов.
12. В круг вписан квадрат. Какова вероятность того, что из 8 точек, брошенных наудачу в круг, три попадут в квадрат, две – в один сегмент, три – в оставшиеся три сегмента?
13. Станок-автомат штампует детали. Вероятность того, что изготовленная деталь окажется бракованной, равна 0,02. Найти вероятность того, что среди 100 деталей бракованных окажется не менее 2 и не более 4.
14. Игральная кость бросается 12000 раз. Найти вероятность того, что число выпадений одного очка будет заключено между 1900 и 2150.
15. Система радиолокационных станций ведёт наблюдение за группой объектов, состоящей из 10 единиц. Каждый из объектов может быть (независимо от других) потерян с вероятностью 0,1. Найти вероятность того, что не менее двух объектов будет потеряно.
16. Прибор состоит из 8 однотипных элементов, но может работать при наличии в исправном состоянии не менее 6 из них. Каждый из элементов за время t работы прибора выходит из строя независимо от других с вероятностью 0,2. Найти вероятность того, что прибор откажет за время t.
17. Вероятность выигрыша по одному лотерейному билету равна 0,005. Найти вероятность того, что, имея 100 билетов, можно выиграть хотя бы по двум из них.
18. Тестовый билет содержит 5 вопросов. На каждый вопрос предлагается два варианта ответа, из которых один является правильным. Для получения зачёта нужно правильно ответить не менее чем на три вопроса. Какова вероятность того, что некто, выбирая ответы наудачу, может получить зачёт?
19. Вероятность хотя бы одного выигрыша по 5 лотерейным билетам равна 0,1. Какова вероятность выигрыша по одному билету, если предполагать, что для всех билетов эта вероятность одинакова?

20. Доля изделий первого сорта в продукции завода составляет 70%. Какова вероятность того, что из отобранных 400 изделий окажется от 270 до 300 изделий первого сорта?
21. Сколько нужно взять случайных цифр, чтобы вероятность появления среди них цифры, кратной трём, была бы не менее 0,9?
22. Вероятность появления события в одном опыте равна 0,6. Какова вероятность того, что это событие появится в большинстве из 60 опытов?
23. Вероятность изготовления детали отличного качества равна 0,9. Какова вероятность того, что среди 8 деталей не более 6 отличного качества?
24. Батарея дала 10 выстрелов по военному объекту, вероятность попадания в который равна 0,3. Найти а) наивероятнейшее число попаданий и его вероятность; б) вероятность разрушения объекта, если для его разрушения требуется не менее 3 попаданий.
25. Шахматист играет в шахматы со своим партнёром 4 партии. Вероятность выиграть партию равна для этого шахматиста 0,4. Найти вероятность того, что он выиграет не более 2 и не менее 3 партий.

Тема 2. Случайные величины

Задание 7. Задан закон распределения дискретной случайной величины *X*. Найти математическое ожидание, дисперсию, среднее квадратическое отклонение. Построить график функции распределения вероятностей случайной величины *X*.

1.

x	10	10,1	10,3	10,6	11
p	0,6	0,1	0,1	0,1	0,1

2.

x	10,1	10,2	10,4	10,7	11,1
p	0,5	0,2	0,1	0,1	0,1

3.

x	10,3	20,3	20,7	21,3	22,1
p	0,4	0,3	0,1	0,1	0,1

4.

x	10,6	20,6	21	21,6	22,4
p	0,3	0,3	0,2	0,1	0,1

5.

x	11	21	21,4	22	22,8
p	0,2	0,2	0,3	0,2	0,1

6.

x	11,5	21,5	21,9	22,5	23,3
p	0,1	0,1	0,3	0,4	0,1

7.

x	12,1	22,1	22,5	23,1	23,9
p	0,1	0,3	0,4	0,1	0,1

8.

x	12,8	22,8	23,2	23,8	24,6
p	0,1	0,2	0,4	0,2	0,1

9.

x	13,6	23,6	24	24,6	25,4
p	0,3	0,3	0,2	0,1	0,1

10.

x	14,5	24,5	24,9	25,5	26,3
p	0,3	0,4	0,1	0,1	0,1

11.

x	15,5	25,5	25,9	26,5	27,3
p	0,2	0,4	0,2	0,1	0,1

12.

x	16,6	26,6	27	27,6	28,4
p	0,1	0,5	0,1	0,2	0,1

13.

x	17,8	27,8	28,2	28,8	29,6
p	0,3	0,2	0,2	0,2	0,1

14.

x	19,1	29,1	29,5	30,1	30,9
p	0,2	0,1	0,1	0,1	0,5

15.

x	20,5	30,5	30,9	31,5	32,3
p	0,1	0,3	0,3	0,2	0,1

16.

x	20,4	30,4	30,8	31,4	32,2
p	0,1	0,2	0,2	0,3	0,2

17.

x	20,2	30,2	30,6	31,2	32
p	0,2	0,1	0,4	0,2	0,1

18.

x	19,9	29,9	30,3	30,9	31,7
p	0,3	0,3	0,2	0,1	0,1

19.

x	19,5	29,5	29,9	30,5	31,3
p	0,1	0,1	0,3	0,4	0,1

20.

x	19	29	29,4	30	30,8
p	0,2	0,3	0,2	0,2	0,1

21.

x	18,4	28,4	28,8	29,4	30,2
p	0,1	0,2	0,2	0,4	0,1

22.

x	17,7	27,7	28,1	28,7	29,5
p	0,2	0,3	0,2	0,2	0,1

23.

x	16,9	26,9	27,3	27,9	28,7
p	0,1	0,1	0,2	0,4	0,2

24.

x	16	26	26,4	27	27,8
p	0,3	0,3	0,2	0,1	0,1

25.

x	15	25	25,4	26	26,8
p	0,2	0,1	0,1	0,3	0,3

Задание 8. Непрерывная случайная величина X задана функцией распределения (задачи 1–14) или плотностью распределения вероятностей (задачи 15–25). Требуется: а) найти плотность распределения (1–14) или функцию распределения вероятностей (15–25); б) найти математическое ожидание, дисперсию, среднее квадратическое отклонение, скошенность и эксцесс распределения; вероятность того, что случайная величина отклонится от своего математического ожидания не более, чем на одну четвёртую длины всего интервала возможных значений этой величины; в) построить графики функций распределения и плотности распределения вероятностей.

1. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \dfrac{x^2}{4} & \text{при } 0<x\le 2,\\ 1 & \text{при } x>2.\end{cases}$

2. $F(x)=\begin{cases}0 & \text{при } x\le \dfrac{3}{4},\\ 4-\dfrac{3}{x} & \text{при } \dfrac{3}{4}<x\le 1,\\ 1 & \text{при } x>1.\end{cases}$

3. $F(x)=\begin{cases}0 & \text{при } x\le \dfrac{2}{3},\\ 3-\dfrac{2}{x} & \text{при } \dfrac{2}{3}<x\le 2,\\ 1 & \text{при } x>2.\end{cases}$

4. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ 2\sin x & \text{при } 0<x\le \dfrac{\pi}{6},\\ 1 & \text{при } x>\dfrac{\pi}{6}.\end{cases}$

5. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ e^x-1 & \text{при } 0<x\le \ln 2,\\ 1 & \text{при } x>2.\end{cases}$

6. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \sqrt{x} & \text{при } 0<x\le 1,\\ 1 & \text{при } x>1.\end{cases}$

7. $F(x)=\begin{cases}0 & \text{при } x\le \dfrac{\pi}{4},\\ -\cos 2x & \text{при } \dfrac{\pi}{4}<x\le \dfrac{\pi}{2},\\ 1 & \text{при } x>\dfrac{\pi}{2}.\end{cases}$

8. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ 2\sin\dfrac{x}{2} & \text{при } 0<x\le \dfrac{\pi}{3},\\ 1 & \text{при } x>\dfrac{\pi}{3}.\end{cases}$

9. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \dfrac{x^3}{8} & \text{при } 0<x\le 2,\\ 1 & \text{при } x>2.\end{cases}$

10. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \dfrac{x^3}{27} & \text{при } 0<x\le 3,\\ 1 & \text{при } x>3.\end{cases}$

11. $F(x)=\begin{cases}0 & \text{при } x\le 1,\\ \ln x & \text{при } 1<x\le e,\\ 1 & \text{при } x>e.\end{cases}$

12. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \sqrt[3]{x} & \text{при } 0<x\le 1,\\ 1 & \text{при } x>1.\end{cases}$

13. $F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \dfrac{x^2}{16} & \text{при } 0<x\le 4,\\ 1 & \text{при } x>4.\end{cases}$

14. $F(x)=\begin{cases}0 & \text{при } x\le 2,\\ \ln\dfrac{x}{2} & \text{при } 2<x\le 2\mathrm{e},\\ 1 & \text{при } x>2\mathrm{e}.\end{cases}$

15. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>\dfrac{\pi}{4},\\ 2\cos 2x & \text{при } 0<x\le\dfrac{\pi}{4}.\end{cases}$

16. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>1,\\ \dfrac{3}{2}\sqrt{x} & \text{при } 0<x\le 1.\end{cases}$

17. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>3,\\ \dfrac{x^2}{9} & \text{при } 0<x\le 3.\end{cases}$

18. $f(x)=\begin{cases}0 & \text{при } x\le 1 \text{ и } x>\sqrt{e},\\ \dfrac{2}{x} & \text{при } 1<x\le\sqrt{e}.\end{cases}$

19. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>1,\\ \dfrac{4}{3}\sqrt[3]{x} & \text{при } 0<x\le 1.\end{cases}$

20. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>3,\\ \dfrac{2}{3}\left(1-\dfrac{x}{3}\right) & \text{при } 0<x\le 3.\end{cases}$

21. $f(x)=\begin{cases}0 & \text{при } x\le\dfrac{3}{4} \text{ и } x>1,\\ \dfrac{6}{x^3}-\dfrac{4}{x^2} & \text{при } \dfrac{3}{4}<x\le 1.\end{cases}$

22. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>\dfrac{\pi}{3},\\ \cos\dfrac{x}{2} & \text{при } 0<x\le\dfrac{\pi}{3}.\end{cases}$

23. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>1,\\ \dfrac{1}{3\sqrt[3]{x^2}} & \text{при } 0<x\le 1.\end{cases}$

24. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>\dfrac{1}{4},\\ \dfrac{1}{\sqrt{x}} & \text{при } 0<x\le\dfrac{1}{4}.\end{cases}$

25. $f(x)=\begin{cases}0 & \text{при } x\le 0 \text{ и } x>4,\\ \dfrac{3x^2}{64} & \text{при } 0<x\le 4.\end{cases}$

Тема 3. Многомерные случайные величины

Задание 9. Закон распределения дискретной двухмерной случайной величины (X,Y) представлен таблицей. Определить одномерные законы распределения случайных величин X и Y. Найти условные плотности распределения вероятностей величин. Вычислить математические ожидания m_x и m_y, дисперсии σ_x и σ_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}.

x_i \ y_j	y_1	y_2	y_3	y_4	y_5
x_1	0,04	0,04	0,03	0,03	0,01
x_2	0,04	0,07	0,06	0,05	0,03
x_3	0,05	0,08	0,09	0,08	0,05
x_4	0,03	0,04	0,04	0,06	0,08

Возможные значения случайных величин выбрать по номеру варианта.

1. $X = (1; 2; 3; 8)$, $Y = (9; 11; 13; 16; 18)$
2. $X = (-2; -1; 1; 3)$, $Y = (6; 8; 12; 14; 16)$
3. $X = (-1; 2; 4; 6)$, $Y = (5; 7; 9; 11; 14)$
4. $X = (2; 3; 5; 8)$, $Y = (7; 9; 10; 11; 13)$
5. $X = (-3; -2; 2; 7)$, $Y = (8; 10; 12; 14; 16)$
6. $X = (-2; -1; 2; 4)$, $Y = (3; 5; 10; 12; 15)$
7. $X = (1; 4; 7; 8)$, $Y = (4; 7; 9; 10; 11)$
8. $X = (-3; -1; 1; 4)$, $Y = (6; 7; 10; 12; 14)$
9. $X = (-1; 1; 3; 6)$, $Y = (3; 5; 8; 10; 12)$
10. $X = (-4; -2; 2; 4)$, $Y = (5; 7; 9; 11; 13)$
11. $X = (1; 3; 5; 7)$, $Y = (2; 4; 10; 12; 14)$
12. $X = (-4; -3; 1; 2)$, $Y = (6; 9; 10; 12; 15)$
13. $X = (-2; 2; 4; 6)$, $Y = (9; 10; 12; 14; 15)$
14. $X = (-5; -3; 1; 3)$, $Y = (5; 8; 11; 14; 15)$
15. $X = (1; 2; 4; 5)$, $Y = (7; 8; 9; 12; 13)$
16. $X = (-3; -1; 3; 5)$, $Y = (3; 6; 8; 10; 12)$
17. $X = (1; 4; 5; 6)$, $Y = (2; 5; 8; 11; 15)$
18. $X = (-4; -1; 1; 2)$, $Y = (1; 4; 7; 12; 15)$
19. $X = (2; 3; 4; 6)$, $Y = (7; 10; 11; 12; 14)$
20. $X = (-1; 1; 4; 5)$, $Y = (6; 7; 10; 14; 16)$
21. $X = (1; 2; 3; 5)$, $Y = (4; 8; 9; 11; 12)$
22. $X = (2; 3; 5; 7)$, $Y = (3; 6; 8; 10; 12)$
23. $X = (2; 3; 5; 8)$, $Y = (7; 9; 10; 11; 13)$
24. $X = (-2; -1; 3; 5)$, $Y = (1; 3; 5; 8; 12)$
25. $X = (-1; 2; 4; 6)$, $Y = (4; 6; 8; 11; 14)$

Задание 10. Двухмерная случайная величина (X,Y) распределена равномерно в треугольнике, ограниченном прямыми $x = 0$, $y = 0$, $ax + by = c$. Найти одномерные плотности распределения вероятностей и условные плотности распределения. Вычислить математические ожидания m_x и m_y, дисперсии σ_x и σ_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}. Коэффициенты a, b, c выбрать по номеру варианта.

1. $a = 1; b = 2; c = 4$ **2.** $a = 2; b = 1; c = 4$ **3.** $a = 1; b = 3; c = 6$ **4.** $a = 3; b = 1; c = 6$ **5.** $a = 1; b = 4; c = 8$
6. $a = 4; b = 1; c = 8$ **7.** $a = 2; b = 3; c = 6$ **8.** $a = 3; b = 2; c = 6$ **9.** $a = 3; b = 4; c = 12$
10. $a = 4; b = 3; c = 12$ **11.** $a = 1; b = 5; c = 5$ **12.** $a = 5; b = 1; c = 2$ **13.** $a = 2; b = 5; c = 10$
14. $a = 5; b = 2; c = 10$ **15.** $a = 1; b = 1; c = 4$ **16.** $a = 1; b = 1; c = 3$ **17.** $a = 1; b = 2; c = 8$
18. $a = 2; b = 1; c = 8$ **19.** $a = 1; b = 2; c = 6$ **20.** $a = 2; b = 1; c = 6$ **21.** $a = 1; b = 1; c = 2$
22. $a = 1; b = 6; c = 6$ **23.** $a = 6; b = 1; c = 6$ **24.** $a = 3; b = 5; c = 15$ **25.** $a = 5; b = 3; c = 15$

Тема 4. Функциональные преобразования случайных величин

Задание 11. Дискретная величина X задана таблично:

x_i	–4	–3	–2	–1	0	1	2	3	4	5
p_i	0,02	0,03	0,05	0,1	0,2	0,3	0,15	0,1	0,05	0,05

Записать в виде таблицы закон распределения заданной функции. Найти математическое ожидание функции. Функцию выбрать по номеру варианта.

1. $Y = 0{,}5X^2$ **2.** $Y = 2X^2 - 40$ **3.** $Y = 2|X| - 10$ **4.** $Y = X^2 - 20$ **5.** $Y = |X| + 10$
6. $Y = 2^{|X|}$ **7.** $Y = (X + 2)^2$ **8.** $Y = 2|X - 3| - 50$ **9.** $Y = (X + 1)^2 - 30$ **10.** $Y = |X - 2| + 10$
11. $Y = 2^{-|X|}$ **12.** $Y = 2(X - 1)^2 - 30$ **13.** $Y = 2|X + 2| - 4$ **14.** $Y = (X + 3)^2 - 50$ **15.** $Y = |X - 3| + 3$
16. $Y = 0{,}5(X + 2)^2$ **17.** $Y = 2^{|X-1|}$ **18.** $Y = 2^{|X+1|}$ **19.** $Y = 20 - X^2$ **20.** $Y = 10 - |X|$
21. $Y = 30 - 2X^2$ **22.** $Y = 4 - 2|X + 2|$ **23.** $Y = 2|X| - 10$ **24.** $Y = 2^{|X|-1}$ **25.** $Y = 2^{|X|-2}$.

Задание 12. Математические ожидания и дисперсии статистически независимых величин X и Y равны m_x, D_x и m_y, D_y. Вычислить математическое ожидание и дисперсию функции $Z = 2XY - 9$. Исходные данные выбрать по номеру варианта.

1. $m_x = 5$, $D_x = 2$; $m_y = 1$, $D_y = 5$ **2.** $m_x = 3$, $D_x = 4$; $m_y = 2$, $D_y = 6$ **3.** $m_x = -2$, $D_x = 3$; $m_y = 7$, $D_y = 8$

4. $m_x = -5$, $D_x = 5$; $m_y = 3$, $D_y = 4$ **5.** $m_x = -2$, $D_x = 9$; $m_y = 1$, $D_y = 4$ **6.** $m_x = 1$, $D_x = 8$; $m_y = -1$, $D_y = 3$

7. $m_x = 9$, $D_x = 5$; $m_y = -1$, $D_y = 6$ **8.** $m_x = 4$, $D_x = 4$; $m_y = 3$, $D_y = 9$ **9.** $m_x = -3$, $D_x = 8$; $m_y = 6$, $D_y = 12$

10. $m_x = 8$, $D_x = 3$; $m_y = 3$, $D_y = 7$ **11.** $m_x = 6$, $D_x = 12$; $m_y = 4$, $D_y = 5$ **12.** $m_x = -6$, $D_x = 10$; $m_y = 2$, $D_y = 4$

13. $m_x = 3$, $D_x = 2$; $m_y = 4$, $D_y = 15$ **14.** $m_x = 1$, $D_x = 4$; $m_y = 3$, $D_y = 7$ **15.** $m_x = -2$, $D_x = 5$; $m_y = 6$, $D_y = 9$

16. $m_x = 4$, $D_x = 6$; $m_y = 5$, $D_y = 4$ **17.** $m_x = 6$, $D_x = 8$; $m_y = 9$, $D_y = 6$ **18.** $m_x = 3$, $D_x = 12$; $m_y = -1$, $D_y = 5$

19. $m_x = 1$, $D_x = 2$; $m_y = 4$, $D_y = 9$ **20.** $m_x = 3$, $D_x = 6$; $m_y = 3$, $D_y = 9$ **21.** $m_x = -2$, $D_x = 4$; $m_y = 7$, $D_y = 6$

22. $m_x = -6$, $D_x = 7$; $m_y = 2$, $D_y = 5$ **23.** $m_x = 2$, $D_x = 8$; $m_y = 1$, $D_y = 6$ **24.** $m_x = 4$, $D_x = 8$; $m_y = -2$, $D_y = 4$

25. $m_x = 7$, $D_x = 2$; $m_y = 5$, $D_y = 6$

Тема 5. Закон больших чисел

Задание 13. Дисперсия случайной величины X равна σ^2. С помощью неравенства Чебышева оценить вероятность того, что случайная величина отклонится от своего математического ожидания не более чем на величину ε. Параметры выбрать по номеру варианта.

1. $\sigma^2 = 1{,}5$; $\varepsilon = 2$	**2.** $\sigma^2 = 1{,}4$; $\varepsilon = 2$	**3.** $\sigma^2 = 1{,}1$; $\varepsilon = 1{,}5$	**4.** $\sigma^2 = 1{,}2$; $\varepsilon = 1{,}8$	**5.** $\sigma^2 = 1$; $\varepsilon = 1{,}8$
6. $\sigma^2 = 1{,}2$; $\varepsilon = 1{,}8$	**7.** $\sigma^2 = 1{,}3$; $\varepsilon = 2{,}2$	**8.** $\sigma^2 = 2{,}5$; $\varepsilon = 3$	**9.** $\sigma^2 = 1{,}8$; $\varepsilon = 2{,}4$	**10.** $\sigma^2 = 1{,}6$; $\varepsilon = 3$
11. $\sigma^2 = 1{,}5$; $\varepsilon = 2{,}2$	**12.** $\sigma^2 = 1{,}4$; $\varepsilon = 1{,}8$	**13.** $\sigma^2 = 2$; $\varepsilon = 3$	**14.** $\sigma^2 = 1{,}7$; $\varepsilon = 2{,}1$	**15.** $\sigma^2 = 3$; $\varepsilon = 2$
16. $\sigma^2 = 2$; $\varepsilon = 2{,}3$	**17.** $\sigma^2 = 1{,}8$; $\varepsilon = 3$	**18.** $\sigma^2 = 1{,}6$; $\varepsilon = 2{,}1$	**19.** $\sigma^2 = 2$; $\varepsilon = 2{,}5$	**20.** $\sigma^2 = 2$; $\varepsilon = 2{,}8$
21. $\sigma^2 = 1{,}8$; $\varepsilon = 2{,}3$	**22.** $\sigma^2 = 2$; $\varepsilon = 2$	**23.** $\sigma^2 = 1{,}3$; $\varepsilon = 2{,}2$	**24.** $\sigma^2 = 1{,}3$; $\varepsilon = 2$	**25.** $\sigma^2 = 2$; $\varepsilon = 2{,}4$

Задание 14. Для случайной величины из задания 13 оценивается математическое ожидание. Сколько нужно сделать измерений, чтобы с вероятностью, не меньшей 0,95, среднее арифметическое этих измерений отклонилось от истинного математического ожидания не более чем на величину ε?

Задание 15. Для оценки процента дефектных деталей обследуются на наличие дкфектов n деталей. С помощью неравенства Чебышева оценить вероятность того, что доля дефектных деталей k/n отклонится от истинной вероятности дефектной детали не более чем на величину ε. Параметры n и ε выбрать по номеру варианта.

1. $n = 64$; $\varepsilon = 0{,}08$	**2.** $n = 49$; $\varepsilon = 0{,}09$	**3.** $n = 64$; $\varepsilon = 0{,}1$	**4.** $n = 36$; $\varepsilon = 0{,}1$	**5.** $n = 25$; $\varepsilon = 0{,}2$
6. $n = 49$; $\varepsilon = 0{,}1$	**7.** $n = 36$; $\varepsilon = 0{,}12$	**8.** $n = 25$; $\varepsilon = 0{,}13$	**9.** $n = 64$; $\varepsilon = 0{,}11$	**10.** $n = 49$; $\varepsilon = 0{,}11$
11. $n = 36$; $\varepsilon = 0{,}13$	**12.** $n = 25$; $\varepsilon = 0{,}14$	**13.** $n = 64$; $\varepsilon = 0{,}12$	**14.** $n = 49$; $\varepsilon = 0{,}12$	**15.** $n = 36$; $\varepsilon = 0{,}14$
16. $n = 25$; $\varepsilon = 0{,}15$	**17.** $n = 64$; $\varepsilon = 0{,}13$	**18.** $n = 49$; $\varepsilon = 0{,}13$	**19.** $n = 36$; $\varepsilon = 0{,}15$	**20.** $n = 64$; $\varepsilon = 0{,}14$
21. $n = 49$; $\varepsilon = 0{,}14$	**22.** $n = 49$; $\varepsilon = 0{,}15$	**23.** $n = 64$; $\varepsilon = 0{,}15$	**24.** $n = 64$; $\varepsilon = 0{,}16$	**25.** $n = 64$; $\varepsilon = 0{,}17$

Тема 6. Стаистическая обработка экспериментальных данных. Оценка параметров

Задание 16. По данной выборке случайной величины X вычислить все основные эмпирические характеристики: математическое ожидание m_x^*, дисперсию D^*, несмещённую дисперсию S^2, среднее квадратическое отклонение σ_x^*, построить доверительный интервал для математического ожидания, построить доверительный интервал для дисперсии (доверительную вероятность положить равной 0,95).

1.	1,6	1,5	2,4	2,6	4,9	3,2	1,0	0,1	0,0	2,8	0,3	2,2	0,8	3,2	8,0
	0,7	4,1	0,2	0,3	0,7	3,3	3,4	4,6	0,6	0,5	4,2	3,7	0,1	0,4	1,2
	4,5	1,6	1,5	9,6	4,0	0,3	0,7	7,3	2,5	2,1	2,7	0,3	0,9	4,9	0,1
	1,2	0,5	0,3	1,4	2,8	0,6	1,4	0,8	1,1	0,9	0,4	1,2	0,2	0,1	0,8
2.	1,4	0,6	3,6	3,6	3,4	3,7	3,7	3,6	5,8	0,6	8,3	0,6	5,6	3,8	3,4
	2,0	3,3	3,6	0,6	7,0	1,2	0,7	2,1	3,0	7,5	1,2	5,1	5,7	4,5	3,0
	1,3	2,1	3,7	6,4	1,0	3,7	3,7	0,9	2,2	2,4	3,4	1,3	5,7	1,4	1,2
	0,6	3,6	3,4	0,7	3,7	1,6	1,1	1,3	2,2	3,7	3,5	2,3	3,2	2,7	1,4
3.	0,1	1,2	0,5	2,4	2,6	4,9	3,2	1,0	0,1	0,0	2,8	0,3	2,2	0,8	3,2
	0,7	1,5	0,2	0,3	0,7	3,3	3,4	4,6	0,6	0,5	4,2	3,7	0,1	0,4	1,2
	4,5	0,6	0,1	1,6	1,5	7,6	4,2	0,3	0,7	7,3	2,5	2,1	2,7	0,3	0,9
	4,9	0,2	1,5	1,8	0,5	2,1	0,9	1,4	0,2	1,1	0,4	5,2	0,5	1,7	1,2
4.	0,0	0,4	1,5	0,7	2,9	0,3	2,1	0,6	0,2	0,3	7,4	0,2	0,1	1,3	1,5
	0,3	1,0	0,1	2,5	1,2	3,5	5,2	1,3	1,0	3,3	2,5	9,6	1,6	0,5	3,1
	0,8	1,9	0,0	0,5	1,5	2,1	3,0	2,3	1,0	2,3	1,5	2,2	1,4	0,3	0,9
	1,2	2,3	0,3	1,1	2,0	0,2	1,3	0,4	0,1	6,2	4,4	1,4	0,9	1,7	0,5
5.	0,2	0,1	1,7	0,8	4,9	0,2	2,5	0,3	2,4	0,2	1,9	0,5	1,6	1,8	0,2
	2,6	1,0	0,8	4,3	1,1	0,9	2,7	0,9	5,8	1,9	0,3	2,6	1,0	0,0	1,2
	1,1	2,6	1,5	2,6	0,4	0,5	0,5	0,2	2,6	1,3	0,4	0,0	2,3	0,3	1,2
	0,2	2,0	1,1	0,8	1,7	3,9	1,8	2,9	0,4	2,3	3,5	0,7	4,1	1,5	0,3
6.	1,8	0,4	1,5	1,7	0,2	2,4	1,0	0,7	4,0	1,1	0,9	2,5	0,8	5,4	1,8
	0,3	2,4	0,9	0,0	1,1	1,0	2,5	1,4	2,5	0,2	0,5	0,4	0,2	2,4	1,2
	0,4	0,0	2,2	0,3	1,1	0,2	1,9	1,0	0,8	1,6	0,8	1,1	1,3	0,9	1,6
	0,3	0,3	0,8	0,1	0,1	3,6	3,0	0,3	0,7	1,3	0,8	1,2	2,6	1,3	1,1
7.	0,1	6,5	0,7	0,2	1,6	2,5	7,0	0,7	4,6	5,2	4,0	2,5	0,8	1,6	3,2
	5,9	0,5	3,2	0,9	0,1	3,1	3,0	2,9	3,2	3,3	3,1	5,3	0,1	7,8	0,2
	5,1	3,3	2,9	1,5	2,8	3,1	3,2	0,4	1,7	1,9	2,9	0,8	5,5	0,9	0,7
	0,1	3,1	2,9	0,2	3,2	1,1	0,6	0,8	1,7	3,0	1,8	2,7	2,2	0,9	5,1
8.	3,3	0,8	2,9	3,2	0,3	4,5	1,8	1,4	7,5	2,0	1,6	4,7	1,5	9,8	3,4
	0,5	4,6	1,7	0,1	2,1	1,9	4,6	2,7	4,6	0,3	0,9	0,8	0,4	4,5	2,3
	0,7	0,0	4,1	0,6	2,0	0,3	3,5	2,0	1,4	3,0	1,4	2,2	2,5	1,8	3,0
	0,5	0,5	1,5	0,3	0,2	6,7	5,6	0,6	1,3	2,5	1,5	2,3	4,8	2,5	2,1
9.	4,1	1,5	3,1	1,8	1,7	1,2	2,9	1,9	3,3	1,6	3,5	2,8	2,3	2,1	2,2
	5,5	3,5	4,6	6,0	2,1	1,4	1,3	1,2	4,7	2,9	3,4	3,5	2,3	5,1	3,2
	1,8	6,1	2,2	5,5	3,4	3,5	2,3	1,2	1,2	1,3	2,5	1,8	7,8	5,6	3,5
	3,0	1,5	1,5	3,4	4,2	5,0	3,5	7,0	6,4	5,0	3,5	2,8	7,3	2,7	3,5

10.	2,9	0,7	2,5	2,8	0,3	4,0	1,6	1,2	6,6	1,8	1,4	4,2	1,4	9,0	3,0
	0,5	4,1	1,5	0,1	1,9	1,7	4,1	2,4	4,1	0,3	0,8	0,7	0,3	4,0	2,0
	0,6	0,0	3,6	0,5	1,8	0,3	3,1	1,7	1,3	2,6	1,3	1,9	2,2	1,6	2,6
	0,5	0,4	1,4	0,2	0,2	6,0	5,0	0,6	1,2	2,2	1,3	2,0	4,3	2,2	1,9
11.	5,1	2,5	4,1	2,8	2,7	2,2	3,9	2,9	4,3	2,6	4,5	3,8	3,3	3,1	3,2
	6,5	4,5	5,6	7,0	3,1	2,4	2,3	2,2	5,7	3,9	4,4	4,5	3,3	6,1	4,2
	2,8	7,1	3,2	6,5	4,4	4,5	3,3	2,2	2,2	2,3	3,5	2,8	8,8	6,6	4,5
	4,0	2,5	2,5	4,4	5,2	6,0	4,5	8,0	7,4	6,0	4,5	3,8	8,3	3,7	4,5
12.	5,5	2,0	4,2	2,4	2,2	1,6	3,9	2,6	4,4	2,1	4,7	3,8	3,1	2,8	2,9
	7,3	4,7	6,2	8,0	2,8	1,9	1,8	1,6	6,2	3,9	4,6	4,7	3,0	6,8	4,3
	2,5	2,1	3,0	7,4	4,5	4,7	3,1	1,6	1,6	1,7	3,3	2,4	9,6	7,5	4,7
	4,0	2,0	1,9	4,6	5,7	4,6	9,3	8,5	6,7	4,7	3,7	9,8	3,6	4,7	2,9
13.	2,5	0,6	2,2	2,5	0,3	3,5	1,4	1,1	5,7	1,5	1,2	3,6	1,2	7,8	2,6
	0,4	3,5	1,3	0,1	1,6	1,5	3,6	2,1	3,6	0,2	0,7	0,6	0,3	3,5	1,8
	0,6	0,0	3,2	0,5	1,5	0,2	2,7	1,5	1,1	2,3	1,1	1,7	1,9	1,3	2,3
	0,4	0,4	1,2	0,2	0,2	5,2	4,3	0,5	1,0	1,9	1,2	1,8	3,7	1,9	1,6
14.	6,1	3,5	5,1	3,8	3,7	3,2	4,9	3,9	5,3	3,6	5,5	4,8	4,3	4,1	4,2
	7,5	5,5	6,6	8,0	4,1	3,4	3,3	3,2	6,7	4,9	5,4	5,5	4,3	7,1	5,2
	3,8	8,1	4,2	7,5	5,4	5,5	4,3	3,2	3,2	3,3	4,5	3,8	9,8	7,6	5,5
	5,0	3,5	3,5	5,4	6,2	7,0	5,5	9,0	8,4	7,0	5,5	4,8	9,3	4,7	5,5
15.	2,1	0,5	1,8	2,0	0,2	2,9	1,2	0,9	4,7	1,3	1,0	3,0	1,0	6,5	2,1
	0,3	2,9	1,1	0,0	1,3	1,2	2,9	1,7	2,9	0,2	0,6	0,5	0,2	2,9	1,5
	0,5	0,0	2,6	0,4	1,3	0,2	2,2	1,2	0,9	1,9	0,9	1,4	1,6	1,1	1,9
	0,3	0,3	1,0	0,2	0,2	4,3	3,6	0,4	0,8	1,6	1,0	1,4	3,1	1,6	1,3
16.	5,9	3,8	5,1	4,0	3,9	3,6	4,9	4,1	5,3	3,9	5,4	4,9	4,5	4,3	4,4
	7,0	5,4	6,3	7,4	4,3	3,7	3,7	3,6	6,3	4,9	5,3	5,4	4,4	6,7	5,2
	4,1	7,5	4,4	7,0	5,3	5,4	4,4	3,6	3,6	3,6	4,6	4,1	8,8	7,1	5,4
	5,0	3,8	3,8	5,3	6,0	5,4	8,2	7,7	6,6	5,4	4,8	8,5	4,7	5,4	5,9
17.	6,3	3,3	5,1	3,7	3,5	3,0	4,9	3,8	5,4	3,4	5,6	4,8	4,2	4,0	4,1
	7,8	5,6	6,8	8,4	4,0	3,2	3,1	2,9	6,9	4,9	5,5	5,6	4,2	7,4	5,3
	3,7	8,5	4,1	7,9	5,4	5,6	4,2	3,0	3,0	3,0	4,4	3,7	9,7	7,9	5,6
	5,0	3,3	3,2	5,5	6,4	7,3	5,5	9,5	8,8	7,3	5,6	4,8	9,9	4,6	5,6
18.	4,1	4,2	3,0	1,9	2,0	2,1	3,2	2,5	8,5	6,3	4,2	3,7	2,2	2,3	4,1
	4,9	5,7	4,2	7,7	7,1	5,7	4,2	3,5	7,8	3,4	4,2	4,8	2,2	3,8	2,5
	2,4	1,9	3,6	2,6	4,0	2,3	4,2	3,5	3,0	2,8	2,9	6,2	4,2	5,3	6,7
	2,8	2,1	2,0	1,9	5,4	3,6	4,1	4,2	3,0	5,8	3,9	2,5	6,8	2,9	6,2
19.	2,4	0,6	2,0	2,3	0,2	3,2	1,3	1,0	5,3	1,4	1,1	3,4	1,1	7,3	2,4
	0,4	3,3	1,2	0,0	1,5	1,4	3,3	1,9	3,3	0,2	0,7	0,6	0,3	3,2	1,6
	0,5	0,0	2,9	0,4	1,4	0,2	2,5	1,4	1,0	2,1	1,0	1,5	1,8	1,3	2,1
	0,4	0,4	1,1	0,2	0,2	4,8	4,0	0,5	0,9	1,8	1,1	1,6	3,4	1,8	1,5
20.	3,8	2,1	4,0	3,3	2,8	2,6	2,7	6,0	4,1	5,1	6,5	2,6	1,9	1,8	1,7
	5,2	3,4	3,9	4,0	2,8	5,6	3,7	2,5	6,6	2,7	6,0	3,9	4,0	2,8	1,7
	1,8	1,9	2,9	2,3	8,3	6,1	4,3	3,5	2,0	2,1	3,9	4,7	5,5	3,9	7,5
	6,9	5,5	4,0	3,3	7,8	3,4	3,8	4,6	1,9	3,6	2,3	2,2	1,7	3,4	2,4

21.	7,2	9,0	4,8	2,9	2,8	2,6	7,2	4,9	5,6	5,7	4,0	5,8	5,3	3,5	3,1
	4,0	8,4	5,5	5,7	4,1	2,6	2,6	2,7	4,3	3,4	9,8	8,5	5,7	5,0	3,0
	2,9	5,6	6,7	5,6	6,5	3,0	5,2	3,4	3,2	2,6	4,9	3,6	5,4	3,1	5,7
	4,8	4,1	3,8	3,9	8,3	5,7	9,6	8,8	7,7	5,7	4,7	9,8	4,6	5,7	3,9
22.	1,6	0,4	1,4	1,6	0,2	2,2	0,9	0,7	3,6	1,0	0,8	2,3	0,7	5,0	1,6
	0,3	2,2	0,8	0,0	1,0	0,9	2,3	1,3	2,3	0,1	0,5	0,4	0,2	2,2	1,1
	0,4	0,0	2,0	0,3	1,0	0,1	1,7	1,0	0,7	1,5	0,7	1,1	1,2	0,9	1,5
	0,3	0,2	0,7	0,1	0,1	3,3	2,7	0,3	0,6	1,2	0,7	1,1	2,4	1,2	1,0
23.	5,8	3,2	4,8	3,5	3,4	2,9	4,6	3,6	5,0	3,3	5,2	4,5	4,0	3,8	3,9
	7,2	5,2	5,1	5,2	4,1	2,9	2,8	3,0	4,2	3,5	9,5	7,3	5,2	4,7	3,2
	3,1	5,1	5,9	6,7	6,3	7,7	3,8	3,1	3,0	2,9	6,4	4,6	5,1	5,2	4,0
	6,8	4,9	3,5	7,8	3,9	7,2	5,2	8,7	8,1	6,7	5,2	4,5	9,0	4,4	5,3
24.	5,3	6,4	3,3	2,7	3,0	2,6	5,3	3,9	4,3	4,4	3,4	5,7	4,2	3,1	6,5
	3,4	6,0	4,9	2,8	4,1	3,0	2,9	2,6	3,9	3,1	4,3	2,9	4,4	3,9	3,5
	3,3	3,4	6,0	4,4	4,3	4,4	3,4	2,7	2,6	2,5	3,6	3,1	7,8	6,1	4,4
	4,0	2,8	3,0	4,3	5,0	4,5	7,2	6,7	5,6	4,4	3,8	7,5	3,7	4,6	4,9
25.	5,8	2,8	4,6	3,2	3,0	2,5	4,4	3,5	4,9	2,9	5,1	4,3	3,7	3,5	3,6
	7,3	5,1	6,3	8,1	3,5	2,7	2,6	2,4	6,4	4,4	5,0	5,1	3,7	6,9	4,8
	3,2	8,0	3,6	7,4	4,9	5,1	3,7	2,5	3,0	2,3	3,9	3,2	9,2	7,4	5,1
	4,5	2,8	2,7	5,0	5,9	6,8	5,0	9,0	8,3	6,8	5,1	4,3	9,4	4,1	5,2

Задание 17. Для оценки вероятности появления дефектов были обследованы детали, выпускаемые некоторой производственной линией. Среди них было обнаружено k дефектных деталей. Построить доверительный интервал для истинной вероятности появления дефектной детали с доверительной вероятностью, равной 0,95. Параметры n и k выбрать по номеру варианта.

1. $n = 64$; $k = 8$	**2.** $n = 49$; $k = 9$	**3.** $n = 64$; $k = 10$	**4.** $n = 100$; $k = 10$	**5.** $n = 81$; $k = 12$
6. $n = 49$; $k = 7$	**7.** $n = 100$; $k = 12$	**8.** $n = 81$; $k = 13$	**9.** $n = 64$; $k = 11$	**10.** $n = 81$; $k = 11$
11. $n = 100$; $k = 13$	**12.** $n = 81$; $k = 14$	**13.** $n = 64$; $k = 12$	**14.** $n = 100$; $k = 11$	**15.** $n = 100$; $k = 14$
16. $n = 81$; $k = 15$	**17.** $n = 81$; $k = 10$	**18.** $n = 100$; $k = 18$	**19.** $n = 100$; $k = 15$	**20.** $n = 64$; $k = 9$
21. $n = 100$; $k = 16$	**22.** $n = 49$; $k = 8$	**23.** $n = 100$; $k = 17$	**24.** $n = 81$; $k = 9$	**25.** $n = 100$; $k = 19$

Тема 7. Проверка статистических гипотез

Задание 18. По представленной в задании 16 выборке построить полигон и гистограмму. Подобрать подходящий теоретический закон распределения вероятностей и проверить гипотезу о соответствии эмпирического закона распределения выбранному теоретическому при уровне значимости $\alpha = 0{,}05$.

Задание 19. Используя данные таблицы задания 16, с помощью критерия квантилей проверить гипотезу о том, что медиана распределения равна эмпирическому математическому ожиданию.

Задание 20. Предположим, что две первые строки таблицы задания 16 являются измерениями случайной величины X, а две последние – измерениями случайной величины Y. Проверить гипотезу о равенстве математических ожиданий величин X и Y.

Тема 8. Дисперсионный анализ

Задание 21. Сделано по 5 измерений случайной величины X на каждом из четырёх уровней фактора A. Полученные результаты представлены матрицей X (строки матрицы соответствуют уровням фактора, столбцы – номеру измерения). Методом дисперсионного анализа проверить гипотезу о том, что фактор A не влияет на математическое ожидание величины X. Уровень значимости α принять равным 0,05. Матрицу выбрать по номеру варианта.

1. $X = \begin{pmatrix} 5 & 3 & 7 & 9 & 3 \\ 4 & 8 & 8 & 3 & 4 \\ 7 & 6 & 2 & 4 & 6 \\ 9 & 5 & 6 & 4 & 6 \end{pmatrix}$ **2.** $X = \begin{pmatrix} 4 & 9 & 6 & 4 & 5 \\ 7 & 3 & 5 & 8 & 4 \\ 6 & 6 & 4 & 7 & 6 \\ 3 & 5 & 9 & 5 & 6 \end{pmatrix}$ **3.** $X = \begin{pmatrix} 2 & 6 & 2 & 4 & 6 \\ 6 & 7 & 3 & 5 & 8 \\ 9 & 5 & 2 & 9 & 6 \\ 4 & 4 & 9 & 8 & 7 \end{pmatrix}$

4. $X = \begin{pmatrix} 3 & 9 & 4 & 2 & 6 \\ 7 & 5 & 5 & 3 & 9 \\ 2 & 6 & 2 & 5 & 5 \\ 5 & 5 & 8 & 4 & 8 \end{pmatrix}$ **5.** $X = \begin{pmatrix} 7 & 5 & 2 & 9 & 4 \\ 5 & 4 & 6 & 6 & 9 \\ 5 & 7 & 6 & 4 & 3 \\ 8 & 5 & 3 & 9 & 5 \end{pmatrix}$ **6.** $X = \begin{pmatrix} 4 & 3 & 4 & 9 & 8 \\ 2 & 4 & 6 & 5 & 3 \\ 8 & 9 & 3 & 6 & 5 \\ 5 & 8 & 4 & 5 & 3 \end{pmatrix}$

7. $X = \begin{pmatrix} 3 & 2 & 2 & 9 & 1 \\ 4 & 5 & 7 & 3 & 2 \\ 5 & 4 & 3 & 1 & 4 \\ 8 & 6 & 7 & 4 & 5 \end{pmatrix}$ **8.** $X = \begin{pmatrix} 3 & 9 & 6 & 4 & 2 \\ 7 & 8 & 5 & 5 & 2 \\ 5 & 4 & 3 & 6 & 5 \\ 7 & 5 & 9 & 3 & 4 \end{pmatrix}$ **9.** $X = \begin{pmatrix} 4 & 3 & 7 & 9 & 4 \\ 3 & 5 & 8 & 3 & 6 \\ 2 & 6 & 2 & 4 & 3 \\ 9 & 5 & 7 & 6 & 5 \end{pmatrix}$

10. $X = \begin{pmatrix} 9 & 3 & 8 & 4 & 7 \\ 4 & 5 & 7 & 4 & 6 \\ 7 & 9 & 2 & 5 & 3 \\ 8 & 4 & 5 & 4 & 6 \end{pmatrix}$ **11.** $X = \begin{pmatrix} 7 & 9 & 3 & 5 & 2 \\ 4 & 7 & 5 & 2 & 4 \\ 3 & 6 & 2 & 5 & 6 \\ 4 & 5 & 9 & 4 & 6 \end{pmatrix}$ **12.** $X = \begin{pmatrix} 2 & 6 & 4 & 2 & 5 \\ 4 & 8 & 6 & 7 & 5 \\ 5 & 6 & 2 & 4 & 6 \\ 3 & 4 & 9 & 8 & 5 \end{pmatrix}$

13. $X = \begin{pmatrix} 9 & 3 & 6 & 9 & 2 \\ 4 & 7 & 5 & 3 & 4 \\ 8 & 6 & 2 & 5 & 6 \\ 9 & 5 & 4 & 4 & 6 \end{pmatrix}$ **14.** $X = \begin{pmatrix} 7 & 9 & 6 & 4 & 5 \\ 7 & 3 & 5 & 8 & 4 \\ 6 & 4 & 9 & 7 & 6 \\ 5 & 5 & 8 & 5 & 4 \end{pmatrix}$ **15.** $X = \begin{pmatrix} 4 & 6 & 2 & 4 & 5 \\ 8 & 7 & 6 & 5 & 8 \\ 9 & 5 & 2 & 9 & 6 \\ 4 & 3 & 9 & 8 & 7 \end{pmatrix}$

16. $X = \begin{pmatrix} 4 & 5 & 4 & 2 & 5 \\ 7 & 5 & 5 & 3 & 8 \\ 2 & 4 & 2 & 5 & 6 \\ 5 & 5 & 8 & 4 & 9 \end{pmatrix}$ **17.** $X = \begin{pmatrix} 8 & 5 & 8 & 9 & 4 \\ 5 & 4 & 6 & 6 & 7 \\ 5 & 6 & 6 & 4 & 8 \\ 7 & 5 & 2 & 9 & 5 \end{pmatrix}$ **18.** $X = \begin{pmatrix} 5 & 6 & 4 & 9 & 7 \\ 3 & 4 & 6 & 5 & 4 \\ 8 & 5 & 3 & 6 & 5 \\ 4 & 7 & 4 & 5 & 6 \end{pmatrix}$

19. $X = \begin{pmatrix} 8 & 6 & 2 & 9 & 5 \\ 4 & 5 & 7 & 3 & 2 \\ 5 & 4 & 3 & 1 & 4 \\ 3 & 2 & 7 & 4 & 1 \end{pmatrix}$ **20.** $X = \begin{pmatrix} 7 & 3 & 5 & 4 & 6 \\ 2 & 8 & 8 & 5 & 3 \\ 5 & 3 & 9 & 6 & 5 \\ 7 & 5 & 6 & 3 & 4 \end{pmatrix}$ **21.** $X = \begin{pmatrix} 6 & 2 & 7 & 9 & 5 \\ 3 & 5 & 8 & 3 & 4 \\ 8 & 4 & 3 & 4 & 3 \\ 4 & 5 & 7 & 6 & 4 \end{pmatrix}$

22. $X = \begin{pmatrix} 8 & 4 & 8 & 3 & 6 \\ 4 & 5 & 7 & 4 & 7 \\ 7 & 9 & 2 & 5 & 8 \\ 9 & 3 & 9 & 4 & 6 \end{pmatrix}$ **23.** $X = \begin{pmatrix} 9 & 4 & 3 & 9 & 8 \\ 4 & 7 & 6 & 2 & 4 \\ 3 & 8 & 2 & 5 & 6 \\ 9 & 5 & 4 & 7 & 9 \end{pmatrix}$ **24.** $X = \begin{pmatrix} 2 & 6 & 8 & 2 & 7 \\ 5 & 8 & 9 & 7 & 5 \\ 6 & 9 & 2 & 4 & 6 \\ 6 & 4 & 9 & 8 & 8 \end{pmatrix}$

25. $X = \begin{pmatrix} 7 & 8 & 4 & 3 & 9 \\ 4 & 8 & 9 & 7 & 5 \\ 5 & 7 & 2 & 4 & 6 \\ 6 & 4 & 9 & 8 & 4 \end{pmatrix}$

Задание 22. Фактор *A* имеет 4 уровня, фактор *B* – 5 уровней. Сделано по одному измерению случайной величины *X* на каждой комбинации уровней факторов. Полученные результаты представлены матрицей *X* (строки матрицы соответствуют уровням фактора *A*, столбцы – уровням фактору *B*). Методом дисперсионного анализа проверить гипотезу о том, что факторы *A* и B не влияют на математическое ожидание величины *X*. Предполагать, что взаимодействия между факторами нет. Уровень значимости α принять равным 0,05. Матрицу выбрать из задания 21 по номеру вырианта.

Задание 23. Проверить гипотезу о влиянии факторов *A*, *B*, и *C* на исследуемую величину *X*, если все факторы имеют по четыре уровня, а измерения сделаны по плану эксперимента «латинский квадрат». Результаты измерений представлены матрицей *X*, исключая пятый столбец (строки матрицы соответствуют уровням фактора *A*, столбцы – уровням фактору *B*). Предполагать, что взаимодействия между факторами нет. Уровень значимости α принять равным 0,05. Матрицу выбрать из задания 21 по номеру вырианта (не забывая исключить пятый столбец).

Тема 9. Регрессионный анализ

Задание 24. При изучении зависимости между величиной *Y* и величиной *X* было получено 15 пар соответствующих значений этих величин. Аппроксимировать статистическую зависимость величины *Y* от *X* линейной функцией $y = ax + b$. Вычислить остаточную дисперсию и оценку коэффициента корреляции. Данные выбрать по номеру варианта.

1. *X*: –1,0 –0,8 –0,6 –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8
Y: –2,1 –2,9 –3,5 –4,1 –4,2 –3,9 –3,7 –3,2 –1,3 0,2 1,5 3,4 5,3 5,7 7,5

2. *X*: –1,0 –0,8 –0,6 –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8
Y: –1,2 –1,8 –2,7 –3,4 –3,6 –2,9 –2,4 –2,1 –0,4 1,1 2,8 4,6 6,5 6,4 8,3

3. *X*: –1,0 –0,8 –0,6 –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8
Y: –0,1 –0,7 –1,4 –2,2 –2,3 –1,8 –1,6 –1,3 1,3 2,4 3,6 5,1 7,4 7,6 9,4

4. *X*: –1,0 –0,8 –0,6 –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8
Y: –0,6 –1,5 –2,0 –2,7 –2,8 –2,4 –2,2 –1,7 0,2 1,7 3,1 4,8 6,8 7,4 9,1

5. *X*: –1,0 –0,8 –0,6 –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8
Y: –0,1 –0,5 –1,1 –1,6 –1,9 –1,3 –1,2 –0,8 1,3 2,6 4,2 5,7 7,7 8,4 9,9

6. *X*: –0,5 –0,3 –0,1 0,1 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3
Y: –0,1 –0,2 –0,8 –1,2 –1,4 –1,5 –1,3 –1,1 –0,7 –0,2 0,4 1,4 2,4 3,6 5,1

7. *X*: –0,5 –0,3 –0,1 0,1 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3
Y: 0,9 0,7 0,2 –0,2 –0,5 –0,6 –0,4 –0,2 0,3 0,7 1,5 2,3 3,5 4,7 6,2

8. *X*: –0,5 –0,3 –0,1 0,1 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3
Y: 1,8 1,7 1,1 0,7 0,5 0,4 0,6 0,9 1,4 1,8 2,4 3,4 4,6 5,8 7,1

9. *X*: –0,5 –0,3 –0,1 0,1 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3
Y: 1,4 1,3 0,7 0,3 0,1 0,1 0,2 0,4 0,8 1,3 1,9 2,9 3,8 5,1 6,6

10. *X*: –0,5 –0,3 –0,1 0,1 0,3 0,5 0,7 0,9 1,1 1,3 1,5 1,7 1,9 2,1 2,3
Y: 2,3 2,2 1,8 1,2 1,1 1,2 1,3 1,5 1,7 2,4 3,0 4,1 4,9 6,3 7,8

11. *X*: –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Y: 4,3 3,5 2,8 2,7 2,5 2,6 2,7 3,1 3,4 4,1 4,9 5,9 7,1 8,3 9,7

12. *X*: –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Y: 5,2 4,6 3,8 3,6 3,4 3,5 3,8 4,0 4,5 5,2 5,8 7,1 8,2 9,3 10,8

13. *X*: –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Y: 4,7 4,1 3,4 3,1 2,8 2,9 3,3 3,5 4,1 4,7 5,4 6,5 7,7 8,8 10,3

14. *X*: –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Y: 6,1 5,5 4,8 4,7 4,5 4,6 4,7 5,1 5,5 6,2 6,7 8,0 9,1 10,1 11,6

15. *X*: –0,4 –0,2 0,0 0,2 0,4 0,6 0,8 1,0 1,2 1,4 1,6 1,8 2,0 2,2 2,4
Y: 5,8 4,9 4,3 4,2 4,1 4,0 4,3 3,6 4,9 5,4 6,3 7,5 8,6 9,2 11,2

16. *X*: 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6 3,9 4,2
Y: 9,1 7,3 5,8 4,4 3,2 2,3 1,4 0,8 0,4 0,2 0,1 0,1 0,4 0,8 1,5

17. *X*: 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6 3,9 4,2
Y: 9,9 8,2 6,9 5,3 4,1 3,1 2,3 1,7 1,3 1,1 1,0 0,9 1,4 1,8 2,6

18. *X*: 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6 3,9 4,2
Y: 9,3 7,4 6,4 4,8 3,6 2,7 1,8 1,2 0,8 0,6 0,5 0,3 0,7 1,3 2,1

19. *X*: 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6 3,9 4,2
Y: 10,2 8,3 7,4 4,7 4,5 3,6 2,9 2,4 1,8 1,5 1,4 1,2 1,8 2,4 3,3

20. *X*: 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6 3,9 4,2
Y: 10,7 8,7 7,9 5,3 5,1 4,1 3,3 2,9 2,2 2,0 1,8 1,7 2,2 3,0 3,9

21. *X*: –0,6 –0,3 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6
Y: 9,4 7,6 6,0 4,6 3,4 2,3 1,4 0,8 0,3 –0,1 –0,3 –0,2 0,1 0,4 1,0

22. *X*: –0,6 –0,3 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6
Y: 9,8 7,9 6,4 5,1 3,9 2,7 1,8 1,2 0,7 0,3 0,1 0,2 0,5 0,8 1,4

23. *X*: –0,6 –0,3 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6
Y: 10,4 8,4 6,9 5,7 4,5 3,2 2,3 1,8 1,2 0,9 0,6 0,8 1,1 1,4 1,9

24. *X*: –0,6 –0,3 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6
Y: 10,8 8,9 7,5 6,1 5,1 3,8 2,8 2,3 1,7 1,4 0,9 1,3 1,7 1,9 2,4

25. *X*: –0,6 –0,3 0,0 0,3 0,6 0,9 1,2 1,5 1,8 2,1 2,4 2,7 3,0 3,3 3,6
Y: 11,3 9,4 8,1 6,6 5,6 4,4 3,3 2,8 2,3 1,9 1,3 1,8 2,2 2,5 3,0

Задание 25. Предполагая, что аппроксимацию задания 24 можно улучшить, аппроксимировать статистическую зависимость величины *Y* от *X* функцией $y = ax^2 + bx + c$ исходя из той же самой таблицы исходных данных. Вычислить остаточную дисперсию и оценку корреляционного отношения. Сравнить полученные результаты с результатами линейного анализа из задания 24.

Задание 26. По представленным данным аппроксимировать статистическую зависимость величины *Y* от *X* и Z функцией $y = ax + bz + c$. Вычислить остаточную дисперсию, найти оценку обобщённого коэффициента корреляции. Значения аргументов *X* и *Z* представлены ниже:

1,2	1,2	1,3	1,4	1,5	1,6	1,7	1,8	2,1	2,2	2,3	2,4	2,6	2,6	2,7
2,1	3,5	2,6	1,3	2,2	1,8	1,5	1,9	1,4	2,5	1,7	2,1	1,6	1,5	2,3

Соответствующие значения функции *Y* выбрать по номеру варианта.

1.	1,1	5,9	3,4	–1,1	2,2	1,2	0,4	1,8	0,5	4,4	2,1	3,6	3,0	1,9	4,7
2.	1,4	4,9	2,8	–0,8	2,3	0,8	0,2	1,8	0,7	4,2	2,2	3,4	2,4	2,4	5,2
3.	0,9	5,6	2,6	–0,5	2,2	1,1	0,5	2,0	0,4	4,8	2,1	3,9	2,3	2,3	4,4
4.	1,0	5,4	3,4	–0,9	2,4	1,2	0,5	1,8	0,6	3,9	2,1	3,3	2,0	2,4	5,1
5.	0,9	5,8	3,2	–0,5	1,8	0,7	0,1	1,6	1,1	4,9	2,2	3,9	2,5	1,9	5,3
6.	5,7	7,5	6,3	5, 8	6,6	6,7	6,7	6,9	7,2	8,0	7,4	8,2	7,9	8,3	9,2
7.	5,9	7,1	6,8	5,5	7,1	6,0	6,6	7,5	7,0	8,5	8,1	8,1	8,0	7,8	8,8
8.	6,3	7,7	6,6	5,9	6,4	6,5	6,4	7,0	6,8	8,7	7,5	8,0	8,3	8,1	9,3
9.	6,2	7,2	6,8	6,0	6,3	6,9	6,5	7,3	7,4	8,4	7,4	8,1	8,7	7,8	9,7
10.	5,7	7,2	7,0	5,9	6,2	6,8	6,8	7,0	7,3	8,9	8,2	8,4	7,9	8,1	9,2
11.	4,9	7,6	5,9	3,5	5,3	4,9	4,6	4,6	4,8	6,9	5,5	5,7	5,8	5,0	7,1
12.	5,1	7,5	6,2	3,6	5,8	5,0	4,7	5,4	4,0	6,2	5,0	6,5	5,0	4,6	7,1
13.	4,8	7,7	6,1	3,6	5,6	4,6	4,4	5,1	4,4	6,5	5,2	5,7	5,3	4,8	6,3
14.	5,2	7,6	6,1	3,9	5,7	4,4	4,2	4,8	4,0	7,0	4,7	6,6	5,5	4,6	6,6
15.	4,5	7,3	5,9	3,1	5,8	5,2	4,4	4,7	4,2	6,9	4,9	6,5	5,1	5,5	6,4
16.	3,8	5,3	4,5	4,2	5,5	5,1	5,3	5,8	6,1	7,9	7,2	8,2	8,1	7,7	8,6
17.	4,0	5,3	4,7	4,4	5,0	5,2	5,6	5,3	6,0	8,0	6,6	8,0	7,9	8,0	8,6
18.	4,5	5,5	4,6	3,8	5,8	4,8	4,8	5,9	6,7	7,7	7,4	7,3	8,3	7,4	8,6
19.	4,7	5,4	4,6	3,7	4,8	4,7	4,9	6,0	6,4	7,4	7,1	8,0	7,8	7,9	9,3
20.	4,4	5,3	5,4	4,4	4,9	5,6	4,9	5,9	5,9	7,9	6,7	7,9	8,2	7,3	8,9
21.	2,7	1,7	3,2	4,4	4,1	4,9	4,9	4,8	6,6	5,4	6,8	6,1	8,0	7,7	7,7
22.	3,1	1,9	2,7	4,2	3,9	4,4	5,0	4,6	6,8	5,9	6,9	7,0	7,3	7,6	7,4
23.	2,8	1,6	2,9	4,6	3,6	4,0	5,1	4,5	6,0	5,7	7,1	6,4	7,4	7,4	7,7
24.	2,6	1,5	3,1	4,2	4,3	4,7	4,8	5,3	6,0	5,2	6,3	6,8	7,9	7,4	7,7
25.	3,4	1,8	2,6	4,1	3,9	4,8	4,7	4,7	6,6	5,9	7,2	6,2	7,8	7,8	7,0

4. Пример выполнения типового расчёта

Тема 1. Случайные события

Задание 1. Судно имеет одно рулевое устройство, четыре котла и две турбины. Событие A означает исправность рулевого устройства, B_k (k = 1, 2, 3, 4) – исправность k-го котла, а C_j (j = 1, 2) – исправность j-ой турбины. Событие D – судно управляемое, что будет в том случае, когда исправны рулевое устройство, хотя бы один котёл и хотя бы одна турбина. Выразить события D и $\overline{D}$ через события A, B_k и C_j.

Решение. Пусть событие E означает исправность хотя бы одного котла, а событие F – исправность хотя бы одной турбины. Тогда, в соответствии с определением операции суммы событий, можно записать:

$$E = B_1 + B_2 + B_3 + B_4 \text{ и } F = C_1 + C_2.$$

С другой стороны, событие D происходит тогда и только тогда, когда происходят и A и E и F одновременно, что соответствует произведению этих событий, а событие $\overline{D}$ выполняется тогда и только тогда, когда не происходит хотя бы одно из этих событий, что соответствует сумме противоположных событий. Таким образом,

$$D = A \cdot E \cdot F \text{ и } \overline{D} = \overline{A} + \overline{E} + \overline{F}$$

Учитывая, что $\overline{E} = \overline{B}_1 \cdot \overline{B}_2 \cdot \overline{B}_3 \cdot \overline{B}_4$ и $\overline{F} = \overline{C}_1 \cdot \overline{C}_2$, получим окончательно:

$$D = A \cdot (B_1 + B_2 + B_3 + B_4) \cdot (C_1 + C_2),\ \overline{D} = \overline{A} + \overline{B}_1 \cdot \overline{B}_2 \cdot \overline{B}_3 \cdot \overline{B}_4 + \overline{C}_1 \cdot \overline{C}_2 .$$

Задание 2. На восьми одинаковых карточках написаны соответственно числа 2, 4, 6, 7, 8, 11, 12 и 13. Наугад берутся две карточки. Определить вероятность того, что образованная из двух полученных чисел дробь сократима.

Решение. Пусть А – событие, состоящее в том, что полученная дробь сократима. Воспользуемся классическим определением вероятности: $P(A) = \frac{N_A}{N}$, где N – общее число исходов опыта, а N_A – число исходов, которые приводят к осуществлению события A. Так как порядок извлечения карточек в данном случае не имеет значения (например, 2, 7 и 7, 2 являются неразличимыми комбинациями по условиям опыта), то подсчёт числа исходов опыта делается по формуле для числа сочетаний. Дробь будет сократимой, если два полученных числа принадлежат множеству из 5 чисел (2, 4, 6, 8, 12), которые имеют хотя бы один общий делитель. Числа 7, 11, 13 являются простыми, и их появление среди выбранных двух не приведёт к осуществлению A. Таким образом,

$$N = C_8^2 = \frac{8!}{6! \cdot 2!} = \frac{7 \cdot 8}{2} = 28;\ N_A = C_5^2 = \frac{5!}{3! \cdot 2!} = 10;\ P(A) = \frac{10}{28} = \frac{5}{14}.$$

Задание 3. На отрезке AB длиной d наудачу поставлены две точки L и M. Найти вероятность того, что точка L будет ближе к точке M, чем к точке A.

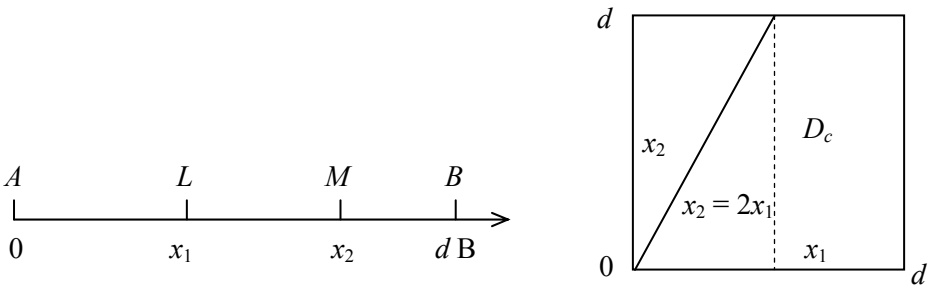

Рис. 1. **Иллюстрация к заданию 3**

Решение. Пусть x_1 и x_2 есть соответственно координаты точек L и M на числовой оси, при условии, что точка A отрезка совмещена с началом отсчёта (рис. 1). Тогда область D возможных исходов опыта будет определяться следующими неравенствами: $0 \le x_1 \le d$ и $0 \le x_2 \le d$. Если по оси абсцисс будем откладывать x_1, а по оси ординат – x_2, то эта область будет представлять собой квадрат со стороной d. Пусть C – событие, состоящее в том, что точка L будет ближе к точке M, чем к точке A. Область D_C значений x_1 и x_2, которые приводят к осуществлению события C, определяется неравенством: $|x_1 - x_2| < x_1$, или $-x_1 < x_1 - x_2 < x_1$. Из этих неравенств находим границы области D_C, полагая $x_1 - x_2 = -x_1$ и $x_1 - x_2 = x_1$. Эти уравнения определяют прямые линии $x_2 = 2x_1$ и $x_2 = 0$, вторая из которых совпадает с осью абсцисс. Таким образом, событие C осуществится, если $0 < x_2 < 2x_1$. Воспользуемся геометрическим определением вероятности события. В данном случае мерой множества исходов опыта является площади соответствующих областей. Очевидно, что площадь области D равна d^2, а площадь области D_C составляет величину $0{,}75d^2$. Следовательно,

$$P(C) = \frac{mes(D_C)}{mes(D)} = \frac{0{,}75d^2}{d^2} = 0{,}75 .$$

Заметим, что возможно $x_1 < x_2$ или $x_2 < x_1$. Это не имеет значения, так как в любом случае расстояние между точками равно модулю разности этих координат.

Задание 4. Двое поочерёдно стреляют по мишени до первого попадания, причём имеется всего 6 пуль. Вероятность попадания в мишень для первого стрелка равна 0,3, для второго – 0,4. Определить вероятность того, что: а) мишень будет поражена первым стрелком; б) мишень будет поражена вторым стрелком; с) мишень не будет поражена.

Решение. Пусть A – событие, состоящее в том, что мишень будет поражена первым стрелком, B – мишень поражена вторым стрелком, C – мишень не будет поражена. Введём дополнительно следующие события: A_k – первый стрелок попадёт при своём k-ом выстреле, B_k – второй стрелок попадёт при k-ом выстреле, $k = 1, 2, 3$ (так как всего можно сделать не более 6 выстрелов). Событие A может произойти в трёх возможных ситуациях: или A_1, или $\overline{A}_1 \cdot \overline{B}_1 \cdot A_2$, или $\overline{A}_1 \cdot \overline{B}_1 \cdot \overline{A}_2 \cdot \overline{B}_2 \cdot A_3$. Это означает, что первый стрелок может поразить мишень или первым выстрелом, или вторым, или третьим. Таким образом,

$$A = A_1 + \overline{A}_1 \cdot \overline{B}_1 \cdot A_2 + \overline{A}_1 \cdot \overline{B}_1 \cdot \overline{A}_2 \cdot \overline{B}_2 \cdot A_3 .$$

Слагаемые в этой сумме являются несовместными событиями. Поэтому

$$P(A) = P(A_1) + P(\overline{A}_1 \cdot \overline{B}_1 \cdot A_2) + P(\overline{A}_1 \cdot \overline{B}_1 \cdot \overline{A}_2 \cdot \overline{B}_2 \cdot A_3) .$$

Вероятность произведения событий в этом выражении равна произведению вероятностей перемножаемых событий, так как результат каждого выстрела не зависит от того, что произошло при других выстрелах. Учитывая, что $P(A_k) = 0{,}3$, $P(B_k) = 0{,}4$, $P(\overline{A}_k) = 0{,}7$, $P(\overline{B}_k) = 0{,}6$, ($k = 1, 2, 3$), получим:

$$P(A) = P(A_1) + P(\overline{A}_1)P(\overline{B}_1)P(A_2) + P(\overline{A}_1)P(\overline{B}_1)P(\overline{A}_2)P(\overline{B}_2)P(A_3) =$$
$$= 0{,}3 \cdot [1 + 0{,}7 \cdot 0{,}6 \cdot [1 + 0{,}7 \cdot 0{,}6]] = 0{,}3 \cdot [1 + 0{,}42 \cdot 1{,}42] \approx 0{,}479.$$

Второй стрелок также может поразить мишень или при первом, или при втором, или при третьем выстреле. Однако он имеет право на очередной выстрел тогда, когда все предыдущие выстрелы были неудачными. Рассуждения, подобные изложенным выше, приводят к следующему результату:

$$P(B) = P(\overline{A}_1)P(B_1) + P(\overline{A}_1)P(\overline{B}_1)P(\overline{A}_2)P(B_2) + (\overline{A}_1)P(\overline{B}_1)P(\overline{A}_2)P(\overline{B}_2)P(\overline{A}_3)P(B_3) =$$
$$= 0{,}7 \cdot 0{,}4[1 + 0{,}7 \cdot 0{,}6[1 + 0{,}7 \cdot 0{,}6]] = 0{,}28[1 + 0{,}42 \cdot 1{,}42] \approx 0{,}447.$$

Наконец, очевидно, что $C = \overline{A}_1 \cdot \overline{B}_1 \cdot \overline{A}_2 \cdot \overline{B}_2 \cdot \overline{A}_3 \cdot \overline{B}_3$, т.е. все 6 выстрелов были неудачными. Тогда

$$P(C) = P(\overline{A}_1)P(\overline{B}_1)P(\overline{A}_2)P(\overline{B}_2)P(\overline{A}_3)P(\overline{B}_3) = [0{,}7 \cdot 0{,}6]^3 \approx 0.074.$$

Заметим, что события *A*, *B*, и *C* являются несовместными и составляют полную группу, т.е. *P*(*A*) + *P*(*B*) + *P*(*C*) = 1. Этим можно было воспользоваться при вычислении *P*(*C*).

Задание 5. Вероятности попадания при каждом выстреле для трёх стрелков равны соответственно $\frac{4}{5}, \frac{3}{4}, \frac{2}{3}$. При одновременном выстреле всех трёх стреков имелось два попадания. Определить вероятность того, что промахнулся третий стрелок.

Решение. Пусть *A* – событие, состоящее в том, что было два попадания, событие *B* – первый стрелок попал в цель, событие *C* – второй стрелок попал в цель. Событие *A* может произойти при двух предположениях: третий стрелок попал в цель – гипотеза H_1 или третий стрелок промахнулся – гипотеза H_2. Априорные вероятности (до опыта) этих гипотез известны: $P(H_1) = \frac{2}{3}$ и $P(H_2) = 1 - \frac{2}{3} = \frac{1}{3}$. Если верна гипотеза H_1, то событие *A* может произойти только тогда, когда один из двух других стрелков промахнётся. Если же верна гипотеза H_2, то событие *A* произойдёт лишь в том случае, когда будет попадание у первого и второго стрелков. Следовательно,

$$P(A \mid H_1) = P(B\overline{C} + \overline{B}C) = P(B\overline{C}) + P(\overline{B}C) = P(B)P(\overline{C}) + P(\overline{B})P(C),$$

$$P(A \mid H_1) = P(BC) = P(B)P(C),$$

так как события $B\overline{C}$ и $\overline{B}C$ несовместны, а события *A* и *B* являются статистически не зависимыми. Воспользуемся исходными данными, получим:

$$P(A \mid H_1) = \frac{4}{5}(1 - \frac{3}{4}) + (1 - \frac{4}{5})\frac{3}{4} = \frac{7}{20},\ P(A \mid H_2) = \frac{4}{5} \cdot \frac{3}{4} = \frac{12}{20}.$$

Вычислим *P*(*A*) по формуле полной вероятности:

$$P(A) = P(A \mid H_1)P(H_1) + P(A \mid H_2)P(H_2) = \frac{7}{20} \cdot \frac{2}{3} + \frac{12}{20} \cdot \frac{1}{3} = \frac{13}{30}.$$

Наконец, вычислим искомую вероятность $P(H_2 \mid A)$:

$$P(H_2 \mid A) = \frac{P(A \mid H_2)P(H_2)}{P(A)} = \frac{12}{20} \cdot \frac{1}{3} \cdot \frac{30}{13} = \frac{6}{13}.$$

Заметим, что эту задачу можно решить другим образом. Так как было только два попадания, то кто-то один промахнулся. Это может быть первый стрелок – гипотеза H_1, второй – гипотеза H_2 или третий – гипотеза H_3. Рекомендуется разобрать эту задачу, исходя из указанных гипотез. Естественно, ответ должен получиться тот же самый.

Задание 6. Проводится 6 испытаний, в каждом из которых событие *A* может произойти с одной и той же вероятностью *p* = 0,4. Найти вероятность того, что в этих испытаниях событие *A* произойдет два раза и вероятность того, что событие *A* произойдет хотя бы два раза. Определить, сколько раз нужно провести испытание, чтобы с вероятностью, не меньшей, чем 0,98, событие произошло хотя бы один раз.

Решение. Пусть *B* – событие, состоящее в том, что *A* произошло два раза, событие *C* – *A* произойдет хотя бы два раза. Очевидно, имеет место схема независимых последовательных испытаний при *n* = 6, *p* = 0,4, *q* = 0,6. Тогда вероятность события *B* определяется непосредственно формулой Бернулли

$$P(B) = P\{k = 2\} = p_n(2) = C_6^2 p^2 q^4 = \frac{6!}{2!\cdot 4!} \cdot 0{,}4^2 \cdot 0{,}6^4 = 15 \cdot 0{,}16 \cdot 0{,}1296 \approx 0{,}311.$$

Вероятность события *C* удобнее, очевидно, вычислять через вероятность противоположного события:

$$P(C) = 1 - P(\overline{C}) = 1 - P\{k < 2\} = 1 - P\{k = 0\} - P\{k = 1\} = 1 - C_6^0 p^0 q^6 - C_6^1 p^1 q^5 =$$

$$= 1 - 0{,}6^6 - 6 \cdot 0{,}4 \cdot 0{,}6^5 \approx 1 - 0{,}047 - 0{,}187 = 0{,}766.$$

Для ответа на последний вопрос задачи достаточно воспользоваться известной формулой:

$$n \geq \frac{|\ln(1-P)|}{|\ln(1-p)|} = \frac{|\ln(1-0{,}98)|}{|\ln(1-0{,}4)|} = \frac{|\ln(0{,}02)|}{|\ln(0{,}6)|} \approx \frac{|-3{,}912|}{|-0{,}511|} \approx 7{,}656.$$

Так как n должно быть целым числом, то следует взять $n = 8$. Таким образом, если испытание провести 8 раз, то с вероятностью $P = 0{,}98$ событие *A* произойдёт хотя бы один раз.

Тема 2. Случайные величины

*Задание 7***.** Задан закон распределения дискретной случайной величины *X*. Найти математическое ожидание, дисперсию, среднее квадратическое отклонение. Построить график функции распределения вероятностей случайной величины.

x	9,2	9,8	10,2	10,8	11,6
p	0,1	0,2	0,4	0,2	0,1

Решение. Найдём сначала начальные моменты ν_k случайной величины *X*:

$$\nu_1 = \sum_{i=1}^{n} x_i \cdot p_i = 9{,}2 \cdot 0{,}1 + 9{,}8 \cdot 0{,}2 + 10{,}2 \cdot 0{,}4 + 10{,}8 \cdot 0{,}2 + 11{,}6 \cdot 0{,}1 = 10{,}28;$$

$$\nu_2 = \sum_{i=1}^{n} x_i^2 \cdot p_i = 9{,}2^2 \cdot 0{,}1 + 9{,}8^2 \cdot 0{,}2 + 10{,}2^2 \cdot 0{,}4 + 10{,}8^2 \cdot 0{,}2 + 11{,}6^2 \cdot 0{,}1 = 106{,}072.$$

Таким образом, математическое ожидание равно $m_x = \nu_1 = 10{,}28$. Дисперсию определим через начальные моменты:

$$D_x = \nu_2 - \nu_1^2 = 106{,}072 - 10.28^2 = 106{,}072 - 105{,}6784 = 0{,}3936.$$

Среднее квадратическое отклонение равно: $\sigma_x = \sqrt{D_x} = \sqrt{0{,}3936} \approx 0{,}627$.

График функции распределения вероятностей дискретной случайной величины *X* имеет горизонтальные участки между точками её возможных значений. Изменение значения функции происходит скачком в точках возможных значений. График функции представлен на рис. 2.

$$F(x) = \begin{cases} 0, \text{ при } x \leq 9{,}2, \\ 0{,}1, \text{ при } 9{,}2 < x \leq 9{,}8, \\ 0{,}3, \text{ при } 9{,}8 < x \leq 10{,}2, \\ 0{,}7, \text{ при } 10{,}2 < x \leq 10{,}8, \\ 0{,}9, \text{ при } 10{,}8 < x \leq 11{,}6, \\ 1{,}0, \text{ при } x \geq 11{,}6. \end{cases}$$

Аналитическое выражение функции распределения вероятностей.

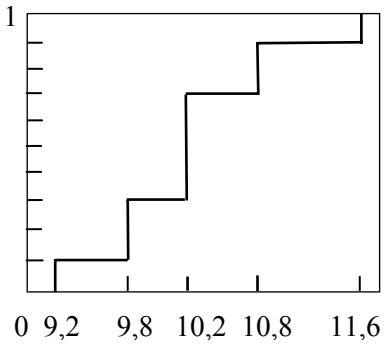

Рис. 2. **График функции распределения вероятностей**

Задание 8. Непрерывная случайная величина X задана плотностью распределения вероятностей. Требуется: А) найти функцию распределения вероятностей; Б) найти математическое ожидание, дисперсию, среднее квадратическое отклонение; скошенность и эксцесс распределения; вероятность того, что случайная величина отклонится от своего математического ожидания не более, чем на 0,5; В) построить графики функций распределения и плотности распределения вероятностей.

$$f(x)=\begin{cases}0 & \text{при } x\le 0,\\ 4x\cdot \mathrm{e}^{-2x} & \text{при } x>0.\end{cases}$$

Решение. Найдём функцию распределения $F(x)$. Если $x \le 0$, то $f(t) = 0$ для $t \in (-\infty; x)$. Следовательно,

$$F(x)=\int\limits_{-\infty}^{x} f(t)dt=\int\limits_{-\infty}^{x} 0\cdot dt=0.$$

Если $x > 0$, то

$$F(x)=\int\limits_{-\infty}^{x} f(t)dt=\int\limits_{-\infty}^{0} 0\cdot dt+\int\limits_{0}^{x} 4t\cdot e^{-2t}dt=\left|z=2t,\ dz=2dt\right|=\int\limits_{0}^{2x} z\cdot e^{-z}dz=$$

$$=\begin{vmatrix} u=z & dv=e^{-z}dz \\ du=dz & v=-e^{-z}\end{vmatrix}=-z\cdot e^{-z}\Big|_0^{2x}+\int\limits_0^{2x} e^{-z}dz=-2x\cdot e^{-2x}-e^{-2z}\Big|_0^{2x}=1-(1+2x)e^{-2x}.$$

Здесь сначала сделана замена переменной, затем применён метод интегрирования по частям. Таким образом,

$$F(x)=\begin{cases}0 & \text{при } x\le 0,\\ 1-(1+2x)\cdot \mathrm{e}^{-2x} & \text{при } x>0.\end{cases}$$

Найдём начальные моменты случайной величины:

$$\nu_k=\int\limits_{-\infty}^{\infty} x^k f(x)dx=\int\limits_{0}^{\infty} x^k\cdot 4x\cdot e^{-2x}dx=\left|z=2x,\ dz=2dx\right|=\frac{1}{2^k}\int\limits_0^{\infty} z^{k+1}\cdot e^{-z}dz=\frac{(k+1)!}{2^k}.$$

Последний интеграл берётся путём последовательного интегрирования по частям $(k + 1)$ раз. Значение этого интеграла можно также найти в справочниках по определённым интегралам. Таким образом,

$$\nu_1=1,\quad \nu_2=\frac{3}{2},\quad \nu_3=3,\quad \nu_4=\frac{15}{2}.$$

Вычислим центральные моменты (заметим, что $\mu_1 = 0$ всегда):

$$\mu_2=\nu_2-\nu_1^2=\frac{3}{2}-1=\frac{1}{2},\quad \mu_3=\nu_3-3\nu_2\nu_1^2+2\nu_1^3=3-3\cdot\frac{3}{2}+2=\frac{1}{2},$$

$$\mu_4=\nu_4-4\nu_3\nu_1+6\nu_2\nu_1^2-3\nu_1^4=\frac{15}{2}-4\cdot 3+6\cdot\frac{3}{2}-3=\frac{3}{2}.$$

Теперь можно определить все числовые характеристики: математическое ожидание $m_x = \nu_1 = 1$; дисперсия $D_x = \mu_2 = \frac{1}{2}$; среднее квадратическое отклонение $\sigma=\sqrt{D_x}=\frac{1}{\sqrt{2}}$; скошенность $Sk_x=\frac{\mu_3}{\sigma^3}=\frac{1}{2}\cdot\left(\frac{1}{\sqrt{2}}\right)^{-3}=\sqrt{2}$; эксцесс $Ex_x=\frac{\mu_4}{\sigma^4}-3=\frac{3}{2}\cdot\left(\frac{1}{\sqrt{2}}\right)^{-4}-3=3.$

В пункте Б) требуется найти вероятность $P\{|X - m_x| \le 0{,}5\}$. Преобразуем неравенство, стоящее под символом вероятности:

$$P\{|X - m_x| \le 0{,}5\} = P\{-0{,}5 \le X - m_x \le 0{,}5\} = P\{m_x - 0{,}5 \le X \le m_x + 0{,}5\}.$$

Подставляя сюда $m_x = 1$ и применяя свойство функции распределения вероятностей, получим:

$$P\{0{,}5 \le X \le 1{,}5\} = F(1{,}5) - F(0{,}5) = 1 - (1 + 2 \cdot 1{,}5) \cdot e^{-2 \cdot 1.5} - [1 - (1 + 2 \cdot 0{,}5) \cdot e^{-2 \cdot 0.5}] =$$
$$= 2 \cdot e^{-1} - 4 \cdot e^{-3} \approx 2 \cdot 0{,}368 - 4 \cdot 0{,}05 = 0{,}536.$$

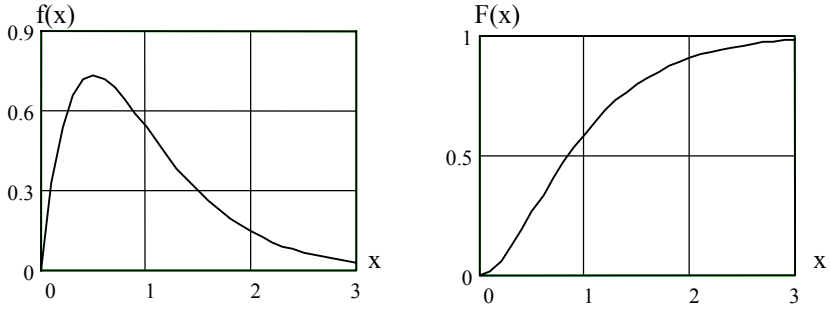

Рис. 3. **Графики плотности и функции распределения вероятностей**

Графики плотности и функции распределения вероятностей можно строить по отдельным точкам, проводя через эти точки плавную кривую. Процесс построения графиков не отличается от обычного процесса, применяемого по отношению к любой функции при её исследовании. Графики представлены на рис. 3. Рассмотренное распределение является *распределением Эрланга* второго порядка.

Тема 3. Многомерные случайные величины

*Задание 9***.** Закон распределения дискретной двухмерной случайной величины (*X*, *Y*) представлен следующей таблицей:

x_i \ y_j	1	3	5	8	9
7	0,04	0,04	0,03	0,03	0,05
9	0,04	0,07	0,06	0,05	0,03
11	0,01	0,08	0,09	0,08	0,05
13	0,03	0,04	0,04	0,06	0,08

Определить одномерные законы распределения случайных величин *X* и *Y*. Найти условные плотности распределения вероятностей величин $P(x_i|y_2)$ и $P(y_k|x_1)$. Вычислить математические ожидания m_x и m_y, дисперсии D_x и D_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}.

Решение. Складывая элементы строк, получим одномерное распределение величины *X*, складывая элементы столбцов, получим одномерное распределение для величины *Y*:

x_i	7	9	11	13
p_{i*}	0,19	0,25	0,31	0,25

y_j	1	3	5	8	9
p_{*k}	0,12	0,23	0,22	0,22	0,21

Поделив элементы второго столбца на их сумму, получим условный закон распределения $P(x_i|y_2)$, а поделив элементы первой строки на их сумму, получим условный закон распределения $P(y_k|x_1)$:

x_i	7	9	11	13
$P(x_i\|y_2)$	4/23	7/23	8/23	4/23

y_j	1	3	5	8	9
$P(y_k\|x_1)$	4/19	4/19	3/19	3/19	5/19

Найдём математические ожидания и дисперсии, исходя из одномерных законов распределения: $m_x = 7\cdot0{,}19 + 9\cdot0{,}25 + 11\cdot0{,}31 + 13\cdot0{,}25 = 10{,}24$; $m_y = 1\cdot0{,}12 + 3\cdot0{,}23 + 5\cdot0{,}22 + 8\cdot0{,}22 + 9\cdot0{,}21 = 5{,}56$; $M\{X^2\} = 49\cdot0{,}19 + 81\cdot0{,}25 + 121\cdot0{,}31 + 169\cdot0{,}25 = 109{,}32$; $D_x = 109{,}32 - (10{,}24)^2 = 4{,}462$; $M\{Y^2\} = 1\cdot0{,}12 + 9\cdot0{,}23 + 25\cdot0{,}22 + 64\cdot0{,}22 + 81\cdot0{,}21 = 38{,}78$; $D_y = 38{,}78 - (5{,}56)^2 = 7{,}866$. Извлекая корни, получим: $\sigma_x = 2{,}112$, $\sigma_y = 2{,}805$.

Корреляционный момент вычисляется по двухмерному закону распределения. Вычислим сначала начальный смешанный момент второго порядка:

$M\{XY\} = 7(1\cdot0{,}04 + 3\cdot0{,}04 + 5\cdot0{,}03 + 8\cdot0{,}03 + 9\cdot0{,}05) + 9(1\cdot0{,}04 + 3\cdot0{,}07 + 5\cdot0{,}06 + 8\cdot0{,}05 + 9\cdot0{,}03) + 11(1\cdot0{,}01 + 3\cdot0{,}08 + 5\cdot0{,}09 + 8\cdot0{,}08 + 9\cdot0{,}05) + 13(1\cdot0{,}03 + 3\cdot0{,}04 + 5\cdot0{,}04 + 8\cdot0{,}06 + 9\cdot0{,}08) = 57{,}82$. Тогда $K_{xy} = M\{XY\} - m_x\, m_y = 57{,}82 - 10{,}24\cdot5{,}56 = 0{,}886$. Таким образом, $r_{xy}^2 = K_{xy}/(\sigma_x\sigma_y) = 0{,}15$. Извлекая корень, получим: $r_{xy} = 0{,}387$.

Задание 10. Двухмерная случайная величина (X, Y) распределена равномерно в треугольнике, ограниченном прямыми $x = 0$, $y = 0$, $4x + 5y = 20$. Найти одномерные плотности распределения вероятностей и условные плотности распределения. Вычислить математические ожидания m_x и m_y, дисперсии D_x и D_y, ковариационный момент K_{xy} и коэффициент корреляции r_{xy}.

Решение. Площадь треугольника равна 10, следовательно $f(x, y) = 1/10$ внутри треугольника и $f(x, y) = 0$ вне треугольника. Найдём одномерные плотности распределения вероятностей:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{(20-4x)/5} \frac{1}{10}dy = \frac{1}{10}\cdot\frac{20-4x}{5} = \frac{20-4x}{50} = \frac{10-2x}{25}.$$

Аналогично,

$$f_Y(x) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^{(20-5y)/4} \frac{1}{10}dy = \frac{1}{10}\cdot\frac{20-5y}{4} = \frac{20-5y}{40} = \frac{4-y}{8}.$$

Найдём начальные моменты:

$$\nu_{kx} = \int_{-\infty}^{\infty} x^k f_X(x)dx = \int_0^5 x^k \frac{10-2x}{25}dx = \frac{1}{25}\cdot\left(\frac{10x^{k+1}}{k+1} - \frac{2x^{k+2}}{k+2}\right)\Bigg|_0^5 = \frac{1}{25}\cdot\left(\frac{10\cdot5^{k+1}}{k+1} - \frac{2\cdot5^{k+2}}{k+2}\right).$$

Аналогично находим: $\nu_{ky} = \frac{1}{8}\cdot\left(\frac{4\cdot4^{k+1}}{k+1} - \frac{4^{k+2}}{k+2}\right)$. Полагая $k = 1$, найдём математические ожидания: $m_x = 5/3$, $m_y = 4/3$. При $k = 2$, получим: $\nu_{2x} = 25/6$, $\nu_{2y} = 8/3$. Следовательно,

$$D_x = 25/6 - (5/3)^2 = 25/18,\ D_y = 8/3 - (4/3)^2 = 8/9.$$

Запишем условные плотности распределения:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{10}\cdot\frac{8}{4-y} = \frac{4}{20-5y};\quad f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{10}\cdot\frac{25}{10-2x} = \frac{5}{20-4x}.$$

Найдём смешанный начальный момент второго порядка:

$$\nu_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \frac{1}{10}\int_0^5 xdx \int_0^{(20-4x)/5} ydy = \frac{1}{10}\int_0^5 \frac{x(20-4x)^2}{50}dx = \frac{10000}{500}\cdot\left(\frac{1}{2}-\frac{2}{3}+\frac{1}{4}\right) = \frac{5}{3}.$$

Таким образом, $K_{xy} = \nu_{11} - m_x\, m_y = 5/3 - (5/3)(4/3) = -5/9$. Тогда

$$r_{xy} = \frac{K_{xy}}{\sigma_x\sigma_y} = -\frac{5}{9\cdot\sqrt{(25/18)\cdot(8/9)}} = -\frac{1}{2}.$$

Тема 4. Функциональные преобразования случайных величин

Задание 11. Дискретная величина X задана таблично:

x_i	–4	–3	–2	–1	0	1	2	3	4	5
p_i	0,02	0,03	0,05	0,1	0,15	0,3	0,15	0,1	0,05	0,05

Записать в виде таблицы закон распределения заданной функции. Найти математическое ожидание функции $Y = 8 - 2|X + 1|$.

Решение. Найдём сначала все возможные значения функции и внесём их в заготовленную таблицу:

y_k	–4	–2	0	2	4	6	8
p_k	0,05	0,05	0,1	0,17	0,33	0,2	0,1

Вероятности подсчитываются следующим образом. Значение 4 функция принимает только при $x = 5$. Тогда вероятность этого значения равна 0,05. Значение 4 функция принимает при $x = -3$ и при $x = 1$. Тогда вероятность этого значения равна 0,03 + 0,3 = 0,33 и т.д.

Математическое ожидание вычисляется как обычно: $m_y = 3{,}36$.

Задание 12. Математические ожидания и дисперсии статистически независимых величин X и Y равны $m_x = 10$, $D_x = 8$ и $m_y = -5$, $D_y = 12$. Вычислить математическое ожидание и дисперсию функции $Z = 2XY - 9$.

Решение. Вычислим математическое ожидание. Так как величины статистически независимы, то математическое ожидание произведения равно произведению математических ожиданий: $m_z = M\{2XY - 9\} = 2\cdot M\{XY\} - 9 = 2\cdot 10\cdot(-5) - 9 = -109$. Для вычисления дисперсии воспользуемся второй теоремой о дисперсиях:

$$D\{Z\} = D\{2XY - 9\} = 4D\{XY\} = 4[D_xD_y + m_x^2D_y + m_y^2D_x] = 4[8\cdot 12 + 100\cdot 12 + 25\cdot 8] = 1496.$$

Тема 5. Закон больших чисел

Задание 13. Дисперсия случайной величины X равна $\sigma^2 = 3{,}5$. С помощью неравенства Чебышева оценить вероятность того, что случайная величина отклонится от своего математического ожидания не более чем на величину $\varepsilon = 2{,}5$.

Решение. Пользуемся неравенством Чебышева:

$$P\{|X - m_x| < \varepsilon\} > 1 - \frac{\sigma^2}{\varepsilon^2} = 1 - \frac{3{,}5}{(2{,}5)^2} = \frac{2{,}75}{6{,}25} = 0{,}44.$$

Задание 14. Для случайной величины из задания 13 оценивается математическое ожидание. Сколько нужно сделать измерений, чтобы с вероятностью не меньшей 0,95,

среднее арифметическое этих измерений отклонилось от истинного математического ожидания не более чем на величину ε?

Решение. Если дисперсия величины X равна σ^2, то дисперсия среднего арифметического n её измерений будет равна σ^2/n. Воспользуемся неравенством Чебышева для среднего арифметического измерений m*: $P\{|m^* - m_x| < \varepsilon\} > 1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \frac{3,5}{n(2,5)^2}$. По условию задачи должно выполняться неравенство $1 - \frac{3,5}{n(2,5)^2} \geq 0,95$. Отсюда находим: $\frac{3,5}{n(2,5)^2} \leq 0,05 \Rightarrow n \geq \frac{3,5}{0,05 \cdot 6,25} = 11,2$. Число измерений должно быть целым, поэтому берём $n = 12$.

Задание 15. Для оценки процента дефектных деталей обследуются на наличие дефектов n деталей. С помощью неравенства Чебышева оценить вероятность того, что доля дефектных деталей k/n отклонится от истинной вероятности дефектной детали не более чем на величину ε. Решить задачу для $n = 81$ и $\varepsilon = 0,1$.

Решение. Запмшем неравенство Чебышева по отношению к доле дефектных деталей $p^* = k/n$: $P\{|p^* - p| < \varepsilon\} > 1 - \frac{pq}{n\varepsilon^2}$. Так как истинная вероятность p появления дефектной детали неизвестна, заменим произведение pq его наибольшим значением (при $0 \leq p \leq 1$), равным 0,25: $P\{|p^* - p| < 0,1\} > 1 - \frac{0,25}{81 \cdot 0,1^2} \approx 0,69$.

Тема 6. Статистическая обработка экспериментальных данных. Оценка параметров

Задание 16. Подобное задание подробно рассмотрено в учебном пособии.

Задание 17. Для оценки вероятности появления дефектных деталей, выпускаемых некоторой производственной линией, было обследовано n деталей. Среди обследованных деталей было обнаружено k дефектных деталей. Построить доверительный интервал для истинной вероятности появления дефектной детали с доверительной вероятностью, равной 0,95. Решить задачу при $n = 121$ и $k = 20$.

Решение. Запишем доверительный интервал для вероятности события: $p \in \left(p^* - \varepsilon_\gamma \sqrt{\frac{pq}{n}}; p^* + \varepsilon_\gamma \sqrt{\frac{pq}{n}} \right)$ с вероятностью γ. В данном случае $p^* = 20/121 \approx$ $\approx 0,165$. По таблицам нормального распределения по заданной доверительной вероятности $\gamma = 0,95$ находим: $\varepsilon_\gamma = 1,96$. Найдём доверительный интервал двумя способами.

Первый способ. Вместо неизвестной вероятности появления дефектной детали подставим её оценку: $p \in \left(0,165 - 1,96\sqrt{\frac{0,165 \cdot (1 - 0,165)}{121}}; 0,165 + 1,96\sqrt{\frac{0,165 \cdot (1 - 0,165)}{121}} \right)$. Вычисляя, получим: $p \in (0,165 - 0,066; 0,165 + 0,066)$. Или $p \in (0,099; 0,231)$ с вероятностью, приближённо равной 0,95.

Второй способ. Вместо неизвестного произведения pq подставим его наибольшее значение (при $0 \leq p \leq 1$), равное 0,25: $p \in \left(0,165 - 1,96\sqrt{\frac{0,25}{121}}; 0,165 + 1,96\sqrt{\frac{0,25}{121}} \right)$. Вычисляя,

получим: $p \in (0{,}165 - 0{,}089;\ 0{,}165 + 0{,}089)$. Или $p \in (0{,}076;\ 0{,}254)$ с вероятностью не меньшей чем 0,95. Второй способ надёжнее, хотя он даёт более широкий интервал.

Тема 7. Проверка статистических гипотез

Задание 18. Подобное задание подробно рассмотрено в учебном пособии.

Задание 19. Используя данные таблицы задания, с помощью критерия квантилей проверить гипотезу о том, что медиана распределения равна эмпирическому математическому ожиданию.

1,4	1,8	2,4	2,6	4,9	2,2	1,3	0,1	0,0	2,8	0,3	2,2	0,8	3,2	8,2
0,9	3,1	0,1	0,3	0,8	3,3	3,4	4,7	0,5	0,5	4,2	3,7	0,1	0,4	1,1
4,8	1,6	1,5	9,6	4,0	0,3	0,7	7,3	2,5	2,1	2,7	0,3	0,9	4,9	0,2
1,7	0,4	0,3	1,3	2,8	0,6	1,4	0,8	1,1	0,9	0,4	1,2	0,2	0,1	0,7

Решение. Сложив все элементы выборки и поделив на число измерений 60, получим оценку математического ожидания m^* = 1,99. Найдём число элементов, которое меньше математического ожидания: k = 37. Для медианы p = 0,5. Следовательно, $t = \dfrac{k - np}{\sqrt{npq}} = \dfrac{37 - 60 \cdot 0{,}5}{\sqrt{60 \cdot 0{,}5 \cdot 0{,}5}} = \dfrac{7}{3{,}873} \approx 1{,}807$. При уровне значимости α = 0,05 по нормальному распределению найдём критическое значение t_α = 1,96. Так как $|t| < t_\alpha$, то гипотезу о равенстве медианы и математического ожидания можно принять.

Задание 20. Подобное задание подробно рассмотрено в учебном пособии.

Тема 8. Дисперсионный анализ

Задание 21. Сделано по 5 измерений случайной величины X на каждом из четырёх уровней фактора A. Полученные результаты представлены матрицей X (строки матрицы соответствуют уровням фактора, столбцы – номеру измерения). Методом дисперсионного анализа проверить гипотезу о том, что фактор A не влияет на математическое ожидание величины X. Уровень значимости α принять равным 0,05.

$$X = \begin{pmatrix} 7 & 6 & 4 & 3 & 9 \\ 5 & 7 & 8 & 8 & 9 \\ 4 & 3 & 4 & 6 & 5 \\ 8 & 9 & 7 & 8 & 6 \end{pmatrix}.$$

Решение. Все вычисления удобно проводить в специальной таблице. Внесём в неё исходные данные, вычислим оценки математическиз ожиданий на каждом уровне фактора, вычислим квадраты измерений и квадраты оценок математических ожиданий. Оценки вычисляются по формуле $m_i^* = \dfrac{1}{5}\sum_{j=1}^{5} x_{ij}$, общая сумма квадратов измерений представлена в последней клетке последней строки. Оценка общего математического ожидания $m^* = \dfrac{1}{20}\sum_{i=1}^{4}\sum_{j=1}^{5} x_{ij} = \dfrac{1}{4}\sum_{i=1}^{4} m_i^* = 6{,}3$, $m^{*2} = 39{,}69$. Сумма квадратов оценок по уровням фактора равна 33,64 + 54,76 + 19,36 + 57,76 = 165,52.

A_i	x_{ij}					m_i*	x_{ij}^2					m_i*2
A_1	7	6	4	3	9	5,8	49	36	16	9	81	33,64
A_2	5	7	8	8	9	7,4	25	49	64	64	81	54,76
A_3	4	3	4	6	5	4,4	16	9	16	36	25	19,36
A_4	8	9	7	8	6	7,6	64	81	49	64	36	57,76
Σ							154	175	145	173	223	870

Таким образом,

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - n \cdot k \cdot m^{*2} = 870 - 5 \cdot 4 \cdot 39{,}69 = 76{,}2,$$

$$Q_1 = n(\sum_{i=1}^{k} m_{i+}^{*2} - k \cdot m^{*2}) = 5 \cdot (165{,}52 - 4 \cdot 39{,}69) = 33{,}8,$$

$$Q_2 = Q - Q_1 = 76{,}2 - 33{,}8 = 42{,}4.$$

Вычислим F:

$$F = \frac{Q_1/(k-1)}{Q_2/[k(n-1)]} = \frac{33{,}8/3}{42{,}4/16} = \frac{11{,}27}{3{,}53} = 4{,}25.$$

По таблице F-распределения при степенях свободы $k_1 = 3$ и $k_2 = 16$ и уровне значимости $\alpha = 0{,}05$ находим критическое значение $F_\alpha = 3{,}24$. Так как вычисленное значение $F = 4{,}25$ больше критического значения F_α, то гипотеза H_0 отвергается. Следует признать влияние фактора A на исследуемую величину при уровне значимости 0,05.

Задание 22. Фактор A имеет 4 уровня, фактор B – 5 уровней. Сделано по одному измерению случайной величины X на каждой комбинации уровней факторов. Полученные результаты представлены матрицей X (строки матрицы соответствуют уровням фактора A, столбцы – уровням фактору B). Методом дисперсионного анализа проверить гипотезу о том, что факторы A и B не влияет на математическое ожидание величины X. Предполагать, что взаимодействия между факторами нет. Уровень значимости принять равным $\alpha = 0{,}05$. Матрицу взять из задания 21.

Решение. Расширим предыдущую таблицу с учётом появления второго фактора.

B_j / A_i	x_{ij}					m_{i+}*	x_{ij}^2					m_{i+}*2
	B_1	B_2	B_3	B_4	B_5		B_1	B_2	B_3	B_4	B_5	
A_1	7	6	4	3	9	5,8	49	36	16	9	81	33,64
A_2	5	7	8	8	9	7,4	25	49	64	64	81	54,76
A_3	4	3	4	6	5	4,4	16	9	16	36	25	19,36
A_4	8	9	7	8	6	7,6	64	81	49	64	36	57,76
m_{+j}*	6	6,25	5,75	6,25	7,25	6,3	Суммы квадратов					
m_{+j}*2	36	39,1	33,1	39,1	52,6	39,69	154	175	145	173	223	870

Все данные и результаты вычислений представлены в таблице. Значения Q и Q_1 вычислены в предыдущей задаче:

$$Q_2 = k(\sum_{i=1}^{n} m_{+j}^{*2} - n \cdot m^{*2}) = 4(36 + 39{,}1 + 33{,}1 + 39{,}1 + 52{,}6 - 5 \cdot 39{,}69) = 5{,}8,$$

$$Q_3 = Q - Q_1 - Q_2 = 76{,}2 - 33{,}8 - 5{,}8 = 36{,}6.$$

Тогда $F_1 = \dfrac{Q_1 \cdot (n-1)}{Q_3} = \dfrac{33{,}8 \cdot 4}{36{,}6} \approx 3{,}69;\quad F_2 = \dfrac{Q_2 \cdot (k-1)}{Q_3} = \dfrac{5{,}8 \cdot 3}{36{,}6} \approx 0{,}48.$

По степеням свободы $k - 1 = 3$ и $(k - 1)(n - 1) = 12$ определяется критическое значение $F_{\alpha 1} = 3{,}49$ при $\alpha = 0{,}05$, а по степеням свободы $n - 1 = 4$ и $(k - 1)(n - 1) = 12$ – критическое значение $F_{\alpha 2} = 3{,}26$. По условию задачи взаимодействия между факторами нет Поэтому можно принять решение по каждому фактору индивидуально. Так как $F_1 > F_{\alpha 1}$, то следует признать влияние фактора А. Так как $F_2 < F_{\alpha 2}$, то следует отвергнуть влияние фактора В на исследуемую велдичину.

Задание 23. Проверить гипотезу о влиянии факторов *A*, *B*, и *C* на исследуемую величину *X*, если все факторы имеют по четыре уровня, а измерения сделаны по плану эксперимента «латинский квадрат». Результаты измерений представлены матрицей *X*, исключая пятый столбец (строки матрицы соответствуют уровням фактора *A*, столбцы – уровням фактору *B*). Предполагать, что взаимодействия между факторами нет. Уровень значимости α принять равным 0,05. Матрицу выбрать из задания 21, исключив пятый столбец.

Решение. Изменим предыдущую таблицу с учётом появления второго фактора и с учётом того, что матрица *X* не имеет пятого столбца. Кроме того, добавим два столбца в таблицу для вычисления оценок математического ожидания на уровнях третьего фактора *C* и квадратов этих оценок.

B_j / A_i	x_{ij}				m_{i+}*	m_t*	x_{ij}^2				m_{i+}*2	m_t*2
	B_1	B_2	B_3	B_4			B_1	B_2	B_3	B_4		
A_1	7	6	4	3	5	7	49	36	16	9	25	49
A_2	5	7	8	8	7	6	25	49	64	64	49	36
A_3	4	3	4	6	4,25	5,75	16	9	16	36	18,06	33,06
A_4	8	9	7	8	8	5,5	64	81	49	64	64	30,25
m_{+j}*	6	6,25	5,75	6,25		6,06	Суммы квадратов					
m_{+j}*2	36	39,1	33,1	39,1		36,72	154	175	145	173		647

Вычисление оценок на уровнях факторов А и В проводится так же, как в двухфакторном анализе. Вычисление оценок на уровнях фактора С проводится по ключевой матрице СК: $CK = \begin{pmatrix} c_1 & c_2 & c_3 & c_4 \\ c_2 & c_3 & c_4 & c_1 \\ c_3 & c_4 & c_1 & c_2 \\ c_4 & c_1 & c_2 & c_3 \end{pmatrix}$. Таким образом,

$$m_1^* = \frac{1}{4}\sum_{t=1} x_{ij} = \frac{1}{4}(x_{11} + x_{42} + x_{33} + x_{24}) = \frac{1}{4}(7 + 9 + 4 + 8) = 7,$$

$$m_2^* = \frac{1}{4}\sum_{t=2} x_{ij} = \frac{1}{4}(x_{21} + x_{12} + x_{43} + x_{34}) = \frac{1}{4}(5 + 6 + 7 + 6) = 6,$$

и т.д. Общая сумма (647) квадратов измерений представлена в последней клетке последней строки. Оценка общего математического ожидания

$$m^* = \frac{1}{16}\sum_{i=1}^{4}\sum_{j=1}^{4} x_{ij} = \frac{1}{4}\sum_{t=1}^{4} m_t^* = 6{,}06,\quad m^{*2} = 36{,}72.$$

Сумма квадратов оценок по уровням фактора A равна 25 + 49 + 18,06 + 64 = 156,06. Сумма квадратов оценок по уровням фактора B равна 36 + 39,1 + 33,1 + 39,1 = 147,3. Сумма квадратов оценок по уровням фактора C равна 49 + 36 + 33,06 + 30,25 = 148,31. Тогда

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n}[x_{ij} - m^*]^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} x_{ij}^2 - 4 \cdot 4 \cdot m^{*2} = 647 - 16 \cdot 36{,}72 = 59{,}48,$$

$$Q_1 = n \cdot \sum_{i=1}^{n}[m_{i+}^* - m^*]^2 = n(\sum_{i=1}^{n} m_{i+}^{*2} - nm^{*2}) = 4(156{,}06 - 4 \cdot 36{,}72) = 36{,}72,$$

$$Q_2 = n \cdot \sum_{j=1}^{n}[m_{+j}^* - m^*]^2 = n(\sum_{j=1}^{n} m_{+j}^{*2} - nm^{*2}) = 4(147{,}3 - 4 \cdot 36{,}72) = 1{,}68,$$

$$Q_3 = n \cdot \sum_{t=1}^{n}[m_t^* - m^*]^2 = n(\sum_{j=1}^{n} m_t^{*2} - nm^{*2}) = 4(148{,}31 - 4 \cdot 36{,}72) = 5{,}72,$$

$$Q_4 = Q - Q_1 - Q_2 - Q_3 = 59{,}48 - 36{,}72 - 1{,}68 - 5{,}72 = 15{,}36.$$

Вычислим отношения F_1, F_1, F_1:

$$F_1 = \frac{Q_1 \cdot (n-2)}{Q_4} = \frac{36{,}72 \cdot 2}{15{,}36} = 4{,}78,$$

$$F_2 = \frac{Q_2 \cdot (n-2)}{Q_4} = \frac{1{,}68 \cdot 2}{15{,}36} = 0{,}22,$$

$$F_3 = \frac{Q_3 \cdot (n-2)}{Q_4}. = \frac{5{,}72 \cdot 2}{15{,}36} = 0{,}74.$$

Критическое значение F_α находится по таблицам F–распределения для $(n - 1) = 3$ и $(n - 1)(n - 2) = 6$ степеней свободы при заданном уровне значимости $\alpha = 0{,}05$. Из таблицы находим: $F_\alpha = 4{,}76$. Таким образом, $F_1 > F_\alpha$, $F_2 < F_\alpha$, $F_3 < F_\alpha$, Учитывая, что взаимодействия между факторами нет (по условию задачи), делаем вывод, что фактор A влияет на исследуемую величину, а факторы B и C не влияют на эту величину (при данном уровне значимости).

Тема 9. Регрессионный анализ

Задание 24. Подобное задание подробно рассмотрено в учебном пособии.

Задание 25. Подобное задание подробно рассмотрено в учебном пособии.

Задание 26. По данным, представленным ниже, аппроксимировать статистическую зависимость величины Y от X и Z функцией $y = ax + bz + c$. Вычислить остаточную дисперсию, найти оценку обобщённого коэффициента корреляции.

X: 1,0 1,8 1,5 0,8 1,8 0,8 1,3 0,8 0,7 0,6 1,7 1,9 0,9 0,7 1,5
Z: 1,9 0,2 1,8 0,9 0,7 0,9 1,9 1,8 2,2 –0,2 –0,4 –0,4 1,0 2,0 1,9
Y: 5,0 6,6 6,0 4,8 6,6 4,6 5,6 5,2 4,4 4,2 6,4 6,8 4,8 4,4 6,0.

Все вычисления будем проводить в таблице. Будем решать задачу не через центральные моменты, как в тексте пособия, а через начальные. В таблице получаем все коэффициенты системы, определяющей коэффициенты аппроксимирующей функции:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i z_i + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i z_i + b\sum_{i=1}^{n} z_i^2 + c\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} z_i + cn = \sum_{i=1}^{n} y_i \end{cases}.$$

Поделив на n, полностью переходим к начальным моментам:

$$\begin{cases} a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i z_i + c\cdot\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i z_i + b\cdot\frac{1}{n}\sum_{i=1}^{n} z_i^2 + c\cdot\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} z_i y_i \\ a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + b\cdot\frac{1}{n}\sum_{i=1}^{n} z_i + c = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}.$$

В таблице представлены все необходимые данные.

x_i	z_i	y_i	x_i^2	z_i^2	y_i^2	$x_i y_i$	$x_i z_i$	$z_i y_i$
1	1,9	5	1	3,61	25	5	1,9	9,5
1,8	0,2	6,6	3,24	0,04	43,56	11,88	0,36	1,32
1,5	1,8	6	2,25	3,24	36	9	2,7	10,8
0,8	0,9	4,8	0,64	0,81	23,04	3,84	0,72	4,32
1,8	0,7	6,6	3,24	0,49	43,56	11,88	1,26	4,62
0,8	0,9	4,6	0,64	0,81	21,16	3,68	0,72	4,14
1,3	1,9	5,6	1,69	3,61	31,36	7,28	2,47	10,64
0,8	1,8	5,2	0,64	3,24	27,04	4,16	1,44	9,36
0,7	2,2	4,4	0,49	4,84	19,36	3,08	1,54	9,68
0,6	–0,2	4,2	0,36	0,04	17,64	2,52	–0,12	–0,84
1,7	–0,4	6,4	2,89	0,16	40,96	10,88	–0,68	–2,56
1,9	–0,4	6,8	3,61	0,16	46,24	12,92	–0,76	–2,72
0,9	1	4,8	0,81	1	23,04	4,32	0,9	4,8
0,7	2	4,4	0,49	4	19,36	3,08	1,4	8,8
1,5	1,9	6	2,25	3,61	36	9	2,85	11,4
17,8	16,2	81,4	24,24	29,66	453,32	102,52	16,7	83,26
1,187	1,08	5,427	1,616	1,977	30,221	6,835	1,113	5,551

В предпоследней строке записаны суммы элементов столбцов, а в последней строке – средние арифметические этих элементов (эмпирические начальные моменты). Запишем систему уравнений, определяющих коэффициенты аппроксимирующей функции:

$$\begin{cases} 1{,}977a + 1{,}113b + 1{,}187c = 6{,}835 \\ 1{,}113a + 1{,}977b + 1{,}08c = 5{,}551 \\ 1{,}187a + 1{,}08b + c = 5{,}427 \end{cases}.$$

Решая систему, получим: $a^* = 0{,}62$, $b^* = -0{,}25$, $c^* = 4{,}97$. Таким образом, аппроксимирующая функция имеет вид: $y = 0{,}62x - 0{,}25z - + 4{,}97$. Вычислим дисперсию D_y^*: $D_y^* = 30{,}221 - (5{,}427)^2 = 0{,}769$. Тогда $S^2 = 0{,}769 \cdot 15/14 = 0{,}824$. В дополнительной таблице вычис-

лим остаточную дисперсию (через y_i^* обозначены спрогнозированные значения функции, вычисленные по аппроксимирующей функции).

x_i	z_i	y_i	y_i^*	$y_i^* - y_i$	$(y_i^* - y_i)^2$
1	1,9	5	5,115	0,115	0,013225
1,8	0,2	6,6	6,036	–0,564	0,318096
1,5	1,8	6	5,45	–0,55	0,3025
0,8	0,9	4,8	5,241	0,441	0,194481
1,8	0,7	6,6	5,911	–0,689	0,474721
0,8	0,9	4,6	5,241	0,641	0,410881
1,3	1,9	5,6	5,301	–0,299	0,089401
0,8	1,8	5,2	5,016	–0,184	0,033856
0,7	2,2	4,4	4,854	0,454	0,206116
0,6	–0,2	4,2	5,392	1,192	1,420864
1,7	–0,4	6,4	6,124	–0,276	0,076176
1,9	–0,4	6,8	6,248	–0,552	0,304704
0,9	1	4,8	5,278	0,478	0,228484
0,7	2	4,4	4,904	0,504	0,254016
1,5	1,9	6	5,425	–0,575	0,330625
17,8	16,2	81,4	81,536	0,136	4,658146

Таким образом, $D_{\text{ост}} = 4{,}658/(n - 3) = 4{,}658/12 = 0{,}388$. Следовательно, $R^2 = 1 - 0{,}388/0{,}824 = 0{,}529$, тогда $|R| = 0{,}727$. Результат можно интерпретировать следующим образом. Пользуясь аппроксимирующей функцией, мы снимаем 52,9% начальной неопределённости относительно значений величины Y. В то же время 47,1% неопределённости сохраняется. Эта неопределённость связана с другими, неучтёнными в данной задаче, факторами. Заметим, что все вычисления удобно проводить в программной среде Microsoft Excel.

Список используемой литературы

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983. – 471 с. ББК 22.172, А11.
2. Вентцель Е.С. Теория вероятностей: Учеб. для вузов. – 7-е изд. стер. – М.: Высш. шк., 2001. – 575 с.: ил.
3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1979.
4. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматгиз, 1961.
5. Горяинов В.Т., Журавлев А.Г., Тихонов В.И. Примеры и задачи по статистической радиотехнике / Под общ. ред. В.И. Тихонова. – М.: Советское радио, 1970.
6. Гурский Е.И. Теория вероятностей с элементами математической статистики: Учеб. пособие для втузов. – М.: Высшая школа, 1971.
7. Емельянов Г.В., Скитович В.П. Задачник по теории вероятностей и математической статистике, Изд. Ленинградского университета, 1967.
8. Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика: Учебник / Под ред. В.А. Колемаева. – М.: ИНФРА-М, 2001. – 302 с. – (Серия «Высшее образование»).
9. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. – М.: ЮНИТИ–ДАНА, 2001. – 543 с.
10. Лисьев В.П. Теория вероятностей и математическая статистика: Учебное пособие / ВКГТУ. – Усть-Каменогорск, 2001.
11. Лозинский С.Н. Сборник задач по теории вероятностей и математической статистике. – М.: Статистика, 1967.
12. Румшинский Л.З. Элементы теории вероятностей. – М.: Наука, 1970.
13. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1969.
14. Феллер В. Введение в теорию вероятностей и её приложения. – М.: Мир, 1964.
15. Чистяков В.П. Курс теории вероятностей. – М.: Наука, 1978.