В. Е. Гмурман

Теория вероятностей и математическая статистика

Издание четвертое, дополненное

Допущено
Министерством высшего
и среднего специального
образования СССР
в качестве учебного пособия
для инженерно-экономических
институтов и факультетов

Издательство «Высшая школа»
Москва — 1972

517.8
Г 55
УДК 519(0.75)

Гмурман В. Е.

Г 55 Теория вероятностей и математическая статистика. Изд. 4-е, доп. Учеб. пособие для вузов. М., «Высш. школа», 1972.

368 с. с илл.

Книга содержит весь материал новой программы по теории вероятностей и математической статистике. Добавлены главы: показательное распределение, проверка статистических гипотез, однофакторный дисперсионный анализ. Большое внимание уделено статистическим методам обработки экспериментальных данных; приведены удобные расчетные таблицы. В конце каждой главы помещены задачи с ответами.

Книга предназначается для студентов инженерно-экономических институтов и факультетов, а также будет полезной инженерам и экономистам, применяющим вероятностные и статистические методы при решении практических задач.

2—2—3 / 47—72 517.8

ПРЕДИСЛОВИЕ

Настоящее издание приведено в соответствие с новой программой. Добавлены три главы: показательное распределение, статистическая проверка статистических гипотез, однофакторный дисперсионный анализ. Включен ряд новых вопросов: поток случайных событий, распределения, связанные с нормальным, математическое ожидание функции и др. Внесены некоторые изменения и отдельные дополнения. Критерий Пирсона изложен заново и перенесен из гл. XVI в гл. XIX. Изменено название книги.

Выражаю благодарность Р. С. Гутеру за помощь и полезные советы

Автор

Гмурман Владимир Ефимович

ТЕОРИЯ ВЕРОЯТНОСТЕЙ
И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Редактор Ж. И. Яковлева
Переплет художника А. И. Демко
Художественный редактор В. И. Пономаренко
Технический редактор З. А. Муслимова
Корректор В. И. Мишанина

Сдано в набор 18/I 1972 г. Подписано к печати 3/V 1972 г. Формат $84\times108^1/_{32}$. Объем 11,5 печ. л. Усл. п. л. 19,32. Уч.-изд. л. 16,46. Изд. № ФМ-102. Тираж 150 000 экз. Зак. 43. Цена 56 коп.

План выпуска литературы издательства «Высшая школа» (вузы и техникумы) на 1972 г. Позиция № 47.

Москва, К-51, Неглинная ул., д. 29/14.
Издательство «Высшая школа»

Ярославский полиграфкомбинат Главполиграфпрома Комитета по печати при Совете Министров СССР
Ярославль, ул. Свободы, 97.

ВВЕДЕНИЕ

Предмет теории вероятностей. Наблюдаемые нами события (явления) можно подразделить на следующие три вида: достоверные, невозможные и случайные.

Достоверным называют событие, которое обязательно произойдет, если будет осуществлена определенная совокупность условий S.

Например, если в сосуде содержится вода при нормальном атмосферном давлении и температуре 20°, то событие «вода в сосуде находится в жидком состоянии» есть достоверное. В этом примере заданные атмосферное давление и температура воды составляют совокупность условий S.

Невозможным называют событие, которое заведомо не произойдет, если будет осуществлена совокупность условий S.

Например, событие «вода в сосуде находится в твердом состоянии» заведомо не произойдет, если будет осуществлена совокупность условий предыдущего примера.

Случайным называют событие, которое при осуществлении совокупности условий S может либо произойти, либо не произойти.

Например, если брошена монета, то она может упасть так, что сверху будет либо герб, либо надпись. Поэтому событие «при бросании монеты выпал герб» — случайное.

Каждое случайное событие, в частности — выпадение герба, есть следствие действия очень многих случайных причин (в нашем примере: сила, с которой брошена монета,

форма монеты и многие другие). Невозможно учесть влияние на результат всех этих причин, поскольку число их очень велико и законы их действия неизвестны. Поэтому теория вероятностей не ставит перед собой задачу предсказать, произойдет единичное событие или нет,— она просто не в силах это сделать.

По-иному обстоит дело, если рассматриваются случайные события, которые могут многократно наблюдаться при осуществлении одних и тех же условий S, т. е. если речь идет о массовых однородных случайных событиях. Оказывается, что достаточно большое число однородных случайных событий, независимо от их конкретной природы, подчиняется определенным закономерностям, а именно — вероятностным закономерностям. Установлением этих закономерностей и занимается теория вероятностей.

Итак, *предметом теории вероятностей является изучение вероятностных закономерностей массовых однородных случайных событий.*

Знание закономерностей, которым подчиняются массовые случайные события, позволяет предвидеть, как эти события будут протекать. Например, хотя, как было уже сказано, нельзя наперед определить результат одного бросания монеты, но можно предсказать, причем с небольшой погрешностью, число появлений герба, если монета будет брошена достаточно большое число раз. При этом предполагается, конечно, что монета бросается в одних и тех же условиях.

Методы теории вероятностей широко применяются в различных отраслях естествознания и техники: в теории надежности, теории массового обслуживания, в теоретической физике, геодезии, астрономии, теории стрельбы, теории ошибок наблюдений, теории автоматического управления, общей теории связи и во многих других теоретических и прикладных науках. Теория вероятностей служит также для обоснования математической и прикладной статистики, которая, в свою очередь, используется при планировании и организации производства, при анализе технологических процессов, предупредительном и приемочном контроле качества продукции и для многих других целей.

В последние годы методы теории вероятностей все шире и шире проникают в различные области науки и техники, способствуя их прогрессу.

Краткая историческая справка. Первые работы, в которых зарождались основные понятия теории вероятностей, представляли собой попытки создания теории азартных игр (Кардано, Гюйгенс, Паскаль, Ферма и др. в XVI—XVII вв.).

Следующий этап развития теории вероятностей связан с именем Якова Бернулли (1654—1705). Доказанная им теорема, получившая впоследствии название «Закона больших чисел», была первым теоретическим обоснованием накопленных ранее фактов.

Дальнейшими успехами теория вероятностей обязана Муавру, Лапласу, Гауссу, Пуассону и др.

Новый, наиболее плодотворный, период связан с именами П. Л. Чебышева (1821—1894) и его учеников А. А. Маркова (1856—1922) и А. М. Ляпунова (1857—1918). В этот период теория вероятностей становится стройной математической наукой. Ее последующее развитие обязано, в первую очередь, русским и советским математикам (С. Н. Бернштейн, В. И. Романовский, А. Н. Колмогоров, А. Я. Хинчин, Б. В. Гнеденко, Н. В. Смирнов и др.). В настоящее время ведущая роль в создании новых ветвей теории вероятностей также принадлежит советским математикам.

Часть первая

СЛУЧАЙНЫЕ СОБЫТИЯ

Глава первая

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

§ 1. Испытания и события

Выше мы назвали событие случайным, если при осуществлении определенной совокупности условий S оно может либо произойти, либо не произойти. В дальнейшем вместо того, чтобы говорить «совокупность условий S осуществлена», мы будем говорить кратко: «произведено испытание». Таким образом, мы будем рассматривать событие как результат испытания.

Пример 1. Стрелок стреляет по мишени, разделенной на четыре области. Выстрел — это испытание. Попадание в определенную область мишени — событие.

Пример 2. В урне имеются цветные шары. Из урны наудачу берут один шар. Извлечение шара из урны есть испытание. Появление шара определенного цвета — событие.

§ 2. Виды случайных событий

События называют *несовместными*, если появление одного из них исключает появление других событий в одном и том же испытании.

Пример 1. Из ящика с деталями наудачу извлечена деталь. Появление стандартной детали исключает появление нестандартной детали. События «появилась стандартная деталь» и «появилась нестандартная деталь» — несовместные

Пример 2. Брошена монета. Появление герба исключает появление надписи. События «появился герб» и «появилась надпись» — несовместные.

События называют *единственно возможными*, если появление в результате испытания одного и только одного из них является достоверным событием.

Очевидно, единственно возможные события попарно несовместны.

Пример 3. Приобретены два билета денежно-вещевой лотереи. Обязательно произойдет одно и только одно из следующих событий: «выигрыш выпал на первый билет и не выпал на второй», «выигрыш не выпал на первый билет и выпал на второй», «выигрыш выпал на оба билета», «на оба билета выигрыш не выпал». Эти события единственно возможные.

Пример 4. Стрелок произвел выстрел по цели. Обязательно произойдет одно из следующих двух событий: попадание или промах. Эти события единственно возможные.

События называют *равновозможными*, если есть основания считать, что ни одно из этих событий не является более возможным, чем другие.

Пример 5. Появление герба и появление надписи при бросании монеты есть события равновозможные. Действительно, предполагается, что монета изготовлена из однородного материала, имеет правильную цилиндрическую форму и наличие чеканки не оказывает влияния на выпадение той или иной стороны монеты.

Пример 6. Появление того или иного числа очков на брошенной игральной кости есть события равновозможные. Действительно, предполагается, что игральная кость изготовлена из однородного материала, имеет форму правильного многогранника и наличие очков не оказывает влияния на выпадение той или иной грани.

§ 3. Классическое определение вероятности

Вероятность является одним из основных понятий теории вероятностей. Существует несколько определений этого понятия. Здесь будет дано определение, которое называют классическим. Далее (§ 6) мы укажем слабые стороны этого определения и приведем другое (статистическое) определение вероятности, позволяющее преодолеть недостатки классического определения.

Рассмотрим пример. Пусть в урне содержится 6 одинаковых, тщательно перемешанных шаров, причем 2 из них — красные, 3 — синие и 1 — белый. Очевидно, возможность вынуть наудачу из урны цветной шар (т. е. красный или синий) больше, чем возможность извлечь белый шар. Можно ли охарактеризовать эту возможность числом? Оказывается, можно. Это число и называют вероятностью события. Таким образом, вероятность есть число, характеризующее возможность появления события.

Поставим своей задачей дать количественную оценку возможности того, что взятый наудачу шар будет цветным. Появление цветного шара будем рассматривать в качестве события A. Каждый из возможных результатов испытания (испытание состоит в извлечении шара из урны), т. е. каждое событие, которое может наступить в испытании, назовем элементарным исходом. Элементарные исходы обозначим через E_1, E_2, E_3 и т. д. В нашем примере возможны следующие 6 элементарных исходов: E_1— появился белый шар; E_2, E_3— появился красный шар; E_4, E_5, E_6— появился синий шар.

Легко видеть, что эти исходы единственно возможны (обязательно появится один шар) и равновозможны (шар вынимают наудачу, шары одинаковы и тщательно перемешаны).

Те элементарные исходы, при которых интересующее нас событие наступает, назовем благоприятствующими этому событию. В нашем примере благоприятствуют событию A (появлению цветного шара) следующие 5 исходов: E_2, E_3, E_4, E_5, E_6.

Отношение числа благоприятствующих событию A элементарных исходов к их общему числу называют вероятностью события A и обозначают $P(A)$. В рассматриваемом примере всего элементарных исходов — 6, из них 5 благоприятствуют событию A. Следовательно, вероятность того, что взятый шар окажется цветным, равна $P(A)=\frac{5}{6}$. Найденное число (вероятность) и дает ту количественную оценку возможности появления цветного шара, которую мы поставили своей задачей найти.

Дадим теперь определение вероятности.

Вероятностью события A называют отношение числа благоприятствующих этому событию исходов к общему числу всех единственно возможных и равновозможных элементарных исходов испытания

Таким образом, вероятность события A определяется формулой

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

где m — число элементарных исходов, благоприятствующих событию A; n — число всех возможных элементарных исходов испытания. Здесь предполагается, что элементарные исходы единственно возможны и равновозможны.

Из определения вероятности вытекают следующие ее свойства:

1. *Вероятность достоверного события равна единице.*

Действительно, если событие достоверно, то каждый элементарный исход испытания благоприятствует событию. В этом случае $m=n$ и, следовательно,

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

2. *Вероятность невозможного события равна нулю.*

Действительно, если событие невозможно, то ни один из элементарных исходов испытания не благоприятствует событию. В этом случае $m=0$ и, следовательно,

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

3. *Вероятность случайного события есть положительное число, заключенное между нулем и единицей.*

Действительно, случайному событию благоприятствует лишь часть из общего числа элементарных исходов испытания. В этом случае $0<m<n$, а, значит, $0<\frac{m}{n}<1$ и, следовательно,

$$0<P(A)<1.$$

Итак, вероятность любого события удовлетворяет неравенствам

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

Далее будут указаны теоремы, которые значительно упрощают решение многих задач. Пока же приведем примеры, при решении которых используется лишь определение вероятности.

4. Примеры непосредственного вычисления вероятностей

Пример 1. Набирая номер телефона, абонент забыл одну цифру и набрал ее наудачу. Найти вероятность того, что набрана нужная цифра.

Решение. Обозначим через A событие — набрана нужная цифра.

Абонент мог набрать любую из 10 цифр, поэтому общее число возможных элементарных исходов равно 10. Эти исходы единственно возможны (одна из цифр набрана обязательно) и равновозможны (цифра набрана наудачу). Благоприятствует событию A лишь один исход (нужная цифра лишь одна).

Искомая вероятность равна отношению числа исходов, благоприятствующих событию, к числу всех элементарных исходов:

$$P(A)=\frac{1}{10}.$$

Пример 2. Набирая номер телефона, абонент забыл последние две цифры и, помня лишь, что эти цифры различны, набрал их наудачу. Найти вероятность того, что набраны нужные цифры.

Решение. Обозначим через B событие — набраны две нужные цифры.

Всего можно набрать столько пар различных цифр, сколько может быть составлено размещений из десяти цифр по две, т. е. $A^2_{10}=10\cdot 9=90$. Таким образом, общее число возможных элементарных исходов равно 90. Эти исходы единственно возможны и равновозможны. Благоприятствует событию B лишь один исход.

Искомая вероятность равна отношению числа исходов, благоприятствующих событию, к числу всех элементарных исходов:

$$P(B)=\frac{1}{90}.$$

Пример 3. Указать ошибку «решения» задачи: «Брошены две игральные кости. Найти вероятность того, что сумма выпавших очков равна 4 (событие A)».

Решение. Всего возможны 2 исхода испытания: сумма выпавших очков равна 4, сумма выпавших очков не равна 4 Поскольку событию A благоприятствует один исход, а общее число исходов равно двум,— искомая вероятность $P(A) = \frac{1}{2}$.

Ошибка этого решения состоит в том, что рассматриваемые исходы не являются равновозможными.

Правильное решение. Общее число равновозможных исходов испытания равно $6 \cdot 6=36$ (каждое число выпавших очков на одной кости может сочетаться со всеми числами очков другой кости). Среди этих исходов благоприятствуют событию A только 3 исхода: (1; 3), (3; 1), (2; 2) (в скобках указаны числа выпавших очков). Следовательно, искомая вероятность

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

Пример 4. В партии из 10 деталей имеется 7 стандартных. Найти вероятность того, что среди шести взятых наудачу деталей, ровно 4 стандартных.

Решение. Общее число возможных элементарных исходов испытания равно числу способов, которыми можно извлечь 6 деталей из 10, т. е. числу сочетаний из 10 элементов по 6 (C^6_{10}).

Подсчитаем число исходов, благоприятствующих интересующему нас событию A — среди шести взятых деталей ровно 4 стандартных: 4 стандартные детали можно взять из 7 стандартных деталей C_7^4 способами; при этом остальные $6-4=2$ детали должны быть нестандартными; взять же 2 нестандартные детали из $10-7=3$ нестандартных деталей можно C_3^2 способами. Следовательно, число благоприятствующих исходов равно $C_7^4 \cdot C_3^2$.

Искомая вероятность равна отношению числа исходов, благоприятствующих событию, к числу всех элементарных исходов:

$$P(A) = \frac{C_7^4 \cdot C_3^2}{C_{10}^6} = \frac{1}{2}.$$

§ 5. Относительная частота. Устойчивость относительной частоты

Относительная частота, наряду с вероятностью, принадлежит к основным понятиям теории вероятностей.

Относительной частотой события называют отношение числа испытаний, в которых событие появилось, к общему числу фактически произведенных испытаний.

Таким образом, относительная частота события A определяется формулой

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

где m — число появлений события, n — общее число испытаний.

Сопоставляя определения вероятности и относительной частоты, заключаем: определение вероятности не требует, чтобы испытания производились в действительности; определение же относительной частоты предполагает, что испытания были произведены фактически. Другими словами, вероятность вычисляют до опыта, а относительную частоту — после опыта.

Пример 1. Отдел технического контроля обнаружил 3 нестандартных детали в партии из 80 случайно отобранных деталей. Относительная частота появления нестандартных деталей

$$W(A) = \frac{3}{80}$$

Пример 2. По цели произвели 24 выстрела, причем было зарегистрировано 19 попаданий. Относительная частота поражения цели

$$W(A) = \frac{19}{24}.$$

Длительные наблюдения показали, что если в одинаковых условиях производятся опыты, в каждом из которых число испытаний достаточно велико, то относительная частота обнаруживает свойство устойчивости. Это свойство состоит в том, что в различных опытах относительная частота изменяется мало (тем меньше, чем больше произведено испытаний), колеблясь около некоторого постоянного числа. Оказалось, что это постоянное число есть вероятность появления события.

Таким образом, если опытным путем установлена относительная частота, то полученное число можно принять за приближенное значение вероятности.

Подробнее и точнее связь между относительной частотой и вероятностью будет изложена далее. Теперь же проиллюстрируем свойство устойчивости на примерах.

Пример 3. По данным шведской статистики относительная частота рождения девочек за 1935 г. по месяцам характеризуется следующими числами (числа расположены в порядке следования месяцев, начиная с января): 0,486; 0,489; 0,490; 0,471; 0,478; 0,482; 0,462; 0,484; 0,485; 0,491; 0,482; 0,473.

Относительная частота колеблется около числа 0,482, которое можно принять за приближенное значение вероятности рождения девочек

Заметим, что статистические данные различных стран дают примерно то же значение относительной частоты.

Пример 4. Многократно проводились опыты бросания монеты, в которых подсчитывали число появления герба. Результаты нескольких опытов приведены в таблице 1.

Таблица 1

Число бросаний	Число появлений герба	Относительная частота
4040	2048	0,5069
12 000	6019	0,5016
24 000	12 012	0,5005

Здесь относительные частоты незначительно отклоняются от числа 0,5, причем тем меньше, чем больше число испытаний. Например, при 4040 испытаниях отклонение равно 0,0069, а при 24 000 испытаний — лишь 0,0005. Приняв во внимание, что вероятность появления герба при бросании монеты равна 0,5, мы вновь убеждаемся, что относительная частота колеблется около вероятности.

§ 6. Ограниченность классического определения вероятности. Статистическая вероятность

«Классическое» определение вероятности предполагает, что число элементарных исходов испытания — конечно. На практике же весьма часто встречаются испытания, число возможных исходов которых — бесконечно. В таких случаях классическое определение неприменимо. Уже это обстоятельство указывает на ограниченность классического определения. Правда, указанный недостаток может быть преодолен путем соответствующего обобщения определения вероятности.

Наиболее слабая сторона классического определения состоит в том, что очень часто невозможно представить результат испытания в виде совокупности элементарных событий. Еще труднее указать основания, позволяющие считать элементарные события равновозможными. Обычно о равновозможности элементарных исходов испытания заключают из соображений симметрии. Так обстоит дело, например, при бросании игральной кости, когда предполагают, что кость имеет форму правильного многогранника (куба). Однако задачи, в которых можно исходить из соображений симметрии, на практике встречаются весьма редко.

По этой причине наряду с классическим определением пользуются также статистическим определением вероятности, принимая за вероятность события относительную частоту или число, близкое к ней. Например, если в результате достаточно большого числа испытаний оказалось, что относительная частота весьма близка к числу 0,4, то это число можно принять за статистическую вероятность события.

Задачи

1. В ящике имеется 50 одинаковых деталей, из них 5 окрашенных. Наудачу вынимают одну деталь. Найти вероятность того, что извлеченная деталь окажется окрашенной.

Отв. $p = 0{,}1$.

2. Брошена игральная кость. Найти вероятность того, что выпадет четное число очков.

Отв. $p = 0{,}5$.

3. Участники жеребьевки тянут из ящика жетоны с номерами от 1 до 100. Найти вероятность того, что номер первого, наудачу извлеченного жетона, не содержит цифры 5.

Отв. $p = 0{,}81$.

4. В мешочке имеется 5 одинаковых кубиков. На всех гранях каждого кубика написана одна из следующих букв: о, п, р, с, т. Найти вероятность того, что на вынутых по одному и расположенных «в одну линию» кубиках можно будет прочесть слово «спорт».

Отв. $p = \frac{1}{120}$.

5. На каждой из шести одинаковых карточек напечатана одна из следующих букв: а, т, м, р, с, о. Карточки тщательно перемешаны. Найти вероятность того, что на четырех, вынутых по одной и расположенных «в одну линию» карточках, можно будет прочесть слово «трос»

Отв. $p = \dfrac{1}{A_6^4} = \dfrac{1}{360}$

6. Куб, все грани которого окрашены, распилен на тысячу кубиков одинакового размера, которые затем тщательно перемешаны. Найти вероятность того, что наудачу извлеченный кубик будет иметь окрашенных граней: а) одну; б) две; в) три.

Отв. а) 0,384; б) 0,096; в) 0,008.

7. Из тщательно перемешанного полного набора 28 костей домино наудачу извлечена кость. Найти вероятность того, что вторую наудачу извлеченную кость можно приставить к первой, если первая кость: а) оказалась дублем; б) не есть дубль.

Отв. а) $\dfrac{2}{9}$; б) $\dfrac{4}{9}$

8. В замке на общей оси пять дисков, каждый из которых разделен на шесть секторов с различными написанными на них буквами. Замок открывается только в том случае, если каждый диск занимает одно определенное положение относительно корпуса замка. Найти вероятность того, что при произвольной установке дисков замок можно будет открыть

Отв. $p = \dfrac{1}{6^5}$

9. Восемь различных книг расставляются наудачу на одной полке. Найти вероятность того, что две определенные книги окажутся поставленными рядом

Отв. $p = \dfrac{7 \cdot 2! \cdot 6!}{8!} = \dfrac{1}{4}$

10. Библиотечка состоит из десяти различных книг, причем пять книг стоят по 4 рубля каждая, три книги — по одному рублю и две книги — по 3 рубля. Найти вероятность того, что взятые наудачу две книги стоят 5 рублей.

Отв $p = \dfrac{C_5^1 \cdot C_3^1 + C_2^1 \cdot C_3^2}{C_{10}^2} = \dfrac{7}{9}$

11. В партии из 100 деталей отдел технического контроля обнаружил 5 нестандартных деталей. Чему равна относительная частота появления нестандартных деталей?

Отв $w = 0{,}05$.

12. При стрельбе из винтовки относительная частота попадания в цель оказалась равной 0,85 Найти число попаданий, если всего было произведено 120 выстрелов.

Отв 102 попадания

Глава вторая

ТЕОРЕМА СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ

§ 1. Теорема сложения вероятностей несовместных событий

Суммой $A+B$ двух событий A и B называют событие, состоящее в появлении события A или события B, или обоих этих событий.

Например, если из орудия произведены два выстрела и A — попадание при первом выстреле, B — попадание при втором выстреле, то $A+B$ — попадание при первом выстреле, или при втором, или в обоих выстрелах.

В частности, если два события A и B — несовместные, то $A+B$ — событие, состоящее в появлении одного из этих событий, безразлично какого.

Суммой нескольких событий называют событие, которое состоит в появлении хотя бы одного из этих событий.

Например, событие $A+B+C$ состоит в появлении одного из следующих событий: A, B, C, A и B, A и C, B и C, A и B и C.

Пусть события A и B — несовместные, причем вероятности этих событий даны. Как найти вероятность того, что наступит либо событие A, либо событие B? Ответ на этот вопрос дает теорема сложения.

Теорема. *Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий:*

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

Доказательство. Введем обозначения:

n — общее число возможных элементарных исходов испытания;

m_1 — число исходов, благоприятствующих событию A;

m_2 — число исходов, благоприятствующих событию B.

Число элементарных исходов, благоприятствующих наступлению либо события A, либо события B, равно m_1+m_2. Следовательно,

$$P(A+B) = \frac{m_1+m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n}$$

Приняв во внимание, что $\dfrac{m_1}{n} = P(A)$ и $\dfrac{m_2}{n} = P(B)$, окончательно получим

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

Следствие. *Вероятность появления одного из нескольких попарно несовместных событий, безразлично какого, равна сумме вероятностей этих событий:*

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

Доказательство. Рассмотрим три события A, B и C. Так как рассматриваемые события попарно несовместны, то появление одного из трех событий A, B и C, равносильно наступлению одного из двух событий $A+B$ и C, поэтому, в силу указанной теоремы,

$$P(A+B+C)=P[(A+B)+C]=P(A+B)+P(C)=$$
$$=P(A)+P(B)+P(C).$$

Для произвольного числа попарно несовместных событий доказательство проводится методом математической индукции.

Пример 1. В урне 30 шаров: 10 красных, 5 синих и 15 белых. Найти вероятность появления цветного шара.

Решение. Появление цветного шара означает появление либо красного, либо синего шара.

Вероятность появления красного шара (событие A)

$$P(A)=\frac{10}{30}=\frac{1}{3}.$$

Вероятность появления синего шара (событие B)

$$P(B)=\frac{5}{30}=\frac{1}{6}.$$

События A и B несовместны (появление шара одного цвета исключает появление шара другого цвета), поэтому теорема сложения применима.

Искомая вероятность

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

Пример 2. Стрелок стреляет по мишени, разделенной на 3 области. Вероятность попадания в первую область равна 0,45, во вторую — 0,35. Найти вероятность того, что стрелок при одном выстреле попадет либо в первую, либо во вторую область.

Решение. События A — «стрелок попал в первую область» и B — «стрелок попал во вторую область» — несовместны (попадание в одну область исключает попадание в другую), поэтому теорема сложения применима.

Искомая вероятность

$$P(A+B)=P(A)+P(B)=0,45+0,35=0,80.$$

§ 2. Полная группа событий

Полной группой называют совокупность единственно возможных событий испытания.

Пример 1. Стрелок производит по мишени 2 выстрела. События A_1 (одно попадание), A_2 (2 попадания) и A_3 (промах) образуют полную группу.

Теорема. *Сумма вероятностей событий $A_1, A_2, \ldots, A_n$, образующих полную группу, равна единице:*

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

Доказательство. Так как появление одного из событий полной группы достоверно, а вероятность достоверного события равна единице, то

$$P(A_1+A_2+\ldots+A_n)=1. \qquad (*)$$

Любые два события полной группы несовместны, поэтому можно применить теорему сложения:

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

Сравнивая (*) и (**), получим

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

Пример 2. Консультационный пункт института получает пакеты с контрольными работами из городов A, B и C. Вероятность получения пакета из города A равна 0,7, из города B — 0,2. Найти вероятность того, что очередной пакет будет получен из города C.

Решение. События «пакет получен из города A», «пакет получен из города B» и «пакет получен из города C» образуют полную группу, поэтому сумма вероятностей этих событий равна единице:

$$0,7+0,2+p=1.$$

Отсюда искомая вероятность

$$p=1-0,9=0,1.$$

§ 3. Противоположные события

Противоположными называют два единственно возможных события, образующих полную группу. Если одно из двух противоположных событий обозначено через A, то другое принято обозначать $\overline{A}$.

Пример 1. Попадание и промах при выстреле по цели — противоположные события. Если A — попадание, то $\overline{A}$ — промах.

Пример 2. Из ящика наудачу взята деталь. События «появилась стандартная деталь» и «появилась нестандартная деталь» — противоположные.

Теорема. *Сумма вероятностей противоположных событий равна единице:*

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

Доказательство. Противоположные события образуют полную группу, а сумма вероятностей событий, образующих полную группу, равна единице (§ 2).

Замечание 1. Если вероятность одного из двух противоположных событий обозначена через p, то вероятность другого события обозначают через q. Таким образом, в силу предыдущей теоремы

$$p+q=1$$

Пример 3. Вероятность того, что день будет дождливым $p=0,7$. Найти вероятность того, что день будет ясным.

Решение. События «день дождливый» и «день ясный» — противоположные, поэтому искомая вероятность

$$q=1-p=1-0,7=0,3.$$

Замечание 2. При решении задач на отыскание вероятности события A часто выгодно сначала вычислить вероятность события $\overline{A}$, а затем найти искомую вероятность по формуле:

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

Пример 4. В ящике имеется n деталей, из которых m стандартных. Найти вероятность того, что среди k наудачу извлеченных деталей есть хотя бы одна стандартная.

Решение. События «среди извлеченных деталей есть хотя бы одна стандартная» и «среди извлеченных деталей нет ни одной стандартной» — противоположные. Обозначим первое событие через A, а второе через $\overline{A}$.

Очевидно

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

Найдем $P(\overline{A})$. Общее число способов, которыми можно извлечь k деталей из n деталей, равно C_n^k. Число нестандартных деталей равно $n-m$; из этого числа деталей можно C_{n-m}^k способами извлечь k нестандартных деталей. Поэтому вероятность того, что среди извлеченных k деталей нет ни одной стандартной, равна $P(\overline{A})=\frac{C_{n-m}^k}{C_n^k}$.

Искомая вероятность

$$P(A)=1-P(\overline{A})=1-\frac{C_{n-m}^k}{C_n^k}.$$

§ 4. Принцип практической невозможности маловероятных событий

При решении многих практических задач приходится иметь дело с событиями, вероятность которых весьма мала, т. е. близка к нулю. Можно ли считать, что маловероятное событие A в единичном испытании не произойдет? Такого заключения сделать нельзя, так как не исключено, хотя и мало вероятно, что событие A наступит.

Казалось бы, появление или непоявление маловероятного события в единичном испытании предсказать невозможно. Однако длительный опыт показывает, что маловероятное событие в единичном испытании в подавляющем большинстве случаев не наступает. На основании этого факта принимают следующий «принцип практической невозможности маловероятных событий»: *если случайное событие имеет очень малую вероятность, то практически можно считать, что в единичном испытании это событие не наступит.*

Естественно возникает вопрос: насколько малой должна быть вероятность события, чтобы можно было считать невозможным его появление в одном испытании? На этот вопрос нельзя ответить однозначно. Для задач, различных по существу, ответы будут разными. Например, если вероятность того, что парашют при прыжке не раскроется, равна 0,01, то было бы недопустимым применять такие парашюты. Если же вероятность того, что поезд дальнего следования прибудет с опозданием, равна 0,01, то можно практически быть уверенным, что поезд прибудет вовремя.

Достаточно малую вероятность, при которой (в данной определенной задаче) событие можно считать практически невозможным, называют *уровнем значимости*. На практике обычно принимают уровни значимости, заключенные между 0,01 и 0,05. Уровень значимости, равный 0,01, называют однопроцентным; уровень значимости, равный 0,02, называют двухпроцентным и т. д.

Подчеркнем, что рассмотренный здесь принцип позволяет делать предсказания не только о событиях, имеющих малую вероятность, но и о событиях, вероятность которых близка к единице. Действительно, если событие A имеет вероятность близкую к нулю, то вероятность противоположного события $\bar{A}$ близка к единице. С другой стороны, непоявление события A означает наступление противоположного события $\bar{A}$. Таким образом, из принципа невозможности маловероятных событий вытекает следующее важное для приложений следствие: *если случайное событие имеет вероятность очень близкую к единице, то практически можно считать, что в единичном испытании это событие наступит*. Разумеется, и здесь ответ на вопрос о том, какую вероятность считать близкой к единице, зависит от существа задачи.

Задачи

1. В денежно-вещевой лотерее на каждые 10 000 билетов разыгрывается 150 вещевых и 50 денежных выигрышей. Чему равна вероятность выигрыша, безразлично денежного или вещевого, для владельца одного лотерейного билета?

Отв. $p = 0{,}02$.

2. Вероятность того, что стрелок при одном выстреле выбьет 10 очков, равна 0,1; вероятность выбить 9 очков равна 0,3; вероятность выбить 8 или меньше очков равна 0,6. Найти вероятность того, что при одном выстреле стрелок выбьет не менее 9 очков.

Отв. $p = 0{,}4$.

3. В партии из 10 деталей 8 стандартных. Найти вероятность того, что среди наудачу извлеченных 2 деталей есть хотя бы одна стандартная.

Отв. $p = \frac{44}{45}$.

4. В ящике 10 деталей, среди которых 2 нестандартных. Найти вероятность того, что в наудачу отобранных 6 деталях окажется не более одной нестандартной детали.

Отв. $p = \frac{2}{3}$.

Указание. Если A — нет ни одной нестандартной детали, B — есть одна нестандартная деталь, то

$$P(A+B) = P(A) + P(B) = \frac{C_8^6}{C_{10}^6} + \frac{C_2^1 \cdot C_8^5}{C_{10}^6}.$$

5. События A, B, C и D образуют полную систему. Вероятности событий таковы: $P(A) = 0{,}1$; $P(B) = 0{,}4$; $P(C) = 0{,}3$. Чему равна вероятность события D?

Отв. $P(D) = 0{,}2$.

6. По статистическим данным ремонтной мастерской в среднем на 20 остановок токарного станка приходится: 10 — для смены резца; 3 — из-за неисправности привода; 2 — из-за несвоевременной подачи заготовок. Остальные остановки происходят по другим причинам. Найти вероятность остановки станка по другим причинам.

Отв. $p = 0{,}25$.

Глава третья

ТЕОРЕМА УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ

§ 1. Независимые и зависимые события

Два события называют *независимыми*, если вероятность одного из них не зависит от появления или непоявления другого.

Пример 1. Монета брошена 2 раза. Вероятность появления герба в первом испытании (событие A) не зависит от появления или непоявления герба во втором испытании (событие B). В свою очередь, вероятность выпадения герба во втором испытании не зависит от результата первого испытания. Таким образом, события A и B — независимые.

Пример 2. В урне 5 белых и 3 черных шара. Из нее наудачу берут один шар. Очевидно, вероятность появления белого шара (событие A) равна $\frac{5}{8}$. Взятый шар возвращают в урну и испытание повторяют. Вероятность появления белого шара при втором испытании (событие B), по-прежнему, равна $\frac{5}{8}$ и не зависит от результата первого испытания. В свою очередь, вероятность извлечения белого шара при первом испытании не зависит от исхода второго испытания. Таким образом, события A и B — независимые.

Несколько событий называют *попарно независимыми*, если каждые два из них независимы.

Пример 3. Монета брошена 3 раза. Пусть A, B, C — события, состоящие в появлении герба соответственно в первом, втором и третьем испытаниях. Ясно, что каждые два из рассматриваемых событий (т. е. A и B, A и C, B и C)— независимы. Таким образом, события A, B и C — попарно независимые.

Два события называют *зависимыми*, если вероятность появления одного из них зависит от наступления или ненаступления другого события.

Пример 4. В ящике 100 деталей: 80 стандартных и 20 нестандартных. Наудачу берут одну деталь, не возвращая ее в ящик. Если появилась стандартная деталь (событие A), то вероятность извлечения стандартной детали при втором испытании (событие B) $P(B)=\frac{79}{99}$; если же в первом испытании вынута нестандартная деталь, то вероятность $P(B)=\frac{80}{99}$.

Таким образом, вероятность появления события B зависит от наступления или ненаступления события A События A и B — зависимые.

§ 2. Теорема умножения вероятностей независимых событий

Произведением двух событий A и B называют событие AB, состоящее в совместном появлении (совмещении) этих событий.

Например, если в ящике содержатся детали, изготовленные заводами № 1 и № 2, и A — появление стандартной детали, B — деталь изготовлена заводом № 1, то AB — появление стандартной детали завода № 1.

Произведением нескольких событий называют событие, состоящее в совместном появлении всех этих событий. Например, событие ABC состоит в совмещении событий A, B и C.

Пусть события A и B независимые, причем вероятности этих событий известны. Как найти вероятность совмещения событий A и B? Ответ на этот вопрос дает теорема умножения.

Теорема. *Вероятность совместного появления двух независимых событий равна произведению вероятностей этих событий:*

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

Доказательство. Введем обозначения:

n — число возможных элементарных исходов испытания, в которых событие A наступает или не наступает;

n_1— число исходов, благоприятствующих событию $A(n_1\leqslant n)$;

m —число возможных элементарных исходов испытания, в которых событие B наступает или не наступает;

m_1— число исходов, благоприятствующих событию $B(m_1\leqslant m)$.

Общее число возможных элементарных исходов испытания (в которых наступает и A и B, либо A и $\overline{B}$, либо $\overline{A}$ и B, либо $\overline{A}$ и $\overline{B}$) равно nm. Действительно, каждый из n исходов, в которых событие A наступает или не наступает, может сочетаться с каждым из m исходов, в которых событие B появляется или не появляется.

Из этого числа n_1m_1 исходов благоприятствуют совмещению событий A и B. Действительно, каждый из n_1 исходов, благоприятствующих событию A, может сочетаться с каждым из m_1 исходов, благоприятствующих событию B.

Вероятность совместного наступления событий A и B

$$P(AB)=\frac{n_1m_1}{nm}=\frac{n_1}{n}\cdot\frac{m_1}{m}.$$

Приняв во внимание, что $\frac{n_1}{n}=P(A)$ и $\frac{m_1}{m}=P(B)$, окончательно получим:

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

Для того чтобы обобщить теорему умножения на несколько событий, введем понятие независимости событий в совокупности.

Несколько событий называют *независимыми в совокупности*, если каждое из них и любая комбинация остальных событий (содержащая либо все остальные события, либо часть из них) есть события независимые. Например, если события A_1, A_2 и A_3 независимые в совокупности, то независимыми являются события: A_1 и A_2, A_1 и A_3, A_2 и A_3, A_1A_2 и A_3, A_1A_3 и A_2, A_2A_3 и A_1.

Подчеркнем, что если несколько событий независимы попарно, то отсюда еще не следует их независимость в совокупности. В этом смысле требование независимости событий в совокупности сильнее требования их попарной независимости.

Поясним сказанное примером. Пусть в урне имеется 4 шара, окрашенные: 1 — в красный цвет (A), 1 — в синий цвет (B), 1 — в черный цвет (C) и 1 — во все эти три цвета (ABC).

Чему равна вероятность $P(A)$ того, что извлеченный из урны шар имеет красный цвет? Так как из четырех шаров два имеют красный цвет, то $P(A)=\frac{2}{4}=\frac{1}{2}$. Рассуждая аналогично, найдем: $P(B)=\frac{1}{2}$, $P(C)=\frac{1}{2}$.

Допустим теперь, что взятый шар имеет синий цвет, т. е. что событие B уже произошло. Изменится ли вероятность того, что извлеченный шар имеет красный цвет, т. е. изменится ли вероятность события A? Из двух шаров, имеющих синий цвет, один шар имеет и красный цвет, поэтому вероятность события A, по-прежнему, равна $\frac{1}{2}$. Таким образом, события A и B независимы.

Рассуждая аналогично, придем к выводу, что события A и C, B и C независимы. Итак, события A, B и C попарно независимы.

Будут ли эти события независимы в совокупности? Оказывается, не будут. Действительно, пусть извлеченный шар имеет два цвета, например, синий и черный. Чему равна вероятность того, что этот шар имеет и красный цвет? Так как лишь один шар окрашен во все три цвета, то взятый шар имеет и красный цвет. Таким образом, допустив, что события B и C произошли, мы пришли к выводу, что событие A обязательно наступит. Следовательно, это событие достоверно и вероятность его равна единице (а не $\frac{1}{2}$).

Итак, попарно независимые события A, B и C не являются независимыми в совокупности.

Приведем теперь следствие из теоремы умножения.

Следствие. *Вероятность совместного появления нескольких событий, независимых в совокупности, равна произведению вероятностей этих событий:*

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

Доказательство. Рассмотрим три события A, B и C. Совмещение событий A, B и C равносильно совмещению событий AB и C, поэтому

$$P(ABC)=P(AB\cdot C).$$

Так как события A, B и C независимы в совокупности, то независимы, в частности события AB и C, а также A и B. По теореме умножения для двух независимых событий будем иметь:

$$P(AB\cdot C)=P(AB)\cdot P(C)$$

и

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

Итак, окончательно получим

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

Для произвольного n доказательство проводится методом математической индукции.

Замечание. Если события $A_1, A_2, \dots, A_n$ независимы в совокупности, то и противоположные им события $\overline{A}_1, \overline{A}_2, \dots, \overline{A}_n$ также независимы в совокупности.

Пример 1. Найти вероятность совместного появления герба при одном бросании двух монет.

Решение. Вероятность появления герба первой монеты (событие A)

$$P(A)=\frac{1}{2}.$$

Вероятность появления герба второй монеты (событие B)

$$P(B)=\frac{1}{2}.$$

Так как события A и B независимые, то искомая вероятность по теореме умножения равна

$$P(AB)=P(A)\cdot P(B)=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}.$$

Пример 2. Имеется 3 ящика, содержащих по 10 деталей. В первом ящике 8, во втором 7 и в третьем 9 стандартных деталей. Из каждого ящика наудачу вынимают по одной детали. Найти вероятность того, что все три вынутые детали окажутся стандартными.

Решение. Вероятность того, что из первого ящика вынута стандартная деталь (событие A)

$$P(A)=\frac{8}{10}=0{,}8.$$

Вероятность того, что из второго ящика вынута стандартная деталь (событие B)

$$P(B)=\frac{7}{10}=0{,}7.$$

Вероятность того, что из третьего ящика вынута стандартная деталь (событие C)

$$P(C)=\frac{9}{10}=0{,}9.$$

Так как события A, B и C независимые в совокупности, то искомая вероятность (по теореме умножения) равна

$$P(ABC)=P(A)\cdot P(B)\cdot P(C)=0{,}8\cdot 0{,}7\cdot 0{,}9=0{,}504.$$

Приведем пример совместного применения теорем сложения и умножения.

Пример 3. Вероятности появления каждого из трех независимых событий A_1, A_2, A_3 соответственно равны p_1, p_2, p_3. Найти вероятность появления только одного из этих событий.

Решение. Заметим, что, например, появление только первого события A_1 равносильно появлению события $A_1\overline{A}_2\overline{A}_3$ (появилось первое и не появились второе и третье события).

Введем обозначения:

B_1— появилось только событие A_1, т. е. $B_1=A_1\overline{A}_2\overline{A}_3$;
B_2— появилось только событие A_2, т. е. $B_2=A_2\overline{A}_1\overline{A}_3$;
B_3— появилось только событие A_3, т. е. $B_3=A_3\overline{A}_1\overline{A}_2$.

Таким образом, чтобы найти вероятность появления только одного из событий A_1, A_2, A_3, будем искать вероятность $P(B_1+B_2+B_3)$ появления одного, безразлично какого из событий B_1, B_2, B_3.

Так как события B_1, B_2, B_3 несовместны, то применима теорема сложения:

$$P(B_1+B_2+B_3)=P(B_1)+P(B_2)+P(B_3). \qquad (*)$$

Остается найти вероятности каждого из событий B_1, B_2, B_3.

События A_1, A_2, A_3 независимы, следовательно, независимы события A_1, $\overline{A}_2$, $\overline{A}_3$, поэтому к ним применима теорема умножения:

$$P(B_1)=P(A_1\overline{A}_2\overline{A}_3)=P(A_1)P(\overline{A}_2)P(\overline{A}_3)=p_1q_2q_3.$$

Аналогично:

$$P(B_2)=P(A_2\overline{A}_1\overline{A}_3)=P(A_2)P(\overline{A}_1)P(\overline{A}_3)=p_2q_1q_3;$$

$$P(B_3)=P(A_3\overline{A}_1\overline{A}_2)=P(A_3)P(\overline{A}_1)P(\overline{A}_2)=p_3q_1q_2.$$

Подставив эти вероятности в (*), найдем искомую вероятность появления только одного из событий A_1, A_2, A_3:

$$P(B_1+B_2+B_3)=p_1q_2q_3+p_2q_1q_3+p_3q_1q_2.$$

§ 3. Вероятность появления хотя бы одного события

Пусть в результате испытания может появиться n событий независимых в совокупности, либо некоторые из них (в частности, только одно или ни одного), причем вероятности появления каждого из событий известны. Как найти вероятность того,что наступит хотя бы одно из этих событий? Например, если в результате испытания могут появиться три события, то появление хотя бы одного из этих событий означает наступление либо одного, либо двух, либо трех событий. Ответ на поставленный вопрос дает следующая теорема.

Теорема. *Вероятность появления хотя бы одного из событий A_1, A_2, ..., A_n, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий $\overline{A}_1\overline{A}_2...\overline{A}_n$:*

$$P(A)=1-q_1q_2...q_n \qquad (*)$$

Доказательство. Обозначим через A событие, состоящее в появлении хотя бы одного из событий A_1, A_2, ..., A_n. События A и $\overline{A}_1\overline{A}_2...\overline{A}_n$ (ни одно из событий не наступило) противоположны, следовательно, сумма их вероятностей равна единице:

$$P(A)+P(\overline{A}_1\overline{A}_2...\overline{A}_n)=1$$

Отсюда, пользуясь теоремой умножения, получим:

$$P(A)=1-P(\overline{A}_1\overline{A}_2\ ...\ \overline{A}_n)=1-P(\overline{A}_1)\cdot P(\overline{A}_2)\ ...\ P(\overline{A}_n),$$

или

$$P(A)=1-q_1q_2\ ...\ q_n.$$

Частный случай. *Если события A_1, A_2, ..., A_n имеют одинаковую вероятность, равную p, то вероятность появления хотя бы одного из этих событий*

$$P(A)=1-q^n. \qquad (**)$$

Пример 1. Вероятности попадания в цель при стрельбе из трех орудий таковы: $p_1=0{,}8$; $p_2=0{,}7$; $p_3=0{,}9$. Найти вероятность хотя бы одного попадания (событие A) при одном залпе из всех орудий.

Решение. Вероятность попадания в цель каждым из орудий не зависит от результатов стрельбы из других орудий, поэтому рассматриваемые события A_1 (попадание первого орудия), A_2 (попадание второго орудия) и A_3 (попадание третьего орудия) независимы в совокупности.

Вероятности событий, противоположных событиям A_1, A_2 и A_3 (т. е. вероятности промахов), соответственно равны:

$$q_1=1-p_1=1-0,8=0,2;$$
$$q_2=1-p_2=1-0,7=0,3;$$
$$q_3=1-p_3=1-0,9=0,1.$$

Искомая вероятность

$$P(A)=1-q_1q_2q_3=1-0,2\cdot0,3\cdot0,1=0,994.$$

Пример 2. В типографии имеется 4 плоскопечатных машины. Для каждой машины вероятность того, что она работает в данный момент, равна 0,9. Найти вероятность того, что в данный момент работает хотя бы одна машина (событие A)

Решение. Так как события «машина работает» и «машина не работает» (в данный момент) противоположные, то сумма их вероятностей равна единице:

$$p+q=1.$$

Отсюда вероятность того, что машина в данный момент не работает, равна

$$q=1-p=1-0,9=0,1.$$

Искомая вероятность

$$P(A)=1-q^4=1-0,1^4=0,9999.$$

Так как полученная вероятность весьма близка к единице, то, на основании следствия из принципа практической невозможности маловероятных событий, мы вправе заключить, что в данный момент работает хотя бы одна из машин.

Пример 3. Вероятность того, что при одном выстреле стрелок попадает в цель, равна 0,4. Сколько выстрелов должен произвести стрелок, чтобы с вероятностью не менее 0,9 он попал в цель хотя бы один раз?

Решение. Обозначим через A событие: при n выстрелах стрелок попадает в цель хотя бы один раз.

События, состоящие в попадании в цель при первом, втором и т. д. выстрелах, независимы в совокупности, поэтому применима формула (**)

$$P(A)=1-q^n.$$

Приняв во внимание, что по условию $P(A)\geqslant0,9$, $p=0,4$ (следовательно, $q=1-0,4=0,6$), получим:

$$1-0,6^n\geqslant0,9.$$

Отсюда

$$0,6^n\leqslant0,1.$$

Прологарифмируем это неравенство по основанию 10:

$$n\lg0,6\leqslant\lg0,1.$$

Отсюда, учитывая, что $\lg 0,6<0$, имеем

$$n\geqslant\frac{\lg0,1}{\lg0,6}=\frac{-1}{\bar{1},7782}=\frac{-1}{-0,2218}=4,5.$$

Итак, $n\geqslant5$, т. е. стрелок должен произвести не менее 5 выстрелов.

Пример 4. Вероятность того, что событие появится хотя бы один раз в трех независимых в совокупности испытаниях, равна 0,936. Найти вероятность появления события в одном испытании (предполагается, что во всех испытаниях вероятность появления события одна и та же).

Решение. Так как рассматриваемые события независимы в совокупности, где применима формула (**)

$$P(A)=1-q^n.$$

По условию $P(A)=0,936$; $n=3$. Следовательно,

$$0,936=1-q^3,$$

или

$$q^3=1-0,936=0,064.$$

Отсюда

$$q=\sqrt[3]{0,064}=0,4.$$

Искомая вероятность

$$p=1-q=1-0,4=0,6.$$

4. Условная вероятность

Пусть события A и B зависимые. Из определения зависимых событий следует, что вероятность одного из событий зависит от появления или непоявления другого. Поэтому, если нас интересует вероятность, например события B, то важно знать, наступило ли событие A.

Условной вероятностью $P_A(B)$ называют вероятность события B, вычисленную в предположении, что событие A уже наступило.

Пример. В урне содержится 3 белых и 3 черных шара. Из урны дважды вынимают наудачу по одному шару, не возвращая их в урну. Найти вероятность появления белого шара при втором испытании (событие B), если при первом испытании был извлечен черный шар (событие A).

Р е ш е н и е. После первого испытания в урне осталось всего 5 шаров, из них 3 белых. Искомая условная вероятность

$$P_A(B) = \frac{3}{5}.$$

З а м е ч а н и е. Из определения независимых событий следует, что появление одного из них не изменяет вероятности наступления другого. Поэтому для независимых событий справедливы равенства:

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

Таким образом, условные вероятности независимых событий равны их безусловным вероятностям.

§ 5. Теорема умножения вероятностей зависимых событий

Пусть события A и B зависимые, причем вероятности $P(A)$ и $P_A(B)$ известны. Как найти вероятность совмещения этих событий, т. е. вероятность того, что появится и событие A и событие B? Ответ на этот вопрос дает теорема умножения.

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

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

Д о к а з а т е л ь с т в о. Введем обозначения:

n — число возможных элементарных исходов испытания, в которых событие A наступает или не наступает;

n_1 — число исходов, благоприятствующих событию $A(n_1 \leqslant n)$;

m — число элементарных исходов испытания, в которых наступает событие B, в предположении, что событие A уже наступило, т. е. эти исходы благоприятствуют наступлению события $AB(m \leqslant n_1)$.

Вероятность совместного появления событий A и B

$$P(AB) = \frac{m}{n} = \frac{n_1}{n} \cdot \frac{m}{n_1}.$$

Приняв во внимание, что $\frac{n_1}{n} = P(A)$ и $\frac{m}{n_1} = P_A(B)$, окончательно получим

$$P(AB) = P(A) \cdot P_A(B). \qquad (*)$$

З а м е ч а н и е 1. Применив формулу (*) к событию BA, имеем: $P(BA) = P(B) \cdot P_B(A)$, или (поскольку событие BA не отличается от события AB)

$$P(AB) = P(B) \cdot P_B(A) \qquad (**)$$

Сопоставляя формулы (*) и (**), заключаем о справедливости равенства

$$P(A) \cdot P_A(B) = P(B) \cdot P_B(A) \qquad (***)$$

Следствие. *Вероятность совместного появления нескольких зависимых событий равна произведению вероятности одного из них на условные вероятности всех остальных, причем вероятность каждого последующего события вычисляется в предположении, что все предыдущие события уже появились:*

$$P(A_1A_2A_3 \ldots A_n) = P(A_1) \cdot P_{A_1}(A_2) \cdot P_{A_1A_2}(A_3) \ldots P_{A_1A_2 \ldots A_{n-1}}(A_n),$$

где $P_{A_1A_2 \ldots A_{n-1}}(A_n)$ — вероятность события A_n, вычисленная в предположении, что события $A_1, A_2, \ldots, A_{n-1}$ наступили.

В частности, для трех зависимых событий будем иметь:

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

Заметим, что порядок, в котором расположены события, может быть выбран любым, т. е. безразлично, какое событие считать первым, вторым и т. д.

Для произвольного n доказательство производится методом математической индукции.

Пример 1. У сборщика имеется 3 конусных и 7 эллиптических валиков. Сборщик наудачу взял один валик, а затем второй. Найти вероятность того, что первый из взятых валиков — конусный, а второй — эллиптический.

Решение. Вероятность того, что первый из взятых валиков окажется конусным (событие A)

$$P(A) = \frac{3}{10}$$

Вероятность того, что второй из валиков окажется эллиптическим (событие B), вычисленная в предположении, что первый валик — конусный, т. е. условная вероятность равна

$$P_A(B) = \frac{7}{9}$$

Искомая вероятность по теореме умножения вероятностей зависимых событий равна

$$P(AB) = P(A) \cdot P_A(B) = \frac{3}{10} \cdot \frac{7}{9} = \frac{7}{30}.$$

Заметим, что сохранив обозначения, легко найдем $P(B) = \frac{7}{10}$, $P_B(A) = \frac{3}{9}$, $P(B) \cdot P_B(A) = \frac{7}{30}$, что наглядно иллюстрирует справедливость равенства (***).

Пример 2. В урне находится 5 белых, 4 черных и 3 синих шара. Каждое испытание состоит в том, что наудачу извлекают один шар, не возвращая его в урну. Найти вероятность того, что при первом испытании появится белый шар (событие A), при втором — черный (событие B) и при третьем — синий (событие C).

Решение. Вероятность появления белого шара при первом испытании

$$P(A) = \frac{5}{12}.$$

Вероятность появления черного шара при втором испытании, вычисленная в предположении, что при первом испытании появился белый шар, т. е. условная вероятность

$$P_A(B) = \frac{4}{11}$$

Вероятность появления синего шара при третьем испытании, вычисленная в предположении, что при первом испытании появился белый шар, а при втором — черный

$$P_{AB}(C) = \frac{3}{10}.$$

Искомая вероятность

$$P(ABC) = P(A) \cdot P_A(B) \cdot P_{AB}(C) = \frac{5}{12} \cdot \frac{4}{11} \cdot \frac{3}{10} = \frac{1}{22}.$$

Замечание 2. Выразим условную вероятность из соотношения (*) считая $P(A) \neq 0$:

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

Это равенство можно принять в качестве определения условной вероятности

Задачи

1. Вероятность того, что стрелок при одном выстреле попадает в мишень, равна $p = 0,9$. Стрелок произвел 3 выстрела. Найти вероятность того, что все 3 выстрела дали попадание

Отв. 0,729.

2. Брошены монета и игральная кость. Найти вероятность совмещения событий: «появился герб», «появилось 6 очков»

Отв. $\frac{1}{12}$

3. В двух ящиках находятся детали: в первом — 10 (из них 3 стандартных), во втором — 15 (из них 6 стандартных). Из каждого ящика наудачу вынимают по одной детали Найти вероятность того, что обе детали окажутся стандартными.

Отв 0,12.

4. В студии телевидения 3 телевизионных камеры Для каждой камеры вероятность того, что она включена в данный момент, равна $p = 0,6$. Найти вероятность того, что в данный момент включена хотя бы одна камера (событие A).

Отв. 0,936

5. Чему равна вероятность того, что при бросании трех игральных костей 6 очков появится хотя бы на одной из костей (событие A)?

Отв. $\frac{91}{216}$.

6. Предприятие изготовляет 95% изделий стандартных, причем из них 86% — первого сорта. Найти вероятность того что взятое наудачу изделие изготовленное на этом предприятии окажется первого сорта.

Отв. 0,817

7. Монета бросается до тех пор, пока 2 раза подряд она не выпадет одной и той же стороной Найти вероятности следующих событий: а) опыт окончится до шестого бросания: б) потребуется четное число бросаний

Отв а) $\frac{15}{16}$. б) $\frac{2}{3}$

8. Из цифр 1, 2, 3, 4, 5 сначала выбирается одна, а затем из оставшихся четырех — вторая цифра. Предполагается, что все 20 возможных исходов равновероятны. Найти вероятность того, что будет выбрана нечетная цифра: а) в первый раз; б) во второй раз; в) в оба раза.

Отв. а) $\frac{3}{5}$; б) $\frac{3}{5}$; в) $\frac{3}{10}$

9. Вероятность того, что при одном выстреле стрелок попадет в десятку, равна 0,6. Сколько выстрелов должен сделать стрелок, чтобы с вероятностью не менее 0,8 он попал в десятку хотя бы один раз?

Отв. $n \geq 2$.

10. Три электрические лампочки последовательно включены в цепь. Вероятность того, что одна (любая) лампочка перегорит, если напряжение в сети превысит номинальное, равна 0,6. Найти вероятность того, что при повышенном напряжении тока в цепи не будет.

Отв. 0,936

11. Вероятность того, что событие A появится хотя бы один раз при двух независимых испытаниях, равна 0,75. Найти вероятность появления события в одном испытании (предполагается, что вероятность появления события в обоих испытаниях одна и та же).

Отв. 0,5.

12. Три команды A_1, A_2, A_3 спортивного общества A состязаются соответственно с тремя командами общества B. Вероятности того, что команды общества A выиграют матчи у команд общества B таковы: при встрече A_1 с B_1 — 0,8; A_2 с B_2 — 0,4; A_3 с B_3 — 0,4. Для победы необходимо выиграть не менее двух матчей из трех (ничьи во внимание не принимаются). Победа какого из обществ вероятнее?

Отв. Общества A $\left(P_A = 0{,}544 > \frac{1}{2}\right)$

13. Вероятность поражения цели первым стрелком при одном выстреле равна 0,8, а вторым стрелком — 0,6. Найти вероятность того, что цель будет поражена только одним стрелком.

Отв. 0,44.

14. Из последовательности чисел 1, 2, ..., n наудачу одно за другим выбираются два числа. Найти вероятность того, что одно из них меньше целого положительного числа k, а другое больше k, где $1 < k < n$.

Отв. $\frac{2(k-1)(n-k)}{n(n-1)}$.

Указание. Сделать допущения: а) первое число $< k$, а второе $> k$; б) первое число $> k$, а второе $< k$.

15. Отдел технического контроля проверяет изделия на стандартность. Вероятность того, что изделие нестандартно, равна 0,1. Найти вероятность того, что: а) из трех проверенных изделий только одно окажется нестандартным; б) нестандартным окажется только четвертое по порядку проверенное изделие.

Отв. а) 0,243; б) 0,0729.

Глава четвертая

СЛЕДСТВИЯ ТЕОРЕМ СЛОЖЕНИЯ И УМНОЖЕНИЯ

§ 1. Теорема сложения вероятностей совместных событий

Нами была рассмотрена теорема сложения для несовместных событий. Здесь будет изложена теорема сложения для совместных событий.

Два события называют *совместными*, если появление одного из них не исключает появления другого в одном и том же испытании.

Пример. A — появление четырех очков при бросании игральной кости; B — появление четного числа очков. События A и B — совместные.

Пусть события A и B совместны, причем даны вероятности этих событий и вероятность их совместного появления. Как найти вероятность события $A+B$, состоящего в том, что появится хотя бы одно из событий A и B? Ответ на этот вопрос дает теорема сложения вероятностей совместных событий.

Теорема. *Вероятность появления хотя бы одного из двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного появления*

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

Доказательство. Поскольку события A и B по условию совместны, то событие $A+B$ наступит, если наступит одно из следующих трех несовместных событий: $A\bar{B}$, $\bar{A}B$ или AB. По теореме сложения вероятностей несовместных событий

$$P(A+B)=P(A\bar{B})+P(\bar{A}B)+P(AB) \qquad (*)$$

Событие A произойдет, если наступит одно из двух несовместных событий: $A\bar{B}$ или AB. По теореме сложения вероятностей несовместных событий имеем:

$$P(A)=P(A\bar{B})+P(AB).$$

Отсюда

$$P(A\bar{B})=P(A)-P(AB). \qquad (**)$$

Аналогично будем иметь:

$$P(B)=P(\overline{A}B)+P(AB)$$

Отсюда

$$P(\overline{A}B)=P(B)-P(AB). \qquad (***)$$

Подставив (**) и (***) в (*), окончательно получим:

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

Замечание 1. При использовании полученной формулы следует иметь в виду, что события A и B могут быть как независимыми, так и зависимыми.

Для независимых событий

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

для зависимых событий

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

Замечание 2. Если события A и B несовместны, то их совмещение есть невозможное событие и, следовательно, $P(AB)=0$. Формула (****) для несовместных событий принимает вид

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

Мы вновь получили теорему сложения для несовместных событий. Таким образом, формула (****) справедлива как для совместных, так и для несовместных событий.

Пример. Вероятности попадания в цель при стрельбе первого и второго орудий соответственно равны: $p_1=0{,}7$; $p_2=0{,}8$. Найти вероятность попадания при одном залпе (из обоих орудий) хотя бы одним из орудий.

Решение. Вероятность попадания в цель каждым из орудий не зависит от результата стрельбы из другого орудия, поэтому события A (попадание первого орудия) и B (попадание второго орудия) независимы.

Вероятность события AB (оба орудия дали попадание)

$$P(AB)=P(A)\cdot P(B)=0{,}7\cdot 0{,}8=0{,}56.$$

Искомая вероятность

$$P(A+B)=P(A)+P(B)-P(AB)=$$
$$=0{,}7+0{,}8-0{,}56=0{,}94$$

Замечание. Так как в настоящем примере события A и B независимые, то можно было воспользоваться формулой $P=1-q_1q_2$ (гл. III, § 3).

В самом деле, вероятности событий, противоположных событиям A и B, т. е. вероятности промахов таковы:

$$q_1=1-p_1=1-0{,}7=0{,}3;$$
$$q_2=1-p_2=1-0{,}8=0{,}2.$$

Искомая вероятность того, что при одном залпе хотя бы одно орудие даст попадание, равна

$$P=1-q_1q_2=1-0{,}3\cdot 0{,}2=0{,}94$$

Как и следовало ожидать, мы получили тот же результат.

Формула полной вероятности

Пусть событие A может наступить при условии появления одного из несовместных событий $B_1, B_2, \ldots, B_n$, которые образуют полную группу. Пусть известны вероятности этих событий и условные вероятности $P_{B_1}(A)$, $P_{B_2}(A)$, …, $P_{B_n}(A)$ события A. Как найти вероятность события A? Ответ на этот вопрос дает следующая теорема.

Теорема. *Вероятность события A, которое может наступить лишь при условии появления одного из несовместных событий $B_1, B_2, \ldots, B_n$, образующих полную группу, равна сумме произведений вероятностей каждого из этих событий на соответствующую условную вероятность события A:*

$$P(A)=P(B_1)\cdot P_{B_1}(A)+P(B_2)\cdot P_{B_2}(A)+\ldots+P(B_n)\cdot P_{B_n}(A)$$

Эту формулу называют «формулой полной вероятности».

Доказательство. По условию событие A может наступить, если наступит одно из несовместных событий $B_1, B_2, \ldots, B_n$. Другими словами, появление события A означает осуществление одного, безразлично какого, из несовместных событий $B_1A, B_2A, \ldots, B_nA$. Пользуясь для вычисления вероятности события A теоремой сложения, получим

$$P(A)=P(B_1A)+P(B_2A)+\ldots+P(B_nA). \qquad (*)$$

Остается вычислить каждое из слагаемых. По теореме умножения вероятностей зависимых событий имеем

$$P(B_1A)=P(B_1)\cdot P_{B_1}(A);\quad P(B_2A)=$$
$$=P(B_2)\cdot P_{B_2}(A);\quad \ldots;\quad P(B_nA)=P(B_n)\cdot P_{B_n}(A)$$

Подставив правые части этих равенств в соотношение (*), получим формулу полной вероятности

$$P(A)=P(B_1)\cdot P_{B_1}(A)+P(B_2)\cdot P_{B_2}(A)+\ldots+P(B_n)\cdot P_{B_n}(A)$$

Пример 1. Имеется два набора деталей. Вероятность того, что деталь первого набора стандартна, равна 0,8, а второго — 0,9. Найти вероятность того, что взятая наудачу деталь (из наудачу взятого набора) — стандартная.

Решение. Обозначим через A событие — извлеченная деталь стандартна.

Деталь может быть извлечена либо из первого набора (событие B_1), либо из второго (событие B_2).

Вероятность того, что деталь будет вынута из первого набора

$$P(B_1) = \frac{1}{2}.$$

Вероятность того, что деталь будет вынута из второго набора

$$P(B_2) = \frac{1}{2}.$$

Условная вероятность того, что из первого набора будет извлечена стандартная деталь

$$P_{B_1}(A) = 0{,}8.$$

Условная вероятность того, что из второго набора будет извлечена стандартная деталь

$$P_{B_2}(A) = 0{,}9.$$

Искомая вероятность того, что извлеченная наудачу деталь — стандартная по формуле полной вероятности равна

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) = \\ = 0{,}5 \cdot 0{,}8 + 0{,}5 \cdot 0{,}9 = 0{,}85.$$

Пример 2. В первой коробке содержится 20 радиоламп, из них 18 стандартных; во второй коробке — 10 ламп, из них 9 стандартных. Из второй коробки наудачу взята лампа и переложена в первую. Найти вероятность того, что лампа, наудачу извлеченная из первой коробки, будет стандартной.

Решение. Обозначим через A событие — из первой коробки извлечена стандартная лампа.

Из второй коробки могла быть извлечена либо стандартная лампа (событие B_1), либо нестандартная (событие B_2).

Вероятность того, что из второй коробки извлечена стандартная лампа,

$$P(B_1) = \frac{9}{10}.$$

Вероятность того, что из второй коробки извлечена нестандартная лампа,

$$P(B_2) = \frac{1}{10}.$$

Условная вероятность того, что из первой коробки извлечена стандартная лампа, при условии, что из второй коробки в первую была переложена стандартная лампа, равна

$$P_{B_1}(A) = \frac{19}{21}.$$

Условная вероятность того, что из первой коробки извлечена стандартная лампа, при условии, что из второй коробки в первую была переложена нестандартная лампа, равна

$$P_{B_2}(A) = \frac{18}{21}.$$

Искомая вероятность того, что из первой коробки будет извлечена стандартная лампа, по формуле полной вероятности равна

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) = \frac{9}{10} \cdot \frac{19}{21} + \\ + \frac{1}{10} \cdot \frac{18}{21} = 0{,}9.$$

Вероятность гипотез. Формулы Бейеса

Пусть событие A может наступить при условии появления одного из несовместных событий B_1, B_2, ..., B_n, образующих полную группу. Поскольку заранее неизвестно, какое из этих событий наступит, их называют *гипотезами*. Вероятность появления события A определяется по формуле полной вероятности (§ 2):

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \ldots + \\ + P(B_n) \cdot P_{B_n}(A) \qquad (*)$$

Допустим, что произведено испытание, в результате которого появилось событие A. Поставим своей задачей определить, как изменились (в связи с тем, что событие A

уже наступило) вероятности гипотез. Другими словами, будем искать условные вероятности

$$P_A(B_1),\ P_A(B_2),\ \ldots,\ P_A(B_n).$$

Найдем сначала условную вероятность $P_A(B_1)$. По теореме умножения имеем

$$P(AB_1)=P(A)\cdot P_A(B_1)=P(B_1)\cdot P_{B_1}(A)$$

Отсюда

$$P_A(B_1)=\frac{P(B_1)\cdot P_{B_1}(A)}{P(A)}$$

Заменив здесь $P(A)$ по формуле (*), получим

$$P_A(B_1)=\frac{P(B_1)\cdot P_{B_1}(A)}{P(B_1)\cdot P_{B_1}(A)+P(B_2)\cdot P_{B_2}(A)+\cdots+P(B_n)\cdot P_{B_n}(A)}.$$

Аналогично выводятся формулы, определяющие условные вероятности остальных гипотез, т. е. условная вероятность любой гипотезы B_i ($i=1, 2, \ldots, n$) может быть вычислена по формуле

$$P_A(B_i)=\frac{P(B_i)\cdot P_{B_i}(A)}{P(B_1)\cdot P_{B_1}(A)+P(B_2)\cdot P_{B_2}(A)+\cdots+P(B_n)\cdot P_{B_n}(A)}$$

Полученные формулы называют *формулами Бейеса* (по имени английского математика, который их вывел; опубликованы в 1764 г.). **Формулы Бейеса позволяют переоценить вероятности гипотез после того, как становится известным результат испытания, в итоге которого появилось событие A.**

Пример. Детали, изготовляемые цехом завода, попадают для проверки их на стандартность к одному из двух контролеров. Вероятность того, что деталь попадает к первому контролеру, равна 0,6, а ко второму — 0,4. Вероятность того, что годная деталь будет признана стандартной первым контролером, равна 0,94, а вторым — 0,98. Годная деталь при проверке была признана стандартной. Найти вероятность того, что эту деталь проверил первый контролер.

Решение. Обозначим через A событие, состоящее в том, что годная деталь признана стандартной. Можно сделать два предположения:

1) деталь проверил первый контролер (гипотеза B_1);
2) деталь проверил второй контролер (гипотеза B_2).

Искомую вероятность того, что деталь проверил первый контролер, найдем по формуле Бейеса

$$P_A(B_1)=\frac{P(B_1)\cdot P_{B_1}(A)}{P(B_1)\cdot P_{B_1}(A)+P(B_2)\cdot P_{B_2}(A)}$$

По условию задачи имеем:

$P(B_1)=0,6$ (вероятность того, что деталь попадет к первому контролеру);
$P(B_2)=0,4$ (вероятность того, что деталь попадет ко второму контролеру);
$P_{B_1}(A)=0,94$ (вероятность того, что годная деталь будет признана первым контролером стандартной);
$P_{B_2}(A)=0,98$ (вероятность того, что годная деталь будет признана вторым контролером стандартной).

Искомая вероятность

$$P_A(B_1)=\frac{0,6\cdot 0,94}{0,6\cdot 0,94+0,4\cdot 0,98}\approx 0,59$$

Как видно, до испытания вероятность гипотезы B_1 равнялась 0,6, а после того, как стал известен результат испытания, вероятность этой гипотезы (точнее, условная вероятность) изменилась и стала равной 0,59. Таким образом, использование формулы Бейеса позволило переоценить вероятность рассматриваемой гипотезы.

Задачи

1. Два стрелка произвели по одному выстрелу. Вероятность попадания в мишень первым стрелком равна 0,7, а вторым — 0,6. Найти вероятность того, что хотя бы один из стрелков попал в мишень.

Отв. 0,88.

2. У сборщика имеется 16 деталей, изготовленных заводом № 1, и 4 детали завода № 2. Наудачу взяты 2 детали. Найти вероятность того, что хотя бы одна из них окажется изготовленной заводом № 1.

Отв. $\frac{92}{95}$.

3. В группе спортсменов 20 лыжников, 6 велосипедистов и 4 бегуна. Вероятность выполнить квалификационную норму такова: для лыжника 0,9, для велосипедиста 0,8 и для бегуна 0,75. Найти вероятность того, что спортсмен, выбранный наудачу, выполнит норму.

Отв. 0,86.

4. Сборщик получил 3 коробки деталей, изготовленных заводом № 1, и 2 коробки деталей, изготовленных заводом № 2. Вероятность того, что деталь завода № 1 стандартна, равна 0,8, а за-

вода № 2 — 0,9. Сборщик наудачу извлек деталь из наудачу взятой коробки. Найти вероятность того, что извлечена стандартная деталь.

Отв. 0,84

5. В первом ящике содержится 20 деталей, из них 15 стандартных; во втором — 30 деталей, из них 24 стандартных; в третьем — 10 деталей, из них 6 стандартных. Найти вероятность того, что наудачу извлеченная деталь из наудачу взятого ящика — стандартная.

Отв. $\frac{43}{60}$

6. В телевизионном ателье имеется 4 кинескопа. Вероятности того, что кинескоп выдержит гарантийный срок службы, соответственно равны 0,8; 0,85; 0,9; 0,95. Найти вероятность того, что взятый наудачу кинескоп выдержит гарантийный срок службы.

Отв. 0,875.

7. В двух ящиках имеются радиолампы. В первом ящике содержится 12 ламп, из них 1 нестандартная; во втором 10 ламп, из них 1 нестандартная. Из первого ящика наудачу взята лампа и переложена во второй. Найти вероятность того, что наудачу извлеченная из второго ящика лампа будет нестандартной.

Отв. $\frac{13}{132}$

8. Из полного набора 28 костей домино наудачу извлечена кость. Найти вероятность того, что вторую извлеченную наудачу кость можно приставить к первой.

Отв. $\frac{7}{18}$

9. Студент знает не все экзаменационные билеты. В каком случае вероятность вытащить неизвестный билет будет для него наименьшей: когда он берет билет первым или последним?

Отв. Вероятности одинаковы в обоих случаях.

10. В ящик, содержащий 3 одинаковых детали, брошена стандартная деталь, а затем наудачу извлечена одна деталь. Найти вероятность того, что извлечена стандартная деталь, если равновероятны все возможные предположения о числе стандартных деталей, первоначально находившихся в ящике.

Отв. 0,625.

11. При отклонении от нормального режима работы автомата срабатывает сигнализатор С-1 с вероятностью 0,8, а сигнализатор С-11 срабатывает с вероятностью 1. Вероятности того, что автомат снабжен сигнализатором С-1 или С-11, соответственно равны 0,6 и 0,4. Получен сигнал о разладке автомата. Что вероятнее: автомат снабжен сигнализатором С-1 или С-11?

Отв. Вероятность того, что автомат снабжен сигнализатором С-1, равна $\frac{6}{11}$, а С-11 — $\frac{5}{11}$.

12. Для участия в студенческих отборочных спортивных соревнованиях выделено из первой группы курса — 4, из второй — 6, из третьей группы — 5 студентов. Вероятности того, что студент первой, второй и третьей группы попадет в сборную института, соответственно равны 0,9; 0,7; и 0,8. Наудачу выбранный студент в итоге соревнования попал в сборную. К какой из групп вероятнее всего принадлежал этот студент?

Отв. Вероятности того, что выбран студент первой, второй, третьей групп соответственно равны: $\frac{18}{59}$, $\frac{21}{59}$, $\frac{20}{59}$.

13. Вероятность для изделий некоторого производства удовлетворять стандарту равна 0,96. Предлагается упрощенная система проверки на стандартность, дающая положительный результат с вероятностью 0,98 для изделий, удовлетворяющих стандарту, а для изделий, которые не удовлетворяют стандарту, с вероятностью 0,05. Найти вероятность того, что изделие, признанное при проверке стандартным, действительно удовлетворяет стандарту.

Отв. 0,998.

Глава пятая

ПОВТОРЕНИЕ ИСПЫТАНИЙ

§ 1. Формула Бернулли

Если производится несколько испытаний, причем вероятность события A в каждом испытании не зависит от исходов других испытаний, то такие испытания называют *независимыми относительно события* A.

В разных независимых испытаниях событие A может иметь либо различные вероятности, либо одну и ту же вероятность. Мы будем далее рассматривать лишь такие независимые испытания, в которых событие A имеет одну и ту же вероятность.

Ниже мы воспользуемся понятием сложного события, понимая под ним совмещение нескольких отдельных событий, которые называют простыми.

Пусть производится n независимых испытаний, в каждом из которых событие A может появиться либо не появиться. Будем считать, что вероятность события A в каждом испытании одна и та же, а именно равна p. Следовательно, вероятность ненаступления события A в каждом испытании также постоянна и равна $q=1-p$.

Поставим своей задачей вычислить вероятность того, что при n испытаниях событие A осуществится ровно k раз и, следовательно, не осуществится $n-k$ раз.

Важно подчеркнуть, что не требуется, чтобы событие A повторилось ровно k раз в определенной последовательности. Например, если речь идет о появлении события A

три раза в четырех испытаниях, то возможны следующие сложные события:

$$AAA\bar{A},\ AA\bar{A}A,\ A\bar{A}AA \text{ и } \bar{A}AAA.$$

Запись $AAA\bar{A}$ означает, что в первом, втором и третьем испытаниях событие A наступило, а в четвертом испытании оно не появилось, т. е. наступило противоположное событие $\bar{A}$; соответственный смысл имеют и другие записи.

Искомую вероятность обозначим $P_n(k)$. Например, символ $P_5(3)$ означает вероятность того, что в пяти испытаниях событие появится ровно 3 раза и, следовательно, не наступит 2 раза.

Поставленную задачу решает так называемая формула Бернулли.

Вывод формулы Бернулли. Вероятность одного сложного события, состоящего в том, что в n испытаниях событие A наступит k раз и не наступит $n-k$ раз, по теореме умножения вероятностей независимых событий, равна

$$p^kq^{n-k}.$$

Таких сложных событий может быть столько, сколько можно составить сочетаний из n элементов по k элементов, т. е. C_n^k. Так как эти сложные события несовместны, то по теореме сложения вероятностей несовместных событий искомая вероятность равна сумме вероятностей всех возможных сложных событий. Поскольку же вероятности всех этих сложных событий одинаковы, то искомая вероятность (появления k раз события A в n испытаниях) равна вероятности одного сложного события, умноженной на их число:

$$P_n(k) = C_n^kp^kq^{n-k}.$$

или

$$P_n(k) = \frac{n!}{k!\,(n-k)!}\,p^kq^{n-k}.$$

Полученную формулу называют формулой Бернулли.

Пример. Вероятность того, что расход электроэнергии на продолжении одних суток не превысит установленной нормы, равна $p=0{,}75$. Найти вероятность того, что в ближайшие 6 суток расход электроэнергии в течение 4 суток не превысит нормы.

Р е ш е н и е. Вероятность нормального расхода электроэнергии на продолжении каждых из 6 суток постоянна

и равна $p=0{,}75$. Следовательно, вероятность перерасхода электроэнергии в каждые сутки также постоянна и равна $q=1-p=1-0{,}75=0{,}25$.

Искомая вероятность по формуле Бернулли равна

$$P_6(4) = C_6^4\,p^4q^2 = C_6^2\,p^4q^2 = \frac{6\cdot 5}{1\cdot 2}\,(0{,}75)^4\cdot(0{,}25)^2 = 0{,}30.$$

§ 2. Локальная теорема Лапласа

Выше мы вывели формулу Бернулли, позволяющую вычислить вероятность того, что событие появится в n испытаниях ровно k раз. При выводе мы предполагали, что вероятность появления события в каждом испытании постоянна.

Легко видеть, что пользоваться формулой Бернулли при больших значениях n достаточно трудно, так как формула требует выполнения действий над громадными числами. Например, если $n=50$, $k=30$, $p=0{,}1$, то для отыскания вероятности $P_{50}(30)$ надо вычислить выражение $P_{50}(30)=\frac{50!}{30!\,20!}\,(0{,}1)^{30}\cdot(0{,}9)^{20}$, где $50! = 30\,414\,093\cdot 10^{57}$, $30!=26\,525\,286\cdot 10^{25}$, $20! = 24\,329\,020\cdot 10^{11}$. Правда, можно несколько упростить вычисления, пользуясь специальными таблицами логарифмов факториалов. Однако и этот путь остается громоздким и, к тому же, имеет существенный недостаток: таблицы содержат приближенные значения логарифмов, поэтому в процессе вычислений накапливаются погрешности; в итоге окончательный результат может значительно отличаться от истинного.

Естественно возникает вопрос: нельзя ли вычислить интересующую нас вероятность, не прибегая к формуле Бернулли? Оказывается, можно. Локальная теорема Лапласа и дает асимптотическую* формулу, которая позволяет приближенно найти вероятность появления события ровно k раз в n испытаниях, если число испытаний достаточно велико.

Заметим, что для частного случая, а именно для $p=\frac{1}{2}$, асимптотическая формула была найдена в 1730 г. Муав-

* Функцию $\varphi(x)$ называют асимптотическим приближением функции $f(x)$, если $\lim\limits_{x\to\infty}\frac{f(x)}{\varphi(x)}=1$.

ром; в 1783 г. Лаплас обобщил формулу Муавра для произвольного p, отличного от 0 и 1. Поэтому теорему, о которой здесь идет речь, иногда называют теоремой Муавра—Лапласа.

Доказательство локальной теоремы Лапласа довольно сложно, поэтому мы приведем лишь формулировку теоремы и примеры, иллюстрирующие ее использование.

Локальная теорема Лапласа. *Если вероятность p появления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность $P_n(k)$ того, что событие A появится в n испытаниях ровно k раз, приближенно равна (тем точнее, чем больше n) значению функции*

$$y = \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{npq}} \cdot \varphi(x)$$

при $x = \dfrac{k-np}{\sqrt{npq}}$.

Имеются таблицы, в которых помещены значения функции $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, соответствующие положительным значениям аргумента x (приложение 1). Для отрицательных значений аргумента пользуются теми же таблицами, так как функция $\varphi(x)$ четна, т. е. $\varphi(-x) = \varphi(x)$.

Итак, вероятность того, что событие A появится в n независимых испытаниях ровно k раз, приближенно равна

$$P_n(k) \approx \frac{1}{\sqrt{npq}} \cdot \varphi(x),$$

где $x = \dfrac{k-np}{\sqrt{npq}}$.

Пример 1. Найти вероятность того, что событие A наступит ровно 80 раз в 400 испытаниях, если вероятность появления этого события в каждом испытании равна 0,2.

Р е ш е н и е. По условию $n=400$; $k=80$; $p=0{,}2$; $q=0{,}8$. Воспользуемся асимптотической формулой Лапласа:

$$P_{400}(80) \approx \frac{1}{\sqrt{400 \cdot 0{,}2 \cdot 0{,}8}} \cdot \varphi(x) = \frac{1}{8} \cdot \varphi(x).$$

Вычислим определяемое данными задачи значение x:

$$x = \frac{k-np}{\sqrt{npq}} = \frac{80 - 400 \cdot 0{,}2}{8} = 0.$$

По таблице (приложение 1) находим $\varphi(0)=0{,}3989$.

Искомая вероятность

$$P_{400}(80) = \frac{1}{8} \cdot 0{,}3989 = 0{,}04986.$$

Формула Бернулли приводит примерно к такому же результату (выкладки, ввиду их громоздкости, опущены):

$$P_{400}(80) = 0{,}0498.$$

Пример 2. Вероятность поражения мишени стрелком при одном выстреле $p=0{,}75$. Найти вероятность того, что при 10 выстрелах стрелок поразит мишень 8 раз.

Р е ш е н и е. По условию $n=10$; $k=8$; $p=0{,}75$; $q=0{,}25$. Воспользуемся асимптотической формулой Лапласа:

$$P_{10}(8) \simeq \frac{1}{\sqrt{10 \cdot 0{,}75 \cdot 0{,}25}} \cdot \varphi(x) = 0{,}7301 \cdot \varphi(x).$$

Вычислим определяемое данными задачи значение x:

$$x = \frac{k-np}{\sqrt{npq}} = \frac{8 - 10 \cdot 0{,}75}{10 \cdot 0{,}75 \cdot 0{,}25} \simeq 0{,}36.$$

По таблице (приложение 1) находим $\varphi(0{,}36)=0{,}3739$.

Искомая вероятность

$$P_{10}(8) = 0{,}7301 \cdot 0{,}3739 = 0{,}273.$$

Формула Бернулли приводит к иному результату, а именно $P_{10}(8) = 0{,}282$. Столь значительное расхождение ответов объясняется тем, что в настоящем примере n имеет малое значение (формула Лапласа дает достаточно хорошие приближения лишь при достаточно больших значениях n).

§ 3. Интегральная теорема Лапласа

Вновь предположим, что производится n испытаний, в каждом из которых вероятность появления события A постоянна и равна $p (0<p<1)$. Как вычислить вероятность $P_n(k_1, k_2)$, того, что событие A появится в n испытаниях не менее k_1 и не более k_2 раз (для краткости будем говорить «от k_1 до k_2 раз»)? На этот вопрос отвечает интегральная теорема Лапласа, которую мы приводим ниже, опустив доказательство.

Теорема. *Если вероятность p наступления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность $P_n(k_1, k_2)$ того, что событие A появится в n испытаниях от k_1 до k_2 раз, приближенно равна определенному интегралу*

$$P_n(k_1, k_2) \simeq \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-\frac{z^2}{2}}\,dz, \qquad (*)$$

где

$$x' = \frac{k_1 - np}{\sqrt{npq}} \quad \text{и} \quad x'' = \frac{k_2 - np}{\sqrt{npq}}.$$

При решении задач, требующих применения интегральной теоремы Лапласа, пользуются специальными таблицами, так как неопределенный интеграл $\int e^{-\frac{z^2}{2}}\,dz$ не выражается через элементарные функции. Таблица для интеграла $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}}\,dz$ приведена в конце книги (приложение 2). В таблице даны значения функции $\Phi(x)$ для положительных значений x и для $x = 0$; для $x < 0$ пользуются той же таблицей (функция $\Phi(x)$ нечетна, т. е. $\Phi(-x) = -\Phi(x)$). В таблице приведены значения интеграла лишь до $x=5$, так как для $x>5$ можно принять $\Phi(x)=0,5$. Функцию $\Phi(x)$ часто называют функцией Лапласа.

Для того чтобы можно было пользоваться таблицей функции Лапласа, преобразуем соотношение (*) так:

$$P_n(k_1, k_2) \simeq \frac{1}{\sqrt{2\pi}} \int_{x'}^{0} e^{-\frac{z^2}{2}}\,dz + \frac{1}{\sqrt{2\pi}} \int_0^{x''} e^{-\frac{z^2}{2}}\,dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{x''} e^{-\frac{z^2}{2}}\,dz - \frac{1}{\sqrt{2\pi}} \int_0^{x'} e^{-\frac{z^2}{2}}\,dz = \Phi(x'') - \Phi(x')$$

Итак, вероятность того, что событие A появится в n независимых испытаниях от k_1 до k_2 раз

$$P_n(k_1, k_2) \simeq \Phi(x'') - \Phi(x'),$$

где $x' = \frac{k_1 - np}{\sqrt{npq}}$ и $x'' = \frac{k_2 - np}{\sqrt{npq}}$

Приведем примеры, иллюстрирующие применение интегральной теоремы Лапласа.

Пример. Вероятность того, что деталь не прошла проверку ОТК, равна $p=0,2$. Найти вероятность того, что среди 400 случайно отобранных деталей окажется непроверенных от 70 до 100 деталей.

Решение. По условию $p=0,2$; $q=0,8$; $n=400$; $k_1=70$; $k_2=100$.

Воспользуемся интегральной теоремой Лапласа.

$$P_{400}(70, 100) \approx \Phi(x'') - \Phi(x').$$

Вычислим нижний и верхний пределы интегрирования:

$$x' = \frac{k_1 - np}{\sqrt{npq}} = \frac{70 - 400 \cdot 0,2}{\sqrt{400 \cdot 0,2 \cdot 0,8}} = -1,25;$$

$$x'' = \frac{k_2 - np}{\sqrt{npq}} = \frac{100 - 400 \cdot 0,2}{\sqrt{400 \cdot 0,2 \cdot 0,8}} = 2,5$$

Таким образом имеем

$$P_{400}(70, 100) = \Phi(2,5) - \Phi(-1,25) = \Phi(2,5) + \Phi(1,25).$$

По таблице (приложение 2) находим

$$\Phi(2,5)=0,4938; \quad \Phi(1,25)=0,3944$$

Искомая вероятность

$$P_{400}(70, 100)=0,4938+0,3944=0,8882.$$

Замечание. Обозначим через m число появлений события при n независимых испытаниях, в каждом из которых вероятность наступления события A постоянна и равна p. Если число m изменяется от k_1 до k_2, то дробь $\frac{m - np}{\sqrt{npq}}$ будет изменяться от $\frac{k_1 - np}{\sqrt{npq}} = x'$ до $\frac{k_2 - np}{\sqrt{npq}} = x''$. Следовательно, интегральную теорему Лапласа можно записать и так

$$P\left(x' \leqslant \frac{m - np}{\sqrt{npq}} \leqslant x''\right) \simeq \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-\frac{z^2}{2}}\,dz.$$

Эту форму записи мы используем ниже

§ 4. Вероятность отклонения относительной частоты от постоянной вероятности в независимых испытаниях

Вновь будем считать, что производится n независимых испытаний, в каждом из которых вероятность появления события A постоянна и равна p $(0<p<1)$. Поставим своей задачей найти вероятность того, что отклонение относительной частоты $\frac{m}{n}$ от постоянной вероятности p по абсолютной величине не превышает заданного числа $\varepsilon>0$. Другими словами, найдем вероятность осуществления неравенства

$$\left|\frac{m}{n}-p\right|\leqslant\varepsilon. \qquad (*)$$

Эту вероятность будем обозначать так: $P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)$.

Заменим неравенство (*) ему равносильными:

$$-\varepsilon\leqslant\frac{m}{n}-p\leqslant\varepsilon, \quad \text{или} \quad -\varepsilon\leqslant\frac{m-np}{n}\leqslant\varepsilon$$

Умножая эти неравенства на положительный множитель $\sqrt{\frac{n}{pq}}$, получим неравенства, равносильные исходному:

$$-\varepsilon\sqrt{\frac{n}{pq}}\leqslant\frac{m-np}{\sqrt{npq}}\leqslant\varepsilon\sqrt{\frac{n}{pq}}$$

Воспользуемся интегральной теоремой Лапласа в форме, указанной в замечании (стр. 51). Положив

$$x'=-\varepsilon\sqrt{\frac{n}{pq}} \text{ и } x''=\varepsilon\sqrt{\frac{n}{pq}}, \text{ имеем:}$$

$$P\left(-\varepsilon\sqrt{\frac{n}{pq}}\leqslant\frac{m-np}{\sqrt{npq}}\leqslant\varepsilon\sqrt{\frac{n}{nq}}\right)\approx$$

$$\approx\frac{1}{\sqrt{2\pi}}\int\limits_{-\varepsilon\sqrt{\frac{n}{pq}}}^{\varepsilon\sqrt{\frac{n}{pq}}}e^{-\frac{z^2}{2}}dz=\frac{2}{\sqrt{2\pi}}\int\limits_{0}^{\varepsilon\sqrt{\frac{n}{pq}}}e^{-\frac{z^2}{2}}dz=$$

$$=2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right).$$

Наконец, заменив неравенства, заключенные в скобках, равносильным им исходным неравенством, окончательно получим:

$$P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)\simeq2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right).$$

Итак, вероятность осуществления неравенства

$$\left|\frac{m}{n}-p\right|\leqslant\varepsilon$$

приближенно равна значению удвоенной функции Лапласа $2\Phi(x)$ при $x=\varepsilon\sqrt{\frac{n}{pq}}$.

Пример 1. Вероятность того, что деталь не стандартна, $p=0{,}1$. Найти вероятность того, что среди случайно отобранных 400 деталей относительная частота появления нестандартных деталей отклонится от вероятности $p=0{,}1$ по абсолютной величине не более, чем на 0,03.

Решение. По условию $n=400$; $p=0{,}1$; $q=0{,}9$; $\varepsilon=0{,}03$.

Требуется найти вероятность $P\left(\left|\frac{m}{400}-0{,}1\right|\leqslant0{,}03\right)$.

Пользуясь формулой $P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)\approx2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)$,

имеем: $P\left(\left|\frac{m}{400}-0{,}1\right|\leqslant0{,}03\right)\approx2\Phi\left(0{,}03\sqrt{\frac{400}{0{,}1\cdot0{,}9}}\right)=$

$=2\Phi(2)$.

По таблице (приложение 2) находим $\Phi(2)=0{,}4772$. Следовательно, $2\Phi(2)=0{,}9544$.

Итак, искомая вероятность приближенно равна 0,9544.

Смысл полученного результата таков: если взять достаточно большое число проб по 400 деталей в каждой, то примерно в 95,44% этих проб отклонение относительной частоты от постоянной вероятности $p=0{,}1$ по абсолютной величине не превысит 0,03.

Пример 2. Вероятность того, что деталь не стандартна, $p=0{,}1$. Найти, сколько деталей надо отобрать, чтобы с вероятностью равной 0,9544 можно было утверждать, что относительная частота появления нестандартных деталей (среди отобранных) отклонится от постоянной вероятности p по абсолютной величине не более, чем на 0,03.

Решение. По условию $p=0,1$; $q=0,9$; $\varepsilon=0,03$;

$$P\left(\left|\frac{m}{n}-0,1\right| \leqslant 0,03\right)=0,9544.$$

Требуется найти n.

Воспользуемся формулой $P\left(\left|\frac{m}{n}-p\right| \leqslant \varepsilon\right) \approx \approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)$

В силу условия,

$$2\Phi\left(0,03\sqrt{\frac{n}{0,1\cdot 0,9}}\right)=2\Phi\left(0,1\sqrt{n}\right)=0,9544.$$

Следовательно, $\Phi(0,1\sqrt{n})=0,4772$.

По таблице (приложение 2) находим

$$\Phi(2)=0,4772.$$

Для отыскания числа n получаем уравнение

$$0,1\sqrt{n}=2.$$

Отсюда искомое число деталей $n=400$.

Смысл полученного результата таков: если взять достаточно большое число проб по 400 деталей, то в 95, 44% этих проб относительная частота появления нестандартных деталей будет отличаться от постоянной вероятности $p=0,1$ по абсолютной величине не более, чем на 0,03, т. е. относительная частота будет заключена в границах от 0,07 (0,1—0,03=0,07) до 0,13 (0,1+0,03=0,13).

Другими словами, число нестандартных деталей в 95,44% проб будет заключено от 28 (7% от 400) до 52 (13% от 400).

Если взять лишь одну пробу из 400 деталей, то с большой уверенностью можно ожидать, что в этой пробе будет нестандартных деталей не менее 28 и не более 52. Возможно, хотя и маловероятно, что нестандартных деталей окажется меньше 28, либо больше 52.

Задачи

1. В цехе 6 моторов. Для каждого мотора вероятность того что он в данный момент включен, равна 0,8. Найти вероятность того, что в данный момент: а) включено 4 мотора, б) включены все моторы, в) выключены все моторы

Отв а) $P_6(4)=0,246$; б) $P_6(6)=0,26$; в) $P_6(0)=0,000064$.

2. Найти вероятность того, что событие A появится в пяти независимых испытаниях не менее двух раз, если в каждом испытании вероятность появления события A равна 0,3.

Отв. $P=1-[P_5(0)+ + P_5(1)]=0,472$

3. Событие B появится в случае, если событие A появится не менее двух раз. Найти вероятность того, что наступит событие B, если будет произведено 6 независимых испытаний, в каждом из которых вероятность появления события A равна 0,4.

Отв. $P=1-[P_6(0)+ + P_6(1)]=0{,}767$.

4. Произведено 8 независимых испытаний, в каждом из которых вероятность появления события A равна 0,1 Найти вероятность того, что событие A появится хотя бы 2 раза.

Отв $P=1-[P_8(0)+ + P_8(1)]=0,19$.

5. Монету бросают 6 раз Найти вероятност того что герб выпадет: а) менее двух раз, б) не менее двух раз.

Отв. а) $P=P_6(0)+ + P_6(1)=\frac{7}{64}$; б) $Q=1-[P_6(0)+ + P_6(1)]=\frac{57}{64}$

6. Вероятность попадания в цель при одном выстреле из орудия $p=0,9$. Вероятность поражения цели при k попаданиях $(k\geqslant 1)$ равна $1-q^k$ Найти вероятность того, что цель будет поражена, если сделано 2 выстрела.

Указание Воспользоваться формулами Бернулли и полной вероятности

Отв 0,9639.

7. Найти приближенно вероятность того что при 400 испытаниях событие наступит ровно 104 раза если вероятность его появления в каждом испытании равна 0,2

Отв $P_{400}(104)=0,0006$

8. Вероятность поражения мишени стрелком при одном выстреле равна 0,75 Найти вероятность того, что при 100 выстрелах мишень будет поражена: а) не менее 70 и не более 80 раз, б) не более 70 раз.

Отв а) $P_{100}(70,\ 80)= =2\Phi(1,15)=0,7498$; б) $P_{100}(0;\ 70)= =-\Phi(1,15)+0,5= =0,1251$

9. Вероятность появления события в каждом из 10 000 независимых испытаний $p = 0,75$. Найти вероятность того, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более, чем на 0,001.

Отв. $P = 2\Phi(0,23) = 0,182$.

10. Вероятность появления события в каждом из независимых испытаний равна 0,2. Найти, какое отклонение относительной частоты появления события от его вероятности можно ожидать с вероятностью 0,9128 при 5000 испытаниях.

Отв. $\varepsilon = 0,00967$.

11. Сколько раз надо бросить монету, чтобы с вероятностью 0,6 можно было ожидать, что отклонение относительной частоты появлений герба от вероятности $p = 0,5$ окажется по абсолютной величине не более 0,01?

Отв. $n = 1764$.

Часть вторая

СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Глава шестая

ВИДЫ СЛУЧАЙНЫХ ВЕЛИЧИН. ЗАДАНИЕ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Случайная величина

Уже в первой части приводились события, состоящие в появлении того или иного числа. Например, при бросании игральной кости могли появиться числа 1, 2, 3, 4, 5 и 6. Наперед определить число выпавших очков невозможно, поскольку оно зависит от многих случайных причин, которые полностью не могут быть учтены. В этом смысле число очков есть величина случайная; числа 1, 2, 3, 4, 5 и 6 есть *возможные значения* этой величины.

Случайной называют величину, которая в результате испытания примет одно и только одно возможное значение, наперед неизвестное и зависящее от случайных причин, которые заранее не могут быть учтены.

Пример 1. Число родившихся мальчиков среди ста новорожденных есть случайная величина, которая имеет следующие возможные значения: 0, 1, 2, ..., 100.

Пример 2. Расстояние, которое пролетит снаряд при выстреле из орудия, есть случайная величина. Действительно, расстояние зависит не только от установки прицела, но и от многих других причин (силы и направления ветра, температуры и т. д.), которые не могут быть полностью учтены. Возможные значения этой величины принадлежат некоторому промежутку (a, b).

Мы будем далее обозначать случайные величины прописными буквами X, Y, Z, а их возможные значения —

соответствующими строчными буквами x, y, z. Например, если случайная величина X имеет три возможных значения, то они будут обозначены так: x_1, x_2, x_3.

§ 2. Дискретные и непрерывные случайные величины

Вернемся к примерам, приведенным выше. В первом из них случайная величина X могла принять одно из следующих возможных значений: 0, 1, 2, ..., 100. Эти значения отделены одно от другого промежутками, в которых нет возможных значений X. Таким образом, в этом примере случайная величина принимает отдельные, изолированные возможные значения.

Во втором примере случайная величина могла принять любое из значений промежутка (a, b). Здесь нельзя отделить одно возможное значение от другого промежутком, не содержащим возможных значений случайной величины.

Уже из сказанного можно заключить о целесообразности различать случайные величины, принимающие лишь отдельные, изолированные значения и случайные величины, возможные значения которых сплошь заполняют некоторый промежуток.

Дискретной (*прерывной*) называют случайную величину, которая принимает отдельные, изолированные возможные значения с определенными вероятностями. Число возможных значений дискретной случайной величины может быть конечным или бесконечным.

Непрерывной называют случайную величину, которая может принимать все значения из некоторого конечного или бесконечного промежутка. Очевидно, число возможных значений непрерывной случайной величины — бесконечно.

З а м е ч а н и е. Настоящее определение непрерывной случайной величины не является точным. Более строгое определение будет дано позднее

§ 3. Закон распределения вероятностей дискретной случайной величины

На первый взгляд может показаться, что для задания дискретной случайной величины достаточно перечислить все ее возможные значения. В действительности это не так: случайные величины могут иметь о д и н а к о в ы е

перечни возможных значений, а вероятности их — р а з л и ч н ы е. Поэтому для задания дискретной случайной величины недостаточно перечислить все возможные ее значения, нужно еще указать их вероятности.

Законом распределения дискретной случайной величины называют соответствие между возможными значениями и их вероятностями; его можно задать таблично, аналитически (в виде формулы) и графически.

При табличном задании закона распределения дискретной случайной величины первая строка таблицы содержит возможные значения, а вторая — их вероятности:

X	x_1	x_2 ...	x_n
p	p_1	p_2 ...	p_n

Приняв во внимание, что в одном испытании случайная величина принимает одно и только одно возможное значение, заключаем, что события $X=x_1$, $X=x_2$, ..., $X=x_n$ образуют полную группу; следовательно, сумма вероятностей этих событий, т. е. сумма вероятностей второй строки таблицы равна единице:

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

Пример. В денежной лотерее выпущено 100 билетов. Разыгрывается один выигрыш в 50 руб. и десять выигрышей по 1 руб. Найти закон распределения случайной величины X — стоимости возможного выигрыша для владельца одного лотерейного билета.

Р е ш е н и е. Напишем возможные значения X:

$$x_1=50,\quad x_2=1,\quad x_3=0.$$

Вероятности этих возможных значений таковы:
$p_1=0{,}01$, $p_2=0{,}1$, $p_3=1-(p_1+p_2)=0{,}89$.

Напишем искомый закон распределения:

X	50	10	0
p	0,01	0,1	0,89

Контроль: $0{,}01+0{,}1+0{,}89=1$.

В целях наглядности закон распределения дискретной случайной величины можно изобразить и графически, для чего в прямоугольной системе координат строят точки (x_i, p_i), а затем соединяют их отрезками прямых. Полученную фигуру называют *многоугольником распределения*.

§ 4. Биноминальное распределение

Пусть производится n независимых испытаний, в каждом из которых событие A может появиться, либо не появиться. Вероятность наступления события во всех испытаниях постоянна и равна p (следовательно, вероятность непоявления $q=1-p$). Рассмотрим в качестве дискретной случайной величины X число появлений события A в этих испытаниях.

Поставим перед собой задачу: найти закон распределения величины X. Для ее решения требуется определить возможные значения X и их вероятности.

Очевидно, событие A в n испытаниях может либо не появиться, либо появиться 1 раз, либо 2 раза, ..., либо n раз. Таким образом, возможные значения X таковы:

$$x_1=0,\ x_2=1,\ x_3=2,\ \ldots,\ x_{n+1}=n.$$

Остается найти вероятности этих возможных значений, для чего достаточно воспользоваться формулой Бернулли:

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

где $k=0,\ 1,\ 2,\ \ldots,\ n$.

Формула (*) и является аналитическим выражением искомого закона распределения.

Биноминальным называют распределение вероятностей, определяемое формулой Бернулли.

Закон назван «биноминальным» потому, что правую часть равенства (*) можно рассматривать как общий член разложения бинома Ньютона

$$(p+q)^n=C_n^n p^n+C_n^{n-1}p^{n-1}q+\cdots+C_n^k p^k q^{n-k}+\cdots+C_n^0 q^n.$$

Таким образом, первый член разложения p^n определяет вероятность наступления рассматриваемого события n раз в n независимых испытаниях; второй член $np^{n-1}q$ определяет вероятность наступления события $n-1$ раз; ...; последний член q^n определяет вероятность того, что событие не появится ни разу.

Напишем биноминальный закон в виде таблицы:

X	n	$n-1$	...	k	...	0
P	p^n	$np^{n-1}q$	...	$C_n^k p^k q^{n-k}$	...	q^n.

Пример. Монета брошена 2 раза. Написать в виде таблицы закон распределения случайной величины X — числа выпадений герба.

Решение. Вероятность появления герба в каждом бросании монеты $p=\frac{1}{2}$, следовательно, вероятность непоявления герба $q=1-\frac{1}{2}=\frac{1}{2}$.

При двух бросаниях монеты герб может появиться либо 2 раза, либо 1 раз, либо совсем не появиться. Таким образом, возможные значения X таковы:

$$x_1=2,\ x_2=1,\ x_3=0.$$

Найдем вероятности этих возможных значений по формуле Бернулли:

$$P_2(2)=C_2^2 p^2=\left(\frac{1}{2}\right)^2=0{,}25.$$

$$P_2(1)=C_2^1 pq=2\cdot\frac{1}{2}\cdot\frac{1}{2}=0{,}5,$$

$$P_2(0)=C_2^0 q^2=\left(\frac{1}{2}\right)^2=0{,}25.$$

Напишем искомый закон распределения:

X	2	1	0
p	0,25	0,5	0,25.

Контроль: $0{,}25+0{,}5+0{,}25=1$.

§ 5. Распределение Пуассона

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A равна p. Для определения вероятности k появлений события в этих испытаниях используют формулу Бернулли. Если же n велико, то пользуются асимптотической формулой Лапласа. Однако, эта формула непригодна, если вероятность события мала ($p\leqslant 0{,}1$). В этих случаях (n велико, p мало) прибегают к асимптотической формуле Пуассона.

Итак, поставим своей задачей найти вероятность того, что при очень большом числе испытаний, в каждом из которых вероятность события очень мала, событие наступит ровно k раз.

Сделаем важное допущение: произведение np сохраняет постоянное значение, а именно $np=\lambda$. Как будет следовать из дальнейшего (гл. VII, § 5) это означает, что среднее число появлений события в различных сериях испытаний, т. е. при различных значениях n, остается неизменным.

Воспользуемся формулой Бернулли для вычисления интересующей нас вероятности:

$$P_n(k) = \frac{n(n-1)(n-2)\ldots[n-(k-1)]}{k!} p^k (1-p)^{n-k}.$$

Так как $pn=\lambda$, то $p=\frac{\lambda}{n}$. Следовательно,

$$P_n(k) = \frac{n(n-1)(n-2)\ldots[n-(k-1)]}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Приняв во внимание, что n имеет очень большое значение, вместо $P_n(k)$ найдем $\lim\limits_{n\to\infty} P_n(k)$. При этом будет найдено лишь приближенное значение отыскиваемой вероятности: n хотя и велико, но конечно, а при отыскании предела мы устремим n к бесконечности.

Итак,

$$P_n(k) \simeq \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots[n-(k-1)]}{k!} \cdot \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k} =$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left[1 \cdot \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \ldots \left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-k}\right] = \frac{\lambda^k}{k!} \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n \cdot \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} =$$

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

Итак (для простоты записи знак приближенного равенства опущен),

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

Эта формула выражает закон распределения Пуассона вероятностей массовых (n велико) редких (p мало) событий.

Замечание. Имеются специальные таблицы пользуясь которыми можно найти $P_n(k)$ зная k и λ.

Пример. Завод отправил на базу 5000 доброкачественных изделий. Вероятность того, что в пути изделие повредится, равна 0,0002. Найти вероятность того, что на базу прибудут 3 негодных изделия.

Решение. По условию $n=5000$, $p=0{,}0002$, $k=3$. Найдем λ:

$$\lambda = np = 5000 \cdot 0{,}0002 = 1$$

Искомая вероятность по формуле Пуассона приближенно равна

$$P_{5000}(3) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{e^{-1}}{3!} = \frac{1}{6e} \simeq 0{,}06.$$

§ 6. Простейший поток событий

Рассмотрим события, которые наступают в случайные моменты времени.

Потоком событий называют последовательность событий, которые наступают в случайные моменты времени. Примерами потоков могут служить: поступление вызовов на АТС, на пункт неотложной медицинской помощи, прибытие самолетов в аэропорт, клиентов на предприятие бытового обслуживания, последовательность отказов элементов и многие другие.

Среди свойств, которыми могут обладать потоки, выделим свойства стационарности, отсутствия последействия и ординарности.

Свойство стационарности характеризуется тем, что вероятность появления k событий на любом промежутке времени зависит только от числа k и от длительности t промежутка и не зависит от начала его отсчета; при этом различные промежутки времени предполагаются непересекающимися. Например, вероятности появления k событий на промежутках времени (1; 7), (10; 16), (T; $T+6$) одинаковой длительности $t=6$ единицам времени, равны между собой.

Итак, если поток обладает свойством стационарности, то вероятность появления k событий за промежуток времени длительности t есть функция, зависящая только от k и t.

Свойство «отсутствия последействия» характеризуется тем, что вероятность появления k событий на любом промежутке времени не зависит от того, появлялись или не появлялись события в моменты времени, предшествующие началу рассматриваемого промежутка. Другими словами, условная вероятность появления k событий на любом промежутке времени, вычисленная при любых предположениях о том, что происходило до начала рассматриваемого промежутка (сколько событий появилось, в какой последовательности), равна безусловной вероятности. Таким образом, предыстория потока не сказывается на вероятности появления событий в ближайшем будущем.

Итак, если поток обладает свойством отсутствия последействия, то имеет место взаимная независимость появлений того или иного числа событий в непересекающиеся промежутки времени.

Свойство ординарности характеризуется тем, что появление двух и более событий за малый промежуток времени практически невозможно. Другими словами, вероятность появления более одного события за малый промежуток времени пренебрежимо мала по сравнению с вероятностью появления только одного события.

Итак, если поток обладает свойством ординарности, то за бесконечно малый промежуток времени может появиться не более одного события.

Простейшим (*пуассоновским*) называют поток событий, который обладает свойствами стационарности, отсутствия последействия и ординарности.

Замечание. Часто на практике трудно установить, обладает ли поток перечисленными выше свойствами. Поэтому были найдены и другие условия, при соблюдении которых поток можно считать простейшим, или близким к простейшему. В частности, установлено, что если поток представляет собой сумму очень большого числа независимых стационарных потоков, влияние каждого из которых на всю сумму (суммарный поток) ничтожно мало, то суммарный поток (при условии его ординарности) близок к простейшему.

Интенсивностью потока λ называют среднее число событий, которые появляются в единицу времени.

Можно доказать, что если постоянная интенсивность потока известна, то вероятность появления k событий простейшего потока за время длительностью t определяется формулой Пуассона

$$P_t(k) = \frac{(\lambda t)^k \cdot e^{-\lambda t}}{k!}.$$

Эта формула отражает все свойства простейшего потока.

Действительно, из формулы видно, что вероятность появления k событий за время t, при заданной интенсивности, является функцией k и t, что характеризует свойство стационарности.

Формула не использует информации о появлении событий до начала рассматриваемого промежутка, что характеризует свойство отсутствия последействия.

Убедимся, что формула отражает свойство ординарности. Положив $k=0$ и $k=1$, найдем соответственно вероятности непоявления событий и появления одного события:

$$P_t(0) = e^{-\lambda t}, \quad P_t(1) = \lambda t e^{-\lambda t}.$$

Следовательно, вероятность появления более одного события

$$P_t(k>1) = 1 - [P_t(0) + P_t(1)] = 1 - [e^{-\lambda t} + \lambda t e^{-\lambda t}].$$

Пользуясь разложением

$$e^{-\lambda t} = 1 - \lambda t + \frac{(\lambda t)^2}{2!} - \cdots,$$

после элементарных преобразований получим

$$P_t(k>1) = \frac{(\lambda t)^2}{2} + \cdots.$$

Сравнивая $P_t(1)$ и $P_t(k>1)$, заключаем, что при малых значениях t вероятность появления более одного события пренебрежимо мала по сравнению с вероятностью наступления одного события, что характеризует свойство ординарности.

Итак, формулу Пуассона можно считать математической моделью простейшего потока событий.

Пример. Среднее число вызовов, поступающих на АТС в одну минуту, равно двум. Найти вероятности того, что за 5 минут поступит: а) 2 вызова, б) менее двух вызовов, в) не менее двух вызовов. Поток вызовов предполагается простейшим.

Решение. По условию $\lambda=2$, $t=5$, $k=2$. Воспользуемся формулой Пуассона

$$P_t(k) = \frac{(\lambda t)^k \cdot e^{-\lambda t}}{k!}.$$

а) Искомая вероятность того, что за 5 минут поступит 2 вызова

$$P_5(2) = \frac{10^2 \cdot e^{-10}}{2!} = \frac{100 \cdot 0{,}000045}{2} = 0{,}000025.$$

Это событие практически невозможно.

б) События «не поступило ни одного вызова» и «поступил один вызов» — несовместны, поэтому искомая вероятность того, что за 5 минут поступит менее двух вызовов, по теореме сложения

$$P_5(k<2) = P_5(0) + P_5(1) = e^{-10} + \frac{10 \cdot e^{-10}}{1!} = 0.000195$$

Это событие практически невозможно.

в) События «поступило менее двух вызовов» и «поступило не менее двух вызовов» противоположны, поэтому искомая вероятность того, что за 5 минут поступит не менее двух вызовов

$$P_5(k \geqslant 2) = 1 - P_5(k<2) = 1 - 0{,}000495 = 0{,}999505.$$

Это событие практически достоверно.

Задачи

1. Возможные значения случайной величины таковы: $x_1 = 2$, $x_2 = 5$, $x_3 = 8$. Известны вероятности первых двух возможных значений: $p_1 = 0{,}4$, $p_2 = 0{,}15$. Найти вероятность x_3.

Отв. $p_3 = 0{,}45$.

2. Игральная кость брошена 3 раза. Написать закон распределения числа появлений шестерки.

Отв.

X	3	2	1	0
p	$\frac{1}{216}$	$\frac{15}{216}$	$\frac{75}{216}$	$\frac{125}{216}$

3. Составить закон распределения вероятностей числа появлений события A в трех независимых испытаниях, если вероятность появления события в каждом испытании равна 0,6.

Отв.

k	0	1	2	3
p	0,064	0,288	0,432	0,216

4. Прядильщица обслуживает 1000 веретен. Вероятность обрыва нити на одном веретене в течение одной минуты равна 0,004. Найти вероятность того, что в течение одной минуты обрыв произойдет в пяти веретенах.

Отв. $P_{1000}(5) = 0{,}1562$.

5. Найти среднее число опечаток на странице рукописи, если вероятность того, что страница рукописи содержит хотя бы одну опечатку, равна 0,95. Предполагается, что число опечаток распределено по закону Пуассона.

Указание: задача сводится к отысканию параметра λ из уравнения $e^{-\lambda} = 0{,}05$.

Отв. 3.

6. Коммутатор учреждения обслуживает 100 абонентов. Вероятность того, что в течение одной минуты абонент позвонит на коммутатор, равна 0,02. Какое из двух событий вероятнее: в течение одной минуты позвонит 3 абонента; позвонит 4 абонента?

Отв. $P_{100}(3) = 0{,}18$; $P_{100}(4) = 0{,}09$.

7. Рукопись объемом в 1000 страниц машинописного текста содержит 1000 опечаток. Найти вероятность того, что наудачу взятая страница содержит: а) хотя бы одну опечатку, б) ровно две опечатки, в) не менее двух опечаток. Предполагается, что число опечаток распределено по закону Пуассона.

Отв. а) $P = 1 - e^{-1} = 0{,}6321$; б) $P_{1000}(2) = 0{,}18395$; в) $P = 0{,}2642$.

8. Среднее число вызовов, поступающих на АТС в одну минуту, равно пяти. Найти вероятность того, что за 2 минуты поступит: а) два вызова, б) менее двух вызовов, в) не менее двух вызовов.

Указание: $e^{-10} = 0{,}000045$.

Отв. а) 0,000025; б) 0,000495; в) 0,999505.

Глава седьмая

МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Числовые характеристики дискретных случайных величин

Мы уже знаем, что закон распределения полностью характеризует случайную величину. Однако часто закон распределения неизвестен и приходится ограничиваться меньшими сведениями. Иногда даже выгоднее пользоваться числами, которые описывают случайную величину суммарно; такие числа называют *числовыми характеристиками случайной величины*. К числу важных числовых характеристик относится математическое ожидание.

Математическое ожидание, как будет показано далее, приближенно равно среднему значению случайной величины.

Для решения многих задач достаточно знать математическое ожидание. Например, если известно, что математическое ожидание числа выбиваемых очков у первого стрелка больше, чем у второго, то первый стрелок в среднем выбивает больше очков, чем второй и, следовательно, стреляет лучше второго.

Хотя математическое ожидание дает о случайной величине значительно меньше сведений, чем закон ее распределения,— для решения задач, подобных приведенной и многих других, знание математического ожидания оказывается достаточным.

§ 2. Математическое ожидание дискретной случайной величины

Математическим ожиданием дискретной случайной величины называют сумму произведений всех ее возможных значений на их вероятности.

Пусть случайная величина X может принимать только значения $x_1, x_2, \ldots, x_n$, вероятности которых соответственно равны $p_1, p_2, \ldots, p_n$. Тогда математическое ожидание $M(X)$ случайной величины X определяется равенством

$$M(X)=x_1p_1+x_2p_2+\ldots+x_np_n.$$

З а м е ч а н и е. Из определения следует, что математическое ожидание дискретной случайной величины есть н е с л у ч а й н а я (постоянная) величина. Рекомендуем запомнить это утверждение, так как далее оно используется многократно. В дальнейшем читатель узнает, что математическое ожидание непрерывной случайной величины также есть постоянная величина.

Пример 1. Найти математическое ожидание случайной величины X, зная закон ее распределения:

X	3	5	2
p	0,1	0,6	0,3.

Р е ш е н и е. Искомое математическое ожидание равно сумме произведений всех возможных значений случайной величины на их вероятности:

$$M(X)=3\cdot 0,1+5\cdot 0,6+2\cdot 0,3=3,9.$$

Пример 2. Найти математическое ожидание числа появлений события A в одном испытании, если вероятность события A равна p.

Р е ш е н и е. Случайная величина X — число появлений события A в одном испытании может принимать только два значения: $x_1=1$ (событие A наступило) с вероятностью p и $x_2=0$ (событие A не наступило) с вероятностью $q=1-p$. Искомое математическое ожидание

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

Итак, м а т е м а т и ч е с к о е о ж и д а н и е ч и с л а п о я в л е н и й с о б ы т и я в о д н о м и с п ы т а н и и р а в н о в е р о я т н о с т и э т о г о с о б ы т и я. Этот результат будет использован ниже.

…роятностный смысл математического ожидания

…усть произведено n испытаний, в которых случайная …чина X приняла m_1 раз значение x_1, m_2 раз значение …, m_k раз значение x_k, причем $m_1+m_2+\ldots+m_k=n$. …ма всех значений, принятых X, равна

$$x_1m_1+x_2m_2+\ldots+x_km_k.$$

…м среднее арифметическое $\overline{X}$ всех значений, приня…случайной величиной, для чего разделим найденную …мму на общее число испытаний:

$$\overline{X}=\frac{x_1m_1+x_2m_2+\cdots+x_km_k}{n},$$

…ли

$$\overline{X}=x_1\cdot\frac{m_1}{n}+x_2\cdot\frac{m_2}{n}+\cdots+x_k\frac{m_k}{n}. \qquad (*)$$

…аметив, что отношение $\frac{m_1}{n}$ есть относительная частота … значения x_1, $\frac{m_2}{n}$ есть относительная частота W_2 значения x_2 и т. д., запишем соотношение (*) так:

$$\overline{X}=x_1W_1+x_2W_2+\ldots+x_kW_k. \qquad (**)$$

Допустим, что число испытаний достаточно велико. …а относительная частота приближенно равна вероят…ти появления события (это будет доказано в гл. IX, …)

$$W_1\simeq p_1,\ W_2\simeq p_2,\ \ldots,\ W_k\simeq p_k.$$

…аменив в соотношении (**) относительные частоты …ответствующими вероятностями, получим

$$\overline{X}\simeq x_1p_1+x_2p_2+\ldots+x_kp_k.$$

…равая часть этого приближенного равенства есть $M(X)$. …так,

$$\overline{X}\simeq M(X).$$

Вероятностный смысл полученного результата таков: …м а т и ч е с к о е о ж и д а н и е п р и б л и …н н о р а в н о (тем точнее, чем больше число испы…ний) с р е д н е м у а р и ф м е т и ч е с к о м у н а…д а е м ы х з н а ч е н и й с л у ч а й н о й в е…л и ч и н ы.

Замечание 1. Легко сообразить, что математическое ожидание больше наименьшего и меньше наибольшего возможных значений. Другими словами, на числовой оси возможные значения расположены слева и справа от математического ожидания. В этом смысле математическое ожидание характеризует **расположение распределения** и поэтому его часто называют *центром распределения*.

Этот термин заимствован из механики: если массы $p_1, p_2, \ldots, p_n$ расположены в точках с абсциссами $x_1, x_2, \ldots x_n$, причем $\sum p_i = 1$, то абсцисса центра тяжести

$$x_c = \frac{\sum x_i p_i}{p_i}$$

Учитывая, что $\sum x_i p_i = M(X)$ и $\sum p_i = 1$, получим

$$M(X) = x_c.$$

Итак, математическое ожидание есть абсцисса центра тяжести системы материальных точек, абсциссы которых равны возможным значениям случайной величины, а массы — их вероятностям.

Замечание 2. Происхождение термина «математическое ожидание» связано с начальным периодом возникновения теории вероятностей (XVI—XVII вв.), когда область ее применения ограничивалась азартными играми. Игрока интересовало среднее значение ожидаемого выигрыша или иными словами — математическое ожидание выигрыша

§ 4. Свойства математического ожидания

Свойство 1. *Математическое ожидание постоянной величины равно самой постоянной:*

$$M(C) = C.$$

Доказательство. Будем рассматривать постоянную C как дискретную случайную величину, которая имеет одно возможное значение C и принимает его с вероятностью $p = 1$. Следовательно,

$$M(C) = C \cdot 1 = C.$$

Замечание 1 Определим *произведение постоянной величины C на дискретную случайную величину X* как дискретную случайную величину CX, возможные значения которой равны произведениям постоянной C на возможные значения X; вероятности возможных значений CX равны вероятностям соответствующих возможных значений X Например, если вероятность возможного значения x_1 равна p_1, то вероятность того, что величина CX примет значение Cx_1 также равна p_1

Свойство 2. *Постоянный множитель можно выносить за знак математического ожидания:*

$$M(CX) = C \cdot M(X)$$

Доказательство. Пусть случайная величина X задана законом распределения вероятностей:

X	x_1	$x_2 \ldots$	x_n
p	p_1	$p_2 \ldots$	p_n.

Учитывая замечание 1, напишем закон распределения случайной величины CX:

CX	Cx_1	Cx_2	...	Cx_n
p	p_1	p_2	...	p_n.

Математическое ожидание случайной величины CX

$$M(CX) = Cx_1p_1 + Cx_2p_2 + \ldots + Cx_np_n = \\ = C(x_1p_1 + x_2p_2 + \ldots + x_np_n) = CM(X).$$

Итак,

$$M(CX) = CM(X).$$

Замечание 2. Прежде, чем перейти к следующему свойству, укажем, что две случайные величины называют *независимыми* если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. Несколько случайных величин называют *взаимно независимыми*, если законы распределения любого числа из них не зависят от того, какие возможные значения приняли остальные величины.

Замечание 3. Определим *произведение независимых случайных величин X и Y* как случайную величину XY, возможные значения которой равны произведениям каждого возможного значения X на каждое возможное значение Y; вероятности возможных значений произведения XY равны произведениям вероятностей возможных значений сомножителей. Например, если вероятность возможного значения x_1 равна p_1, вероятность возможного значения y_1 равна g_1, то вероятность возможного значения x_1y_1 равна p_1g_1.

Свойство 3. *Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:*

$$M(XY) = M(X)M(Y).$$

Доказательство. Пусть независимые случайные величины X и Y заданы своими законами распределения вероятностей*:

X	x_1	x_2	Y	y_1	y_2
P	p_1	p_2	g	g_1	g_2.

---

* Мы ограничились малым числом возможных значений, чтобы упростить выкладки. В общем случае доказательство аналогичное.

Составим все значения, которые может принимать случайная величина XY, для чего перемножим все возможные значения X на каждое возможное значение Y; в итоге получим: x_1y_1, x_2y_1, x_1y_2 и x_2y_2.

Учитывая замечание 3, напишем «закон распределения» произведения XY:

X	x_1y_1	x_2y_1	x_1y_2	x_2y_2
P	p_1g_1	p_2g_1	p_1g_2	p_2g_2.

Математическое ожидание равно сумме произведений всех возможных значений на их вероятности:

$$M(XY)=x_1y_1\cdot p_1g_1+x_2y_1\cdot p_2g_1+x_1y_2\cdot p_1g_2+x_2y_2\cdot p_2g_2,$$

или

$$M(XY)=y_1g_1(x_1p_1+x_2p_2)+y_2g_2(x_1p_1+x_2p_2)=$$
$$=(x_1p_1+x_2p_2)(y_1g_1+y_2g_2)=M(X)\cdot M(Y).$$

Итак, $M(XY)=M(X)\cdot M(Y)$.

Следствие. *Математическое ожидание произведения нескольких взаимно независимых случайных величин равно произведению их математических ожиданий.*

Например, для трех случайных величин имеем:

$$M(XYZ)=M(XY\cdot Z)=M(XY)M(Z)=M(X)M(Y)M(Z).$$

Для произвольного числа случайных величин доказательство проводится методом математической индукции.

Пример 1. Независимые случайные величины X и Y заданы следующими законами распределения:

X	5	2	4		Y	7	9
p	0,6	0,1	0,3		p	0,8	0,2.

Найти математическое ожидание случайной величины XY.

Решение. Найдем математические ожидания каждой из данных величин:

$$M(X)=5\cdot 0,6+2\cdot 0,1+4\cdot 0,3=4,4;$$
$$M(Y)=7\cdot 0,8+9\cdot 0,2=7,4.$$

Так как случайные величины X и Y независимые, то искомое математическое ожидание равно:

$$M(XY)=M(X)M(Y)=4,4\cdot 7,4=32,56.$$

Замечание 4. Определим *сумму случайных величин X и Y* как случайную величину $X+Y$, возможные значения которой равны суммам каждого возможного значения X с каждым возможным значением Y; вероятности возможных значений $X+Y$ для независимых величин X и Y равны произведениям вероятностей слагаемых; для зависимых величин — произведениям вероятности одного слагаемого на условную вероятность второго.

Следующее ниже свойство справедливо как для независимых, так и для зависимых случайных величин.

Свойство 4. *Математическое ожидание суммы двух случайных величин равно сумме математических ожиданий слагаемых:*

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

Доказательство. Пусть случайные величины X и Y заданы следующими законами распределения*:

X	x_1	x_2		Y	y_1	y_2
p	p_1	p_2		g	g_1	g_2.

Составим все возможные значения величины $X+Y$, для чего к каждому возможному значению X прибавим каждое возможное значение Y; получим x_1+y_1, x_1+y_2, x_2+y_1 и x_2+y_2. Обозначим вероятности этих значений соответственно через p_{11}, p_{12}, p_{21} и p_{22}.

Математическое ожидание величины $X+Y$ равно сумме произведений возможных значений на их вероятности:

$$M(X+Y)=(x_1+y_1)p_{11}+(x_1+y_2)p_{12}+(x_2+y_1)p_{21}+$$
$$+(x_2+y_2)p_{22},$$

или

$$M(X+Y)=x_1(p_{11}+p_{12})+x_2(p_{21}+p_{22})+y_1(p_{11}+p_{21})+$$
$$+y_2(p_{12}+p_{22}). \qquad (*)$$

Докажем, что $p_{11}+p_{12}=p_1$. Событие, состоящее в том что X примет значение x_1 (вероятность этого события равна p_1), влечет за собой событие, которое состоит в том, что $X+Y$ примет значение x_1+y_1 или x_1+y_2 (вероятность этого события по теореме сложения равна $p_{11}+p_{12}$) и обратно. Отсюда и следует, что $p_{11}+p_{12}=p_1$.

Аналогично доказываются равенства

$$p_{21}+p_{22}=p_2,\ \ p_{11}+p_{21}=g_1 \text{ и } p_{12}+p_{22}=g_2.$$

* Чтобы упростить вывод, мы ограничились лишь двумя возможными значениями каждой из величин. В общем случае доказательство аналогичное.

Подставляя правые части этих равенств в соотношение (*), получим:

$$M(X+Y)=(x_1p_1+x_2p_2)+(y_1g_1+y_2g_2),$$

или окончательно

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

Следствие. *Математическое ожидание суммы нескольких случайных величин равно сумме математических ожиданий слагаемых.*

Например, для трех слагаемых величин имеем:

$$M(X+Y+Z)=M[(X+Y)+Z]=$$
$$=M(X+Y)+M(Z)=M(X)+M(Y)+M(Z).$$

Для произвольного числа слагаемых величин доказательство проводится методом математической индукции.

Пример 1. Производится 3 выстрела с вероятностями попадания в цель, равными $p_1=0{,}4$; $p_2=0{,}3$ и $p_3=0{,}6$. Найти математическое ожидание общего числа попаданий.

Решение. Число попаданий при первом выстреле есть случайная величина X_1, которая может принимать только два значения: 1 (попадание) с вероятностью $p_1=0{,}4$ и 0 (промах) с вероятностью $q=1-0{,}4=0{,}6$.

Математическое ожидание числа попаданий при первом выстреле равно вероятности попадания (см. пример 2, стр. 68), т. е. $M(X_1)=0{,}4$.

Аналогично найдем математические ожидания числа попаданий при втором и третьем выстрелах:

$$M(X_2)=0{,}3,\quad M(X_3)=0{,}6.$$

Общее число попаданий есть также случайная величина, состоящая из суммы попаданий в каждом из трех выстрелов:

$$X=X_1+X_2+X_3.$$

Искомое математическое ожидание находим по теореме о математическом ожидании суммы:

$$M(X)=M(X_1+X_2+X_3)=M(X_1)+M(X_2)+M(X_3)=$$
$$=0{,}4+0{,}3+0{,}6=1{,}3 \text{ (попаданий).}$$

Пример 2. Найти математическое ожидание суммы числа очков, которые могут выпасть при бросании двух игральных костей.

Решение. Обозначим число очков, которое может выпасть на первой кости, через X и на второй — через Y. Возможные значения этих величин одинаковы и равны 1, 2, 3, 4, 5 и 6, причем вероятность каждого из этих значений равна $\frac{1}{6}$.

Найдем математическое ожидание числа очков, которые могут выпасть на первой кости:

$$M(X)=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+$$
$$+6\cdot\frac{1}{6}=\frac{7}{2}.$$

Очевидно, что и $M(Y)=\frac{7}{2}$.

Искомое математическое ожидание

$$M(X+Y)=M(X)+M(Y)=\frac{7}{2}+\frac{7}{2}=7.$$

§ 5. Математическое ожидание числа появлений события в независимых испытаниях

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A постоянна и равна p. Чему равно среднее число появлений события A в этих испытаниях? Ответ на этот вопрос дает следующая теорема.

Теорема. *Математическое ожидание $M(X)$ числа появлений события A в n независимых испытаниях равно произведению числа испытаний на вероятность появления события в каждом испытании:*

$$M(X)=np.$$

Доказательство. Будем рассматривать в качестве случайной величины X — число наступлений события A в n независимых испытаниях.

Очевидно, общее число X появлений события A в этих испытаниях складывается из чисел появлений события в отдельных испытаниях. Поэтому, если X_1 — число появлений события в первом испытании, X_2 — во втором, ..., X_n — в n-м, то общее число появлений события $X=X_1+X_2+\ldots+X_n$.

По третьему свойству математического ожидания имеем

$$M(X)=M(X_1)+M(X_2)+\ldots+M(X_n). \qquad (*)$$

Каждое из слагаемых правой части равенства есть математическое ожидание числа появлений события в одном испытании: $M(X_1)$ — в первом, $M(X_2)$ — во втором и т. д. Так как математическое ожидание числа появлений события в одном испытании равно вероятности события (§ 2, пример 2), то $M(X_1)=M(X_2)=M(X_n)=p$. Подставляя в правую часть равенства (*) вместо каждого слагаемого p, получим

$$M(X) = np. \qquad (**)$$

Замечание. Так как величина X распределена по биномиальному закону, то доказанную теорему можно сформулировать и так: *математическое ожидание биномиального распределения с параметрами n и p равно произведению np.*

Пример. Вероятность попадания в цель при стрельбе из орудия $p=0{,}6$. Найти математическое ожидание общего числа попаданий, если будет произведено 10 выстрелов.

Решение. Попадание при каждом выстреле не зависит от исходов других выстрелов, поэтому рассматриваемые события независимы и, следовательно, искомое математическое ожидание

$$M(X)=np=10\cdot 0{,}6=6 \text{ (попаданий).}$$

Задачи

1. Найти математическое ожидание дискретной случайной величины, зная закон ее распределения:

X	6	3	1
p	0,2	0,3	0,5.

Отв. 2,6.

2. Производится 4 выстрела с вероятностями попадания в цель $p_1 = 0{,}6$, $p_2 = 0{,}4$, $p_3 = 0{,}5$ и $p_4 = 0{,}7$. Найти математическое ожидание общего числа попаданий.

Отв. 2,2 попадания

3. Дискретные независимые случайные величины заданы законами распределения:

X	1	2
p	0,2	0,8

Y	0,5	1
p	0,3	0,7.

Найти математическое ожидание произведения XY двумя способами: 1) составив закон распределения XY; 2) пользуясь свойством 3.

Отв. 1,53.

4. Дискретные случайные величины X и Y заданы законами распределения, указанными в задаче 3. Найти математическое ожидание суммы $X+Y$ двумя способами: 1) составив закон распределения $X+Y$; 2) пользуясь свойством 4.

Отв. 2,65.

5. Вероятность отказа детали за время испытания на надежность равна 0,2. Найти математическое ожидание числа отказавших деталей, если испытанию будут подвергнуты 10 деталей.

Отв. 2 детали.

6. Найти математическое ожидание произведения числа очков, которые могут выпасть при одном бросании двух игральных костей.

Отв. 12,25 очков.

7. Найти математическое ожидание числа лотерейных билетов, на которые выпадут выигрыши, если приобретено 20 билетов, причем вероятность выигрыша по одному билету равна 0,3.

Отв. 6 билетов.

Глава восьмая

ДИСПЕРСИЯ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Целесообразность введения числовой характеристики рассеяния случайной величины

Легко указать такие случайные величины, которые имеют одинаковые математические ожидания, но различные возможные значения.

Рассмотрим, например, дискретные случайные величины X и Y, заданные следующими законами распределения:

X	—0,01	0,01
p	0,5	0,5

Y	—100	100
p	0,5	0,5.

Найдем математические ожидания этих величин:

$$M(X)=-0{,}01\cdot 0{,}5+0{,}01\cdot 0{,}5=0,$$
$$M(Y)=-100\cdot 0{,}5+100\cdot 0{,}5=0.$$

Здесь математические ожидания обеих величин одинаковы, а возможные значения различны, причем X имеет возможные значения, близкие к математическому ожиданию, а Y — далекие от своего математического ожидания. Таким образом, зная лишь математическое ожидание случайной величины, еще нельзя судить ни о том, какие возможные значения она может принимать, ни о том, как они рассеяны вокруг математического ожидания. Другими словами, математическое ожидание полностью случайную величину не характеризует.

По этой причине, наряду с математическим ожиданием, вводят и другие числовые характеристики. Так например, для того, чтобы оценить, как рассеяны возможные значения случайной величины вокруг ее математического ожидания, пользуются, в частности, числовой характеристикой, которую называют дисперсией.

Прежде чем перейти к определению и свойствам дисперсии, введем понятие отклонения случайной величины от ее математического ожидания.

§ 2. Отклонение случайной величины от ее математического ожидания

Пусть X — случайная величина и $M(X)$ — ее математическое ожидание. Рассмотрим в качестве новой случайной величины разность $X-M(X)$.

Отклонением называют разность между случайной величиной ее математическим ожиданиям.

Пусть закон распределения X известен:

$$\begin{array}{ccccc} X & x_1 & x_2 & \ldots & x_n \\ p & p_1 & p_2 & \ldots & p_n. \end{array}$$

Напишем закон распределения отклонения. Для того чтобы отклонение приняло значение $x_1-M(X)$ достаточно, чтобы случайная величина приняла значение x_1. Вероятность же этого события равна p_1; следовательно, и вероятность того, что отклонение примет значение $x_1-M(X)$, также равна p_1. Аналогично обстоит дело и для остальных возможных значений отклонения.

Таким образом, отклонение имеет следующий закон распределения

$$\begin{array}{ccccc} X-M(X) & x_1-M(X) & x_2-M(X) & \ldots & X_n-M(X) \\ p & p_1 & p_2 & \ldots & p_n. \end{array}$$

Приведем важное свойство отклонения, которое будет использовано далее.

Теорема. *Математическое ожидание отклонения равно нулю:*

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

Доказательство. Пользуясь свойствами математического ожидания (математическое ожидание разности равно разности математических ожиданий, математическое ожидание постоянной равно самой постоянной) и приняв во внимание, что $M(X)$ есть постоянная величина, имеем:

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

§ 3. Дисперсия дискретной случайной величины

На практике часто требуется оценить рассеяние возможных значений случайной величины вокруг ее среднего значения. Например, в артиллерии важно знать, насколько кучно лягут снаряды вблизи цели, которая должна быть поражена.

На первый взгляд может показаться, что для оценки рассеяния проще всего вычислить все возможные значения отклонения случайной величины и затем найти их среднее значение. Однако такой путь ничего не даст, так как среднее значение отклонения, т. е. $M[X-M(X)]$, для любой случайной величины равно нулю. Это свойство уже было доказано в предыдущем параграфе и объясняется тем, что одни возможные отклонения положительны, а другие отрицательны; в результате их взаимного погашения среднее значение отклонения равно нулю.

Эти соображения говорят о целесообразности заменить возможные отклонения их абсолютными значениями или их квадратами. Так и поступают на деле. Правда, в случае, когда возможные отклонения заменяют их абсолютными значениями, приходится оперировать с абсолютными величинами, что приводит иногда к серьезным затруднениям. Поэтому чаще всего идут по другому пути, т. е. вычисляют среднее значение квадрата отклонения, которое и называют дисперсией.

Дисперсией (рассеянием) дискретной случайной величины называют математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

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

Пусть случайная величина задана законом распределения

$$\begin{array}{ccccc} X & x_1 & x_2 & \ldots & x_n \\ p & p_1 & p_2 & \ldots & p_n. \end{array}$$

Тогда квадрат отклонения имеет следующий закон распределения:

$$\begin{array}{lllll} [X-M(X)]^2 & [x_1-M(X)]^2 & [x_2-M(X)]^2 & \ldots & [x_n-M(X)]^2 \\ p & p_1 & p_2 & \ldots & p_n. \end{array}$$

По определению дисперсия равна

$$D(X)=M[X-M(X)]^2=[x_1-M(X)]^2\cdot p_1+[x_2-M(X)]^2\times$$
$$\times p_2+ \ \ldots \ +[x_n-M(X)]^2\cdot p_n.$$

Таким образом, для того, чтобы найти дисперсию, достаточно вычислить сумму произведений возможных значений квадрата отклонения на их вероятности.

Замечание. Из определения следует, что дисперсия дискретной случайной величины есть неслучайная (постоянная) величина. В дальнейшем читатель узнает, что дисперсия непрерывной случайной величины также есть постоянная величина.

Пример. Найти дисперсию случайной величины X, которая задана следующим законом распределения:

X	1	2	5
p	0,3	0,5	0,2.

Решение. Найдем математическое ожидание

$$M(X)=1\cdot0,3+2\cdot0,5+5\cdot0,2=2,3.$$

Найдем все возможные значения квадрата отклонения

$$[x_1-M(X)]^2=(1-2,3)^2=1,69;$$
$$[x_2-M(X)]^2=(2-2,3)^2=0,09;$$
$$[x_3-M(X)]^2=(5-2,3)^2=7,29.$$

Напишем закон распределения квадрата отклонения:

$[X-M(X)]^2$	1,69	0,09	7,29
p	0,3	0,5	0,2.

По определению дисперсия

$$D(X)=1,69\cdot0,3+0,09\cdot0,5+7,29\cdot0,2=2,01.$$

Мы видим, что вычисление, основанное на определении дисперсии, оказалось относительно громоздким. Далее будет указана формула, которая приводит к цели значительно быстрее.

4. Формула для вычисления дисперсии

Для вычисления дисперсии часто бывает удобно пользоваться следующей теоремой.

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

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

Доказательство. Математическое ожидание $M(X)$ есть постоянная величина, следовательно, $2M(X)$ и $M^2(X)$ есть также постоянные величины. Приняв это во внимание и пользуясь свойствами математического ожидания (постоянный множитель можно вынести за знак математического ожидания, математическое ожидание суммы равно сумме математических ожиданий слагаемых), упростим формулу, выражающую определение дисперсии:

$$D(X)=M[X-M(X)]^2=M[X^2-2XM(X)+M^2(X)]=$$
$$=M(X^2)-2M(X)\cdot M(X)+M^2(X)=$$
$$=M(X^2)-2M^2(X)+M^2(X)=M(X^2)-M^2(X)$$

Итак,

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

Квадратная скобка введена в запись формулы для удобства ее запоминания.

Пример 1. Найти дисперсию случайной величины X, которая задана следующим законом распределения:

X	2	3	5
p	0,1	0,6	0,3.

Решение. Найдем математическое ожидание $M(X)$:

$$M(X)=2\cdot0,1+3\cdot0,6+5\cdot0,3=3,5.$$

Напишем закон распределения случайной величины X^2:

X^2	4	9	25
p	0,1	0,6	0,3.

Найдем математическое ожидание $M(X^2)$:

$$M(X^2)=4\cdot0,1+9\cdot0,6+25\cdot0,3=13,3.$$

Искомая дисперсия

$$D(X)=M(X^2)-[M(X)]^2=13,3-(3,5)^2=1,05.$$

Замечание. Казалось бы, что если X и Y имеют одинаковые возможные значения и одно и то же математическое ожидание, то и дисперсии этих величин равны (ведь возможные значения обеих величин одинаково рассеяны вокруг своих математических ожиданий!). Однако, в общем случае, это не так. Дело в том, что одинаковые возможные значения рассматриваемых величин имеют, вообще говоря, различные вероятности, а величина дисперсии определяется не только самими возможными значениями, но и их вероятностями. Например, если вероятности «далеких» от математического ожидания возможных значений X больше, чем вероятности этих же значений Y и вероятности «близких» значений X меньше, чем вероятности тех же значений Y, то, очевидно, дисперсия X больше дисперсии Y.

Приведем иллюстрирующий пример.

Пример 2. Сравнить дисперсии случайных величин, заданных законами распределения:

X	-1	1	2	3
p	0,48	0,01	0,09	0,42

Y	-1	1	2	3
p	0,19	0,51	0,25	0,05.

Решение. Легко убедиться, что

$$M(X)=M(Y)=0,97;$$
$$D(X)\simeq 3,69, \quad D(Y)\simeq 1,21.$$

Таким образом, возможные значения и математические ожидания X и Y одинаковы, а дисперсии различны, причем $D(X)>D(Y)$.

Этот результат можно было предвидеть без вычислений, глядя лишь на законы распределений.

§ 5. Свойства дисперсии

Свойство 1. *Дисперсия постоянной величины C равна нулю:*

$$D(C)=0.$$

Доказательство. По определению дисперсии имеем

$$D(C)=M\{[C-M(C)]^2\}.$$

Пользуясь первым свойством математического ожидания (математическое ожидание постоянной равно самой постоянной), получим

$$D(C)=M[(C-C)^2]=M(0)=0.$$

Итак,

$$D(C)=0.$$

Свойство становится ясным, если учесть, что постоянная величина сохраняет одно и то же значение и рассеяния, конечно, не имеет.

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

$$D(CX)=C^2D(X).$$

Доказательство. По определению дисперсии имеем

$$D(CX)=M\{[CX-M(CX)]^2\}.$$

Пользуясь вторым свойством математического ожидания (постоянный множитель можно выносить за знак математического ожидания), получим

$$D(CX)=M\{[CX-CM(X)]^2\}=M\{C^2[X-M(X)]^2\}=$$
$$=C^2M\{[X-M(X)]^2\}=C^2D(X).$$

Итак,

$$D(CX)=C^2D(X).$$

Свойство становится ясным, если принять во внимание, что при $|C|>1$ величина CX имеет возможные значения (по абсолютной величине) большие, чем величина X. Отсюда следует, что эти значения рассеяны вокруг математического ожидания $M(CX)$ больше, чем возможные значения X вокруг $M(X)$, т. е. $D(CX)>D(X)$. Напротив, если $0<|C|<1$, то $D(CX)<D(X)$.

Свойство 3. *Дисперсия суммы двух независимых случайных величин равна сумме дисперсий этих величин:*

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

Доказательство. По формуле для вычисления дисперсии имеем

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

Раскрыв скобки и пользуясь свойствами математического ожидания суммы нескольких величин и произведения двух независимых случайных величин, получим

$$D(X+Y)=M[X^2+2XY+Y^2]-[M(X)+M(Y)]^2=$$
$$=M(X^2)+2M(X)\cdot M(Y)+M(Y^2)-M^2(X)-$$

$$-2M(X)\cdot M(Y)-M^2(Y)=[M(X^2)-[M(X)]^2]+$$
$$+[M(Y^2)-[M(Y)]^2]=D(X)+D(Y).$$

Итак,

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

Следствие 1. *Дисперсия суммы нескольких взаимно независимых случайных величин равна сумме дисперсий этих величин.*

Например, для трех слагаемых имеем

$$D(X+Y+Z)=D[X+(Y+Z)]=D(X)+D(Y+Z)=$$
$$=D(X)+D(Y)+D(Z).$$

Для произвольного числа слагаемых доказательство проводится методом математической индукции.

Следствие 2. *Дисперсия суммы постоянной величины и случайной равна дисперсии случайной величины:*

$$D(C+X)=D(X).$$

Доказательство. Величины C и X независимы, поэтому по третьему свойству

$$D(C+X)=D(C)+D(X).$$

В силу первого свойства $D(C)=0$. Следовательно,

$$D(C+X)=D(X).$$

Свойство становится понятным, если учесть, что величины X и $X+C$ отличаются лишь началом отсчета и, значит, рассеяны вокруг своих математических ожиданий одинаково.

Свойство 4. Дисперсия разности двух независимых случайных величин равна сумме их дисперсий:

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

Доказательство. В силу третьего свойства

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

По второму свойству

$$D(X-Y)=D(X)+(-1)^2D(Y),$$

или

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

6. Дисперсия числа появлений события в независимых испытаниях

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A постоянна. Чему равна дисперсия числа появлений события в этих испытаниях? Ответ на этот вопрос дает следующая теорема.

Теорема. *Дисперсия числа появлений события A в n независимых испытаниях, в каждом из которых вероятность p появления события постоянна, равна произведению числа испытаний на вероятности появления и непоявления события в одном испытании:*

$$D(X)=npq.$$

Доказательство. Рассмотрим случайную величину X — число появлений события A в n независимых испытаниях. Очевидно, общее число появлений события в этих испытаниях равно сумме появлений события в отдельных испытаниях:

$$X=X_1+X_2+\ldots+X_n,$$

где X_1— число наступлений события в первом испытании, X_2— во втором, ..., X_n— в n-м.

Величины X_1, X_2, ..., X_n взаимно независимы, так как исход каждого испытания не зависит от исходов остальных, поэтому мы вправе воспользоваться следствием 1 (§ 5):

$$D(X)=D(X_1)+D(X_2)+\ldots+D(X_n). \qquad (*)$$

Вычислим дисперсию X_1 по формуле

$$D(X_1)=M(X_1^2)-[M(X_1)]^2. \qquad (**)$$

Величина X_1 есть число появлений события A в первом испытании, поэтому (гл. VII, § 2, пример 2) $M(X_1)=p$.

Найдем математическое ожидание величины X_1^2, которая может принимать только два значения, а именно 1^2 с вероятностью p и 0^2 с вероятностью q:

$$M(X_1^2)=1^2\cdot p+0^2\cdot q=p.$$

Подставляя найденные результаты в соотношение (**), имеем

$$D(X_1)=p-p^2=p(1-p)=pq.$$

Очевидно, дисперсия каждой из остальных случайных величин также равна p. Заменив каждое слагаемое правой части (*) через p, окончательно получим

$$D(X) = npq.$$

Замечание. Так как величина X распределена по биномиальному закону, то доказанную теорему можно сформулировать и так: *дисперсия биномиального распределения с параметрами n и p равна произведению npq.*

Пример. Производятся 10 независимых испытаний, в каждом из которых вероятность появления события равна 0,6. Найти дисперсию случайной величины X — числа появлений события в этих испытаниях.

Решение. По условию $n = 10$; $p = 0,6$. Очевидно, вероятность непоявления события

$$q = 1 - 0,6 = 0,4.$$

Искомая дисперсия

$$D(X) = npq = 10 \cdot 0,6 \cdot 0,4 = 2,4.$$

§ 7. Среднее квадратическое отклонение

Для оценки рассеяния возможных значений случайной величины вокруг ее среднего значения кроме дисперсии служат и некоторые другие характеристики. К их числу относится среднее квадратическое отклонение.

Средним квадратическим отклонением случайной величины X называют квадратный корень из дисперсии:

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

Легко показать, что дисперсия имеет размерность, равную квадрату размерности случайной величины. Так как среднее квадратическое отклонение равно квадратному корню из дисперсии, то размерность $\sigma(X)$ совпадает с размерностью X. Поэтому в тех случаях, когда желательно, чтобы оценка рассеяния имела размерность случайной величины, вычисляют среднее квадратическое отклонение, а не дисперсию. Например, если X выражается в линейных метрах, то $\sigma(X)$ будет выражаться также в линейных метрах, а $D(X)$ — в квадратных метрах.

Пример. Случайная величина X задана законом распределения

X	2	3	10
p	0,1	0,4	0,5

Найти среднее квадратическое отклонение $\sigma(X)$.

Решение. Найдем математическое ожидание X:

$$M(X) = 2 \cdot 0,1 + 3 \cdot 0,4 + 10 \cdot 0,5 = 6,4.$$

Найдем математическое ожидание X^2:

$$M(X^2) = 2^2 \cdot 0,1 + 3^2 \cdot 0,4 + 10^2 \cdot 0,5 = 54.$$

Найдем дисперсию:

$$D(X) = M(X^2) - [M(X)]^2 = 54 - 6,4^2 = 13,04.$$

Искомое среднее квадратическое отклонение

$$\sigma(X) = \sqrt{D(X)} = \sqrt{13,04} \simeq 3,61.$$

8. Среднее квадратическое отклонение суммы взаимно независимых случайных величин

Пусть известны средние квадратические отклонения нескольких взаимно независимых случайных величин. Как найти среднее квадратическое отклонение суммы этих величин? Ответ на этот вопрос дает следующая теорема.

Теорема. *Среднее квадратическое отклонение суммы конечного числа взаимно независимых случайных величин равно квадратному корню из суммы квадратов средних квадратических отклонений этих величин:*

$$\sigma(X_1 + X_2 + \ldots + X_n) = \sqrt{\sigma^2(X_1) + \sigma^2(X_2) + \ldots + \sigma^2(X_n)}.$$

Доказательство. Обозначим через X сумму рассматриваемых взаимно независимых величин:

$$X = X_1 + X_2 + \ldots + X_n.$$

Так как дисперсия суммы нескольких взаимно независимых случайных величин равна сумме дисперсий слагаемых (§ 5, следствие 1), то

$$D(X) = D(X_1) + D(X_2) + \ldots + D(X_n).$$

Отсюда

$$\sqrt{D(X)} = \sqrt{D(X_1) + D(X_2) + \ldots + D(X_n)},$$

или окончательно

$$\sigma(X)=\sqrt{\sigma^2(X_1)+\sigma^2(X_2)+\ldots+\sigma^2(X_n)}\,.$$

§ 9. Одинаково распределенные взаимно независимые случайные величины

Уже известно, что по закону распределения можно найти числовые характеристики случайной величины. Отсюда следует, что если несколько случайных величин имеют одинаковые распределения, то их числовые характеристики одинаковы.

Рассмотрим n взаимно независимых случайных величин $X_1, X_2, \ldots, X_n$, которые имеют одинаковые распределения, а следовательно, и одинаковые характеристики (математическое ожидание, дисперсию и др.). Наибольший интерес представляет изучение числовых характеристик среднего арифметического этих величин, чем мы и займемся в настоящем параграфе.

Обозначим среднее арифметическое рассматриваемых случайных величин через $\overline{X}$:

$$\overline{X}=\frac{X_1+X_2+\ldots+X_n}{n}\,.$$

Следующие ниже три положения устанавливают связь между числовыми характеристиками среднего арифметического $\overline{X}$ и соответствующими характеристиками каждой отдельной величины.

1. *Математическое ожидание среднего арифметического одинаково распределенных взаимно независимых случайных величин равно математическому ожиданию a каждой из величин:*

$$M(\overline{X})=a.$$

Доказательство. Пользуясь свойствами математического ожидания (постоянный множитель можно вынести за знак математического ожидания; математическое ожидание суммы равно сумме математических ожиданий слагаемых), имеем:

$$M(\overline{X})=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=$$

$$=\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\,.$$

Приняв во внимание, что математическое ожидание каждой из величин по условию равно a, получим

$$M(\overline{X})=\frac{na}{n}=a.$$

2. *Дисперсия среднего арифметического n одинаково распределенных взаимно независимых случайных величин в n раз меньше дисперсии D каждой из величин:*

$$D(\overline{X})=\frac{D}{n}\,. \qquad (*)$$

Доказательство. Пользуясь свойствами дисперсии (постоянный множитель можно вынести за знак дисперсии, возведя его в квадрат; дисперсия суммы независимых величин равна сумме дисперсий слагаемых), имеем

$$D(\overline{X})=D\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=$$

$$=\frac{D(X_1)+D(X_2)+\ldots+D(X_n)}{n^2}\,.$$

Приняв во внимание, что дисперсия каждой из величин по условию равна D, получим:

$$D(\overline{X})=\frac{nD}{n^2}=\frac{D}{n}\,.$$

3. *Среднее квадратическое отклонение среднего арифметического n одинаково распределенных взаимно независимых случайных величин в $\sqrt{n}$ раз меньше среднего квадратического отклонения σ каждой из величин:*

$$\sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}\,. \qquad (**)$$

Доказательство. Так как $D(\overline{X})=\frac{D}{n}$, то среднее квадратическое отклонение $\overline{X}$ равно

$$\sigma(\overline{X})=\sqrt{D(\overline{X})}=\sqrt{\frac{D}{n}}=\frac{\sqrt{D}}{\sqrt{n}}=\frac{\sigma}{\sqrt{n}}\,.$$

Общий вывод из формул (*) и (**): вспоминая, что дисперсия и среднее квадратическое отклонение служат мерами рассеяния случайной величины, заключаем, что среднее арифметическое достаточно большого числа взаимно независимых случайных величин имеет значительно меньшее рассеяние, чем каждая отдельная величина.

Поясним на примере значение этого вывода для практики.

Пример. Обычно для измерения некоторой физической величины производят несколько измерений, а затем находят среднее арифметическое полученных чисел, которое принимают за приближенное значение измеряемой величины. Предполагая, что измерения производятся в одних и тех же условиях, доказать:

а) среднее арифметическое даст результат более надежный, чем отдельные измерения;

б) с увеличением числа измерений надежность этого результата возрастает.

Решение. а) Известно, что отдельные измерения дают не одинаковые значения измеряемой величины. Результат каждого измерения зависит от многих случайных причин (изменение температуры, колебания прибора и т. п.), которые не могут быть заранее полностью учтены.

Поэтому мы вправе рассматривать возможные результаты n отдельных измерений в качестве случайных величин $X_1, X_2, \ldots, X_n$ (индекс указывает номер измерения). Эти величины имеют одинаковое распределение вероятностей (измерения производятся по одной и той же методике и теми же приборами), а следовательно, и одинаковые числовые характеристики; кроме того, они взаимно независимы (результат каждого отдельного измерения не зависит от остальных измерений).

Мы уже знаем, что среднее арифметическое таких величин имеет меньшее рассеяние, чем каждая отдельная величина. Иначе говоря, среднее арифметическое оказывается более близким к истинному значению измеряемой величины, чем результат отдельного измерения. Это и означает, что среднее арифметическое нескольких измерений дает более надежный результат, чем отдельное измерение.

б) Нам уже известно, что при возрастании числа отдельных случайных величин, рассеяние среднего арифметического убывает. Это значит, что с увеличением числа измерений среднее арифметическое нескольких измерений все менее отличается от истинного значения измеряемой величины. Таким образом, увеличивая число измерений, получают более надежный результат.

Например, если среднее квадратическое отклонение отдельного измерения $\sigma = 6$ *м*, а всего произведено $n=36$ измерений, то среднее квадратическое отклонение среднего арифметического этих измерений равно лишь 1 *м*. Действительно,

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1.$$

Мы видим, что среднее арифметическое нескольких измерений, как и следовало ожидать, оказалось более близким к истинному значению измеряемой величины, чем результат отдельного измерения.

§ 10. Понятие о моментах распределения

Рассмотрим дискретную случайную величину X, заданную законом распределения:

X	1	2	5	100
p	0,6	0,2	0,19	0,01.

Найдем математическое ожидание X:

$$M(X) = 1\cdot 0{,}6 + 2\cdot 0{,}2 + 5\cdot 0{,}19 + 100\cdot 0{,}01 = 2{,}95.$$

Напишем закон распределения X^2:

X^2	1	4	25	10 000
p	0,06	0,02	0,19	0,01.

Найдем математическое ожидание X^2:

$$M(X^2) = 1\cdot 0{,}6 + 4\cdot 0{,}2 + 25\cdot 0{,}19 + 10\,000\cdot 0{,}01 = 106{,}15.$$

Мы видим, что $M(X^2)$ значительно больше $M(X)$. Это объясняется тем, что после возведения в квадрат возможное значение величины X^2, соответствующее значению $x=100$ величины X, стало равным 10 000, т. е. значительно увеличилось; вероятность же этого значения мала (0,01).

Таким образом, переход от $M(X)$ к $M(X^2)$ позволил лучше учесть влияние на математическое ожидание того возможного значения, которое велико и имеет малую вероятность. Разумеется, если бы величина X имела несколько больших и маловероятных значений, то переход к величине X^2, а тем более к величинам X^3, X^4 и т. д., позволил бы еще больше «усилить роль» этих больших, но маловероятных возможных значений. Вот почему оказывается целесообразным рассматривать математическое ожидание целой положительной степени случайной величины (не только дискретной, но и непрерывной).

Начальным моментом порядка k случайной величины X называют математическое ожидание величины X^k:

$$\nu_k = M(X^k).$$

В частности,

$$\nu_1 = M(X).$$

$$\nu_2 = M(X^2).$$

Пользуясь этими моментами, формулу для вычисления дисперсии $D(X)=M(X^2)-[M(X)]^2$ можно записать так:

$$D(X) = \nu_2 - \nu_1^2. \qquad (*)$$

Кроме моментов случайной величины X, целесообразно рассматривать моменты отклонения $X-M(X)$.

Центральным моментом порядка k случайной величины X называют математическое ожидание величины $(X-M(X))^k$:

$$\mu_k=M[(X-M(X))^k].$$

В частности,

$$\mu_1=M[X-M(X))]=0, \qquad (**)$$

$$\mu_2=M[(X-M(X))^2]=D(X). \qquad (***)$$

Легко выводятся соотношения, связывающие начальные и центральные моменты.

Например, сравнивая (*) и (***), получим

$$\mu_2 = \nu_2 - \nu_1^2.$$

Нетрудно, исходя из определения центрального момента и пользуясь свойствами математического ожидания, получить формулы:

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3,$$

$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4.$$

Моменты более высоких порядков применяются редко.

Замечание. Моменты, рассмотренные здесь, называют *теоретическими*. В отличие от теоретических моментов, моменты, которые вычисляются по данным наблюдений, называют *эмпирическими*. Определения эмпирических моментов даны далее (гл. XVII, § 2).

Задачи

1. Известны дисперсии двух независимых случайных величин: $D(X) = 4$, $D(Y) = 3$. Найти дисперсию суммы этих величин.
Отв. 7.

2. Дисперсия случайной величины X равна 5. Найти дисперсию следующих величин: а) $X - 1$; б) $-2X$; в) $3X + 6$.
Отв. а) 5; б) 20; в) 45.

3. Случайная величина X принимает только два значения $+C$ и $-C$, каждое с вероятностью 0,5. Найти дисперсию этой величины.
Отв. C^2.

4. Найти дисперсию случайной величины, зная закон ее распределения

X	0,1	2	10	20
p	0,4	0,2	0,15	0,25.

Отв. 67,6404.

5. Случайная величина X может принимать два возможных значения: x_1 с вероятностью 0,3 и x_2 с вероятностью 0,7, причем $x_2 > x_1$. Найти x_1 и x_2, зная что $M(X) = 2,7$ и $D(X) = 0,21$.
Отв. $x_1 = 2$; $x_2 = 3$.

6. Найти дисперсию случайной величины X — числа появлений события A в двух независимых испытаниях, если $M(X) = 0,8$.
Указание. Написать биномиальный закон распределения вероятностей числа появлений события A в двух независимых испытаниях.
Отв. 0,48.

7. Испытывается устройство, состоящее из четырех независимо работающих приборов. Вероятности отказа приборов таковы: $p_1 = 0,3$; $p_2 = 0,4$; $p_3 = 0,5$; $p_4 = 0,6$. Найти математическое ожидание и дисперсию числа отказавших приборов.
Отв. 1,8; 0,94.

8. Найти дисперсию случайной величины X — числа появлений события в 100 независимых испытаниях, в каждом из которых вероятность наступления события равна 0,7.
Отв. 21.

9. Дисперсия случайной величины $D(X) = 6,25$. Найти среднее квадратическое отклонение $\sigma(X)$.
Отв. 2,5.

10. Случайная величина задана законом распределения

X	2	4	8
p	0,1	0,5	0,4.

Найти среднее квадратическое отклонение этой величины.
Отв. 2,2.

11. Дисперсия каждой из 9 одинаково распределенных взаимно независимых случайных величин равна 36. Найти дисперсию среднего арифметического этих величин.
Отв. 4.

12. Среднее квадратическое отклонение каждой из 16 одинаково распределенных взаимно независимых случайных величин равно 10. Найти среднее квадратическое отклонение среднего арифметического этих величин.

Отв. 2,5.

Глава девятая

ЗАКОН БОЛЬШИХ ЧИСЕЛ

§ 1. Предварительные замечания

Как мы знаем, нельзя заранее уверенно предвидеть, какое из возможных значений примет случайная величина в итоге испытания; это зависит от многих случайных причин, учесть которые мы не в состоянии. Казалось бы, что поскольку о каждой случайной величине мы располагаем в этом смысле весьма скромными сведениями, то вряд ли можно установить закономерности поведения и суммы достаточно большого числа случайных величин. На самом деле это не так. Оказывается, что при некоторых сравнительно широких условиях суммарное поведение достаточно большого числа случайных величин почти утрачивает случайный характер и становится закономерным.

Для практики очень важно знание условий, при выполнении которых совокупное действие очень многих случайных причин приводит к результату, почти не зависящему от случая, так как позволяет предвидеть ход явлений. Эти условия и указываются в теоремах, носящих общее название закона больших чисел. К ним относятся теоремы Чебышева и Бернулли (имеются и другие теоремы, которые здесь не рассматриваются). Теорема Чебышева является наиболее общим законом больших чисел, теорема Бернулли — простейшим.

Для доказательства этих теорем мы воспользуемся неравенством Чебышева.

§ 2. Неравенство Чебышева

Неравенство Чебышева справедливо для дискретных и непрерывных случайных величин. В целях упрощения мы ограничимся доказательством этого неравенства для дискретных величин.

Рассмотрим дискретную случайную величину X, заданную таблицей распределения:

$$\begin{array}{ccccc} X & x_1 & x_2 & \dots & x_n \\ p & p_1 & p_2 & \dots & p_n. \end{array}$$

Поставим своей задачей оценить вероятность того, что отклонение случайной величины от ее математического ожидания не превышает по абсолютной величине положительного числа ε. Если ε достаточно мало, то мы оценим, таким образом, вероятность того, что X примет значения, достаточно близкие к своему математическому ожиданию. П. Л. Чебышев доказал неравенство, позволяющее дать интересующую нас оценку.

Неравенство Чебышева. *Вероятность того, что отклонение случайной величины X от ее математического ожидания по абсолютной величине меньше положительного числа ε не меньше, чем $1-\frac{D(X)}{\varepsilon^2}$:*

$$P(|X-M(X)|<\varepsilon)\geqslant 1-\frac{D(X)}{\varepsilon^2}.$$

Доказательство. Так как события, состоящие в осуществлении неравенств $|X-M(X)|<\varepsilon$ и $|X-M(X)|\geqslant\varepsilon$ противоположны, то сумма их вероятностей равна единице, т. е.

$$P(|X-M(X)|<\varepsilon)+P(|X-M(X)|\geqslant\varepsilon)=1.$$

Отсюда интересующая нас вероятность

$$P(|X-M(X)|<\varepsilon)=1-P(|X-M(X)|\geqslant\varepsilon). \qquad (*)$$

Мы видим, что задача сведена к вычислению вероятности $P(|X-M(X)|\geqslant\varepsilon)$.

Напишем выражение дисперсии случайной величины X:

$$D(X)=[x_1-M(X)]^2p_1+[x_2-M(X)]^2p_2+\dots$$
$$\dots+[x_n-M(X)]^2p_n.$$

Очевидно, все слагаемые этой суммы не отрицательны.

Отбросим те слагаемые, у которых $|x_i-M(X)|<\varepsilon$ (для оставшихся слагаемых $|x_j-M(X)|\geqslant\varepsilon$), вследствие чего сумма может только уменьшиться. Будем считать для определенности, что отброшено k первых слагаемых (не нарушая общности, можно считать, что в таблице распределения возможные значения занумерованы именно в таком порядке). Таким образом,

$$D(X) \geqslant [x_{k+1} - M(X)]^2 p_{k+1} + [x_{k+2} - M(X)]^2 p_{k+2} + \ldots + [x_n - M(X)]^2 p_n.$$

Заметим, что обе части неравенства $|x_j - M(X)| \geqslant \varepsilon$ ($j = k+1, k+2, \ldots, n$) положительны, поэтому, возведя их в квадрат, получим равносильное неравенство $|x_j - M(X)|^2 \geqslant \varepsilon^2$. Воспользуемся этим замечанием и, заменяя в оставшейся сумме каждый из множителей $|x_j - M(X)|^2$ числом ε^2 (при этом неравенство может лишь усилиться), получим

$$D(X) \geqslant \varepsilon^2 (p_{k+1} + p_{k+2} + \ldots + p_n). \qquad (**)$$

По теореме сложения сумма вероятностей $p_{k+1} + p_{k+2} + \ldots + p_n$ есть вероятность того, что X примет одно, безразлично какое, из значений $x_{k+1}, x_{k+2}, \ldots, x_n$, а при любом из них отклонение будет удовлетворять неравенству $|x_j - M(X)| \geqslant \varepsilon$. Отсюда следует, что сумма $p_{k+1} + p_{k+2} + \ldots + p_n$ выражает вероятность

$$P(|X - M(X)| \geqslant \varepsilon)$$

Это соображение позволяет переписать неравенство (**) так:

$$D(X) \geqslant \varepsilon^2 \cdot P(|X - M(X)| \geqslant \varepsilon),$$

или

$$P(|X - M(X)| \geqslant \varepsilon) \leqslant \frac{D(X)}{\varepsilon^2}. \qquad (***)$$

Подставляя (***) в (*), окончательно получим

$$P(|X - M(X)| < \varepsilon) \geqslant 1 - \frac{D(X)}{\varepsilon^2},$$

что и требовалось доказать.

Замечание. Неравенство Чебышева имеет для практики ограниченное значение, поскольку часто дает грубую, а иногда и тривиальную (не представляющую интереса) оценку. Например, если $D(X) > \varepsilon^2$ и, следовательно, $\frac{D(X)}{\varepsilon^2} > 1$, то $1 - \frac{D(X)}{\varepsilon^2} < 0$;

таким образом, в этом случае неравенство Чебышева указывает лишь на то, что вероятность отклонения не отрицательна, а это и без того очевидно, так как любая вероятность выражается неотрицательным числом.

Теоретическое же значение неравенства Чебышева весьма велико. Ниже мы воспользуемся этим неравенством для вывода теоремы Чебышева.

§ 3. Теорема Чебышева

Теорема Чебышева. *Если $X_1, X_2, \ldots, X_n$ попарно независимые случайные величины, причем дисперсии их равномерно ограничены (не превышают постоянного числа C), то как бы мало ни было положительное число ε, вероятность неравенства*

$$\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{M(X_1) + M(X_2) + \cdots + M(X_n)}{n} \right| < \varepsilon$$

будет как угодно близка к единице, если число случайных величин достаточно велико.

Другими словами, в условиях теоремы

$$\lim_{n \to \infty} P\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{M(X_1) + M(X_2) + \cdots + M(X_n)}{n} \right| < \varepsilon \right) = 1.$$

Таким образом, теорема Чебышева утверждает, что если рассматривается достаточно большое число независимых случайных величин, имеющих ограниченные дисперсии, то почти достоверным можно считать событие, состоящее в том, что отклонение среднего арифметического случайных величин от среднего арифметического их математических ожиданий будет по абсолютной величине сколь угодно малым.

Доказательство. Введем в рассмотрение новую случайную величину — среднее арифметическое случайных величин

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Найдем математическое ожидание $\overline{X}$. Пользуясь свойствами математического ожидания (постоянный множитель можно вынести за знак математического ожидания, математическое ожидание суммы равно сумме математических ожиданий слагаемых), получим

$$M\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{M(X_1)+M(X_2)+\cdots+M(X_n)}{n}. \quad (*)$$

Применяя к величине $\overline{X}$ неравенство Чебышева, имеем

$$P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-M\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)\right|<\varepsilon\right)\geqslant$$

$$\geqslant 1-\frac{D\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)}{\varepsilon^2},$$

или, учитывая соотношение (*),

$$P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\right.\right.$$

$$\left.\left.-\frac{M(X_1)+M(X_2)+\cdots+M(X_n)}{n}\right|<\varepsilon\right)\geqslant 1-$$

$$-\frac{D\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)}{\varepsilon^2}. \quad (**)$$

Пользуясь свойствами дисперсии (постоянный множитель можно вынести за знак дисперсии, возведя его в квадрат; дисперсия суммы независимых случайных величин равна сумме дисперсий слагаемых), получим

$$D\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{D(X_1)+D(X_2)+\cdots+D(X_n)}{n^2}.$$

По условию дисперсии всех случайных величин ограничены постоянным числом C, т. е. имеют место неравенства:

$$D(X_1)\leqslant C;\quad D(X_2)\leqslant C;\ldots;\ D(X_n)\leqslant C,$$

поэтому

$$\frac{D(X_1)+D(X_2)+\cdots+D(X_n)}{n^2}\leqslant\frac{C+C+\cdots+C}{n^2}=$$

$$=\frac{nC}{n^2}=\frac{C}{n}.$$

Итак,

$$D\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)\leqslant\frac{C}{n}. \quad (***)$$

Подставляя правую часть (***) в неравенство (**) (отчего последнее может быть лишь усилено), имеем

$$P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\right.\right.$$

$$\left.\left.-\frac{M(X_1)+M(X_2)+\cdots+M(X_n)}{n}\right|<\varepsilon\right)\geqslant 1-\frac{C}{n\varepsilon^2}.$$

Отсюда переходя к пределу при $n\to\infty$, получим

$$\lim_{n\to\infty}P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\right.\right.$$

$$\left.\left.-\frac{M(X_1)+M(X_2)+\cdots+M(X_n)}{n}\right|<\varepsilon\right)\geqslant 1.$$

Наконец, учитывая, что вероятность не может превышать единицу, окончательно можем написать

$$\lim_{n\to\infty}P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\right.\right.$$

$$\left.\left.-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|<\varepsilon\right)=1.$$

Теорема доказана.

Выше, формулируя теорему Чебышева, мы предполагали, что случайные величины имеют различные математические ожидания. На практике часто бывает, что случайные величины имеют одно и то же математическое ожидание. Очевидно, что если вновь допустить, что дисперсии этих величин ограничены, то к ним будет применима теорема Чебышева.

Обозначим математическое ожидание каждой из случайных величин через a; в рассматриваемом случае среднее арифметическое математических ожиданий, как легко видеть, также равно a.

Мы можем сформулировать теорему Чебышева для рассматриваемого частного случая.

Если $X_1, X_2, \ldots, X_n$ — попарно независимые случайные величины, имеющие одно и то же математическое ожидание a, и если дисперсия этих величин равномерно ограниче-

ны, то как бы мало ни было число $\varepsilon > 0$, вероятность неравенства

$$\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon$$

будет как угодно близка к единице, если число случайных величин достаточно велико.

Другими словами, в условиях теоремы будет иметь место равенство

$$\lim_{n \to \infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon\right) = 1.$$

§ 4. Сущность теоремы Чебышева

Сущность доказанной теоремы такова: хотя отдельные независимые случайные величины могут принимать значения далекие от своих математических ожиданий,— среднее арифметическое достаточно большого числа случайных величин с большой вероятностью принимает значения близкие к определенному постоянному числу, а именно к числу $\frac{M(X_1) + M(X_2) + \ldots + M(X_n)}{n}$ (или к числу a в частном случае). Иными словами, отдельные случайные величины могут иметь значительный разброс, а их среднее арифметическое рассеянно мало.

Таким образом, нельзя уверенно предсказать, какое возможное значение примет каждая из случайных величин, но можно предвидеть какое значение примет их среднее арифметическое.

Итак, **среднее арифметическое достаточно большого числа независимых случайных величин** (дисперсии которых равномерно ограничены) **утрачивает характер случайной величины**. Объясняется это тем, что отклонения каждой из величин от своих математических ожиданий могут быть как положительными, так и отрицательными, а в среднем арифметическом они взаимно погашаются.

Теорема Чебышева справедлива не только для дискретных, но и для непрерывных случайных величин; она является ярким примером, подтверждающим справедливость учения диалектического материализма о связи между случайностью и необходимостью.

§ 5. Значение теоремы Чебышева для практики

Приведем примеры применения теоремы Чебышева к решению практических задач.

Обычно для измерения некоторой физической величины производят несколько измерений и их среднее арифметическое принимают в качестве искомого размера. При каких условиях этот способ измерения можно считать правильным? Ответ на этот вопрос дает теорема Чебышева (ее частный случай).

Действительно, рассмотрим результаты каждого измерения как случайные величины $X_1, X_2, \ldots, X_n$. К этим величинам можно применить теорему Чебышева, если: 1) они попарно независимы, 2) имеют одно и то же математическое ожидание, 3) дисперсии их равномерно ограничены.

Первое требование выполняется, если результат каждого измерения не зависит от результатов остальных.

Второе требование выполняется, если измерения произведены без систематических (одного знака) ошибок. В этом случае математические ожидания всех случайных величин одинаковы и равны истинному размеру a.

Третье требование выполняется, если прибор обеспечивает определенную точность измерений. Хотя при этом результаты отдельных измерений различны, но рассеяние их ограниченно.

Если все указанные требования выполнены, мы вправе применить к результатам измерений теорему Чебышева: при достаточно большом n вероятность неравенства

$$\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon$$

как угодно близка к единице. Другими словами, при достаточно большом числе измерений почти достоверно, что их среднее арифметическое как угодно мало отличается от истинного значения измеряемой величины.

Итак, теорема Чебышева указывает условия, при которых описанный способ измерения может быть применен.

Однако ошибочно думать, что увеличивая число измерений можно достичь сколь угодно большой точности. Дело в том, что сам прибор дает показания лишь с точностью $\pm\alpha$, поэтому каждый из результатов измерений, а следовательно и их среднее арифметическое, будут получены лишь с точностью, не превышающей точности прибора.

На теореме Чебышева основан широко применяемый в статистике выборочный метод, суть которого состоит в том, что по сравнительно небольшой случайной выборке судят о всей совокупности (генеральной совокупности) исследуемых объектов. Например, о качестве кипы хлопка заключают по небольшому пучку, состоящему из волокон, наудачу отобранных из разных мест кипы. Хотя число волокон в пучке значительно меньше, чем в кипе, сам пучок содержит достаточно большое количество волокон, исчисляемое сотнями.

В качестве другого примера можно указать на определение качества зерна по небольшой его пробе. И в этом случае число наудачу отобранных зерен мало сравнительно со всей массой зерна, но само по себе оно достаточно велико.

Уже из приведенных примеров можно заключить, что для практики теорема Чебышева имеет неоценимое значение.

§ 6. Теорема Бернулли

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A равна p. Можно ли предвидеть какова будет примерно относительная частота появлений события? Положительный ответ на этот вопрос дает теорема, доказанная Яковом Бернулли (опубликована в 1713 г.), которая получила название «закона больших чисел» и положила начало теории вероятностей как науке. Доказательство Бернулли было сложным; простое доказательство дано П. Л. Чебышевым в 1846 г.

Теорема Бернулли. *Если в каждом из n независимых испытаний вероятность p появления события A постоянна, то как угодно близка к единице вероятность того, что отклонение относительной частоты от вероятности p по абсолютной величине будет сколь угодно малым, если число испытаний достаточно велико.*

Другими словами, если ε сколь угодно малое положительное число, то при соблюдении условий теоремы будет иметь место равенство

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1.$$

Доказательство. Обозначим через X_1 дискретную случайную величину — число появлений события в первом испытании, через X_2 — во втором, ..., через X_n — в n-м испытании.

Ясно, что каждая из величин может принять лишь два значения: 1 (событие A наступило) с вероятностью p и 0 (событие не появилось) с вероятностью $1-p=q$.

Можно ли применить к рассматриваемым величинам теорему Чебышева? Можно, если случайные величины попарно независимы и дисперсии их ограничены. Оба условия выполняются. Действительно, попарная независимость величин $X_1, X_2, \ldots, X_n$ следует из того, что испытания независимы. Дисперсия любой величины X_i $(i=1, 2, \ldots, n)$ равна произведению pq*, так как $p+q=1$, то произведение pq не превышает $\frac{1}{4}$** и, следовательно, дисперсии всех величин ограничены, например числом $C=\frac{1}{4}$.

Применяя теорему Чебышева (частный случай) к рассматриваемым величинам, имеем

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-a\right|<\varepsilon\right)=1.$$

Приняв во внимание, что математическое ожидание a каждой из величин X_i (т. е. математическое ожидание числа появлений события в одном испытании) равно вероятности p наступления события (пример 2, стр. 68), получим

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-p\right|<\varepsilon\right)=1.$$

Остается показать, что дробь $\frac{X_1+X_2+\ldots+X_n}{n}$ равна относительной частоте $\frac{m}{n}$ появлений события A в n испыта-

---

* Это следует из § 6, гл. VIII, если принять $n=1$.

** Известно, что произведение двух сомножителей, сумма которых есть величина постоянная, имеет наибольшее значение при равенстве сомножителей. Здесь сумма $p_i+q_i=1$, т. е. постоянна, поэтому при $p_i=q_i=\frac{1}{2}$ произведение p_iq_i имеет наибольшее значение и равно $\frac{1}{4}$.

ниях. Действительно, каждая из величин $X_1, X_2, \ldots, X_n$ при появлении события в соответствующем испытании принимает значение, равное единице; следовательно, сумма $X_1+X_2+\ldots+X_n$ равна числу m появлений события в n испытаниях, а значит,

$$\frac{X_1+X_2+\ldots+X_n}{n}=\frac{m}{n}.$$

Учитывая это равенство, окончательно получим

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1.$$

Замечание. Было бы неправильным на основании теоремы Бернулли сделать вывод, что с ростом числа испытаний относительная частота неуклонно стремится к вероятности p; другими словами, из теоремы Бернулли не вытекает равенство $\lim\limits_{n\to\infty}\frac{m}{n}=p$. В теореме речь идет лишь о вероятности того, что при достаточно большом числе испытаний относительная частота будет как угодно мало отличаться от постоянной вероятности появления события в каждом испытании.

Таким образом, сходимость относительной частоты $\frac{m}{n}$ к вероятности p отличается от сходимости в смысле обычного анализа. Для того чтобы подчеркнуть это различие, вводят понятие «сходимости по вероятности». Точнее, различие между указанными видами сходимости состоит в следующем: если $\frac{m}{n}$ стремится при $n\to\infty$ к p как пределу в смысле обычного анализа, то, начиная с некоторого $n=N$ и для всех последующих значений n, неуклонно выполняется неравенство $\left|\frac{m}{n}-p\right|<\varepsilon$; если же $\frac{m}{n}$ стремится по вероятности к p при $n\to\infty$, то для отдельных значений n неравенство может не выполняться.

Итак, теорема Бернулли утверждает, что при $n\to\infty$ относительная частота стремится по вероятности к p. Коротко теорему Бернулли записывают так:

$$\frac{m}{n}\xrightarrow[n\to\infty]{\text{вер}} p.$$

Как видим, теорема Бернулли объясняет, почему относительная частота при достаточно большом числе испытаний обладает свойством устойчивости и оправдывает статистическое определение вероятности (гл. I, § 5—6).

Задачи

1. Сформулировать и записать теорему Чебышева используя понятие «сходимости по вероятности»

2. Пользуясь неравенством Чебышева, оценить вероятность того, что $|X-M(X)|<0{,}1$, если $D(X)=0{,}001$.
Отв. $P\geqslant 0{,}9$.

3. Дано: $P(|X-M(X)|<\varepsilon)\geqslant 0{,}9$; $D(X)=0{,}004$. Пользуясь неравенством Чебышева, найти ε.
Отв. 0,2.

Глава десятая

ИНТЕГРАЛЬНАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Определение интегральной функции распределения*

Вспомним, что дискретная случайная величина задается перечнем всех ее возможных значений и их вероятностей. Такой способ задания не является общим: он неприменим, например, для непрерывных случайных величин.

Действительно, рассмотрим случайную величину X, возможные значения которой сплошь заполняют интервал (a, b). Можно ли составить перечень всех возможных значений X? Очевидно, что этого сделать нельзя. Этот пример указывает на целесообразность дать общий способ задания любых типов случайных величин. С этой целью и вводят интегральную функцию распределения.

Пусть x — действительное число. Вероятность события, состоящего в том, что X примет значение меньшее x, т. е. вероятность события $X<x$ обозначим через $F(x)$. Разумеется, если x будет изменяться, то вообще говоря, будет изменяться и $F(x)$, т. е. $F(x)$ есть функция от x.

Интегральной функцией распределения называют функцию $F(x)$, определяющую для каждого значения x вероятность того, что случайная величина X примет значение, меньшее x, т. е.

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

Геометрически это равенство можно истолковать так: $F(x)$ есть вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки x.

* Часто вместо термина «интегральная функция» пользуются термином «функция распределения».

Теперь мы можем дать более точное определение непрерывной случайной величины: случайную величину будем называть *непрерывной*, если ее интегральная функция распределения $F(x)$ непрерывно дифференцируема.

§ 2. Свойства интегральной функции

Свойство 1. *Значения интегральной функции принадлежат отрезку* $[0;\ 1]$:

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

Доказательство. Свойство вытекает из определения интегральной функции как вероятности: вероятность всегда есть неотрицательное число, не превышающее единицы.

Свойство 2. $F(x)$ — *неубывающая функция, т. е*

$$F(x_2) \geqslant F(x_1), \text{ если } x_2 > x_1.$$

Доказательство. Пусть $x_2 > x_1$. Событие, состоящее в том, что X примет значение меньшее x_2, можно подразделить на следующие два несовместных события: 1) X примет значение меньшее x_1 с вероятностью $P(X<x_1)$; 2) X примет значение, удовлетворяющее неравенству $x_1 \leqslant X < x_2$, с вероятностью $P(x_1 \leqslant X < x_2)$. По теореме сложения имеем

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

Отсюда

$$P(X<x_2) - P(X<x_1) = P(x_1 \leqslant X < x_2),$$

или

$$F(x_2) - F(x_1) = P(x_1 \leqslant X < x_2). \qquad (*)$$

Так как любая вероятность есть число неотрицательное, то $F(x_2) - F(x_1) \geqslant 0$, или $F(x_2) \geqslant F(x_1)$, что и требовалось доказать.

Следствие 1. *Вероятность того, что случайная величина примет значение, заключенное в интервале* $(a,\ b)$, *равна приращению интегральной функции на этом интервале*:

$$P(a \leqslant X < b) = F(b) - F(a). \qquad (**)$$

Это важное следствие вытекает из формулы (*), если положить $x_2 = b$ и $x_1 = a$.

Пример. Случайная величина X задана интегральной функцией:

$$F(x) = \begin{cases} 0 & \text{при } x \leqslant -1; \\ \frac{1}{4}x + \frac{1}{4} & \text{при } -1 < x \leqslant 3; \\ 1 & \text{при } x > 3. \end{cases}$$

Найти вероятность того, что в результате испытания X примет значение, принадлежащее интервалу $(0;\ 2)$

$$P(0<X<2) = F(2) - F(0).$$

Так как на интервале $(0;\ 2)$ по условию

$$F(x) = \frac{1}{4}x + \frac{1}{4},$$

то

$$F(2) - F(0) = \left[\frac{1}{4}\cdot 2 + \frac{1}{4}\right] - \left[\frac{1}{4}\cdot 0 + \frac{1}{4}\right] = \frac{1}{2}.$$

Итак

$$P(0<X<2) = \frac{1}{2}.$$

Следствие 2. *Вероятность того, что непрерывная случайная величина X примет одно определенное значение, равна нулю.*

Действительно, положив в формуле (**), $a = x_1$, $b = x_1 + \Delta x$, имеем

$$P(x_1 \leqslant X < x_1 + \Delta x) = F(x_1 + \Delta x) - F(x_1).$$

Устремим Δx к нулю. Так как X — непрерывная случайная величина, то функция $F(x)$ непрерывна. В силу непрерывности $F(x)$ в точке x_1 разность $F(x_1+\Delta x) - F(x_1)$ также будет стремиться к нулю, следовательно, $P(X = x_1) = 0$.

Используя это положение, легко убедиться в справедливости равенств

$$P(a \leqslant X < b) = P(a < X < b) = P(a < X \leqslant b) = P(a \leqslant X \leqslant b). \qquad (***)$$

Например, равенство $P(a < X \leqslant b) = P(a < X < b)$ доказывается так:

$$P(a < X \leqslant b) = P(a < X < b) + P(X = b) = P(a < X < b).$$

Таким образом, не представляет интереса говорить о вероятности того, что непрерывная случайная величина примет одно определенное значение, но имеет смысл рассматривать вероятность попадания ее в интервал, пусть даже сколь угодно малый.

Этот факт полностью соответствует требованиям практических задач. Например, интересуются вероятностью того, что размеры деталей не выходят за дозволенные границы, но не ставят вопроса о вероятности их совпадения с проектным размером.

Заметим, что было бы неправильным думать, что равенство нулю вероятности $P(X=x_1)$ означает, что событие $X=x_1$ невозможно (если, конечно, не ограничиваться классическим определением вероятности). Действительно, в результате испытания случайная величина обязательно примет одно из возможных значений; в частности, это значение может оказаться равным x_1.

Свойство 3. *Если возможные значения случайной величины принадлежат интервалу* (a, b), *то*

$$1)\ F(x)=0 \text{ при } x\leqslant a;$$
$$2)\ F(x)=1 \text{ при } x\geqslant b.$$

Доказательство. 1) Пусть $x_1\leqslant a$. Тогда событие $X<x_1$ невозможно (так как значений меньших x_1 величина X по условию не принимает) и, следовательно, вероятность его равна нулю.

2) Пусть $x_2\geqslant b$. Тогда событие $X<x_2$ достоверно (так как все возможные значения X меньше x_2) и, следовательно, вероятность его равна единице.

Следствие. *Если возможные значения непрерывной случайной величины расположены на всей оси* x, *то справедливы следующие предельные соотношения:*

$$\lim_{x\to-\infty} F(x)=0; \quad \lim_{x\to\infty} F(x)=1.$$

§ 3. График интегральной функции

Доказанные свойства позволяют представить, как выглядит график интегральной функции непрерывной случайной величины.

График расположен в полосе, ограниченной прямыми $y=0$, $y=1$ (первое свойство).

При возрастании x в интервале (a, b), в котором заключены все возможные значения случайной величины, график «подымается вверх» (второе свойство).

При $x\leqslant a$ ординаты графика равны нулю; при $x\geqslant b$ ординаты графика равны единице (третье свойство).

График интегральной функции непрерывной случайной величины изображен на рис. 1.

Замечание. Для дискретной случайной величины график интегральной функции имеет ступенчатый вид. Убедимся в этом на примере.

Пример. Дискретная случайная величина X задана следующей таблицей распределения:

X	1	4	8
p	0,3	0,1	0,6.

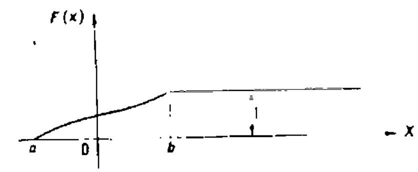

Рис. 1.

Найти интегральную функцию и вычертить ее график.

Решение. 1°. Если $x\leqslant 1$, то $F(x)=0$ (третье свойство).

2°. Если $1<x\leqslant 4$, то $F(x)=0{,}3$. Действительно, X может принять значение 1 с вероятностью 0,3.

3°. Если $4<x\leqslant 8$, то $F(x)=0{,}4$. Действительно, если x_1 удовлетворяет неравенству $4<x_1\leqslant 8$, то $F(x_1)$ равно вероятности события $X<x_1$, которое может быть осуществлено, когда X примет значение 1 (вероятность этого события равна 0,3) или значение 4 (вероятность этого события равна 0,1). Поскольку эти два события несовместны, то по теореме сложения вероятность события $X<x_1$ равна сумме вероятностей $0{,}3+0{,}1=0{,}4$.

4°. Если $x>8$, то $F(x)=1$. Действительно, событие $X\leqslant 8$ достоверно и, следовательно, его вероятность равна единице.

Итак, интегральная функция аналитически может быть записана так:

$$F(x)=\begin{cases}0 & \text{при } x\leqslant 1,\\ 0{,}3 & \text{при } 1<x\leqslant 4,\\ 0{,}4 & \text{при } 4<x\leqslant 8\\ 1 & \text{при } x>8.\end{cases}$$

График этой функции приведен на рис. 2.

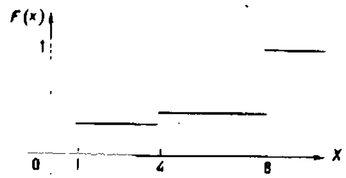

Рис. 2.

Задачи

1. Случайная величина X задана интегральной функцией:

$$F(x)=\begin{cases}0 & \text{при } x<-1;\\ \frac{1}{3}x+\frac{1}{3} & \text{при } -1<x<2;\\ 1 & \text{при } x>2\end{cases}$$

Найти вероятность того, что в результате испытания X примет значение, заключенное в интервале (0; 1).

Отв. $\frac{1}{3}$.

2. Случайная величина X задана интегральной функцией

$$F(x)=\begin{cases}0 & \text{при } x<2;\\ \frac{1}{2}x-1 & \text{при } 2<x<4;\\ 1 & \text{при } x>4.\end{cases}$$

Найти вероятность того, что в результате испытания X примет значение, заключенное в интервале (2; 3)

Отв $\frac{1}{2}$.

3. Дискретная случайная величина X задана следующим законом распределения:

X	2	6	10
p	0,5	0,4	0,1

Построить график интегральной функции этой величины

Глава одиннадцатая

ДИФФЕРЕНЦИАЛЬНАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 1. Определение дифференциальной функции распределения

Выше мы задавали непрерывную случайную величину при помощи интегральной функции. Этот способ задания не является единственным. Непрерывную случайную величину можно также задать, пользуясь дифференциальной функцией распределения вероятностей.

*Дифференциальной функцией распределения** $f(x)$ называют первую производную от интегральной функции:

$$f(x)=F'(x).$$

Из приведенного определения следует, что интегральная функция является первообразной для дифференциальной функции.

Заметим, что для описания распределения вероятностей дискретной случайной величины дифференциальная функция неприменима.

§ 2. Вероятность попадания непрерывной случайной величины в заданный интервал

Зная дифференциальную функцию, можно вычислить вероятность того, что непрерывная случайная величина примет значение, принадлежащее заданному интервалу. Вычисление основано на следующей теореме.

Теорема. *Вероятность того, что непрерывная случайная величина X примет значение, принадлежащее интервалу (a, b), равна определенному интегралу от дифференциальной функции, взятому в пределах от a до b:*

$$P(a<X<b)=\int_a^b f(x)dx.$$

* Часто вместо термина «дифференциальная функция» пользуются термином «плотность вероятности».

Доказательство. Воспользуемся соотношением (**) (стр. 106):

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

По формуле Ньютона—Лейбница

$$F(b) - F(a) = \int_a^b F'(x)dx = \int_a^b f(x)dx.$$

Таким образом,

$$P(a \leqslant X < b) = \int_a^b f(x)dx.$$

Так как $P(a \leqslant X < b) = P(a < X < b)$, то окончательно получим

$$P(a < X < b) = \int_a^b f(x)dx. \qquad (*)$$

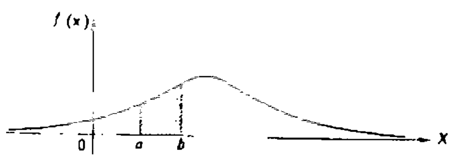

Рис. 3.

Геометрически полученный результат можно истолковать так: вероятность того, что непрерывная случайная величина примет значение, принадлежащее интервалу (a, b), равна площади криволинейной трапеции, ограниченной осью x, кривой распределения $f(x)$ и прямыми $x=a$ и $x=b$ (рис. 3).

Замечание. В частности, если $f(x)$ — четная функция и концы интервала симметричны относительно начала координат, то

$$P(-a < X < a) = P(|X| < a) = 2\int_0^a f(x)\,dx$$

Пример. Дана дифференциальная функция случайной величины X:

$$f(x) = \begin{cases} 0 & \text{при } x \leqslant 0; \\ 2x & \text{при } 0 < x \leqslant 1; \\ 0 & \text{при } x > 1. \end{cases}$$

Найти вероятность того, что в результате испытания X примет значение, принадлежащее интервалу (0,5; 1).

Решение. Искомая вероятность

$$P(0{,}5 < X < 1) = 2\int_{0{,}5}^{1} x\,dx = x^2\Big|_{0{,}5}^{1} = 1 - 0{,}25 = 0{,}75.$$

§ 3. Нахождение интегральной функции распределения по известной дифференциальной функции

Зная дифференциальную функцию $f(x)$, можно найти интегральную функцию $F(x)$ по формуле

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

Действительно, мы обозначили через $F(x)$ вероятность того, что случайная величина примет значение, меньше x, т. е.

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

Очевидно, неравенство $X < x$ можно записать в виде двойного неравенства $-\infty < X < x$, следовательно,

$$F(x) = P(-\infty < X < x). \qquad (*)$$

Полагая в формуле (*) (§ 2) $a = -\infty$, $b = x$, имеем

$$P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx.$$

Наконец, заменив $P(-\infty < X < x)$ через $F(x)$, в силу (*), окончательно получим

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

Таким образом, зная дифференциальную функцию распределения, можно найти интегральную функцию. Разумеется, по известной интегральной функции может быть найдена дифференциальная функция, а именно

$$f(x) = F'(x).$$

Пример. Найти интегральную функцию по данной дифференциальной функции

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

Построить график найденной функции.

Решение. Воспользуемся формулой $F(x) = \int\limits_{-\infty}^{x} f(x)dx$. Если $x \leqslant a$, то $f(x) = 0$ и, следовательно, $F(x) = 0$. Если $a < x \leqslant b$, то $f(x) = \frac{1}{b-a}$ и, следовательно,

$$F(x) = \int\limits_{-\infty}^{x} f(x)\,dx = \int\limits_{-\infty}^{a} 0\,dx + \int\limits_{a}^{x} \frac{1}{b-a}\,dx = \frac{x-a}{b-a}.$$

Если $x > b$, то

$$F(x) = \int\limits_{-\infty}^{a} 0\,dx + \int\limits_{a}^{b} \frac{dx}{b-a} + \int\limits_{b}^{x} 0\,dx = \frac{b-a}{b-a} = 1.$$

Итак, искомая интегральная функция аналитически может быть записана так:

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

График этой функции изображен на рис. 4.

§ 4. Свойства дифференциальной функции

Свойство 1. *Дифференциальная функция неотрицательна:*

$$f(x) \geqslant 0.$$

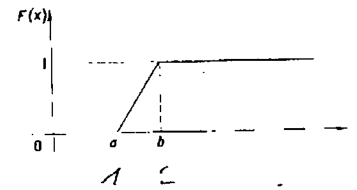

Рис. 4.

Доказательство. Интегральная функция есть не убывающая функция, следовательно, ее производная $F'(x) = f(x)$ есть функция неотрицательная.

Геометрически это свойство означает, что точки, принадлежащие графику дифференциальной функции, расположены либо над осью x, либо на этой оси.

Заметим, что график дифференциальной функции принято называть *кривой распределения.*

Свойство 2. *Несобственный интеграл от дифференциальной функции в пределах от $-\infty$ до ∞ равен единице:*

$$\int\limits_{-\infty}^{\infty} f(x)dx = 1.$$

Доказательство. Несобственный интеграл $\int\limits_{-\infty}^{\infty} f(x)dx$ выражает вероятность события, состоящего в том, что случайная величина примет значение, принадлежащее интервалу $(-\infty, \infty)$. Очевидно, такое событие достоверно и, следовательно, вероятность его равна единице.

Геометрически это означает, что вся площадь криволинейной трапеции, ограниченной осью x и кривой распределения, равна единице.

В частности, если все возможные значения случайной величины принадлежат интервалу (a, b), то

$$\int\limits_{a}^{b} f(x)\,dx = 1.$$

Пример. Дифференциальная функция распределения случайной величины x задана равенством

$$fx = \frac{a}{e^x + e^{-x}}.$$

Найти постоянный параметр a.

Решение. Дифференциальная функция должна удовлетворять условию $\int\limits_{-\infty}^{\infty} f(x)dx=1$, поэтому потребуем, чтобы выполнялось равенство

$$a\int\limits_{-\infty}^{\infty}\frac{dx}{e^{-x}+e^{x}}=1.$$

Отсюда

$$a=\frac{1}{\int\limits_{-\infty}^{\infty}\frac{dx}{e^{-x}+e^{x}}}.$$

Найдем неопределенный интеграл

$$\int\frac{dx}{e^{-x}+e^{x}}=\int\frac{e^{x}\,dx}{1+e^{2x}}=\operatorname{arctg} e^{x}.$$

Вычислим несобственный интеграл

$$\int\limits_{-\infty}^{\infty}\frac{dx}{e^{-x}+e^{x}}=\lim_{b\to-\infty}\int\limits_{b}^{0}\frac{dx}{e^{-x}+e^{x}}+\lim_{c\to\infty}\int\limits_{0}^{c}\frac{dx}{e^{-x}+e^{x}}=$$

$$=\lim_{b\to-\infty}(-\operatorname{arctg} e^{b})+\lim_{c\to\infty}(\operatorname{arctg} e^{c})=\frac{\pi}{2}.$$

Таким образом, искомый параметр

$$a=\frac{1}{\frac{\pi}{2}}=\frac{2}{\pi}.$$

§ 5. Вероятностный смысл дифференциальной функции

Пусть $F(x)$ — интегральная функция непрерывной случайной величины X. По определению дифференциальной функции $f(x)=F'(x)$, или в иной форме

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

Как мы уже знаем, разность $F(x+\Delta x)-F(x)$ определяет вероятность того, что X примет значение, принадлежащее интервалу $(x,\ x+\Delta x)$. Таким образом, предел отношения вероятности того, что непрерывная случайная величина примет значение, принадлежащее интервалу $(x,\ x+\Delta x)$, к длине этого интервала (при $\Delta x\to 0$), равен значению дифференциальной функции в точке x.

По аналогии с определением плотности массы в точке* целесообразно рассматривать значение функции $f(x)$ в точке x как плотность вероятности в этой точке.

Итак, дифференциальная функция определяет плотность распределения вероятности для каждой точки x.

Из дифференциального исчисления известно, что приращение функции приближенно равно дифференциалу функции, т. е.

$$F(x+\Delta x)-F(x)\simeq dF(x),$$

или

$$F(x+\Delta x)-F(x)\simeq F'(x)dx.$$

Так как $F'(x)=f(x)$ и $dx=\Delta x$, то

$$F(x+\Delta x)-F(x)\simeq f(x)\Delta x.$$

1 Вероятностный смысл этого равенства таков: вероятность того, что случайная величина примет значение, принадлежащее интервалу $(x,\ x+\Delta x)$, приближенно равна (с точностью до бесконечно малых высшего порядка относительно Δx) произведению плотности вероятности в точке x на длину интервала Δx.

2 Геометрически этот результат можно истолковать так: вероятность того, что случайная величина примет значение, принадлежащее интервалу $(x,\ x+\Delta x)$, приближенно равна площади прямоугольника с основанием Δx и высотой $f(x)$.

На рис. 5 видно, что площадь заштрихованного прямоугольника равна произведению $f(x)\Delta x$, лишь приближенно равна площади криволинейной трапеции (истинной вероятности, определяемой определенным интегралом $\int\limits_{x}^{x+\Delta x} f(x)dx$). Допущенная при этом погрешность равна площади криволинейного треугольника ABC.

---

* Если масса непрерывно распределена вдоль оси x по некоторому закону, например $F(x)$, то плотностью $\rho(x)$ массы в точке x называют предел отношения массы интервала $(x,\ x+\Delta x)$ к длине интервала при $\Delta x\to 0$, т. е. $\rho(x)=\lim\limits_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}$.

§ 6. Закон равномерного распределения вероятностей

При решении задач, которые выдвигает практика, сталкиваются с различными распределениями непрерывных случайных величин. Дифференциальные функции этих распределений называют также законами распределений. Часто встречаются, например, законы равномерного и нормального распределений. В настоящем параграфе рассматривается закон равномерного распределения. Закону нормального распределения посвящена следующая глава.

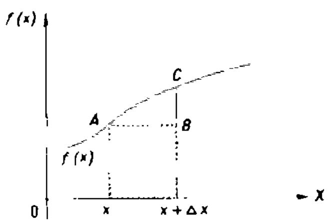

Рис. 5.

Распределение вероятностей называют *равномерным*, если на интервале, которому принадлежат все возможные значения случайной величины, дифференциальная фукция имеет постоянное значение.

Приведем пример равномерно распределенной непрерывной случайной величины.

Пример. Шкала измерительного прибора проградуирована в некоторых единицах. Ошибку при округлении отсчета до ближайшего целого деления можно рассматривать как случайную величину X, которая может принимать с постоянной плотностью вероятности любое значение между двумя соседними целыми делениями. Таким образом, X имеет равномерное распределение.

Найдем дифференциальную функцию равномерного распределения, считая, что все возможные значения случайной величины заключены в интервале (a, b), на котором дифференциальная функция сохраняет постоянное значение $f(x)=C$.

По условию X не принимает значений вне интервала (a, b), поэтому $f(x)=0$ при $x<a$ и $x>b$.

Найдем значение постоянной C. Так как все возможные значения случайной величины принадлежат интервалу (a, b), то должно выполняться равенство

$$\int_a^b f(x)\,dx = 1, \quad \text{или} \quad \int_a^b C\,dx = 1.$$

Отсюда

$$C = \frac{1}{\int_a^b dx} = \frac{1}{b-a}.$$

Итак, закон равномерного распределения аналитически можно записать так:

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

График дифференциальной функции равномерного распределения изображен на рис. 6, а график интегральной функции — на рис. 4.

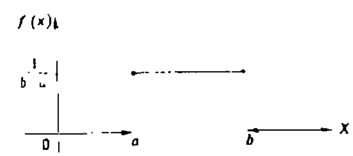

Рис. 6.

Задачи

1. Случайная величина задана дифференциальной функцией

$$f(x) = \begin{cases} 0 & \text{при } x < -\dfrac{\pi}{2}; \\ a\cos x & \text{при } -\dfrac{\pi}{2} < x < \dfrac{\pi}{2}; \\ 0 & \text{при } x > \dfrac{\pi}{2}. \end{cases}$$

Найти коэффициент a.

Отв. $a = \dfrac{1}{2}$.

2. Случайная величина задана дифференциальной функцией

$$f(x) = \begin{cases} 0 & \text{при } x < 0; \\ \frac{1}{2}\sin x & \text{при } 0 < x < \pi; \\ 0 & \text{при } x > \pi. \end{cases}$$

Найти: а) интегральную функцию; б) вероятность того, что в результате испытания случайная величина примет значение, заключенное в интервале $(0; \frac{\pi}{4})$.

Отв. а) $$F(x) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{1}{2}(1 - \cos x) & \text{при } 0 < x < \pi; \\ 1 & \text{при } x > \pi; \end{cases}$$

б) $\frac{2-\sqrt{2}}{4}$.

3. Случайная величина X задана интегральной функцией:

$$F(x) = \begin{cases} 0 & \text{при } x < 0, \\ x & \text{при } 0 < x < 1, \\ 1 & \text{при } x > 1. \end{cases}$$

Найти дифференциальную функцию.

Отв. $$f(x) = \begin{cases} 0 & \text{при } x < 0, \\ 1 & \text{при } 0 < x < 1; \\ 0 & \text{при } x > 1. \end{cases}$$

4. Случайная величина X задана интегральной функцией:

$$F(x) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{1}{2}(1 - \cos x) & \text{при } 0 < x < \pi, \\ 1 & \text{при } x > \pi. \end{cases}$$

Найти дифференциальную функцию.

Отв. $$f(x) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{1}{2}\sin x & \text{при } 0 < x \leqslant \pi, \\ 0 & \text{при } x > \pi. \end{cases}$$

Глава двенадцатая

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

§ 1. Числовые характеристики непрерывных случайных величин

Распространим определения числовых характеристик дискретных величин на величины непрерывные. Начнем с математического ожидания.

Пусть непрерывная случайная величина X задана дифференциальной функцией $f(x)$. Допустим, что все возможные значения X принадлежат отрезку $[a, b]$. Разобьем этот отрезок на n частичных отрезков длиною $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$ и выберем в каждом из них произвольную точку x_i $(i = 1, 2, \ldots, n)$. Имея в виду определить математическое ожидание непрерывной величины по аналогии с дискретной, составим сумму произведений возможных значений x_i на вероятности попадания их в интервал Δx_i (напомним, что произведение $f(x)\Delta x$ приближенно равно вероятности попадания X в интервал Δx):

$$\sum x_i \cdot f(x_i)\Delta x_i.$$

Перейдя к пределу при стремлении к нулю длины наибольшего из частичных отрезков, получим определенный интеграл $\int_a^b xf(x)dx$.

Математическим ожиданием непрерывной случайной величины X, возможные значения которой принадлежат отрезку $[a, b]$, называют определенный интеграл:

$$M(X) = \int_a^b xf(x)\,dx.$$

Если возможные значения принадлежат всей оси x, то

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Предполагается, что несобственный интеграл сходится абсолютно, т. е. существует интеграл $\int_{-\infty}^{\infty} |x|\, f(x)dx$. Если бы это требование не выполнялось, то значение интеграла зависело бы от скорости стремления (в отдельности) нижнего предела к $-\infty$, а верхнего к $+\infty$.

По аналогии с дисперсией дискретной величины определяется и дисперсия непрерывной величины.

Дисперсией непрерывной случайной величины называют математическое ожидание квадрата ее отклонения.

Если возможные значения X принадлежат отрезку $[a, b]$, то

$$D(X)=\int_a^b [x-M(X)]^2 f(x)\,dx;$$

если возможные значения принадлежат всей оси x, то

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

Среднее квадратическое отклонение непрерывной случайной величины определяется, как и для величины дискретной, равенством

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

Замечание 1. Можно доказать, что свойства математического ожидания и дисперсии дискретных величин сохраняются и для непрерывных величин.

Замечание 2. Легко получить для вычисления дисперсии более удобные формулы:

$$D(X)=\int_a^b x^2 f(x)\,dx-[M(X)]^2,$$

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

Пример. Найти математическое ожидание и дисперсию случайной величины X, заданной интегральной функцией

$$F(x)=\begin{cases}0 & \text{при } x\leqslant 0,\\ x & \text{при } 0<x\leqslant 1,\\ 1 & \text{при } x>1.\end{cases}$$

Решение. Найдем дифференциальную функцию

$$f(x)=F'(x)=\begin{cases}0 & \text{при } x\leqslant 0,\\ 1 & \text{при } 0<x<1,\\ 0 & \text{при } x>1.\end{cases}$$

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

Эта функция табулирована (приложение 1).

Замечание 2. Интегральная функция общего нормального распределения (гл. XI, § 3)

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(z-a)^2}{2\sigma^2}}\,dz,$$

а нормированного

$$F_0(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz.$$

Функция $F_0(x)$ табулирована. Легко проверить, что

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

Замечание 3. Вероятность попадания нормированной нормальной величины X в интервал $(0, x)$ можно найти пользуясь функцией Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}\,dz$. Действительно (гл. XI, § 2), $P(0<X<x)=\int_0^x \varphi(x)\,dx=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}\,dz=\Phi(x)$.

Замечание 4. Учитывая, что $\int_{-\infty}^{\infty}\varphi(x)dx=1$ (гл. XI, § 4, свойство 2) и, следовательно, в силу симметрии $\varphi(x)$ относительно нуля,

$$\int_{-\infty}^{0}\varphi(x)dx=0{,}5, \text{ а значит и } P(-\infty<X<0)=0{,}5,$$

легко получить, что

$$F_0(x)=0{,}5+\Phi(x).$$

Действительно,

$$F_0(x)=P(-\infty<X<x)=P(-\infty<X<0)+P(0<X<x)=0{,}5+\Phi(x).$$

§ 3. Нормальная кривая

График дифференциальной функции нормального распределения называют *нормальной кривой* (*кривой Гаусса*).

Исследуем функцию

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

методами дифференциального исчисления.

1. Очевидно, функция определена на всей оси x.

2. При всех значениях x функция принимает положительные значения, т. е. нормальная кривая расположена над осью x.

3. Предел функции при неограниченном возрастании x (по абсолютной величине) равен нулю: $\lim\limits_{|x| \to \infty} y = 0$, т. е. ось x служит горизонтальной асимптотой графика.

4. Исследуем функцию на экстремум. Найдем первую производную:

$$y' = -\frac{x-a}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Легко видеть, что $y'=0$ при $x=a$, $y'>0$ при $x<a$, $y'<0$ при $x>a$.

Следовательно, при $x=a$ функция имеет максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$.

5. Разность $x-a$ содержится в аналитическом выражении функции в квадрате, т. е. график функции симметричен относительно прямой $x=a$.

6. Исследуем функцию на точки перегиба. Найдем вторую производную

$$y'' = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \cdot \left[1 - \frac{(x-a)^2}{\sigma^2}\right].$$

Легко видеть, что при $x=a+\sigma$ и $x=a-\sigma$ вторая производная равна нулю, а при переходе через эти точки она меняет знак (в обеих этих точках значение функции равно $\left.\frac{1}{\sigma\sqrt{2\pi e}}\right)$. Таким образом, точки графика

$$\left(a-\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right) \text{ и } \left(a+\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$$

являются точками перегиба.

На рис. 7 изображена нормальная кривая при $a=1$ и $\sigma=2$.

§ 4. Влияние параметров нормального распределения на форму нормальной кривой

Выясним как влияют на форму и расположение нормальной кривой значения параметров a и σ.

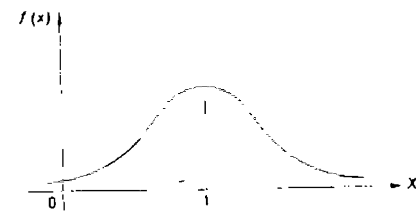

Рис. 7.

Известно, что графики функций $f(x)$ и $f(x-a)$ имеют одинаковую форму; сдвинув график $f(x)$ в положительном направлении оси x на a единиц масштаба при $a>0$, или в отрицательном направлении при $a<0$, получим график $f(x-a)$. Отсюда следует, что и з м е н е н и е в е л и ч и н ы п а р а м е т р а a (математического ожидания) н е и з м е н я е т ф о р м ы н о р м а л ь н о й к р и в о й, а п р и в о д и т л и ш ь к е е с д в и г у в д о л ь о с и x: в п р а в о, е с л и a в о з р а с т а е т и в л е в о, е с л и a у б ы в а е т.

По-иному обстоит дело, если изменяется параметр σ (среднее квадратическое отклонение). Как было указано в предыдущем параграфе, максимум дифференциальной функции нормального распределения равен $\frac{1}{\sigma\sqrt{2\pi}}$. Отсюда следует, что с в о з р а с т а н и е м σ м а к с и м а л ь н а я о р д и н а т а н о р м а л ь н о й к р и в о й у б ы в а е т, а с а м а к р и в а я с т а н о в и т с я б о л е е п о л о г о й, т. е. с ж и м а е т с я к о с и x; п р и у б ы в а н и и σ н о р м а л ь н а я к р и в а я с т а н о в и т с я б о л е е «о с т р о в е р ш и н-

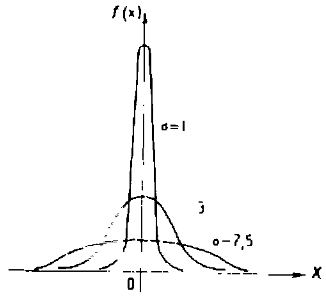

Рис. 8.

ной» и растягивается в положительном направлении оси y.

Подчеркнем, что при любых значениях параметров a и σ площадь, ограниченная нормальной кривой и осью x, остается равной единице (гл. XI, § 4, второе свойство дифференциальной функции).

На рис. 8 изображены нормальные кривые при различных значениях σ и $a=0$. Чертеж наглядно иллюстрирует как изменение параметра σ сказывается на форме нормальной кривой.

Заметим, что при $a=0$ и $\sigma=1$ нормальную кривую $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ называют *нормированной*.

§ 5. Вероятность попадания в заданный интервал нормальной случайной величины

Мы уже знаем, что если случайная величина X задана дифференциальной функцией $f(x)$, то вероятность того, что X примет значение, принадлежащее интервалу (α, β), такова:

$$P(\alpha<X<\beta)=\int_{\alpha}^{\beta}f(x)\,dx.$$

Пусть случайная величина X распределена по нормальному закону. Тогда вероятность того, что X примет значение, принадлежащее интервалу (α, β), равна

$$P(\alpha<X<\beta)=\frac{1}{\sigma\sqrt{2\pi}}\int_{\alpha}^{\beta}e^{-\frac{(x-a)^2}{2\sigma^2}}\,dx.$$

Преобразуем эту формулу так, чтобы можно было пользоваться готовыми таблицами. Введем новую переменную $z=\frac{x-a}{\sigma}$. Отсюда $x=\sigma z+a$, $dx=\sigma dz$. Найдем новые пределы интегрирования. Если $x=\alpha$, то $z=\frac{\alpha-a}{\sigma}$; если $x=\beta$, то $z=\frac{\beta-a}{\sigma}$.

Таким образом, имеем

$$P(\alpha<X<\beta)=\frac{1}{\sigma\sqrt{2\pi}}\int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}}e^{-\frac{z^2}{2}}(\sigma dz)=\frac{1}{\sqrt{2\pi}}\int_{\frac{\alpha-a}{\sigma}}^{0}e^{-\frac{z^2}{2}}dz+$$

$$+\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{\beta-a}{\sigma}}e^{-\frac{z^2}{2}}dz=\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{\beta-a}{\sigma}}e^{-\frac{z^2}{2}}dz-\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{\alpha-a}{\sigma}}e^{-\frac{z^2}{2}}dz.$$

Пользуясь функцией Лапласа

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{z^2}{2}}dz,$$

окончательно получим

$$P(\alpha<X<\beta)=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right). \qquad (*)$$

Пример. Случайная величина X распределена по нормальному закону. Математическое ожидание и среднее квадратическое отклонение этой величины соответственно равны 30 и 10. Найти вероятность того, что X примет значение, принадлежащее интервалу (10, 50).

Решение. Воспользуемся формулой (*). По условию $\alpha=10$, $\beta=50$, $a=30$, $\sigma=10$, следовательно,

$$P(10<X<50)=\Phi\left(\frac{50-30}{10}\right)-\Phi\left(\frac{10-30}{10}\right)=2\Phi(2).$$

По таблице (приложение 2) находим

$$\Phi(2)=0{,}4772.$$

Отсюда искомая вероятность

$$P(10<X<50)=2\cdot 0{,}4772=0{,}9544.$$

§ 6. Вычисление вероятности заданного отклонения

Часто требуется вычислить вероятность того, что отклонение нормально распределенной случайной величины X по абсолютной величине меньше заданного положительного числа α, т. е. требуется найти вероятность осуществления неравенства $|X-a|<\sigma$.

Заменим это неравенство равносильным ему двойным неравенством

$$-\delta<X-a<\delta,$$

или

$$a-\delta<X<a+\delta.$$

Пользуясь формулой (*) (§ 5), получим

$$P(|X-a|<\delta)=P(a-\delta<X<a+\delta)=$$

$$=\Phi\left[\frac{(a+\delta)-a}{\sigma}\right]-\Phi\left[\frac{(a-\delta)-a}{\sigma}\right]=\Phi\left(\frac{\delta}{\sigma}\right)-\Phi\left(-\frac{\delta}{\sigma}\right).$$

Приняв во внимание равенство

$$\Phi\left(-\frac{\delta}{\sigma}\right)=-\Phi\left(\frac{\delta}{\sigma}\right)$$

(функция Лапласа — нечетная), окончательно имеем

$$P(|X-a|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

В частности, при $a=0$

$$P(|X|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

На рис. 9 наглядно показано, что если две случайные величины нормально распределены и $a=0$, то вероятность принять значение, принадлежащее интервалу $(-\delta,\ \delta)$, больше у той величины, которая имеет меньшее значение σ. Этот факт полностью соответствует вероятностному смыслу параметра σ (σ есть среднее квадратическое отклонение; оно характеризует рассеяние случайной величины вокруг ее математического ожидания).

Замечание. Очевидно, события, состоящие в осуществлении неравенств $|X-a|<\delta$ и $|X-a|\geqslant\delta$, — противоположные. Поэтому, если вероятность осуществления неравенства $|X-a|<\delta$ равна p, то вероятность неравенства $|X-a|\geqslant\delta$ равна $1-p$

Пример. Случайная величина X распределена нормально. Математическое ожидание и среднее квадратическое отклонение X соответственно равны 20 и 10. Найти вероятность того, что отклонение по абсолютной величине будет меньше трех.

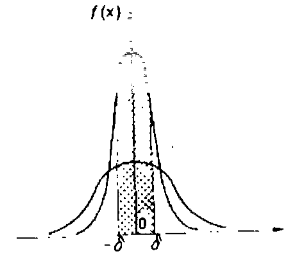

Рис. 9.

Решение. Воспользуемся формулой

$$P(|X-a|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

По условию $\delta=3$, $a=20$, $\sigma=10$. Следовательно,

$$P(|X-20|<3)=2\Phi\left(\frac{3}{10}\right)=2\Phi(0,3).$$

По таблице (приложение 2) находим $\Phi(0,3)=0,1179$.

Искомая вероятность

$$P(|X-20|<3)=0,2358.$$

§ 7. Правило трех сигм

Преобразуем формулу (§ 6)

$$P(|X-a|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right),$$

положив $\delta=\sigma t$. В итоге получим

$$P(|X-a|<\sigma t)=2\Phi(t).$$

Если $t=3$ и, следовательно, $\sigma t=3\sigma$, то

$$P(|X-a|<3\sigma)=2\Phi(3)=2\cdot 0,49865=0,9973,$$

т. е. вероятность того, что отклонение по абсолютной величине будет меньше утроенного среднего квадратического отклонения, равна 0,9973.

Другими словами, вероятность того, что абсолютная величина отклонения превысит утроенное среднее квадратическое отклонение, очень мала, а именно равна 0,0027. Это означает, что лишь в 0,27% случаев так может произойти. Такие события, исходя из принципа невозможности маловероятных событий, можно считать практически невозможными. В этом и состоит сущность правила трех сигм: *если случайная величина распределена нормально, то абсолютная величина ее отклонения от математического ожидания не превосходит утроенного среднего квадратического отклонения.*

На практике правило трех сигм применяется так: если распределение изучаемой случайной величины неизвестно, но условие, указанное в приведенном правиле, выполняется, то имеются основания предполагать, что изучаемая величина распределена нормально; в противном случае она не распределена нормально.

§ 8. Понятие о теореме Ляпунова

Известно, что нормально распределенные случайные величины широко распространены на практике. Чем это объясняется? Ответ на этот вопрос был дан выдающимся русским математиком А. М. Ляпуновым (центральная предельная теорема теории вероятностей). Приведем лишь *следствие из теоремы Ляпунова*: если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то X имеет распределение, близкое к нормальному.

На практике наиболее часто встречаются именно такие случайные величины.

Приведем пример, поясняющий сказанное.

Пример. Пусть производится измерение некоторой физической величины. Любое измерение дает лишь приближенное значение измеряемой величины, так как на результат измерения оказывают влияние очень многие независимые случайные факторы (температура, колебания прибора, влажность и др.). Каждый из этих факторов порождает ничтожную «частную ошибку». Однако, поскольку число этих факторов очень велико, совокупное их действие порождает уже заметную «суммарную ошибку».

Рассматривая суммарную ошибку как сумму очень большого числа взаимно независимых частных ошибок, мы вправе заключить, что суммарная ошибка имеет распределение, близкое к нормальному. Опыт подтверждает справедливость такого заключения.

§ 9. Оценка отклонения теоретического распределения от нормального. Асимметрия и эксцесс

Эмпирическим называют распределение относительных частот. Эмпирические распределения изучает математическая статистика.

Теоретическим называют распределение вероятностей. Теоретические распределения изучает теория вероятностей. В этом параграфе рассматриваются теоретические распределения.

При изучении распределений, отличных от нормального, возникает необходимость количественно оценить это различие. С этой целью вводят специальные характеристики, в частности, асимметрию и эксцесс. Для нормального распределения эти характеристики равны нулю. Поэтому, если для изучаемого распределения асимметрия и эксцесс имеют небольшие значения, то можно предположить близость этого распределения к нормальному. Наоборот, большие значения асимметрии и эксцесса указывают на значительное отклонение от нормального.

Как оценить асимметрию? Можно доказать, что для симметричного распределения (график такого распределения симметричен относительно прямой $x=M(X)$) каждый центральный момент нечетного порядка равен нулю. Для несимметричных распределений центральные моменты нечетного порядка отличны от нуля. Поэтому любой из этих моментов (кроме момента первого порядка, который равен нулю для любого распределения) может служить для оценки асимметрии; естественно выбрать простейший из них, т. е. момент третьего порядка μ_3. Однако принять этот момент для оценки асимметрии неудобно потому, что его величина зависит от единиц, в которых измеряется случайная величина. Чтобы устранить этот недостаток, μ_3 делят на σ^3 и таким образом получают безразмерную характеристику.

Асимметрией теоретического распределения называют отношение центрального момента третьего порядка к кубу среднего квадратического отклонения:

$$A_s = \frac{\mu_3}{\sigma^3}.$$

Асимметрия положительна, если «длинная часть» кривой распределения расположена справа от математического ожидания; асимметрия отрицательна, если «длинная часть»

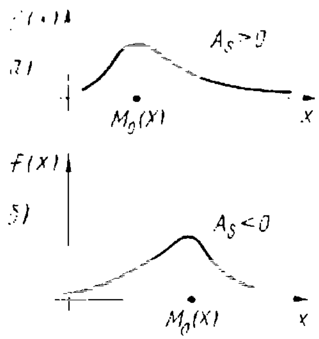

Рис. 10.

кривой расположена слева от математического ожидания. Практически определяют знак асимметрии по расположению кривой распределения относительно моды (точки максимума дифференциальной функции): если длинная часть кривой расположена правее моды, то асимметрия положительна (рис. 10, *а*), если слева — отрицательна (рис. 10, *б*).

Для оценки «крутости», т. е. большего или меньшего подъема кривой теоретического распределения по сравнению с нормальной кривой, пользуются характеристикой—эксцессом.

Эксцессом теоретического распределения называют характеристику, которая определяется равенством

$$E_k = \frac{\mu_4}{\sigma^4} - 3.$$

Для нормального распределения $\frac{\mu_4}{\sigma^4} = 3$, и, следовательно, эксцесс равен нулю. Поэтому, если эксцесс некоторого распределения отличен от нуля, то кривая этого распределения отличается от нормальной кривой: если эксцесс положительный, то кривая имеет более высокую и «острую» вершину, чем нормальная кривая (рис. 11, *а*); если эксцесс отрицательный, то сравниваемая кривая имеет более низкую и «плоскую» вершину, чем нормальная кривая (рис. 11, *б*). При этом предполагается, что нормальное и теоретическое распределения имеют одинаковые математические ожидания и дисперсии.

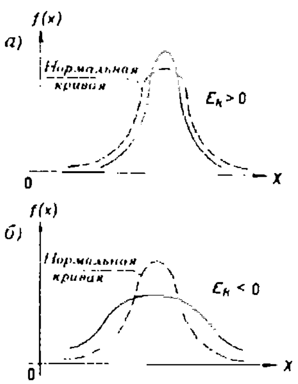

Рис. 11.

§ 10. Функция одного случайного аргумента и ее распределение

Предварительно заметим, что далее вместо того, чтобы говорить «закон распределения вероятностей», мы будем часто говорить кратко — «распределение».

Если каждому возможному значению случайной величины X соответствует одно возможное значение случайной величины Y, то Y называют функцией случайного аргумента X:

$$Y = \varphi(X).$$

Далее показано, как найти распределение функции по известному распределению дискретного и непрерывного аргумента.

1. Пусть аргумент X — дискретная случайная величина.

а) Если различным возможным значениям аргумента X соответствуют различные возможные значения функции Y, то вероятности соответствующих значений X и Y между собой равны.

Пример 1. Дискретная случайная величина X задана распределением:

X	2	3
p	0,6	0,4.

Найти распределение функции $Y = X^2$.

Решение. Найдем возможные значения Y:

$$y_1 = 2^2 = 4; \quad y_2 = 3^2 = 9.$$

Напишем искомое распределение Y:

Y	4	9
p	0,6	0,4.

б) Если различным возможным значениям X соответствуют значения Y, среди которых есть равные между собой, то следует складывать вероятности повторяющихся значений Y.

Пример 2. Дискретная случайная величина X задана распределением:

X	−2	2	3
p	0,4	0,5	0,1.

Найти распределение функции $Y = X^2$.

Решение. Вероятность возможного значения $y_1 = 4$ равна сумме вероятностей несовместных событий $X = -2$,

$X=2$, т. е. $0{,}4+0{,}5=0{,}9$. Вероятность возможного значения $y_2=9$ равна 0,1. Напишем искомое распределение Y:

$$\begin{array}{lll} Y & 4 & 9 \\ p & 0{,}9 & 0{,}1. \end{array}$$

2. Пусть аргумент X — непрерывная случайная величина. Как найти распределение функции $Y=\varphi(X)$, зная дифференциальную функцию $f(x)$ случайного аргумента X? Доказано: если $y=\varphi(x)$ дифференцируемая строго возрастающая или строго убывающая функция, обратная функция которой $x=\varphi(y)$, то дифференциальная функция $g(y)$ случайной величины Y находится по равенству

$$g(y)=f[\psi(y)]\cdot|\psi'(y)|.$$

Пример 3. Случайная величина X распределена нормально, причем ее математическое ожидание $a=0$. Найти распределение функции $Y=X^3$.

Решение. Так как функция $y=x^3$ дифференцируема и строго возрастает, то можно применить формулу:

$$g(y)=f[\psi(y)]\cdot|\psi'(y)|. \qquad (*)$$

Найдем функцию, обратную функции $y=x^3$:

$$\psi(y)=x=y^{\frac{1}{3}}.$$

Найдем $f[\psi(y)]$. По условию

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

поэтому

$$f[\psi(y)]=f\left[y^{\frac{1}{3}}\right]=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{y^{\frac{2}{3}}}{2\sigma^2}}. \qquad (**)$$

Найдем производную обратной функции по y:

$$\psi'(y)=\left(y^{\frac{1}{3}}\right)'=\frac{1}{3y^{\frac{2}{3}}}. \qquad (***)$$

Найдем искомую дифференциальную функцию, для чего подставим (**) и (***) в (*):

$$g(y)=\frac{1}{3\sigma y^{\frac{2}{3}}\sqrt{2\pi}}e^{-\frac{y^{\frac{2}{3}}}{2\sigma^2}}.$$

Замечание. Пользуясь формулой (*), можно доказать, что линейная функция $Y=AX+B$ нормально распределенного аргумента X также распределена нормально, причем для того чтобы найти математическое ожидание Y, надо в выражение функции подставить вместо аргумента X его математическое ожидание a:

$$M(Y)=Aa+B;$$

для того чтобы найти среднее квадратическое отклонение Y, надо среднее квадратическое отклонение аргумента X умножить на модуль коэффициента при X:

$$\sigma(Y)=|A|\cdot\sigma(X).$$

Пример 4. Найти дифференциальную функцию линейной функции $Y=3X+1$, если аргумент распределен нормально, причем математическое ожидание X равно 2 и среднее квадратическое отклонение равно 0,5.

Решение. Найдем математическое ожидание Y

$$M(Y)=3\cdot 2+1=7.$$

Найдем среднее квадратическое отклонение Y

$$\sigma(Y)=3\cdot 0{,}5=1{,}5.$$

Искомая дифференциальная функция имеет вид

$$g(y)=\frac{1}{1{,}5\sqrt{2\pi}}e^{-\frac{(y-7)^2}{2\cdot(1{,}5)^2}}.$$

§ 11. Математическое ожидание функции одного случайного аргумента

Задана функция $Y=\varphi(X)$ случайного аргумента X. Требуется найти математическое ожидание этой функции, зная закон распределения аргумента.

1. Пусть аргумент X — дискретная случайная величина с возможными значениями $x_1, x_2, \ldots, x_n$, вероятности которых соответственно равны $p_1, p_2, \ldots, p_n$. Очевидно, Y также дискретная случайная величина с возможными значениями $y_1=\varphi(x_1)$, $y_2=\varphi(x_2)$, ..., $y_n=\varphi(x_n)$. Так как событие «величина X приняла значение x_i» влечет за собой событие «величина Y приняла значение $\varphi(x_i)$», то вероятности возможных зна-

чений Y соответственно равны $p_1, p_2, \ldots, p_n$. Следовательно, математическое ожидание функции

$$M[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i) \cdot p_i. \qquad (*)$$

Пример 1. Дискретная случайная величина X задана распределением

X	1	3	5
p	0,2	0,5	0,3.

Найти математическое ожидание функции $Y=\varphi(X)=X^2+1$.

Решение. Найдем возможные значения Y:

$\varphi(1)=1^2+1=2$; $\varphi(3)=3^2+1=10$; $\varphi(5)=5^2+1=26$.

Искомое математическое ожидание функции

$$M[X^2+1]=2\cdot 0,2+10\cdot 0,5+26\cdot 0,3=13,2.$$

2. Пусть аргумент X — непрерывная случайная величина, заданная дифференциальной функцией $f(x)$. Для отыскания математического ожидания функции $Y=\varphi(X)$ можно сначала найти дифференциальную функцию $g(y)$ величины Y, а затем воспользоваться формулой

$$M(Y) = \int_{-\infty}^{\infty} y \cdot g(y)dy.$$

Однако, если отыскание дифференциальной функции $g(y)$ является затруднительным, то можно непосредственно найти математическое ожидание функции $\varphi(X)$ по формуле

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

В частности, если возможные значения X принадлежат интервалу (a, b),

$$M[\varphi(X)] = \int_{a}^{b} \varphi(x) f(x) dx. \qquad (**)$$

Опуская доказательство, заметим, что оно аналогично доказательству формулы (*), если заменить суммирование интегрированием, а вероятность — элементом вероятности $f(x)\Delta x$.

Пример 2. Непрерывная случайная величина X задана дифференциальной функцией $f(x)=\sin x$ в интервале $(0; \frac{\pi}{2})$; $f(x)=0$ вне этого интервала. Найти математическое ожидание функции $Y=\varphi(X)=X^2$.

Решение. Воспользуемся формулой (**). По условию $f(x)=\sin x$, $\varphi(x)=x^2$, $a=0$, $b=\frac{\pi}{2}$.

Следовательно,

$$M[\varphi(X)] = \int_{0}^{\frac{\pi}{2}} x^2 \sin x dx.$$

Интегрируя по частям, получим искомое математическое ожидание

$$M[X^2] = \pi - 2.$$

§ 12. Функция двух случайных аргументов. Распределение суммы независимых слагаемых. Устойчивость нормального распределения

Если каждой паре возможных значений случайных величин X и Y соответствует одно возможное значение случайной величины Z, то Z называют *функцией двух случайных аргументов* X и Y:

$$Z=\varphi(X, Y).$$

Далее на примерах будет показано, как найти распределение функции

$$Z=X+Y$$

по известным распределениям слагаемых. Такая задача часто встречается на практике. Например, если X — погрешность показаний измерительного прибора (распределена нормально), Y — погрешность округления показаний до ближайшего деления шкалы (распределена равномерно), то возникает задача — найти закон распределения суммы погрешностей $Z=X+Y$.

1. Пусть X и Y — дискретные независимые случайные величины. Для того чтобы составить закон распределения функции $Z=X+Y$, надо найти все возможные значения Z и их вероятности.

Пример 1. Дискретные независимые случайные величины заданы распределениями:

X	1	2
p	0,4	0,6

Y	3	4
p	0,2	0,8.

Составить распределение случайной величины $Z=X+Y$.

Решение. Возможные значения Z есть суммы каждого возможного значения X со всеми возможными значениями Y:

$$z_1=1+3=4;\quad z_2=1+4=5;\quad z_3=2+3=5;\quad z_4=2+4=6.$$

Найдем вероятности этих возможных значений. Для того чтобы $Z=4$, достаточно, чтобы величина X приняла значение $x_1=1$ и величина Y — значение $y_1=3$. Вероятности этих возможных значений, как следует из данных законов распределения, соответственно равны 0,4 и 0,2.

Так как аргументы X и Y независимы, то события $X=1$ и $Y=3$ независимы и, следовательно, вероятность их совместного наступления (т. е. вероятность события $Z=1+3=4$) по теореме умножения равна $0,4\cdot 0,2=0,08$.

Аналогично найдем:

$$P(Z=1+4=5)=0,4\cdot 0,8=0,32;$$
$$P(Z=2+3=5)=0,6\cdot 0,2=0,12;$$
$$P(Z=2+4=6)=0,6\cdot 0,8=0,48.$$

Напишем искомое распределение, сложив предварительно вероятности несовместных событий $Z=z_2$, $Z=z_3$ $(0,32+0,12=0,44)$:

Z	4	5	6
p	0,08	0,44	0,48.

Контроль: $0,08+0,44+0,48=1$.

2. Пусть X и Y — непрерывные случайные величины. Доказано: если X и Y независимы, то дифференциальная функция $g(z)$ суммы $Z=X+Y$ (при условии, что дифференциальная функция хотя бы одного из аргументов задана на интервале $(-\infty, \infty)$ одной формулой) может быть найдена по равенству

$$g(z)=\int_{-\infty}^{\infty} f_1(x)\, f_2(z-x)\, dx \qquad (*)$$

либо по равносильному равенству

$$g(z)=\int_{-\infty}^{\infty} f_1(z-y)\, f_2(y)\, dy, \qquad (**)$$

где f_1, f_2 — дифференциальные функции аргументов.

Если возможные значения аргументов неотрицательны, то $g(z)$ находят по формуле

$$g(z)=\int_{0}^{z} f_1(x)\, f_2(z-x)\, dx, \qquad (***)$$

либо по равносильной формуле

$$g(z)=\int_{0}^{z} f_1(z-y)\, f_2(y)\, dy. \qquad (****)$$

Дифференциальную функцию суммы независимых случайных величин называют *композицией*.

Закон распределения вероятностей называют *устойчивым*, если композиция таких законов есть тот же закон (отличающийся, вообще говоря, параметрами). Нормальный закон обладает свойством устойчивости: композиция нормальных законов также имеет нормальное распределение (математическое ожидание и дисперсия этой композиции равны соответственно суммам математических ожиданий и дисперсий слагаемых). Например, если X и Y — независимые случайные величины, распределенные нормально с математическими ожиданиями и дисперсиями, соответственно равными $a_1=3$, $a_2=4$, $D_1=1$, $D_2=0,5$, то композиция этих величин (т. е. дифференциальная функция суммы $Z=X+Y$) также распределена нормально, причем математическое ожидание и дисперсия композиции соответственно равны $a=3+4=7$; $D=1+0,5=1,5$.

Пример 2. Независимые случайные величины X и Y заданы дифференциальными функциями:

$$f(x)=\frac{1}{3}e^{-\frac{x}{3}} \qquad (0\leqslant x<\infty);$$
$$f(y)=\frac{1}{4}e^{-\frac{y}{4}} \qquad (0\leqslant y<\infty).$$

Найти композицию этих законов, т. е. дифференциальную функцию случайной величины $Z=X+Y$.

Решение. Так как возможные значения аргументов неотрицательны, воспользуемся формулой (***)

$$g(z)=\int_0^z f_1(x)f_2(z-x)\,dx=\int_0^z\left[\frac{1}{3}e^{-\frac{x}{3}}\right]\left[\frac{1}{4}e^{-\frac{z-x}{4}}\right]dx=$$

$$=\frac{1}{12}e^{-\frac{z}{4}}\int_0^z e^{-\frac{x}{12}}dx=e^{-\frac{z}{4}}\left(1-e^{-\frac{z}{12}}\right).$$

Заметим, что здесь $z\geqslant 0$, так как $Z=X+Y$ и по условию возможные значения X и Y неотрицательны.

Рекомендуем читателю для контроля убедиться, что

$$\int_0^\infty g(z)\,dz=1.$$

В следующих далее параграфах кратко описаны распределения, связанные с нормальным, которые будут использованы при изложении математической статистики.

§ 13. Распределение χ^2

Пусть X_i ($i=1, 2, \ldots, n$) — нормальные независимые случайные величины, причем математическое ожидание каждой из них равно нулю, а среднее квадратическое отклонение — единице. Тогда сумма квадратов этих величин

$$\chi^2=\sum_{i=1}^{n}X_i^2$$

распределена по закону χ^2 («хи квадрат») с $k=n$ степенями свободы; если же эти величины связаны одним линейным соотношением, например $\sum X_i=n\overline{X}$, то число степеней свободы $k=n-1$.

Дифференциальная функция этого распределения

$$f(x)=\begin{cases}0 & \text{при } x\leqslant 0,\\ \dfrac{1}{2^{\frac{k}{2}}\cdot\Gamma\left(\frac{k}{2}\right)}e^{-\frac{x}{2}}x^{\frac{k}{2}-1} & \text{при } x>0,\end{cases}$$

где $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$ — гамма-функция; в частности

$\Gamma(n+1)=n!$

Отсюда видно, что распределение «хи квадрат» определяется одним параметром — числом степеней свободы k.

С увеличением числа степеней свободы распределение медленно приближается к нормальному.

§ 14. Распределение Стьюдента

Пусть Z — нормальная случайная величина, причем $M(Z)=0$, $\sigma(Z)=1$, а V — независимая от Z величина, которая распределена по закону χ^2 с k степенями свободы. Тогда величина

$$T=\frac{Z}{\sqrt{\frac{V}{k}}}\qquad(*)$$

имеет распределение, которое называют t-распределением, или распределением Стьюдента (псевдоним английского статистика В. Госсета) с k степенями свободы.

Итак, отношение нормированной нормальной величины к квадратному корню из независимой случайной величины, распределенной по закону «хи квадрат» с k степенями свободы, деленной на k, распределено по закону Стьюдента с k степенями свободы.

С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормальному. Дополнительные сведения об этом распределении приведены далее (гл. XV, § 16).

§ 15. Распределение F Фишера—Снедекора

Если U и V независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2, то величина

$$F=\frac{\frac{U}{k_1}}{\frac{V}{k_2}}\qquad(*)$$

имеет распределение, которое называют распределением F Фишера — Снедекора со степенями свободы k_1 и k_2 (иногда его обозначают через V^2).

Дифференциальная функция

$$f(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ C_0 \dfrac{x^{\frac{k_1-2}{2}}}{(k_2 + k_1 x)^{\frac{k_1+k_2}{2}}} & \text{при } x > 0, \end{cases}$$

где $$C_0 = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right) \cdot k_1^{\frac{k_1}{2}} \cdot k_2^{\frac{k_2}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}.$$

Мы видим, что распределение F определяется двумя параметрами — числами степеней свободы. Дополнительные сведения об этом распределении приведены далее (гл. XVIII, § 8).

Задачи

1. Найти математическое ожидание и дисперсию случайной величины X, зная ее дифференциальную функцию:

а) $f(x) = \dfrac{1}{\pi\sqrt{1-x^2}}$ при $-1 < x < 1$, $f(x) = 0$ при остальных значениях x;

б) $f(x) = \dfrac{1}{2 \cdot l}$ при $a - l < x < a + l$, $f(x) = 0$ при остальных значениях x.

Отв. а) $M(X) = 0$, $D(X) = \dfrac{1}{2}$; б) $M(X) = a$, $DX = \dfrac{l^2}{3}$.

2. Случайная величина X распределена нормально. Математическое ожидание и среднее квадратическое отклонение этой величины соответственно равны 6 и 2. Найти вероятность того, что в результате испытания X примет значение, заключенное в интервале (4; 8).

Отв. 0,6826.

3. Случайная величина распределена нормально. Среднее квадратическое отклонение этой величины равно 0,4. Найти вероятность того, что отклонение случайной величины от ее математического ожидания по абсолютной величине будет меньше 0,3.

Отв 0,5468.

4. Случайные ошибки измерения подчинены нормальному закону со средним квадратическим отклонением $\sigma = 1$ *мм* и математическим ожиданием $a = 0$. Найти вероятность того, что из двух независимых наблюдений ошибка хотя бы одного из них не превзойдет по абсолютной величине 1,28 *мм*.

Отв 0,96.

5. Валики, изготовляемые автоматом, считаются стандартными, если отклонение диаметра валика от проектного размера не превышает 2 *мм*. Случайные отклонения диаметра валиков подчиняются нормальному закону со средним квадратическим отклонением $\sigma = 1,6$ *мм* и математическим ожиданием $a = 0$. Сколько процентов стандартных валиков изготовляет автомат?

Отв. Примерно 79%.

6. Дискретная случайная величина X задана законом распределения:

а)

X	1	2	3
p	0,2	0,1	0,7;

б)

X	–1	1	2
p	0,1	0,2	0,7

Найти закон распределения случайной величины $Y = X^4$.

Отв. а)

Y	1	16	81
p	0,2	0,1	0,7;

б)

Y	1	16
p	0,3	0,7.

7. Непрерывная случайная величина X задана дифференциальной функцией $f(x)$. Найти дифференциальную функцию $g(y)$ случайной величины Y, если:

а) $Y = X + 1$ $(-\infty < x < \infty)$;
б) $Y = 2X$ $(-a < x < a)$.

Отв а) $g(y) = f(y-1)$ $(-\infty < y < \infty)$;
б) $g(y) = \dfrac{1}{2} f\left(\dfrac{y}{2}\right)$ $(-2a < y < 2a)$.

8. Независимые дискретные случайные величины заданы следующими законами распределения:

X	2	3	5
p	0,3	0,5	0,2;

Y	1	4
p	0,2	0,8.

Найти законы распределения функций: а) $Z = X + Y$; б) $Z = XY$.

Отв. а)

Z	3	4	6	7	9
p	0,06	0,10	0,28	0,40	0,16;

б)

Z	2	3	5	8	12	20
p	0,06	0,10	0,04	0,24	0,40	0,16.

9. Независимые случайные величины X и Y заданы дифференциальными функциями:

$$f_1(x) = \frac{1}{3} e^{-\frac{x}{3}} \quad (0 < x < \infty);$$

$$f_2(x) = \frac{1}{5} e^{-\frac{y}{5}} \quad (0 < y < \infty).$$

Найти композицию этих законов, т. е. дифференциальную функцию случайной величины $Z = X + Y$.

$$Отв.\ g(z) = \begin{cases} \frac{1}{2} e^{-\frac{z}{5}} \left(1 - e^{-\frac{2z}{15}}\right) & \text{при } z \geqslant 0; \\ 0 & \text{при } z < 0. \end{cases}$$

Глава тринадцатая

ПОКАЗАТЕЛЬНОЕ РАСПРЕДЕЛЕНИЕ

§ 1. Определение показательного распределения

Показательным (*экспоненциальным*) называют распределение вероятностей, которое описывается дифференциальной функцией

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

где λ — постоянная положительная величина.

Мы видим, что показательное распределение определяется одним параметром λ. Эта особенность показательного распределения указывает на его преимущество, по сравнению с распределениями, зависящими от большего числа параметров. Обычно параметры неизвестны и приходится находить их оценки (приближенные значения); разумеется, проще оценить один параметр, чем два, или три и т. д.

Примером непрерывной случайной величины, распределенной по показательному закону, может служить время между появлениями двух последовательных событий простейшего потока (см. § 5).

Найдем интегральную функцию показательного распределения (гл. XI, § 3).

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

Мы определили показательное распределение при помощи дифференциальной функции; ясно, что его можно определить, пользуясь интегральной функцией.

Графики дифференциальной и интегральной функций изображены на рис. 12.

Пример. Написать дифференциальную и интегральную функции показательного распределения, если параметр $\lambda=8$.

Решение. Очевидно,

$$f(x)=8\,e^{-8x} \text{ при } x\geqslant 0;\ f(x)=0 \text{ при } x<0;$$
$$F(x)=1-e^{-8x}.$$

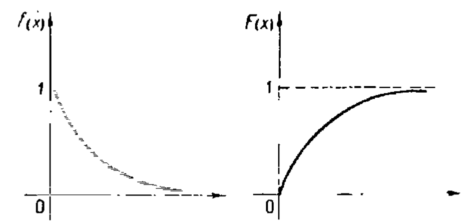

Рис. 12.

§ 2. Вероятность попадания в заданный интервал показательно распределенной случайной величины

Найдем вероятность попадания в интервал (a, b) непрерывной случайной величины X, распределенной по показательному закону, заданному интегральной функцией

$$F(x)=1-e^{-\lambda x}.$$

Воспользуемся формулой (гл. X, § 2, следствие 1)

$$P(a<X<b)=F(b)-F(a).$$

Учитывая, что $F(a)=1-e^{-a\lambda}$, $F(b)=1-e^{-b\lambda}$, получим

$$P(a<X<b)=e^{-a\lambda}-e^{-b\lambda}. \qquad (*)$$

Значения функции e^{-x} находят по таблице.

Пример. Непрерывная случайная величина X распределена по показательному закону

$$f(x)=2e^{-2x} \text{ при } x\geqslant 0;\ f(x)=0 \text{ при } x<0.$$

Найти вероятность того, что в результате испытания X попадет в интервал (0,3; 1).

Решение. По условию $\lambda=2$. Воспользуемся формулой (*):

$$P(0{,}3<X<1)=e^{-(2\cdot 0{,}3)}-e^{-(2\cdot 1)}=e^{-0{,}6}-e^{-2}=$$
$$=0{,}54881-0{,}13534\simeq 0{,}41.$$

§ 3. Числовые характеристики показательного распределения

Пусть непрерывная случайная величина X распределена по показательному закону

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

Найдем математическое ожидание (гл. XII, § 1):

$$M(X)=\int_0^\infty xf(x)\,dx=\lambda\int_0^\infty xe^{-\lambda x}\,dx.$$

Интегрируя по частям, получим

$$M(X)=\frac{1}{\lambda}. \qquad (*)$$

Таким образом, математическое ожидание показательного распределения равно обратной величине параметра λ.

Найдем дисперсию (гл. XII, § 1):

$$D(X)=\int_0^\infty x^2f(x)\,dx-[M(X)]^2=\lambda\int_0^\infty x^2e^{-\lambda x}\,dx-\frac{1}{\lambda^2}.$$

Интегрируя по частям, получим

$$\lambda\int_0^\infty x^2e^{-\lambda x}\,dx=\frac{2}{\lambda^2}.$$

Следовательно,

$$D(X)=\frac{1}{\lambda^2}.$$

Найдем среднее квадратическое отклонение, для чего извлечем квадратный корень из дисперсии:

$$\sigma(X)=\frac{1}{\lambda}. \qquad (**)$$

Сравнивая (*) и (**), заключаем, что

$$M(X)=\sigma(X)=\frac{1}{\lambda},$$

т. е. математическое ожидание и среднее квадратическое отклонение показательного распределения равны между собой.

Пример. Непрерывная случайная величина X распределена по показательному закону

$$f(x)=5e^{-5x} \text{ при } x\geqslant 0;\ f(x)=0 \text{ при } x<0.$$

Найти математическое ожидание, среднее квадратическое отклонение и дисперсию X.

Решение. По условию $\lambda=5$. Следовательно,

$$M(X)=\sigma(X)=\frac{1}{\lambda}=\frac{1}{5}=0{,}2;$$

$$D(X)=\frac{1}{\lambda^2}=\frac{1}{5^2}=0{,}04.$$

Замечание 1. Пусть на практике изучается показательно распределенная случайная величина, причем параметр λ неизвестен. Если математическое ожидание также неизвестно, то находят его оценку (приближенное значение), в качестве которой принимают выборочную среднюю $\bar{x}$ (гл. XVI, § 5). Тогда приближенное значение параметра λ находят по равенству

$$\lambda^*=\frac{1}{\bar{x}}.$$

Замечание 2. Допустим, что имеются основания предположить, что изучаемая на практике случайная величина имеет показательное распределение. Для того чтобы проверить эту гипотезу, находят оценки математического ожидания и среднего квадратического отклонения, т. е. находят выборочную среднюю и выборочное среднее квадратическое отклонение (гл. XVI, § 5, 9). Поскольку математическое ожидание и среднее квадратическое отклонение показательного распределения равны между собой, их оценки должны различаться незначительно. Если оценки окажутся близкими одна к другой, то данные наблюдений подтверждают гипотезу о показательном распределении изучаемой величины;

если же оценки различаются существенно, то гипотезу следует отвергнуть.

Показательное распределение широко применяется в приложениях, в частности, в теории надежности, одним из основных понятий которой является функция надежности.

§ 4. Функция надежности

Будем называть элементом некоторое устройство, независимо от того «простое» оно, или «сложное».

Пусть элемент начинает работать в момент времени $t_0=0$, а по истечении времени длительностью t происходит отказ.

Обозначим через T непрерывную случайную величину — длительность времени безотказной работы элемента. Если элемент проработал безотказно (до наступления отказа) время, меньшее чем t, то, следовательно, за время длительностью t наступит отказ.

Таким образом, интегральная функция

$$F(t)=P(T<t)$$

определяет вероятность отказа за время длительностью t. Следовательно, вероятность безотказной работы за это же время, длительностью t, т. е. вероятность противоположного события $T>t$, равна

$$R(t)=P(T>t)=1-F(t). \qquad (*)$$

Функцией надежности $R(t)$ называют функцию, определяющую вероятность безотказной работы элемента за время длительностью t:

$$R(t)=P(T>t).$$

§ 5. Показательный закон надежности

Часто длительность времени безотказной работы элемента имеет показательное распределение, интегральная функция которого

$$F(t)=1-e^{-\lambda t}.$$

Следовательно, в силу соотношения (*) предыдущего параграфа, функция надежности, в случае показательного

распределения времени безотказной работы элемента, имеет вид

$$R(t)=1-F(t)=1-\left(1-e^{-\lambda t}\right)=e^{-\lambda t}.$$

Показательным законом надежности называют функцию надежности, определяемую равенством

$$R(t)=e^{-\lambda t}, \qquad (*)$$

где λ — интенсивность отказов.

Как следует из определения функции надежности (§ 4), эта формула позволяет найти вероятность безотказной работы элемента на интервале времени, длительностью t, если время безотказной работы имеет показательное распределение.

Пример. Время безотказной работы элемента распределено по показательному закону $f(t)=0{,}02\, e^{-0{,}02t}$ при $t\geqslant 0$ (t — время в часах). Найти вероятность того, что элемент проработает безотказно 100 часов.

Решение. По условию постоянная интенсивность отказов $\lambda=0{,}02$. Воспользуемся формулой (*):

$$R(100)=e^{-0{,}02\cdot 100}=e^{-2}=0{,}13534.$$

Искомая вероятность того, что элемент проработает безотказно 100 *ч*, приближенно равна 0,14.

Замечание. Если отказы элементов в случайные моменты времени образуют простейший поток, то вероятность того, что за время длительностью t не наступит ни одного отказа (гл. VI, § 6)

$$P_t(0)=e^{-\lambda t},$$

что согласуется с равенством (*), поскольку λ в обеих формулах имеет один и тот же смысл (постоянная интенсивность отказов).

§ 6. Характеристическое свойство показательного закона надежности

Показательный закон надежности весьма прост и удобен для решения задач, возникающих на практике. Очень многие формулы теории надежности значительно упрощаются. Объясняется это тем, что этот закон обладает следующим важным свойством: *вероятность безотказной работы элемента на интервале времени длительностью t не зависит от времени предшествующей работы до начала рассматриваемого интервала, а зависит только от* *длительности времени* t (при заданной интенсивности отказов λ).

Для доказательства свойства введем обозначения событий:

A — безотказная работа элемента на интервале $(0,\ t_0)$ длительностью t_0;

B — безотказная работа на интервале $(t_0,\ t_0+t)$ длительностью t.

Тогда AB — безотказная работа на интервале $(0,\ t_0+t)$ длительностью t_0+t.

Найдем вероятности этих событий по формуле (*) (§ 5):

$$P(A)=e^{-\lambda t_0},\quad P(B)=e^{-\lambda t},$$

$$P(AB)=e^{-\lambda(t_0+t)}=e^{-\lambda t_0}\cdot e^{-\lambda t}.$$

Найдем условную вероятность того, что элемент будет работать безотказно на интервале $(t_0,\ t_0+t)$ при условии, что он уже проработал безотказно на предшествующем интервале $(0,\ t_0)$ (гл. III, § 5, замечание 2):

$$P_A(B)=\frac{P(AB)}{P(A)}=\frac{e^{-\lambda t_0}\cdot e^{-\lambda t}}{e^{-\lambda t_0}}=e^{-\lambda t}.$$

Мы видим, что полученная формула не содержит t_0, а содержит только t. Это и означает, что время работы на предшествующем интервале не сказывается на величине вероятности безотказной работы на последующем интервале, а зависит только от длины последующего интервала, что и требовалось доказать.

Полученный результат можно сформулировать несколько иначе. Сравнив вероятности $P(B)=e^{-\lambda t}$ и $P_A(B)=e^{-\lambda t}$, заключаем: условная вероятность безотказной работы элемента на интервале длительностью t, вычисленная в предположении, что элемент проработал безотказно на предшествующем интервале, равна безусловной вероятности.

Итак, в случае показательного закона надежности, безотказная работа элемента «в прошлом» не сказывается на величине вероятности его безотказной работы «в ближайшем будущем».

Замечание. Можно доказать, что рассматриваемым свойством обладает **только** показательное распределение. Поэтому, если на практике изучаемая случайная величина этим свойством обладает, то она распределена по показательному закону. Например, при допущении, что метеориты распределены равномерно в пространстве и во времени, — вероятность попадания метеорита в космический корабль не зависит от того, попадали или не попадали метеориты в корабль до начала рассматриваемого интервала времени. Следовательно, случайные моменты времени попадания метеоритов в космический корабль распределены по показательному закону.

Задачи

1. Написать дифференциальную и интегральную функции показательного распределения, если параметр $\lambda=5$.

Отв. $f(x)=5e^{-5x}$ при $x\geq 0$, $f(x)=0$ при $x<0$; $F(x)=1-e^{-5x}$.

2. Непрерывная случайная величина X распределена по показательному закону: $f(x)=5e^{-5x}$ при $x\geq 0$, $f(x)=0$ при $x<0$. Найти вероятность того, что в результате испытания X попадет в интервал $(0{,}4;\ 1)$.

Отв. $P(0{,}4<X<1)=0{,}13$.

3. Непрерывная случайная величина X распределена по показательному закону $f(x)=4e^{-4x}$ $(x>0)$. Найти математическое ожидание, среднее квадратическое отклонение и дисперсию X.

Отв. $M(X)=\sigma(X)=0{,}25$; $D(X)=0{,}0625$.

4. Время безотказной работы элемента распределено по показательному закону $f(t)=0{,}01\cdot e^{-0{,}01t}$ $(t>0)$, где t — время в часах. Найти вероятность того, что элемент проработает безотказно 100 ч.

Отв. $R(100)=0{,}37$.

Глава четырнадцатая

СИСТЕМА ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН

§ 1. Понятие о системе нескольких случайных величин

До сих пор рассматривались случайные величины, возможные значения которых определялись одним числом. Такие величины называют одномерными. Например, число очков, которое может выпасть при бросании игральной кости — дискретная одномерная величина; расстояние от орудия до места падения снаряда — непрерывная одномерная случайная величина.

Кроме одномерных случайных величин, изучают величины, возможные значения которых определяются двумя, тремя, ..., n числами. Такие величины называются соответственно двумерными, трехмерными, ..., n-мерными.

Будем обозначать через (X, Y) двумерную случайную величину. Каждую из величин X и Y называют составляющей (компонентой); обе величины X и Y, рассматриваемые одновременно, образуют систему двух случайных величин. Аналогично n-мерную величину можно рассматривать как систему n случайных величин. Например, трехмерная величина (X, Y, Z) определяет систему трех случайных величин X, Y и Z.

Пример. Станок-автомат штампует стальные плитки. Если контролируемыми размерами являются длина X и ширина Y, то имеем двумерную случайную величину (X, Y); если же контролируется и высота Z, то имеем трехмерную величину (X, Y, Z).

Двумерную случайную величину (X, Y) геометрически можно истолковать либо как случайную точку $M(X, Y)$ на плоскости (т. е. как точку со случайными координатами), либо как случайный вектор $\overline{OM}$. Трехмерную случайную величину геометрически можно истолковать как точку $M(X, Y, Z)$ в трехмерном пространстве, или как вектор $\overline{OM}$.

Целесообразно различать дискретные (составляющие этих величин дискретны) и непрерывные (составляющие этих величин непрерывны) многомерные случайные величины.

§ 2. Закон распределения вероятностей дискретной двумерной случайной величины

Законом распределения дискретной двумерной случайной величины называют перечень возможных значений этой величины (т. е. пар чисел (x_i, y_j) и их вероятностей $p(x_i, y_j)$ $(i=1, 2, ..., n;\ j=1, 2, ..., m)$. Обычно закон распределения задают в виде таблицы с двойным входом (табл. 2).

Первая строка таблицы содержит все возможные значения составляющей X, а первый столбец — все возможные значения составляющей Y. В клетке, стоящей на пересечении «столбца x_i» и «строки y_j», указана вероятность $p(x_i, y_j)$ того, что двумерная случайная величина примет значение (x_i, y_j).

Таблица 2

Y \ X	x_1	x_2	...	x_i	...	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$		$p(x_i, y_1)$		$p(x_n, y_1)$
...	...	...	...	...	...	...
y_j	$p(x_1, y_j)$	$p(x_2, y_j)$	...	$p(x_i, y_j)$	...	$p(x_n, y_j)$
...	...	...	...	...	...	...
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	...	$p(x_i, y_m)$	...	$p(x_n, y_m)$

Так как события $(X=x_i, Y=y_j)$, $(i=1, 2, ..., n;\ j=1, 2, ..., m)$ образуют полную группу (гл. II, § 2), то сумма вероятностей, помещенных во всех клетках таблицы, равна единице.

Зная закон распределения двумерной дискретной случайной величины, можно найти законы распределения каждой из составляющих. Действительно, например события $(X=x_1;\ Y=y_1)$, $(X=x_1;\ Y=y_2)$, ..., $(X=x_1;\ Y=y_m)$ несовместны, поэтому вероятность $P(x_1)$ того, что X примет значение x_1, по теореме сложения такова:

$$P(x_1)=p(x_1, y_1)+p(x_1, y_2)+ \ldots +p(x_1, y_m).$$

Таким образом, вероятность того, что X примет значение x_1, равна сумме вероятностей «столбца x_1». В общем случае для того, чтобы найти вероятность $P(X=x_i)$, надо просуммировать вероятности столба x_i. Аналогично сложив вероятности «строки y_j», получим вероятность $P(Y=y_j)$.

Пример. Найти законы распределения составляющих двумерной случайной величины, заданной законом распределения (табл. 3).

Таблица 3

Y \ X	x_1	x_2	x_3
y_1	0,10	0,30	0,20
y_2	0,06	0,18	0,16

Решение. Сложив вероятности по столбцам, получим вероятности возможных значений X: $p(x_1)=0,16$;

$p(x_2)=0{,}48$; $p(x_3)=0{,}36$. Напишем закон распределения составляющей X:

X	x_1	x_2	x_3
p	0,16	0,48	0,36.

Контроль: $0{,}16+0{,}48+0{,}36=1$.

Сложив вероятности по строкам, получим вероятности возможных значений Y: $p(y_1)=0{,}60$; $p(y_2)=0{,}40$. Напишем закон распределения составляющей Y:

Y	y_1	y_2
p	0,60	0,40.

Контроль: $0{,}60+0{,}40=1$.

§ 3. Интегральная функция распределения двумерной случайной величины

Рассмотрим двумерную случайную величину (X, Y) (безразлично дискретную или непрерывную). Пусть x, y — пара действительных чисел. Вероятность события, состоящего в том, что X примет значение, меньше x, и при этом Y примет значение, меньшее y, обозначим через $F(x, y)$. Если x и y будут изменяться, то, вообще говоря, будет изменяться и $F(x, y)$, т. е. $F(x, y)$ есть функция от x и y.

Интегральной функцией распределения двумерной случайной величины (X, Y) называют функцию $F(x, y)$, определяющую для каждой пары чисел x, y вероятность того, что X примет значение, меньшее x и при этом Y примет значение, меньшее y:

$$F(x, y)=P(X<x, \ Y<y).$$

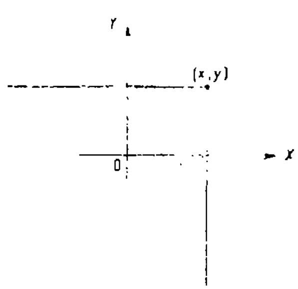

Рис. 13.

Геометрически это равенство можно истолковать так: $F(x, y)$ есть вероятность того, что случайная точка (X, Y) попадет в бесконечный квадрант с вершиной (x, y), расположенный левее и ниже этой вершины (рис. 13).

Пример. Найти вероятность того, что в результате испытания составляющая X двумерной случайной величины (X, Y) примет значение $X<2$ и при этом составляющая Y примет значение $Y<3$, если известна интегральная функция системы

$$F(x, y)=\left(\frac{1}{\pi}\operatorname{arctg}\frac{x}{2}+\frac{1}{2}\right)\cdot\left(\frac{1}{\pi}\operatorname{arctg}\frac{y}{3}+\frac{1}{2}\right).$$

Решение. По определению интегральной функции двумерной случайной величины

$$F(x, y)=P(X<x, \ Y<y).$$

Положив $x=2$, $y=3$, получим искомую вероятность

$$P(X<2, Y<3)=F(2, 3)=\left(\frac{1}{\pi}\operatorname{arctg}\frac{2}{2}+\frac{1}{2}\right)\times$$

$$\times\left(\frac{1}{\pi}\operatorname{arctg}\frac{3}{3}+\frac{1}{2}\right)=\left(\frac{1}{\pi}\cdot\frac{\pi}{4}+\frac{1}{2}\right)\cdot\left(\frac{1}{\pi}\cdot\frac{\pi}{4}+\frac{1}{2}\right)=$$

$$=\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{16}.$$

§ 4. Свойства интегральной функции двумерной случайной величины

Свойство 1. *Значения интегральной функции удовлетворяют двойному неравенству*

$$0\leqslant F(x, y)\leqslant 1.$$

Доказательство. Свойство вытекает из определения интегральной функции как вероятности: вероятность всегда неотрицательное число, не превышающее единицу.

Свойство 2. *$F(x, y)$ есть неубывающая функция по каждому аргументу*, т. е.

$$F(x_2, y)\geqslant F(x_1, y), \text{ если } x_2>x_1;$$
$$F(x, y_2)\geqslant F(x, y_1), \text{ если } y_2>y_1.$$

Доказательство. Докажем, что $F(x, y)$ есть неубывающая функция по аргументу x. Событие, состоящее в том, что составляющая примет значение, меньшее x_2, и при этом составляющая $Y<y$, можно подразделить на следующие два несовместных события:

1) X примет значение, меньшее x_1, и при этом $Y<y$ с вероятностью $P(X<x_1, Y<y)$;

2) X примет значение, удовлетворяющее неравенству $x_1 \leqslant X < x_2$, и при этом $Y < y$ с вероятностью $P(x_1 \leqslant X < x_2,\ Y < y)$.

По теореме сложения имеем

$$P(X < x_2,\ Y < y) = P(X < x_1,\ Y < y) + P(x_1 \leqslant X < x_2,\ Y < y).$$

Отсюда

$$P(X < x_2,\ Y < y) - P(X < x_1,\ Y < y) = P(x_1 \leqslant X < x_2,\ Y < y),$$

или

$$F(x_2,\ y) - F(x_1,\ y) = P(x_1 \leqslant X < x_2,\ Y < y).$$

Так как любая вероятность есть число неотрицательное, то

$$F(x_2,\ y) - F(x_1,\ y) \geqslant 0,$$

или

$$F(x_2,\ y) \geqslant F(x_1,\ y),$$

что и требовалось доказать.

Свойство становится наглядно ясным, если воспользоваться геометрическим истолкованием интегральной функции как вероятности попадания случайной точки в бесконечный квадрант с вершиной (x, y) (рис. 13). При возрастании x правая граница этого квадранта сдвигается вправо; при этом вероятность попадания случайной точки в «новый» квадрант, очевидно, не может уменьшиться.

Аналогично доказывается, что $F(x, y)$ есть неубывающая функция по аргументу y.

Свойство 3. *Имеют место предельные соотношения:*

1) $F(-\infty,\ y) = 0,$
2) $F(x,\ -\infty) = 0,$
3) $F(-\infty,\ -\infty) = 0,$
4) $F(\infty,\ \infty) = 1.$

Доказательство. 1) $F(-\infty, y)$ есть вероятность события $X < -\infty$ и $Y < y$; но такое событие невозможно (поскольку невозможно событие $X < -\infty$), следовательно, вероятность этого события равна нулю.

Свойство становится наглядно ясным, если прибегнуть к геометрической интерпретации: при $x \to -\infty$ правая граница бесконечного квадранта (рис. 13) неограниченно сдвигается влево и при этом вероятность попадания случайной точки в квадрант стремится к нулю.

2) Событие $Y < -\infty$ невозможно, поэтому $F(x, -\infty) = 0$.

3) Событие $X < -\infty$ и $Y < -\infty$ невозможно, поэтому $F(-\infty,\ -\infty) = 0$.

4) Событие $X < \infty$ и $Y < \infty$ достоверно, следовательно, вероятность этого события $F(\infty,\ \infty) = 1$.

Свойство становится наглядно ясным, если принять во внимание, что при $x \to \infty$ и $y \to \infty$ бесконечный квадрант (рис. 13) превращается во всю плоскость XOY и, следовательно, попадание случайной точки (X, Y) в эту плоскость в результате испытания есть достоверное событие.

Свойство 4. а) *При $y = \infty$ интегральная функция системы становится интегральной функцией составляющей X:*

$$F(x,\ \infty) = F_1(x).$$

б) *При $x = \infty$ интегральная функция системы становится интегральной функцией составляющей Y:*

$$F(\infty,\ y) = F_2(y).$$

Доказательство. а) Так как событие $Y < \infty$ достоверно, то $F(x,\ \infty)$ определяет вероятность события $X < x$, т. е. представляет собой интегральную функцию составляющей X.

б) Доказывается аналогично.

§ 5. Вероятность попадания случайной точки в полуполосу

Пользуясь интегральной функцией системы случайных величин X и Y, легко найти вероятность того, что в результате испытания случайная точка попадает в полуполосу $x_1 < X < x_2$ и $Y < y$ (рис. 14, *а*), или в полуполосу $X < x$ и $y_1 < Y < y_2$ (рис. 14, *б*).

Вычитая из вероятности попадания случайной точки в квадрант с вершиной (x_2, y) вероятность попадания точки в квадрант с вершиной (x_1, y) (рис. 14, *а*) получим

$$P(x_1 < X < x_2,\ Y < y) = F(x_2,\ y) - F(x_1,\ y).$$

Аналогично имеем

$$P(X < x,\ y_1 < Y < y_2) = F(x,\ y_2) - F(x,\ y_1).$$

Таким образом, вероятность попадания случайной точки в полуполосу равна приращению интегральной функции по одному из аргументов.

§ 6. Вероятность попадания случайной точки в прямоугольник

Рассмотрим прямоугольник $ABCD$ со сторонами, параллельными координатным осям (рис. 15). Пусть уравнения сторон таковы:

$$X=x_1,\quad X=x_2,\quad Y=y_1 \text{ и } Y=y_2.$$

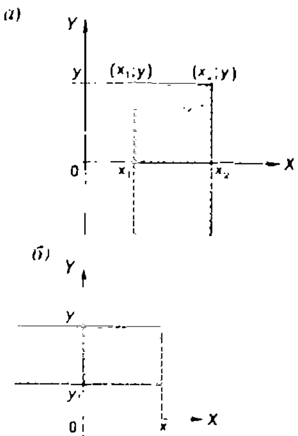

Рис. 14.

Найдем вероятность попадания случайной точки (X, Y) в этот прямоугольник. Искомую вероятность можно найти, например, так: из вероятности попадания случайной точки в полуполосу AB с вертикальной штриховкой (эта вероятность равна $F(x_2, y_2)-F(x_1, y_2)$) вычесть вероятность попадания точки в полуполосу CD с горизонтальной штриховкой (эта вероятность равна $F(x_2, y_1)-F(x_1, y_1)$):

$$P(x_1<X<x_2,\ y_1<Y<y_2) = [F(x_2, y_2)-F(x_1, y_2)]-[F(x_2, y_1)-F(x_1, y_1)] \quad (*)$$

Пример. Найти вероятность попадания случайной точки (X, Y) в прямоугольник, ограниченный прямыми $x=\frac{\pi}{6}$, $x=\frac{\pi}{2}$, $y=\frac{\pi}{4}$, $y=\frac{\pi}{3}$, если известна интегральная функция

$$F(x, y)=\sin x\cdot\sin y\left(0\leqslant x\leqslant\frac{\pi}{2},\ 0\leqslant y\leqslant\frac{\pi}{2}\right).$$

Решение. Положив $x_1=\frac{\pi}{6}$, $x_2=\frac{\pi}{2}$, $y_1=\frac{\pi}{4}$,

$y=\frac{\pi}{3}$ в формуле (*), получим:

$$P\left(\frac{\pi}{6}<X<\frac{\pi}{2},\ \frac{\pi}{4}<Y<\frac{\pi}{3}\right)=\left[F\left(\frac{\pi}{2},\ \frac{\pi}{3}\right)-F\left(\frac{\pi}{6},\ \frac{\pi}{3}\right)\right]-\left[F\left(\frac{\pi}{2},\ \frac{\pi}{4}\right)-F\left(\frac{\pi}{6},\ \frac{\pi}{4}\right)\right]=$$

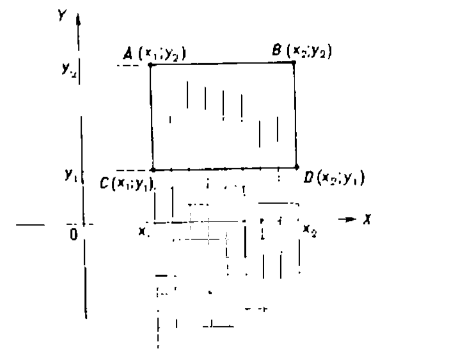

Рис. 15.

$$=\left[\sin\frac{\pi}{2}\cdot\sin\frac{\pi}{3}-\sin\frac{\pi}{6}\cdot\sin\frac{\pi}{3}\right]-\left[\sin\frac{\pi}{2}\cdot\sin\frac{\pi}{4}-\sin\frac{\pi}{6}\cdot\sin\frac{\pi}{4}\right]=\left[\frac{\sqrt{3}}{2}-\frac{1}{2}\cdot\frac{\sqrt{3}}{2}\right]-\left[\frac{\sqrt{2}}{2}-\frac{1}{2}\cdot\frac{\sqrt{2}}{2}\right]=$$

$$=\frac{\sqrt{3}-\sqrt{2}}{4}=0{,}08.$$

§ 7. Дифференциальная функция непрерывной двумерной случайной величины (двумерная плотность вероятности)

Мы задавали двумерную случайную величину при помощи интегральной функции. Непрерывную двумерную величину можно также задать, пользуясь дифференциальной функцией распределения. Здесь и далее мы будем предполагать, что интегральная функция всюду непрерывна и имеет всюду (за исключением, быть может, конечного числа кривых) непрерывную смешанную частную производную второго порядка.

Дифференциальной функцией распределения $f(x, y)$ двумерной непрерывной случайной величины (X, Y) называют вторую смешанную частную производную от интегральной функции:

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

Геометрически эту функцию можно истолковать как поверхность, которую называют *поверхностью распределения.*

Пример. Найти дифференциальную функцию $f(x, y)$ системы случайных величин (X, Y) по известной интегральной функции

$$F(x, y) = \sin x \cdot \sin y \quad \left(0 \leqslant x \leqslant \frac{\pi}{2},\quad 0 \leqslant y \leqslant \frac{\pi}{2}\right).$$

Р е ш е н и е. По определению дифференциальной функции системы случайных величин

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

Найдем частную производную по x от интегральной функции

$$\frac{\partial F}{\partial x} = \cos x \cdot \sin y.$$

Найдем от полученного результата частную производную по y, в итоге чего получим искомую дифференциальную функцию

$$f(x, y) = \frac{\partial^2 F}{\partial x\,\partial y} = \cos x \cdot \cos y \left(0 \leqslant x \leqslant \frac{\pi}{2},\ 0 \leqslant y \leqslant \frac{\pi}{2}\right).$$

§ 8. Нахождение интегральной функции распределения по известной дифференциальной функции

Зная дифференциальную функцию $f(x, y)$, можно найти интегральную функцию $F(x, y)$ по формуле

$$F(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x, y)\,dx\,dy,$$

что непосредственно следует из определения дифференциальной функции.

Пример. Найти интегральную функцию распределения двумерной случайной величины по данной дифференциальной функции $f(x, y) = \dfrac{1}{\pi^2(1 + x^2)(1 + y^2)}$.

Р е ш е н и е. Воспользуемся формулой

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

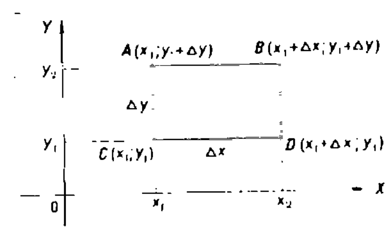

Рис. 16.

Положив здесь $f(x, y) = \dfrac{1}{\pi^2(1 + x^2)(1 + y^2)}$, получим

$$F(x, y) = \frac{1}{\pi^2}\int_{-\infty}^{y}\left(\frac{1}{1 + y^2}\int_{-\infty}^{x}\frac{dx}{1 + x^2}\right)dy =$$

$$= \frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1 + y^2}\left(\operatorname{arctg} x + \frac{\pi}{2}\right)dy = \left(\frac{1}{\pi}\operatorname{arctg} x + \right.$$

$$\left. + \frac{1}{2}\right)\frac{1}{\pi}\int_{-\infty}^{y}\frac{dy}{1 + y^2} = \left(\frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}\right)\left(\frac{1}{\pi}\operatorname{arctg} y + \frac{1}{2}\right).$$

§ 9. Вероятностный смысл дифференциальной функции двумерной случайной величины

Вероятность попадания случайной точки (X, Y) в прямоугольник $ABCD$ (рис. 16) равна (§ 6)

$$P(x_1 < X < x_2,\ y_1 < Y < y_2) = [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)].$$

Обозначив для краткости левую часть равенства через P_{ABCD} и применив к правой части теорему Лагранжа, получим

$$P_{ABCD} = F''_{xy}(\xi, \eta) \cdot \Delta x \cdot \Delta y,$$

где

$$x_1 < \xi < x_2, \quad \Delta x = x_2 - x_1,$$
$$y_1 < \eta < y_2, \quad \Delta y = y_2 - y_1.$$

Отсюда

$$F''_{xy}(\xi, \eta) = \frac{P_{ABCD}}{\Delta x \cdot \Delta y}, \qquad (*)$$

или

$$f(\xi, \eta) = \frac{P_{ABCD}}{\Delta x \cdot \Delta y}. \qquad (**)$$

Приняв во внимание, что произведение $\Delta x \cdot \Delta y$ равно площади прямоугольника $ABCD$, заключаем: $f(\xi, \eta)$ есть отношение вероятности попадания случайной точки в прямоугольник $ABCD$ к площади этого прямоугольника.

Перейдем теперь в равенстве (**) к пределу при $\Delta x \to 0$ и $\Delta y \to 0$. Тогда $\xi \to x$, $\eta \to y$ и, следовательно, $f(\xi, \eta) \to f(x, y)$.

Итак, функцию $f(x, y)$ можно рассматривать как предел отношения вероятности попадания случайной точки в прямоугольник (со сторонами Δx и Δy) к площади этого прямоугольника, когда обе стороны прямоугольника стремятся к нулю.

§ 10. Вероятность попадания случайной точки в произвольную область

Перепишем соотношение (**) § 9 так:

$$f(\xi, \eta) \cdot \Delta x \cdot \Delta y = P_{ABCD}.$$

Отсюда заключаем: произведение $f(\xi, \eta) \cdot \Delta x \cdot \Delta y$ есть вероятность попадания случайной точки в прямоугольник со сторонами Δx и Δy.

Пусть в плоскости XOY дана произвольная область D. Обозначим событие, состоящее в попадании случайной точки в эту область, так: $(X, Y) \subset D$.

Разобьем область D на n элементарных областей прямыми, параллельными оси OY, находящимися на расстоянии Δx одна от другой и прямыми, параллельными оси OX, находящимися на расстоянии Δy одна от другой (рис. 17) (для простоты предполагается, что эти прямые пересекают контур области не более, чем в двух точках).

Так как события, состоящие в попадании случайной точки в элементарные области, несовместны, то вероятность попадания в область D приближенно (сумма элементарных областей приближенно равна области D!) равна сумме вероятностей попаданий точки в элементарные области:

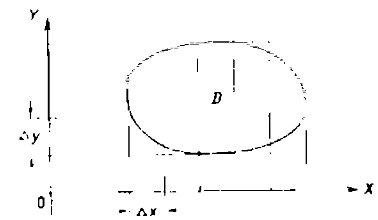

Рис. 17.

$$P((X, Y) \subset D) \approx \sum_{i=1}^{n} f(\xi_i, \eta_i) \cdot \Delta x \cdot \Delta y.$$

Переходя к пределу при $\Delta x \to 0$ и $\Delta y \to 0$, получим

$$P((X, Y) \subset D) = \iint\limits_{(D)} f(x, y)\, dx\, dy. \qquad (*)$$

Итак, для того чтобы вычислить вероятность попадания случайной точки (X, Y) в область D, достаточно найти двойной интеграл по области D от дифференциальной функции.

Геометрически равенство (*) можно истолковать так: вероятность попадания случайной точки (X, Y) в область D равна объему тела, ограниченного сверху поверхностью $z = f(x, y)$, основанием которого служит проекция этой поверхности на плоскость XOY.

Замечание. Подынтегральное выражение $f(x, y)dxdy$ называют *элементом вероятности*. Как следует из предыдущего, элемент вероятности определяет вероятность попадания случайной точки в элементарный прямоугольник со сторонами dx и dy.

Пример. Дана дифференциальная функция двумерной случайной величины

$$f(x, y) = \frac{1}{\pi^2 (1 + x^2)(1 + y^2)}.$$

Найти вероятность попадания случайной точки в прямоугольник (рис. 18) с вершинами K (1; 1), L ($\sqrt{3}$; 1), M (1; 0) и N ($\sqrt{3}$; 0).

Решение. Искомая вероятность

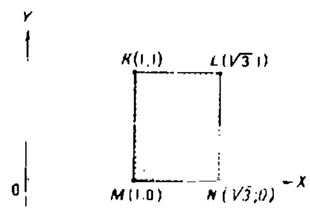

Рис. 18.

$$P((X, Y) \subset D) = \int\limits_{(D)}\int \frac{1}{\pi^2 (1 + x^2)(1 + y^2)}\, dx\, dy =$$

$$= \frac{1}{\pi^2} \int\limits_0^1 \left[\frac{1}{1 + y^2} \int\limits_1^{\sqrt{3}} \frac{dx}{1 + x^2} \right] dy = \frac{1}{\pi^2} \cdot \operatorname{arctg} x \Big|_1^{\sqrt{3}} \int\limits_0^1 \frac{dy}{1 + y^2} =$$

$$= \frac{1}{\pi^2} \left(\frac{\pi}{3} - \frac{\pi}{4} \right) \cdot \operatorname{arctg} y \Big|_0^1 = \frac{1}{\pi^2} \cdot \frac{\pi}{12} \cdot \frac{\pi}{4} = \frac{1}{48}.$$

§ 11. Свойства дифференциальной функции двумерной случайной величины

Свойство 1. *Дифференциальная функция неотрицательна*:

$$f(x, y) \geqslant 0.$$

Доказательство. Вероятность попадания случайной точки в прямоугольник со сторонами Δx и Δy есть неотрицательное число; площадь этого прямоугольника — положительное число. Следовательно, отношение этих двух чисел, а значит, и их предел (при $\Delta x \to 0$ и $\Delta y \to 0$), который равен $f(x, y)$ (§ 9), есть неотрицательное число, т. е.

$$f(x, y) \geqslant 0.$$

Заметим, что свойство непосредственно следует из того, что $F(x, y)$ есть неубывающая функция своих аргументов (§ 4).

Свойство 2. *Двойной несобственный интеграл с бесконечными пределами от дифференциальной функции равен единице*:

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

Доказательство. Бесконечные пределы интегрирования указывают, что областью интегрирования служит вся плоскость xOy; поскольку событие, состоящее в том, что случайная точка попадет при испытании на плоскость xOy достоверно, то вероятность этого события (она и определяется двойным несобственным интегралом от дифференциальной функции) равна единице, т. е.

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

§ 12. Отыскание дифференциальных функций составляющих двумерной случайной величины

Пусть известна дифференциальная функция системы двух случайных величин. Поставим своей задачей — найти дифференциальные функции каждой из составляющих.

Найдем сначала дифференциальную функцию $f_1(x)$ составляющей X. Обозначим через $F_1(x)$ интегральную функцию составляющей X. По определению дифференциальной функции одномерной случайной величины

$$f_1(x) = \frac{dF_1(x)}{dx}.$$

Приняв во внимание соотношения

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\,dx\,dy \qquad (\S\ 8),$$

$$F_1(x) = F(x, \infty) \qquad (\S\ 4),$$

найдем

$$F_1(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y)\,dx\,dy.$$

Продифференцировав обе части этого равенства по x, получим

$$\frac{dF_1}{dx} = \int_{-\infty}^{\infty} f(x, y)\,dy,$$

или

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\,dy. \qquad (*)$$

Аналогично находится дифференциальная функция составляющей Y:

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \qquad (**)$$

Итак, *дифференциальная функция одной из составляющих равна несобственному интегралу с бесконечными пределами от дифференциальной функции системы, причем переменная интегрирования соответствует другой составляющей.*

Пример. Двумерная случайная величина (X, Y) задана дифференциальной функцией

$$f(x, y) = \begin{cases} \dfrac{1}{6\pi} & \text{при } \dfrac{x^2}{9} + \dfrac{y^2}{4} < 1, \\ 0 & \text{при } \dfrac{x^2}{9} + \dfrac{y^2}{4} > 1. \end{cases}$$

Найти дифференциальные функции составляющих X и Y.

Решение. Найдем дифференциальную функцию составляющей X по формуле (*)

$$f_1(x) = \int_{-\infty}^{\infty} \frac{1}{6\pi}\,dy = \frac{1}{6\pi} \int_{-2\sqrt{1-\frac{x^2}{9}}}^{2\sqrt{1-\frac{x^2}{9}}} dy = \frac{2}{6\pi} \int_{0}^{2\sqrt{1-\frac{x^2}{9}}} dy = \frac{2}{9\pi}\sqrt{9-x^2}.$$

Итак,

$$f_1(x) = \begin{cases} \dfrac{2}{9\pi}\sqrt{9-x^2} & \text{при } |x| < 3, \\ 0 & \text{при } |x| \geqslant 3. \end{cases}$$

Аналогично, пользуясь формулой (**), найдем дифференциальную функцию составляющей Y:

$$f_2(y) = \begin{cases} \dfrac{1}{2\pi}\sqrt{4-y^2} & \text{при } |y| < 2, \\ 0 & \text{при } |y| \geqslant 2. \end{cases}$$

Рекомендуем читателю для контроля самостоятельно убедиться в том, что найденные функции удовлетворяют соотношениям

$$\int_{-\infty}^{\infty} f_1(x)\,dx = 1 \quad \text{и} \quad \int_{-\infty}^{\infty} f_2(y)\,dy = 1.$$

§ 13. Условные законы распределения составляющих системы дискретных случайных величин

Мы установили, что если события A и B зависимы, то условная вероятность события B отличается от его безусловной вероятности. В этом случае (гл. III, § 5, замечание 2)

$$P_A(B) = \frac{P(AB)}{P(A)}. \qquad (*)$$

Аналогичное положение имеет место и для случайных величин. Для того чтобы охарактеризовать зависимость между составляющими двумерной случайной величины, введем понятие условного распределения.

Рассмотрим дискретную двумерную случайную величину (X, Y). Пусть возможные значения составляющих таковы

$$x_1, x_2, \ldots, x_n; \; y_1, y_2, \ldots, y_m.$$

Допустим, что в результате испытания величина Y приняла значение $Y=y_1$; при этом X примет одно из своих возможных значений x_1 или x_2, ..., или x_n. Обозначим условную вероятность того, что X примет, например значение x_1 при условии, что $Y=y_1$, через $p(x_1|y_1)$. Эта вероятность, вообще говоря, не будет равна безусловной вероятности $p(x_1)$.

В общем случае условные вероятности составляющей будем обозначать так:

$$p(x_i|y_j)(i=1, 2, \ldots, n; \; j=1, 2, \ldots, m).$$

Условным распределением составляющей X при $Y=y_j$ называют совокупность условных вероятностей

$$p(x_1|y_j), \; p(x_2|y_j), \ldots, p(x_n|y_j),$$

вычисленных в предположении, что событие $Y=y_j$ (j имеет одно и то же значение при всех значениях X) уже наступило.

Аналогично определяется условное распределение составляющей Y.

Зная закон распределения двумерной дискретной случайной величины, можно, пользуясь формулой (*), вычислить условные законы распределения составляющих. Например, условный закон распределения X, в предположении, что событие $Y=y_1$ уже произошло, может быть найден по формуле

$$p(x_i|y_1) = \frac{p(x_i, y_1)}{p(y_1)}, \quad (i=1, 2, \ldots, n).$$

В общем случае условные законы распределения составляющей X определяются соотношением

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}. \qquad (**)$$

Аналогично находят условные законы распределения составляющей Y:

$$p(y_j|x_i) = \frac{p(x_i, y_j)}{p(x_i)}. \qquad (***)$$

Замечание. Сумма вероятностей условного распределения равна единице. Действительно, так как при фиксированном y_j имеем (§ 2) $\sum_{i=1}^{n} p(x_i, y_j) = p(y_j)$, то

$$\sum_{i=1}^{n} p(x_i|y_j) = \sum_{i=1}^{n} \frac{p(x_i, y_j)}{p(y_j)} = \frac{p(y_j)}{p(y_j)} = 1.$$

Аналогично доказывается, что при фиксированном x_i

$$\sum_{j=1}^{m} p(y_j|x_i) = 1.$$

Это свойство условных распределений используют для контроля вычислений.

Пример. Двумерная случайная величина задана таблицей 4.

Таблица 4

Y \ X	x_1	x_2	x_3
y_1	0,10	0,30	0,20
y_2	0,06	0,18	0,16

Найти условный закон распределения составляющей X при условии, что составляющая Y приняла значение y_1.

Решение. Искомый закон определяется совокупностью следующих условных вероятностей:

$$p(x_1|y_1), \quad p(x_2|y_1), \quad p(x_3|y_1).$$

Воспользовавшись формулой (*), и приняв во внимание, что $p(y_1)=0{,}60$ (стр. 156), имеем:

$$p(x_1|y_1) = \frac{p(x_1, y_1)}{p(y_1)} = \frac{0{,}10}{0{,}60} = \frac{1}{6};$$

$$p(x_2|y_1) = \frac{p(x_2, y_1)}{p(y_1)} = \frac{0{,}30}{0{,}60} = \frac{1}{2};$$

$$p(x_3|y_1) = \frac{p(x_3, y_1)}{p(y_1)} = \frac{0{,}20}{0{,}60} = \frac{1}{3}.$$

Сложив для контроля найденные условные вероятности, убедимся, что их сумма равна единице, как и должно быть (в соответствии с замечанием на стр. 171):

$$\frac{1}{6}+\frac{1}{2}+\frac{1}{3}=1.$$

§ 14. Условные законы распределения составляющих системы непрерывных случайных величин

Пусть (X, Y) — непрерывная двумерная случайная величина. *Условной дифференциальной функцией* $\varphi(x|y)$ составляющей X при данном значении $Y=y$ называют отношение дифференциальной функции $f(x, y)$ системы к дифференциальной функции $f_2(y)$ составляющей Y:

$$\varphi(x|y)=\frac{f(x, y)}{f_2(y)}. \qquad (*)$$

Подчеркнем, что отличие условной функции $\varphi(x|y)$ от безусловной дифференциальной функции $f_1(x)$ состоит в том, что $\varphi(x|y)$ дает распределение X при условии, что составляющая Y приняла значение $Y=y$; функция же $f_1(x)$ дает распределение X независимо от того, какие из возможных значений приняла составляющая Y.

Аналогично определяется условная дифференциальная функция составляющей Y при данном значении $X=x$:

$$\psi(y|x)=\frac{f(x, y)}{f_1(x)}. \qquad (**)$$

Если известна дифференциальная функция $f(x, y)$ системы, то условные дифференциальные функции составляющих могут быть найдены, в силу (*) и (**) (стр. 168), по формулам:

$$\varphi(x|y)=\frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y)\,dx}, \qquad (***)$$

$$\psi(y|x)=\frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y)\,dy}. \qquad (****)$$

Запишем формулы (*) и (**) в виде

$$f(x, y)=f_2(y)\cdot\varphi(x|y),$$
$$f(x, y)=f_1(x)\cdot\psi(y|x).$$

Отсюда заключаем: умножая закон распределения одной из составляющих на условный закон распределения другой составляющей, найдем закон распределения системы случайных величин.

Как и любая дифференциальная функция, условные дифференциальные функции обладают следующими свойствами:

$$\varphi(x|y)\geqslant 0, \qquad \int\limits_{-\infty}^{\infty}\varphi(x|y)\,dx=1;$$

$$\psi(y|x)\geqslant 0, \qquad \int\limits_{-\infty}^{\infty}\psi(y|x)\,dy=1.$$

Пример. Двумерная случайная величина (X, Y) задана дифференциальной функцией

$$f(x, y)=\begin{cases}\dfrac{1}{\pi r^2} & \text{при } x^2+y^2<r^2,\\ 0 & \text{при } x^2+y^2>r^2.\end{cases}$$

Найти условные дифференциальные законы распределения вероятностей составляющих.

Решение. Найдем условную дифференциальную функцию составляющей X по формуле (***):

$$\varphi(x|y)=\frac{f(x, y)}{\int\limits_{-\infty}^{\infty} f(x, y)\,dx}=\frac{\frac{1}{\pi r^2}}{\frac{1}{\pi r^2}\int\limits_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} dx}=\frac{1}{2\sqrt{r^2-y^2}}$$

при $|x|<\sqrt{r^2-y^2}$.

Так как $f(x, y)=0$ при $x^2+y^2>r^2$, то $\varphi(x|y)=0$ при $|x|>\sqrt{r^2-y^2}$.

Пользуясь формулой (****), аналогично найдем условную дифференциальную функцию составляющей Y:

$$\psi(y \mid x)=\begin{cases}\dfrac{1}{2\sqrt{r^2-x^2}} & \text{при } |y|<\sqrt{r^2-x^2}, \\ 0 & \text{при } |y|>\sqrt{r^2-x^2}.\end{cases}$$

§ 15. Условное математическое ожидание

Важной характеристикой условного распределения вероятностей является условное математическое ожидание.

Условным математическим ожиданием дискретной случайной величины Y при $X=x$ (x — определенное возможное значение X) называют произведение возможных значений Y на их условные вероятности:

$$M(Y \mid X=x)=\sum_{j=1}^{m} y_j\, p(y_j \mid x). \qquad (*)$$

Для непрерывных величин

$$M(Y \mid X=x)=\int_{-\infty}^{\infty} y \cdot \psi(y \mid x)\, dy,$$

где $\psi(y|x)$ — условная дифференциальная функция случайной величины Y при $X=x$.

Аналогично определяется условное математическое ожидание случайной величины X.

Пример. Дискретная двумерная случайная величина задана таблицей 5.

Таблица 5

Y \ X	$x_1=1$	$x_2=3$	$x_3=4$	$x_4=8$
$y_1=3$	0,15	0,06	0,25	0,04
$y_2=6$	0,30	0,10	0,03	0,07

Найти условное математическое ожидание составляющей Y при $X=x_1=1$.

Решение. Найдем $p(x_1)$, для чего сложим вероятности, помещенные в первом столбце табл. 5.

$$p(x_1)=0{,}15+0{,}30=0{,}45.$$

Найдем условное распределение вероятностей величины Y при $X=x_1=1$ (§ 13):

$$p(y_1 \mid x_1)=\frac{p(x_1, y_1)}{p(x_1)}=\frac{0{,}15}{0{,}45}=\frac{1}{3};$$

$$p(y_2 \mid x_1)=\frac{p(x_2, y_2)}{p(x_1)}=\frac{0{,}30}{0{,}45}=\frac{2}{3}.$$

Найдем искомое условное математическое ожидание по формуле (*):

$$M(Y \mid X=x_1)=\sum_{j=1}^{2} y_j\, p(y_j \mid x_1)=y_1 \cdot p(y_1 \mid x_1)+$$
$$+y_2 \cdot p(y_2 \mid x_1)=3 \cdot \frac{1}{3}+6 \cdot \frac{2}{3}=5.$$

§ 16. Зависимые и независимые случайные величины

Мы назвали две случайные величины независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. Из этого определения следует, что условные распределения независимых величин равны их безусловным распределениям.

Выведем необходимые и достаточные условия независимости случайных величин.

Теорема. *Для того чтобы случайные величины X и Y были независимыми, необходимо и достаточно, чтобы интегральная функция системы (X, Y) была равна произведению интегральных функций составляющих:*

$$F(x, y)=F_1(x) \cdot F_2(y).$$

Доказательство. а) Необходимость. Пусть X и Y независимы. Тогда события $X<x$ и $Y<y$ независимы, следовательно, вероятность совмещения этих

событий равна произведению их вероятностей

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

или

$$F(x,\ y)=F_1(x)\cdot F_2(y).$$

б) Достаточность. Пусть $F(x,\ y)=F_1(x)\cdot F_2(y)$. Отсюда

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

т. е. вероятность совмещения событий $X<x$ и $Y<y$ равна произведению вероятностей этих событий. Следовательно, случайные величины X и Y независимы.

Следствие. *Для того чтобы непрерывные случайные величины X и Y были независимыми, необходимо и достаточно, чтобы дифференциальная функция системы $(X,\ Y)$ была равна произведению дифференциальных функций составляющих:*

$$f(x,\ y)=f_1(x)\cdot f_2(y).$$

Доказательство. а) Необходимость. Пусть X и Y — независимые непрерывные случайные величины. Тогда (на основании предыдущей теоремы)

$$F(x,\ y)=F_1(x)\cdot F_2(y).$$

Дифференцируя это равенство по x, затем по y, имеем

$$\frac{\partial^2 F}{\partial x\,\partial y}=\frac{\partial F_1}{\partial x}\cdot\frac{\partial F_2}{\partial y},$$

или (по определению дифференциальных функций двумерной и одномерной величин)

$$f(x,\ y)=f_1(x)\cdot f_2(y).$$

б) Достаточность. Пусть

$$f(x,\ y)=f_1(x)\cdot f_2(y).$$

Интегрируя это равенство по x и по y, получим

$$\int\limits_{-\infty}^{y}\int\limits_{-\infty}^{x} f(x,y)\,dx\,dy=\int\limits_{-\infty}^{x} f_1(x)\,dx\int\limits_{-\infty}^{y} f_2(y)\,dy,$$

или (§ 8 гл. XIV и § 3 гл. 11)

$$F(x,\ y)=F_1(x)\cdot F_2(y).$$

Отсюда (на основании предыдущей теоремы) заключаем, что X и Y независимы.

Замечание. Так как приведенные выше условия являются необходимыми и достаточными, то можно дать новые определения независимых случайных величин:

1) две случайные величины называют независимыми, если интегральная функция системы этих величин равна произведению интегральных функций составляющих;

2) две непрерывные случайные величины называют независимыми, если дифференциальная функция систем этих величин равна произведению дифференциальных функций составляющих

§ 17. Числовые характеристики системы двух случайных величин. Корреляционный момент. Коэффициент корреляции

Для описания системы двух случайных величин, кроме математических ожиданий и дисперсий составляющих, пользуются и другими характеристиками, к числу которых относятся корреляционный момент и коэффициент корреляции.

Корреляционным моментом μ_{xy} случайных величин X и Y называют математическое ожидание произведения отклонений этих величин:

$$\mu_{xy}=M[(X-M(X))(Y-M(Y))].$$

Для вычисления корреляционного момента дискретных величин пользуются формулой

$$\mu_{xy}=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i-M(X))(y_j-M(Y))\,p(x_i,y_j),$$

а для непрерывных величин —

$$\mu_{xy}=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x-M(X))(y-M(Y))\,f(x,y)\,dx\,dy.$$

Корреляционный момент служит для характеристики связи между величинами X и Y. Как будет показано ниже, корреляционный момент равен нулю, если X и Y независимы; следовательно, если корреляционный момент не равен нулю, то X и Y — зависимые случайные величины.

Теорема. *Корреляционный момент двух независимых случайных величин X и Y равен нулю.*

Доказательство. Так как X и Y независимые случайные величины, то их отклонения $X-M(X)$ и $Y-M(Y)$ также независимы. Пользуясь свойствами математического ожидания (математическое ожидание произведения независимых случайных величин равно произведению математических ожиданий сомножителей) и отклонения (математическое ожидание отклонения равно нулю), получим

$$\mu_{xy} = M[(X - M(X)) \cdot (Y - M(Y))] =$$

$$= M[X - M(X)] \cdot M[Y - M(Y)] = 0.$$

Из определения корреляционного момента следует, что он имеет размерность, равную произведению размерностей величин X и Y. Другими словами, величина корреляционного момента зависит от единиц измерения случайных величин. По этой причине для одних и тех же двух величин величина корреляционного момента будет иметь различные значения в зависимости от того, в каких единицах были измерены величины.

Пусть например X и Y были измерены в сантиметрах и $\mu_{xy}=2$ *см*^{2}; если измерить X и Y в миллиметрах, то $\mu_{xy}=200$ *мм*^{2}. Такая особенность корреляционного момента является недостатком этой числовой характеристики, поскольку сравнение корреляционных моментов различных систем случайных величин становится затруднительным. Для того чтобы устранить этот недостаток, вводят новую числовую характеристику — коэффициент корреляции.

Коэффициентом корреляции r_{xy} случайных величин X и Y называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y}.$$

Так как размерность μ_{xy} равна произведению размерностей величин X и Y, σ_x имеет размерность величины X, σ_y имеет размерность величины Y (гл. VIII, § 7), то r_{xy} есть безразмерная величина. Таким образом, величина коэффициента корреляции не зависит от выбора единиц измерения случайных величин. В этом состоит преимущество коэффициента корреляции перед корреляционным моментом.

Очевидно, коэффициент корреляции независимых случайных величин равен нулю (так как $\mu_{xy}=0$).

Замечание. Во многих вопросах теории вероятностей целесообразно вместо случайной величины X рассматривать нормированную величину X', которая определяется как отношение отклонения к среднему квадратическому отклонению:

$$X' = \frac{X - M(X)}{\sigma_x}$$

Нормированная величина имеет математическое ожидание, равное нулю, и дисперсию, равную единице. Действительно, пользуясь свойствами математического ожидания и дисперсии имеем:

$$M(X') = M\left(\frac{X - M(X)}{\sigma_x}\right) = \frac{1}{\sigma_x} \cdot M[X - M(X)] = \frac{1}{\sigma_x} \cdot 0 = 0;$$

$$D(X') = D\left(\frac{X - M(X)}{\sigma_x}\right) = \frac{1}{\sigma_x^2} \cdot D(X - M(X)) = \frac{D(X)}{\sigma_x^2} = 1.$$

Легко убедиться в том, что коэффициент корреляции r_{xy} равен корреляционному моменту нормированных величин X' и Y':

$$r_{xy} = \frac{M[(X - M(X)) \cdot (Y - M(Y))]}{\sigma_x \sigma_y} = M\left[\frac{X - M(X)}{\sigma_x} \times\right.$$

$$\left.\times \frac{Y - M(Y)}{\sigma_y}\right] = M(X' \cdot Y') = \mu_{x' y'}.$$

§ 18. Коррелированность и зависимость случайных величин

Две случайные величины X и Y называют *коррелированными*, если их корреляционный момент (или что то же — коэффициент корреляции) отличен от нуля; X и Y называют *некоррелированными* величинами, если их корреляционный момент равен нулю.

Две коррелированные величины также и зависимы. Действительно, допустив противное, мы должны заключить, что $\mu_{xy}=0$, а это противоречит условию, так как для коррелированных величин $\mu_{xy}\neq 0$.

Обратное предложение не всегда имеет место, т. е. если две величины зависимы, то они могут быть как коррелированными, так и некоррелированными. Другими словами, корреляционный момент двух зависимых величин может быть не равен нулю, но может и равняться нулю.

Убедимся на примере, что две зависимые величины могут быть некоррелированными.

Пример. Двумерная случайная величина (X, Y) задана дифференциальной функцией:

$$f(x, y) = \frac{1}{6\pi} \text{ внутри эллипса } \frac{x^2}{9} + \frac{y^2}{4} = 1,$$

$$f(x, y) = 0 \text{ вне этого эллипса}$$

Доказать, что X и Y зависимые некоррелированные величины.

Р е ш е н и е. Воспользуемся ранее вычисленными дифференциальными функциями X и Y (§ 12):

$$f_1(x) = \frac{2}{9\pi}\sqrt{9-x^2}, \quad f_2(y) = \frac{1}{2\pi}\sqrt{4-y^2} \text{ внутри заданного эллипса и } f_1(x) = 0, \; f_2(y) = 0 \text{ вне его.}$$

Так как $f(x, y)\neq f_1(x)\cdot f_2(y)$, то X и Y зависимые величины (§ 16).

Для того, чтобы доказать некоррелированность X и Y, достаточно убедиться в том, что $\mu_{xy}=0$.

Найдем корреляционный момент по формуле (§ 17):

$$\mu_{xy} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} [x - M(X)][y - M(Y)]\, f(x, y)\, dx dy$$

Поскольку дифференциальная функция $f_1(x)$ симметрична относительно оси OY, то $M(x)=0$; аналогично $M(Y)=0$, в силу симметрии $f_2(y)$ относительно оси OX. Следовательно

$$\mu_{xy} = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} xy \cdot f(x, y)\, dx dy$$

Вынеся постоянный множитель $f(x, y)$ за знак интеграла, получим

$$\mu_{xy} = f(x, y)\int\limits_{-\infty}^{\infty} y \left(\int\limits_{-\infty}^{\infty} x dx\right) dy.$$

Внутренний интеграл равен нулю (подынтегральная функция нечетна, пределы интегрирования симметричны относительно начала координат), следовательно, $\mu_{xy}=0$, т. е. зависимые случайные величины X и Y некоррелированы.

Итак, из коррелированности двух случайных величин следует их зависимость, но из зависимости еще не вытекает коррелированность. Из независимости двух величин следует их некоррелированность, но из некоррелированности еще нельзя заключить о независимости этих величин.

Заметим, однако, что из некоррелированности нормально распределенных величин вытекает их независимость. Это утверждение будет доказано в следующем параграфе.

§ 19. Нормальный закон распределения на плоскости

На практике часто встречаются двумерные случайные величины, распределение которых нормально.

Нормальным законом распределения на плоскости называют распределение вероятностей двумерной случайной величины (X, Y), если

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \times$$

$$\times e^{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\frac{(x-a_1)^2}{\sigma_x^2} + \frac{(y-a_2)^2}{\sigma_y^2} - 2\rho_{xy}\frac{x-a_1}{\sigma_x}\frac{y-a_2}{\sigma_y}\right]}. \quad (*)$$

Мы видим, что нормальный закон на плоскости определяется пятью параметрами: a_1, a_2, σ_x, σ_y и ρ_{xy}. Можно доказать, что эти параметры имеют следующий вероятностный смысл:

a_1, a_2— математические ожидания,

σ_x, σ_y— средние квадратические отклонения,

ρ_{xy}— коэффициент корреляции величин X и Y.

Убедимся в том, что если составляющие двумерной нормально распределенной случайной величины некоррелированны, то они и независимы. Действительно, пусть X

и Y некоррелированны. Тогда, полагая в формуле (*) $\rho_{xy}=0$, получим

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot e^{-\frac{1}{2}\left[\frac{(x-a_1)^2}{\sigma_x^2} + \frac{(y-a_2)^2}{\sigma_y^2}\right]} =$$

$$= \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{(x-a_1)^2}{2\sigma_x^2}} \cdot \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-a_2)^2}{2\sigma_y^2}} = f_1(x)\cdot f_2(y).$$

Таким образом, если составляющие нормально распределенной случайной величины некоррелированны, то дифференциальная функция системы равна произведению дифференциальных функций составляющих, а отсюда и следует независимость составляющих (§ 16). Справедливо и обратное утверждение (§ 18).

Итак, для нормально распределенных составляющих двумерной случайной величины понятия независимости и некоррелированности — равносильны.

Задачи

1. Найти законы распределения составляющих дискретной двумерной случайной величины, заданной законом распределения

Y \ X	x_1	x_2	x_3
y_1	0,12	0,18	0,10
y_2	0,10	0,11	0,39

Отв.

X	x_1	x_2	x_3
p	0,22	0,29	0,49

Y	y_1	y_2
p	0,40	0,60.

2. Найти вероятность того, что составляющая X двумерной случайной величины примет значение $X < \frac{1}{2}$ и при этом составляющая Y примет значение $y < \frac{1}{3}$, если известна интегральная функция системы

$$F(x, y) = \left(\frac{1}{\pi}\operatorname{arctg} 2x + \frac{1}{2}\right)\left(\frac{1}{\pi}\operatorname{arctg} 3y + \frac{1}{2}\right).$$

Отв. $P\left(X < \frac{1}{2},\ \ Y < \frac{1}{3}\right) = \frac{9}{16}$.

3. Найти вероятность попадания случайной точки $(X\ Y)$ в прямоугольник, ограниченный прямыми $x = \frac{\pi}{4}$, $x = \frac{\pi}{2}$, $y = \frac{\pi}{6}$, $y = \frac{\pi}{3}$, если известна интегральная функция

$$F(x, y) = \sin x \sin y \left(0 < x < \frac{\pi}{2},\ \ 0 < y < \frac{\pi}{2}\right).$$

Отв $P\left(\frac{\pi}{4} < X < \frac{\pi}{2},\ \frac{\pi}{6} < Y < \frac{\pi}{3}\right) = 0{,}11.$

4. Найти дифференциальную функцию системы двух случайных величин по известной интегральной функции

$$F(x, y) = (1 - e^{-2x})(1 - e^{-3y})\ (x \geqslant 0,\ y \geqslant 0).$$

Отв. $f(x, y) = \frac{\partial^2 F}{\partial x \partial y} = 6e^{-(2x+3y)}$.

5. Внутри прямоугольника, ограниченного прямыми $x = 0$, $x = \frac{\pi}{2}$, $y = 0$, $y = \frac{\pi}{2}$, дифференциальная функция системы двух случайных величин $f(x, y) = C \sin(x + y)$ вне прямоугольника $f(x, y) = 0$. Найти: а) величину C, б) интегральную функцию системы

Отв а) $C = 0{,}5$; б) $F(x, y) = 0{,}5\,[\sin x + \sin y - \sin(x + y)]\left(0 < x < \frac{\pi}{2},\ 0 < y < \frac{\pi}{2}\right)$.

6. Система двух случайных величин распределена равномерно: в прямоугольнике, ограниченном прямыми $x = 4$, $x = 6$, $y = 10$, $y = 15$, дифференциальная функция сохраняет постоянное значение, а вне этого прямоугольника она равна нулю. Найти: а) дифференциальную функцию, б) интегральную функцию системы

Отв а) $f(x, y) = \begin{cases} 0{,}1 & \text{внутри прямоугольника,} \\ 0 & \text{вне прямоугольника;} \end{cases}$

б) $F(x, y) = \frac{(x-4)(y-10)}{10}$.

7. Дифференциальная функция системы двух случайных величин $f(x, y) = \frac{C}{(4+x^2)(9+y^2)}$. Найти: а) величину C. б) интегральную функцию системы

Отв а) $C = \frac{6}{\pi^2}$;

б) $F(x, y) =$

$$= \left(\frac{1}{\pi} \operatorname{arctg} \frac{x}{2} + \frac{1}{2}\right)\left(\frac{1}{\pi} \operatorname{arctg} \frac{y}{3} + \frac{1}{2}\right).$$

8. Двумерная случайная величина задана дифференциальной функцией

$$f(x, y) = \frac{3\sqrt{3}}{\pi} e^{-4x^2 - 6xy - 9y^2}.$$

Найти условные законы распределения составляющих

Отв. $\varphi(x \mid y) =$

$$= \frac{2}{\sqrt{\pi}} e^{-\left(2x + \frac{3}{2} y\right)^2},$$

$\psi(x \mid y) =$

$$= \frac{3}{\sqrt{\pi}} e^{-(x+3y)^2}.$$

Часть третья

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Глава пятнадцатая

ВЫБОРОЧНЫЙ МЕТОД

§ 1. Задача математической статистики

Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении статистических данных — результатах наблюдений. Первая задача математической статистики — указать способы сбора и группировки (если данных очень много) статистических сведений.

Вторая задача математической статистики — разработать методы анализа статистических данных, в зависимости от целей исследования.

Изучение тех или иных явлений методами математической статистики служит средством решения многих вопросов, выдвигаемых наукой и практикой (правильная организация технологического процесса, наиболее целесообразное планирование и др.).

Итак, задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов

§ 2. Краткая историческая справка

Математическая статистика возникла (XVII в.) и создавалась параллельно с теорией вероятностей Дальнейшее развитие математической статистики (вторая половина XIX и начало XX вв.) обязано, в первую очередь, П. Л. Че бышеву, А. А. Маркову, А. М Ляпунову, а также К. Гауссу, А. Кетле, Ф. Гальтону, К. Пирсону и др.

В XX в. наиболее существенный вклад в математическую статистику был сделан советскими математиками (В. И. Романовский, Е. Е. Слуцкий, А. Н. Колмогоров, Н. В. Смирнов), а также английскими (Стьюдент, Р. Фишер, Э. Пирсон) и американскими (Ю. Нейман, А. Вальд) учеными.

§ 3. Генеральная и выборочная совокупности

Пусть требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака, характеризующего эти объекты. Например, если имеется партия деталей, то качественным признаком может служить стандартность детали, а количественным — контролируемый размер детали.

Иногда проводят сплошное обследование, т. е. обследуют каждый из объектов совокупности относительно признака, которым интересуются. На практике, однако, сплошное обследование применяется сравнительно редко. Например, если совокупность содержит очень большое число объектов, то провести сплошное обследование физически невозможно. Если обследование объекта связано с его уничтожением или требует больших материальных затрат, то проводить сплошное обследование практически не имеет смысла. В таких случаях случайно отбирают из всей совокупности ограниченное число объектов и подвергают их изучению.

Выборочной совокупностью, или просто *выборкой*, называют совокупность случайно отобранных объектов.

Генеральной совокупностью называют совокупность объектов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности. Например, если из 1000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности $N=1000$, а объем выборки $n=100$.

Замечание. Часто генеральная совокупность содержит конечное число объектов Однако, если это число достаточно велико, то иногда в целях упрощения вычислений, или для облегчения теоретических выводов, допускают, что генеральная совокупность состоит из бесчисленного множества объектов. Такое допущение оправдывается тем, что увеличение объема генеральной совокупности (достаточно большого объема) практически не сказывается на результатах обработки данных выборки.

§ 4. Повторная и бесповторная выборки. Репрезентативная выборка

При составлении выборки можно поступать двояко: после того, как объект отобран и над ним произведено наблюдение, он может быть возвращен, либо не возвращен в генеральную совокупность. В соответствии со сказанным, выборки подразделяют на повторные и бесповторные.

Повторной называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

На практике обычно пользуются бесповторным случайным отбором.

Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем нас признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно его представляли. Это требование коротко формулируют так: выборка должна быть *репрезентативной* (представительной).

В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществить случайно: каждый объект выборки отобран случайно из генеральной совокупности, если все объекты имеют одинаковую вероятность попасть в выборку.

Если объем генеральной совокупности достаточно велик, а выборка составляет лишь незначительную часть этой совокупности, то различие между повторной и бесповторной выборками стирается; в предельном случае, когда рассматривается бесконечная генеральная совокупность, а выборка имеет конечный объем, это различие исчезает.

§ 5. Способы отбора

На практике применяются различные способы отбора. Принципиально эти способы можно подразделить на два вида:

1. Отбор, не требующий расчленения генеральной совокупности на части, сюда относятся:

а) простой случайный бесповторный отбор;

б) простой случайный повторный отбор.

2. Отбор, при котором генеральная совокупность разбивается на части, сюда относятся:

а) типический отбор;

б) механический отбор;

в) серийный отбор.

Простым случайным называют такой отбор, при котором объекты извлекают по одному из всей генеральной совокупности. Осуществить простой отбор можно различными способами. Например, для извлечения n объектов из генеральной совокупности объема N поступают так: выписывают номера от 1 до N на карточках, которые тщательно перемешивают и наугад вынимают одну карточку; объект, имеющий одинаковый номер с извлеченной карточкой, подвергают обследованию; затем карточка возвращается в пачку и процесс повторяется, т. е. карточки перемешиваются, наугад вынимают одну из них и т. д. Так поступают n раз; в итоге получают простую случайную повторную выборку объема n.

Если извлеченные карточки не возвращать в пачку, то выборка будет простой случайной бесповторной.

При большом объеме генеральной совокупности описанный процесс оказывается очень трудоемким. В этом случае пользуются готовыми таблицами «случайных чисел», в которых числа расположены в случайном порядке. Для того чтобы отобрать, например 50 объектов из пронумерованной генеральной совокупности, открывают любую страницу таблицы случайных чисел и выписывают подряд 50 чисел; в выборку попадают те объекты, номера которых совпадают с выписанными случайными числами. Если бы оказалось, что случайное число таблицы превышает число N, то такое случайное число пропускают. При осуществлении бесповторной выборки случайные числа таблицы, уже встречавшиеся ранее, следует также пропустить.

Типическим называют отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части. Например, если детали изготовляют на нескольких станках, то отбор производят не из всей совокупности деталей, произведенных всеми станками, а из продукции каждого станка в отдельности. Типическим отбором пользуются тогда, когда обследуемый признак

заметно колеблется в различных типических частях генеральной совокупности. Например, если продукция изготовляется на нескольких машинах, среди которых есть более и менее изношенные, то здесь типический отбор целесообразен.

Механическим называют отбор, при котором генеральная совокупность «механически» делится на столько групп, сколько объектов должно войти в выборку, и из каждой группы отбирается один объект.

Например, если нужно отобрать 20% изготовленных станком деталей, то отбирают каждую пятую деталь; если требуется отобрать 5% деталей, то отбирают каждую двадцатую деталь и т. д.

Следует указать, что иногда механический отбор может не обеспечить репрезентативности выборки. Например, если отбирается каждый двадцатый обтачиваемый валик, причем сразу же после отбора производят замену резца, то отобранными окажутся все валики, обточенные затупленными резцами. В таком случае надо устранить совпадение ритма отбора с ритмом замены резца, для чего надо отбирать, скажем, каждый десятый валик из двадцати обточенных.

Серийным называют отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Например, если изделия изготовляются большой группой станков-автоматов, то подвергают сплошному обследованию продукцию только нескольких станков. Серийным отбором пользуются тогда, когда обследуемый признак колеблется в различных сериях незначительно.

Подчеркнем, что на практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы.

Например, иногда разбивают генеральную совокупность на серии одинакового объема, затем простым случайным отбором выбирают несколько серий и, наконец, из каждой серии простым случайным отбором извлекают отдельные объекты.

§ 6. Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 — n_2 раз, x_k — n_k раз и $\Sigma n_i = n$ — объем выборки. Наблюдаемые значения x_i называют *вариантами*, а последовательность вариант, записанных в возрастающем порядке — вариационным рядом. Числа наблюдений называют *частотами*, а их отношения к объему выборки $\frac{n_i}{n}=W_i$ — *относительными частотами*.

Статическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал).

Заметим, что в теории вероятностей под *распределением* понимают соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике — соответствие между наблюдаемыми вариантами и их частотами, или относительными частотами.

Пример. Задано распределение частот выборки объема =20:

x_i	2	6	12
n_i	3	10	7.

Написать распределение относительных частот.

Решение. Найдем относительные частоты, для чего разделим частоты на объем выборки:

$$W_1=\frac{3}{20}=0{,}15,\quad W_2=\frac{10}{20}=0{,}50\quad W_3=\frac{7}{20}=0{,}35.$$

Напишем распределение относительных частот:

x_i	2	6	12
W_i	0,15	0,6	0,35

Контроль: $0{,}15+0{,}5+0{,}35=1$.

§ 7. Эмпирическая функция распределения

Пусть известно статистическое распределение частот количественного признака X. Введем обозначения:

n_x — число наблюдений, при которых наблюдалось значение признака меньшее x,

n — общее число наблюдений (объем выборки).

Ясно, что относительная частота события $X<x$ равна $\frac{n_x}{n}$. Если x будет изменяться, то вообще говоря, будет изменяться и относительная частота, т. е. относительная частота $\frac{n_x}{n}$ есть функция от x. Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X<x$.

Итак, по определению

$$F^*(x)=\frac{n_x}{n},$$

где n_x— число вариант, меньших x,

n — объем выборки.

Таким образом, для того чтобы найти, например $F^*(x_2)$, надо число вариант, меньших x_2, разделить на объем выборки:

$$F^*(x_2)=\frac{n_{x_2}}{n}.$$

В отличие от эмпирической функции распределения выборки, интегральную функцию $F(x)$ распределения генеральной совокупности называют *теоретической функцией распределения*. Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X<x$, а эмпирическая функция $F^*(x)$ определяет относительную частоту этого же события. Из теоремы Бернулли следует, что относительная частота события $X<x$, т. е. $F^*(x)$ стремится по вероятности к вероятности $F(x)$ этого события. Другими словами, числа $F^*(x)$ и $F(x)$ мало отличаются одно от другого. Уже отсюда следует целесообразность использования эмпирической функции распределения выборки для приближенного представления теоретической (интегральной) функции распределения генеральной совокупности.

Такое заключение подтверждается и тем, что $F^*(x)$ обладает всеми свойствами $F(x)$. Действительно, из определения функции $F^*(x)$ вытекают следующие ее свойства:

1) значения эмпирической функции принадлежат отрезку [0,1];

2) $F^*(x)$ — неубывающая функция;

3) если x_1 — наименьшая варианта, то $F^*(x)=0$ при $x \leqslant x_1$;

если x_k — наибольшая варианта, то $F^*(x)=1$ при $x>x_k$.

Итак, эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

Пример. Построить эмпирическую функцию по данному распределению выборки:

варианты x_i	2	6	10
частоты n_i	12	18	30.

Решение. Найдем объем выборки: $12+18+30=60$.

Наименьшая варианта равна 2, следовательно,

$$F^*(x)=0, \text{ при } x \leqslant 2.$$

Значение $X<6$, а именно $x_1=2$ наблюдалось 12 раз; следовательно,

$$F^*(x)=\frac{12}{60}=0{,}2 \text{ при } 2<x \leqslant 6.$$

Значения $X<10$, а именно $x_1=2$ и $x_2=6$ наблюдались $12+18=30$ раз; следовательно,

$$F^*(x)=\frac{30}{60}=0{,}5 \text{ при } 6<x \leqslant 10.$$

Так как $x=10$ — наибольшая варианта, то

$$F^*(x)=1 \text{ при } x>10.$$

Искомая эмпирическая функция

$$F^*(x)=\begin{cases} 0 & \text{при } x \leqslant 2, \\ 0{,}2 & \text{при } 2<x \leqslant 6, \\ 0{,}5 & \text{при } 6<x \leqslant 10, \\ 1 & \text{при } x>10. \end{cases}$$

График этой функции изображен на рис. 19.

§ 8. Полигон и гистограмма

В целях наглядности строят различные графики статистического распределения и, в частности, полигон и гистограмму.

Полигоном частот называют ломаную, отрезки которой соединяют точки (x_1, n_1), (x_2, n_2), ..., (x_k, n_k). Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат — соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки (x_1, W_1), (x_2, W_2), ..., (x_k, W_k). Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i, а на оси ординат

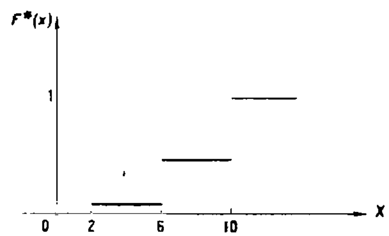

Рис. 19

соответствующие им относительные частоты W_i. Точки (x_i, W_i) соединяют отрезками прямых и получают полигон относительных частот.

В случае непрерывного признака целесообразно строить гистограмму, для чего интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько частичных интервалов длиною h и находят для каждого частичного интервала n_i — сумму частот вариант, попавших в i-й интервал.

На рис. 20 изображен полигон относительных частот следующего распределения:

X	1,5	3,5	5,5	7,5
W	0,1	0,2	0,4	0,3.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частоты).

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{n_i}{h}$.

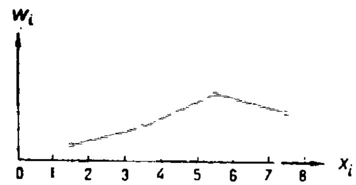

Рис. 20.

Площадь i-го частичного прямоугольника равна $h \cdot \frac{n_i}{h} = n_i$ — сумме частот вариант i-го интервала; следовательно, **площадь гистограммы частот равна сумме всех частот, т. е. объему выборки.**

На рис. 21 изображена гистограмма частот распределения объема $n=100$, приведенного в таблице 6.

Таблица 6

Частичный интервал длиною $h=5$	Сумма частот вариант частичного интервала n_i	Плотность частоты $\frac{n_i}{h}$
5—10	4	0,8
10—15	6	1,2
15—20	16	3,2
20—25	36	7,2
25—30	24	4,8
30—35	10	2,0
35—40	4	0,8

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{W_i}{h}$ (плотность относительной частоты).

Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{W_i}{h}$. Площадь i-го частичного прямоугольника равна $h \cdot \frac{W_i}{h} = W_i$ — относительной частоте вариант, попавших

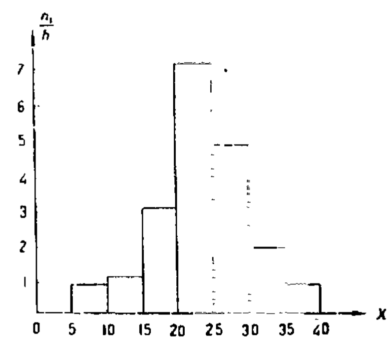

Рис. 21

в i-й интервал. Следовательно, **площадь гистограммы относительных частот равна сумме всех относительных частот, т. е. единице.**

Задачи

1. Построить график эмпирической функции распределения

x_i	5	7	10	15
n_i	2	3	8	7.

2. Построить полигоны частот и относительных частот распределения

x_i	1	3	5	7	9
n_i	10	15	30	33	12.

3. Построить гистограммы частот и относительных частот распределения (в первом столбце указан частичный интервал во втором — сумма частот вариант частичного интервала):

2—5	9
5—8	10
8—11	25
11—14	6

Глава шестнадцатая

СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

§ 1. Статистические оценки параметров распределения

Пусть требуется изучить количественный признак генеральной совокупности Допустим, что из теоретических соображений удалось установить, какое именно распределение имеет признак. Естественно возникает задача оценки параметров, которыми определяется это распределение. Например, если наперед известно, что изучаемый признак распределен в генеральной совокупности нормально, то необходимо оценить (приближенно найти) математическое ожидание и среднее квадратическое отклонение, так как эти два параметра полностью определяют нормальное распределение; если же есть основания считать, что признак имеет, например распределение Пуассона, то необходимо оценить параметр λ, которым это распределение определяется.

Обычно в распоряжении исследователя имеются лишь данные выборки, например, значения количественного признака $x_1, x_2, \ldots, x_n$, полученные в результате n наблюдений (здесь и далее наблюдения предполагаются независимыми). Через эти данные и выражают оцениваемый параметр.

Рассматривая $x_1, x_2, \ldots, x_n$ как независимые случайные величины $X_1, X_2, \ldots, X_n$, можно сказать, что найти статистическую оценку неизвестного параметра теоретического распределения — это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра. Например как будет показано далее, для оценки математического ожидания нормального распределения служит функция (среднее арифметическое наблюдаемых значений признака):

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Итак, статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин.

§ 2. Несмещенные, эффективные и состоятельные оценки

Для того чтобы статистические оценки давали «хорошие» приближения оцениваемых параметров, они должны удовлетворять определенным требованиям. Ниже указаны эти требования.

Пусть Θ^* есть статистическая оценка неизвестного параметра Θ теоретического распределения. Допустим, что по выборке объема n найдена оценка Θ_1^*. Повторим опыт, т. е. извлечем из генеральной совокупности другую выборку того же объема и по ее данным найдем оценку Θ_2^*. Повторяя опыт многократно, получим числа $\Theta_1^*, \Theta_2^*, \ldots, \Theta_k^*$, которые, вообще говоря, будут различны между собой. Таким образом, оценку Θ^* можно рассматривать как случайную величину, а числа $\Theta_1^*, \Theta_2^*, \ldots, \Theta_k^*$, — как ее возможные значения.

Представим себе, что оценка Θ^* дает приближенное значение Θ с избытком; тогда каждое, найденное по данным выборок, число Θ_i^* $(i=1, 2, \ldots, k)$ будет больше истинного значения Θ. Ясно, что в этом случае и математическое ожидание (среднее значение) случайной величины Θ^* будет больше, чем Θ, т. е. $M(\Theta^*) > \Theta$. Очевидно, что если Θ^* дает оценку с недостатком, то $M(\Theta^*) < \Theta$.

Таким образом, использование статистической оценки, математическое ожидание которой не равно оцениваемому параметру, привело бы к систематическим (одного знака) ошибкам. По этой причине естественно потребовать, чтобы математическое ожидание оценки Θ^* было равно оцениваемому параметру. Хотя соблюдение этого требования не устранит ошибок (одни значения Θ^* больше, а другие меньше Θ), однако ошибки разных знаков будут встречаться одинаково часто. Иными словами, соблюдение требований $M(\Theta^*) = \Theta$ гарантирует от получения систематических ошибок.

Несмещенной называют статистическую ошибку Θ^*, математическое ожидание которой равно оцениваемому параметру Θ при любом объеме выборки, т. е.

$$M(\Theta^*) = \Theta.$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако было бы ошибочным считать, что несмещенная оценка всегда дает хорошее приближение оцениваемого параметра. Действительно, возможные значения Θ^* могут быть сильно рассеяны вокруг своего среднего значения, т. е. дисперсия $D(\Theta^*)$ может быть значительной. В этом случае, найденная по данным одной выборки оценка, например Θ_1^*, может оказаться весьма удаленной от среднего значения $\overline{\Theta^*}$, а значит, и от самого оцениваемого параметра Θ; приняв Θ_1^* в качестве приближенного значения Θ, мы допустили бы большую ошибку. Если же потребовать, чтобы дисперсия Θ^* была малой, то возможность допустить большую ошибку будет исключена. По этой причине к статистической оценке предъявляется требование эффективности.

Эффективной называют статистическую оценку, которая (при заданном обьеме выборки n) имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема (n велико!) к статистическим оценкам предъявляется требование состоятельности.

Состоятельной называют статистическую оценку, которая при $n\to\infty$ стремится по вероятности к оцениваемому параметру. Например, если дисперсия несмещенной оценки при $n\to\infty$ стремится к нулю, то такая оценка оказывается и состоятельной.

§ 3. Генеральная средняя

Пусть изучается дискретная генеральная совокупность относительно количественного признака X.

Генеральной средней $\overline{x}_\Gamma$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N р а з л и ч н ы, то

$$\overline{x}_\Gamma = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причем $N_1+N_2+\ldots+N_k=N$,

$$\overline{x}_\Gamma = \frac{x_1N_1 + x_2N_2 + \cdots + x_kN_k}{N},$$

т. е. генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

З а м е ч а н и е. Пусть генеральная совокупность объема N содержит объекты с различными значениями признака X, равными $x_1, x_2, \ldots, x_N$. Представим себе, что из этой совокупности наудачу извлекается один объект. Вероятность того, что будет извлечен объект со значением признака, например x_1, очевидно, равна $\frac{1}{N}$. С этой же вероятностью может быть извлечен и любой другой объект. Таким образом, величину признака X можно рассматривать как случайную величину, возможные значения которой $x_1, x_2, \ldots, x_N$ имеют одинаковые вероятности, равные $\frac{1}{N}$. Найдем математическое ожидание $M(X)$:

$$M(X) = x_1\cdot\frac{1}{N} + x_2\cdot\frac{1}{N} + \cdots + x_N\cdot\frac{1}{N} =$$

$$= \frac{x_1 + x_2 + \cdots + x_N}{N} = \overline{x}_\Gamma.$$

Итак, если рассматривать обследуемый признак X генеральной совокупности как случайную величину, то математическое ожидание признака равно генеральной средней этого признака:

$$M(X)=\overline{x}_\Gamma.$$

Этот вывод мы получили, считая, что все объекты генеральной совокупности имеют различные значения признака. Такой же итог будет получен, если допустить, что генеральная совокупность содержит по несколько объектов с одинаковым значением признака.

Обобщая полученный результат на генеральную совокупность с непрерывным распределением признака X, определим генеральную среднюю, и в этом случае как математическое ожидание признака:

$$\overline{x}_\Gamma = M(X).$$

§ 4 Выборочная средняя

Пусть для изучения генеральной совокупности относительно количественного признака X извлечена выборка объема n.

Выборочной средней $\overline{x}_в$ называют среднее арифметическое значение признака выборочной совокупности

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$\bar{x}_в = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем

$$n_1+n_2+\ldots+n_k=n,$$

$$\bar{x}_В = \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{n},$$

или

$$\bar{x}_в = \frac{\sum_{i=1}^{k} n_i x_i}{n},$$

т. е. выборочная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

Замечание. Выборочная средняя найденная по данным одной выборки есть, очевидно, определенное число. Если же извлекать другие выборки того же объема из той же генеральной совокупности, то выборочная средняя будет изменяться от выборки к выборке. Таким образом, выборочную среднюю можно рассматривать как случайную величину, а следовательно, можно говорить о распределениях (теоретическом и эмпирическом) выборочной средней и о числовых характеристиках этого распределения (его называют выборочным), в частности, о математическом ожидании и дисперсии выборочного распределения.

Заметим, что в теоретических рассуждениях выборочные значения $x_1, x_2, \ldots, x_n$ признака X, полученные в итоге независимых наблюдений, также рассматривают как случайные величины $x_1, x_2, \ldots, x_n$, имеющие то же распределение и, следовательно, те же числовые характеристики, которые имеют X

§ 5. Оценка генеральной средней по выборочной средней. Устойчивость выборочных средних

Пусть из генеральной совокупности (в результате независимых наблюдений над количественным признаком X) извлечена повторная выборка объема n со значениями признака $x_1, x_2, \ldots, x_n$. Не уменьшая общности рассуждений, будем считать эти значения признака различными. Пусть генеральная средняя $\bar{x}_г$ неизвестна и требуется оценить ее по данным выборки. В качестве оценки генеральной средней принимают выборочную среднюю

$$\bar{x}_в = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Убедимся, что $\bar{x}_в$ есть несмещенная оценка, т. е. покажем, что математическое ожидание этой оценки равно $\bar{x}_г$. Будем рассматривать $\bar{x}_в$ как случайную величину и $x_1, x_2, \ldots, x_n$, как независимые, одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$. Поскольку эти величины одинаково распределены, то они имеют одинаковые числовые характеристики, в частности, одинаковое математическое ожидание, которое обозначим через a. Так как математическое ожидание среднего арифметического одинаково распределенных случайных величин равно математическому ожиданию каждой из величин (гл. VIII, § 9), то

$$M(\bar{X}_в) = M\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = a. \quad (*)$$

Приняв во внимание, что каждая из величин $X_1, X_2, \ldots, X_n$ имеет то же распределение, что и генеральная совокупность (которую мы также рассматриваем как случайную величину), заключаем, что и числовые характеристики этих величин и генеральной совокупности одинаковы. В частности, математическое ожидание a каждой из величин равно математическому ожиданию признака X генеральной совокупности, т. е.

$$M(X) = \bar{x}_г = a.$$

Заменив в формуле (*) математическое ожидание a через $\bar{x}_г$, окончательно получим

$$M(\bar{X}_в) = \bar{x}_г$$

Тем самым доказано, что выборочная средняя есть несмещенная оценка генеральной средней.

Легко показать, что выборочная средняя является и состоятельной оценкой генеральной средней. Действительно, допустим, что случайные величины $X_1, X_2, \ldots, X_n$ имеют ограниченные дисперсии, мы вправе применить к этим величинам теорему Чебышева (частный случай), в силу которой при увеличении n среднее арифметическое рассматриваемых величин, т. е. $\overline{X}_в$ стремится по вероятности к математическому ожиданию a каждой из величин, или, что то же, к генеральной средней $\overline{x}_г$ (так как $\overline{x}_г = a$).

Итак, при увеличении объема выборки n выборочная средняя стремится по вероятности к генеральной средней, а это и означает, что выборочная средняя есть состоятельная оценка генеральной средней.

Из сказанного следует также, что если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приближенно равны между собой. В этом и состоит свойство *устойчивости выборочных средних*.

Заметим, что если дисперсии двух совокупностей одинаковы, то близость выборочных средних к генеральным не зависит от отношения объема выборки к объему генеральной совокупности. Она зависит от объема выборки: чем объем выборки больше, тем меньше выборочная средняя отличается от генеральной. Например, если из одной совокупности отобран 1% объектов, а из другой совокупности отобрано 4% объектов, причем объем первой выборки оказался большим, чем второй, то первая выборочная средняя будет меньше отличаться от соответствующей генеральной средней, чем вторая.

Замечание. Мы предполагали выборку повторной. Однако полученные выводы применимы и для бесповторной выборки, если ее объем значительно меньше объема генеральной совокупности. Это положение часто используется на практике.

§ 6. Групповая и общая средние

Допустим, что все значения количественного признака X совокупности, безразлично генеральной или выборочной, разбиты на несколько групп. Рассматривая каждую группу как самостоятельную совокупность, можно найти ее среднюю арифметическую.

Групповой средней называют среднее арифметическое значений признака, принадлежащих группе.

Теперь целесообразно ввести специальный термин для средней всей совокупности.

Общей средней $\overline{x}$ называют среднее арифметическое значений признака, принадлежащих всей совокупности.

Зная групповые средние и объемы групп, можно найти общую среднюю: *общая средняя равна средней арифметической групповых средних, взвешенной по объемам групп.*

Опуская доказательство, приведем иллюстрирующий пример.

Пример. Найти общую среднюю совокупности, состоящей из следующих двух групп:

Группа	первая		вторая	
Значение признака	1	6	1	5
Частота	10	15	20	30
Объем	10+15=25		20+30=50.	

Решение. Найдем групповые средние:

$$\overline{x}_1 = \frac{10 \cdot 1 + 15 \cdot 6}{25} = 4;$$

$$\overline{x}_2 = \frac{20 \cdot 1 + 30 \cdot 5}{50} = 3{,}4.$$

Найдем общую среднюю по групповым средним

$$\overline{x} = \frac{25 \cdot 4 + 50 \cdot 3{,}4}{25 + 50} = 3{,}6.$$

Замечание. Для упрощения расчета общей средней совокупности большого объема целесообразно разбить ее на несколько групп, найти групповые средние и по ним общую среднюю.

§ 7. Отклонение от общей средней и его свойство

Рассмотрим совокупность, безразлично генеральную или выборочную, значений количественного признака X объема n:

значение признака	x_1	x_2	...	x_k
частота	n_1	n_2	...	n_k,

причем $\sum\limits_{i=1}^{k} n_i = n$.

Далее для удобства записи знак суммы $\sum\limits_{i=1}^{k}$ будет заменен знаком Σ.

Найдем общую среднюю

$$\overline{x} = \frac{\sum n_i x_i}{n}.$$

Отсюда

$$\sum n_i x_i = n\overline{x}. \qquad (*)$$

Заметим, что поскольку $\overline{x}$ — постоянная величина, то

$$\sum n_i \overline{x} = \overline{x} \sum n_i = n\overline{x}. \qquad (**)$$

Отклонением называют разность $x_i - \overline{x}$ между значением признака и общей средней.

Теорема. *Сумма произведений отклонений на соответствующие частоты равна нулю*

$$\sum n_i (x_i - \overline{x}) = 0.$$

Доказательство. Учитывая (*) и (**), получим

$$\sum n_i (x_i - \overline{x}) = \sum n_i x_i - \sum n_i \overline{x} = n\overline{x} - n\overline{x} = 0.$$

Пример. Дано распределение количественного признака X:

x_i	1	2	3
n_i	10	4	6.

Убедиться, что сумма произведений отклонений на соответствующие частоты равны нулю.

Решение. Найдем общую среднюю

$$\overline{x} = \frac{10 \cdot 1 + 4 \cdot 2 + 6 \cdot 3}{20} = 1{,}8.$$

Найдем сумму произведений отклонений на соответствующие частоты:

$$\sum n_i (x_i - \overline{x}) = 10 \cdot (1 - 1{,}8) + 4 \cdot (2 - 1{,}8) + 6 \cdot (3 - 1{,}8) = 8 - 8 = 0.$$

§ 8. Генеральная дисперсия

Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику — генеральную дисперсию.

Генеральной дисперсией $D_г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\overline{x}_г$.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то

$$D_г = \frac{\sum_{i=1}^{N} (x_i - \overline{x}_г)^2}{N}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то

$$D_г = \frac{\sum_{i=1}^{k} N_i (x_i - \overline{x}_г)^2}{N},$$

т. е. генеральная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Пример. Генеральная совокупность задана таблицей распределения:

x_i	2	4	5	6
N_i	8	9	10	3.

Найти генеральную дисперсию.

Решение. Найдем генеральную среднюю (§ 3):

$$\overline{x}_г = \frac{8 \cdot 2 + 9 \cdot 4 + 10 \cdot 5 + 3 \cdot 6}{8 + 9 + 10 + 3} = \frac{120}{30} = 4$$

Найдем генеральную дисперсию:

$$D_г = \frac{8 \cdot (2-4)^2 + 9 \cdot (4-4)^2 + 10 \cdot (5-4)^2 + 3 \cdot (6-4)^2}{30} =$$

$$= \frac{54}{30} = 1{,}8$$

Кроме дисперсии, для характеристики рассеяния значений признака генеральной совокупности вокруг своего среднего значения пользуются сводной характеристикой — средним квадратическим отклонением

Генеральным средним квадратическим отклонением (стандартом) называют квадратный корень из генеральной дисперсии:

$$\sigma_г = \sqrt{D_г}.$$

§ 9. Выборочная дисперсия

Для того чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $\overline{x}_в$ вводят сводную характеристику — выборочную дисперсию.

Выборочной дисперсией $D_в$ называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\overline{x}_в$.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^2}{n}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем $n_1+n_2+\ldots+n_k=n$, то

$$D_в = \frac{\sum_{i=1}^{k} n_i(x_i - \overline{x}_в)^2}{n},$$

т. е. выборочная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Пример. Выборочная совокупность задана таблицей распределения

x_i	1	2	3	4
n_i	20	15	10	5

Найти выборочную дисперсию.

Решение. Найдем выборочную среднюю (§ 4):

$$\overline{x}_в = \frac{20\cdot 1 + 15\cdot 2 + 10\cdot 3 + 5\cdot 4}{20+15+10+5} = \frac{100}{50} = 2$$

Найдем выборочную дисперсию:

$$D_в = \frac{20(1-2)^2 + 15\cdot(2-2)^2 + 10\cdot(3-2)^2 + 5\cdot(4-2)^2}{50} =$$

$$= \frac{50}{50} = 1.$$

Кроме дисперсии, для характеристики рассеяния значений признака выборочной совокупности вокруг своего среднего значения пользуются сводной характеристикой — средним квадратическим отклонением.

Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии:

$$\sigma_в = \sqrt{D_в}.$$

§ 10. Формула для вычисления дисперсии

Вычисление дисперсии, безразлично, выборочной или генеральной, можно упростить, используя следующую теорему.

Теорема. *Дисперсия равна среднему квадратов значений признака минус квадрат общей средней*

$$D = \overline{x^2} - [\overline{x}]^2.$$

Доказательство. Справедливость теоремы вытекает из преобразований:

$$D = \frac{\sum n_i(x_i - \overline{x})^2}{n} = \frac{\sum n_i(x_i^2 - 2x_i\overline{x} + [\overline{x}]^2)}{n} =$$

$$= \frac{\sum n_i x_i^2}{n} - 2\overline{x}\frac{\sum n_i x_i}{n} + [\overline{x}]^2\frac{\sum n_i}{n} = \overline{x^2} - 2\overline{x}\cdot\overline{x} + [\overline{x}]^2 =$$

$$= \overline{x^2} - [\overline{x}]^2.$$

Итак,

$$D = \overline{x^2} - [\overline{x}]^2,$$

где

$$\overline{x} = \frac{\sum n_i x_i}{n}, \quad \overline{x^2} = \frac{\sum n_i x_i^2}{n}.$$

Пример. Найти дисперсию по данному распределению:

x_i	1	2	3	4
n_i	20	15	10	5

Решение. Найдем общую среднюю:

$$\bar{x} = \frac{20\cdot 1+15\cdot 2+10\cdot 3+5\cdot 4}{20+15+10+5} = \frac{100}{50} = 2$$

Найдем среднюю квадратов значений признака:

$$\overline{x^2} = \frac{20\cdot 1^2+15\cdot 2^2+10\cdot 3^2+5\cdot 4^2}{50} = 5.$$

Искомая дисперсия

$$D = \overline{x^2} - [\bar{x}]^2 = 5 - 2^2 = 1.$$

§ 11. Групповая, внутригрупповая, межгрупповая и общая дисперсии

Допустим, что все значения количественного признака X совокупности, безразлично, генеральной или выборочной, разбиты на k групп. Рассматривая каждую группу как самостоятельную совокупность, можно найти групповую среднюю (§ 6) и дисперсию значений признака, принадлежащих группе, относительно групповой средней.

Групповой дисперсией называют дисперсию значений признака, принадлежащих группе, относительно групповой средней:

$$D_{j\,гр} = \frac{\sum n_i (x_i - \bar{x}_j)^2}{N_j}.$$

где n_i — частота значения x_i;
j — номер группы;
$\bar{x}_j$ — групповая средняя группы j;
$N_j = \sum n_i$ — объем группы j.

Пример 1. Найти групповые дисперсии совокупности, состоящей из следующих двух групп:

Первая группа

x_i	n_i
2	1
4	7
5	2

$N_1 = \Sigma n_i = 10$

Вторая группа

x_i	n_i
3	2
8	3

$N_2 = \Sigma n_i = 5.$

Решение. Найдем групповые средние:

$$\bar{x}_1 = \frac{\sum n_i x_i}{\sum n_i} = \frac{1\cdot 2+7\cdot 4+2\cdot 5}{10} = 4;$$

$$\bar{x}_2 = \frac{2\cdot 3+3\cdot 8}{5} = 6$$

Найдем искомые групповые дисперсии:

$$D_{1гр} = \frac{\sum n_i (x_i - \bar{x}_1)^2}{N_1} = \frac{1\cdot(2-4)^2+7\cdot(4-4)^2+2\cdot(5-4)^2}{10} = 0{,}6;$$

$$D_{2гр} = \frac{2\cdot(3-6)^2+3\cdot(8-6)^2}{5} = 6.$$

Зная дисперсию каждой группы, можно найти их среднюю арифметическую.

Внутригрупповой дисперсией называют среднюю арифметическую групповых дисперсий, взвешенную по объемам групп:

$$D_{внгр} = \frac{\sum N_j D_{j\,гр}}{n}.$$

где N_j — объем группы j;
$n = \sum_{j=1}^{k} N_j$ — объем всей совокупности.

Пример 2. Найти внутригрупповую дисперсию по данным примера 1.

Решение. Искомая внутригрупповая дисперсия равна

$$D_{внгр} = \frac{N_1 D_{1гр} + N_2 D_{2гр}}{n} = \frac{10\cdot 0{,}6+5\cdot 6}{15} = \frac{12}{5}.$$

Зная групповые средние и общую среднюю, можно найти дисперсию групповых средних относительно общей средней.

Межгрупповой дисперсией называют дисперсию групповых средних относительно общей средней:

$$D_{\text{межгр}} = \frac{\sum N_j (\bar{x}_j - \bar{x})^2}{n},$$

где $\bar{x}_j$ — групповая средняя группы j;
N_j — объем группы j;
$\bar{x}$ — общая средняя;
$n=\sum_{j=1}^{k} N_j$ — объем всей совокупности.

Пример 3. Найти межгрупповую дисперсию по данным примера 1.

Решение. Найдем общую среднюю

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{1 \cdot 2 + 7 \cdot 4 + 2 \cdot 5 + 2 \cdot 3 + 3 \cdot 8}{15} = \frac{14}{3}.$$

Используя вычисленные выше величины $\bar{x}_1=4$, $\bar{x}_2=6$, найдем искомую межгрупповую дисперсию

$$D_{\text{межгр}} = \frac{N_1 (\bar{x}_1 - \bar{x})^2 + N_2 (\bar{x}_2 - \bar{x})^2}{n} =$$

$$= \frac{10 \cdot \left(4 - \frac{14}{3}\right)^2 + 5 \cdot \left(6 - \frac{14}{3}\right)^2}{15} = \frac{8}{9}.$$

Теперь целесообразно ввести специальный термин для дисперсии всей совокупности.

Общей дисперсией называют дисперсию значений признака всей совокупности относительно общей средней:

$$D_{\text{общ}} = \frac{\sum n_i (x_i - \bar{x})^2}{n},$$

где n_i — частота значения x_i;
$\bar{x}$ — общая средняя;
n — объем всей совокупности.

Пример 4. Найти общую дисперсию по данным примера 1.

Решение. Найдем искомую общую дисперсию учитывая, что общая средняя равна $\frac{14}{3}$:

$$D_{\text{общ}} = \frac{1 \cdot \left(2 - \frac{14}{3}\right)^2 + 7 \cdot \left(4 - \frac{14}{3}\right)^2 + 2 \cdot \left(5 - \frac{14}{3}\right)^2}{15} +$$

$$+ \frac{2 \cdot \left(3 - \frac{14}{3}\right)^2 + 3 \cdot \left(8 - \frac{14}{3}\right)^2}{15} = \frac{148}{45}.$$

Замечание. Найденная общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий:

$$D_{\text{общ}} = \frac{148}{45};$$

$$D_{\text{внгр}} + D_{\text{межгр}} = \frac{12}{5} + \frac{8}{9} = \frac{148}{45}.$$

В следующем параграфе будет доказано, что такая закономерность справедлива для любой совокупности.

§ 12. Сложение дисперсий

Теорема. *Если совокупность состоит из нескольких групп, то общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий:*

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}}.$$

Доказательство. Для упрощения доказательства предположим, что вся совокупность значений количественного признака X разбита на две следующие группы:

Группа	первая	вторая
Значение признака	$x_1 \ x_2$	$x_1 \ x_2$
Частота	$m_1 \ m_2$	$n_1 \ n_2$
Объем группы	$N_1=m_1+m_2$	$N_2=n_1+n_2$
Групповая средняя	$\bar{x}_1$	$\bar{x}_2$
Групповая дисперсия	$D_{1\text{гр}}$	$D_{2\text{гр}}$
Объем всей совокупности	$n=N_1+N_2$	

Далее для удобства записи вместо знака суммы $\sum_{i=1}^{2}$ будем писать знак $\sum$. Например, $\sum m_i = \sum_{i=1}^{2} m_i = m_1 + m_2 = N_1$.

Следует также иметь в виду, что если под знаком суммы стоит постоянная величина, то ее целесообразно выносить за знак суммы. Например, $\sum m_i(\overline{x}_1-\overline{x})^2=(\overline{x}_1-\overline{x})^2\sum m_i=$ $=(\overline{x}_1-\overline{x})^2N_1$.

Найдем общую дисперсию:

$$D_{\text{общ}}=\frac{\sum m_i(x_i-\overline{x})^2+\sum n_i(x_i-\overline{x})^2}{n}. \qquad (*)$$

Преобразуем первое слагаемое числителя, вычтя и прибавив $\overline{x}_1$:

$$\sum m_i(x_i-\overline{x})^2=\sum m_i\left[(x_i-\overline{x}_1)+(\overline{x}_1-\overline{x})\right]^2=$$
$$=\sum m_i(x_i-\overline{x}_1)^2+2(\overline{x}_1-\overline{x})\sum m_i(x_i-\overline{x}_1)+$$
$$+\sum m_i(\overline{x}_1-\overline{x})^2.$$

Так как

$$\sum m_i(x_i-\overline{x}_1)^2=N_1D_{1\text{гр}}$$

$\left(\text{равенство следует из соотношения } D_{1\text{гр}}=\frac{\sum m_i(x_i-\overline{x}_1)^2}{N_1}\right)$ и, в силу § 7,

$$\sum m_i(x_i-\overline{x}_1)=0,$$

то первое слагаемое принимает вид

$$\sum m_i(x_i-\overline{x}_1)^2=N_1D_{1\text{гр}}+N_1(\overline{x}_1-\overline{x})^2. \qquad (**)$$

Аналогично можно представить второе слагаемое числителя (*) (вычтя и прибавив $\overline{x}_2$):

$$\sum n_i(x_i-\overline{x})^2=N_2D_{2\text{гр}}+N_2(\overline{x}_2-\overline{x})^2. \qquad (***)$$

Подставим (**) и (***) в (*):

$$D_{\text{общ}}=\frac{N_1D_{1\text{гр}}+N_2D_{2\text{гр}}}{n}+\frac{N_1(\overline{x}_1-\overline{x})^2+N_2(\overline{x}_2-\overline{x})^2}{n}=$$
$$=D_{\text{внгр}}+D_{\text{межгр}}.$$

Итак,

$$D_{\text{общ}}=D_{\text{внгр}}+D_{\text{межгр}}.$$

Пример, иллюстрирующий доказанную теорему, приведен в предыдущем параграфе.

Замечание. Теорема имеет не только теоретическое, но и важное практическое значение. Например, если в результате наблюдений получены несколько групп значений признака, то для вычисления общей дисперсии можно группы в единую совокупность не объединять. С другой стороны, если совокупность имеет большой объем, то целесообразно разбить ее на несколько групп. В том и другом случаях непосредственное вычисление общей дисперсии заменяется вычислением дисперсий отдельных групп, что облегчает расчеты.

§ 13. Оценка генеральной дисперсии по исправленной выборочной

Пусть из генеральной совокупности в результате n независимых наблюдений над количественным признаком X извлечена повторная выборка объема n:

значения признака	x_1	x_2	...	x_k,
частота	n_1	n_2	...	n_k,

причем $n_1+n_2+\ldots+n_k=n$.

Требуется по данным выборки оценить (приближенно найти) неизвестную генеральную дисперсию $D_{\text{г}}$. Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка будет приводить к систематическим ошибкам, давая заниженное значение генеральной дисперсии. Объясняется это тем, что как можно доказать, выборочная дисперсия является смещенной оценкой $D_{\text{г}}$, другими словами, математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно

$$M[D_{\text{в}}]=\frac{n-1}{n}D_{\text{г}}.$$

Легко «исправить» выборочную дисперсию так, чтобы ее математическое ожидание было равно генеральной дисперсии. Достаточно для этого умножить $D_{\text{в}}$ на дробь $\frac{n}{n-1}$. Сделав это, получим «исправленную дисперсию», которую обычно обозначают через s^2:

$$s^2=\frac{n}{n-1}D_{\text{в}}=\frac{n}{n-1}\cdot\frac{\sum\limits_{i=1}^{k}n_i(x_i-\overline{x}_{\text{в}})^2}{n}=\frac{\sum\limits_{i=1}^{k}n_i(x_i-\overline{x}_{\text{в}})^2}{n-1}.$$

Исправленная дисперсия является, конечно, несмещенной оценкой генеральной дисперсии. Действительно,

$$M[s^2] = M\left[\frac{n}{n-1} D_в\right] = \frac{n}{n-1} M[D_в] =$$

$$= \frac{n}{n-1} \cdot \frac{n-1}{n} D_г = D_г$$

Итак, в качестве оценки генеральной дисперсии принимают исправленную дисперсию

$$s^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_в)^2}{n-1}.$$

Для оценки же среднего квадратического отклонения генеральной совокупности используют «исправленное» среднее квадратическое отклонение, которое равно квадратному корню из исправленной дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_в)^2}{n-1}}.$$

Подчеркнем, что s не является несмещенной оценкой; чтобы отразить этот факт мы написали и будем писать далее так: «исправленное» среднее квадратическое отклонение.

Замечание. Сравнивая формулы

$$D_в = \frac{\sum n_i (x_i - \bar{x}_в)^2}{n} \quad и \quad s^2 = \frac{\sum n_i (x_i - \bar{x}_в)^2}{n-1}$$

видим, что они отличаются лишь знаменателями. Очевидно, при достаточно больших значениях n объема выборки, выборочная и исправленная дисперсия различаются мало. На практике пользуются исправленной дисперсией, если примерно $n < 30$.

§ 14. Точность оценки, доверительная вероятность (надежность). Доверительный интервал

Точечной называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше — точечные. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами — концами интервала. Интервальные оценки позволяют установить точность и надежность оценок (смысл этих понятий выясняется ниже).

Пусть, найденная по данным выборки, статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Будем считать Θ постоянным числом (Θ может быть и случайной величиной). Ясно, что Θ^* тем точнее определяет параметр Θ, чем меньше абсолютная величина разности $|\Theta-\Theta^*|$. Другими словами, если $\delta>0$ и $|\Theta-\Theta^*|<\delta$, то, чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует *точность оценки.*

Однако статистические методы не позволяют категорически утверждать, что оценка Θ^* удовлетворяет неравенству $|\Theta-\Theta^*|<\delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (*доверительной вероятностью*) оценки Θ по Θ^* называют вероятность γ, с которой осуществляется неравенство $|\Theta-\Theta^*|<\delta$. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что $|\Theta-\Theta^*|<\delta$ равна γ:

$$P[|\Theta-\Theta^*|<\delta]=\gamma.$$

Заменив неравенство $|\Theta-\Theta^*|<\delta$ равносильным ему двойным неравенством $-\delta<\Theta-\Theta^*<\delta$, или

$$\Theta^*-\delta<\Theta<\Theta^*+\delta,$$

имеем

$$P[\Theta^*-\delta<\Theta<\Theta^*+\delta]=\gamma.$$

Это соотношение следует понимать так: вероятность того, что интервал $(\Theta^*-\delta, \Theta^*+\delta)$ заключает в себе (покрывает) неизвестный параметр Θ, равна γ.

Доверительным называют интервал $(\Theta^*-\delta, \Theta^*+\delta)$, который покрывает неизвестный параметр с заданной надежностью γ.

Замечание. Интервал $(\Theta^*-\delta, \Theta^*+\delta)$ имеет случайные концы (их называют доверительными границами). Действительно, в разных выборках, получаются различные значения Θ. Следовательно, от выборки к выборке будут изменяться и концы доверительного интервала, т. е. доверительные границы сами являются случайными величинами — функциями от $x_1, x_2, \dots, x_n$.

Так как случайной величиной является не оцениваемый параметр Θ, а доверительный интервал, то более правильно говорить не о вероятности попадания Θ в доверительный интервал, а о вероятности того, что доверительный интервал покроет Θ.

Метод доверительных интервалов разработан американским статистиком Ю. Нейманом, исходя из идей английского статистика Р. Фишера.

§ 15. Доверительные интервалы для оценки математического ожидания нормального распределения при известном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}$. Поставим своей задачей найти доверительные интервалы, покрывающие параметр a с надежностью γ.

Будем рассматривать выборочную среднюю $\bar{x}$, как случайную величину $\bar{X}$ ($\bar{x}$ изменяется от выборки к выборке) и выборочные значения признака $x_1, x_2, \dots, x_n$, как одинаково распределенные независимые случайные величины $X_1, X_2, \dots, X_n$ (эти числа также изменяются от выборки к выборке). Другими словами, математическое ожидание каждой из этих величин равно a и среднее квадратическое отклонение — σ.

Примем без доказательства, что если случайная величина X распределена нормально, то выборочная средняя $\bar{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\bar{X}$ таковы (гл. VIII, § 9):

$$M(\bar{X}) = a, \quad \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Потребуем, чтобы выполнялось соотношение

$$P(|\bar{X} - a| < \delta) = \gamma,$$

где γ — заданная надежность.

Пользуясь формулой (гл. XII, § 6)

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right),$$

заменив X через $\bar{X}$ и σ через $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, получим

$$P(|\bar{X} - a| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t),$$

где $t = \frac{\delta\sqrt{n}}{\sigma}$.

Найдя из последнего равенства $\delta = t\frac{\sigma}{\sqrt{n}}$, можем написать

$$P\left(|\bar{X} - a| < t\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(t).$$

Приняв во внимание, что вероятность P задана и равна γ, окончательно имеем (чтобы получить рабочую формулу выборочную среднюю вновь обозначим через $\bar{x}$):

$$P\left(\bar{x} - t\frac{\sigma}{\sqrt{n}} < a < \bar{x} + t\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

Смысл полученного соотношения таков: с надежностью γ можно утверждать, что доверительный интервал $\left(\bar{x} - t\frac{\sigma}{\sqrt{n}}, \ \bar{x} + t\frac{\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a; точность оценки $\delta = t\cdot\frac{\sigma}{\sqrt{n}}$.

Итак, поставленная выше задача полностью решена. Укажем еще, что число t определяется из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \frac{\gamma}{2}$; по таблице функции Лапласа (приложение 2) находят аргумент t, которому соответствует значение функции Лапласа, равное $\frac{\gamma}{2}$.

Замечание 1. Оценку $|\bar{x} - a| < t\cdot\frac{\sigma}{\sqrt{n}}$ называют классической. Из формулы $\delta = t\cdot\frac{\sigma}{\sqrt{n}}$, определяющей точность классической оценки, можно сделать следующие выводы:

1) при возрастании объема выборки n число δ убывает и, следовательно, точность оценки увеличивается;

2) увеличение надежности оценки $\gamma = 2\Phi(t)$ приводит к увеличению t ($\Phi(t)$ — возрастающая функция), а следовательно, и к возрастанию δ; другими словами, увеличение надежности классической оценки влечет за собой уменьшение ее точности.

Пример. Случайная величина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma=3$. Найти доверительные интервалы для оценки неизвестного математического ожидания a по выборочным средним $\overline{x}$, если объем выборки $n=36$ и задана надежность оценки $\gamma=0{,}95$.

Р е ш е н и е. Найдем t. Из соотношения $2\Phi(t)=0{,}95$ получим $\Phi(t)=0{,}475$. По таблице (приложение 2) находим

$$t=1{,}96.$$

Найдем точность оценки:

$$\delta=\frac{t\cdot\sigma}{\sqrt{n}}=\frac{1{,}96\cdot 3}{\sqrt{36}}=0{,}98.$$

Доверительные интервалы таковы:

$$(\overline{x}-0{,}98;\ \overline{x}+0{,}98).$$

Например, если $\overline{x}=4{,}1$, то доверительный интервал имеет следующие доверительные границы:

$$\overline{x}-0{,}98=4{,}1-0{,}98=3{,}12;$$
$$\overline{x}+0{,}98=4{,}1+0{,}98=5{,}08.$$

Таким образом, значения неизвестного параметра a, согласующиеся с данными выборки, удовлетворяют неравенству

$$3{,}12<a<5{,}08.$$

Подчеркнем, что было бы ошибочным написать:

$$P\ (3{,}12<a<5{,}08)=0{,}95.$$

Действительно, так как a — постоянная величина, то либо она заключена в найденном интервале (тогда событие $3{,}12<a<5{,}08$ достоверно и его вероятность равна единице), либо в нем не заключена (в этом случае событие $3{,}12<a<5{,}08$ невозможно и его вероятность равна нулю). Другими словами, доверительную вероятность не следует связывать с оцениваемым параметром; она связана лишь с границами доверительного интервала, которые, как уже было указано, изменяются от выборки к выборке.

Поясним смысл, который имеет заданная надежность. Надежность $\gamma=0{,}95$ указывает, что если произведено достаточно большое число выборок, то 95% из них определяет

такие доверительные интервалы, в которых параметр действительно заключен; лишь в 5% случаев он может выйти за границы доверительного интервала.

З а м е ч а н и е 2. Если требуется оценить математическое ожидание с наперед заданной точностью δ и надежностью γ, то минимальный объем выборки, который обеспечит эту точность, находят по формуле

$$n=\frac{t^2\sigma^2}{\delta^2}$$

$\left(\text{следствие равенства } \delta=t\frac{\sigma}{\sqrt{n}}\right)$

§ 16. Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание a при помощи доверительных интервалов. Разумеется, невозможно воспользоваться результатами предыдущего параграфа, в котором σ предполагалось известным.

Оказывается, что по данным выборки можно построить случайную величину (ее возможные значения будем обозначать через t),

$$T=\frac{\overline{X}-a}{\frac{S}{\sqrt{n}}},$$

которая имеет распределение Стьюдента с $k=n-1$ степенями свободы (см. пояснение в конце параграфа); здесь $\overline{X}$ — выборочная средняя, S — «исправленное» среднее квадратическое отклонение, n — объем выборки.

Дифференциальная функция

$$S(t,n)=B_n\left[1+\frac{t^2}{n-1}\right]^{-\frac{n}{2}},$$

где $$B_n=\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}.$$

Мы видим, что распределение Стьюдента определяется параметром n — объемом выборки (или, что то же, числом степеней свободы $k=n-1$) и не зависит от неизвестных параметров a и σ; эта особенность является его большим достоинством). Поскольку $S(t, n)$ — четная функция от t, вероятность осуществления неравенства $\left|\frac{\overline{X}-a}{\frac{S}{\sqrt{n}}}\right| < \gamma$ определяется так (гл. XI, § 2, замечание):

$$P\left(\left|\frac{\overline{X}-a}{\frac{S}{\sqrt{n}}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} S(t, n)\,dt = \gamma.$$

Заменив неравенство в круглых скобках равносильным ему двойным неравенством, получим

$$P\left(\overline{X} - t_\gamma \frac{S}{\sqrt{n}} < a < \overline{X} + t_\gamma \frac{S}{\sqrt{n}}\right) = \gamma.$$

Итак, пользуясь распределением Стьюдента, мы нашли доверительный интервал $\left(\overline{x}-t_\gamma \frac{s}{\sqrt{n}},\ \overline{x}+t_\gamma \frac{s}{\sqrt{n}}\right)$, покрывающий неизвестный параметр a с надежностью γ. Здесь случайные величины $\overline{X}$ и S заменены неслучайными величинами $\overline{x}$ и s, найденными по выборке. По таблице (приложение 3), по заданным n и γ можно найти t_γ.

Пример. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n=16$ найдены выборочная средняя $\overline{x}=20{,}2$ и «исправленное» среднее квадратическое отклонение $s=0{,}8$. Оценить неизвестное математическое ожидание при помощи доверительного интервала с надежностью 0,95.

Решение. Найдем t_γ. Пользуясь таблицей (приложение 3), по $\gamma=0{,}95$ и $n=16$, находим $t_\gamma=2{,}13$.

Найдем доверительные границы:

$$\overline{x} - t_\gamma \frac{s}{\sqrt{n}} = 20{,}2 - 2{,}13\ \frac{0{,}8}{\sqrt{16}} = 19{,}774,$$

$$\overline{x} + t_\gamma \frac{s}{\sqrt{n}} = 20{,}2 + 2{,}13\ \frac{0{,}8}{\sqrt{16}} = 20{,}626.$$

Итак, с надежностью 0,95 неизвестный параметр a заключен в доверительном интервале $19{,}774<a<20{,}626$.

Замечание. Из предельных соотношений

$$\lim_{n\to\infty} B_n = \frac{1}{\sqrt{2\pi}}, \quad \lim_{n\to\infty}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} = e^{-\frac{t^2}{2}}$$

следует, что при неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Поэтому при $n>30$ можно вместо распределения Стьюдента пользоваться нормальным распределением.

Однако важно подчеркнуть, что для малых выборок ($n<30$), в особенности для малых значений n, замена распределения нормальным приводит к грубым ошибкам, а именно — к неоправданному сужению доверительного интервала, т. е. к повышению точности оценки. Например, если $n=5$ и $\gamma=0{,}99$, то пользуясь распределением Стьюдента, найдем $t_\gamma=4{,}6$, а используя функцию Лапласа, найдем $t_\gamma=2{,}58$, т. е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стьюдента.

То обстоятельство, что распределение Стьюдента при малой выборке дает не вполне определенные результаты (широкий доверительный интервал), вовсе не свидетельствует о слабости метода Стьюдента, а объясняется тем, что малая выборка, разумеется, содержит малую информацию об интересующем нас признаке.

Пояснение. Ранее было указано (гл. XII, § 14), что если Z — нормальная величина, причем $M(Z)=0$, $\sigma(Z)=1$, а V — независимая от Z величина, распределенная по закону χ^2 с k степенями свободы, то величина

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \qquad (*)$$

распределена по закону Стьюдента с k степенями свободы.

Пусть количественный признак X генеральной совокупности распределен нормально, причем $M(X)=a$, $\sigma(X)=\sigma$. Если из этой совокупности извлекать выборки объема n и по ним находить выборочные средние, то можно доказать, что выборочная средняя распределена нормально, причем (гл. VIII, § 9)

$$M(\overline{X}_в) = a, \quad \sigma(\overline{X}_в) = \frac{\sigma}{\sqrt{n}}.$$

Тогда случайная величина

$$Z = \frac{\overline{X}_в - a}{\frac{\sigma}{\sqrt{n}}} \qquad (**)$$

также имеет нормальное распределение, как линейная функция нормального аргумента $\overline{X}_в$ (гл XII, § 10 замечание), причем $M(Z)=0$, $\sigma(Z)=1$.

Доказано, что независимая от Z случайная величина

$$V = \frac{(n-1)S^2}{\sigma^2} \qquad (***)$$

(S^2— исправленная выборочная дисперсия) распределена по закону χ^2 с $k=n-1$ степенями свободы.

Следовательно, подставив (**) и (***) в (*) получим величину

$$T = \frac{(\overline{X}_в - a)\sqrt{n}}{S},$$

которая распределена по закону Стьюдента с $k=n-1$ степенями свободы.

§ 17. Оценка истинного значения измеряемой величины

Пусть производится n независимых равноточных измерений некоторой физической величины, истинное значение a которой неизвестно. Будем рассматривать результаты отдельных измерений как случайные величины $X_1, X_2, \ldots, X_n$. Эти величины независимы (измерения независимы), имеют одно и то же математическое ожидание a (истинное значение измеряемой величины), одинаковые дисперсии σ^2 (измерения равноточны) и распределены нормально (такое допущение подтверждается опытом). Таким образом, все предположения, которые были сделаны при выводе доверительных интервалов в двух предыдущих параграфах, выполняются и, следовательно, мы вправе использовать полученные в них формулы. Другими словами, истинное значение измеряемой величины можно оценивать по среднему арифметическому результатов отдельных измерений при помощи доверительных интервалов Поскольку обычно σ неизвестно, следует пользоваться формулами § 16.

Пример. По данным девяти независимых равноточных измерений физической величины найдены среднее арифметическое результатов отдельных измерений $\overline{x}=42{,}319$ и «исправленное» среднее квадратическое отклонение $s=5{,}0$. Требуется оценить истинное значение измеряемой величины с надежностью $\gamma=0{,}95$.

Решение. Истинное значение измеряемой величины равно ее математическому ожиданию. Поэтому задача сводится к оценке математического ожидания (при неизвестном σ) при помощи доверительного интервала

$$\overline{x} - t_\gamma \frac{s}{\sqrt{n}} < a < \overline{x} + t_\gamma \frac{s}{\sqrt{n}},$$

покрывающего a с заданной надежностью $\gamma=0{,}95$

Пользуясь таблицей (приложение 3) по $\gamma=0{,}95$ и $n=9$, находим $t_\gamma=2{,}31$.

Найдем точность оценки:

$$t_\gamma \cdot \frac{s}{\sqrt{n}} = 2{,}31 \cdot \frac{5}{9} = 3{,}85.$$

Найдем доверительные границы:

$$\overline{x} - t_\gamma \frac{s}{\sqrt{n}} = 42{,}319 - 3{,}85 = 36{,}469;$$

$$\overline{x} + t_\gamma \frac{s}{\sqrt{n}} = 42{,}319 + 3{,}85 = 46{,}169.$$

Итак, с надежностью 0,95 истинное значение измеряемой величины заключено в доверительном интервале

$$38{,}469 < a < 46{,}169.$$

§ 18. Доверительные интервалы для оценки среднего квадратического отклонения σ нормального распределения

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по «исправленному» выборочному среднему квадратическому отклонению s. Поставим перед собой задачу найти доверительные интервалы, покрывающие параметр σ с заданной надежностью γ.

Потребуем, чтобы выполнялось соотношение

$$P(|\sigma - s| < \delta) = \gamma,$$

или

$$P(s-\delta<\sigma<s+\delta)=\gamma$$

Для того чтобы можно было пользоваться готовой таблицей, преобразуем двойное неравенство

$$s-\delta<\sigma<s+\delta$$

в равносильное неравенство

$$s\left(1-\frac{\delta}{s}\right)<\sigma<s\left(1+\frac{\delta}{s}\right).$$

Положив $\frac{\delta}{s}=q$, получим

$$s(1-q)<\sigma<s(1+q). \qquad (*)$$

Остается найти q. С этой целью введем в рассмотрение случайную величину «хи»:

$$\chi=\frac{S}{\sigma}\sqrt{n-1},$$

где n — объем выборки.

Как было указано (§ 16, пояснение, соотношение (***), величина $\frac{S^2(n-1)}{\sigma^2}$ распределена по закону χ^2, поэтому квадратный корень из нее обозначают через χ.

Дифференциальная функция распределения χ имеет вид (см пояснение в конце параграфа)

$$R(\chi, n)=\frac{\chi^{n-2}e^{-\frac{\chi^2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}. \qquad (**)$$

Мы видим, что это распределение не зависит от оцениваемого параметра σ, а зависит лишь от объема выборки n.

Преобразуем неравенство (*) так, чтобы оно приняло вид

$$\chi_1<\chi<\chi_2.$$

Вероятность этого неравенства (гл. XI, § 2) равна заданной вероятности γ, т. е.

$$\int_{\chi_1}^{\chi_2} R(\chi, n)\,d\chi=\gamma.$$

Предполагая, что $q<1$, перепишем неравенство (*) так:

$$\frac{1}{S(1+q)}<\frac{1}{\sigma}<\frac{1}{S(1-q)}.$$

Умножив все члены неравенства на $S\sqrt{n-1}$, получим

$$\frac{\sqrt{n-1}}{1+q}<\frac{S\sqrt{n-1}}{\sigma}<\frac{\sqrt{n-1}}{1-q},$$

или

$$\frac{\sqrt{n-1}}{1+q}<\chi<\frac{\sqrt{n-1}}{1-q}.$$

Вероятность того, что это неравенство, а следовательно, и равносильное ему неравенство (*) будет осуществлено, равна

$$\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} R(\chi, n)\,d\chi=\gamma.$$

Из этого уравнения можно по заданным n и γ найти q. Практически для отыскания q пользуются таблицей (приложение 4).

Вычислив по выборке s и найдя по таблице q, получим искомый доверительный интервал (*), покрывающий σ с заданной надежностью γ, т. е. интервал

$$s(1-q)<\sigma<s(1+q)$$

Пример 1. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n=25$ найдено «исправленное» среднее квадратическое отклонение $s=0{,}8$. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,95.

Решение. По таблице (приложение 4) по данным $\gamma=0{,}95$ и $n=25$ найдем $q=0{,}32$.

Искомый доверительный интервал (*) таков:

$$0{,}8(1-0{,}32)<\sigma<0{,}8(1+0{,}32),$$

или

$$0{,}544<\sigma<1{,}056$$

Замечание. Выше предполагалось, что $q < 1$. Если $q > 1$ то неравенство (*) примет вид (учитывая, что $\sigma > 0$)

$$0 < \sigma < s(1 + q).$$

или (после преобразований, аналогичных случаю $q < 1$)

$$\frac{\sqrt{n-1}}{1+q} < \chi < \infty$$

Следовательно, значения $q > 1$ могут быть найдены из уравнения

$$\int\limits_{\frac{\sqrt{n-1}}{1+q}}^{\infty} R(\chi, n)\, d\chi = \gamma.$$

Практически для отыскания значений $q > 1$, соответствующих различным заданным n и γ, пользуются таблицей (приложение 4).

Пример 2. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n=10$ найдено «исправленное» среднее квадратическое отклонение $s=0,16$. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,999.

Решение. По таблице (приложение 4) по данным $\gamma=0,999$ и $n=10$ найдем $q=1,80$ ($q>1$). Искомый доверительный интервал таков:

$$0<\sigma<0,16(1+1,80)$$

или

$$0<\sigma<0,448.$$

Пояснение. Покажем, что дифференциальная функция распределения χ имеет вид (**).

Если случайная величина X распределена по закону χ^2 с $k=n-1$ степенями свободы, то ее дифференциальная функция (гл. XII, § 13)

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\, \Gamma\left(\frac{k}{2}\right)},$$

или, после подстановки $k=n-1$,

$$f(x) = \frac{x^{\frac{n-3}{2}} e^{-\frac{x}{2}}}{2^{\frac{n-1}{2}}\, \Gamma\left(\frac{n-1}{2}\right)}.$$

Воспользуемся формулой (гл. XII, § 10)

$$g(y) = f[\psi(y)] \cdot |\psi'(y)|.$$

чтобы найти распределение функции $\chi=\varphi(X)=\sqrt{X}$ ($\chi>0$). Отсюда обратная функция

$$x = \psi(\chi) = \chi^2$$

и

$$\psi'(\chi) = 2\chi.$$

Так как

$$\chi>0, \text{ то } |\psi'(\chi)| = 2\chi.$$

Следовательно,

$$g(\chi) = f[\psi(\chi)] \cdot |\psi'(\chi)| = \frac{(\chi^2)^{\frac{n-3}{2}} \cdot e^{-\frac{\chi^2}{2}}}{2^{\frac{n-1}{2}}\, \Gamma\left(\frac{n-1}{2}\right)} \cdot 2\chi.$$

Выполнив элементарные преобразования и изменив обозначения ($g(\chi)$ заменим на $R(\chi, n)$), окончательно получим

$$R(\chi, n) = \frac{\chi^{n-2} e^{-\frac{\chi^2}{2}}}{2^{\frac{n-3}{2}}\, \Gamma\left(\frac{n-1}{2}\right)}.$$

§ 19. Оценки точности измерений

В теории ошибок принято точность измерений (точность прибора) характеризовать при помощи среднего квадратического отклонения σ случайных ошибок измерений. Для оценки σ используют «исправленное» среднее квадратическое отклонение s.

Поскольку обычно результаты измерений взаимно независимы, имеют одно и то же математическое ожидание (истинное значение измеряемой величины) и одинаковую дисперсию (в случае равноточных измерений), то теория, изложенная в предыдущем параграфе, применима для оценки измерений.

Пример. По 15 равноточным измерениям найдено «исправленное» среднее квадратическое отклонение $s=0,12$. Найти точность измерений с надежностью 0,99.

Решение. Точность измерений характеризуется средним квадратическим отклонением σ случайных ошибок, поэтому задача сводится к отысканию доверительного интервала (*) (§ 18), покрывающего σ с заданной надежностью 0,99.

По таблице (приложение 4) по $\gamma=0,99$ и $n=15$ найдем $q=0,73$. Искомый доверительный интервал таков:

$$0,12(1-0,73)<\sigma<0,12(1+0,73),$$

или

$$0,03<\sigma<0,21.$$

§ 20. Другие характеристики вариационного ряда

Кроме выборочной средней и выборочной дисперсии применяются и другие характеристики вариационного ряда. Укажем главные из них.

Модой M_0 называют варианту, которая имеет наибольшую частоту. Например, для ряда

варианта	1	4	7	9
частота	5	1	20	6

мода равна 7.

Медианой m_e называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т. е. $n=2k+1$, то $m_e=x_{k+1}$; при четном $n=2k$ медиана

$$m_e=\frac{x_k+x_{k+1}}{2}.$$

Например, для ряда

$$2\quad 3\quad 5\quad 6\quad 7$$

медиана равна 5; для ряда

$$2\quad 3\quad 5\quad 6\quad 7\quad 9$$

медиана равна $\frac{5+6}{2}=5,5$.

Размахом варьирования R называют разность между наибольшей и наименьшей вариантами:

$$R=x_{max}-x_{min}.$$

Например, для ряда

$$1\quad 3\quad 4\quad 5\quad 6\quad 10$$

размах равен $10-1=9$.

Размах является простейшей характеристикой рассеяния вариационного ряда.

Средним абсолютным отклонением θ называют среднее арифметическое абсолютных отклонений:

$$\theta=\frac{\sum n_i|x_i-\bar{x}_в|}{\sum n_i}.$$

Например, для ряда

x_i	1	3	6	16
n_i	4	10	5	1

имеем

$$\bar{x}_в=\frac{4\cdot 1+10\cdot 3+5\cdot 6+1\cdot 16}{4+10+5+1}=\frac{80}{20}=4;$$

$$\theta=\frac{4\cdot|1-4|+10\cdot|3-4|+5\cdot|6-4|+1\cdot|16-4|}{20}=2,2.$$

Среднее абсолютное отклонение служит для характеристики рассеяния вариационного ряда.

Коэффициентом вариации V называют выраженное в процентах отношение выборочного среднего квадратического отклонения к выборочной средней:

$$V=\frac{\sigma_в}{\bar{x}_в}\cdot 100\%.$$

Коэффициент вариации служит для сравнения величин рассеяния двух вариационных рядов: тот из рядов имеет большее рассеяние, у которого коэффициент вариации больше.

Замечание. Выше предполагалось, что вариационный ряд составлен по данным выборки, поэтому все описанные характеристики называют *выборочными*; если вариационный ряд составлен по данным генеральной совокупности, то характеристики называют *генеральными*.

Задачи

1. Найти групповые средние совокупности состоящей из двух групп:

первая группа	x_i	0,1	0,4	0,6
	n_i	3	2	5;
вторая группа	x_i	0,1	0,3	0,4
	n_i	10	4	6.

Отв. $\overline{x_1} = 0{,}41$; $\overline{x_2} = 0{,}23$.

2. Найти общую среднюю по данным задачи 1 двумя способами: а) объединить обе группы в одну совокупность; б) использовать найденные в задаче 1 групповые средние.

Отв. $\overline{x} = 0{,}29$

3. Дано распределение статистической совокупности·

x_i	1	4	5
n_i	6	11	3.

Убедиться, что сумма произведений отклонений на соответствующие частоты равна нулю.

4. Дано распределение статистической совокупности:

x_i	4	7	10	15
n_i	10	15	20	5.

Найти дисперсию совокупности: а) исходя из определения дисперсии; б) пользуясь формулой $D = \overline{x^2} - [\overline{x}]^2$.

Отв. $D = 9{,}84$.

5. Найти внутригрупповую, межгрупповую и общую дисперсии совокупности, состоящей из трех групп:

первая группа	x_i	1	2	8
	n_i	30	15	5;
вторая группа	x_i	1	6	
	n_i	10	15;	
третья группа	x_i	3	8	
	n_i	20	5.	

Отв. $D_{\text{внгр}} = 4{,}6$; $D_{\text{межгр}} = 1$; $D_{\text{общ}} = 5{,}6$.

6. Найти внутригрупповую, межгрупповую и общую дисперсии совокупности, состоящей из двух групп:

первая группа	x_i	2	7
	n_i	6	4;
вторая группа	x_i	2	7
	n_i	2	8.

Отв. $D_{\text{внгр}} = 5$; $D_{\text{межгр}} = 1$; $D_{\text{общ}} = 6$.

7. Найти выборочную и исправленную дисперсии вариационного ряда, составленного по данным выборкам:

варианта	1	2	5	8	9
частота	3	4	6	4	3.

Отв. $\sigma_{\text{в}}^2 = 8{,}4$; $s^2 = 8{,}84$

В задачах 8—9 даны среднее квадратическое отклонение, выборочная средняя и объем выборки нормально распределенного признака. Найти доверительные интервалы для оценки неизвестного математического ожидания с заданной надежностью.

8. $\sigma = 2$, $\overline{x}_{\text{в}} = 5{,}40$, $n = 10$, $\gamma = 0{,}95$.

Отв. $4{,}16 < a < 6{,}64$.

9. $\sigma = 3$, $\overline{x}_{\text{в}} = 20{,}12$, $n = 25$, $\gamma = 0{,}99$.

Отв. $18{,}57 < a < 21{,}67$.

10. Найти минимальный объем выборки, при котором с надежностью 0,95 точность оценки математического ожидания нормально распределенного признака по выборочной средней будет равна 0,2, если среднее квадратическое отклонение равно 2

Указание. См замечание 2, § 15.

Отв. $n = 385$.

В задачах 11—12 даны «исправленное» среднее квадратическое отклонение, выборочная средняя и объем малой выборки нормально распределенного признака. Найти, пользуясь распределением Стьюдента, доверительные интервалы для оценки неизвестного математического ожидания с заданной надежностью.

11. $s = 1{,}5$, $\overline{x}_{\text{в}} = 16{,}8$, $n = 12$, $\gamma = 0{,}95$.

Отв. $15{,}85 < a < 17{,}75$.

12. $s = 2{,}4$, $\overline{x}_{\text{в}} = 14{,}2$, $n = 9$, $\gamma = 0{,}99$.

Отв. $11{,}512 < a < 16{,}888$.

13. По данным 16 независимых равноточных измерений физической величины найдены $\overline{x}_{\text{в}} = 23{,}161$ и $s = 0{,}400$. Требуется оценить истинное значение a измеряемой величины и точность измерений σ с надежностью 0,95.

Отв. $22{,}948 < a < 23{,}374$; $0{,}224 < \sigma < 0{,}576$.

Глава семнадцатая

МЕТОДЫ РАСЧЕТА СВОДНЫХ ХАРАКТЕРИСТИК ВЫБОРКИ

§ 1. Условные варианты

Предположим, что варианты выборки расположены в возрастающем порядке, т. е. в виде вариационного ряда.

Равноотстоящими называют варианты, которые образуют арифметическую прогрессию с разностью h.

Условными называют варианты, определяемые равенством:

$$u_i = \frac{x_i - C}{h},$$

где C — ложный нуль (новое начало отсчета);

h — шаг, т. е. разность между любыми двумя соседними первоначальными вариантами (новая единица масштаба).

Упрощенные методы расчета сводных характеристик выборки основаны на замене первоначальных вариант условными.

Покажем, что если вариационный ряд состоит из равноотстоящих вариант с шагом h, то условные варианты есть целые числа. Действительно, выберем в качестве ложного нуля произвольную варианту, например x_m. Тогда

$$u_i = \frac{x_i - x_m}{h} = \frac{x_1 + (i-1)h - [x_1 + (m-1)h]}{h} = i - m.$$

Так как i и m — целые числа, то их разность $i-m=u_i$ также есть целое число.

Замечание 1. В качестве ложного нуля можно принять любую варианту. Максимальная простота вычислений достигается, если выбрать в качестве ложного нуля варианту, которая расположена примерно в середине вариационного ряда (часто такая варианта имеет наибольшую частоту).

Замечание 2. Варианте, которая принята в качестве ложного нуля, соответствует условная варианта, равная нулю.

Пример. Найти условные варианты статистического распределения

варианты	23,6	28,6	33,6	38,6	43,6
частоты	5	20	50	15	10.

Решение. Выберем в качестве ложного нуля варианту 33,6 (эта варианта расположена в середине вариационного ряда).

Найдем шаг

$$h=28{,}6-23{,}6=5.$$

Найдем условную варианту

$$u_1 = \frac{x_1 - C}{h} = \frac{23{,}6 - 33{,}6}{5} = -2.$$

Аналогично получим

$$u_2=-1,\ u_3=0,\ u_4=1,\ u_5=2.$$

Мы видим, что условные варианты — небольшие целые числа. Разумеется, оперировать с ними проще, чем с первоначальными вариантами.

§ 2. Обычные, начальные и центральные эмпирические моменты

Для вычисления сводных характеристик выборки удобно пользоваться эмпирическими моментами, определения которых аналогичны определениям соответствующих теоретических моментов (гл. VIII, § 10). В отличие от теоретических, эмпирические моменты вычисляют по данным наблюдений.

Обычным эмпирическим моментом порядка k называют среднее значение k-х степеней разностей $x_i - C$:

$$M'_k = \frac{\sum n_i (x_i - C)^k}{n},$$

где x_i — наблюдаемая варианта,

n_i — частота варианты,

$n=\sum n_i$ — объем выборки,

C — произвольное постоянное число (ложный нуль).

Начальным эмпирическим моментом порядка k называют обычный момент порядка k при $C=0$

$$M_k = \frac{\sum n_i x_i^k}{n}.$$

В частности,

$$M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_в,$$

т. е. начальный эмпирический момент первого порядка равен выборочной средней.

Центральным эмпирическим моментом порядка k называют обычный момент порядка k при $C=\bar{x}_в$

$$m_k = \frac{\sum n_i (x_i - \bar{x}_в)^k}{n}.$$

В частности,

$$m_2 = \frac{\sum n_i (x_i - \bar{x}_в)^2}{n} = D_в, \qquad (*)$$

т. е. центральный эмпирический момент второго порядка равен выборочной дисперсии.

Легко выразить центральные моменты через обычные (рекомендуем читателю сделать это самостоятельно):

$$m_2 = M_2' - (M_1')^2; \qquad (**)$$

$$\left.\begin{aligned} m_3 &= M_3' - 3M_2'M_1' + 2(M_1')^3; \\ m_4 &= M_4' - 4M_3'M_1' + 6M_2'(M_1')^2 - 3(M_1')^4. \end{aligned}\right\} \qquad (***)$$

§ 3. Условные эмпирические моменты. Отыскание центральных моментов по условным

Вычисление центральных моментов требует довольно громоздких вычислений. Чтобы упростить расчеты, заменяют первоначальные варианты условными.

Условным эмпирическим моментом порядка k называют начальный момент порядка k, вычисленный для условных вариант:

$$M_k^* = \frac{\sum n_i u_i^k}{n} = \frac{\sum n_i \left(\frac{x_i - c}{h}\right)^k}{n}.$$

В частности,

$$M_1^* = \frac{\sum n_i \left(\frac{x_i - c}{h}\right)}{n} = \frac{1}{h}\left[\frac{\sum n_i x_i}{n} - c\,\frac{\sum n_i}{n}\right] =$$

$$= \frac{1}{h}(\bar{x}_в - c).$$

Отсюда

$$\bar{x}_в = M_1^* h + c. \qquad (*)$$

Таким образом, для того чтобы найти выборочную среднюю, достаточно вычислить условный момент первого порядка, умножить его на h и к результату прибавить ложный нуль c.

Выразим обычные моменты через условные:

$$M_k^* = \frac{1}{h^k} \cdot \frac{\sum n_i (x_i - c)^k}{n} = \frac{M_k'}{h^k}.$$

Отсюда

$$M_k' = M_k^* h^k.$$

Таким образом, для того чтобы найти обычный момент порядка k, достаточно условный момент того же порядка умножить на h^k.

Найдя же обычные моменты, легко найти центральные моменты по равенствам (**) и (***) предыдущего параграфа. В итоге получим удобные для вычислений формулы, выражающие центральные моменты через условные:

$$m_2 = \left[M_2^* - (M_1^*)^2\right] h^2; \qquad (**)$$

$$\left.\begin{aligned} m_3 &= \left[M_3^* - 3M_2^*M_1^* + 2(M_1^*)^3\right] h^3; \\ m_4 &= \left[M_4^* - 4M_3^*M_1^* + 6M_2^*(M_1^*)^2 - 3(M_1^*)^4\right] h^4. \end{aligned}\right\} \qquad (***)$$

В частности, в силу (**) и соотношения (*) предыдущего параграфа получим формулу для вычисления выборочной дисперсии по условным моментам первого и второго порядков

$$D_в = \left[M_2^* - (M_1^*)^2\right] h^2. \qquad (****)$$

Техника вычислений центральных моментов по условным описана далее.

§ 4. Метод произведений вычисления выборочных средней и дисперсии

Метод произведений дает удобный способ вычисления условных моментов различных порядков вариационного ряда с равностоящими вариантами. Зная же условные моменты, нетрудно найти интересующие нас начальные и центральные эмпирические моменты. В частности, методом произведений удобно вычислять выборочную среднюю и выборочную дисперсию. Целесообразно пользоваться расчетной таблицей, которая составляется так:

1) в первый столбец таблицы записывают выборочные (первоначальные) варианты, располагая их в возрастающем порядке;

2) во второй столбец записывают частоты вариант; складывают все частоты и их сумму (объем выборки n) помещают в нижнюю клетку столбца;

3) в третий столбец записывают условные варианты $u_i = \frac{x_i - C}{h}$, причем в качестве ложного нуля C выбирают варианту с наибольшей частотой и полагают h равным разности между любыми двумя соседними вариантами; практически же третий столбец заполняется так: в клетке строки, содержащей наибольшую частоту, пишут 0; в клетках над нулем пишут последовательно —1, —2, —3 и т. д., а под нулем 1, 2, 3 и т. д.;

4) умножают частоты на условные варианты и записывают их произведения $n_i u_i$ в четвертый столбец; сложив все полученные числа, их сумму $\sum n_i u_i$ помещают в нижнюю клетку столбца;

5) умножают частоты на квадраты условных вариант и записывают их произведения $n_i u_i^2$ в пятый столбец; сложив все полученные числа, их сумму $\sum n_i u_i^2$, помещают в нижнюю клетку столбца;

6) умножают частоты на квадраты условных вариант, увеличенных каждая на единицу, и записывают произведения $n_i (u_i + 1)^2$ в шестой контрольный столбец; сложив все полученные числа, их сумму $\sum n_i (u_i + 1)^2$ помещают в нижнюю клетку столбца.

Замечание 1. Целесообразно отдельно складывать отрицательные числа четвертого столбца (их сумму A_1 записывают в клетку строки, содержащей наибольшую частоту) и отдельно положительные (их сумму A_2 записывают в предпоследнюю клетку столбца); тогда $\sum n_i u_i = A_1 + A_2$.

Замечание 2. При вычислении произведений $n_i u_i^2$ пятого столбца, целесообразно числа $n_i u_i$ четвертого столбца умножать на u_i.

Замечание 3. Шестой столбец служит для контроля вычислений: если сумма $\sum n_i (u_i + 1)^2$ окажется равной сумме $\sum n_i u_i^2 + 2 \sum n_i u_i + n$ (как и должно быть в соответствии с тождеством $\sum n_i (u_i + 1)^2 = \sum n_i u_i^2 + 2 \sum n_i u_i + n$), то вычисления проведены правильно.

Замечание 4. В качестве ложного нуля может быть взята любая варианта, т. е. не обязательно брать варианту, имеющую наибольшую частоту, как указано в п. 3. Например, если варианта, которая имеет наибольшую частоту, расположена в первых, или последних строках «столбца x_i», то *выгоднее принять в качестве ложного нуля варианту, которая находится примерно в середине столбца.*

После того, как расчетная таблица заполнена и проверена правильность вычислений, вычисляют условные моменты:

$$M_1^* = \frac{\sum n_i u_i}{n}, \qquad M_2^* = \frac{\sum n_i u_i^2}{n}.$$

Наконец, вычисляют выборочные среднюю и дисперсию по формулам (*) и (****) § 3:

$$\bar{x}_B = M_1^* \cdot h + C,$$

$$D_B = \left[M_2^* - (M_1^*)^2\right] h^2.$$

Пример. Найти методом произведений выборочные среднюю и дисперсию следующего статистического распределения:

варианты:	10,2	10,4	10,6	10,8	11,0	11,2	11,4	11,6	11,8	12,0
частота:	2	3	8	13	25	20	12	10	6	1

Решение. Составим расчетную таблицу, для чего:

1) запишем варианты в первый столбец;

2) запишем частоты во второй столбец; сумму частот (100) поместим в нижнюю клетку столбца;

3) в качестве ложного нуля выберем варианту 11,0 (эта варианта имеет наибольшую частоту); в клетке третьего столбца, которая принадлежит строке, содержащей наибольшую частоту, пишем 0; над нулем последовательно —1, —2, —3, —4, а под нулем 1, 2, 3, 4, 5;

4) произведения частот на условные варианты записываем в четвертый столбец; отдельно находим сумму (—46) отрицательных и отдельно сумму (103) положительных чисел; сложив эти числа, их сумму (57) помещаем в нижнюю клетку столбца;

5) произведения частот на квадраты условных вариант запишем в пятый столбец; сумму чисел столбца (383) помещаем в нижнюю клетку столбца;

6) произведения частот на квадраты условных вариант, увеличенных на единицу, запишем в шестой контрольный столбец; сумму (597) чисел столбца помещаем в нижнюю клетку столбца.

В итоге получим расчетную таблицу 7.

Таблица 7

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i (u_i+1)^2$
10,2	2	−4	−8	32	18
10,4	3	−3	−9	27	12
10,6	8	−2	−16	32	8
10,8	13	−1	−13	13	0
11,0	25	0	$A_1 = -46$		25
11,2	20	1	20	20	80
11,4	12	2	24	48	108
11,6	10	3	30	90	160
11,8	6	4	24	96	150
12,0	1	5	5	25	36
			$A_2 = 103$		
	$n=100$		$\Sigma n_i u_i = 57$	$\Sigma n_i u_i^2 = 383$	$\Sigma n_i (u_i+1)^2 = 597$

Контроль: $\sum n_i u_i^2 + 2\sum n_i u_i + n = 383 + 2\cdot 57 + 100 = 597$

$$\sum n_i (u_i+1)^2 = 597.$$

Вычисления произведены правильно.

Вычислим условные моменты первого и второго порядков:

$$M_1^* = \frac{\sum n_i u_i}{n} = \frac{57}{100} = 0{,}57;$$

$$M_2^* = \frac{\sum n_i u_i^2}{n} = \frac{383}{100} = 3{,}83$$

Найдем шаг: $h = 10{,}4 - 10{,}2 = 0{,}2$.

Вычислим искомые выборочные среднюю и дисперсию:

$$\bar{x}_в = M_1^* \cdot h + C = 0{,}57 \cdot 0{,}2 + 11{,}0 = 11{,}1,$$

$$D_в = [M_2^* - (M_1^*)^2] \cdot h^2 = [3{,}83 - (0{,}57)^2] \cdot 0{,}2^2 = 0{,}14.$$

§ 5. Сведение первоначальных вариант к равноотстоящим

Выше изложена методика расчета выборочных характеристик для равноотстоящих вариант. На практике, как правило, данные наблюдений не будут равноотстоящими числами. Естественно, возникает вопрос: нельзя ли соответствующей обработкой наблюдаемых значений признака свести вычисления к случаю равноотстоящих вариант? Оказывается, можно. С этой целью интервал, в котором заключены все наблюдаемые значения признака (первоначальные варианты), делят на несколько равных частичных интервалов. (Практически в каждый частичный интервал должно попасть не менее 8—10 первоначальных вариант.) Затем находят середины частичных интервалов, которые и образуют последовательность равноотстоящих вариант.

В качестве частоты каждой «новой» варианты (середины частичного интервала) принимают общее число первоначальных вариант, попавших в соответствующий частичный интервал.

Ясно, что замена первоначальных вариант серединами частичных интервалов сопровождается ошибками (первоначальные варианты левой половины частичного интервала будут увеличены, а варианты правой половины уменьшены), однако эти ошибки будут в основном погашаться, поскольку они имеют разные знаки.

Пример. Выборочная совокупность объема $n=100$ задана таблицей 8.

Таблица 8

x_i	n_i	x_i	n_i	x_i	n_i
1,00	1	1,19	2	1,37	6
1,03	3	1,20	4	1,38	2
1,05	6	1,23	4	1,39	1
1,06	4	1,25	8	1,40	2
1,08	2	1,26	4	1,44	3
1,10	4	1,29	4	1,45	3
1,12	3	1,30	6	1,46	2
1,15	6	1,32	4	1,49	4
1,16	5	1,33	5	1,50	2

Составить распределение равноотстоящих вариант.

Решение. Разобьем интервал 1,00—1,50, например, на следующие 5 частичных интервалов:

1,00—1,10; 1,10—1,20; 1,20—1,30; 1,30—1,40; 1,40—1,50. Приняв середины частичных интервалов в качестве новых вариант y_i, получим равноотстоящие варианты:

$$y_1=1{,}05;\quad y_2=1{,}15;\quad y_3=1{,}25;\quad y_4=1{,}35;\quad y_5=1{,}45$$

Найдем частоту варианты y_1:

$$n_1=1+3+6+4+2+\frac{4}{2}=18$$

(Поскольку первоначальная варианта 1,10 одновременно является концом первого частичного интервала и началом второго, частота 4 этой варианты поровну распределена между обоими частичными интервалами.)

Найдем частоту варианты y_2:

$$n_2=\frac{4}{2}+3+6+5+2+\frac{4}{2}=20.$$

Аналогично вычислим частоты остальных вариант:

$$n_3=25;\quad n_4=22;\quad n_5=15$$

В итоге получим следующее распределение равноотстоящих вариант:

y_i	1,05	1,15	1,25	1,35	1,45
n_i	18	20	25	22	15

Рекомендуем читателю в порядке упражнения убедиться, что выборочные средние и дисперсии, вычисленные по первоначальным и равноотстоящим вариантам, окажутся соответственно равными:

$$\bar{x}_в=1{,}250;\quad \bar{y}_в=1{,}246;$$
$$D_x=0{,}018;\quad D_y=0{,}017$$

Как видим, замена первоначальных вариант равноотстоящими не привела к существенным ошибкам; при этом объем вычислительной работы значительно уменьшается.

§ 6 Эмпирические и выравнивающие (теоретические) частоты

А. Дискретное распределение

Рассмотрим дискретную случайную величину X, закон распределения которой неизвестен. Пусть произведено n испытаний, в которых величина X приняла n_1 раз значение x_1, n_2 раз значение x_2, ..., n_k раз значение x_k, причем $\sum n_i=n$.

Эмпирическими частотами называют фактически наблюдаемые частоты n_i.

Пусть имеются основания предположить, что изучаемая величина X распределена по некоторому определенному закону. Для того чтобы проверить, согласуется ли это предположение с данными наблюдений, в ы ч и с л я ю т частоты наблюдаемых значений, т. е. н а х о д я т т е о р е т и ч е с к и сколько раз величина X должна была принять каждое из наблюдаемых значений, если она распределена по предполагаемому закону.

Выравнивающими (теоретическими), в отличие от фактически наблюдаемых эмпирических частот, называют частоты n'_i, найденные теоретически (вычислением).

Выравнивающие частоты находят по равенству

$$n'_i=nP_i,$$

где n — число испытаний,

P_i — вероятность наблюдаемого значения x_i, вычисленная при допущении, что X имеет предполагаемое распределение. Эта формула следует из теоремы о математическом ожидании числа появлений события в независимых испытаниях (гл. VII, § 5).

Итак, в ы р а в н и в а ю щ а я ч а с т о т а н а б л ю д а е м о г о з н а ч е н и я x_i, д и с к р е т н о г о р а с п р е д е л е н и я р а в н а п р о и з в е д е н и ю ч и с л а и с п ы т а н и й н а в е р о я т н о с т ь э т о г о н а б л ю д а е м о г о з н а ч е н и я.

Пример. В результате эксперимента, состоящего из $n=520$ испытаний, в каждом из которых регистрировалось число x_i появлений некоторого события, получено следующее эмпирическое распределение:

набл. знач. x_i	0	1	2	3	4	5	6	7
эмп. частота n_i	120	167	130	69	27	5	1	1.

Найти выравнивающие частоты n'_i, в предположении, что случайная величина X (генеральная совокупность) распределена по закону Пуассона.

Р е ш е н и е. Известно, что параметр λ, которым определяется распределение Пуассона, равен математическому ожиданию этого распределения. Поскольку в качестве оценки математического ожидания принимают выборочную среднюю (гл. XVI, § 5), то и в качестве оценки λ можно принять выборочную среднюю $\bar{x}_в$. Легко найти по условию, что выборочная средняя равна 1,5; следовательно, можно принять $\lambda=1,5$.

Таким образом, формула Пуассона

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

принимает вид:

$$P_{520}(k)=\frac{1,5^k\cdot e^{-1,5}}{k!}.$$

Пользуясь этой формулой, найдем вероятности $P_{520}(k)$ при $k=0$, 1, 2, 3, 4, 5, 6, 7 (для простоты записи индекс 520 далее опущен): $P(0)=0,22313$, $P(1)=0,33469$, $P(2)=0,251021$, $P(3)=0,125511$, $P(4)=0,047066$, $P(5)=0,014120$, $P(6)=0,003530$, $P(7)=0,000755$.

Найдем выравнивающие частоты (результаты умножения округлены до единицы):

$$n_1'=n\cdot P(0)=520\cdot 0,22313=116,$$
$$n_2'=n\cdot P(1)=520\cdot 0,33469=174.$$

Аналогично находят и остальные выравнивающие частоты. В итоге получим:

эмп. частота	123	167	130	69	27	5	1	1
выр. частота	116	174	131	65	25	7	2	0.

Сравнительно небольшое расхождение эмпирических и выравнивающих частот подтверждает предположение, что рассматриваемое распределение подчинено закону Пуассона.

Заметим, что если подсчитать выборочную дисперсию по данному распределению, то окажется, что она равна выборочной средней, т. е. 1,5. Это служит еще одним подтверждением сделанного предположения, поскольку для распределения Пуассона $\lambda=M(X)=D(X)$.

Б. Непрерывное распределение

В случае непрерывного распределения, вероятности отдельных возможных значений равны нулю (гл. X, § 2, следствие 2). Поэтому весь интервал возможных значений делят на k непересекающихся интервалов и вычисляют вероятности P_i попадания X в i-й частичный интервал, а затем, как и для дискретного распределения, умножают число испытаний на эти вероятности.

Итак, выравнивающие частоты непрерывного распределения находят по равенству

$$n_i'=nP_i,$$

где n — число испытаний,

P_i — вероятность попадания X в i-й частичный интервал, вычисленная при допущении, что X имеет предполагаемое распределение.

В частности, если имеются основания предположить, что случайная величина X (генеральная совокупность) распределена нормально, то выравнивающие частоты могут быть найдены по формуле

$$n_i'=\frac{nh}{\sigma_в}\varphi(u_i), \qquad (*)$$

где n — число испытаний (объем выборки),

h — длина частичного интервала,

$\sigma_в$ — выборочное среднее квадратическое отклонение,

$u_i=\frac{x_i-\bar{x}_в}{\sigma_в}$ (x_i — середина i-го частичного интервала),

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}.$$

Пример на применение формулы (*) приведен в § 7.

Пояснение. Поясним происхождение формулы (*). Напишем дифференциальную функцию общего нормального распределения

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

При $a=0$ и $\sigma=1$ получим дифференциальную функцию нормированного распределения

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

или, изменив обозначение аргумента,

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}.$$

Положив $u = \frac{x-a}{\sigma}$, имеем

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

Сравнивая (**) и (***), заключаем, что

$$f(x) = \frac{1}{\sigma}\,\varphi(u)$$

Если математическое ожидание a и среднее квадратическое отклонение σ неизвестны, то в качестве оценок этих параметров принимают соответственно выборочную среднюю $\bar{x}_в$ и выборочное среднее квадратическое отклонение $\sigma_в$ (гл. XVI, § 5, 9)
Тогда

$$f(x) = \frac{1}{\sigma_в}\,\varphi(u),$$

где $\quad u = \frac{x - \bar{x}_в}{\sigma_в}$.

Пусть x_i — середина i-го интервала (на которые разбита совокупность всех наблюдаемых значений нормально распределенной случайной величины X) длиною h. Тогда вероятность попадания X в этот интервал приближенно равна произведению длины интервала на значение дифференциальной функции $f(x)$ в любой точке интервала и, в частности, при $x = x_i$ (гл. XI, § 5):

$$P_i = hf(x_i) = h \cdot \frac{1}{\sigma_в}\,\varphi(u_i).$$

Следовательно, выравнивающая частота

$$n_i' = nP_i = \frac{nh}{\sigma_в}\,\varphi(u_i),$$

где

$$u_i = \frac{x_i - \bar{x}_в}{\sigma_в}.$$

Мы получили формулу (*)

§ 7. Построение нормальной кривой по опытным данным

Один из способов построения нормальной кривой по данным наблюдений состоит в следующем.

1) находят $\bar{x}_в$ и $\sigma_в$, например по методу произведений;
2) находят ординаты y_i (выравнивающие частоты) теоретической кривой по формуле $y_i = \frac{n \cdot h}{\sigma_в} \cdot \varphi(u_i)$, где n — сумма наблюдаемых частот, h — разность между двумя соседними вариантами $u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$ и $\varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$;
3) строят точки (x_i, y_i) в прямоугольной системе координат и соединяют их плавной кривой.

Близость выравнивающих частот к наблюдаемым подтверждает правильность допущения о том, что обследуемый признак распределен нормально.

Пример. Построить нормальную кривую по данному распределению:

варианта x_i 15 20 25 30 35 40 45 50 55
частота n_i 6 13 38 74 106 85 30 10 4.

Решение. Пользуясь методом произведений (§ 4), найдем $\bar{x}_в = 34{,}7$, $\sigma_в = 7{,}38$.

Вычислим выравнивающие частоты (см. таблицу 9).

Таблица 9

x_i	n_i	$x_i - \bar{x}_в$	$u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$	$\varphi(u_i)$	$y_i = \frac{n \cdot h}{\sigma_в} \cdot \varphi(u_i) = 248 \cdot \varphi(u_i)$
15	6	−19,7	−2,67	0,0113	3
20	13	−14,7	−1,99	0,0551	14
25	38	−9,7	−1,31	0,1691	42
30	74	−4,7	−0,63	0,3271	82
35	106	0,3	0,05	0,3984	99
40	85	5,3	0,73	0,3056	76
45	30	10,3	1,41	0,1476	37
50	10	15,3	2,09	0,0449	11
55	4	20,3	2,77	0,0086	2
	$n = 366$				$\Sigma y_i = 366$

На рис. 22 построены нормальная (теоретическая) кривая по выравнивающим частотам (они отмечены кружками) и полигон наблюдаемых частот (они отмечены крестиками). Сравнение графиков наглядно показывает, что построенная теоретическая кривая удовлетворительно отражает данные наблюдений.

Для того чтобы более уверенно считать, что данные наблюдений свидетельствуют о нормальном распределении

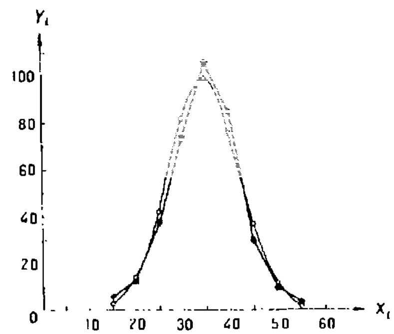

Рис. 22.

признака, пользуются специальными правилами (их называют критериями согласия), понятие о которых читатель найдет далее (гл. XIX, § 22).

§ 8. Оценка отклонения эмпирического распределения от нормального. Асимметрия и эксцесс

Для оценки отклонения эмпирического распределения от нормального используют различные характеристики, к числу которых относятся асимметрия и эксцесс. Определения этих характеристик аналогичны определениям асимметрии и эксцесса теоретического распределения (гл. XII, § 9).

Асимметрия эмпирического распределения определяется равенством:

$$a_s = \frac{m_3}{\sigma_в^3}.$$

где m_3 — центральный эмпирический момент третьего порядка (§ 2).

Эксцесс эмпирического распределения определяется равенством:

$$e_k = \frac{m_4}{\sigma_в^4} - 3,$$

где m_4 — центральный эмпирический момент четвертого порядка.

Моменты m_3 и m_4 удобно вычислять методом произведений (§ 4), используя формулы (***) § 3.

Пример. Найти асимметрию и эксцесс эмпирического распределения:

варианта	10,2	10,4	10,6	10,8	11,0	11,2	11,4	11,6	11,8	12,0
частота	2	3	8	13	25	20	12	10	6	1

Решение. Воспользуемся методом произведений, для чего составим расчетную таблицу. Поскольку в § 4 указано, как заполняются столбцы 1—5 таблицы, ограничимся краткими пояснениями: для заполнения столбца 6 удобно перемножать числа каждой строки столбцов 3 и 5; для заполнения столбца 7 удобно перемножать числа каждой строки столбцов 3 и 6. Столбец 8 служит для контроля вычислений по тождеству:

$$\sum n_i (u_i + 1)^4 = \sum n_i u_i^4 + 4 \sum n_i u_i^3 + 6 \sum n_i u_i^2 + 4 \sum n_i u_i + n.$$

Приведем расчетную таблицу 10.

Контроль: $\sum n_i (u_i + 1)^4 = 9141$;

$$\sum n_i u_i^4 + 4 \sum n_i u_i^3 + 6 \sum n_i u_i^2 + 4 \sum n_i u_i + n = 4079 + 4 \cdot 609 + 6 \cdot 383 + 4 \cdot 57 + 100 = 9141.$$

Совпадение сумм свидетельствует о том, что вычисления произведены правильно.

В примере § 4 для рассматриваемого распределения было найдено:

$$M_1^* = 0{,}57; \quad M_2^* = 3{,}83; \quad D_в = 0{,}14,$$

следовательно,

$$\sigma_в = \sqrt{0.14}.$$

Таблица 10

1	2	3	4	5	6	7	8
x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	$n_i (u_i+1)^4$
10,2	2	—4	—8	32	—128	512	162
10,4	3	—3	—9	27	—81	243	48
10,6	8	—2	—16	32	—64	128	8
10,8	13	—1	—13	13	—13	13	—
11,0	25	0	—46		—286		25
11,2	20	1	20	20	20	20	320
11,4	12	2	24	48	96	192	972
11,6	10	3	30	90	270	810	2560
11,8	6	4	24	96	384	1536	3750
12,0	1	5	5	25	125	625	1296
			103		895		
	$n=100$		$\sum n_i u_i = 57$	$\sum n_i u_i^2 = 383$	$\sum n_i u_i^3 = 609$	$\sum n_i u_i^4 = 4079$	$\sum n_i (u_i + 1)^4 = 9141$

Найдем условные моменты третьего и четвертого порядка:

$$M_3^* = \frac{\sum n_i u_i^3}{n} = \frac{609}{100} = 6,09; \quad M_4^* = \frac{\sum n_i u_i^4}{n} = \frac{4079}{100} = 40,79.$$

Найдем центральные эмпирические моменты третьего и четвертого порядка:

$$m_3 = \left[M_3^* - 3M_1^* M_2^* + 2\left(M_1^*\right)^3\right] h^3 =$$
$$= [6,09 - 3 \cdot 0,57 \cdot 3,83 + 2 \cdot (0,57)^3] \cdot 0,2^3 = -0,0007;$$
$$m_4 = \left[M_4^* - 4M_1^* M_3^* + 6\left(M_1^*\right)^2 M_2^* - 3\left(M_1^*\right)^4\right] h^4 =$$
$$= [40,79 - 4 \cdot 0,57 \cdot 6,09 + 6\,(0,57)^2 \cdot 3,83 - 3 \cdot (0,57)^4] \cdot 0,2^4 = 0,054.$$

Найдем асимметрию и эксцесс:

$$a_s = \frac{m_3}{\sigma_в^3} = -\frac{0,0007}{(\sqrt{0,14})^3} = -0,01;$$
$$e_k = \frac{m_4}{\sigma_в^4} - 3 = \frac{0,054}{(\sqrt{0,14})^4} - 3 = -0,24.$$

Замечание. В случае малых выборок к оценкам асимметрии и эксцесса следует относиться с осторожностью и находить точность этих оценок (см. Н. В. Смирнов и И. В. Дунин-Барковский. Курс теории вероятностей и математической статистики. Наука, 1965, стр. 277).

Задачи

В задачах 1—2 даны выборочные варианты и их частоты. Найти, пользуясь методом произведений, выборочные среднюю и дисперсию.

1.

x_i	10,3	10,5	10,7	10,9	11,1	11,3	11,5	11,7	11,9	12,1
n_i	4	7	8	10	25	15	12	10	4	5.

Отв. $\bar{x}_в = 11,19$, $D_в = 0,19$.

2.

x_i	83	85	87	89	91	93	95	97	99	101
n_i	6	7	12	15	30	10	8	6	4	2.

Отв. $\bar{x}_в = 90,72$, $D_в = 17,20$.

3. Найти асимметрию и эксцесс эмпирического распределения

x_i	10,6	10,8	11,0	11,2	11,4	11,6	11,8
n_i	5	10	17	30	20	12	6.

Отв. $a_s = -0,0006$, $e_k = 0,00004$.

Глава восемнадцатая

ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

§ 1. Функциональная, статистическая и корреляционная зависимости

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин. Рассмотрим сначала зависимость Y от одной случайной (или неслучайной) величины X, а затем от нескольких величин (§ 15).

Две случайные величины могут быть связаны функциональной зависимостью (гл. XII, § 10), либо зависимостью другого рода, называемой статистической, либо быть независимыми.

Строгая функциональная зависимость реализуется редко, так как обе величины или одна из них подвержены еще действию случайных факторов, причем среди них могут быть и общие для обеих величин (под «общими» здесь подразумеваются такие факторы, которые воздействуют и на Y и на X). В этом случае возникает статистическая зависимость.

Например, если Y зависит от случайных факторов

$$Z_1,\ Z_2,\ V_1,\ V_2,$$

а X зависит от случайных факторов

$$Z_1,\ Z_2,\ U_1,$$

то между Y и X имеется статистическая зависимость, так как среди случайных факторов есть общие, а именно Z_1 и Z_2.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют *корреляционной*.

Приведем пример случайной величины Y, которая не связана с величиной X функционально, а связана корреляционно. Пусть Y — урожай зерна, X — количество удобрений. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т. е. Y не является функцией от X. Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, *средний* урожай является функцией от количества удобрений, т. е. Y связан с X корреляционной зависимостью.

§ 2. Условные средние. Корреляционная зависимость

Уточним определение корреляционной зависимости, для чего введем понятие условной средней.

Предположим, что изучается связь между случайной величиной Y и случайной величиной X. Пусть каждому значению X соответствует несколько значений Y. Например, пусть при $x_1=2$ величина Y приняла значения: $y_1=5$, $y_2=6$, $y_3=10$. Найдем среднее арифметическое этих чисел:

$$\overline{y}_2 = \frac{5+6+10}{3} = 7.$$

Число $\overline{y}_2$ называют условным средним; черточка над буквой y служит обозначением среднего арифметического, а число 2 указывает, что рассматриваются те значения Y, которые соответствуют $x_1=2$.

Применительно к примеру предыдущего параграфа эти данные можно истолковать так: на каждый из трех одинаковых участков земли внесли по 2 единицы удобрений и сняли соответственно 5; 6 и 10 единиц зерна; средний урожай составил 7 соответствующих единиц.

Условным средним $\overline{y}_x$ называют среднее арифметическое значений Y, соответствующих значению $X=x$.

Если каждому значению x соответствует одно значение условной средней, то, очевидно, условная средняя есть функция от x; в этом случае говорят, что случайная величина Y зависит от X корреляционно.

Корреляционной зависимостью Y от X называют функциональную зависимость условной средней $\overline{y}_x$ от x:

$$\overline{y}_x = f(x). \qquad (*)$$

Уравнение (*) называют *уравнением регрессии* Y на X; функцию $f(x)$ называют *регрессией* Y на X, а ее график — *линией регрессии* Y на X.

Аналогично определяется условная средняя $\overline{x}_y$, и корреляционная зависимость X от Y.

Условным средним $\overline{x}_y$ называют среднее арифметическое значений X, соответствующих $Y=y$.

Корреляционной зависимостью X от Y называют функциональную зависимость условной средней $\overline{x}_y$ от y:

$$\overline{x}_y = \varphi(y). \qquad (**)$$

Уравнение (**) называют *уравнением регрессии* X на Y; функцию $\varphi(y)$ называют *регрессией* X на Y, а ее график — *линией регрессии* X на Y.

§ 3. Две основные задачи теории корреляции

Первая задача теории корреляции — установить форму корреляционной связи, т. е. вид функции регрессии (линейная, квадратичная показательная и т. д.). Наиболее часто функции регрессии оказываются линейными. Если обе функции регрессии $f(x)$ и $\varphi(y)$ линейны, то корреляцию называют *линейной*; в противном случае — *нелинейной*. Очевидно, при линейной корреляции обе линии регрессии являются прямыми линиями.

Вторая задача теории корреляции — оценить тесноту (силу) корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине рассеяния значений Y вокруг условного среднего $\overline{y}_x$. Большое рассеяние свидетельствует о слабой зависимости Y от X либо об отсутствии зависимости. Малое рассеяние указывает наличие достаточно сильной зависимости; возможно даже, что Y и X связаны функционально, но под воздействием второстепенных случайных факторов эта связь оказалась размытой, в результате чего при одном и том же значении x величина Y принимает различные значения.

Аналогично (по величине рассеяния значений X вокруг условного среднего $\overline{x}_y$) оценивается теснота корреляционной связи X от Y.

§ 4. Отыскание параметров выборочного уравнения прямой линии регрессии по несгруппированным данным

Допустим, что количественные признаки X и Y связаны линейной корреляционной зависимостью. В этом случае обе линии регрессии будут прямыми.

Предположим, что для отыскания уравнений этих прямых проведено n независимых испытаний, в результате которых получены n пар чисел:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Поскольку наблюдаемые пары чисел можно рассматривать как случайную выборку из генеральной совокупности всех возможных значений случайной величины (X, Y), то величины и уравнения, найденные по этим данным, называют *выборочными*.

Для определенности будем искать выборочное уравнение прямой линии регрессии Y на X.

Рассмотрим простейший случай: различные значения x признака X и соответствующие им значения y признака Y наблюдались по одному разу. Очевидно, что группировать данные нет необходимости. Также нет надобности использовать понятие условной средней, поэтому искомое уравнение

$$\overline{y}_x = kx + b$$

можно записать так:

$$Y = kx + b.$$

Угловой коэффициент прямой линии регрессии Y на X принято называть *выборочным коэффициентом регрессии* Y на X и обозначать через ρ_{yx}.

Итак, будем искать выборочное уравнение прямой линии регрессии Y на X вида:

$$Y = \rho_{yx} x + b. \qquad (*)$$

Поставим своей задачей подобрать параметры ρ_{yx} и b так, чтобы точки (x_1, y_1) (x_2, y_2), ..., (x_n, y_n), построенные по данным наблюдений на плоскости XOY, как можно ближе лежали вблизи прямой (*).

Уточним смысл этого требования. Назовем отклонением разность

$$Y_i - y_i \quad (i = 1, 2, \ldots, n),$$

где Y_i — вычисленная по уравнению (*) ордината, соответствующая наблюдаемому значению x_i;

y_i — наблюдаемая ордината, соответствующая x_i.

Подберем параметры ρ_{yx} и b так, чтобы сумма квадратов отклонений была минимальной (в этом состоит сущность метода наименьших квадратов).

Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих параметров (временно вместо ρ_{yx} будем писать ρ):

$$F(\rho, b) = \sum_{i=1}^{n} (Y_i - y_i)^2,$$

или

$$F(\rho, b) = \sum_{i=1}^{n} (\rho x_i + b - y_i)^2.$$

Для отыскания минимума приравняем нулю соответствующие частные производные:

$$\frac{\partial F}{\partial \rho} = 2 \sum_{i=1}^{n} (\rho x_i + b - y_i)\, x_i = 0;$$

$$\frac{\partial F}{\partial b} = 2 \sum_{i=1}^{n} (\rho x_i + b - y_i) = 0.$$

Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно ρ и b*

* Для простоты записи вместо $\sum\limits_{i=1}^{n}$ будем писать $\sum$.

$$\left(\sum x^2\right)\rho+\left(\sum x\right)b=\sum xy;\quad \left(\sum x\right)\rho+nb=\sum y. \qquad (**)$$

Решив эту систему, найдем искомые параметры:

$$\rho_y=\frac{n\sum xy-\sum x\cdot\sum y}{n\sum x^2-\left(\sum x\right)^2};$$

$$b=\frac{\sum x^2\cdot\sum y-\sum x\cdot\sum xy}{n\sum x^2-\left(\sum x\right)^2}. \qquad (***)$$

Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y:

$$\bar{x}_y=\rho_{xy}x+C,$$

где ρ_{xy}— выборочный коэффициент регрессии X на Y.

Пример. Найти выборочное уравнение прямой линии регрессии Y на X по данным $n=5$ наблюдений:

x 1,00 1,50 3,00 4,50 5,00;
y 1,25 1,40 1,50 1,75 2,25.

Решение. Составим расчетную таблицу 11.

Таблица 11

x_i	y_i	x_i^2	$x_i\cdot y_i$
1,00	1,25	1,00	1,250
1,50	1,40	2,25	2,100
3,00	1,50	9,00	4,500
4,50	1,75	20,25	4,875
5,00	2,25	25,00	11,250
$\sum x_i=15$	$\sum y_i=8,15$	$\sum x_i^2=57,50$	$\sum x_iy_i=26,975$

Найдем искомые параметры, для чего подставим вычисленные по таблице суммы в соотношения (***):

$$\rho_{yx}=\frac{5\cdot 26,975-15\cdot 8,15}{5\cdot 57,5-15^2}=0,202;$$

$$b=\frac{57,5\cdot 8,15-15\cdot 26,975}{62,5}=1,024.$$

Напишем искомое уравнение регрессии:

$$Y=0,202x+1,024.$$

Для того чтобы получить представление, насколько хорошо вычисленные по этому уравнению значения Y_i согласуются с наблюдаемыми значениями y_i, найдем отклонения Y_i-y_i. Результаты вычислений сведены в таблицу 12.

Таблица 12

x_i	Y_i	y_i	Y_i-y_i
1,00	1,226	1,25	−0,024
1,50	1,327	1,40	−0,073
3,00	1,630	1,50	0,130
4,50	1,933	1,75	0,083
5,00	2,034	2,25	−0,216

Как видно из таблицы, не все отклонения достаточно малы. Это объясняется малым числом наблюдений.

§ 5. Корреляционная таблица

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение y может встретиться n_y раз, одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т. е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют *корреляционной*.

Поясним устройство корреляционной таблицы на примере (табл. 13).

Таблица 13

Y \ X	10	20	30	40	n_y
0,4	5	—	7	14	26
0,6	—	2	6	4	12
0,8	3	19	—	—	22
n_x	8	21	13	18	$n=60$

В первой строке таблицы указаны наблюдаемые значения (10; 20; 30; 40) признака X, а в первом столбце — наблюдаемые значения (0,4; 0,6; 0,8) признака Y. На пересечении строк и столбцов вписаны частоты n_{xy} наблюдаемых пар значений признаков. Например, частота 5 указывает, что пара чисел (10; 0,4) наблюдалась 5 раз. Все частоты помещены в прямоугольнике, стороны которого проведены жирными отрезками. Черточка означает, что соответственная пара чисел, например (20; 0,4), не наблюдалась.

В последнем столбце записаны суммы частот строк. Например, сумма частот первой строки жирного прямоугольника равна $n_y=5+7+14=26$; это число указывает, что значение признака Y, равное 0,4 (в сочетании с различными значениями признака X) наблюдалось 26 раз.

В последней строке записаны суммы частот столбцов. Например, число 8 указывает, что значение признака X, равное 10 (в сочетании с различными значениями признака Y) наблюдалось 8 раз.

В клетке, расположенной в нижнем правом углу таблицы, помещена сумма всех частот (общее число всех наблюдений n). Очевидно $\Sigma n_x=\Sigma n_y=n$. В нашем примере

$$\Sigma n_x=8+21+13+18=60 \quad \text{и} \quad \Sigma n_y=26+12+22=60.$$

§ 6. Отыскание параметров выборочного уравнения прямой линии регрессии по сгруппированным данным. Выборочный коэффициент корреляции

В § 4 для определения параметров уравнения прямой линии регрессии Y на X была получена система уравнений:

$$\left.\begin{aligned}(\textstyle\sum x^2)\rho_{yx} + (\sum x)b &= \textstyle\sum xy;\\ (\textstyle\sum x)\rho_{yx} + nb &= \textstyle\sum y.\end{aligned}\right\} \qquad (*)$$

Предполагалось, что значения X и соответствующие им значения Y наблюдались по одному разу. Теперь же допустим, что получено большое число данных (практически для удовлетворительной оценки искомых параметров должно быть хотя бы 50 наблюдений), среди них есть повторяющиеся, и они сгруппированы в виде корреляционной таблицы. Запишем систему (*) так, чтобы она

отражала данные корреляционной таблицы. Воспользуемся тождествами:

$$\sum x = n\bar{x} \quad \left(\text{следствие из } \bar{x} = \frac{\sum x}{n}\right);$$

$$\sum y = n\bar{y} \quad \left(\text{следствие из } \bar{y} = \frac{\sum y}{n}\right);$$

$$\sum x^2 = n\overline{x^2} \left(\text{следствие из } \overline{x^2} = \frac{\sum x^2}{n}\right)$$

$\sum xy = \sum n_{xy}xy$ (учтено, что пара чисел (x, y) наблюдалась n_{xy} раз).

Подставив правые части тождеств в систему (*) и сократив обе части второго уравнения на n, получим:

$$\left.\begin{aligned}(n\overline{x^2})\rho_{yx} + (n\bar{x})b &= \textstyle\sum n_{xy}xy;\\ (\bar{x})\rho_{yx} + b &= \bar{y}.\end{aligned}\right\} \qquad (**)$$

Решив эту систему, найдем параметры ρ_{yx} и b и, следовательно, искомое уравнение:

$$\bar{y}_x=\rho_{yx}x+b.$$

Однако более целесообразно, введя новую величину — коэффициент корреляции, написать уравнение регрессии в ином виде. Сделаем это.

Найдем b из второго уравнения (**):

$$b=\bar{y}-\rho_{yx}\bar{x}.$$

Подставив правую часть этого равенства в уравнение $\bar{y}_x=\rho_{yx}x+b$, получим:

$$\bar{y}_x-\bar{y}=\rho_{yx}(x-\bar{x}). \qquad (***)$$

Найдем из системы (*) коэффициент регрессии, учитывая, что $\overline{x^2}-(\bar{x})^2=\sigma_x^2$ (гл. XVI, § 10):

$$\rho_{yx} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n[\overline{x^2}-(\bar{x})^2]} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\sigma_x^2}.$$

Умножим обе части равенства на дробь $\frac{\sigma_x}{\sigma_y}$:

$$\rho_{yx}\cdot\frac{\sigma_x}{\sigma_y} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\sigma_x\sigma_y}.$$

Обозначим правую часть равенства через $r_В$ и назовем ее *выборочным коэффициентом корреляции*:

$$\rho_{yx}\frac{\sigma_x}{\sigma_y}=r_В,$$

или

$$\rho_{yx}=r_В\frac{\sigma_y}{\sigma_x}.$$

Подставив правую часть этого равенства в (***), окончательно получим выборочное уравнение прямой линии регрессии Y на X вида

$$\overline{y}_x-\overline{y}=r_В\frac{\sigma_y}{\sigma_x}(x-\overline{x})$$

Замечание 1. Аналогично находят выборочное уравнение прямой линии регрессии X на Y вида

$$\overline{x}_y-\overline{x}=r_В\frac{\sigma_x}{\sigma_y}(y-\overline{y}),$$

где

$$r_В\frac{\sigma_x}{\sigma_y}=\rho_{xy}.$$

Замечание 2. Уравнения прямых регрессии могут быть записаны в более симметричной форме:

$$\frac{\overline{y}_x-\overline{y}}{\sigma_y}=r_В\frac{x-\overline{x}}{\sigma_x};$$

$$\frac{\overline{x}_y-\overline{x}}{\sigma_x}=r_В\frac{y-\overline{y}}{\sigma_y}.$$

Замечание 3. Выборочный коэффициент корреляции имеет важное самостоятельное значение. Как следует из предыдущего, выборочный коэффициент корреляции определяется равенством

$$r_В=\frac{\sum n_{xy}xy-n\overline{x}\,\overline{y}}{n\sigma_x\sigma_y},$$

где x, y — варианты (наблюдавшиеся значения) признаков X и Y;
n_{xy} — частота наблюдавшейся пары вариант (x, y);
n — объем выборки (сумма всех частот);
$\overline{x}, \overline{y}$ — выборочные средние;
σ_x, σ_y — выборочные средние квадратические отклонения

§ 7. Свойства выборочного коэффициента корреляции

Приведем свойства выборочного коэффициента корреляции, из которых следует, что он служит для оценки тесноты линейной корреляционной зависимости.

Воспользуемся формулами (вывод опускаем):

$$S_y=D_y(1-r_В^2);\quad S_x=D_x(1-r_В^2),$$

где S_y — дисперсия наблюдавшихся значений y вокруг соответствующих условных средних $\overline{y}_x$;
D_y — дисперсия наблюдавшихся значений y вокруг общей средней $\overline{y}$.

Аналогичный смысл имеют дисперсии S_x, D_x.

1. *Абсолютная величина выборочного коэффициента корреляции не превосходит единицы.*

Доказательство. Любая дисперсия неотрицательна. В частности,

$$S_y=D_y(1-r_В^2)\geqslant 0.$$

Следовательно,

$$1-r_В^2\geqslant 0.$$

Отсюда

$$-1\leqslant r_В\leqslant 1,$$

или

$$|r_В|\leqslant 1.$$

2. *Если выборочный коэффициент корреляции равен нулю и выборочные линии регрессии — прямые, то X и Y не связаны линейной корреляционной зависимостью.*

Доказательство. При $r_В=0$ уравнение выборочной прямой регрессии Y на X

$$\overline{y}_x-\overline{y}=r_В\frac{\sigma_y}{\sigma_x}(x-\overline{x})$$

имеет вид:

$$\overline{y}_x-\overline{y}=0,$$

или

$$\overline{y}_x=\overline{y}.$$

При $r_В=0$ уравнение прямой регрессии X на Y имеет вид

$$\overline{x}_y = \overline{x}.$$

Таким образом, при $r_В=0$ условные средние сохраняют постоянное значение при изменении соответствующих аргументов; в этом смысле можно считать, что X и Y не связаны линейной корреляционной зависимостью. Очевидно, в рассматриваемом случае прямые регрессии параллельны соответствующим координатным осям.

Замечание. Если выборочный коэффициент корреляции равен нулю, то признаки X и Y могут быть связаны нелинейной корреляционной или даже функциональной зависимостью.

3. *Если абсолютная величина выборочного коэффициента корреляции равна единице, то наблюдаемые значения признаков связаны линейной функциональной зависимостью.*

$$\text{Если } |r_В| = 1, \text{ то } S_y = D_y(1 - r_В^2) = 0.$$

Можно показать, что отсюда следует равенство:

$$y - \overline{y} - r_В \frac{\sigma_y}{\sigma_x}(x - \overline{x}) = 0.$$

Как видим, любая наблюдаемая пара чисел (x, y) удовлетворяет этому *линейному* относительно x и y уравнению, т. е. значения признаков в выборке связаны линейной функциональной зависимостью. Заметим, что отсюда еще нельзя уверенно заключить, что и в генеральной совокупности признаки связаны линейной функциональной зависимостью (при репрезентативной выборке большого объема зависимость между признаками нормально распределенной генеральной совокупности будет близка к линейной, или даже будет линейной).

4. *С возрастанием абсолютной величины выборочного коэффициента корреляции линейная корреляционная зависимость становится более тесной и при $|r_В|=1$ переходит в функциональную зависимость.*

Доказательство. Из формул

$$S_y = D_y(1 - r_В^2), \quad S_x = D_x(1 - r_В^2)$$

видно, что с возрастанием абсолютной величины $r_В$ дисперсии S_y и S_x убывают, т. е. уменьшается рассеяние

наблюдаемых значений признаков вокруг условных средних, а это и означает, что связь между признаками становится более тесной и при $|r_В|=1$, как следует из свойства 3, переходит в функциональную.

Из приведенных свойств вытекает смысл $r_В$: *выборочный коэффициент корреляции характеризует тесноту* линейной *связи между количественными признаками в выборке: чем ближе $|r_В|$ к 1, тем связь сильнее; чем ближе $|r_В|$ к 0, тем связь слабее.*

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте линейной зависимости между признаками, полученное по данным *выборки*, в известной степени может быть распространено и на *генеральную совокупность*. Например, для оценки коэффициента корреляции $r_г$ нормально распределенной генеральной совокупности (при $n \geqslant 50$) можно воспользоваться формулой

$$r_В - 3\frac{1 - r_В^2}{\sqrt{n}} \leqslant r_г \leqslant r_В + 3\frac{1 + r_В^2}{\sqrt{n}}.$$

Замечание 1. Знак выборочного коэффициента корреляции совпадает со знаком выборочных коэффициентов регрессии что следует из формул (§ 4):

$$\rho_{yx} = r_В \frac{\sigma_y}{\sigma_x}, \quad \rho_{xy} = r_В \frac{\sigma_x}{\sigma_y}. \qquad (*)$$

Замечание 2. Выборочный коэффициент корреляции равен среднему геометрическому выборочных коэффициентов регрессии. Действительно, перемножив левые и правые части равенств(*) получим:

$$\rho_{yx}\rho_{xy} = r_В^2.$$

Отсюда

$$r_В = \pm\sqrt{\rho_{yx}\rho_{xy}}.$$

Знак при радикале, в соответствии с замечанием 1, должен совпадать со знаком коэффициентов регрессии

§ 8. Метод четырех полей вычисления выборочного коэффициента корреляции

Пусть требуется по данным корреляционной таблицы вычислить выборочный коэффициент корреляции. Можно значительно упростить вычисления, если перейти к условным вариантам:

$$u_i = \frac{x_i - C_1}{h_1} \text{ и } v_j = \frac{y_j - C_2}{h_2}.$$

В этом случае выборочный коэффициент корреляции вычисляется по формуле (переход к условным вариантам не изменяет величины r_B):

$$r_B = \frac{\sum n_{uv} uv - n\bar{u}\bar{v}}{n\sigma_u\sigma_v}.$$

Величины $\bar{u}$, $\bar{v}$, σ_u и σ_v могут быть вычислены по методу произведений (гл. XVII, § 4). Остается указать способ вычисления $\sum n_{uv}uv$. Этой цели и служит *метод четырех полей*. Название метода связано с тем, что строка и столбец, пересекающиеся в клетке, содержащей наибольшую частоту, делят корреляционную таблицу на 4 части, которые называют *полями*. Поля нумеруются так, как указано в таблице 14.

Таблица 14

v \ u		0	
	I		II
0		наиб. частота	
	III		IV

Покажем, как ведется расчет, для чего ограничимся пока полем I. Пусть часть таблицы 14, содержащая первое поле, представлена в виде таблицы 15.

Таблица 15

v \ u	—3	—2	—1
—2	5		—
—1	—	20	23

Найдем произведения пар вариант u и v и поместим их в верхние правые углы клеток, содержащих соответственные частоты. Например, пара вариант $u=-3$ и $v=-2$ наблюдалась 5 раз; произведение $uv=(-3)\cdot(-2)=6$ помещаем в верхний правый угол клетки, содержащей частоту 5. Заполнив подобным образом остальные клетки первого поля, получим таблицу 16.

Таблица 16

v \ u	—3	—2	—1
—2	5 $\boxed{6}$	7 $\boxed{4}$	—
—1	—	20 $\boxed{2}$	23 $\boxed{1}$

Аналогично заполняются клетки и остальных полей. Таким образом, в каждой клетке (содержащей частоту n_{uv}), оказывается записанным и произведение uv, остается перемножить два числа n_{uv} и uv каждой клетки и результаты сложить; в итоге получим искомое число $\sum n_{uv}uv$.

Для удобства контроля вычислений найденные произведения чисел n_{uv} и uv каждой клетки суммируются отдельно по каждому полю, причем подсчет ведется и по строкам и по столбцам каждого поля: сумму чисел $n_{uv}\cdot uv$ строки поля выписывают в тот из дополнительных столбцов, помещенных справа, который имеет номер того поля, числа которого складывались. Сумму чисел $n_{uv}\cdot uv$ столбца поля выписывают в ту из дополнительных строк, помещенных внизу, которая имеет номер того поля, числа которого складывались. Суммы чисел отдельно по каждому полю записывают в правом нижнем углу таблицы в четырех итоговых клетках. Наконец, складывая все числа итоговых клеток, находят искомое число.

Схематически расчетная таблица представлена в виде таблицы 17.

Поясним, как заполнена таблица 17 (для большей наглядности расчет ведется лишь для первого поля).

$v \backslash u$	−3	−2	−1	0		I	II
−2	$5\ \vert\underline{6}$	$7\ \vert\underline{4}$	—			58	
−1	—	$20\ \vert\underline{2}$	$23\ \vert\underline{1}$		II	63	
0				наиб. частота		III	IV
		III			IV		
I	30	68	23	II		121	II
III				IV		III	IV

Найдем суммы произведений n_{uv} и uv по строкам первого поля ($5\cdot 6+7\cdot 4=58$; $20\cdot 2+23\cdot 1=63$) и поместим их в дополнительный столбец I.

Найдем суммы произведений n_{uv} и uv по столбцам первого поля ($5\cdot 6=30$; $7\cdot 4+20\cdot 2=68$; $23\cdot 1=23$) и поместим их в дополнительную строку I.

Найдем сумму чисел дополнительного столбца I ($58+63=121$) и запишем ее в первую итоговую клетку (в правом нижнем углу таблицы).

Для контроля сложим все числа дополнительной строки ($30+68+23=121$).

Аналогично ведется расчет и по остальным полям.

Пример. Вычислить выборочный коэффициент корреляции по данным корреляционной таблицы 18.

Р е ш е н и е. Перейдем к условным вариантам:

$u=\dfrac{x-c_1}{h_1}=\dfrac{x-40}{10}$ (в качестве ложного нуля c_1 взята варианта $x=40$, имеющая наибольшую частоту; шаг h_1 равен разности между двумя соседними вариантами:

$y \backslash x$	10	20	30	40	50	60	n_y
15	5	7	—	—	—	—	12
25	—	20	23	—	—	—	43
35	—	—	30	47	2	—	79
45	—	—	10	11	20	6	47
55	—	—	—	9	7	3	19
n_x	5	27	63	67	29	9	$n=200$

$20-10=10$), и $v=\dfrac{y-c_2}{h_2}=\dfrac{y-35}{10}$ (в качестве ложного нуля c_2 взята варианта $y=35$, имеющая наибольшую частоту; шаг h_2 равен разности между двумя соседними вариантами $25-15=10$).

Составим корреляционную таблицу в условных вариантах. Практически это делают так: в первом столбце вместо варианты (35), имеющей наибольшую частоту, пишут 0; над нулем пишут последовательно −1, −2; под нулем пишут 1, 2. В первой строке вместо варианты (40), имеющей наибольшую частоту, пишут 0; слева от нуля последовательно записывают −1, −2, −3; справа от нуля пишут 1, 2. Все остальные данные переписывают из первоначальной корреляционной таблицы. В итоге получим корреляционную таблицу 19 в условных вариантах.

Величины $\bar{u}$, $\bar{v}$, σ_u и σ_v можно найти методом произведений; однако поскольку числа u_i, v_j малы, вычислим $\bar{u}$ и $\bar{v}$, исходя из определения средней, а σ_u и σ_v, пользуясь формулами (гл. XVI, § 10):

$$\sigma_u=\sqrt{\overline{u^2}-(\bar{u})^2}, \qquad \sigma_v=\sqrt{\overline{v^2}-(\bar{v})^2}.$$

Таблица 19

v \ u	−3	−2	−1	0	1	2	n_v
−2	5	7	—	—	—	—	12
−1	—	20	23	—	—	—	43
0	—	—	30	47	2	—	79
1	—	—	10	11	20	6	47
2	—	—	—	9	7	3	19
n_u	5	27	63	67	29	9	$n=200$

Найдем $\bar{u}$ и $\bar{v}$:

$$\bar{u}=\frac{\sum n_u u}{n}=\frac{5\cdot(-3)+27\cdot(-1)+63\cdot(-2)+29\cdot 1+9\cdot 2}{200}=$$

$$=-0{,}425;$$

$$\bar{v}=\frac{\sum n_v v}{n}=\frac{12\cdot(-2)+43\cdot(-1)+47\cdot 1+19\cdot 2}{200}=0{,}09.$$

Вычислим вспомогательную величину $\overline{u^2}$, а затем σ_u:

$$\overline{u^2}=\frac{5\cdot 9+27\cdot 4+63\cdot 1+29\cdot 1+9\cdot 4}{200}=1{,}405;$$

$$\sigma_u=\sqrt{\overline{u^2}-(\bar{u})^2}=\sqrt{1{,}405-0{,}425^2}=1{,}106.$$

Аналогично получим $\sigma_v=1{,}209$.

Найдем $\sum n_{uv}uv$ методом четырех полей, для чего составим расчетную таблицу 20.

Таблица 20

	−3	−2	−1	0	1	2	I	II
−2	5 | 6	7 | 4	—		—	—	58	—
−1	—	20 | 2	23 | 1		—	—	63	—
0							III	IV
1	—	—	10 | −1		20 | 1	6 | 2	−10	32
2	—	—	—		7 | 2	3 | 4	—	26
I	30	68	23	11	—	—	121	—
III	—	—	−10	IV	34	24	−10	58

Сложив числа итоговых клеток (4 клетки в нижнем правом углу таблицы 20), получим

$$\sum n_{uv}uv=121-10+58=169.$$

Найдем искомый коэффициент корреляции:

$$r_в=\frac{\sum n_{uv}uv-n\bar{u}\bar{v}}{n\sigma_u\sigma_v}=\frac{169-200\cdot(-0{,}425)\cdot 0{,}09}{200\cdot 1{,}106\cdot 1{,}209}=0{,}603.$$

Итак,

$$r_в=0{,}603.$$

§ 9. Пример на отыскание выборочного уравнения прямой линии регрессии

Теперь, когда известно как вычисляют $r_в$, уместно привести пример на отыскание уравнения прямой линии регрессии.

Поскольку при нахождении $r_в$ уже вычислены $\bar{u}$, $\bar{v}$, σ_u, σ_v, то целесообразно пользоваться формулами:

$$\sigma_x = h_1\sigma_u, \quad \sigma_y = h_2\sigma_v, \quad \bar{x} = \bar{u}h_1 + c_1, \quad \bar{y} = \bar{v}h_2 + c_2.$$

Здесь сохранены обозначения предыдущего параграфа. Рекомендуем читателю самостоятельно вывести эти формулы.

Пример. Найти выборочное уравнение прямой линии регрессии Y на X по данным корреляционной таблицы 18 примера предыдущего параграфа.

Р е ш е н и е. Напишем искомое уравнение в общем виде:

$$\bar{y}_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad (*)$$

Коэффициент корреляции уже вычислен в предыдущем параграфе. Остается найти $\bar{x}$, $\bar{y}$, σ_x и σ_y:

$$\bar{x} = \bar{u}h_1 + c_1 = -0{,}425 \cdot 10 + 40 = 35{,}75;$$
$$\bar{y} = \bar{v}h_2 + c_2 = 0{,}09 \cdot 10 + 35 = 35{,}9;$$
$$\sigma_x = \sigma_u h_1 = 1{,}106 \cdot 10 = 11{,}06;$$
$$\sigma_y = \sigma_v h_2 = 1{,}209 \cdot 10 = 12{,}09.$$

Подставив найденные величины в (*), получим искомое уравнение

$$\bar{y}_x - 35{,}9 = 0{,}603 \frac{12{,}09}{11{,}06}(x - 35{,}75),$$

или окончательно

$$\bar{y}_x = 0{,}659x + 12{,}34.$$

Сравним условные средние, вычисленные: а) по этому уравнению; б) по данным корреляционной таблицы 18. Например, при $x = 30$:

а) $\bar{y}_{30} = 0{,}659 \cdot 30 + 12{,}34 = 32{,}11$;

б) $\bar{y}_{30} = \dfrac{23 \cdot 25 + 30 \cdot 35 + 10 \cdot 45}{63} = 32{,}94.$

Как видим, согласование расчетного и наблюдаемого условных средних — удовлетворительное.

§ 10. Предварительные соображения к введению меры любой корреляционной связи

Выше рассматривалась оценка тесноты *линейной* корреляционной связи. Как оценить тесноту *любой* корреляционной связи?

Пусть данные наблюдений над количественными признаками X и Y сведены в корреляционную таблицу. Можно считать, что тем самым наблюдаемые значения Y разбиты на группы; каждая группа содержит те значения Y, которые соответствуют определенному значению X.

Например, дана корреляционная таблица 21.

Т а б л и ц а 21

Y \ X	8	9
3	4	13
5	6	7
n_x	10	20
$\bar{y}_x$	4,2	3,7

К первой группе относятся те 10 значений Y (4 раза наблюдалось $y_1 = 3$ и 6 раз $y_2 = 5$), которые соответствуют $x_1 = 8$.

Ко второй группе относятся те 20 значений Y (13 раз наблюдалось $y_1 = 3$ и 7 раз $y_2 = 5$), которые соответствуют $x_2 = 9$.

Условные средние теперь можно назвать групповыми средними: групповая средняя первой группы $\bar{y}_8 = \dfrac{4 \cdot 3 + 6 \cdot 5}{10} = 4{,}2$; групповая средняя второй группы

$$\bar{y}_9 = \frac{13 \cdot 3 + 7 \cdot 5}{20} = 3{,}7.$$

Поскольку все значения признака Y разбиты на группы, можно представить общую дисперсию признака в виде суммы внутригрупповой и межгрупповой дисперсий (гл. XVI, § 12):

$$D_{общ} = D_{внгр} + D_{межгр}. \qquad (*)$$

Покажем справедливость следующих утверждений:

1) если Y связан с X функциональной зависимостью, то

$$\frac{D_{межгр}}{D_{общ}} = 1;$$

2) если Y связан с X корреляционной зависимостью, то

$$\frac{D_{межгр}}{D_{общ}} < 1.$$

Доказательство. 1) Если Y связан с X функциональной зависимостью, то определенному значению X соответствует одно значение Y. В этом случае в каждой группе содержатся равные между собой значения Y*, поэтому групповая дисперсия каждой группы равна нулю. Следовательно, средняя арифметическая групповых дисперсий (взвешенная по объемам групп), т. е. внутригрупповая дисперсия $D_{внгр}=0$ и равенство (*) имеет вид

$$D_{общ} = D_{межгр}.$$

Отсюда

$$\frac{D_{межгр}}{D_{общ}} = 1.$$

2) Если Y связан с X корреляционной зависимостью, то определенному значению X соответствуют, вообще говоря, различные значения Y (образующие группу). В этом случае групповая дисперсия каждой группы отлична от нуля. Следовательно, средняя арифметическая групповых дисперсий (взвешенная по объемам групп) $D_{внгр} \neq 0$. Тогда (одно положительное слагаемое — $D_{межгр}$ меньше суммы двух положительных слагаемых $D_{внгр} + D_{межгр} = D_{общ}$):

$$D_{межгр} < D_{общ}.$$

Отсюда

$$\frac{D_{межгр}}{D_{общ}} < 1.$$

Уже из приведенных рассуждений видно, что чем связь между признаками ближе к функциональной, тем меньше $D_{внгр}$ и, следовательно, тем больше будет приближаться $D_{межгр}$ к $D_{общ}$, а значит отношение $\frac{D_{межгр}}{D_{общ}}$ к единице. Отсюда ясно, что целесообразно рассматривать в качестве меры тесноты корреляционной зависимости отношение межгрупповой дисперсии к общей или, что то же, отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению.

§ 11. Выборочное корреляционное отношение

Для оценки тесноты *линейной* корреляционной связи между признаками в выборке служит выборочный коэффициент корреляции. Для оценки тесноты *нелинейной* корреляционной связи вводят новые сводные характеристики:

η_{yx} — выборочное корреляционное отношение Y к X;
η_{xy} — выборочное корреляционное отношение X к Y.

Выборочным корреляционным отношением Y к X называют отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y

$$\eta_{yx} = \frac{\sigma_{межгр}}{\sigma_{общ}},$$

или в других обозначениях:

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y}.$$

* Например, если значению $x_1 = 3$ соответствует $y_1 = 7$, причем $x_1 = 3$ наблюдалось 5 раз, то в группе содержится 5 значений $y_1 = 7$.

Здесь

$$\sigma_{\bar{y}_x} = \sqrt{D_{\text{межгр}}} = \sqrt{\frac{\sum n_x (\bar{y}_x - \bar{y})^2}{n}};$$

$$\sigma_y = \sqrt{D_{\text{общ}}} = \sqrt{\frac{\sum n_y (y - \bar{y})^2}{n}},$$

где n — объем выборки (сумма всех частот);
n_x — частота значения x признака X;
n_y — частота значения y признака Y;
$\bar{y}$ — общая средняя признака Y;
$\bar{y}_x$ — условная средняя признака Y.

Аналогично определяется выборочное корреляционное отношение X к Y:

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}.$$

Пример. Найти η_{yx} по данным корреляционной таблицы 22.

Таблица 22

Y \ X	10	20	30	n_y
15	4	28	6	38
25	6	—	6	12
n_x	10	28	12	$n=50$
$\bar{y}_x$	21	15	20	

Решение. Найдем общую среднюю

$$\bar{y} = \frac{\sum n_y y}{n} = \frac{38 \cdot 15 + 12 \cdot 25}{50} = 17{,}4.$$

Найдем общее среднее квадратическое отклонение

$$\sigma_y = \sqrt{\frac{\sum n_y (y - \bar{y})^2}{n}} =$$

$$= \sqrt{\frac{38\,(15 - 17{,}4)^2 + 12\,(25 - 17{,}4)^2}{50}} = 4{,}27.$$

Найдем межгрупповое среднее квадратическое отклонение

$$\sigma_{\bar{y}_x} = \sqrt{\frac{\sum n_x (\bar{y}_x - \bar{y})^2}{n}} =$$

$$= \sqrt{\frac{10\,(21 - 17{,}4)^2 + 28\,(15 - 17{,}4)^2 + 12\,(20 - 16{,}4)^2}{50}} = 2{,}73.$$

Искомое корреляционное отношение

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y} = \frac{2{,}73}{4{,}27} = 0{,}64.$$

§ 12. Свойства выборочного корреляционного отношения

Поскольку η_{xy} обладает теми же свойствами, что и η_{yx}, перечислим свойства только выборочного корреляционного отношения η_{yx}, которое далее для упрощения записи будем обозначать через η и для простоты речи называть «корреляционным отношением».

1. *Корреляционное отношение удовлетворяет двойному неравенству*:

$$0 \leqslant \eta \leqslant 1.$$

Доказательство. Неравенство $\eta \geqslant 0$ следует из того, что η есть отношение неотрицательных чисел — средних квадратических отклонений (межгруппового к общему).

Для доказательства неравенства $\eta \leqslant 1$ воспользуемся формулой

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}}.$$

Разделив обе части равенства на $D_{\text{общ}}$, получим:

$$1 = \frac{D_{\text{внгр}}}{D_{\text{общ}}} + \frac{D_{\text{межгр}}}{D_{\text{общ}}},$$

или

$$1 = \frac{D_{\text{внгр}}}{D_{\text{общ}}} + \eta^2.$$

Так как оба слагаемых неотрицательны и сумма их равна единице, то каждое из них не превышает единицы; в частности

$$\eta^2 \leqslant 1.$$

Приняв во внимание, что $\eta \geqslant 0$, заключаем:

$$0 \leqslant \eta \leqslant 1.$$

2. *Если $\eta = 0$, то и признак Y с признаком X корреляционной зависимостью не связан.*

Доказательство. По условию

$$\eta = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}} = 0.$$

Отсюда

$$\sigma_{\text{межгр}} = 0$$

и, следовательно,

$$D_{\text{межгр}} = 0.$$

Межгрупповая дисперсия есть дисперсия условных (групповых) средних $\overline{y}_x$ относительно общей средней $\overline{y}$. Равенство нулю межгрупповой дисперсии означает, что при всех значениях X условные средние сохраняют постоянное значение (равное общей средней). Иными словами, при $\eta = 0$ условная средняя не является функцией от X, а значит признак Y не связан корреляционной зависимостью с признаком X.

Замечание 1. Можно доказать и обратное предложение: если признак Y не связан с признаком X корреляционной зависимостью, то $\eta = 0$.

3. *Если $\eta = 1$, то признак Y связан с признаком X функциональной зависимостью.*

Доказательство. По условию

$$\eta = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}} = 1.$$

Отсюда

$$\sigma_{\text{общ}} = \sigma_{\text{межгр}}.$$

Возведя обе части равенства в квадрат, получим

$$D_{\text{общ}} = D_{\text{межгр}}. \qquad (*)$$

Так как

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}},$$

то, в силу (*)

$$D_{\text{внгр}} = 0. \qquad (**)$$

Поскольку внутригрупповая дисперсия есть средняя арифметическая групповых дисперсий (взвешенная по объемам групп), то из (**) следует, что дисперсия каждой группы (значений Y, соответствующих определенному значению X) равна нулю. А это означает, что в группе содержатся равные значения Y, т. е. каждому значению X соответствует одно значение Y. Следовательно, при $\eta = 1$ признак Y связан с признаком X функциональной зависимостью.

Замечание 2. Можно доказать и обратное предложение: если признак Y связан с признаком X функциональной зависимостью то $\eta = 1$.

Приведем еще два свойства, опустив доказательства.

4. *Выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции:*

$$\eta \geqslant |r_{\text{в}}|.$$

5. *Если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.*

Другими словами, если $\eta = |r_{\text{в}}|$, то точки

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

лежат на прямой линии регрессии, найденной способом наименьших квадратов.

§ 13. Корреляционное отношение как мера корреляционной связи. Достоинства и недостатки этой меры

В предыдущем параграфе установлено: при $\eta=0$ признаки не связаны корреляционной зависимостью; при $\eta=1$ имеет место функциональная зависимость.

Убедимся, что с возрастанием η корреляционная связь становится более тесной. С этой целью преобразуем соотношение

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}}$$

так:

$$D_{\text{внгр}} = D_{\text{общ}}\left(1 - \frac{D_{\text{межгр}}}{D_{\text{общ}}}\right),$$

или

$$D_{\text{внгр}} = D_{\text{общ}}(1 - \eta^2).$$

Если $\eta \to 1$, то $D_{\text{внгр}} \to 0$, следовательно, стремится к нулю и каждая из групповых дисперсий. Другими словами, при возрастании η значения Y, соответствующие определенному значению X, все меньше различаются между собой и связь Y с X становится более тесной, переходя в функциональную, при $\eta=1$.

Поскольку в рассуждениях не делалось никаких допущений о форме корреляционной связи, η *служит мерой тесноты связи любой, в том числе нелинейной, формы*. В этом состоит преимущество корреляционного отношения перед коэффициентом корреляции, который оценивает тесноту лишь линейной зависимости. Вместе с тем, корреляционное отношение обладает *недостатком*: оно не позволяет судить, насколько близко расположены точки, найденные по данным наблюдений, к кривой определенного вида, например к параболе, гиперболе и т. д. Это объясняется тем, что при определении корреляционного отношения форма связи во внимание не принималась

§ 14. Простейшие случаи криволинейной корреляции

Если график регрессии $\overline{y}_x=f(x)$ или $\overline{x}_y=\varphi(y)$ изображается кривой линией, то корреляцию называют *криволинейной*

Например, функции регрессии Y на X могут иметь вид:

$\overline{y}_x=ax^2+bx+c$ (параболическая корреляция второго порядка);

$\overline{y}_x=ax^3+bx^2+cx+d$ (параболическая корреляция третьего порядка);

$$\overline{y}_x = \frac{a}{x} + b \text{ (гиперболическая корреляция).}$$

Теория криволинейной корреляции решает те же задачи, что и теория линейной корреляции (установление формы и тесноты корреляционной связи).

Неизвестные параметры уравнения регрессии ищут методом наименьших квадратов. Для оценки тесноты криволинейной корреляции служат выборочные корреляционные отношения (§ 11).

Чтобы выяснить суть дела, ограничимся параболической корреляцией второго порядка, предположив, что данные n наблюдений (выборки) позволяют считать, что имеет место именно такая корреляция. В этом случае выборочное уравнение регрессии Y на X имеет вид:

$$\overline{y}_x = Ax^2 + Bx + C, \qquad (*)$$

где A, B, C — неизвестные параметры.

Пользуясь методом наименьших квадратов, получают систему линейных уравнений относительно неизвестных параметров (вывод опущен, поскольку он не содержит ничего нового сравнительно с § 4):

$$\left.\begin{aligned} (\sum n_x x^4)A + (\sum n_x x^3)B + (\sum n_x x^2)C &= \sum n_x \overline{y}_x x^2; \\ (\sum n_x x^3)A + (\sum n_x x^2)B + (\sum n_x x)C &= \sum n_x \overline{y}_x x; \\ (\sum n_x x^2)A + (\sum n_x x)B + \quad nC &= \sum n_x \overline{y}_x. \end{aligned}\right\} \quad (**)$$

Найденные из этой системы параметры A, B, C подставляют в (*) и итоге получают искомое уравнение регрессии.

Пример. Найти выборочное уравнение регрессии Y на X вида $\overline{y}_x=Ax^2+Bx+C$ по данным корреляционной таблицы 23.

Составим расчетную таблицу 24.

Подставив числа (суммы) нижней строки таблицы 24 в (**), получим систему:

$$\left.\begin{aligned} 74{,}98\ A+67{,}48\ B+60{,}89\ C&=413{,}93, \\ 67{,}48\ A+60{,}89\ B+55{,}10\ C&=373{,}30, \\ 60{,}89\ A+55{,}10\ B+50\ C&=337{,}59. \end{aligned}\right\}$$

Таблица 23

Y \ X	1	1,1	1,2	n_y
6	8	2	—	10
7	—	30	—	30
7,5	—	1	9	10
n_x	8	33	9	$n=50$
$\bar{y}_x$	6	6,73	7,5	

Таблица 24

x	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
1	8	6	8	8	8	8	48	48	48
1,1	33	6,73	36,3	39,93	43,93	48,32	222,09	244,30	268,73
1,2	9	7,5	10,8	12,96	15,55	18,66	67,50	81	97,20
Σ	50	—	55,1	60,89	67,48	74,98	337,59	373,30	413,93

Решив эту систему, найдем:

$$A=1{,}94,\quad B=2{,}98,\quad C=1{,}10.$$

Напишем искомое уравнение регрессии:

$$\bar{y}_x=1{,}94\; x^2+2{,}98\; x+1{,}10.$$

Легко убедиться, что условные средние, вычисленные по этому уравнению, незначительно отличаются от условных средних корреляционной таблицы. Например, при $x_1=1$ найдем: по таблице $\bar{y}_1=6$; по уравнению $\bar{y}_1=$

$=1{,}94+2{,}98+1{,}10=6{,}02$. Таким образом, найденное уравнение хорошо согласуется с данными наблюдений (выборки).

§ 15. Понятие о множественной корреляции

До настоящего параграфа рассматривалась корреляционная связь между двумя признаками. Если же исследуется связь между несколькими признаками, то корреляцию называют *множественной*.

В простейшем случае число признаков равно трем, и связь между ними линейная:

$$z=ax+by+c.$$

В этом случае возникают задачи:

1) найти по данным наблюдений выборочное уравнение связи вида

$$z=Ax+By+C, \qquad (*)$$

т. е. требуется найти коэффициенты регрессии A и B и параметр C;

2) оценить тесноту связи между Z и обоими признаками X, Y;

3) оценить тесноту связи между Z и X (при постоянном Y), между Z и Y (при постоянном X).

Первая задача решается методом наименьших квадратов, причем вместо уравнения (*) удобнее искать уравнение связи вида

$$z-\bar{z}=A(x-\bar{x})+B(y-\bar{y}),$$

где

$$A=\frac{r_{xz}-r_{yz}r_{xy}}{1-r_{xy}^2}\cdot\frac{\sigma_z}{\sigma_x};\quad B=\frac{r_{yz}-r_{xz}r_{xy}}{1-r_{xy}^2}\cdot\frac{\sigma_z}{\sigma_y}.$$

Здесь r_{xz}, r_{yz}, r_{xy} — коэффициенты корреляции соответственно между признаками X и Z, Y и Z, X и Y; σ_x, σ_y, σ_z — средние квадратические отклонения.

Теснота связи признака Z с признаками X, Y оценивается выборочным совокупным коэффициентом корреляции:

$$R=\sqrt{\frac{r_{xz}^2-2r_{xy}r_{xz}r_{yz}+r_{yz}^2}{1-r_{xy}^2}},$$

причем $0\leqslant R\leqslant 1$.

Теснота связи между Z и X (при постоянном Y), между Z и Y (при постоянном X) оценивается соответственно *частными выборочными коэффициентами корреляции:*

$$r_{xz(y)}=\frac{r_{xz}-r_{xy}r_{yz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{yz}^2\right)}};$$

$$r_{yz(x)}=\frac{r_{yz}-r_{xy}r_{xz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{xz}^2\right)}}.$$

Эти коэффициенты имеют те же свойства и тот же смысл, что и обыкновенный выборочный коэффициент корреляции, т. е. служат для оценки линейной связи между признаками.

Задачи

В задачах 1—2 даны корреляционные таблицы. Найти: а) r_B; б) выборочные уравнения прямых регрессии, в) η_{yx} и η_{xy}.

1.

Y \ X	5	10	15	20	n_y	$\bar{x}_y$
10	2	—	—	—	2	5
20	5	4	1	—	10	8
30	3	8	6	3	20	12,25
40	—	3	6	6	15	16
50	—	—	2	1	3	16,67
n_x	10	15	15	10	$n=50$	
$\bar{y}_x$	21	29,33	36	38		

Отв. а) 0,636; б) $\bar{y}_x=1,17x+16,78$; $\bar{x}_y=0,345y+1,67$; в) $\eta_{yx}=0,656$, $\eta_{xy}=0,651$.

Y \ X	65	95	125	155	185	215	n_y	$\bar{x}_y$
30	5	—	—	—	—	—	5	65
40	4	12	—	—	—	—	16	87,5
50	—	8	5	4	—	—	17	101,18
60	—	1	5	7	2	—	15	145
70	—	—	—	—	1	1	2	200
n_x	9	21	10	11	3	1	$n=55$	
$\bar{y}_x$	34,14	44,76	55	56,36	63,33	70		

Отв. а) 0,825; б) $\bar{y}_x=0,23x+21,78$, $\bar{x}_y=2,92y-27,25$; в) $\eta_{yx}=0,859$, $\eta_{xy}=0,875$.

В задачах 3—4 найти выборочные уравнения регрессии $\bar{y}_x=Ax^2+Bx+C$ по данным корреляционной таблицы.

3.

Y \ X	2	3	5	n_y
25	20	—	—	20
45	—	30	1	31
110	—	1	48	49
	20	31	49	$n=100$

Отв. $\bar{y}_x=2,94x^2+7,27x-1,25$.

4.

X / Y	1	2	n_y
2	30	1	31
6	1	18	19
n_x	31	19	$n=50$

Отв. $\overline{y}_x=0{,}39x^2+2{,}49x-0{,}75$.

Глава девятнадцатая

СТАТИСТИЧЕСКАЯ ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

§ 1. Статистическая гипотеза. Нулевая и конкурирующая, простая и сложная гипотезы

Часто необходимо знать закон распределения генеральной совокупности. Если закон распределения неизвестен, но имеются основания предположить, что он имеет определенный вид (назовем его A), выдвигают гипотезу: генеральная совокупность распределена по закону A. Таким образом, в этой гипотезе речь идет о виде предполагаемого распределения.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр Θ равен определенному значению Θ_0, выдвигают гипотезу: $\Theta=\Theta_0$. Таким образом, в этой гипотезе речь идет о предполагаемой величине параметра одного известного распределения.

Возможны и другие гипотезы: о равенстве параметров двух или нескольких распределений, о независимости выборок и многие другие.

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений.

Например, статистическими будут гипотезы:

1) генеральная совокупность распределена по закону Пуассона;

2) дисперсии двух нормальных совокупностей равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения, во второй — о параметрах двух известных распределений.

Гипотеза «в 1980 г. не будет войны» не является статистической, поскольку в ней не идет речь ни о виде, ни о параметрах распределения.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (*основной*) называют выдвинутую гипотезу H_0.

Конкурирующей (*альтернативной*) называют гипотезу H_1, которая противоречит нулевой.

Например, если нулевая гипотеза состоит в предположении, что математическое ожидание a нормального распределения равно 10, то конкурирующая гипотеза, в частности, может состоять в предположении, что $a\neq10$. Коротко это записывают так:

$$H_0: a=10;\quad H_1: a\neq10.$$

Различают гипотезы, которые содержат только одно и более одного предположений.

Простой называют гипотезу, содержащую только одно предположение. Например, если λ — параметр показательного распределения, то гипотеза $H_0: \lambda=5$ — простая. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ известно) — простая.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез. Например, сложная гипотеза $H: \lambda>5$ состоит из бесчисленного множества простых вида $H_i: \lambda=b_i$, где b_i — любое число, большее 5. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ неизвестно) — сложная.

§ 2. Ошибки первого и второго рода

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют *статистической*. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т. е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Подчеркнем, что последствия этих ошибок могут оказаться весьма различными. Например, если отвергнуто правильное решение «продолжать строительство жилого дома», то эта ошибка первого рода повлечет материальный ущерб; если же принято неправильное решение «продолжать строительство», несмотря на опасность обвала стройки, то эта ошибка второго рода может повлечь гибель людей. Разумеется, можно привести примеры, когда ошибка первого рода влечет более тяжелые последствия, чем ошибка второго рода.

Замечание 1. Правильное решение может быть принято также в двух случаях:

1) гипотеза принимается, причем и в действительности она правильная;

2) гипотеза отвергается, причем и в действительности она неверна.

Замечание 2. Вероятность совершить ошибку первого рода принято обозначать через α; ее называют *уровнем значимости*. Наиболее часто уровень значимости принимают равным 0,05 или 0,01. Если, например, принят уровень значимости равный 0,05, то это означает, что в пяти случаях из ста мы рискуем допустить ошибку первого рода (отвергнуть правильную гипотезу).

§ 3. Статистический критерий проверки нулевой гипотезы. Наблюдаемое значение критерия

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближенное распределение которой известно. Эту величину обозначают через U или Z, если она распределена нормально, F или v^2 — по закону Фишера — Снедекора, T — по закону Стьюдента, χ^2 — по закону «хи квадрат» и т. д. Поскольку в этом параграфе вид распределения во внимание приниматься не будет, обозначим эту величину, в целях общности, через K.

Статистическим критерием (или просто *критерием*) называют случайную величину K, которая служит для проверки нулевой гипотезы.

Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве критерия K принимают отношение исправленных выборочных дисперсий:

$$F=\frac{s_1^2}{s_2^2}.$$

Эта величина случайная, потому что в различных опытах дисперсии будут принимать различные, наперед неизвестные значения и распределена по закону Фишера — Снедекора.

Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин, и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением $K_{набл}$ назначают значение критерия, вычисленное по выборкам.

Например, если по двум выборкам, извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии $s_1^2=20$ и $s_2^2=5$, то наблюдаемое значение критерия F

$$F_{набл}=\frac{s_1^2}{s_2^2}=\frac{20}{5}=4.$$

§ 4. Критическая область. Область принятия гипотезы. Критические точки

После выбора определенного критерия, множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, а другое — при которых она принимается.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значений критерия, при которых гипотезу принимают.

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия принадлежит критической области — гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы — гипотезу принимают.

Поскольку критерий K — одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы также являются интервалами и, следовательно, существуют точки, которые их разделяют.

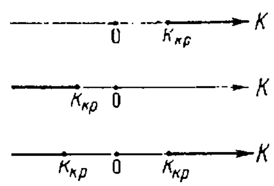

Рис. 23.

Критическими точками (границами) $k_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую неравенством $K>k_{кр}$, где $k_{кр}$ — положительное число (рис. 23, *а*).

Левосторонней называют критическую область, определяемую неравенством $K<k_{кр}$, где $k_{кр}$ — отрицательное число (рис. 23, *б*).

Односторонней называют правостороннюю или левостороннюю критическую область.

Двусторонней называют критическую область, определяемую неравенствами $K<k_1$, $K>k_2$, где $k_2>k_1$.

В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами (в предположении, что $k_{кр}>0$):

$$K<-k_{кр}, \quad K>k_{кр},$$

или равносильным неравенством $|K|>k_{кр}$ (рис. 23, *в*).

§ 5. Отыскание правосторонней критической области

Как найти критическую область? Обоснованный ответ на этот вопрос требует привлечения довольно сложной теории. Ограничимся ее элементами. Для определенности начнем с нахождения правосторонней критической области, которая определяется неравенством

$$K>k_{кр},$$

где $k_{кр}>0$.

Мы видим, что для отыскания правосторонней критической области достаточно найти критическую точку. Следовательно, возникает новый вопрос: как ее найти?

С этой целью задаются достаточно малой вероятностью — уровнем значимости α. Затем ищут критическую точку $k_{кр}$, исходя из требования, чтобы, при условии справедливости нулевой гипотезы, вероятность того, что критерий K примет значение, большее $k_{кр}$, была равна принятому уровню значимости:

$$P(K>k_{кр})=\alpha.$$

Для каждого критерия имеются соответствующие таблицы, по которым и находят критическую точку, удовлетворяющую этому требованию.

Замечание 1. Когда критическая точка уже найдена, вычисляют по данным выборок наблюденное значение критерия и, если окажется, что $K_{набл}>k_{кр}$, то нулевую гипотезу отвергают; если же $K_{набл}<k_{кр}$, то нет оснований, чтобы отвергнуть нулевую гипотезу.

Пояснение. Почему правосторонняя критическая область была определена, исходя из требования, чтобы при справедливости нулевой гипотезы выполнялось соотношение

$$P(K>k_{кр})=\alpha? \qquad (*)$$

Поскольку вероятность события $K>k_{кр}$ мала (α — малая вероятность), такое событие, при справедливости нулевой гипотезы, в силу принципа практической невозможности маловероятных событий, в единичном испытании не должно наступить (гл. II, § 4). Если все же оно произошло, т. е. наблюдаемое значение критерия оказалось больше $k_{кр}$, то это можно объяснить тем, что нулевая гипотеза ложна и, следовательно, должна быть отвергнута. Таким образом, требование (*) определяет такие значения критерия, при которых нулевая гипотеза отвергается, а они и составляют правостороннюю критическую область.

Замечание 2. Наблюдаемое значение критерия может оказаться большим $k_{кр}$ не потому, что нулевая гипотеза ложна, а по другим причинам (малый объем выборки, недостатки методики эксперимента и др.). В этом случае, отвергнув правильную нулевую гипотезу, совершают ошибку первого рода. Вероятность этой ошибки равна уровню значимости α. Итак, пользуясь требованием (*), мы с вероятностью α рискуем совершить ошибку первого рода.

Заметим кстати, что в книгах по контролю качества продукции, вероятность признать негодной партию годных изделий называют «риском производителя», а вероятность принять негодную партию — «риском потребителя».

Замечание 3. Пусть нулевая гипотеза принята; ошибочно думать, что тем самым она доказана. Действительно, известно, что один пример, подтверждающий справедливость некоторого общего утверждения еще не доказывает его. Поэтому более правильно говорить «данные наблюдений согласуются с нулевой гипотезой и, следовательно, не дают оснований ее отвергнуть».

На практике для большей уверенности принятия гипотезы, ее проверяют другими способами, или повторяют эксперимент, увеличив объем выборки.

Отвергают гипотезу более категорично, чем принимают. Действительно, известно, что достаточно привести один пример, противоречащий некоторому общему утверждению, чтобы это утверждение отвергнуть. Если оказалось, что наблюдаемое значение критерия принадлежит критической области, то этот факт и служит примером, противоречащим нулевой гипотезе, что позволяет ее отклонить.

§ 6. Отыскание левосторонней и двусторонней критических областей

Отыскание левосторонней и двусторонней критических областей сводится (так же, как и для правосторонней) к нахождению соответствующих критических точек.

Левосторонняя критическая область определяется (§ 4) неравенством $K<k_{кр}(k_{кр}<0)$.

Критическую точку находят, исходя из требования, чтобы при справедливости нулевой гипотезы, вероятность того, что критерий примет значение, меньшее $k_{кр}$, была равна принятому уровню значимости:

$$P(K<k_{кр})=\alpha.$$

Двусторонняя критическая область определяется (§ 4) неравенствами $K<k_1$, $K>k_2$.

Критические точки находят, исходя из требования, чтобы, при справедливости нулевой гипотезы, сумма вероятностей того, что критерий примет значение меньшее k_1 или большее k_2, была равна принятому уровню значимости:

$$P(K<k_1)+P(K>k_2)=\alpha. \qquad (*)$$

Ясно, что критические точки могут быть выбраны бесчисленным множеством способов. Если же распределение критерия симметрично относительно нуля и имеются основания (например, для увеличения мощности*) выбрать симметричные относительно нуля точки — $k_{кр}$ и $k_{кр}$ ($k_{кр}>0$), то

$$P(K<-k_{кр})=P(K>k_{кр}).$$

Учитывая (*), получим

$$P(K>k_{кр})=\frac{\alpha}{2}.$$

Это соотношение и служит для отыскания критических точек двусторонней критической области.

Как уже было указано (§ 5), критические точки находят по соответствующим таблицам.

§ 7. Дополнительные сведения о выборе критической области. Мощность критерия

Мы строили критическую область, исходя из требования, чтобы вероятность попадания в нее критерия была равна α, при условии, что нулевая гипотеза справедлива. Оказывается целесообразным ввести в рассмотрение вероятность попадания критерия в критическую область при условии, что нулевая гипотеза неверна и, следовательно, справедлива конкурирующая.

Мощностью критерия называют вероятность попадания критерия в критическую область, при условии, что справедлива конкурирующая гипотеза. Другими словами, мощность критерия есть вероятность того, что нулевая гипотеза будет отвергнута, если верна конкурирующая гипотеза.

Пусть для проверки гипотезы принят определенный уровень значимости и выборка имеет фиксированный объем. Остается произвол в выборе критической области. Покажем, что ее целесообразно построить так, чтобы мощность критерия была максимальной.

Предварительно убедимся, что если вероятность ошибки второго рода (принять неправильную гипотезу) равна β,

* Определение мощности дано в § 7.

то мощность равна 1—β. Действительно, если β — вероятность ошибки второго рода, т. е. события «принята нулевая гипотеза, причем справедлива конкурирующая», то вероятность противоположного события «отвергнута нулевая гипотеза, причем справедлива конкурирующая», т. е. мощность критерия равна 1—β.

Пусть мощность 1—β возрастает; следовательно, уменьшается вероятность β совершить ошибку второго рода. Таким образом, чем мощность больше, тем вероятность ошибки второго рода меньше.

Итак, если уровень значимости уже выбран, то *критическую область следует строить так, чтобы мощность критерия была максимальной*. Выполнение этого требования обеспечит минимальную ошибку второго рода, что, конечно, желательно.

Замечание 1. Поскольку вероятность события «ошибка второго рода допущена» равна β, то вероятность противоположного события «ошибка второго рода не допущена» равна 1—β, т. е. мощности критерия. Отсюда следует, что мощность критерия есть вероятность того, что не будет допущена ошибка второго рода.

Замечание 2. Ясно, что чем меньше вероятности ошибок первого и второго рода, тем критическая область «лучше». Однако, при заданном объеме выборки, *уменьшить одновременно α и β невозможно*: если уменьшать α, то β будет возрастать. Например, если принять α = 0, то будут приниматься все гипотезы, в том числе и неправильные, т. е. возрастает вероятность β ошибки второго рода.

Как же выбрать α наиболее целесообразно? Ответ на этот вопрос зависит от «тяжести последствий» ошибок для каждой конкретной задачи. Например, если ошибка первого рода повлечет большие потери, а второго рода — малые, то следует принять возможно меньшее α.

Если α уже выбрано, то пользуясь теоремой Ю. Неймана и Э. Пирсона, изложенной в более полных курсах, можно построить критическую область, для которой β будет минимальным и, следовательно, мощность критерия максимальной

Замечание 3. Единственный способ *одновременного* уменьшения вероятностей ошибок первого и второго рода состоит в увеличении объема выборок.

§ 8. Сравнение двух дисперсий нормальных генеральных совокупностей

На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, инструментов, самих методов измерений и т. д. Очевидно, предпочтительнее

тот прибор, инструмент и метод, который обеспечивает наименьшее рассеяние результатов измерений, т. е. наименьшую дисперсию.

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам объемов n_1 и n_2, извлеченным из этих совокупностей, найдены исправленные выборочные дисперсии s_X^2 и s_Y^2. Требуется по исправленным дисперсиям, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0 : D(X)=D(Y).$$

Учитывая, что исправленные дисперсии являются несмещенными оценками генеральных дисперсий (гл. XVI, § 13), т. е.

$$M[s_X^2]=D(X), \quad M[s_Y^2]=D(Y),$$

нулевую гипотезу можно записать так:

$$H_0 : M[s_X^2] = M[s_Y^2].$$

Таким образом, требуется проверить, что математические ожидания исправленных выборочных дисперсий равны между собой. Такая задача ставится потому, что обычно исправленные дисперсии оказываются различными. Возникает вопрос: *значимо* (существенно) *или незначимо, различаются исправленные дисперсии*?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные дисперсии одинаковы, то различие исправленных дисперсий незначимо и объясняется случайными причинами, в частности, случайным отбором объектов выборки. Например, если различие исправленных выборочных дисперсий результатов измерений, выполненных двумя приборами, оказалось незначимым, то приборы имеют одинаковую точность.

Если нулевая гипотеза будет отвергнута, т. е. генеральные дисперсии неодинаковы, то различие исправленных дисперсий значимо и не может быть объяснено случайными причинами, а является следствием того, что сами генеральные дисперсии различны. Например, если различие исправленных выборочных дисперсий результатов измерений, произведенных двумя приборами, оказалось значимым, то точность приборов различна.

В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий, примем отношение большей исправленной дисперсии к меньшей, т. е. случайную величину

$$F = \frac{s_б^2}{s_м^2}.$$

Величина F, при условии справедливости нулевой гипотезы имеет распределение Фишера—Снедекора (гл. XII, § 15) со степенями свободы $k_1=n_1-1$ и $k_2=n_2-1$, где n_1 — объем выборки, по которой вычислена большая исправленная дисперсия, n_2 — объем выборки, по которой найдена меньшая дисперсия.

Напомним, что распределение Фишера—Снедекора зависит только от чисел степеней свободы и не зависит от других параметров.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза H_0: $D(X)=D(Y)$. Конкурирующая гипотеза $H_1 : D(X)>D(Y)$.

В этом случае строят одностороннюю, а именно правостороннюю, критическую область, исходя из требования, чтобы вероятность попадания критерия F в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости:

$$P[F>F_{кр}(\alpha,\ k_1,\ k_2)]=\alpha.$$

Критическую точку $F_{кр}(\alpha,\ k_1,\ k_2)$ находят по таблице критических точек распределения Фишера—Снедекора (приложение 7) и тогда правосторонняя критическая область определяется неравенством

$$F>F_{кр},$$

а область принятия нулевой гипотезы неравенством

$$F<F_{кр}.$$

Обозначим отношение большей исправленной дисперсии к меньшей, вычисленное по данным наблюдений, через $F_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило 1. Для того чтобы, при заданном уровне значимости, проверить нулевую гипотезу $H_0 : D(X)=D(Y)$

о равенстве генеральных дисперсий нормальных совокупностей, при конкурирующей гипотезе $H_1 : D(X)>D(Y)$, надо вычислить отношение большей исправленной дисперсии к меньшей, т. е.

$$F_{набл} = \frac{s_б^2}{s_м^2}$$

и по таблице критических точек распределения Фишера — Снедекора, по заданному уровню значимости α и числам степеней свободы k_1 и k_2 (k_1 — число степеней свободы большей исправленной дисперсии), найти критическую точку $F_{набл}(\alpha,\ k_1,\ k_2)$.

Если $F_{набл}<F_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $F_{набл}>F_{кр}$ — нулевую гипотезу отвергают.

Пример 1. По двум независимым выборкам объемов $n_1=12$ и $n_2=15$, извлеченным из нормальных генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии $s_X^2=11,41$ и $s_Y^2=6,52$. При уровне значимости 0,05, проверить нулевую гипотезу $H_0 : D(X)=D(Y)$ о равенстве генеральных дисперсий, при конкурирующей гипотезе $H_1 : D(X)>D(Y)$.

Решение. Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{набл} = \frac{11,41}{6,52} = 1,75.$$

Так как конкурирующая гипотеза имеет вид $D(X)>D(Y)$, критическая область — правосторонняя.

По таблице (приложение 7), по уровню значимости $\alpha=0,05$ и числам степеней свободы $k_1=12-1=11$ и $k_2=15-1=14$, находим критическую точку $F_{кр}(0,05;\ 11;\ 14)=2,57$.

Так как $F_{набл}<F_{кр}$ — нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий. Здесь и далее критические точки для уровня значимости 0; 0,5 взяты из табл. VI книги, указанной в сноске на стр 331; на уровне значимости 0; 0,1 критические точки помещены в приложении 7 настоящей книги.

Второй случай. Нулевая гипотеза $H_0 : D(X)=D(Y)$. Конкурирующая гипотеза $H_1 : D(X)\neq D(Y)$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Как выбрать границы критической области? Оказывается, что наибольшая мощность (вероятность попадания критерия в критическую область, при справедливости конкурирующей гипотезы) достигается тогда, когда вероятность попадания критерия в каждый из двух интервалов критической области равна $\frac{\alpha}{2}$.

Таким образом, если обозначить через F_1 левую границу критической области и через F_2 — правую, то должны иметь место соотношения (рис. 24):

$$P(F < F_1) = \frac{\alpha}{2}, \quad P(F > F_2) = \frac{\alpha}{2}.$$

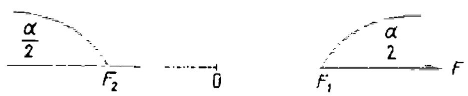

Рис. 24.

Мы видим, что достаточно найти критические точки, чтобы найти саму критическую область:

$$F<F_1, \quad F>F_2,$$

а также область принятия нулевой гипотезы:

$$F_1<F<F_2.$$

Как практически отыскать критические точки?

Правую критическую точку $F_2=F_{кр}(\frac{\alpha}{2}, k_1, k_2)$ находят непосредственно по таблице критических точек распределения Фишера—Снедекора по уровню значимости $\frac{\alpha}{2}$ и степеням свободы k_1 и k_2.

Однако, левых критических точек эта таблица не содержит и поэтому найти F_1 непосредственно по таблице невозможно.

Существует способ, позволяющий преодолеть это затруднение. Однако, мы не будем его описывать, поскольку можно левую критическую точку и не отыскивать. Ограничимся изложением того, как обеспечить попадание критерия F в двустороннюю критическую область с вероятностью, равной принятому уровню значимости α.

Оказывается достаточно найти правую критическую точку F_2 при уровне значимости, вдвое меньшем заданного. Тогда не только вероятность попадания критерия в «правую часть» критической области (т. е. правее F_2) равна $\frac{\alpha}{2}$, но и вероятность попадания этого критерия в «левую часть» критической области (т. е. левее F_1) будет также равна $\frac{\alpha}{2}$. Так как эти события несовместны, то вероятность попадания рассматриваемого критерия во всю двустороннюю критическую область будет равна $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$.

Таким образом, в случае конкурирующей гипотезы $H_1 : D(X) \neq D(Y)$, достаточно найти критическую точку $F_2=F_{кр}(\frac{\alpha}{2}, k_1, k_2)$.

Правило 2. Для того чтобы, при заданном уровне значимости α проверить нулевую гипотезу о равенстве генеральных дисперсий нормально распределенных совокупностей, при конкурирующей гипотезе $H_1 : D(X) \neq D(Y)$, надо вычислить отношение большей исправленной дисперсии к меньшей, т. е. $F_{набл} = \frac{s_б^2}{s_м^2}$ и по таблице критических точек распределения Фишера—Снедекора по уровню значимости $\frac{\alpha}{2}$ (вдвое меньшем заданного) и числам степеней свободы k_1 и k_2^2 (k_1 — число степеней свободы большей дисперсии) найти критическую точку $F_{кр}(\frac{\alpha}{2}, k_1, k_2)$.

Если $F_{набл} < F_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $F_{набл} > F_{кр}$ — нулевую гипотезу отвергают.

Пример 2. По двум независимым выборкам объемов $n_1=10$ и $n_2=18$, извлеченным из нормальных генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии $s_X^2 = 1{,}23$ и $s_Y^2 = 0{,}41$. При уровне значимости $\alpha=0{,}1$ проверить нулевую гипотезу о равенстве генеральных дисперсий, при конкурирующей гипотезе $H_1 : D(X) \neq D(Y)$.

Р е ш е н и е. Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{набл} = \frac{1{,}23}{0{,}41} = 3.$$

По условию конкурирующая гипотеза имеет вид $D(X) \neq D(Y)$, поэтому критическая область — двусторонняя.

По таблице, по уровню значимости, вдвое меньшем заданного, т. е. при $\frac{\alpha}{2}=\frac{0,1}{2}=0,05$ и числам степеней свободы $k_1=10-1=9$, $k_2=18-1=17$, находим критическую точку $F_{кр}$ (0,05; 9; 17)=2,50.

Так как $F_{набл} > F_{кр}$, нулевую гипотезу о равенстве генеральных дисперсий отвергаем. Другими словами, выборочные исправленные дисперсии различаются значимо. Например, если бы рассматриваемые дисперсии характеризовали точность двух методов измерений, то следует предпочесть тот метод, который имеет меньшую дисперсию (0,41).

§ 9. Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

Пусть генеральная совокупность распределена нормально, причем генеральная дисперсия, хотя и неизвестна, но имеются основания предполагать, что она равна гипотетическому (предполагаемому) значению σ_0^2. На практике σ_0^2 устанавливается на основании предшествующего опыта, или теоретически.

Пусть из генеральной совокупности извлечена выборка объема n и по ней найдена исправленная выборочная дисперсия S^2 с $k=n-1$ степенями свободы. Требуется по исправленной дисперсии, при заданном уровне значимости, проверить нулевую гипотезу, состоящую в том, что генеральная дисперсия рассматриваемой совокупности равна гипотетическому значению σ_0^2.

Учитывая, что S^2 является несмещенной оценкой генеральной дисперсии, нулевую гипотезу можно записать так:

$$H_0: M(S^2)=\sigma_0^2.$$

Итак, требуется проверить, что математическое ожидание исправленной дисперсии равно гипотетическому значению генеральной дисперсии Другими словами, **требуется установить значимо, или незначимо, различаются исправленная выборочная и гипотетическая генеральная дисперсии.**

На практике рассматриваемая гипотеза проверяется, если нужно проверить точность приборов, инструментов, станков, методов исследования и устойчивость технологических процессов. Например, если известна допустимая характеристика рассеяния контролируемого размера деталей, изготавливаемых станком-автоматом, равная σ_0^2, а найденная по выборке исправленная дисперсия окажется значимо больше σ_0^2, то станок требует подналадки.

В качестве критерия проверки нулевой гипотезы примем случайную величину $\frac{(n-1)S^2}{\sigma_0^2}$. Эта величина случайная, потому что в разных опытах S^2 будет принимать различные, наперед неизвестные значения. Поскольку можно доказать, что она имеет распределение χ^2 с $k=n-1$ степенями свободы (гл. XII, § 13), обозначим ее через χ^2.

Итак, критерий проверки нулевой гипотезы

$$\chi^2=\frac{(n-1)\,S^2}{\sigma_0^2}.$$

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: \sigma^2=\sigma_0^2$. Конкурирующая гипотеза $H_1: \sigma^2>\sigma_0^2$.

В этом случае строят правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости:

$$P\left[\chi^2>\chi^2_{кр}(\alpha, k)\right]=\alpha.$$

Критическую точку $\chi^2_{кр}(\alpha, k)$ находят по таблице критических точек распределения χ^2 (приложение 5) и тогда правосторонняя критическая область определяется неравенством

$$\chi^2>\chi^2_{кр},$$

а область принятия нулевой гипотезы неравенством

$$\chi^2<\chi^2_{кр}.$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило 1. Для того чтобы, при заданном уровне значимости α проверить нулевую гипотезу $H_0: \sigma^2=\sigma_0^2$ о равенстве неизвестной генеральной дисперсии нормальной совокупности гипотетическому значению, при конкурирующей гипотезе $H_1: \sigma^2 > \sigma_0^2$, надо вычислить наблюдаемое значение критерия $\chi^2_{набл}=\frac{(n-1)S^2}{\sigma_0^2}$ и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k=n-1$, найти критическую точку $\chi^2_{кр}(\alpha,\ k)$.

Если $\chi^2_{набл}<\chi^2_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл}>\chi^2_{кр}$ — нулевую гипотезу отвергают.

Пример 1. Из нормальной генеральной совокупности извлечена выборка объема $n=13$ и по ней найдена исправленная выборочная дисперсия $s^2=14{,}6$. Требуется, при уровне значимости 0,01, проверить нулевую гипотезу $H_0: \sigma^2=\sigma_0^2=12$, приняв в качестве конкурирующей гипотезы $H_1: \sigma^2>12$.

Решение. Найдем наблюденное значение критерия

$$\chi^2_{набл}=\frac{(n-1)\,S^2}{\sigma_0^2}=\frac{(13-1)\cdot 14{,}6}{12}=14{,}6.$$

По условию конкурирующая гипотеза имеет вид $\sigma^2>12$, поэтому критическая область правосторонняя.

По таблице (приложение 5), по уровню значимости 0,01 и числу степеней свободы $k=n-1=13-1=12$, находим критическую точку $\chi^2_{кр}(0{,}01;\ 12)=26{,}2$

Так как $\chi^2_{набл}<\chi^2_{кр}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, различие между исправленной дисперсией (14,6) и гипотетической генеральной дисперсией (12) — незначимое

Второй случай. Нулевая гипотеза $H_0: \sigma^2=\sigma_0^2$. Конкурирующая гипотеза $H_1: \sigma^2\neq\sigma_0^2$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Критические точки — левую и правую границы критической области — находят, требуя, чтобы вероятность попадания критерия в каждый из двух интервалов критической области была равна $\frac{\alpha}{2}$:

$$P\left[\chi^2<\chi^2_{лев.\ кр}\left(\frac{\alpha}{2},\ k\right)\right]=\frac{\alpha}{2},$$

$$P\left[\chi^2>\chi^2_{прав.\ кр}\left(\frac{\alpha}{2},\ k\right)\right]=\frac{\alpha}{2}.$$

В таблице критических точек распределения χ^2 указаны лишь «правые» критические точки, поэтому возникает кажущееся затруднение в отыскании «левой» критической точки. Это затруднение легко преодолеть, если принять во внимание, что события

$$\chi^2<\chi^2_{лев.\ кр} \quad и \quad \chi^2>\chi^2_{лев.\ кр}$$

противоположны и, следовательно, сумма их вероятностей равна единице:

$$P\left(\chi^2<\chi^2_{лев.\ кр}\right)+P\left(\chi^2>\chi^2_{лев.\ кр}\right)=1.$$

Отсюда

$$P\left(\chi^2>\chi^2_{лев.\ кр}\right)=1-P\left(\chi^2<\chi^2_{лев.\ кр}\right)=1-\frac{\alpha}{2}.$$

Мы видим, что левую критическую точку можно искать как правую (и значит ее можно найти по таблице), исходя из требования, чтобы вероятность попадания критерия в интервал, расположенный правее этой точки, была равна $1-\frac{\alpha}{2}$.

Правило 2. Для того чтобы, при заданном уровне значимости α, проверить нулевую гипотезу о равенстве неизвестной генеральной дисперсии σ^2 нормальной совокупности гипотетическому значению σ_0^2, при конкурирующей гипотезе $H_1: \sigma^2\neq\sigma_0^2$, надо вычислить наблюдаемое значение критерия $\chi^2_{набл}=\frac{(n-1)\,s^2}{\sigma_0^2}$ и по таблице найти левую критическую точку $\chi^2_{кр}\left(1-\frac{\alpha}{2},\ k\right)$ и правую — $\chi^2_{кр}\left(\frac{\alpha}{2},\ k\right)$.

Если $\chi^2_{лев.\ кр}<\chi^2_{набл.}<\chi^2_{прав.\ кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл} < \chi^2_{лев.\ кр}$, или $\chi^2_{набл} > \chi^2_{прав.\ кр}$, — нулевую гипотезу отвергают.

Пример 2. Из нормальной генеральной совокупности извлечена выборка объема $n=13$ и по ней найдена исправленная выборочная дисперсия $s^2=10,3$. Требуется, при уровне значимости 0,02, проверить нулевую гипотезу $H_0: \sigma^2=\sigma_0^2=12$, приняв в качестве конкурирующей гипотезы $H_1: \sigma^2 \neq 12$.

Решение. Найдем наблюденное значение критерия

$$\chi^2_{набл} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(13-1)\cdot 10,3}{12} = 10,3.$$

Так как конкурирующая гипотеза имеет вид $\sigma^2 \neq 12$, то критическая область — двусторонняя.

По таблице (приложение 5) находим критические точки: левую — $\chi^2_{кр}\left(1-\frac{\alpha}{2}, k\right) = \chi^2_{кр}\left(1-\frac{0,02}{2},\ 12\right) = \chi^2_{кр}(0,99;\ 12) = 3,57$ и правую — $\chi^2_{кр}\left(\frac{\alpha}{2}, k\right) = \chi^2_{кр}(0,01;\ 12) = 26,2$.

Так как наблюденное значение критерия принадлежит области принятия гипотезы $(3,57 < 10,3 < 26,2)$ — нет оснований ее отвергнуть. Другими словами, исправленная выборочная дисперсия (10,3) незначимо отличается от гипотетической генеральной дисперсии (12).

Случай 3. Конкурирующая гипотеза $H_1: \sigma^2 < \sigma_0^2$.

Правило 3. При конкурирующей гипотезе $H_1: \sigma^2 < \sigma_0^2$, находят критическую точку $\chi^2_{кр}(1-\alpha, k)$.

Если $\chi^2_{набл} > \chi^2_{кр}(1-\alpha, k)$ — нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл} < \chi^2_{кр}(1-\alpha, k)$ — нулевую гипотезу отвергают.

Замечание. В случае, если найдена выборочная дисперсия $D_в$, в качестве критерия принимают случайную величину $\chi^2 = \frac{nD_в}{\sigma_0^2}$, которая имеет распределение χ^2 с $k=n-1$ степенями свободы, либо переходят к $s^2 = \frac{n}{n-1}D_в$.

§ 10. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых известны (независимые выборки)

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии известны (например, из предшествующего опыта, или найдены теоретически). По независимым выборкам объемов n и m, извлеченным из этих совокупностей, найдены выборочные средние $\overline{x}$ и $\overline{y}$.

Требуется по выборочным средним, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что генеральные средние (математические ожидания) рассматриваемых совокупностей равны между собой, т. е.

$$H_0: M(X)=M(Y).$$

Учитывая, что выборочные средние являются несмещенными оценками генеральных средних (гл. XV, § 5), т. е. $M(\overline{X})=M(X)$ и $M(\overline{Y})=M(Y)$, нулевую гипотезу можно записать так:

$$H_0: M(\overline{X})=M(\overline{Y}).$$

Таким образом, требуется проверить, что математические ожидания выборочных средних равны между собой. Такая задача ставится потому, что, как правило, выборочные средние оказываются различными. Возникает вопрос: значимо, или незначимо различаются выборочные средние?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные средние одинаковы, то различие выборочных средних незначимо и объясняется случайными причинами и, в частности, случайным отбором объектов выборки.

Например, если физические величины A и B имеют одинаковые истинные размеры, а средние арифметические $\overline{x}$ и $\overline{y}$ результатов измерений этих величин различны, то это различие незначимое.

Если нулевая гипотеза будет отвергнута, т. е. генеральные средние не одинаковы, то различие выборочных средних значимо и не может быть объяснено случайными причинами, а объясняется тем, что сами генеральные средние (математические ожидания) различны. Например, если среднее арифметическое $\overline{x}$ результатов измерений физической величины A значимо отличается от среднего арифметического $\overline{y}$ результатов измерений физической величины B, то это означает, что истинные размеры (математические ожидания) этих величин — различны.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$Z=\frac{\overline{X}-\overline{Y}}{\sigma(\overline{X}-\overline{Y})}=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{D(X)}{n}+\frac{D(Y)}{m}}}.$$

Эта величина случайная, потому что в различных опытах $\overline{x}$ и $\overline{y}$ принимают различные, наперед неизвестные значения.

Пояснение. По определению среднего квадратического отклонения $\sigma(\overline{X}-\overline{Y})=\sqrt{D(\overline{X}-\overline{Y})}$.

На основании свойства 4 (гл. VIII, § 5) $D(\overline{X}-\overline{Y})=D(\overline{X})+D(\overline{Y})$.

По формуле (*) (гл. VIII, § 9): $D(\overline{X})=\frac{D(X)}{n}$, $D(\overline{Y})=\frac{D(Y)}{n}$. Следовательно,

$$\sigma(\overline{X}-\overline{Y})=\sqrt{\frac{D(X)}{n}+\frac{D(Y)}{m}}.$$

Критерий Z — нормированная нормальная случайная величина. Действительно, величина Z распределена нормально, так как является линейной комбинацией нормально распределенных величин $\overline{X}$ и $\overline{Y}$; сами эти величины распределены нормально как выборочные средние, найденные по выборкам, извлеченным из нормальных генеральных совокупностей; Z — нормированная величина потому, что $M(Z)=0$, при справедливости нулевой гипотезы, $\sigma(Z)=1$, поскольку выборки независимы.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: M(X)=M(Y)$. Конкурирующая гипотеза $H_1: M(X)\neq M(Y)$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Наибольшая мощность критерия (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы) достигается тогда, когда «левая» и «правая» критические точки выбраны так, что

вероятность попадания критерия в каждый из двух интервалов критической области, равна $\frac{\alpha}{2}$:

$$P(Z<z_{\text{лев. кр}})=\frac{\alpha}{2}, \qquad (*)$$

$$P(Z>z_{\text{прав. кр}})=\frac{\alpha}{2}.$$

Поскольку Z нормированная нормальная величина, а распределение такой величины симметрично относительно нуля,— критические точки симметричны относительно нуля.

Таким образом, если обозначить правую границу дву-

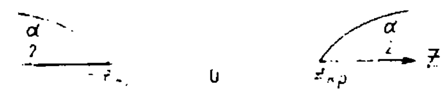

Рис. 25.

сторонней критической области через $z_{\text{кр}}$, то левая граница равна $-z_{\text{кр}}$ (рис. 25).

Итак, достаточно найти правую границу, чтобы найти саму двустороннюю критическую область

$$Z<-z_{\text{кр}}, \quad Z>z_{\text{кр}}$$

и область принятия нулевой гипотезы

$$(-z_{\text{кр}}, z_{\text{кр}}).$$

Покажем как найти $z_{\text{кр}}$ — правую границу двусторонней критической области, пользуясь функцией Лапласа $\Phi(Z)$. Известно, что функция Лапласа определяет вероятность попадания нормированной нормальной случайной величины, например Z, в интервал $(0, z)$:

$$P(0<Z<z)=\Phi(Z). \qquad (**)$$

Так как распределение Z симметрично относительно нуля, то вероятность попадания Z в интервал $(0, \infty)$ равна $\frac{1}{2}$. Следовательно, если разбить этот интервал точкой $z_{\text{кр}}$, на интервалы $(0, z_{\text{кр}})$ и $(z_{\text{кр}}, \infty)$, то по теореме сложения

$$P(0<Z<z_{\text{кр}})+P(Z>z_{\text{кр}})=\frac{1}{2}. \qquad (***)$$

В силу (*) и (**) получим

$$\Phi(z_{кр}) + \frac{\alpha}{2} = \frac{1}{2}.$$

Следовательно,

$$\Phi(z_{кр}) = \frac{1-\alpha}{2}.$$

Отсюда заключаем: для того чтобы найти правую границу двусторонней критической области ($z_{кр}$) достаточно найти значение аргумента функции Лапласа, которому соответствует значение функции, равное $\frac{1-\alpha}{2}$.

Тогда двусторонняя критическая область определяется неравенствами

$$Z < -z_{кр}, \quad Z > z_{кр}$$

или равносильным неравенством

$$|Z| > z_{кр},$$

а область принятия нулевой гипотезы неравенством

$$-z_{кр} < Z < z_{кр},$$

или равносильным неравенством

$$|Z| < z_{кр}.$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $Z_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило 1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X)=M(Y)$ о равенстве математических ожиданий двух нормальных генеральных совокупностей с известными дисперсиями, при конкурирующей гипотезе $H_1: M(X)\neq M(Y)$, надо вычислить наблюденное значение критерия $Z_{набл} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{D(X)}{n}+\frac{D(Y)}{m}}}$ и по таблице функции Лапласа найти критическую точку по равенству $\Phi(z_{кр}) = \frac{1-\alpha}{2}$.

Если $|Z_{набл}| < z_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|Z_{набл}| > z_{кр}$ — нулевую гипотезу отвергают.

Пример 1. По двум независимым выборкам объемов $n=60$ и $m=50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}=1250$ и $\bar{y}=1275$. Генеральные дисперсии известны: $D(X)=120$, $D(Y)=100$. При уровне значимости 0,01, проверить нулевую гипотезу $H_0: M(X)=M(Y)$, при конкурирующей гипотезе $H_1: M(X)\neq M(Y)$.

Решение. Найдем наблюдаемое значение критерия

$$Z_{набл} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{D(X)}{n}+\frac{D(Y)}{m}}} = \frac{1250-1275}{\sqrt{\frac{120}{60}+\frac{100}{50}}} = -12{,}5.$$

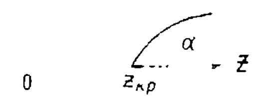

Рис. 26.

По условию конкурирующая гипотеза имеет вид $M(X)\neq M(Y)$, поэтому критическая область — двусторонняя

Найдем правую критическую точку по равенству

$$\Phi(z_{кр}) = \frac{1-\alpha}{2} = \frac{1-0{,}01}{2} = 0{,}495.$$

По таблице функции Лапласа (приложение 2) находим $z_{кр} = 2{,}58$.

Так как $|Z_{набл}| > z_{кр}$ — нулевую гипотезу отвергаем. Другими словами, выборочные средние различаются значимо.

Второй случай. Нулевая гипотеза $H_0: M(X)=M(Y)$. Конкурирующая гипотеза $H_1: M(X)>M(Y)$.

На практике такой случай имеет место, если профессиональные соображения позволяют предположить, что генеральная средняя одной совокупности больше генеральной средней другой. Например, если введено усовершенствование технологического процесса, то естественно допустить, что оно приведет к увеличению выпуска продукции.

В этом случае строят правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости (рис. 26):

$$P(Z > z_{кр}) = \alpha. \qquad (****)$$

Покажем как найти критическую точку при помощи функции Лапласа. Воспользуемся соотношением (***):

$$P(0 < Z < z_{кр}) + P(Z > z_{кр}) = \frac{1}{2}.$$

В силу (**) и (****) имеем

$$\Phi(z_{кр}) + \alpha = \frac{1}{2}.$$

Следовательно,

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2}.$$

Отсюда заключаем, для того чтобы найти границу правосторонней критической области ($z_{кр}$), достаточно найти значение аргумента функции Лапласа, которому соответствует значение функции, равное $\frac{1-2\alpha}{2}$. Тогда правосторонняя критическая область определяется неравенством $Z > z_{кр}$, а область принятия нулевой гипотезы — неравенством $Z < z_{кр}$.

Правило 2. Для того чтобы, при заданном уровне значимости α, проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве математических ожиданий двух нормальных генеральных совокупностей с известными дисперсиями, при конкурирующей гипотезе $H_1: M(X) > M(Y)$, надо вычислить наблюденное значение критерия $Z_{набл} = \dfrac{\bar{x}-\bar{y}}{\sqrt{\frac{D(X)}{n}+\frac{D(Y)}{m}}}$ и по таблице функции Лапласа найти критическую точку из равенства $\Phi(z_{кр}) = \frac{1-2\alpha}{2}$.

Если $Z_{набл} < z_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $Z_{набл} > z_{кр}$ — нулевую гипотезу отвергают.

Пример 2. По двум независимым выборкам объемов $n=10$ и $m=10$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 14{,}3$ и $\bar{y} = 12{,}2$. Генеральные дисперсии известны: $D(X) = 22$,

$D(Y) = 18$. При уровне значимости 0,05, проверить нулевую гипотезу $H_0: M(X) = M(Y)$, при конкурирующей гипотезе $H_1: M(X) > M(Y)$.

Решение. Найдем наблюдаемое значение критерия

$$Z_{набл} = \frac{14{,}3 - 12{,}2}{\sqrt{\frac{22}{10}+\frac{18}{10}}} = 1{,}05.$$

По условию конкурирующая гипотеза имеет вид $M(X) > M(Y)$, поэтому критическая область — правосторонняя.

По таблице функции Лапласа находим $z_{кр} = 1{,}64$.

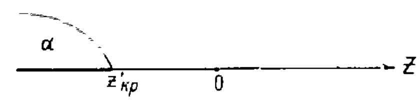

Рис. 27.

Так как $Z_{набл} < z_{кр}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, выборочные средние различаются незначимо.

Третий случай. Нулевая гипотеза $H_0: M(X) = M(Y)$. Конкурирующая гипотеза $H_1: M(X) < M(Y)$.

В этом случае строят левостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости (рис. 27):

$$P(Z < z'_{кр}) = \alpha.$$

Приняв во внимание, что критерий Z распределен симметрично относительно нуля, заключаем, что искомая критическая точка $z'_{кр}$ симметрична такой точке $z_{кр} > 0$, для которой $P(Z > z_{кр}) = \alpha$, т. е. $z'_{кр} = -z_{кр}$. Таким образом, для того, чтобы найти точку $z'_{кр}$, достаточно сначала найти «вспомогательную точку» $z_{кр}$ так, как описано во втором случае, а затем взять найденное значение со знаком минус. Тогда левосторонняя критическая область определяется неравенством $Z < -z_{кр}$, а область принятия нулевой гипотезы неравенством $Z > -z_{кр}$.

Правило 3. При конкурирующей гипотезе $H_1: M(X) < M(Y)$ надо вычислить $Z_{набл}$ и сначала по таблице функции Лапласа найти «вспомогательную точку» $z_{кр}$ по равенству $\Phi(z_{кр}) = \frac{1-2\alpha}{2}$, а затем положить $z'_{кр} = -z_{кр}$.

Если $Z_{набл} > -z_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $Z_{набл} < -z_{кр}$ — нулевую гипотезу отвергают.

Пример 3. По двум независимым выборкам объемов $n=50$ и $m=50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}=142$ и $\bar{y}=150$. Генеральные дисперсии известны: $D(X)=28{,}2$, $D(Y)=22{,}8$. При уровне значимости 0,01, проверить нулевую гипотезу $H_0: M(X)=M(Y)$, при конкурирующей гипотезе $H_1: M(X)<M(Y)$.

Решение. Подставив данные задачи в формулу для вычисления наблюдаемого значения критерия, получим $Z_{набл} = -8$.

По условию конкурирующая гипотеза имеет вид $M(X) < M(Y)$, поэтому критическая область — левосторонняя.

Найдем «вспомогательную точку» $z_{кр}$ по равенству

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}01}{2} = 0{,}49.$$

По таблице функции Лапласа находим $z_{кр}=2{,}33$. Следовательно, $z'_{кр} = -z_{кр} = -2{,}33$.

Так как $Z_{набл} < -z_{кр}$ — нулевую гипотезу отвергаем. Другими словами, выборочная средняя $\bar{x}$ значимо меньше выборочной средней $\bar{y}$.

§ 11. Сравнение двух средних произвольно распределенных генеральных совокупностей (большие независимые выборки)

В предыдущем параграфе предполагалось, что генеральные совокупности X и Y распределены нормально, а их дисперсии известны. При этих предположениях, в случае справедливости нулевой гипотезы о равенстве средних и независимых выборках, критерий Z распределен точно нормально с параметрами 0 и 1.

Если хотя бы одно из приведенных требований не выполняется, метод сравнения средних, описанный в § 10, неприменим.

Однако, если независимые выборки имеют большой объем (не менее 30 каждая), то выборочные средние распределены приближенно нормально, а выборочные дисперсии являются достаточно хорошими оценками генеральных дисперсий и в этом смысле их можно считать известными приближенно. В итоге критерий

$$Z' = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{D_в(X)}{n} + \frac{D_в(Y)}{m}}}$$

распределен приближенно нормально с параметрами $M(Z')=0$ (при условии справедливости нулевой гипотезы) и $\sigma(Z')=1$ (если выборки независимы).

Итак, если: 1) генеральные совокупности распределены нормально, а дисперсии их неизвестны; 2) генеральные совокупности не распределены нормально, а дисперсии их известны; 3) генеральные совокупности не распределены нормально и дисперсии их неизвестны, причем выборки имеют большой объем и независимы,— можно сравнивать средние так, как описано в § 10, заменив точный критерий Z приближенным критерием Z'. В этом случае наблюдаемое значение приближенного критерия таково:

$$Z'_{набл} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{D_в(X)}{n} + \frac{D_в(Y)}{m}}}.$$

Замечание. Поскольку рассматриваемый критерий — приближенный, к выводам, полученным по этому критерию, следует относиться осторожно.

Пример. По двум независимым выборкам объемов $n=100$ и $m=120$, найдены выборочные средние $\bar{x}=32{,}4$, $\bar{y}=30{,}1$ и выборочные дисперсии $D_в(X)=15{,}0$, $D_в(Y)=25{,}2$. При уровне значимости 0,05, проверить нулевую гипотезу $H_0: M(X)=M(Y)$, при конкурирующей гипотезе $H_1: M(X)\neq M(Y)$.

Решение. Подставив данные задачи в формулу для вычисления наблюдаемого значения приближенного критерия, получим $Z'_{набл} = 3{,}83$

По условию конкурирующая гипотеза имеет вид $M(X) > M(Y)$, поэтому критическая область — правосторонняя.

Найдем критическую точку по равенству

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}05}{2} = 0{,}45.$$

По таблице функции Лапласа находим $z_{кр}=1{,}64$.

Так как $Z_{набл} > z_{кр}$ — нулевую гипотезу отвергаем. Другими словами, выборочные средние различаются значимо.

§ 12. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки)

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. Например, по выборкам малого объема нельзя получить хорошие оценки генеральных дисперсий. По этой причине метод сравнения средних, изложенный в § 11, применить нельзя.

Однако, если дополнительно предположить, что неизвестные генеральные дисперсии равны между собой, то можно построить критерий (Стьюдента) сравнения средних. Например, если сравниваются средние размеры двух партий деталей, изготовленных на одном и том же станке, то естественно допустить, что дисперсии контролируемых размеров одинаковы.

Если же нет оснований считать дисперсии одинаковыми, то прежде чем сравнивать средние, следует, пользуясь критерием Фишера—Снедекора (§ 8), предварительно проверить гипотезу о равенстве генеральных дисперсий.

Итак, в предположении, что генеральные дисперсии одинаковы, требуется проверить нулевую гипотезу $H_0: M(X)=M(Y)$. Другими словами, требуется установить значимо или незначимо, различаются выборочные средние $\overline{x}$ и $\overline{y}$, найденные по независимым малым выборкам объемов n и m.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$T = \frac{\overline{X}-\overline{Y}}{\sqrt{(n-1)S_x^2+(m-1)S_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}.$$

Доказано, что величина T, при справедливости нулевой гипотезы, имеет t-распределение Стьюдента с $k=n+m-2$ степенями свободы.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Нулевая гипотеза $H_0: M(X)=M(Y)$. Конкурирующая гипотеза $H_1: M(X)\neq M(Y)$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия T в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Наибольшая мощность критерия (вероятность попадания критерия в критическую область, при справедливости конкурирующей гипотезы) достигается тогда, когда «левая» и «правая» критические точки выбраны так, что вероятность попадания критерия в каждый из двух интервалов двусторонней критической области равна $\frac{\alpha}{2}$:

$$P(T < t_{лев.\,кр}) = \frac{\alpha}{2}, \quad P(T > t_{прав.\,кр}) = \frac{\alpha}{2}.$$

Поскольку величина T имеет распределение Стьюдента, а оно симметрично относительно нуля, то и критические точки симметричны относительно нуля. Таким образом, если обозначить правую границу двусторонней критической области через $t_{двуст.\,кр}(\alpha, k)$, то левая граница равна $-t_{двуст.\,кр}(\alpha, k)$. Итак, достаточно найти правую границу двусторонней критической области, чтобы найти саму двустороннюю критическую область,

$$T < -t_{двуст.\,кр}(\alpha, k), \quad T > t_{двуст.\,кр}(\alpha, k)$$

и область принятия нулевой гипотезы

$$[-t_{двуст.\,кр}(\alpha, k),\ t_{двуст.\,кр}(\alpha, k)].$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $T_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило 1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X)=M(Y)$ о равенстве математических ожиданий двух нормальных совокупностей с неизвестными, но одинаковыми дисперсиями (в случае независимых малых выборок), при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$, надо вычислить наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α (помещенному в верхней строке таблицы) и числу степеней свободы $k=n+m-2$, найти критическую точку $t_{\text{двуст. кр}}(\alpha, k)$

Если $|T_{\text{набл}}| < t_{\text{двуст. кр}}(\alpha, k)$ — отвергнуть нулевую гипотезу — нет оснований.

Если $|T_{\text{набл}}| > t_{\text{двуст. кр}}(\alpha, k)$ — нулевую гипотезу отвергают.

Пример. По двум независимым малым выборкам объемов $n=5$ и $m=6$, извлеченным из нормальных генеральных совокупностей X и Y, найдены выборочные средние $\bar{x}=3{,}3$, $\bar{y}=2{,}48$ и исправленные дисперсии $s_X^2=0{,}25$ и $s_Y^2=0{,}108$. При уровне значимости 0,05, проверить нулевую гипотезу $H_0: M(X)=M(Y)$, при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$.

Решение. Так как выборочные дисперсии различны, проверим предварительно нулевую гипотезу о равенстве генеральных дисперсий, пользуясь критерием Фишера — Снедекора (§ 8).

Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{\text{набл}} = \frac{0{,}25}{0{,}108} = 2{,}31$$

Дисперсия s_X^2 значительно больше дисперсии s_Y^2, поэтому в качестве конкурирующей примем гипотезу $H_1: D(X) > D(Y)$. В этом случае критическая область — правосторонняя. По таблице, по уровню значимости $\alpha=0{,}05$ и числам степеней свободы $k_1=5-1=4$, $k_2=6-1=5$, находим критическую точку $F_{\text{кр}}(0{,}05; 4; 5)=5{,}19$.

Так как $F_{\text{набл}} < F_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий.

Поскольку предположение о равенстве генеральных дисперсий выполняется, сравним средние

Вычислим наблюдаемое значение критерия Стьюдента:

$$T_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{ns_X^2 + ms_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

Подставив числовые значения величин, входящих в эту формулу, получим: $T_{\text{набл}} = 3{,}27$.

По условию конкурирующая гипотеза имеет вид $M(X) \neq M(Y)$, поэтому критическая область — двусторонняя. По уровню значимости 0,05 и числу степеней свободы $k=5+6-2=9$, находим по таблице (приложение 6) критическую точку $t_{\text{двуст. кр}}(0{,}05; 9)=2{,}26$

Так как $T_{\text{набл}} > t_{\text{двуст. кр}}$ — нулевую гипотезу о равенстве генеральных средних отвергаем. Другими словами выборочные средние различаются значимо.

Второй случай. Нулевая гипотеза $H_0: M(X)=M(Y)$. Конкурирующая гипотеза $H_1: M(X) > M(Y)$.

В этом случае строят правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия T в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости:

$$P(T > t_{\text{правост. кр}}) = \alpha$$

Критическую точку $t_{\text{правост. кр}}(\alpha, k)$ находят по таблице (приложение 6) по уровню значимости α, помещенному в нижней строке таблицы и по числу степеней свободы $k=n+m-2$.

Если $T_{\text{набл}} < t_{\text{правост. кр}}$ — нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост. кр}}$ — нулевую гипотезу отвергают.

Третий случай. Нулевая гипотеза $H_0: M(X)=M(Y)$. Конкурирующая гипотеза $H_1: M(X) < M(Y)$

В этом случае строят левостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы была равна принятому уровню значимости:

$$P(T < t_{\text{левост. кр}}) = \alpha$$

В силу симметрии распределения Стьюдента относительно нуля

$$t_{\text{левост. кр}} = -t_{\text{правост. кр}}$$

Поэтому сначала находят «вспомогательную» критическую точку $t_{\text{правост. кр}}$ так, как описано во втором случае и полагают $t_{\text{левост. кр}} = -t_{\text{правост. кр}}$.

Если $T_{\text{набл}} > -t_{\text{правост. кр}}$ — отвергнуть нулевую гипотезу, нет оснований.

Если $T_{\text{набл}} < t_{\text{правост. кр}}$ = — нулевую гипотезу отвергают.

§ 13. Сравнение выборочной средней и гипотетической генеральной средней нормальной совокупности

А. Дисперсия генеральной совокупности известна. Пусть генеральная совокупность X распределена нормально, причем генеральная средняя a, хотя и неизвестна, но имеются основания предполагать, что она равна гипотетическому (предполагаемому) значению a_0. Например, если X — совокупность размеров x_i партии деталей, изготовляемых станком-автоматом, то можно предположить, что генеральная средняя a этих размеров равна проектному размеру a_0. Чтобы проверить это предположение, находят выборочную среднюю $\overline{x}$ и устанавливают значимо, или незначимо, различаются $\overline{x}$ и a_0. Если различие окажется незначимым, то станок обеспечивает в среднем проектный размер; если различие значимое, то станок требует подналадки.

Предположим, что дисперсия генеральной совокупности известна, например, из предшествующего опыта, или найдена теоретически, или вычислена по выборке большого объема (по большой выборке можно получить достаточно хорошую оценку дисперсии).

Итак, пусть из нормальной генеральной совокупности извлечена выборка объема n и по ней найдена выборочная средняя $\overline{x}$, причем генеральная дисперсия σ^2 известна. Требуется по выборочной средней, при заданном уровне значимости, проверить нулевую гипотезу H_0: $a=a_0$ о равенстве генеральной средней a гипотетическому значению a_0.

Учитывая, что выборочная средняя является несмещенной оценкой генеральной средней (гл. XVI, § 5), т. е. $M(\overline{X})=a$, нулевую гипотезу можно записать так: $M(\overline{X})=a_0$.

Таким образом, требуется проверить, что математическое ожидание выборочной средней равно гипотетической генеральной средней. Другими словами, надо установить значимо, или незначимо, различаются выборочная и генеральная средние.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$U = \frac{\overline{X} - a_0}{\sigma(\overline{X})} = \frac{(\overline{X} - a_0)\sqrt{n}}{\sigma},$$

которая распределена нормально, причем, при справедливости нулевой гипотезы, $M(U)=0$, $\sigma(U)=1$.

Поскольку здесь критическая область строится в зависимости от вида конкурирующей гипотезы так же, как в § 10, ограничимся формулировкой правил проверки нулевой гипотезы, обозначив значение критерия U, вычисленное по данным наблюдений через $U_{\text{набл}}$.

Правило 1. Для того чтобы, при заданном уровне значимости проверить нулевую гипотезу $H_0: a=a_0$ о равенстве генеральной средней a нормальной совокупности с известной дисперсией σ^2 гипотетическому значению a_0, при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{(\overline{x} - a_0)\sqrt{n}}{\sigma}$$

и по таблице функции Лапласа найти критическую точку двусторонней критической области по равенству

$$\Phi(u_{\text{кр}}) = \frac{1-\alpha}{2}.$$

Если $|U_{\text{набл}}| < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|U_{\text{набл}}| > u_{\text{кр}}$ — нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: a > a_0$, критическую точку правосторонней критической области находят по равенству

$$\Phi(u_{\text{кр}}) = \frac{1-2\alpha}{2}.$$

Если $U_{\text{набл}} < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} > u_{\text{кр}}$ — нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a < a_0$ сначала находят критическую точку $u_{\text{кр}}$ по правилу 2, а затем полагают границу левосторонней критической области

$$u'_{\text{кр}} = -u_{\text{кр}}$$

Если $U_{набл} > -u_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} < -u_{кр}$ — нулевую гипотезу отвергают

Пример 1. Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma=0,36$ извлечена выборка объема $n=36$ и по ней найдена выборочная средняя $\overline{x}=21,6$. Требуется, при уровне значимости 0,05, проверить нулевую гипотезу $H_0: a=a_0=21$, при конкурирующей гипотезе $H_1: a \neq 21$

Решение. Найдем наблюдаемое значение критерия

$$U_{набл} = \frac{(\overline{x}-a_0)\sqrt{n}}{\sigma} = \frac{(21,6-21)\sqrt{36}}{0,36} = 10.$$

По условию конкурирующая гипотеза имеет вид $a \neq a$, поэтому критическая область — двусторонняя

Найдем критическую точку по равенству

$$\Phi(u_{кр}) = \frac{1-\alpha}{2} = \frac{1-0,05}{2} = 0,475$$

По таблице функции Лапласа находим $u_{кр}=1,96$.

Так как $U_{набл} > u_{кр}$ — нулевую гипотезу отвергаем. Другими словами, выборочная и гипотетическая генеральная средние различаются значимо.

Пример 2. По данным примера 1 проверить нулевую гипотезу $H_0: a=21$, при конкурирующей гипотезе $a>21$.

Решение. Так как конкурирующая гипотеза имеет вид $a>21$, критическая область — правосторонняя

Найдем критическую точку из равенства

$$\Phi(u_{кр}) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0,05}{2} = 0,45$$

По таблице функции Лапласа находим $u_{кр}=1,65$.

Так как $U_{набл}=10 > u_{кр}$, — нулевую гипотезу отвергаем; различие между выборочной и гипотетической генеральной средней — значимое.

Заметим, что в примере 2 нулевую гипотезу можно было отвергнуть сразу, поскольку она была отвергнута в примере 1, при двусторонней критической области Мы привели полное решение в учебных целях.

Б. Дисперсия генеральной совокупности неизвестна. Если дисперсия генеральной совокупности неизвестна (например, в случае малых выборок), то в качестве критерия проверки нулевой гипотезы принимают случайную величину

$$T = \frac{(\overline{X}-a_0)\sqrt{n}}{s}$$

где s — «исправленное» среднее квадратическое отклонение. Величина T имеет распределение Стьюдента с $k=n-1$ степенями свободы.

Критическая область строится в зависимости от вида конкурирующей гипотезы Поскольку это делается так, как описано выше, ограничимся правилами проверки нулевой гипотезы.

Правило 1. Для того чтобы, при заданном уровне значимости α, проверить нулевую гипотезу $H_0: a=a_0$ о равенстве неизвестной генеральной средней a (нормальной совокупности с неизвестной дисперсией) гипотетическому значению a_0, при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия:

$$T_{набл} = \frac{(\overline{x}-a_0)\sqrt{n}}{s}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α, помещенному в верхней строке таблицы, и числу степеней свободы $k=n-1$, найти критическую точку $t_{двуст. кр}(\alpha, k)$.

Если $|T_{набл}| < t_{двуст. кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст. кр}$ — нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: a>a_0$, по уровню значимости α, помещенному в нижней строке таблицы (приложение 6), и числу степеней свободы $k=n-1$, находят критическую точку $t_{правост. кр}(\alpha, k)$ правосторонней критической области.

Если $T_{набл} < t_{правост. кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} > t_{правост. кр}$ — нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a<a_0$, сначала находят «вспомогательную» критическую точку $t_{правост. кр}(\alpha, k)$ и полагают границу левосторонней критической области $t_{левост. кр} = -t_{правост. кр}$.

Если $T_{набл} > -t_{правост. кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} < -t_{правост. кр}$ — нулевую гипотезу отвергают.

Пример 3. По выборке объема $n=20$ извлеченной из нормальной генеральной совокупности, найдены выборочная средняя $\bar{x}=16$ и «исправленное» среднее квадратическое отклонение $s=4,5$. Требуется при уровне значимости 0,05, проверить нулевую гипотезу $H_0: a=a_0=15$, при конкурирующей гипотезе $H_1: a \neq 15$.

Решение. Вычислим наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{(\bar{x}-a_0)\sqrt{n}}{s} = \frac{(16-15)\cdot\sqrt{20}}{4,5} = 0,99$$

По условию конкурирующая гипотеза имеет вид $a \neq a_0$, поэтому критическая область — двусторонняя.

По таблице критических точек распределения Стьюдента по уровню значимости $\alpha=0,05$, помещенному в верхней строке таблицы, и по числу степеней свободы $k=20-1=19$, находим критическую точку $t_{\text{двуст. кр}}(0,05;\ 19)=2,09$.

Так как $|T_{\text{набл}}| < t_{\text{двуст. кр}}$ — нет оснований, чтобы отвергнуть нулевую гипотезу, выборочная средняя незначимо отличается от гипотетической генеральной средней.

§ 14. Связь между двусторонней критической областью и доверительным интервалом

Легко показать, что отыскивая двустороннюю критическую область при уровне значимости α, тем самым находят и соответствующий доверительный интервал с надежностью $\gamma=1-\alpha$. Например, в § 13, проверяя нулевую гипотезу $H_0: a=a_0$, при $H_1: a \neq a_0$, мы требовали, чтобы вероятность попадания критерия $U=\frac{(\bar{x}-a)\sqrt{n}}{\sigma}$ в двустороннюю критическую область была равна уровню значимости α, следовательно, вероятность попадания критерия в область принятия гипотезы $(-u_{\text{кр}},\ u_{\text{кр}})$ равна $1-\alpha=\gamma$. Другими словами, с надежностью γ выполняется неравенство

$$-u_{\text{кр}} < \frac{(\bar{x}-a)\sqrt{n}}{\sigma} < u_{\text{кр}},$$

или равносильное неравенство

$$\bar{x}-u_{\text{кр}}\frac{\sigma}{\sqrt{n}} < a < \bar{x}+u_{\text{кр}}\frac{\sigma}{\sqrt{n}}, \qquad (*)$$

где $\Phi(u_{\text{кр}})=\frac{\gamma}{2}$.

Мы получили доверительный интервал для оценки математического ожидания a нормального распределения при известном σ, с надежностью γ (гл. XVI, § 15).

Замечание. Хотя отыскание двусторонней критической области и доверительного интервала приводит к одинаковым результатам, — их истолкование различно: двусторонняя критическая область определяет границы (критические точки), между которыми заключено $(1-\alpha)\%$ числа наблюдаемых критериев, найденных при повторении опытов; доверительный же интервал определяет границы (концы интервала), между которыми в $\gamma=(1-\alpha)\%$ опытов заключено истинное значение оцениваемого параметра.

§ 15. Определение минимального объема выборки при сравнении выборочной и гипотетической генеральной средних

На практике часто известна величина (точность) $\delta>0$, которую не должна превышать абсолютная величина разности между выборочной и гипотетической генеральной средними. Например, обычно требуют, чтобы средний размер изготовляемых деталей отличался от проектного не более, чем на заданное δ.

Возникает вопрос: каким должен быть минимальный объем выборки, чтобы это требование с вероятностью $\gamma=1-\alpha$ (α — уровень значимости) выполнялось?

Поскольку задача отыскания доверительного интервала для оценки математического ожидания нормального распределения при известном σ и задача отыскания двусторонней критической области для проверки гипотезы о равенстве математического ожидания (генеральной средней) гипотетическому значению (§ 13, А) сводится одна к другой (§ 14), воспользуемся формулой (гл. XVI, § 15)

$$n=\frac{u_{\text{кр}}^2\sigma^2}{\delta^2},$$

где $u_{\text{кр}}$ находят по равенству $\Phi(u_{\text{кр}})=\frac{\gamma}{2}=\frac{1-\alpha}{2}$

Если же σ неизвестно, а найдена его оценка s, то (§ 13, Б)

$$n=\frac{t_{\text{двуст. кр}}^2(\alpha,\ k)\cdot s^2}{\delta^2}.$$

§ 16. Пример на отыскание мощности критерия

Приведем решение примера на нахождение мощности критерия.

Пример. По выборке объема $n=25$, извлеченной из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma=10$, найдена выборочная средняя $\bar{x}=18$. При уровне значимости 0,05 требуется:

а) найти критическую область, если проверяется нулевая гипотеза $H_0: a=a_0=20$ о равенстве генеральной средней гипотетическому значению, при конкурирующей гипотезе $H_1: a<20$;

б) найти мощность критерия проверки, при $a_0=16$.

Решение. а) Так как конкурирующая гипотеза имеет вид $a<a_0$, критическая область — левосторонняя.

Пользуясь правилом 3 (§ 13, А), найдем критическую точку: $u'_{кр}=-1,65$. Следовательно левосторонняя критическая область определяется неравенством $U<-1,65$, или подробнее

$$\frac{(\bar{x}-20)\sqrt{25}}{10}<-1,65.$$

Отсюда $\bar{x}<16,7$.

При этих значениях выборочной средней нулевая гипотеза отвергается; в этом смысле $\bar{x}=16,7$ можно рассматривать как критическое значение выборочной средней.

б) Для того чтобы вычислить мощность рассматриваемого критерия, предварительно найдем его значение, при условии справедливости конкурирующей гипотезы (т. е при $a_0=16$), положив $\bar{x}=16,7$:

$$U=\frac{(\bar{x}-a_0)\sqrt{n}}{\sigma}=\frac{(16,7-16)\sqrt{25}}{10}=0,35$$

Отсюда видно, что если $\bar{x}<16,7$, то $U<0,35$. Поскольку при $\bar{x}<16,7$ нулевая гипотеза отвергается, то и при $U<0,35$ она также отвергается (при этом конкурирующая гипотеза справедлива, так как мы положили $a_0=16$).

Найдем теперь, пользуясь функцией Лапласа, мощность критерия, т. е. вероятность того, что нулевая гипотеза будет отвергнута, если справедлива конкурирующая гипотеза (§ 7):

$$P(U<0,35)=P(-\infty<U<0,35)=P(-\infty<U<0)+$$
$$+P(0<U<0,35)=0,5+\Phi(0,35)=0,5+0,1368=0,6368.$$

Итак, искомая мощность рассматриваемого критерия приближенно равна 0,64. Если увеличить объем выборки, то мощность увеличится.

Например, при $n=64$ мощность равна 0,71. Если увеличить α, то мощность также увеличится. Например, при $\alpha=0,1$ мощность равна 0,7642.

Замечание. Зная мощность, легко найти вероятность ошибки второго рода: $\beta=1-0,64$. (Разумеется, при решении примера можно было сначала найти β, а затем мощность, равную $1-\beta$.)

§ 17. Сравнение двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки)

В предыдущих параграфах выборки предполагались независимыми. Здесь рассматриваются выборки одинакового объема, варианты которых попарно зависимы. Например, если $x_i (i=1, 2, \ldots, n)$ результаты измерений деталей первым прибором, а y_i— результаты измерений этих же деталей, произведенные в том же порядке вторым прибором, то x_i и y_i попарно зависимы, и в этом смысле сами выборки зависимые. Поскольку, как правило $x_i \neq y_i$, возникает необходимость установить, значимо или незначимо различаются пары этих чисел.

Аналогичная задача ставится при сравнении двух методов исследования, осуществленных одной лабораторией, или если исследование произведено одним и тем же методом двумя различными лабораториями.

Итак, пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. Требуется, при уровне значимости α, проверить нулевую гипотезу $H_0: M(\bar{X})=M(\bar{Y})$ о равенстве генеральных средних нормальных совокупностей с неизвестными дисперсиями, при конкурирующей гипотезе $H_1: M(\bar{X})\neq M(\bar{Y})$, по двум зависимым выборкам одинакового объема.

Сведем эту задачу сравнения двух средних к задаче сравнения одной выборочной средней с гипотетическим значением генеральной средней, решенной в § 13, Б.

С этой целью введем в рассмотрение случайные величины — разности $D_i=X_i-Y_i$ и их среднюю:

$$\overline{D}=\frac{\sum D_i}{n}=\frac{\sum (X_i-Y_i)}{n}=\frac{\sum X_i}{n}-\frac{\sum Y_i}{n}=\overline{X}-\overline{Y}.$$

Если нулевая гипотеза справедлива, т. е. $M(\overline{X})=M(\overline{Y})$, то $M(\overline{X})-M(\overline{Y})=0$ и, следовательно,

$$M(\overline{D})=M(\overline{X}-\overline{Y})=M(\overline{X})-M(\overline{Y})=0.$$

Таким образом, нулевую гипотезу $H_0: M(\overline{X})=M(\overline{Y})$ можно записать так:

$$H_0: M(\overline{D})=0.$$

Тогда конкурирующая гипотеза примет вид:

$$H_1: M(\overline{D})\neq 0.$$

Замечание 1. Далее наблюдаемые неслучайные разности x_i-y_i будем обозначать через d_i, в отличие от случайных разностей $D_i=X_i-Y_i$. Аналогично выборочную среднюю этих разностей $\frac{\sum d_i}{n}$ обозначим через $\overline{d}$, в отличие от случайной величины $\overline{D}$.

Итак, задача сравнения двух средних $\overline{x}$ и $\overline{y}$ сведена к задаче сравнения одной выборочной средней $\overline{d}$ с гипотетическим значением генеральной средней $M(\overline{D})=a_0=0$. Эта задача решена ранее в § 13, Б, поэтому приведем лишь правило проверки нулевой гипотезы и иллюстрирующий пример.

Замечание 2. Как следует из изложенного выше, в формуле (§ 13, Б)

$$T_{\text{набл}}=\frac{(\overline{x}-a_0)\sqrt{n}}{s}$$

надо положить

$$\overline{x}=\overline{d},\quad a_0=0,\quad s=s_d=\sqrt{\frac{\sum d_i^2-\frac{[\sum d_i]^2}{n}}{n-1}}.$$

Тогда $T_{\text{набл}}=\frac{\overline{d}\sqrt{n}}{s_d}$.

Правило. Для того чтобы, при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(\overline{X})=M(\overline{Y})$ о равенстве двух средних нормальных совокупностей с

неизвестными дисперсиями (в случае зависимых выборок одинакового объема), при конкурирующей гипотезе $M(\overline{X})\neq M(\overline{Y})$ надо вычислить наблюдаемое значение критерия

$$T_{\text{набл}}=\frac{\overline{d}\sqrt{n}}{s_d}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α, помещенному в верхней строке таблицы, и по числу степеней свободы $k=n-1$, найти критическую точку $t_{\text{двуст. кр}}(\alpha, k)$.

Если $|T_{\text{набл}}|<t_{\text{двуст. кр}}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|T_{\text{набл}}|>t_{\text{двуст. кр}}$ — нулевую гипотезу отвергают.

Пример. Двумя приборами измерены 5 деталей и получены следующие результаты (в сотых долях *мм*):

$x_1=6,\ x_2=7,\ x_3=8,\ x_4=5,\ x_5=7;$

$y_1=7,\ y_2=6,\ y_3=8,\ y_4=7,\ y_5=8.$

При уровне значимости 0,05 установить, значимо или незначимо различаются результаты измерений.

Решение. Вычитая из чисел первой строки числа второй, получим

$$d_1=-1,\ d_2=1,\ d_3=0,\ d_4=-2,\ d_5=-1.$$

Найдем выборочную среднюю:

$$\overline{d}=\frac{\sum d_i}{n}=\frac{-1+1+0-2+-1}{5}=-0{,}6.$$

Учитывая, что $\sum d_i^2=1+1+4+1=7$ и $\sum d_i=-3$, найдем «исправленное» среднее квадратическое отклонение:

$$s_d=\sqrt{\frac{\sum d_i^2-\frac{[\sum d_i]^2}{n}}{n-1}}=\sqrt{\frac{7-\frac{9}{5}}{5-1}}=\sqrt{1{,}3}.$$

Вычислим наблюдаемое значение критерия

$$T_{\text{набл}}=\frac{\overline{d}\sqrt{n}}{s_d}=-\frac{0{,}6\sqrt{5}}{\sqrt{1{,}3}}=-1{,}18.$$

По таблице критических точек распределения Стьюдента, по уровню значимости 0,05, помещенному в верхней строке таблицы, и числу степеней свободы $k=5-1=4$, находим критическую точку $t_{\text{двуст. кр}}(0{,}05;\ 4)=2{,}78$.

Так как $|T_{набл}| < t_{двуст. кр}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, результаты измерений различаются незначимо.

§ 18. Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события

Пусть по достаточно большому числу n независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, найдена относительная частота $\frac{m}{n}$. Пусть имеются основания предполагать, что неизвестная вероятность равна гипотетическому значению p_0. Требуется, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0.

Поскольку вероятность оценивается по относительной частоте, рассматриваемую задачу можно сформулировать и так: требуется установить значимо или незначимо различаются наблюдаемая относительная частота и гипотетическая вероятность.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}},$$

где $q_0 = 1 - p_0$.

Величина U, при справедливости нулевой гипотезы, распределена приближенно нормально с параметрами $M(U) = 0$, $\sigma(U) = 1$.

Пояснение. Доказано (теорема Лапласа), что при достаточно больших значениях n относительная частота имеет приближенно нормальное распределение с математическим ожиданием p и средним квадратическим отклонением $\frac{\sqrt{pq}}{n}$. Нормируя относительную частоту (вычтя математическое ожидание и разделив на среднее квадратическое отклонение) получим

$$U = \frac{\frac{M}{n} - p}{\sqrt{\frac{pq}{n}}} = \frac{\left(\frac{M}{n} - p\right)\sqrt{n}}{\sqrt{pq}},$$

причем $M(U) = 0$, $\sigma(U) = 1$.

При справедливости нулевой гипотезы, т. е. при $p = p_0$

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}.$$

Замечание 1. Далее наблюдаемая частота обозначается через $\frac{m}{n}$ в отличие от случайной величины $\frac{M}{n}$.

Поскольку здесь критическая область строится так же, как и в § 10, приведем лишь правила проверки нулевой гипотезы и иллюстрирующий пример.

Правило 1. Для того чтобы, при заданном уровне значимости, проверить нулевую гипотезу $H_0: p = p_0$ о равенстве неизвестной вероятности гипотетической вероятности, при конкурирующей гипотезе $H_1: p \neq p_0$, надо вычислить наблюденное значение критерия

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}$$

и по таблице функции Лапласа найти критическую точку по равенству $\Phi(u_{кр}) = \frac{1-\alpha}{2}$.

Если $|U_{набл}| < u_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|U_{набл}| > u_{кр}$ — нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: p > p_0$, находят критическую точку правосторонней критической области из равенства $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$.

Если $U_{набл} < u_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} > u_{кр}$ — нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1: p < p_0$, находят критическую точку $u_{кр}$ по правилу 2, а затем полагают границу левосторонней критической области $u'_{кр} = -u_{кр}$.

Если $U_{набл} > -u_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} < -u_{кр}$ — нулевую гипотезу отвергают.

Замечание 2. Удовлетворительные результаты обеспечивает выполнение неравенства $np_0q_0 > 9$.

Пример. По 100 независимым выборкам найдена относительная частота 0,08. При уровне значимости 0,05, проверить нулевую гипотезу $H_0: p=p_0=0,12$, при конкурирующей гипотезе $H_1: p\neq 0,12$.

Решение. Найдем наблюдаемое значение критерия

$$U_{набл}=\frac{\left(\frac{m}{n}-p_0\right)\sqrt{n}}{\sqrt{p_0q_0}}=\frac{(0,08-0,12)\sqrt{100}}{\sqrt{0,12\cdot 0,88}}=-1,23.$$

По условию конкурирующая гипотеза имеет вид $p\neq p_0$, поэтому критическая область — двусторонняя.

Найдем критическую точку из равенства

$$\Phi(u_{кр})=\frac{1-\alpha}{2}=\frac{1-0,05}{2}=0,475.$$

По таблице функции Лапласа находим $u_{кр}=1,96$.

Так как $|U_{набл}|<u_{кр}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, наблюдаемая относительная частота незначимо отличается от гипотетической вероятности.

§ 19. Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам различного объема. Критерий Бартлета

Пусть генеральные совокупности $X_1, X_2, \ldots, X_l$ распределены нормально. Из этих совокупностей извлечены независимые выборки, вообще говоря, различных объемов $n_1, n_2, \ldots, n_l$ (некоторые объемы могут быть одинаковыми; если все выборки имеют одинаковый объем, то предпочтительнее пользоваться критерием Кочрена, который описан в следующем параграфе). По выборкам найдены исправленные выборочные дисперсии $s_1^2, s_2^2, \ldots, s_l^2$.

Требуется по исправленным выборочным дисперсиям, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0: D(X_1)=D(X_2)=\ldots=D(X_l).$$

Другими словами, требуется установить, значимо или незначимо различаются исправленные выборочные дисперсии.

Рассматриваемую здесь гипотезу о равенстве нескольких дисперсий называют *гипотезой об однородности дисперсий*.

Заметим, что числом степеней свободы дисперсии s_i^2 называют число $k_i=n_i-1$, т. е. число на единицу меньшее объема выборки, по которой вычислена дисперсия.

Обозначим через $\overline{s^2}$ — среднюю арифметическую исправленных дисперсий, взвешенную по числам степеней свободы:

$$\overline{s^2}=\frac{\sum_{i=1}^{l}k_is_i^2}{k},$$

где $k=\sum_{i=1}^{l}k_i$.

В качестве критерия проверки нулевой гипотезы об однородности дисперсий примем критерий Бартлета — случайную величину

$$B=\frac{V}{C},$$

где $V=2,303\left[k=\lg\overline{s^2}-\sum_{i=1}^{l}k_i\lg s_i^2\right]$,

$$C=1+\frac{1}{3(l-1)}\left[\sum_{i=1}^{l}\frac{1}{k_i}-\frac{1}{k}\right].$$

Бартлет установил, что случайная величина B, при условии справедливости нулевой гипотезы, распределена приближенно как χ^2 с $l-1$ степенями свободы, если все $k_i>2$. Учитывая, что $k_i=n_i-1$, заключаем, что $n_i-1>2$, или $n_i>3$, т. е. объем каждой из выборок должен быть не меньше 4.

Критическую область строят правостороннюю, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости:

$$P\left[B>\chi^2_{кр}(\alpha, l-1)\right]=\alpha$$

Критическую точку $\chi^2_{кр}(\alpha,\ l-1)$ находят по таблице (приложение 5) по уровню значимости α и числу степеней свободы $k=l-1$ и тогда правосторонняя критическая область определяется неравенством

$$B > \chi^2_{кр}$$

а область принятия гипотезы — неравенством

$$B < \chi^2_{кр}$$

Обозначим значение критерия Бартлета, вычисленное по данным наблюдений, через $B_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы, при заданном уровне значимости α, проверить нулевую гипотезу об однородности дисперсий нормальных совокупностей, надо вычислить наблюдаемое значение критерия Бартлета и по таблице критических точек распределения χ^2 найти критическую точку $\chi^2_{кр}$ $(\alpha,\ l-1)$.

Если $B_{набл} < \chi^2_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $B_{набл} > \chi^2_{кр}$ — нулевую гипотезу отвергают

Замечание 1. Не следует торопиться вычислять постоянную C. Сначала надо найти V и сравнить с $\chi^2_{кр}$: если окажется что $V < \chi^2_{кр}$, то подавно (так как $C > 1$) $B = \dfrac{V}{C} < \chi^2_{кр}$ и следовательно, C вычислять не нужно.

Если же $V > \chi^2_{кр}$, то надо вычислить C и затем сравнить B с $\chi^2_{кр}$.

Замечание 2 Критерий Бартлета весьма чувствителен к отклонениям распределений от нормального, поэтому к выводам, полученным по этому критерию надо относиться с осторожностью

Пример. По четырем независимым выборкам объемов $n_1=10$, $n_2=12$, $n_3=15$, $n_4=16$, извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии, соответственно равные 0,25; 0,40; 0,36; 0,46. При уровне значимости 0,05. проверить гипотезу об однородности дисперсий (критическая область — правосторонняя).

Решение. Составим расчетную таблицу 25 (столбец 8 пока заполнять не будем, поскольку еще неизвестно понадобится ли вычислять C):

Таблица 25

1	2	3	4	5	6	7	8
i	Объем выборки n_i	Число степеней свободы k_i	Дисперсии s_i^2	$k_i s_i^2$	$\lg s_i^2$	$k_i \lg s_i^2$	$\frac{1}{k_i}$
1	10	9	0,25	2 25	$\bar{1},3979$	$\bar{6},5811$	
2	13	12	0,40	4,80	$\bar{1},6021$	$\bar{5},2252$	
3	15	14	0,36	5.04	$\bar{1},5563$	$\bar{7},7822$	
4	16	15	0,46	6,90	$\bar{1},6628$	$\bar{6},9420$	
Σ		$k=50$		18,99		$\overline{22},5305$	

Пользуясь расчетной таблицей, найдем:

$$\overline{s^2} = \frac{\sum k_i s_i^2}{k} = \frac{18,99}{50} = 0,3798; \quad \lg 0,3798 = \bar{1},5795.$$

$$V = 2,303\left[k \lg \overline{s^2} - \sum k_i \lg s_i^2\right] =$$
$$= 2,303\,[50 \cdot \bar{1},5795 - \overline{22},5305] = 1,02$$

По таблице (приложение 5), по уровню значимости 0,05 и числу степеней свободы $l-1=4-1=3$, находим критическую точку $\chi^2_{кр}$ $(0,05;\ 3)=7,8$

Так как $V < \chi^2_{кр}$, то подавно (поскольку $C > 1$) $B_{набл} = \dfrac{V}{C} < \chi^2_{кр}$ и, следовательно, отвергнуть нулевую гипотезу об однородности дисперсий нет оснований. Другими словами, исправленные выборочные дисперсии различаются незначимо.

Замечание 3. Если требуется оценить генеральную дисперсию, то, при условии однородности дисперсий, целесообразно принять в качестве ее оценки среднюю арифметическую исправленных дисперсий взвешенную по числам степеней свободы т е.

$$\overline{s^2} = \frac{\sum k_i s_i^2}{k}$$

Например, в рассмотренной задаче в качестве оценки генеральной дисперсии целесообразно принять 0,3798

§ 20. Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам одинакового объема. Критерий Кочрена

Пусть генеральные совокупности $X_1, X_2, \ldots, X_l$ распределены нормально. Из этих совокупностей извлечено l выборок одинакового объема n и по ним найдены исправленные выборочные дисперсии $s_1^2, s_2^2, \ldots, s_l^2$, все с одинаковым числом степеней свободы $k=n-1$.

Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0: D(X_1)=D(X_2)=\ldots=D(X_l).$$

Другими словами, требуется проверить, значимо или незначимо различаются исправленные выборочные дисперсии.

В рассматриваемом случае выборок одинакового объема, можно по критерию Фишера—Снедекора (§ 8) сравнить наибольшую и наименьшую дисперсии; если окажется, что различие между ними незначимо, то подавно незначимо и различие между остальными дисперсиями. Недостаток этого метода состоит в том, что информация, которую содержат остальные дисперсии, кроме наименьшей и наибольшей, учтена не будет.

Можно также применить критерий Бартлета Однако, как указано в § 19, известно лишь приближенное распределение этого критерия, поэтому предпочтительнее использовать критерий Кочрена, распределение которого найдено точно.

Итак, в качестве критерия проверки нулевой гипотезы примем критерий Кочрена — отношение максимальной исправленной дисперсии к сумме всех исправленных дисперсий:

$$G=\frac{s^2_{\max}}{s_1^2+s_2^2+\ldots+s_l^2}.$$

Распределение этой случайной величины зависит только от числа степеней свободы $k=n-1$ и количества выборок l.

Критическую область строят правостороннюю, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости

$$P[G>G_{\text{кр}}(\alpha, k, l)]=\alpha.$$

Критическую точку $G_{\text{кр}}(\alpha, k, l)$ находят по таблице * и тогда правосторонняя критическая область определяется неравенством

$$G>G_{\text{кр}},$$

а область принятия нулевой гипотезы — неравенством

$$G<G_{\text{кр}}.$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $G_{\text{набл}}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы, при заданном уровне значимости α, проверить гипотезу об однородности дисперсий нормально распределенных совокупностей, надо вычислить наблюдаемое значение критерия и по таблице найти критическую точку.

Если $G_{\text{набл}}<G_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу.

Если $G_{\text{набл}}>G_{\text{кр}}$—нулевую гипотезу отвергают.

Замечание. Если требуется оценить генеральную дисперсию, то, при условии однородности дисперсий, целесообразно принять в качестве ее оценки среднюю арифметическую исправленных выборочных дисперсий.

Пример. По четырем независимым выборкам одинакового объема $n=17$, извлеченным из нормальных генеральных совокупностей, найдены исправленные дисперсии: 0,26; 0,36; 0,40; 0,42. Требуется: а) при уровне значимости 0,05, проверить нулевую гипотезу об однородности генеральных дисперсий (критическая область — правосторонняя), б) оценить генеральную дисперсию.

Решение. а) Найдем наблюдаемое значение критерия Кочрена — отношение максимальной исправленной дисперсии к сумме всех дисперсий:

$$G_{\text{набл}}=\frac{0,42}{0,26+0,36+0,40+0,42}=0,2917.$$

* Н. В. Смирнов, И. В. Дунин-Барковский. Курс теории вероятностей и математической статистики для технических приложений. Табл. VIII, «Наука», 1965

Найдем по таблице (см. сноску на стр. 331) по уровню значимости 0,05, числу степеней свободы $k=17-1=16$ и числу выборок $l=4$, критическую точку $G_{кр}$ (0,05; 16; 4)$=$ $=0,4366$.

Так как $G_{набл}<G_{кр}$ — нет оснований отвергнуть нулевую гипотезу об однородности дисперсий. Другими словами, исправленные выборочные дисперсии различаются незначимо.

б) Поскольку нулевая гипотеза справедлива, в качестве оценки генеральной дисперсии примем среднюю арифметическую исправленных дисперсий:

$$\sigma^2=\frac{0,26+0,36+0,40+0,42}{4}=0,36.$$

§ 21. Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная генеральная совокупность (X, Y) распределена нормально. Из этой совокупности извлечена выборка объема n и по ней найден выборочный коэффициент корреляции $r_в$, который оказался отличным от нуля. Так как выборка отобрана случайно, то еще нельзя заключить, что коэффициент корреляции генеральной совокупности $r_г$ также отличен от нуля. В конечном счете, нас интересует именно этот коэффициент, поэтому возникает необходимость, при заданном уровне значимости α, проверить нулевую гипотезу $H_0: r_г=0$ о равенстве нулю генерального коэффициента корреляции, при конкурирующей гипотезе $H_1: r_г\neq 0$.

Если нулевая гипотеза будет отвергнута, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля (коротко: значим), а X и Y коррелированы, т. е. связаны линейной зависимостью.

Если нулевая гипотеза будет принята, то выборочный коэффициент корреляции незначим, а X и Y некоррелированы, т. е. не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$T=\frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}}.$$

Величина T, при справедливости нулевой гипотезы, имеет распределение Стьюдента с $k=n-2$ степенями свободы.

Поскольку конкурирующая гипотеза имеет вид $r_г\neq 0$, критическая область — двусторонняя; она строится так же, как в § 12 (первый случай).

Обозначим значение критерия, вычисленное по данным наблюдений, через $T_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы при заданном уровне значимости α, проверить нулевую гипотезу H_0: $r_г=0$ о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины, при конкурирующей гипотезе $H_1: r_г\neq 0$, надо вычислить наблюдаемое значение критерия

$$T_{набл}=\frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}}$$

и по таблице критических точек распределения Стьюдента по заданному уровню значимости и числу степеней свободы $k=n-2$, найти критическую точку $t_{кр}$ (α, k) для двусторонней критической области.

Если $|T_{набл}|<t_{кр}$— нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}|>t_{кр}$ — нулевую гипотезу отвергают.

Пример. По выборке объема $n=122$, извлеченной из нормальной двумерной совокупности (X, Y), найден выборочный коэффициент корреляции $r_в=0,4$. При уровне значимости 0,05, проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: r_г\neq 0$.

Р е ш е н и е. Найдем наблюдаемое значение критерия

$$T_{набл}=\frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}}=\frac{0,4\sqrt{122-2}}{\sqrt{1-0,4^2}}=4,78$$

По условию конкурирующая гипотеза имеет вид $r_г\neq 0$, поэтому критическая область — двусторонняя.

По уровню значимости 0,05 и числу степеней свободы $k=122-2=120$, находим по таблице (приложение 6) для двусторонней критической области критическую точку $t_{кр}$ (0,05; 120)$=1,98$

Поскольку $T_{\text{набл}} > t_{\text{кр}}$ — нулевую гипотезу отвергаем. Другими словами, выборочный коэффициент корреляции значимо отличается от нуля, т. е. X и Y коррелированы.

§ 22. Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

В предыдущих параграфах закон распределения генеральной совокупности предполагался известным.

Если закон распределения неизвестен, но есть основания предположить, что он имеет определенный вид (назовем его A), то проверяют нулевую гипотезу: генеральная совокупность распределена по закону A.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т. е. при помощи специально подобранной случайной величины — критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: χ^2 («хи квадрат») К. Пирсона, Колмогорова, Смирнова и др.

Ограничимся описанием применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений, в этом состоит его достоинство). С этой целью будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

Обычно эмпирические и теоретические частоты различаются. Например (гл. XVII, § 7).

эмп. частоты	6	13	38	74	106	85	30	10	4
теорет. частоты	3	14	42	82	99	76	37	11	2.

Случайно ли расхождение частот? Возможно, что расхождение случайно (незначимо) и объясняется малым числом наблюдений, либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены, исходя из неверной гипотезы о нормальном распределении генеральной совокупности.

Критерий Пирсона отвечает на поставленный выше вопрос. Правда, как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает, на принятом уровне значимости, ее согласие или несогласие с данными наблюдений.

Итак, пусть по выборке объема n получено эмпирическое распределение:

варианты	x_i	x_1	x_2	...	x_s,
эмп. частоты	n_i	n_1	n_2	...	n_s.

Допустим, что в предположении нормального распределения генеральной совокупности, вычислены теоретические частоты n_i' (например так, как в следующем параграфе) При уровне значимости α, требуется проверить нулевую гипотезу; генеральная совокупность распределена нормально.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$\chi^2 = \sum \frac{(n_i - n_i')^2}{n_i'}. \qquad (*)$$

Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения. Ясно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия (*) и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределений.

Заметим, что возведением в квадрат разностей частот устраняют возможность взаимного погашения положительных и отрицательных разностей. Делением на n_i' достигают уменьшения каждого из слагаемых; в противном случае сумма была бы настолько велика, что приводила бы к отклонению нулевой гипотезы даже и тогда, когда она справедлива. Разумеется, приведенные соображения не являются обоснованием выбранного критерия, а лишь пояснением.

Доказано, что при $n \to \infty$ закон распределения случайной величины (*), независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с k степенями свободы. Поэтому случайная величина (*) обозначена через χ^2, а сам критерий называют критерием согласия «хи квадрат».

Число степеней свободы находят по равенству $k=s-1-r$,

где s — число групп (частичных интервалов) выборки;
r — число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение — нормальное, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение) поэтому $r=2$ и число степеней свободы $k=s-1-r=s-1-2=s-3$.

Если, например, предполагают, что генеральная совокупность распределена по закону Пуассона, то оценивают один параметр λ, поэтому $r=1$ и $k=s-2$.

Поскольку односторонний критерий более «жестко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α:

$$P\left[\chi^2 > \chi^2_{кр}(\alpha; k)\right] = \alpha.$$

Таким образом, правосторонняя критическая область определяется неравенством

$$\chi^2 > \chi^2_{кр}(\alpha; k),$$

а область принятия нулевой гипотезы — неравенством

$$\chi^2 < \chi^2_{кр}(\alpha; k).$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы, при заданном уровне значимости, проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия

$$\chi^2_{набл} = \sum \frac{(n_i - n'_i)^2}{n'_i} \qquad (**)$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α, и числу степеней свободы $k=s-3$, найти критическую точку $\chi^2_{кр}(\alpha; k)$.

Если $\chi^2_{набл} < \chi^2_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл} > \chi^2_{кр}$ — нулевую гипотезу отвергают.

Замечание 1. Объем выборки должен быть достаточно велик, во всяком случае не менее 50. Каждая группа должна содержать не менее 5—8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Замечание 2. Поскольку возможны ошибки первого и второго рода, в особенности, если согласование теоретических и эмпирических частот «слишком хорошее», следует проявлять осторожность. Например, можно повторить опыт, увеличить число наблюдений, воспользоваться другими критериями, построить график распределения, вычислить асимметрию и эксцесс (гл. XVII, § 8).

Замечание 3. В целях контроля вычислений, формулу (**) преобразуют к виду

$$\chi^2_{набл} = \sum \frac{n_i^2}{n'_i} - n.$$

Рекомендуем читателю выполнить это преобразование самостоятельно, для чего надо в (**) возвести в квадрат разность частот, сократить результат на n'_i и учесть, что $\sum n_i = n$, $\sum n'_i = n$.

Пример. При уровне значимости 0,05, проверить гипотезу о нормальном распределении генеральной совокупности, если известны эмпирические и теоретические частоты:

эмп. частоты 6 13 38 74 106 85 30 14
теорет. частоты 3 14 42 82 99 76 37 13.

Решение. Вычислим $\chi^2_{набл}$, для чего составим расчетную таблицу 26.

Таблица 26

1	2	3	4	5	6	7	8
i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$	n_i^2	$\frac{n_i^2}{n'_i}$
1	6	3	3	9	3	36	12
2	13	14	—1	1	0,07	169	12,07
3	38	42	—4	16	0,38	1444	34,38
4	74	82	—8	64	0,78	5476	66,78
5	106	99	7	49	0,49	11236	113,49
6	85	76	9	81	1,07	7225	95,07
7	30	37	—7	49	1,32	900	24,32
8	14	13	1	1	0,08	196	15,08
Σ	366	366			$\chi^2_{набл}=7.19$		373,19

Контроль: $\chi^2_{\text{набл}} = 7{,}19$:

$$\sum \frac{n_i^2}{n_i'} - n = 373{,}19 - 366 = 7{,}19.$$

Вычисления произведены правильно.

Найдем число степеней свободы, учитывая, что число групп выборки (число различных вариант) $s=8$: $k=8-3=5$.

По таблице критических точек распределения χ^2 (приложение 5), по уровню значимости $\alpha=0{,}05$ и числу степеней свободы $k=5$, находим $\chi^2_{\text{кр}}(0{,}05;\ 5)=11{,}1$.

Так как $\chi^2_{\text{набл}}<\chi^2_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, расхождение эмпирических и теоретических частот незначимое. Следовательно, данные наблюдений согласуются с гипотезой о нормальном распределении генеральной совокупности.

§ 23. Методика вычисления теоретических частот нормального распределения

Как следует из предыдущего параграфа, сущность критерия согласия Пирсона состоит в сравнении эмпирических и теоретических частот. Ясно, что эмпирические частоты находят из опыта. Как найти теоретические частоты, если предполагается, что генеральная совокупность распределена нормально? Ниже указан один из способов решения этой задачи.

1. Весь интервал наблюдаемых значений X (выборки объема n) делят на s частичных интервалов $(x_i,\ x_{i+1})$ одинаковой длины. Находят середины частичных интервалов $x_i^* = \frac{x_i+x_{i+1}}{2}$; в качестве частоты n_i варианты x_i^* принимают число вариант, которые попали в i-й интервал. В итоге получают последовательность равноотстоящих вариант и соответствующих им частот:

$$\begin{matrix} x_1^* & x_2^* & \ldots & x_s^*, \\ n_1 & n_2 & \ldots & n_s, \end{matrix}$$

причем $\Sigma n_i=n$.

2. Вычисляют, например методом произведений, выборочную среднюю $\overline{x^*}$ и выборочное среднее квадратическое отклонение σ^*.

3. Нормируют случайную величину X, т. е. переходят к величине

$Z = \frac{X-\overline{x^*}}{\sigma^*}$ и вычисляют концы интервалов $(z_i,\ z_{i+1})$:

$$z_i = \frac{x_i-\overline{x^*}}{\sigma^*},\quad z_{i+1} = \frac{x_{i+1}-\overline{x^*}}{\sigma^*},$$

причем наименьшее значение Z, т. е. z_1 полагают равным $-\infty$, а наибольшее, т. е. z_s полагают равным ∞.

4. Вычисляют теоретические вероятности p_i попадания X в интервалы $(x_i,\ x_{i+1})$ по равенству ($\Phi(z)$ — функция Лапласа)

$$P_i = \Phi(z_{i+1}) - \Phi(z_i)$$

и, наконец, находят искомые теоретические частоты $n_i' = np_i$.

Пример. Найти теоретические частоты по заданному интервальному распределению выборки объема $n=200$, предполагая, что генеральная совокупность распределена нормально (табл. 27).

Таблица 27

Номер интервала	Границы интервала		Частота	Номер интервала	Границы интервала		Частота
i	x_i	x_{i+1}	n_i	i	x_i	x_{i+1}	n_i
1	4	6	15	6	14	16	21
2	6	8	26	7	16	18	24
3	8	10	25	8	18	20	20
4	10	12	30	9	20	22	13
5	12	14	26				
							$n=200$

Решение. 1. Найдем середины интервалов $x_i^* = \frac{x_i+x_{i+1}}{2}$. Например, $x_1^*=\frac{4+6}{2}=5$. Поступая аналогично, получим последовательность равноотстоящих вариант x_i^* и соответствующих им частот n_i:

$$\begin{matrix} x_i^* & 5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 & 21 \\ n_i & 15 & 26 & 25 & 30 & 26 & 21 & 24 & 20 & 13 \end{matrix}$$

2. Пользуясь методом произведений, найдем выборочную среднюю и выборочное среднее квадратическое отклонение:

$$\overline{x^*}=12{,}63,\quad \sigma^*=4{,}695$$

3. Найдем интервалы (z_i, z_{i+1}), учитывая, что $\overline{x}^*=12{,}63$ $\sigma^*=4{,}695$, $\frac{1}{\sigma^*}=0{,}213$, для чего составим расчетную таблицу 28.

Таблица 28

i	Границы интервала x_i	x_{i+1}	$x_i-\overline{x}^*$	$x_{i+1}-\overline{x}^*$	Границы интервала $z_i=\frac{x_i-\overline{x}^*}{\sigma^*}$	$z_{i+1}=\frac{x_{i+1}-\overline{x}^*}{\sigma^*}$
1	4	6	—	−6,63	−∞	−1,41
2	6	8	−6,63	−4,63	−1,41	−0,99
3	8	10	−4,63	−2,63	−0,99	−0,56
4	10	12	−2,63	−0,63	−0,156	−0,13
5	12	14	−0,63	1,37	−0,13	0,29
6	14	16	1,37	3,37	0,29	0,72
7	16	18	3,37	5,37	0,72	1,14
8	18	20	5,37	7,37	1,14	1,57
9	20	22	7,37	—	1,57	∞

4. Найдем теоретические вероятности p_i и искомые теоретические частоты $n_i'=np_i$, для чего составим расчетную таблицу 29.

Таблица 29

i	Границы интервала z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	$p_i=\Phi(z_{i+1})-\Phi(z_i)$	$n_i'=np_i=200p_i$
1	−∞	−1,41	−0,5	−0,4207	0,0793	15,86
2	−1,41	−0,99	−0,4207	−0,3389	0,0818	16,36
3	−0,99	−0,56	−0,3389	−0,2123	0,1266	25,32
4	−0,56	−0,13	−0,2123	−0,0517	0,1606	32,16
5	−0,13	0,29	−0,0517	0,1141	0,1658	33,16
6	0,29	0,72	0,1141	0,2642	0,1501	30,02
7	0,72	1,14	0,2642	0,3729	0,1087	21,74
8	1,14	1,57	0,3729	0,4418	0,0689	13,78
9	1,57	∞	0,4418	0,5	0,0582	11,64
					$\sum p_i=1$	$\sum n_i'=200$

Искомые теоретические частоты помещены в последнем столбце таблицы 29.

Задачи

1. По двум независимым выборкам объемов n_1 и n_2, извлеченным из нормальных генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии s_X^2 и s_Y^2. При уровне значимости α проверить нулевую гипотезу H_0: $D(X)=D(Y)$ о равенстве генеральных дисперсий, при конкурирующей гипотезе $H_1: D(X)>D(Y)$, если:

а) $n_1=21$, $n_2=16$, $s_X^2=3{,}6$, $s_Y^2=2{,}4$, $\alpha=0{,}05$;

б) $n_1=13$, $n_2=18$, $s_X^2=0{,}72$, $s_Y^2=0{,}20$, $\alpha=0{,}01$

Отв. а) $F_{\text{набл}}=1{,}5$; $F_{\text{кр}}(0{,}05;\ 20;\ 15)=2{,}33$. Нет оснований отвергнуть нулевую гипотезу; б) $F_{\text{набл}}=3{,}6$; $F_{\text{кр}}(0{,}01;\ 12;\ 17)=3{,}46$. Нулевая гипотеза отвергается.

2. По двум независимым выборкам объемов n и m, извлеченным из нормальных генеральных совокупностей X и Y, найдены выборочные средние $\overline{x}$ и $\overline{y}$. Генеральные дисперсии $D(X)$ и $D(Y)$ известны. При уровне значимости α проверить нулевую гипотезу H_0: $M(X)=M(Y)$ о равенстве математических ожиданий, при конкурирующей гипотезе $H_1: M(X)\neq M(Y)$, если:

а) $n=30$, $m=20$, $D(X)=120$, $D(Y)=100$, $\alpha=0{,}05$;

б) $n=50$, $m=40$, $D(X)=50$, $D(Y)=120$, $\alpha=0{,}01$.

Отв. а) $Z_{\text{набл}}=1$, $z_{\text{кр}}=1{,}96$. Нет оснований отвергнуть нулевую гипотезу; б) $Z_{\text{набл}}=10$, $z_{\text{кр}}=2{,}58$. Нулевая гипотеза отвергается.

3. По двум независимым выборкам объемов $n=5$ и $m=6$, извлеченным из нормальных генеральных совокупностей X и Y, найдены выборочные средние $\overline{x}=15{,}9$, $\overline{y}=14{,}1$ и исправленные выборочные дисперсии $s_X^2=14{,}76$, $s_Y^2=4{,}92$. При уровне значимости 0,05 проверить нулевую гипотезу H_0: $M(X)=M(Y)$ о равенстве математических ожиданий при конкурирующей гипотезе $H_1: M(X)\neq M(Y)$.

Указание. Предварительно сравнить дисперсии.

Отв. $T_{\text{набл}}=0{,}88$, $t_{\text{кр}}(0{,}05;\ 9)=2{,}26$. Нет оснований отвергнуть нулевую гипотезу.

4. Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma=2{,}1$ извлечена выборка объема $n=49$ и по ней найдена выборочная средняя $\overline{x}=4{,}5$. Требуется, при уровне значимости 0,05 проверить нулевую гипотезу

$H_0: a = 3$ о равенстве математического ожидания гипотетическому значению, при конкурирующей гипотезе $H_1: a \neq 3$.

Отв. $U_{набл} = 5$, $u_{кр} = 1{,}96$. Нулевая гипотеза отвергается

5. По выборке объема $n = 16$, извлеченной из нормальной генеральной совокупности, найдены выборочная средняя $\overline{x} = 12{,}4$ и «исправленное» среднее квадратическое отклонение $s = 1{,}2$. Требуется, при уровне значимости 0,05, проверить нулевую гипотезу $H_0: a = 11{,}8$ о равенстве математического ожидания гипотетическому значению, при конкурирующей гипотезе $H_1: a \neq 11{,}8$.

Отв. $T_{набл} = 2$, $t_{кр}$ (0,05; 15) = 2,13. Нет оснований отвергнуть нулевую гипотезу.

6. Двумя приборами измерены 5 деталей. Получены следующие результаты (в *мм*):

$$x_1 = 4,\ x_2 = 5,\ x_3 = 6,\ x_4 = 7,\ x_5 = 8;$$
$$y_1 = 5,\ y_2 = 5,\ y_3 = 9,\ y_4 = 4,\ y_5 = 6.$$

При уровне значимости 0,05, проверить, значимо или незначимо различаются результаты измерений.

Отв. $T_{набл} = 10{,}54$, $t_{кр}$ (0,05; 4) = 2,78. Различие результатов измерений значимое.

7. По 100 независимым испытаниям найдена относительная частота $\frac{m}{n} = 0{,}15$. При уровне значимости 0,05 проверить нулевую гипотезу $H_0: p = 0{,}17$ о равенстве относительной частоты гипотетической вероятности, при конкурирующей гипотезе $H_1: p \neq 0{,}17$.

Отв. $|U_{набл}| = 0{,}53$, $u_{кр} = 1{,}96$. Нет оснований отвергнуть нулевую гипотезу.

8. По пяти независимым выборкам объемов $n_1 = 7$, $n_2 = 9$, $n_3 = 10$, $n_4 = 12$, $n_5 = 12$, извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии: 0,27; 0,32; 0,40; 0,42; 0,48. При уровне значимости 0,05, проверить нулевую гипотезу об однородности дисперсий (критическая область — правосторонняя).

Указание. Воспользоваться критерием Бартлета (§ 19).

Отв. $V = 6{,}63$, $\chi^2_{кр}$ (0,05; 4) = 9,5. Нет оснований отвергнуть нулевую гипотезу.

9. По четырем независимым выборкам одинакового объема $n = 17$, извлеченным из нормальных совокупностей, найдены исправленные выборочные дисперсии: 2,12; 2,32; 3,24; 4,32. Требуется: а) при уровне значимости 0,05, проверить нулевую гипотезу о равенстве генеральных дисперсий (критическая область — правосторонняя), б) оценить генеральную дисперсию.

Указание. Воспользоваться критерием Кочрена (§ 20).

Отв. а) $G_{набл} = 0{,}36$, $G_{кр}$(0,05; 16,4) = 0,4366. Нет оснований отвергнуть нулевую гипотезу; б) $\sigma = 3$

10. По выборке объема $n = 62$, извлеченной из двумерной нормальной совокупности (X, Y), найден выборочный коэффициент корреляции $r_в = 0{,}6$. При уровне значимости 0,05, проверить нулевую гипотезу $H_0: r_г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $r_г \neq 0$.

Отв. $T_{набл} = 5{,}81$, $t_{кр}$(0,05; 60) = 2,0. Нулевая гипотеза отвергается

11. При уровне значимости 0,05, проверить гипотезу о нормальном распределении генеральной совокупности, если известны эмпирические и теоретические частоты:

а)	эмпир. частоты	6	12	16	40	13	8	5			
	теорет. частоты	4	11	15	43	15	6	6;			
б)	эмпир. частоты	5	6	14	32	43	39	30	20	6	5
	теорет. частоты	4	7	12	29	48	35	34	18	7	6.
в)	эмпир. частоты	5	13	12	44	8	12	6			
	теорет. частоты	2	20	12	35	15	10	6			

Отв. а) $\chi^2_{набл} = 2{,}5$, $\chi^2_{кр}$ (0,05; 4) = 9,5. Нет оснований отвергнуть гипотезу; б) $\chi^2_{набл} = 3$, $\chi^2_{кр}$(0,05; 7) = 14,1. Нет оснований отвергнуть гипотезу; в) $\chi^2_{набл} = 13$, $\chi^2_{кр}$ (0,05; 4) = 9,5. Гипотеза отвергается

Глава двадцатая

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

§ 1. Сравнение нескольких средних. Понятие о дисперсионном анализе

Пусть генеральные совокупности $X_1, X_2, \ldots X_p$ распределены нормально и имеют одинаковую, хотя и неизвестную, дисперсию; математические ожидания также неизвестны, но могут быть различными. Требуется при заданном уровне значимости, по выборочным средним проверить нулевую гипотезу

$$H_0 : M(X_1)=M(X_2)=\ldots=M(X_p)$$

о равенстве всех математических ожиданий. Другими словами, требуется установить, значимо или незначимо различаются выборочные средние. Казалось бы, для сравнения нескольких средних ($p>2$) можно сравнить их попарно. Однако, с возрастанием числа средних, возрастает и наибольшее различие между ними: среднее новой выборки может оказаться больше наибольшего или меньше наименьшего из средних, полученных до нового опыта. По этой причине для сравнения нескольких средних пользуются другим методом, который основан на сравнении дисперсий и поэтому назван *дисперсионным анализом* (в основном развит английским статистиком Р. Фишером).

На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет p уровней $F_1, F_2, \ldots, F_p$ на изучаемую величину X. Например, если требуется выяснить, какой вид удобрений наиболее эффективен для получения наибольшего урожая, то фактор F — удобрение, а его уровни — виды удобрений.

Основная идея дисперсионного анализа состоит в сравнении «факторной дисперсии», порождаемой воздействием фактора и «остаточной дисперсии», обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на X; в этом случае средние наблюдаемых значений на каждом уровне (групповые средние) будут различаться также значимо.

Если уже установлено, что фактор существенно влияет на X, а требуется выяснить, какой из уровней оказывает наибольшее воздействие, то дополнительно производят попарное сравнение средних.

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей (дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны). Однородные же совокупности можно объединить в одну и тем самым получить о ней более полную информацию, а следовательно и более надежные выводы.

В более сложных случаях исследуют воздействие нескольких факторов на нескольких постоянных или случайных уровнях и выясняют влияние отдельных уровней и их комбинаций (*многофакторный анализ*).

Мы ограничимся простейшим случаем однофакторного анализа, когда на X воздействует только один фактор, который имеет p постоянных уровней.

§ 2. Общая, факторная и остаточная суммы квадратов отклонений

Пусть на количественный нормально распределенный признак X воздействует фактор F, который имеет p постоянных уровней. Будем предполагать, что число наблюдений на каждом уровне одинаково и равно q.

Пусть наблюдалось pq значений x_{ij} признака X, где i — номер испытания ($i=1, 2, \ldots, q$), j — номер уровня фактора ($j=1, 2, \ldots, p$). Результаты наблюдений представлены в таблице 30.

Таблица 30

Номер испытания	Уровни фактора F_j			
	F_1	F_2	...	F_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
q	x_{q1}	x_{q2}	...	x_{qp}
Групповая средняя	$\overline{x}_{\text{гр}1}$	$\overline{x}_{\text{гр}2}$	...	$\overline{x}_{\text{гр}\,p}$

Введем по определению:

$$S_{\text{общ}} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \overline{x})^2$$

(*общая сумма* квадратов отклонений наблюдаемых значений от общей средней $\overline{x}$),

$$S_{\text{факт}} = q \sum_{j=1}^{p} (\overline{x}_{\text{гр}\,j} - \overline{x})^2$$

(*факторная сумма* квадратов отклонений групповых средний от общей средней, которая характеризует рассеяние «между группами»)

$$S_{ост} = \sum_{i=1}^{q} (x_{i1} - \bar{x}_{гр\,1})^2 + \sum_{i=1}^{q} (x_{i2} - \bar{x}_{гр\,2})^2 + \ldots +$$

$$+ \sum_{i=1}^{q} (x_{ip} - \bar{x}_{гр\,p})^2$$

(*остаточная сумма* квадратов отклонений наблюдаемых значений группы от своей групповой средней, которая характеризует рассеяние «внутри групп»).

Практически остаточную сумму находят по равенству (§ 3, следствие):

$$S_{ост} = S_{общ} - S_{факт}.$$

Элементарными преобразованиями можно получить формулы, более удобные для расчетов:

$$S_{общ} = \sum_{j=1}^{p} P_j - \frac{\left[\sum_{j=1}^{p} R_j\right]^2}{pq}, \qquad (*)$$

$$S_{факт} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left[\sum_{j=1}^{p} R_j\right]^2}{pq}, \qquad (**)$$

где $P_j = \sum_{i=1}^{q} x_{ij}^2$ — сумма квадратов значений признака на уровне F_j;

$R_j = \sum_{i=1}^{q} x_{ij}$ — сумма значений признака на уровне F_j.

Замечание. Для упрощения вычислений вычитают из каждого наблюдаемого значения одно и то же число C, примерно равное общей средней. Если уменьшенные значения $y_{ij} = x_{ij} - C$, то

$$S_{общ} = \sum_{j=1}^{p} Q_j - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq}, \qquad (***)$$

$$S_{факт} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq}. \qquad (****)$$

где $Q_j = \sum_{i=1}^{q} y_{ij}^2$ — сумма квадратов уменьшенных значений признака на уровне F_j;

$T_j = \sum_{i=1}^{q} y_{ij}$ — сумма уменьшенных значений признака на уровне F_j.

Для вывода формул (***) и (****) достаточно подставить $x_{ij} = y_{ij} + C$ в соотношение (*) и $R_j = \sum_{i=1}^{q} x_{ij} = \sum_{i=1}^{q} (y_{ij} + C) = \sum_{i=1}^{q} y_{ij} + qC = T_j + qC$ (в соотношение (**)).

Пояснения.

1. Убедимся, что $S_{факт}$ характеризует воздействие фактора F. Допустим, что фактор оказывает существенное влияние на X. Тогда группа наблюдаемых значений признака на одном определенном уровне, будет, вообще говоря, отличаться от групп наблюдений на других уровнях. Следовательно, будут различаться и групповые средние, причем они тем больше рассеяны вокруг общей средней, чем большим окажется воздействие фактора. Отсюда следует, что для оценки воздействия фактора целесообразно составить сумму квадратов отклонений групповых средних от общей средней (отклонение возводят в квадрат, чтобы исключить погашение положительных и отрицательных отклонений). Умножив эту сумму на q, получим $S_{факт}$. Итак, $S_{факт}$ характеризует воздействие фактора.

2. Убедимся, что $S_{ост}$ отражает влияние случайных причин. Казалось бы наблюдения одной группы не должны различаться. Однако, поскольку на X, кроме фактора F, воздействуют и случайные причины,— наблюдения одной и той же группы, вообще говоря, различны и, значит, рассеяны вокруг своей групповой средней. Отсюда следует, что для оценки влияния случайных причин целесообразно составить сумму квадратов отклонений наблюдаемых значений каждой группы от своей групповой средней, т. е. $S_{ост}$. Итак, $S_{ост}$ характеризует воздействие случайных причин.

3. Убедимся, что $S_{общ}$ отражает влияние и фактора и случайных причин. Будем рассматривать все наблюдения как единую совокупность. Наблюдаемые значения признака различны вследствие воздействия фактора и случайных причин. Для оценки этого воздействия целесообразно составить сумму квадратов отклонений наблюдаемых значений от общей средней, т. е. $S_{общ}$.

Итак, $S_{общ}$ характеризует влияние фактора и случайных причин.

Приведем пример, который наглядно показывает, что факторная сумма отражает влияние фактора, а остаточная — влияние случайных причин.

Пример. Двумя приборами произведены по 2 измерения физической величины, истинный размер которой равен x. Рассматривая в качестве фактора систематическую ошибку C, а в качестве его уровней—систематические ошибки C_1 и C_2 соответственно первого и второго прибора, показать, что $S_{факт}$ определяется систематическими, а $S_{ост}$ — случайными ошибками измерений.

Р е ш е н и е. Введем обозначения:

α_1, α_2— случайные ошибки первого и второго измерений первым прибором,

β_1, β_2— случайные ошибки первого и второго измерений вторым прибором.

Тогда наблюдаемые значения результатов измерений соответственно равны (первый индекс при x указывает номер измерения, а второй — номер прибора):

$$x_{11}=x+C_1+\alpha_1,\quad x_{21}=x+C_1+\alpha_2;$$
$$x_{12}=x+C_2+\beta_1,\quad x_{22}=x+C_2+\beta_2.$$

Средние значения измерений первым и вторым приборами соответственно равны:

$$\bar{x}_{гр\,1}=x+C_1+\frac{\alpha_1+\alpha_2}{2}=x+C_1+\alpha,$$

$$\bar{x}_{гр\,2}=x+C_2+\frac{\beta_1+\beta_2}{2}=x+C_2+\beta.$$

Общая средняя

$$\bar{x}=\frac{\bar{x}_{гр\,1}+\bar{x}_{гр\,2}}{2}=x+\frac{C_1+C_2}{2}+\frac{\alpha+\beta}{2},$$

факторная сумма

$$S_{факт}=(\bar{x}_{гр\,1}-\bar{x})^2+(\bar{x}_{гр\,2}-\bar{x})^2.$$

Подставив величины, заключенные в скобках, после элементарных преобразований получим

$$S_{факт}=\frac{(C_1-C_2)^2}{2}+(C_1-C_2)(\alpha-\beta)+\frac{(\alpha-\beta)^2}{2}.$$

Мы видим, что $S_{факт}$ определяется, главным образом, первым слагаемым (поскольку случайные ошибки измерений малы) и, следовательно, действительно отражает влияние фактора C.

Остаточная сумма

$$S_{ост}=(x_{11}-\bar{x}_{гр\,1})^2+(x_{21}-\bar{x}_{гр\,1})^2+(x_{12}-\bar{x}_{гр\,2})^2+$$
$$+(x_{22}-\bar{x}_{гр\,2})^2.$$

Подставив величины, заключенные в скобках, получим

$$S_{ост}=[(\alpha_1-\alpha)^2+(\alpha_2-\alpha)^2]+[(\beta_1-\beta)^2+(\beta_2-\beta)^2].$$

Мы видим, что $S_{ост}$ определяется случайными ошибками измерений и, следовательно, действительно отражает влияние случайных причин.

З а м е ч а н и е. То, что $S_{ост}$ порождается случайными причинами, следует также из равенства (§ 3, следствие)

$$S_{ост}=S_{общ}-S_{факт}.$$

Действительно $S_{общ}$ является результатом воздействия фактора и случайных причин; вычитая $S_{факт}$, мы исключаем влияние фактора. Следовательно, «оставшаяся часть» отражает влияние случайных причин

§ 3. Связь между общей, факторной и остаточной суммами

Покажем, что

$$S_{общ}=S_{факт}+S_{ост}.$$

Для упрощения вывода ограничимся двумя уровнями ($p=2$) и двумя испытаниями на каждом уровне ($q=2$). Результаты испытаний представим в виде таблицы 31.

Т а б л и ц а 31

Номер испытания i	Уровни фактора F_j	
	F_1	F_2
1	x_{11}	x_{12}
2	x_{21}	x_{22}
$\bar{x}_{гр\,j}$	$\bar{x}_{гр\,1}$	$\bar{x}_{гр\,2}$

Тогда

$$S_{общ}=(x_{11}-\bar{x})^2+(x_{21}-\bar{x})^2+(x_{12}-\bar{x})^2+(x_{22}-\bar{x})^2.$$

Вычтем и прибавим к каждому наблюдаемому значению на первом уровне групповую среднюю $\overline{x}_{гр1}$, а на втором — $\overline{x}_{гр2}$.

Выполнив возведение в квадрат и учитывая, что сумма всех удвоенных произведений равна нулю (рекомендуем читателю убедиться в этом самостоятельно), получим

$$S_{общ}=2\left[(\overline{x}_{гр1}-\overline{x})^2+(\overline{x}_{гр2}-\overline{x})^2\right]+\left[(x_{11}-\overline{x}_{гр1})^2+\right.$$
$$\left.+(x_{21}-\overline{x}_{гр1})^2+(x_{12}-\overline{x}_{гр2})^2+(x_{22}-\overline{x}_{гр2})^2\right]=$$
$$=S_{факт}+S_{ост}$$

Итак

$$S_{общ}=S_{факт}+S_{ост}$$

Следствие. Из полученного равенства вытекает важное следствие:

$$S_{ост}=S_{общ}-S_{факт}$$

Отсюда видно, что нет надобности непосредственно вычислять остаточную сумму: достаточно найти общую и факторную суммы, а затем их разность

§ 4. Общая, факторная и остаточная дисперсии

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s^2_{общ}=\frac{S_{общ}}{pq-1},\quad s^2_{факт}=\frac{S_{факт}}{p-1},\quad s^2_{ост}=\frac{S_{ост}}{p(q-1)}$$

где p — число уровней фактора,

q — число наблюдений на каждом уровне.

Если нулевая гипотеза о равенстве средних справедлива, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Например, учитывая что объем выборки $n=pq$, заключаем, что

$$s^2_{общ}=\frac{S_{общ}}{pq-1}=\frac{S_{общ}}{n-1}$$ — исправленная выборочная дисперсия, которая известно, является несмещенной оценкой генеральной дисперсии.

Замечание. Число степеней свободы $p(q-1)$ остаточной дисперсии равно разности между числами степеней свободы общей и факторной дисперсий. Действительно,

$$(pq-1)-(p-1)=pq-p=p(q-1).$$

§ 5. Сравнение нескольких средних методом дисперсионного анализа

Вернемся к задаче, поставленной в § 1: проверить, при заданном уровне значимости, нулевую гипотезу о равенстве нескольких ($p>2$) средних нормальных совокупностей с неизвестными, но одинаковыми дисперсиями. Покажем, что решение этой задачи сводится к сравнению факторной и остаточной дисперсий по критерию Фишера—Снедекора (гл. XIX, § 8).

1. Пусть нулевая гипотеза о равенстве нескольких средних (далее будем называть их групповыми) правильна. В этом случае факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии (§ 4) и, следовательно, различаются незначимо. Если сравнить эти оценки по критерию F, то, очевидно, критерий укажет, что нулевую гипотезу о равенстве факторной и остаточной дисперсий следует принять.

Таким образом, если гипотеза о равенстве групповых средних правильна, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2. Пусть нулевая гипотеза о равенстве групповых средних ложна. В этом случае с возрастанием расхождения между групповыми средними будет увеличиваться факторная дисперсия, а вместе с ней и отношение $F_{набл}=\frac{s^2_{факт}}{s^2_{ост}}$. В итоге $F_{набл}$ окажется больше $F_{кр}$ и, следовательно, гипотеза о равенстве дисперсий будет отвергнута.

Таким образом, если гипотеза о равенстве групповых средних ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Легко доказать от противного справедливость обратных утверждений: из правильности (ложности) гипотезы о дисперсиях следует правильность (ложность) гипотезы о средних.

Итак, *для того чтобы проверить нулевую гипотезу о равенстве групповых средних нормальных совокупностей с одинаковыми дисперсиями, достаточно проверить по критерию F нулевую гипотезу о равенстве факторной и* *остаточной дисперсий.* В этом и состоит метод дисперсионного анализа.

Замечание 1. Если факторная дисперсия окажется меньше остаточной, то уже отсюда следует справедливость гипотезы о равенстве групповых средних и, значит, нет надобности прибегать к критерию F.

Замечание 2. Если нет уверенности в справедливости предположения о равенстве дисперсий рассматриваемых p совокупностей, то это предположение следует проверить предварительно, например по критерию Кочрена.

Пример. Произведено по 4 испытания на каждом из трех уровней. Результаты испытаний приведены в таблице 32. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Таблица 32

Номер испытания i	Уровни фактора F_j		
	F_1	F_2	F_3
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52
$\bar{x}_{гр\,j}$	54	55	47

Решение. Для упрощения расчета вычтем $C=52$ из каждого наблюдаемого значения: $y_{ij}=x_{ij}-52$. Составим расчетную таблицу 33.

Пользуясь таблицей и учитывая, что число уровней фактора $p=3$, число испытаний на каждом уровне $q=4$, найдем общую и факторную суммы квадратов отклонений [§ 2, формулы (***) и (****)]:

$$S_{общ}=\sum_{j=1}^{p} S_j-\frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq}=266-0=266;$$

$$S_{факт}=\frac{\sum_{j=1}^{p} T_j^2}{q}-\frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq}=\frac{608}{4}-0=152.$$

Таблица 33

Номер испытания i	Уровни фактора F_j						
	F_1		F_2		F_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	−1	1	0	0	−10	100	
2	0	0	2	4	−8	64	
3	4	16	4	16	−2	4	
4	5	25	6	36	0	0	
$S_j=\sum_{i=1}^{4} y_{ij}^2$		42		56		168	$\sum S_j=266$
T_j	8		12		−20		$\sum T_j=0$
T_j^2	64		144		400		$\sum T_j^2=608$

Найдем остаточную сумму квадратов отклонений:

$$S_{ост}=S_{общ}-S_{факт}=266-152=114.$$

Найдем факторную и остаточную дисперсии:

$$s_{факт}^2=\frac{S_{факт}}{p-1}=\frac{152}{3-1}=76;$$

$$s_{ост}^2=\frac{S_{ост}}{p(q-1)}=\frac{114}{3(4-1)}=\frac{114}{9}=12,67.$$

Сравним факторную и остаточную дисперсии по критерию F (гл. XIX, § 8), для чего найдем наблюдаемое значение критерия:

$$F_{набл}=\frac{s_{факт}^2}{s_{ост}^2}=\frac{76}{12,67}=6$$

Учитывая, что число степеней свободы числителя $k_1=2$, а знаменателя $k_2=9$ и уровень значимости $\alpha=0,05$, по таблице находим критическую точку $F_{кр}(0,05;\ 2;\ 9)=4,26$.

Так как $F_{набл}>F_{кр}$, нулевую гипотезу о равенстве групповых средних отвергаем. Другими словами, групповые средние «в целом» различаются значимо. Если требуется сравнить средние попарно, то следует воспользоваться критерием Стьюдента.

Замечание 3. Если наблюдаемые значения x_{ij} — десятичные дроби с одним знаком после запятой, то целесообразно перейти к числам $y_{ij} = 10x_{ij} - C$, где C — примерно среднее значение чисел $10x_{ij}$. В итоге получим сравнительно небольшие **целые** числа. Хотя при этом факторная и остаточная дисперсия увеличиваются в 10^2 раз — их отношение не изменится. Например, если

$x_{11} = 12{,}1$, $x_{21} = 12{,}2$, $x_{31} = 12{,}6$, то приняв $y_{ij} = 10x_{ij} - 123$, получим:

$y_{11} = 121 - 123 = -2$, $y_{21} = 122 - 123 = -1$, $y_{31} = 126 - 123 = 3$.

Аналогично поступают, если после запятой имеется k знаков:

$$y_{ij} = 10^k \cdot x_{ij} - C.$$

Задачи

В задачах 1—2 требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми генеральными дисперсиями.

1.

Номер испытания i	Уровни фактора F_j				
	F_1	F_2	F_3	F_4	F_5
1	42	66	35	64	70
2	55	91	50	70	79
3	67	96	60	79	88
4	67	98	69	81	90
$\overline{x}_{\text{гр}\,j}$	57,75	87,75	53,50	73,50	81,75

Отв. $F_{\text{набл}} = 6{,}13$; $F_{\text{кр}}(0{,}05;\ 4;\ 15) = 3{,}06$. Нулевая гипотеза отвергается.

2.

Номер испытания i	Уровни фактора F_j			
	F_1	F_2	F_3	F_4
1	6	6	9	7
2	7	7	12	9
3	8	11	13	10
4	11	12	14	10
$\overline{x}_{\text{гр}\,j}$	8	9	12	9

Отв. $F_{\text{набл}} = 2{,}4$; $F_{\text{кр}}(0{,}05;\ 3;\ 12) = 3{,}49$. Нет оснований отвергнуть нулевую гипотезу

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2913	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0149
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0012	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Приложение 2

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,45	0,1736	0,90	0,3159	1,35	0,4115
0,01	0,0040	0,46	0,1772	0,91	0,3186	1,36	0,4131
0,02	0,0080	0,47	0,1808	0,92	0,3212	1,37	0,4147
0,03	0,0120	0,48	0,1844	0,93	0,3238	1,38	0,4162
0,04	0,0160	0,49	0,1879	0,94	0,3264	1,39	0,4177
0,05	0,0199	0,50	0,1915	0,95	0,3289	1,40	0,4192
0,06	0,0239	0,51	0,1950	0,96	0,3315	1,41	0,4207
0,07	0,0279	0,52	0,1985	0,97	0,3340	1,42	0,4222
0,08	0,0319	0,53	0,2019	0,98	0,3365	1,43	0,4236
0,09	0,0359	0,54	0,2054	0,99	0,3389	1,44	0,4251
0,10	0,0398	0,55	0,2088	1,00	0,3413	1,45	0,4265
0,11	0,0438	0,56	0,2123	1,01	0,3438	1,46	0,4279
0,12	0,0478	0,57	0,2157	1,02	0,3461	1,47	0,4292
0,13	0,0517	0,58	0,2190	1,03	0,3485	1,48	0,4306
0,14	0,0557	0,59	0,2224	1,04	0,3508	1,49	0,4319
0,15	0,0596	0,60	0,2257	1,05	0,3531	1,50	0,4332
0,16	0,0636	0,61	0,2291	1,06	0,3554	1,51	0,4345
0,17	0,0675	0,62	0,2324	1,07	0,3577	1,52	0,4357
0,18	0,0714	0,63	0,2357	1,08	0,3599	1,53	0,4370
0,19	0,0753	0,64	0,2389	1,09	0,3621	1,54	0,4382
0,20	0,0793	0,65	0,2422	1,10	0,3643	1,55	0,4394
0,21	0,0832	0,66	0,2454	1,11	0,3665	1,56	0,4406
0,22	0,0871	0,67	0,2486	1,12	0,3686	1,57	0,4418
0,23	0,0910	0,68	0,2517	1,13	0,3708	1,58	0,4429
0,24	0,0948	0,69	0,2549	1,14	0,3729	1,59	0,4441
0,25	0,0987	0,70	0,2580	1,15	0,3749	1,60	0,4452
0,26	0,1026	0,71	0,2611	1,16	0,3770	1,61	0,4463
0,27	0,1064	0,72	0,2642	1,17	0,3790	1,62	0,4474
0,28	0,1103	0,73	0,2673	1,18	0,3810	1,63	0,4484
0,29	0,1141	0,74	0,2703	1,19	0,3830	1,64	0,4495
0,30	0,1179	0,75	0,2734	1,20	0,3849	1,65	0,4505
0,31	0,1217	0,76	0,2764	1,21	0,3869	1,66	0,4515
0,32	0,1255	0,77	0,2794	1,22	0,3883	1,67	0,4525
0,33	0,1293	0,78	0,2823	1,23	0,3907	1,68	0,4535
0,34	0,1331	0,79	0,2852	1,24	0,3925	1,69	0,4545
0,35	0,1368	0,80	0,2881	1,25	0,3944	1,70	0,4554
0,36	0,1406	0,81	0,2910	1,26	0,3962	1,71	0,4564
0,37	0,1443	0,82	0,2939	1,27	0,3980	1,72	0,4573
0,38	0,1480	0,83	0,2967	1,28	0,3997	1,73	0,4582
0,39	0,1517	0,84	0,2995	1,29	0,4015	1,74	0,4591
0,40	0,1554	0,85	0,3023	1,30	0,4032	1,75	0,4599
0,41	0,1591	0,86	0,3051	1,31	0,4049	1,76	0,4608
0,42	0,1628	0,87	0,3078	1,32	0,4066	1,77	0,4616
0,43	0,1664	0,88	0,3106	1,33	0,4082	1,78	0,4625
0,44	0,1700	0,89	0,3133	1,34	0,4099	1,79	0,4633

Продолжение

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,80	0,4641	2,00	0,4772	2,40	0,4918	2,80	0,4974
1,81	0,4649	2,02	0,4783	2,42	0,4922	2,82	0,4976
1,82	0,4656	2,04	0,4793	2,44	0,4927	2,84	0,4977
1,83	0,4664	2,06	0,4803	2,46	0,4931	2,86	0,4979
1,84	0,4671	2,08	0,4812	2,48	0,4934	2,88	0,4980
1,85	0,4678	2,10	0,4821	2,50	0,4938	2,90	0,4981
1,86	0,4686	2,12	0,4830	2,52	0,4941	2,92	0,4982
1,87	0,4693	2,14	0,4838	2,54	0,4945	2,94	0,4984
1,88	0,4699	2,16	0,4846	2,56	0,4948	2,96	0,4985
1,89	0,4706	2,18	0,4854	2,58	0,4951	2,98	0,4986
1,90	0,4713	2,20	0,4861	2,60	0,4953	3,00	0,49865
1,91	0,4719	2,22	0,4868	2,62	0,4956	3,20	0,49931
1,92	0,4726	2,24	0,4875	2,64	0,4959	3,40	0,49966
1,93	0,4732	2,26	0,4881	2,66	0,4961	3,60	0,499841
1,94	0,4738	2,28	0,4887	2,68	0,4963	3,80	0,499928
1,95	0,4744	2,30	0,4893	2,70	0,4965	4,00	0,499968
1,96	0,4750	2,32	0,4898	2,72	0,4967	4,50	0,499997
1,97	0,4756	2,34	0,4904	2,74	0,4969	5,00	0,499997
1,98	0,4761	2,36	0,4909	2,76	0,4971		
1,99	0,4767	2,38	0,4913	2,78	0,4973		

Приложение 3

Таблица значений $t_\gamma = t(\gamma, n)$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,729	3,600
9	2,31	2,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,001	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Приложение 4

Таблица значений $q = q(\gamma, n)$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение 5

Критические точки распределения χ^2

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63

Продолжение

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 6

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79

Продолжение

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 7

Критические точки распределения F Фишера — Снедекора

(k_1 — число степеней свободы большей дисперсии
k_2 — число степеней свободы меньшей дисперсии)

Уровень значимости $\alpha = 0.01$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	99,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40	99,41	99,42
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,85	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45

ОГЛАВЛЕНИЕ

Часть третья

Элементы математической статистики

Глава двадцатая. Однофакторный дисперсионный анализ