Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования
«Рязанский государственный университет имени С.А. Есенина»

А.А. Дунаев

Основы статистических методов компьютерной обработки результатов наблюдений

Учебное пособие

Рекомендовано УМО
по специальностям педагогического образования
в качестве учебного пособия
для студентов высших учебных заведений,
обучающихся по специальности 050202 — информатика

Рязань 2007

ББК 32.97я73
Д83

Научный редактор *В.Л. Григорьев*, д-р техн. наук, проф. (РГУ)

Рецензенты: *Е.М. Прошин*, д-р техн. наук, проф. (РРТУ)
Е.Н. Моос, д-р техн. наук, проф.
(СПб. АУ и Э; Ряз. филиал)

Дунаев, А.А.

Д83 Основы статистических методов компьютерной обработки результатов наблюдений : учебное пособие / А.А. Дунаев ; Ряз. гос. ун-т им. С.А. Есенина. — Рязань, 2007. — 180 с. : ил.

ISBN 978–5–88006–493–9

В учебном пособии изложены основные понятия и определения теории вероятностей, случайных величин, элементы математической статистики, теории корреляции, анализ временных рядов. Содержится теоретический курс и подробный разбор профильных задач.

Предназначено для студентов, обучающихся по следующим специальностям: «Математика с дополнительной специальностью информатика», «Информатика с дополнительной специальностью английский язык», «Математика с дополнительной специальностью физика», «Математическое обеспечение и администрирование информационных систем», «Информатика». Может быть полезно преподавателям и аспирантам, а также студентам гуманитарных специальностей: экономика, юриспруденция, биология.

Ключевые слова: *вероятность, статистика, корреляция, временные ряды, компьютеры*

ББК 32.97я73

ISBN 978–5–88006–493–9

ВВЕДЕНИЕ

Теория вероятностей и математическая статистика — науки, изучающие количественную сторону массовых случайных явлений информационного, технического, экономического, социального, общественного характера и закономерности, проявляющиеся в их изменениях. Методы теории вероятностей и статистики широко применяются в различных разделах информационных технологий, физики, биологии, экономики и др.

Настоящее учебное пособие представляет собой краткое изложение основ теории вероятностей и математической статистики, основано на курсе лекций, в течение ряда лет читаемых автором для студентов вузов технических и педагогических специальностей и специальности «Менеджмент».

Пособие содержит основы теории вероятностей, случайных величин и математической статистики, наиболее важные современные методы расчетов. В нем изложены такие темы, как определения вероятности, теоремы сложения и произведения событий, формула Байеса, локальная и интегральная теоремы Лапласа, законы распределения случайных величин и их характеристики, центральная предельная теорема, закон больших чисел, методы традиционной и непараметрической статистики, статистическая проверка гипотез, корреляционный анализ, одно- и двухфакторный анализ, анализ временных рядов.

Изложение теоретического материала сопровождается примерами с подробным решением. Это помогает студентам заочного отделения самостоятельно изучать материал.

1. ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

1.1. Испытания, события, их виды

Теория вероятностей изучает закономерности, проявляющиеся в массовых случайных явлениях или событиях. Массовые явления и процессы характеризуются многократным повторением при постоянных условиях некоторых опытов, экспериментов, операций.

Испытанием называется каждое осуществление заданного комплекса условий, который в принципе может быть воспроизведен сколь угодно большое количество раз, например, бросание игрального кубика, проведение эксперимента и т. п.

Явления, происходящие в результате испытания, называются событиями. События обозначают большими начальными буквами латинского алфавита ***A, B, C*** и т.д. Например, событие ***A*** — появление сверху герба при бросании монеты, событие ***B*** — сдача экзамена, событие ***C*** — выпадение определенной грани при бросании игрального кубика. Если событий много, их можно обозначать буквами с индексами: $A_1, A_2, \ldots, A_n$.

Появление события зависит от многих условий и причин. Например, при бросании монеты в помещении на результат бросания монеты влияют многие факторы. Первую группу условий или факторов мы можем задать: температуру, освещение, характер поверхности пола. Вторую группу факторов и условий трудно, но можно задать, например, влажность и давление воздуха. Третью группу составляют условия и факторы, о влиянии которых на результат эксперимента науке ничего не известно. Вторая и третья группы условий и факторов в совокупности образуют фактор случайности. Вследствие взаимодействия заданных факторов и фактора случайности в результате испытания может появиться то или иное событие из совокупности событий, возможных при данном испытании, причем заранее нельзя предсказать, какое именно из них произойдет. События этой совокупности называют случайными. Например, при бросании игрального кубика нельзя заранее предсказать, какая грань окажется сверху, поэтому появление определенной грани является случайным событием.

Некоторые события происходят неизбежно в результате каждого испытания, такие события называются достоверными. События, которые не могут произойти в результате испытания, называются невозможными.

Например, при извлечении шара из урны, в которой все шары белые, событие ***D*** — вынут белый шар — является достоверным, событие ***H*** — вынут черный шар — невозможным.

События $A_1, A_2, \ldots, A_n$ называются несовместимыми (совместимыми), если появление одного из них исключает (не исключает) появление

других событий в одном и том же испытании, то есть если невозможно (возможно) их совместное появление. Например, при бросании монеты событие ***A*** — выпадение герба и событие ***B*** — выпадение цифры — несовместимые события, так как осуществление одного из них исключает осуществление другого. Если при бросании игрального кубика событие ***A*** — появление четырех очков и событие ***B*** — появление четного числа очков, то события ***A*** и ***B*** — совместимые.

События $A_1, A_2, \ldots, A_n$ образуют полную группу событий, если в результате испытания появится хотя бы одно из них. Примером полной группы событий является выпадение одного, двух, трех, четырех, пяти или шести очков при одном бросании игрального кубика.

События $A_1, A_2, \ldots, A_n$ называются равновозможными, если нет оснований считать, что одно из них происходит чаще, чем другие. Например, выпадение граней игрального кубика — равновозможные события, появление любой цифры при розыгрыше лотерей также являются равновозможными событиями.

События $A_1, A_2, \ldots, A_n$, образующие полную группу попарно несовместимых и равновозможных событий, называются элементарными событиями, или исходами. В примере с игральным кубиком исход A_i состоит в том, что кубик при подбрасывании выпадает гранью с цифрой $i\ (i=1,2,3,4,5,6)$, так как события $A_1, A_2, \ldots, A_6$ образуют полную группу попарно несовместимых равновозможных событий.

Исход A_i называют благоприятствующим событию ***B***, если появление исхода A_i влечет за собой наступление события ***B***. Например, при подбрасывании игрального кубика событию ***B*** — появлению четной цифры на грани кубика благоприятствуют три исхода A_2, A_4 и A_6.

Вероятность случайного события ***A*** — это количественная мера степени объективной возможности наступления этого события и обозначается ***P(A)*** или ***p***. Существуют несколько методов определения вероятности случайного события.

1.2. Классическое определение вероятности

Вероятность ***P(A)*** случайного события ***A*** равна отношению числа ***m*** исходов, благоприятствующих событию ***A***, к общему числу ***n*** всех возможных исходов испытания:

$$P(A) = \frac{m}{n}. \qquad (1.1)$$

Свойства вероятности:

— вероятность достоверного события ***D*** равна ***1***: $P(D)=1$;

— вероятность невозможного события ***H*** равна ***0***: $P(H)=0$;

— вероятность случайного события ***A*** заключается между ***0*** и ***1***:

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

Задачи на вычисление вероятности случайного события **A** по его классическому определению целесообразно решать по следующей схеме:

• По условию задачи выяснить, что собой представляет испытание; какие события могут появиться в результате испытания; будут ли эти события исходами, то есть образуют ли они полную группу несовместимых и равновозможных событий. Если исходы есть, то переходим к п. 2, если исходов нет, то применять классическое определение вероятности нельзя, следует использовать другие определения вероятности или же использовать теоремы сложения и умножения вероятностей.

• Подсчитать число n всех исходов.

• Определить, какие исходы благоприятствуют случайному событию A и подсчитать их число m.

• Вычислить вероятность события по формуле: $P(A) = \frac{m}{n}$.

Пример 1. Из цифр от *1* до *9* включительно наугад выбирают одну. Найти вероятность выбора четного числа.

Решение

1. В данной задаче испытание — это выбор цифры наугад. В результате возможно наступление одного из следующих событий $A_1, A_2, \ldots, A_9$: A_1 — появление цифры *«1»*, A_2 — появление *«2»*, …., A_9 — появление *«9»*. Эти события будут несовместимыми (появление одной цифры исключает появление остальных цифр в том же испытании) и равновозможными (выбор производится наугад и все цифры имеют равную возможность появиться в испытании). Кроме того, эти события образуют полную группу, так как в результате выбора одна из цифр обязательно появится. Следовательно, события $A_1, A_2, ..., A_9$ являются исходами.

2. Всего исходов $n = 9$.

3. Событие A (появление четного числа) — наступает при появлении четырех исходов — A_2, A_4, A_6 и A_8, то есть событию A благоприятствуют m=4 исхода.

4. Вероятность случайного события A равна: $P(A) = \frac{m}{n} = 4/9$.

1.3. Статистическое определение вероятности

Частотой $m(A)$ события A называется число испытаний, в которых появилось это событие.

Относительной частотой $w(A)$ события A называется отношение числа $m(A)$ испытаний, в которых появилось событие, к числу n всех испытаний, в которых оно могло появиться:

$$w(A) = \frac{m(A)}{n}. \qquad (1.2)$$

Вероятностью ***P(A)*** события ***A*** в данном испытании называется число, к которому приближается (или около которого колеблется) относительная частота события ***A*** при сохранении неизменных условий опыта и при неограниченном увеличении числа ***n*** испытаний:

$$w(A) \underset{n\to\infty}{\to} P(A). \qquad (1.3)$$

Практически за вероятность принимают относительную частоту события при достаточно большом числе испытаний:

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

Между классическим и статистическим определением вероятности существуют различия и общие черты.

Первое различие состоит в том, что классическая вероятность получается в результате расчета (то есть мысленного эксперимента), а статистическая вероятность определяется по результатам опыта, по экспериментальным данным.

Второе различие состоит в том, что классическая вероятность получается абсолютно точной, без погрешности, а статистическая вероятность является приближенной оценкой вероятности, то есть содержит погрешность, так как определяется по результатам эксперимента.

Третье различие состоит в области применения этих способов определения вероятности. Классическое определение вероятности применимо только тогда, когда сложное событие, вероятность которого надо определить, сводится к схеме равновозможных и несовместимых событий, составляющих полную группу, то есть к схеме, состоящей из элементарных событий (исходов).

Статистическое определение вероятности таких ограничений не имеет и применимо практически для любых событий.

Общим в классическом и статистическом определениях вероятности является то, что это два различных способа оценки одной той же величины: вероятности, то есть степени объективной возможности наступления события.

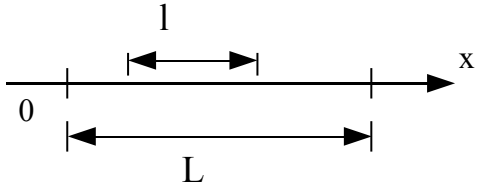
1.4. Геометрическая вероятность

Пусть в область ***G*** бросается наудачу точка. Выражение «бросается наудачу» понимается в том смысле, что брошенная точка может попасть в любую точку области ***G.*** Вероятность попадания в какую-либо часть области ***g*** пропорциональна мере части (длине, площади, объему) и не зависит от ее расположения и формы.

Таким образом, если g — часть области G, то вероятность попадания в область g по определению равна

$$P(g) = \frac{\text{мера } g}{\text{мера } G}. \qquad (1.4)$$

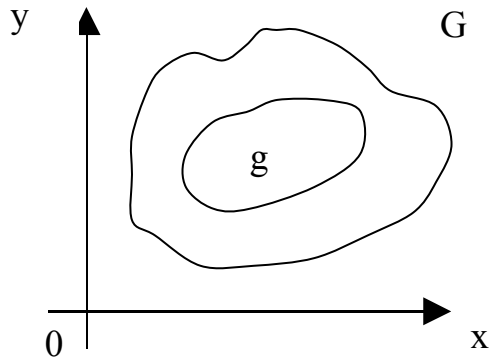
Например, если область G представляет собой прямую длиной L, то вероятность попасть в отрезок длиной $l \le L$ определяется по формуле (рис. 1.1):



$$P(l) = \frac{l}{L}. \qquad (1.5)$$

Рис. 1.1

Если область G представляет собой замкнутую фигуру на плоскости (рис. 1.2), то вероятность попасть в область g равна отношению площади S_g области g и площади S_G области G.



$$P(g) = \frac{S_g}{S_G}. \qquad (1.6)$$

Рис. 1.2

Заметим, что здесь пространство G представляет собой совокупность всех точек области G и, значит, состоит из бесконечного множества элементарных событий. Следовательно, понятие «геометрическая вероятность» можно рассматривать как обобщение понятия «классическая вероятность» на случай опытов с бесконечным числом исходов.

Пример 2. Из отрезка [0;2] наугад выбраны два числа x и y. Найти вероятность того, что эти числа удовлетворяют неравенствам $x^2 \le 4y \le 4x$.

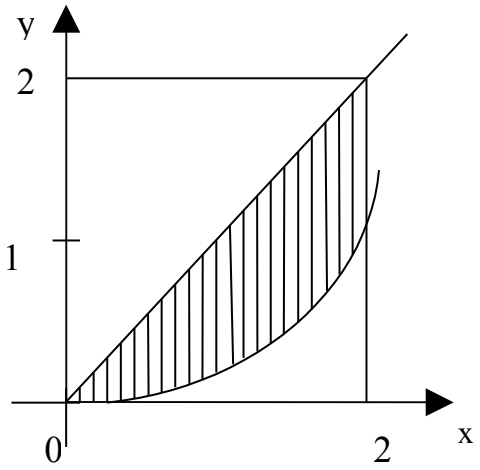


Рис. 1.3

Решение

По условиям опыта координаты точки M(x,y) удовлетворяют системе неравенств:

$$\begin{cases} 0 \le x \le 2, \\ 0 \le y \le 2. \end{cases}$$

Это означает, что точка ***M(x,y)*** выбирается наугад из множества **G** точек квадрата со стороной 2 (рис. 1). Множество ***g*** точек, координаты которых удовлетворяют неравенствам $x^2 \le 4y \le 4x$ или $\frac{x^2}{4} \le y < x$, образуют заштрихованную фигуру **g**. Следовательно, искомая вероятность события ***A*** — удовлетворение неравенствам $\boldsymbol{x^2 \le 4y \le 4x}$ будет равна отношению площади заштрихованной фигуры g к площади квадрата G:

Рис.1.3.

$$P(A) = \frac{\text{площадь } g}{\text{площадь } G} = \frac{\int_0^2 \left(x - \frac{1}{4}x^2 \right) dx}{2 \cdot 2} = \frac{\left(\frac{x^2}{2} - \frac{x^3}{12} \right)\Big|_0^2}{4} = \frac{\frac{4}{2} - \frac{8}{12}}{4} = \frac{1}{3}\cdot$$

1.5. Практически невозможные и практически достоверные события. Принцип практической уверенности

На практике часто приходится иметь дело не с невозможными и достоверными событиями, а с так называемыми «практически невозможными» и «практически достоверными» событиями.

Практически невозможным событием называется событие, вероятность которого очень близка к нулю, но не равна нулю (например, угадать 6 из 49 номеров в спортлото).

Практически достоверным событием называется событие, вероятность которого не в точности равна единице, но весьма близка к единице (например, не угадать 6 номеров из 49 в спортлото).

С точки зрения теории вероятностей, все равно, о каких событиях говорить: о практически невозможных или о практически достоверных, так как они всегда сопутствуют друг другу.

Если, например, известно, что вероятность события в данном опыте равна 0,3, это еще не дает возможности предсказать результат опыта, но если вероятность события A в данном опыте ничтожно мала или весьма близка к единице, это дает возможность предсказать результат опыта; в первом случае нельзя ожидать появления в результате опыта события A, во втором случае можно ожидать его с достаточным основанием.

Принцип практической уверенности: если вероятность некоторого события A в данном опыте близка к единице, то можно быть практически уверенным в том, что при однократном выполнении опыта событие A произойдет.

Принцип практической уверенности подтверждается всем опытом человечества, хотя не может быть доказан математическими средствами. Вопрос о том, насколько мала должна быть вероятность события, чтобы его можно было считать практически невозможным, в каждом отдельном случае решается из практических соображений в соответствии с той важностью, которую имеет для нас желаемый результат опыта.

Например, если вероятность отказа парашюта при прыжке равна $0{,}01$, то нельзя считать его отказ практически невозможным событием, а если поезд опоздает на станцию назначения с той же вероятностью, то мы этой вероятностью пренебрежем.

1.6. Элементы комбинаторики

При подсчете числа исходов в классическом определении вероятностей часто используют понятия и формулы комбинаторики.

Размещениями из n различных элементов по m элементов называются комбинации, составленные из данных n элементов по m элементов, которые отличаются либо набором элементов, либо порядком их следования. Число всех размещений из n элементов по m вычисляется по формуле:

$$A_n^m = \frac{n!}{(n-m)!} = n(n-1)(n-2)...(n-m+1), \qquad (1.7)$$

где $n!=1\cdot 2\cdot 3\cdot \ldots \cdot n$, причем полагают, что $0!=1$.

Перестановками из n различных элементов называются размещения из этих n элементов по n. Число всех перестановок из n элементов вычисляется по формуле:

$$P_n = n! \qquad (1.8)$$

Сочетаниями из n различных элементов по m элементов называются комбинации, составленные из данных n элементов по m элементам, которые отличаются только набором элементов, то есть хотя бы одним элементом. Порядок расположения элементов роли не играет. Число всех сочетаний из n элементов по m вычисляется по формуле:

$$C_n^m = \frac{A_n^m}{P_m} = \frac{n!}{m!(n-m)!} = \frac{n(n-1)(n-2)\cdot ...\cdot (n-m+1)}{1\cdot 2\cdot 3\cdot ...\cdot (m-1)m}, \qquad (1.9)$$

причем:

— $C_n^m = C_n^{n-m}$;

— $C_n^0 = C_n^n = 1$;

— $C_n^0 + C_n^1 + \cdots + C_n^n = 2^n$.

Пример 3. Десять книг наугад расставляют на книжной полке. Какова вероятность события A — три конкретные книги из этих 10 книг окажутся стоящими рядом?

Решение

1. Испытание — расстановка книг наугад. Исходом будет являться любая расстановка десяти книг на полке.

2. Число n всех исходов будет равно числу всех расстановок из 10 книг P_{10}, то есть числу размещений A_{10}^{10} из 10 книг по 10, отличающихся только порядком расположения книг:

$$n = A_{10}^{10} = P_{10} = 10!.$$

3. Благоприятными для события A являются расстановки, в которых данные три книги стоят рядом. Подсчитаем число m таких расстановок.

Для этого представим, что данные три книги объединены в связку. Условимся рассматривать эту связку как одну «большую» книгу, то есть можно считать, что имеется всего 8 книг. Их расстановка возможна $8!$ способами. Кроме того, внутри «большой» книги данные три книги могут переставляться $3!$ способами. Следовательно, число m благоприятных перестановок будет равно произведениям числа этих способов $8!\cdot 3!$.

4. Искомая вероятность равна:

$$P(A) = \frac{m}{n} = \frac{8! \cdot 3!}{10!} = \frac{1 \cdot 2 \cdot 3}{9 \cdot 10} = \frac{1}{15}.$$

Пример 4. Из ящика, в котором находятся ***3*** бракованных и ***12*** годных деталей, наугад берут две детали. Найти вероятность появления двух годных деталей.

Решение

1. Испытание — выбор наугад двух деталей. В результате испытания возможно появление следующих событий: выбор любого сочетания по две детали из ***15*** деталей. Все эти события будут несовместимыми (отличаются хотя бы одной деталью) и равновозможными (выбор производится наугад). Кроме того, они образуют полную группу (какие-то две детали обязательно выберут), то есть являются исходами.

2. Общее число ***n*** всех исходов равно числу способов извлечения из ящика двух деталей из ***15***, то есть числу сочетаний C_{15}^2:

$$n = C_{15}^2 = \frac{15!}{2!(15-2)!} = \frac{15!}{2!13!} = \frac{14 \cdot 15}{1 \cdot 2} = 105.$$

3. Пусть ***A*** — событие, состоящее в появлении двух годных деталей. Число ***m*** исходов, благоприятствующих этому событию, будет равно числу способов выбора двух годных деталей из ***12*** годных, то есть числу сочетаний C_{12}^2:

$$m = C_{12}^2 = \frac{12!}{2!(12-2)!} = \frac{12!}{2!10!} = \frac{11 \cdot 12}{1 \cdot 2} = 66.$$

4. Искомая вероятность $P(A) = \frac{m}{n} = \frac{66}{105} = \frac{22}{35}$.

1.7. Сумма и произведение событий

Суммой двух событий ***A*** и ***B*** называют событие ***C=A+B***, заключающееся в наступлении события ***A***, или события ***B***, или событий ***A*** и ***B*** одновременно.

Например, два стрелка производят по одному выстрелу. Событие ***A*** — попадание в мишень первым стрелком, событие ***B*** — попадание в мишень вторым стрелком. Событие ***C=A+B*** — попадание при выстреле первым стрелком или вторым, или первым и вторым стрелками вместе.

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

Пусть события ***A*** и ***B*** — несовместимые и известны их вероятности.

Теорема. Вероятность суммы двух несовместимых событий ***A*** и ***B*** равна сумме вероятностей этих событий

$$P(A+B)=P(A)+P(B). \qquad (1.10)$$

Доказательство. Пусть n — общее число возможных элементарных событий, m_1 — число элементарных событий, благоприятствующих событию A; m_2 – число элементарных событий, благоприятствующих событию B. Так как A и B — несовместимые события, то событию $A+B$ будет благоприятствовать m_1+m_2 элементарных событий. Согласно классическому определению вероятности

$$P(A+B)=\frac{m_1+m_2}{n}=\frac{m_1}{n}+\frac{m_2}{n}. \qquad (1.11)$$

Учитывая, что $P(A)=\frac{m_1}{n}$ и $P(B)=\frac{m_2}{n}$, окончательно получим

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

Аналогично вероятность суммы попарно несовместимых $A_1, A_2, \ldots, A_n$ равна:

$$P(A_1+A_2+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n). \qquad (1.12)$$

Теорема. Сумма вероятностей несовместимых событий, образующих полную группу, равна единице:

$$P(A_1)+P(A_2)+\ldots+P(A_n)=1. \qquad (1.13)$$

Доказательство. Так как события $A_1, A_2, \ldots, A_n$ образуют полную группу событий, то наступление хотя бы одного из них есть событие достоверное. Следовательно, $P(A_1+A_2+\ldots+A_n)=1$. Отсюда по теореме сложения несовместимых событий получим исходное равенство.

Два несовместимых события A и $\overline{A}$ (не A), образующих полную группу, называются противоположными, причем $P(A)+P(\overline{A})=1$ или $p+q=1$ ($q=P(\overline{A})$), откуда $P(A)=1-P(\overline{A})$ или $p=1-q$.

Например, для события A — выпадение герба при бросании монеты противоположным является событие $\overline{A}$ — выпадение цифры.

Произведением событий A и B называется сложное событие $C=A\cdot B$, состоящее в совместном осуществлении событий A и B. События A и B не обязательно появляются одновременно и в одном месте.

События A и B называются независимыми (зависимыми) друг от друга, если вероятность появления любого из них не зависит (зависит) от того, наступило другое событие или нет. Например, при подбрасывании монеты два раза вероятность появления герба во втором испытании (событие A) не зависит от появления или непоявления цифры в первом испытании (событие B). События A и B независимые.

Несколько событий называются попарно независимыми, если любые два из них независимые.

Несколько событий называются независимыми в совокупности, если каждое из них и любая комбинация остальных событий есть события независимые.

Теорема. Вероятность произведения двух независимых событий ***A*** и ***B*** равна произведению их вероятностей:

$$P(A\cdot B)=P(A)\cdot P(B). \qquad (1.14)$$

Доказательство. Пусть n_1 — число возможных исходов, в которых событие ***A*** осуществляется или нет; m_1 — число исходов, благоприятствующих событию ***A***; n_2 — число возможных исходов, в которых событие ***B*** осуществляется или нет; m_2 — число исходов, благоприятствующих событию ***B***. Общее число возможных исходов испытания равно $n_1\cdot n_2$. Из общего числа возможных исходов число $m_1\cdot m_2$ событий благоприятствует совместному осуществлению событий ***A*** и ***B***. Вероятность совместного наступления событий ***A*** и ***B*** поэтому равна:

$$P(A\cdot B)=\frac{m_1m_2}{n_1n_2}=\frac{m_1}{n_1}\cdot\frac{m_2}{n_2}. \qquad (1.15)$$

Учитывая, что $P(A)=\frac{m_1}{n_1}$, а $P(B)=\frac{m_2}{n_2}$, получаем $P(A\cdot B)=P(A)\cdot P(B)$.

Аналогично можно доказать, что вероятность произведения n попарно независимых событий равна:

$$P(A_1\cdot A_2\cdot\ldots\cdot A_n)=P(A_1)\cdot P(A_2)\cdot\ldots\cdot P(A_n). \qquad (1.16)$$

Условной называется вероятность осуществления события ***B*** при условии, что событие ***A*** уже наступило, обозначается это $P(B/A)$ или, если первым наступило событие ***B***: $P(A/B)$.

Теорема. Вероятность произведения двух зависимых событий ***A*** и ***B*** равна произведению вероятности одного из них на условную вероятность другого при условии, что первое уже наступило:

$$P(A\cdot B)=P(A)\cdot P(B/A)=P(B)\cdot P(A/B). \qquad (1.17)$$

Доказательство. Пусть из общего числа n исходов m_1 благоприятствуют событию ***A***, а из этих m_1 событий m_2 благоприятствуют событию ***B***, а значит, и событию ***AB***. Тогда

$$P(A\cdot B)=\frac{m_2}{n}=\frac{m_1}{n}\cdot\frac{m_2}{m_1}=P(A)P(B/A). \qquad (1.18)$$

Аналогично вероятность произведения n зависимых событий равна:

$$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = \\ = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cdot A_2) \cdot \ldots \cdot P(A_n/A_1 \cdot A_2 \cdot \ldots \cdot A_{n-1}). \quad (1.19)$$

Вероятность появления хотя бы одного из n событий A_i $(i=1, 2, 3, \ldots, n)$, то есть вероятность суммы этих событий равна:

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A_1 \cdot A_2 \cdot \ldots \cdot A_n}) = 1 - P(\overline{A}_1 \cdot \overline{A}_2 \cdot \ldots \cdot \overline{A}_n). \quad (1.20)$$

В частности, если события A_i независимы, то:

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A}_1) \cdot P(\overline{A}_2) \cdot \ldots \cdot P(\overline{A}_n), \quad (1.21)$$

где

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

Если вероятности всех событий равны $P(A_i) = p$ и $P(\overline{A}_i) = 1 - p = q$, то

$$P(A_1 + A_2 + \ldots + A_n) = 1 - (1 - p)^n = 1 - q^n. \quad (1.22)$$

Задачи на вычисление вероятности сложного события, являющегося некоторой комбинацией других событий, целесообразно решать по следующей схеме:

- Обозначить все упоминаемые в условии задачи случайные события.
- Событие, вероятность которого нужно найти, представить в виде суммы или (и) произведения других, более простых событий, вероятности которых или известны по условию задачи, или легко вычисляются.
- Вычислить искомую вероятность, используя формулы вероятности суммы и произведения событий или формулу вероятности противоположного события. Предварительно надо выяснить, являются ли события несовместимыми (совместимыми) при сложении событий или же зависимыми (независимыми) при умножении событий.

Пример 5. В одну и ту же мишень независимо друг от друга стреляют два человека. Вероятность попадания в мишень для первого стрелка 0,7, а для второго — 0,6. Найти вероятность того, что в мишень попадет: а) хотя бы один человек; б) только один человек.

Решение

1. Обозначим: событие A — попадание в мишень первого стрелка; событие B — попадание второго стрелка, событие C — попадание хотя бы одного стрелка. Тогда событие C будет являться суммой событий A и B:

$$C = (A \text{ или } B) = A + B.$$

Найдем вероятность ***P(C)*** через вероятность противоположного события $\overline{C}$ — в мишень не попадет ни один стрелок. Очевидно, событие $\overline{C}$ заключается в непопадании первого и второго стрелков: $\overline{C} = \overline{A}$ и $\overline{B} = \overline{A} \cdot \overline{B}$. Так как события ***A*** и ***B*** — независимые, то независимы и события $\overline{A}$ и $\overline{B}$. Тогда $P(\overline{C}) = P(\overline{A} \cdot \overline{B}) = P(\overline{A}) \cdot P(\overline{B}) = (1 - P(A)) \cdot (1 - P(B))$, откуда $P(C) = 1 - P(\overline{C}) = 1 - (1 - P(A)) \cdot (1 - P(B))$.

Подставляя численные значения, получим:

$$P(C) = 1 - (1 - 0,7) \cdot (1 - 0,6) = 1 - 0,3 \cdot 0,4 = 1 - 0,12 = 0,88.$$

2. Обозначим: событие ***E*** — попадание в мишень только одного стрелка. Это событие произойдет, если в мишень попадет только первый стрелок (второй стрелок при этом не попадет) или только второй стрелок (первый стрелок при этом не попадет), то есть $E = A \cdot \overline{B}$ или $\overline{A} \cdot B = A \cdot \overline{B} + \overline{A} \cdot B$.

События $(A \cdot \overline{B})$ (попадание только первого стрелка) и $(\overline{A} \cdot B)$ (попадание только второго стрелка) несовместимы, поэтому по формуле сложения для несовместимых событий получим: $P(E) = P(A\overline{B} + \overline{A} \cdot B) = P(A \cdot \overline{B}) + P(\overline{A} \cdot B)$.

События ***A*** и ***B*** — независимы, следовательно, независимыми будут пары событий A и $\overline{B}$ и $\overline{A}$ и B. Тогда

$$P(E) = P(A) \cdot P(\overline{B}) + P(\overline{A}) \cdot P(B) = P(A) \cdot (1 - P(\overline{B})) + (1 - P(A)) \cdot P(B).$$

Подставляя $P(A) = 0,7$ и $P(B) = 0,6$, получим:

$$P(E) = 0,7 \cdot (1 - 0,6) + (1 - 0,7) \cdot 0,6 = 0,7 \cdot 0,4 + 0,3 \cdot 0,6 = 0,28 + 0,18 = 0,46.$$

1.8. Следствия теорем сложения и умножения вероятностей

1.8.1. Теорема сложения вероятностей совместимых событий

Пусть известны вероятности появления совместимых событий ***P(A)*** и ***P(B)*** и вероятность их совместного появления ***P(A·B).***

Теорема. Вероятность появления хотя бы одного из двух совместимых событий равна сумма вероятностей этих событий без вероятности их совместного появления

$$P(A+B)=P(A)+P(B)-P(A \cdot B). \qquad (1.23)$$

Аналогично вероятность суммы трех совместимых событий вычисляется по формуле

$$P(A+B+C)=P(A)+P(B)+P(C)-P(AB)-P(AC)-P(BC)-P(ABC). \qquad (1.24)$$

Пример 6. Одну и ту же задачу независимо друг от друга решают два студента. Вероятность решения задачи первым студентом равна ***0,7***,

вторым — ***0,9***. Найти вероятность того, что задачу решит хотя бы один студент.

Решение

1. Обозначим: событие ***A*** — решение задачи первым студентом; событие ***B*** — решение задачи вторым студентом; событие ***C*** — решение задачи хотя бы одним студентом. Тогда событие **C** будет являться суммой событий ***A*** и ***B***:

$$C = (A \text{ или } B) = A + B.$$

Так как события ***A*** и ***B*** совместимы, то используем формулу (1.23) для вероятности суммы совместимых событий***: P(C) = P(A+B) = P(A) + P(B) – P(A·B).***

События ***A*** и ***B*** независимые, поэтому $P(A \cdot B) = P(A) \cdot P(B)$, и

$$P(C) = P(A) + P(B) — P(A) \cdot P(B).$$

Подставляя численные значения, получим: $P(C) = 0{,}7 + 0{,}6 — 0{,}7 \cdot 0{,}6 = 0{,}88$.

1.8.2. Формула полной вероятности

Пусть событие ***A*** может наступить при условии появления одного из несовместимых событий (гипотез) $B_1, B_2, \ldots, B_n$, которые образуют полную группу. Пусть известны вероятности этих гипотез и условные вероятности $P(A/B_1), P(A/B_2), \ldots, P(A/B_n)$ события ***A*** при этих гипотезах.

Теорема. Вероятность события ***A***, которое может наступить лишь при появлении одного из несовместимых 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***:

$$P(A)=P(B_1) \cdot P(A/B_1)+P(B_2) \cdot P(A/B_2)+\ldots+P(B_n) \cdot P(A/B_n) \qquad (1.25)$$

или $$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A / B_i). \qquad (1.26)$$

Доказательство. Осуществление события ***A*** требует появления одного из следующих событий:

$$B_1 \cdot A \text{ или } B_2 \cdot A \text{ или } \ldots \text{ или } B_n \cdot A.$$

Эти события несовместимы, поэтому к ним применима теорема сложения вероятностей несовместимых событий:

$$P(A)=P(B_1A)+P(B_2A)+ \ldots +P(B_nA). \qquad (1.27)$$

События B_i и ***A*** зависимы, поэтому по теореме умножения зависимых событий, получим

$$P(A) = P(B_1)\cdot P(A/B_1) + P(B_2)\cdot P(A/B_2) + \ldots + P(B_n)P(A/B_n) = \sum_{i=1}^{n} P(B_i)\cdot P(A/B_i) \quad (1.28)$$

1.8.3. Формула Байеса

Пусть событие ***A*** может наступить совместно с одним из ***n*** несовместимых событий (гипотез) $B_1, B_2, \ldots, B_n$, образующих полную группу. Вероятность события ***A*** можно определить по формуле полной вероятности (1.25).

Пусть в результате испытания появилось событие ***A***. Определим, как при этом условии изменились вероятности гипотез $P(B_i)$, то есть будем искать $P(B_i/A)$.

По теореме умножения имеем:

$$P(A\cdot B_i) = P(A)\cdot P(B_i/A) = P(A/B_i)\cdot P(B_i). \quad (1.29)$$

Отсюда $$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{P(A)}. \quad (1.30)$$

Выразим $P(A)$ по формуле полной вероятности:

$$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{P(B_1)\cdot P(A/B_1) + P(B_2)\cdot P(A/B_2) + \ldots + P(B_n)\cdot P(A/B_n)}$$

или в общем виде

$$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{\sum_{i=1}^{n} P(B_i)\cdot P(A/B_i)}. \quad (1.31)$$

Пример 7. Первая, вторая и третья смены в цеху завода производят ***50, 30*** и ***20*** % всей продукции. Доля бракованной продукции составляет для каждой смены ***2***, ***1*** и ***3*** % соответственно. Для проверки наугад выбирается единица продукции.

- Какова вероятность, что она окажется бракованной (событие ***A***)?
- Если единица продукции оказалась бракованной, то какова вероятность, что она сделана во вторую смену?

Решение

1. Единица продукции может быть изготовлена в первую смену, или во вторую, или в третью; обозначим первое из этих событий (гипотез) B_1, второе — B_2, третье — B_3. События B_1, B_2 и B_3 несовместимы и образуют полную группу. По условию задачи, вероятность изготовления единицы продукции в первую смену равна ***0,5***: $P(B_1) = 0{,}5$. Аналогично $P(B_2) = 0{,}3$ и $P(B_3) = 0{,}2$.

Кроме того, известны *условные* вероятности изготовления бракованной продукции (события A) в каждой смене:

$$P(A/B_1) = 0{,}02; \quad P(A/B_2) = 0{,}01; \quad P(A/B_3) = 0{,}03.$$

Тогда по формуле полной вероятности получим:

$$P(A) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + P(B_3) \cdot P(A/B_3) =$$

$$= 0{,}5 \cdot 0{,}02 + 0{,}3 \cdot 0{,}01 + 0{,}2 \cdot 0{,}03 = 0{,}01 + 0{,}003 + 0{,}006 = 0{,}019.$$

2. Для нахождения вероятности $P(A/B_2)$ используем формулу Байеса (1.31):

$$P(B_2 / A) = P(B_2) \cdot \frac{P(A / B_2)}{P(A)} = 0{,}3 \cdot \frac{0{,}01}{0{,}019} = \frac{3}{19} \approx 0{,}16 = 16\,\%.$$

Таким образом, из всей бракованной продукции цеха в среднем ***16 %*** изготавливается во вторую смену, что почти в два раза меньше доли (***30 %***) всей продукции, изготовляемой в этой смене.

Задачи для самостоятельного решения

Задачи 1—10 решать, используя классическое определение вероятности.

1. Наудачу выбрано двузначное число. Какова вероятность того, что это число окажется: а) кратным 3; б) кратным 5; в) большим 40; г) не более 50; д) с равными цифрами?
2. Из пяти карточек с буквами «***А***», «***А***», «***В***», «***И***», «***Д***» наугад одна за другой выбирают три и располагают в ряд в порядке появления. Какова вероятность того, что получится слово: а) «***ДВА***»; б) «***ВИД***»?
3. В бригаде рабочих 10 мужчин и 5 женщин. Для выполнения некоторой работы наугад выбирают 3 человек. Чему равна вероятность того, что будут выбраны: а) только мужчины; б) только женщины; в) двое мужчин и женщина?
4. При наборе телефонного номера абонент забыл две последние цифры и набрал их наугад. Найти вероятность того, что номер набран правильно, если известно, что: а) эти цифры нечетные и разные; б) первая цифра (слева) больше второй.
5. В партии из 20 изделий 5 бракованных. Из партии выбирают наугад 3 изделия. Найти вероятность того, что среди этих 3 изделий окажутся бракованными: а) одно изделие; б) два изделия; в) три изделия.
6. Случайным образом выбирают для проведения некоторых работ два дня в апреле 1999 года. Какова вероятность, что: а) они окажутся

выходными; б) они окажутся рабочими; в) они окажутся идущими друг за другом?

7. Из последовательности чисел 1, 2, 3, …, 15 наугад выбирают два числа. Какова вероятность, что одно из них меньше 5, а другое не менее 10?
8. Коля и Миша и еще 8 человек стоят в очереди. Определить вероятность того, что Коля и Миша окажутся: а) крайними в очереди;

   б) рядом.
9. Для проверки 10 предприятий отрасли комиссия случайным образом выбирает очередность проверки. Какова вероятность того, что какие-то конкретные три предприятия будут проверены: а) первыми; б) последними?
10. В студенческой группе 20 человек, из которых 12 девушек. Группу разделили на две равные части. Какова вероятность того, что в каждой части равное число: а) девушек; б) юношей?

Задачи 11—13 решать, используя классическое и статистическое определения вероятности.

11. Число N животных в стаде неизвестно. Из этого стада наугад отбирают M животных, которых клеймят и возвращают в стадо. Затем случайным образом отбирается n животных, среди которых m оказываются клеймеными. Чему приближенно равно значение N?
12. Выполните следующий эксперимент: бросьте 20 раз по две одинаковых монеты одновременно. Подсчитайте вероятность появления следующих событий: ГГ (герб, герб), ГЦ (герб, цифра), ЦГ, ЦЦ по статистическому и классическому определению вероятности. Подтверждают ли ваши вычисления гипотезу о том, что вероятности этих событий равны по 0,25?
13. Двое по очереди бросают монету, причем выигрывает тот, у кого раньше появится герб. Воспроизведите эту игру 20 раз и найдите приближенно вероятность выигрыша для начинающего игру. Сравните ее с точным значением, полученным по классическому определению вероятности.

Задачи 14—23 решать с использованием геометрического определения вероятности.

14. В круге радиуса R наудачу выбирается точка $M(x,y)$. Найти вероятность того, что эта точка окажется внутри вписанного в этот круг правильного треугольника.
15. Стержень длиной l произвольным образом ломают на три части. Какова вероятность того, что из этих частей можно составить треугольник?

16. Точка $M(x,y)$ выбирается наугад в круге радиуса R. Какова вероятность того, что она окажется внутри квадрата: а) описанного около круга; б) вписанного в круг?
17. Плоскость разграфлена параллельными прямыми, отстоящими друг от друга на расстоянии $2a$. На плоскость наудачу брошена монета радиуса $r<a$. Найти вероятность того, что монета не пересечет ни одной из прямой.
18. Стержень длиной l ломают случайным образом на три части. Найти вероятность того, что длина каждой части окажется больше $l/4$.
19. Найти вероятность того, что сумма двух наудачу взятых чисел из отрезка $[-1;1]$ больше нуля, а их произведение отрицательно.
20. В квадрате с вершинами в точках (0;0), (0;1), (1,1), (1,0) наудачу выбирается точка $M(x,y)$. Найти вероятность того, что координаты этой точки удовлетворяют неравенству $y<2x$.
21. Два лица договорились встретиться в определенном месте между 12 и 13 часами, причем каждый пришедший на встречу ждет другого в течение 20 минут, после чего уходит. Найти вероятность встречи этих лиц, если каждый из них приходит на встречу в случайный момент времени, не согласованный с моментом прихода другого.
22. Точка $M(x,y)$ выбирается наугад в круге радиуса R. Какова вероятность того, что она окажется внутри вписанного в круг: а) квадрата; б) правильного треугольника?
23. На паркет, составленный из правильных треугольников со стороной a, случайно брошена монета радиусом r $(r<a)$. Найти вероятность того, что монета не заденет границы ни одного из треугольников.

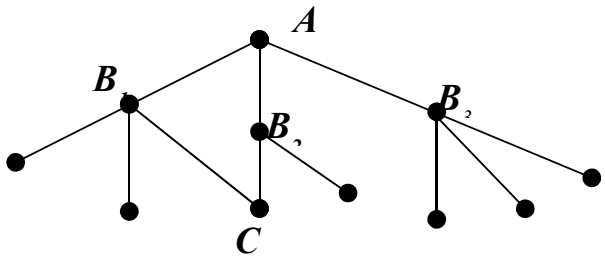
Задачи 24—33 решать, используя сумму и произведение событий.

24. Ящик содержит 90 годных и 10 дефектных деталей. Сборщик последовательно без возвращения достает из ящика 3 детали. Найти вероятность того, что среди взятых деталей: а) нет дефектных; б) хотя бы одна дефектная; в) только одна дефектная.
25. На прием к врачу записались случайным образом 12 человек, из них 7 женщин. Какова вероятность того, что первые двое пациентов по очереди будут: а) женщинами; б) мужчинами; в) лицами разного пола; в) лицами одного пола.
26. Детали проходят три операции обработки. Вероятность получения брака на первой операции равна 0,02; на второй — 0,03; на третьей — 0,01. Найти вероятность получения бракованной детали после: а) первой и второй операций; б) трех операций. Считать, что получение брака на каждой операции являются независимыми событиями.

27. Экзаменационный билет содержит три вопроса. Вероятность правильного ответа студента на первый и второй вопросы билета равны 0,9; на третий — 0,8. Найти вероятность того, что студент сдаст экзамен, если для этого необходимо правильно ответить: а) на все вопросы; б) хотя бы на два вопроса.
28. Среди изготовляемых станком-автоматом деталей в среднем 2 % брака. Какова вероятность, что среди взятых на проверку 4 деталей: а) не найдется ни одной бракованной; б) найдется не менее 3 бракованных; в) найдется не более одной бракованной?
29. Студент пришел на экзамен, зная только 30 вопросов из 40. Какова вероятность сдать экзамен, если для этого студент должен правильно ответить на: а) на все два вопроса; б) на все три вопроса; в) на два вопроса из трех.
30. В приборе имеется три микросхемы. Вероятность выхода из строя в течение года первой микросхемы равна 0,1; второй — 0,05 и третьей — 0,02. Найти вероятность того, что в течение года окажутся неисправными: а) только одна микросхема; б) хотя бы одна из микросхем; в) не более одной микросхемы.
31. В двух ящиках находятся 30 пуговиц, отличающихся только цветом: в первом ящике — 5 синих и 10 красных, во втором ящике — 10 синих и 5 красных. Наугад берут по одной пуговице из каждого ящика. Какова вероятность, что будут вынуты пуговицы: а) одного цвета; б) разного цвета; в) красного цвета?
32. Вероятность того, что в течение смены возникнет неполадка станка, равна 0,05. В цеху 4 станка, работающих независимо друг от друга. Какова вероятность того, что в течение смены произойдет: а) не более одной неполадки; б) хотя бы одна неполадка; в) только одна неполадка?
33. Прибор состоит из двух блоков I типа и трех блоков II типа. Событие $A_к$ (к = 1,2) — неисправен ***к***-й блок I типа; B_i (i = 1, 2, 3) — неисправен ***i***-й блок II типа. Прибор работает, если исправен хотя бы один блок I типа и не менее 2 блоков II типа. Найти вероятность работы прибора, если $P(A_к) = 0{,}05$ и $P(B_i) = 0{,}02$.

Задачи 34—43 решать, используя формулы полной вероятности и Байеса.

34. На рисунке изображена схема дорог, идущих из пункта ***A***. Турист вышел из пункта ***A*** и на каждом разветвлении дорог выбирает наугад один из возможных путей. а) Какова вероятность того, что турист попадет в пункт ***C***? б) Если турист все же пришел в пункт ***C***, то какова вероятность того, что он шел через пункт B_1?



35. Имеются два комплекта экзаменационных билетов разной сложности, то есть вероятность того, что студент правильно ответит на вопросы билетов из первого комплекта, равна 0,7, а из второго — 0,9. Преподаватель наугад берет на экзамен один из комплектов билетов, а студент наудачу вытаскивает из него билет. Какова вероятность, что билет окажется из первого комплекта, если студент ответил правильно на вопросы билета?
36. С первого станка-автомата на сборку поступают 40 %, со второго — 30 %, с третьего — 20 %, с четвертого — 10 % всех изготавливаемых ими деталей. Среди деталей, выпущенных первым станком,
    2 % бракованных, вторым станком — 1%, третьим — 0,5 % и четвертым — 0,2 %. Какова вероятность того, что поступившая на сборку деталь не бракованная? Какова вероятность того, что не бракованная деталь изготовлена первым станком?
37. Известно, что 96 % выпускаемых заводом изделий отвечает стандарту. Упрощенная схема контроля признает пригодной стандартную продукцию с вероятностью 0,98 и нестандартную — с вероятностью 0,05. Найти вероятность того, что изделие, прошедшее упрощенный контроль, отвечает стандарту.
38. Агентство по страхованию автомобилей разделяет водителей на 3 класса: класс B_1 (мало рискует), класс B_2 (рискует средне), класс B_3 (рискует сильно). Агентство предполагает, что из всех водителей, застраховавших автомобили, 30 % принадлежит к классу B_1, 50 % — к классу B_2, и 20 % — к классу B_3. Вероятность совершить хотя бы одну аварию в течение года для водителя класса B_1 равна 0,01, для водителя класса B_2 — 0,02, для водителя класса B_3 — 0,08. Водитель страхует свою машину и в течение года попадает в аварию. Какова вероятность того, что он относится к классу: а) B_1; б) B_2; в) B_3?
39. В специализированную больницу поступают в среднем 60 % больных с заболеванием ***K***, 30 % — с заболеванием ***L***, 10 % — с заболеванием ***M***. Вероятность полного излечения болезни ***K*** равна 0,7; для болезней ***L*** и ***M*** эти вероятности соответственно равны 0,8 и 0,9. Больной, поступивший в эту больницу, был выписан здоровым. Какова

вероятность того, что у этого больного было заболевание: а) ***K***; б) ***L***; в) ***M***?

40. По шоссе, на котором стоит автозаправочная станция, проезжает легковых машин в 1,5 раза больше, чем грузовых. Вероятность того, что проезжающая грузовая машина будет заправляться, равна 0,1; для легковой автомашины эта вероятность равна 0,05. На автозаправочную станцию подъехала машина. Какова вероятность того, что эта машина: а) грузовая; б) легковая?
41. Статистикой установлено, что курящие мужчины в возрасте свыше 40 лет заболевают раком легких в 9 раз чаще, чем некурящие мужчины, и что 60 % всех мужчин курят. Какова вероятность того, что мужчина, заболевший раком легких, был: а) курящим; б) некурящим?
42. Некоторое заболевание, встречающееся у 5 % населения, с трудом поддается диагностике. Один грубый тест на это заболевание дает положительный результат (указывающий на наличие заболевания) в 70 % случаев, когда пациент действительно болен, и в 25 % случаев, когда у пациента нет этого заболевания. Пусть для конкретного пациента этот тест дал положительный результат. Какова вероятность того, что у него есть это заболевание?
43. Для сдачи экзамена студентам было необходимо подготовить 40 вопросов. Из 25 студентов 10 подготовили все вопросы, 8—35 вопросов, 5—30 вопросов и 2—20 вопросов. Вызванный студент ответил на предложенный вопрос. Какова вероятность того, что этот студент подготовил: а) все вопросы; б) только половину вопросов?

2. ПОСЛЕДОВАТЕЛЬНЫЕ НЕЗАВИСИМЫЕ ИСПЫТАНИЯ

2.1. Формула Бернулли

Повторными независимыми называются испытания, если:

— одно и то же испытание проводится конечное число раз;

— в каждом испытании может появиться или не появиться событие A;

— вероятность появления события A в каждом испытании не зависит от результатов других испытаний.

Примерами повторных испытаний являются многократное подбрасывание монеты или игрального кубика.

Если при выполнении n независимых повторных испытаний вероятность осуществления события A в каждом отдельном испытании ***постоянна P(A)=p=const*** (когда вероятность противоположного события $P(\overline{A})= 1 - p = q$), то вероятность появления ровно m раз события A в n испытаниях выражается формулой Бернулли:

$$P_n(m) = C_n^m \cdot p^m \cdot q^{n-m} = \frac{n!}{m!(n-m)!} \cdot p^m \cdot q^{n-m}, \qquad (2.1)$$

где $n!=1\cdot 2\cdot 3\cdot \ldots \cdot n$.

Рассмотрим частные случаи формулы Бернулли.

1. Вероятность осуществления события A в n испытаниях ровно n раз:

$$P_n(n) = C_n^n \cdot p^n \cdot q^{n-n} = \frac{n!}{n!\cdot 0!} \cdot p^n = p^n. \qquad (2.2)$$

2. Вероятность осуществления события A в n испытаниях ноль раз:

$$P_n(0) = \frac{n!}{0!\,n!} p^0 \cdot q^n = q^n. \qquad (2.3)$$

3. Вероятность осуществления события A в n испытаниях не более m_o раз:

$$P_n(m \le m_o) = P_n(0) + P_n(1) + P_n(2) + \ldots + P_n(m_o)$$

или

$$P_n(m \le m_o) = \sum_{i=0}^{m_o} P_n(i). \qquad (2.4)$$

4. Вероятность осуществления события A в n испытаниях не менее m раз

$$P_n(m \ge m_o) = P_n(m_o) + P_n(m_o + 1) + \ldots + P_n(n)$$

или $$P_n(m \geq m_o) = \sum_{i=m_o}^{n} P_n(i). \qquad (2.5)$$

Наиболее вероятное число m_0 наступления события A в n независимых испытаниях определяется из неравенства:

$$np-q \leq m_0 \leq np+p. \qquad (2.6)$$

В частности, если $np+p$ — не целое число, то m_0 равно целой части этого числа; если же $np+p$ — целое число, то m_0 имеет два значения $np-q$ и $np+p$.

Пример 1. Определить вероятность выиграть в шахматах (ничья исключена) у равносильного соперника:

— ровно три партии из четырех;

— не менее трех партий из четырех;

— хотя бы одну партию из четырех.

Решение. В данной задаче одно и то же испытание (партия игры в шахматы) проводится $n = 4$ раза, причем вероятности события A (выиграть партию) в каждом испытании одинаковы и равны $p = P(A) = 0,5$ (вследствие равносильности противников). Следовательно, это случай независимых повторных испытаний и для вычисления искомых вероятностей можно использовать формулу Бернулли.

• Вероятность выиграть ровно три ($m = 3$) партии из четырех:

$$P_4(m = 3) = P_4(3) = C_4^3 \cdot p^3 \cdot q^1 = \frac{4!}{3!(4-3)!} \cdot 0,5^3 \cdot 0,5^1 = 4 \cdot 0,0625 = 0,25.$$

• По формуле (2.5) имеем:

$$P_4(m \geq 3) = P_4(3) + P_4(4) = 0,25 + C_4^4 \cdot 0,5^4 \cdot (1-0,5)^0 =$$
$$= 0,25 + 1 \cdot 0,0625 \cdot 1 = 0,3125.$$

в) Обозначим случайные события: ($m \geq 1$) — выиграть хотя бы одну партию из четырех, ($m = 0$) — не выиграть ни одной партии из четырех. Тогда событие ($m = 0$) является противоположным событию ($m \geq 1$), поэтому

$$P_4(m \geq 1) = 1 - P_4(m = 0) = 1 - P_4(0) = 1 - C_4^0 \cdot 0,5^0 \cdot (1 - 0,5)^4 =$$
$$= 1 - \frac{4!}{0!4!} \cdot 1 \cdot 0,0625 = 1 - 0,0625 = 0,9375.$$

2.2. Локальная и интегральная теоремы Лапласа

При большом числе n повторных независимых испытаний пользоваться формулой Бернулли затруднительно, так как это связано с выполнением действий над большими числами. При больших n и m и не очень малых p, таких что $p>0,1$ и $n \cdot p \cdot q \geq 10$, хорошую точность дает формула, полученная Лапласом, которая получила название локальной теоремы Лапласа.

Теорема. Если вероятность появления некоторого события A в n испытаниях постоянна и отлична от **0** и **1**, то вероятность $P_n(m)$ того, что в n независимых испытаниях событие A появится ровно m раз, приближенно равна (тем точнее, чем больше n) значению функции

$$P_n(m) \approx \frac{1}{\sqrt{npq}} \cdot \varphi(x), \tag{2.7}$$

где $x = \dfrac{m - np}{\sqrt{npq}}$ и $\varphi(x) = \varphi(-x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$. (2.8)

Значения функции $\varphi(x)$ для различных x заранее вычислены и приводятся в таблицах.

Если необходимо определить вероятность того, что событие A может появиться в пределах от m_1 до m_2 раз, то используют интегральную теорему Лапласа. Условия ее применения те же, что и для локальной теоремы Лапласа. Формула вероятности того, что событие A появится от m_1 до m_2 раз, имеет вид:

$$P_n(m_1 \leq m \leq m_2) = \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-z^2/2} \cdot dz = \Phi(x_2) - \Phi(x_1), \tag{2.9}$$

где $x_1 = \dfrac{m_1 - np}{\sqrt{npq}}$; $x_2 = \dfrac{m_2 - np}{\sqrt{npq}}$

и $\Phi(x) = \int_0^x \varphi(t)dt = \dfrac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$. (2.10)

Функцию $\Phi(x)$ называют функцией Лапласа или интегралом вероятностей. Значения этого интеграла для различных x вычислены и приведены в таблицах, причем только для $x \geq 0$. Для нахождения $\Phi(x)$ для отрицательных значений $x<0$ пользуются той же таблицей, учитывая, что $\Phi(x)$ — нечетная функция, то есть $\Phi(-x) = — \Phi(x)$. Кроме того, в таблице приведены значения лишь до $x= 4$, так как для $x>4$ можно принять $\Phi(x) = 0,5$. Поэтому вычисление вероятности

сводится к расчету x_1 и x_2 и дальнейшему определению по таблице $\Phi(x_1)$ и $\Phi(x_2)$.

Пример 2. Всхожесть семян оценивается вероятностью ***0,8.*** Найти вероятность того, что из ***100*** высеянных семян взойдет: а) ровно ***90***; б) от ***76*** до ***90*** семян.

Решение

1. Пусть событие A — семя взошло. Рассматривая посев каждого семени как отдельное испытание, можно сказать, что проводится ***100*** независимых испытаний (в каждом из них событие A наступает с постоянной вероятностью $p = p(A) = 0{,}8$). По формуле Бернулли имеем:

$$P_{100}(90) = C_{100}^{90}(0{,}8)^{90} \cdot (1 - 0{,}8)^{10}.$$

Понятно, что непосредственный расчет по этой формуле окажется трудным. В данной задаче произведение npq равно:

$$npq = 100 \cdot 0{,}8 \cdot (1-0{,}8) = 100 \cdot 0{,}8 \cdot 0{,}2 = 16 > 10,$$

поэтому можно воспользоваться приближенной локальной формулой Лапласа:

$$P_{100}(90) \approx \frac{1}{\sqrt{16}} \cdot \varphi\left(\frac{90 - 100 \cdot 0{,}8}{\sqrt{16}}\right) = \frac{1}{4}\varphi(2{,}5).$$

По таблице значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$ найдем: $\varphi(2{,}5) = 0{,}0175$.

Тогда $P_{100}(90) \approx \frac{1}{4} \cdot 0{,}0175 \approx 0{,}00438$.

2. Обозначим как ($76 \le m \le 90$) событие, заключающееся в том, что число m взошедших семян заключено между 76 и 90. Если для вычисления вероятности этого события использовать формулу Бернулли, то придется считать следующую сумму вероятностей:

$$P_{100}(76 \le m \le 90) = \sum_{k=76}^{90} P_{100}(k) = \sum_{k=76}^{90} C_{100}^{k} \cdot (0{,}8)^k \cdot 0{,}2^{100-k}.$$

Так как $np=16>10$, то хорошую точность расчета искомой вероятности можно получить при использовании приближенной интегральной формулы Лапласа:

$$P_{100}(76 \le m \le 90) \approx \Phi\left(\frac{90 - 100 \cdot 0{,}8}{\sqrt{16}}\right) - \Phi\left(\frac{76 - 100 \cdot 0{,}8}{\sqrt{16}}\right) =$$

$$= \Phi(2{,}5) - \Phi(-1) = \Phi(2{,}5) + \Phi(1),$$

так как функция Лапласа нечетна и $\Phi(-1) = -\Phi(-1)$.

По таблице значений $\Phi(x)$ найдем: $\Phi(2,5) = 0,49379$; $\Phi(1) = 0,34134$.

Тогда $P_{100}(76 \le m \le 90) \approx 0,49379 + 0,34134 = 0,83513$.

2.3. Формула Пуассона

При больших n и очень малых вероятностях $p<0,1$ наступления события A в одном испытании («редкие» события) формула Лапласа дает большую погрешность, поэтому целесообразно в этих случаях использовать другую приближенную формулу, полученную Пуассоном:

$$P_{n,m} \approx \frac{\lambda^m \cdot e^{-\lambda}}{m!}, \qquad (2.11)$$

где $\lambda=np$ — среднее число появления события A в n испытаниях. Формулой Пуассона обычно пользуются при $p<0,1$ и $npq<10$.

Пример 3. Завод отправил в магазин 5000 изделий. Вероятность того, что в пути изделие повредится, равно $0,0004$. Найти вероятность того, что в пути повредятся: а) ровно 3 изделия; б) не более 2 изделий.

Решение

1. Рассматривая перевоз каждого изделия как отдельное испытание, можно утверждать, что производится $n=5000$ повторных испытаний. Событие A — повреждения изделия в пути. Так как вероятности наступления события A в каждом испытании одинаковы ($p = 0,0004$), то эти испытания независимы. Следовательно, для вычисления вероятности повреждения ровно 3 изделия в пути можно использовать формулу Бернулли:

$$P_{5000}(3) = C_{5000}^3 \cdot (0,0004)^3 \cdot (1 - 0,0004)^{4997}.$$

Прямой расчет вероятности по данной формуле достаточно сложен, поэтому воспользуемся приближенной формулой Пуассона для «редких» событий. Действительно $p = 0,0004 < 0,1$ и $npq = 5000 \cdot 0,0004 \cdot 0,9994 \approx 2 < 10$, поэтому $P_n(m) \approx \frac{\lambda^m \cdot e^{-\lambda}}{m!}$, где $\lambda = n \cdot p = 5000 \cdot 0,0004 = = 2$ — среднее число появления события A при 5000 испытаниях.

$$P_{5000}(3) \approx \frac{2^3 \cdot e^{-2}}{3!} = \frac{8 \cdot e^{-2}}{1 \cdot 2 \cdot 3} \approx 0,18.$$

2. Событие ($m \le 2$) является суммой трех несовместимых событий ($m = 0$), ($m = 1$) и ($m = 2$): ($m \le 2$) = ($m = 0$) +($m = 1$) + ($m = 2$). Тогда

$P(m \ge 3) = P(m = 0) + P(m = 1) + P(m = 2) = P_{5000}(0) + P_{5000}(1) + P_{5000}(2) \approx$

$$\approx \frac{2^0 \cdot e^{-2}}{0!} + \frac{2^1 \cdot e^{-2}}{1!} + \frac{2^2 \cdot e^{-2}}{2!} \approx e^{-2}(1+2+2) \approx 0{,}135 \cdot 5 \approx 0{,}677.$$

Задачи для самостоятельного решения

1. Известно, что из каждых 100 электролампочек 10 не отвечают требованиям стандарта. Какова вероятность того, что из четырех взятых наугад лампочек окажутся нестандартными: а) две лампочки; б) не более двух лампочек?
2. Дежурная аптека обслуживает 20 000 населения. Вероятность того, что в ночное время в аптеку придет посетитель, равна 0,0002. Какова вероятность того, что в ночное время в аптеку: а) придут не более 3 посетителей; б) придет хотя бы один посетитель?
3. На автобазе имеется 12 автомашин. Вероятность выхода на линию каждой из них равна 0,85. Найти вероятность нормальной работы автобазы в ближайший день, если для этого необходимо иметь на линии не меньше 8 автомашин.
4. Вероятность того, что изделие, изготовленное данным заводом, является бракованным, равна 0,02. Для контроля отобрано наудачу а) 10 изделий; б) 100 изделий. Найти вероятность того, что среди них окажется от 3 до 8 бракованных изделий.
5. Автоматическая телефонная станция обслуживает 400 абонентов. Для каждого абонента вероятность того, что он в течение часа позвонит на станцию, равна 0,01. Какова вероятность того, что в течение часа на станцию позвонят: а) 5 абонентов; б) не менее 3 абонентов?
6. Средняя плотность болезнетворных микробов в 1 $м^3$ воздуха равна 10. Сколько $м^3$ воздуха необходимо взять для пробы, чтобы с вероятностью, не меньшей 0,9, в ней обнаружить хотя бы один болезнетворный микроб?
7. Вероятность того, что покупателю требуется обувь 41 размера, равна 0,15. Найти вероятность, что среди первых 100 покупателей потребуют обувь 41 размера: а) 20 человек; б) не более 20 человек.
8. Наблюдениями установлено, что в некоторой местности в сентябре в среднем бывают 12 дождливых дней. Какова вероятность того, что из случайно взятых в этом месяце 8 дней окажутся дождливыми: а) 3 дня; б) не более 3 дней? (Предполагается, что вероятность быть дню дождливым не зависит от предшествующей погоды).
9. Школа должна принять в первые классы 200 детей. Определить вероятность того, что девочек среди них окажется: а) не менее 50 и не более 100; б) ровно 100, если вероятность рождения мальчика равна 0,515.

10. Лечение одного заболевания приводит к выздоровлению в 75 % случаев. Лечилось шесть больных. Какова вероятность того, что: а) выздоровят все шестеро; б) выздоровят не менее четырех больных?

3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

3.1. Дискретные случайные величины

Случайной называют величину, которая в результате испытания примет одно и только одно заранее неизвестное возможное значение. Это значение зависит от случайных обстоятельств и причин, которые заранее не могут быть учтены. Таким образом, по смыслу случайной называется величина, принимающая в результате опыта числовое значение, которое заранее предсказать невозможно.

Случайными величинами являются: показатели социально-экономической деятельности предприятий, биологических объектов, природной экосферы, например, продолжительность человеческой жизни, число листьев на дереве и площадь каждого листа.

Обозначают случайные величины большими латинскими буквами: ***X, Y, Z, …,*** а их возможные конкретные значения — малыми буквами: ***x; y; z …***

Вероятности случайных величин обозначают буквами с соответствующими индексами: $P(X=x_i)=p(x_i)=p_i$ или $q(Y=y_i)=q(y_i)=q_i$. Случайные величины делятся на дискретные и непрерывные.

Дискретной (прерывной) называют случайную величину, если она принимает некоторые определенные изолированные значения из данного конечного или бесконечного счетного множества значений.

Дискретная случайная величина (ДСВ) X полностью задана, если указаны все ее возможные значения x_i и вероятности их появления $p_i=P(X=x_i)$.

Законом распределения дискретной случайной величины X называется соотношение, устанавливающее связь между возможными значениями случайной величины x_i ($i=1, 2 \dots n$, где n — число возможных значений — может быть как конечным, так и бесконечным) и вероятностями $p_i=P(X=x_i)$ этих значений. Его удобно представлять в виде таблицы (ряда распределения):

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

Так как события $(X=x_1), (X=x_2), \dots, (X=x_n)$ несовместимы, и одно из них обязательно наступает, то они образуют полную группу событий. Следовательно, вероятности p_i должны удовлетворять условию нормировки:

$$\sum_{i=1}^{n} P_i = 1. \qquad (3.1)$$

Равенство (3.1) используется для проверки правильности составления закона распределения дискретной случайной величины.

Для придания ряду распределения более наглядного вида используют его графическое изображение: на оси абсцисс откладывают возможные значения случайной величины, а на оси ординат — вероятности этих значений. Полученные точки соединяют отрезками прямых. Такая фигура называется многоугольником распределения (рис. 3.1).

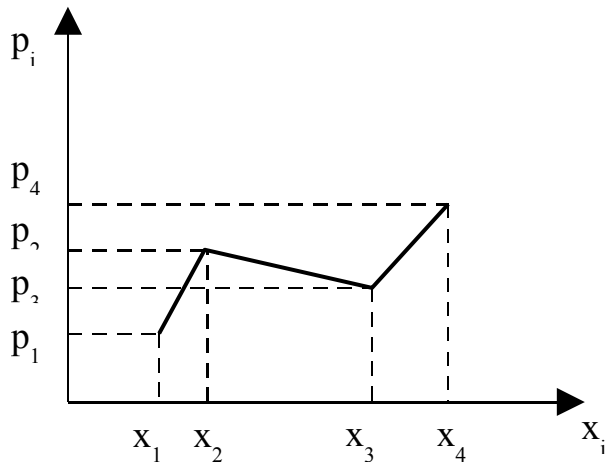


Рис. 3.1

Для количественной характеристики распределения вероятностей удобно воспользоваться не вероятностью события $X=x$, а вероятностью события $X<x$, где x — некоторая текущая переменная. Вероятность этого события зависит, очевидно, от x и поэтому есть некоторая функция от x. Эта функция называется функцией распределения случайной величины X и обозначается $F(x)$. Функцию распределения $F(x)$ называют также интегральной функцией распределения или интегральным законом распределения. Функция распределения является универсальной характеристикой случайных величин, как дискретных, так и непрерывных. Функция распределения полностью характеризует случайную величину с вероятностной точки зрения, то есть является одной из форм закона распределения.

$$F(x) = p(X<x) = p(-\infty<X<x). \qquad (3.2)$$

Для ДСВ X эту вероятность вычисляют как сумму вероятностей p_i появления значений величины X, меньших чем заданное x:

$$F(x) = \sum_{x_i < x} p_i \,. \qquad (3.3)$$

Графически функция распределения ДСВ X изображается в виде последовательности отрезков прямых (рис. 3.2), параллельных оси OX, с ординатами, вычисленными по формуле (3.3). Скачки функции $F(x)$

в точках x_1, x_2 … равны соответствующим вероятностям p_1, p_2, …, взятым из закона распределения.

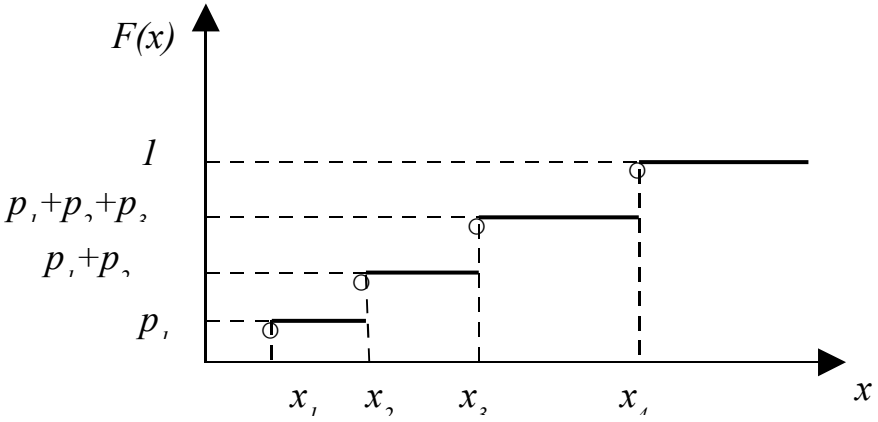


Рис. 3.2

3.1.1. Биномиальное распределение

Пусть проводятся n независимых испытаний, в каждом из которых вероятность появления случайного события A постоянна и равна p (следовательно, вероятность непоявления $q=1-p$). Рассмотрим дискретную случайную величину X — число появлений события A в этих n испытаниях. Ее возможными значениями являются числа **0, 1, 2, …, *n***, а соответствующие им вероятности вычисляются по формуле Бернулли:

$$P(X=m)=P_n(m)=C_n^m\cdot p^m\cdot q^{n-m}. \qquad (3.4)$$

В этом случае говорят, что случайная величина X имеет биномиальный закон распределения:

X	0	1	…	m	…	n
P	q^n	npq^{n-1}	…	$c^m{}_n\cdot p^m\cdot q^{n-m}$	…	p^n

Пример 1. Монета брошена два раза. Написать в виде таблицы закон распределения случайной величины X — числа выпадений «герба», найти функцию распределения и построить ее график.

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 появится. Таким образом, воз-

можные значения X таковы: $x_1=0,\ x_2=1,\ x_3=2.$ Найдем вероятность этих возможных значений по формуле Бернулли:

$$P_1 = P_2(0) = C_2^0 \cdot p^0 \cdot q^2 = 1 \cdot \left(\frac{1}{2}\right)^2 = 0{,}25;$$

$$P_2 = P_2(1) = C_2^1 \cdot p^1 \cdot q^1 = 2 \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = 0{,}5;$$

$$P_3 = P_2(2) = C_2^2 \cdot p^2 \cdot q^0 = 1 \cdot \left(\frac{1}{2}\right)^2 = 0{,}25.$$

Напишем искомый закон распределения:

x_i	0	1	2
p_i	0,25	0,5	0,25

Контроль:
$0{,}25+0{,}5+0{,}25=1.$

Используя данные таблицы и формулу (3.3), получим функцию $F(x)$ распределения:

$$F(x) = \begin{cases} 0, \text{ если } x \le 0; \\ 0{,}25, \text{ если } 0 < x \le 1; \\ 0{,}75, \text{ если } 1 < x \le 2; \\ 1, \text{ если } x > 2. \end{cases}$$

Построим график найденной функции распределения:

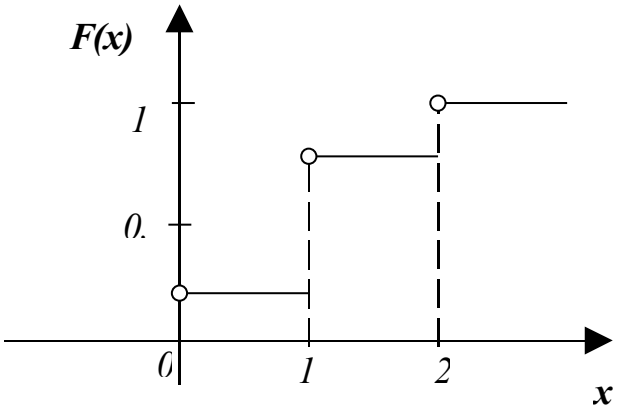


Рис. 3.3

Распределение Пуассона

Если возможными значениями ДСВ x являются **0, 1, 2, …, *m*, …,** а соответствующие им вероятности вычисляются по формуле Пуассона:

$$P(X = m) = \frac{\lambda^m \cdot e^{-\lambda}}{m!}, \qquad (3.5)$$

где $\lambda = const > 0$, то говорят, что эта величина имеет закон распределения Пуассона.

Распределение Пуассона является предельным для биномиального при $n \to \infty$ и $p \to 0$ так, чтобы их произведение np оставалось *постоянным*: $np = \lambda = const$, где λ — это среднее число появления события A в серии n испытаний, поэтому при массовых (n велико) и очень редких ($p < 0,1$ и $npq < 10$) событиях биномиальные вероятности вычисляют приближенно по формуле Пуассона.

3.1.3. Геометрическое распределение

Пусть проводятся независимые испытания, в каждом из которых вероятность появления события A равна p ($0 < p < 1$). Испытания заканчиваются, как только появится событие A. Таким образом, если событие
A появилось в k-м испытании, то в предшествующих $(k-1)$ испытаниях оно не появилось.

Обозначим X — число испытаний до появления события A. Возможные значения: $x_1 = 1;\ x_2 = 2...$ Если в первых $(k-1)$ испытаниях событие не появлялось, а в k-м появилось, то по теореме умножения независимых событий вероятность такой ситуации равна

$$P(X = k) = p_k = q^{k-1} \cdot p. \qquad (3.6)$$

Полагая $k = 1, 2\ldots$ в этой формуле, получим, что эти вероятности p_k образуют геометрическую прогрессию с первым членом p и знаменателем q $(0 < q < 1)$:

$$p,\ qp,\ q^2p,\ \ldots,\ q^{k-1}p,\ \ldots$$

Как известно, сумма всех членов геометрической прогрессии со знаменателем $q < 1$ равна:

$$p + qp + q^2p + \ldots + q^{k-1}p + \ldots = p/(1-q) = p/p = 1,$$

поэтому условие нормировки $\sum_{k=1}^{\infty} P_k = 1$ выполняется.

3.1.4. Гипергеометрическое распределение

Пусть из N изделий имеется M стандартных. Из партии случайно отбирают n изделий. Обозначим ДСВ X — число m стандартных изделий среди n отобранных. Ее возможные значения x: **0, 1, 2 … *min (M, n)***. Найдем вероятность того, что $X = m$, то есть что среди n отобранных изделий ровно m стандартных.

Так как отобранные изделия обратно в партию не возвращаются, то формула Бернулли здесь неприменима, поэтому для нахождения вероятности ***P(X=m)*** используем классическое определение вероятности.

Общее число исходов равно числу сочетаний C_N^n. Найдем число исходов, благоприятствующих событию ***X=m*** (среди ***n*** изделий ровно ***m*** стандартных). Такое (***m***) количество стандартных изделий из ***M*** стандартных можно выбрать C_M^m способами, при этом остальные (***n-m***) изделий должны быть нестандартными, взять же (***n-m***) нестандартных изделий из (***N-M***) нестандартных изделий можно C_{N-M}^{n-m} способами. Следовательно, число благоприятствующих исходов равно $C_M^m \cdot C_{N-M}^{n-m}$, так как согласно правилу произведений в комбинаторике: если объект ***A*** можно выбрать из совокупности объектов ***m*** способами и после каждого такого выбора объект ***B*** можно выбрать ***n*** способами, то пара объектов (***A,*** ***B***)
в указанном порядке может быть выбрана ***m·n*** способами.

Итак, вероятность ***(X=m)*** будет равна:

$$P(X=m)=\frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}. \qquad (3.7)$$

Формула (3.7) определяет распределение вероятностей, называемое *гипергеометрическим* распределением, оно определяется тремя параметрами: ***N, M, n***. Иногда в качестве параметров этого распределения рассматривают ***N***, ***n*** и ***p=M/N*** — вероятность того, что первое извлеченное изделие является стандартным. При ***n<0,1N*** гипергеометрическое распределение дает вероятности, близкие к вероятности, рассчитанной по формуле Бернулли, то есть по биномиальному закону.

3.2. Числовые характеристики дискретной случайной величины

Закон распределения или функция распределения полностью описывает дискретную случайную величину. Однако во многих случаях он неизвестен и тогда для описания случайной величины ***X*** вполне достаточно указать отдельные параметры, выражающие наиболее существенные характеристики ее распределения. Эти параметры называют числовыми характеристиками случайной величины.

Среди числовых характеристик случайных величин прежде всего нужно отметить те, которые характеризуют положение случайной величины на числовой оси, то есть указывают некоторое среднее, около которого группируется все возможные значения случайной величины. Например, среднее время работы лампы — ***100*** часов, в среднем ***50 %*** рабочих имеют стаж более ***4*** лет, наиболее часто в среднем встречаются

семьи с двумя детьми и т.п. Характеристиками положения СВ являются математическое ожидание, мода и медиана, среди которых важнейшую роль играет математическое ожидание случайной величины.

Математическим ожиданием (средним значением) дискретной случайной величины X называется сумма произведений всех ее возможных значений x_i на вероятности p_i этих значений:

$$M[X] = m_x = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i \ . \qquad (3.8)$$

Из определения М[х] следует, что оно является неслучайной (постоянной) величиной. Математическое ожидание случайной величины часто называют центром распределения.

Вероятностный смысл математического ожидания: при большом числе **N** испытаний среднее арифметическое наблюдаемых значений случайной величины близко к ее математическому ожиданию. Действительно, предположим, что в результате N независимых испытаний случайная величина значение x_1 приняла m_1 раз, значение x_2 – m_2 раз, …, значение x_n – m_n раз, причем $m_1+m_2+\ldots+m_n=N$. Среднее арифметическое всех значений, принятых случайной величиной, равно:

$$\overline{X} = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_n x_n}{N},$$

или

$$\overline{X} = \frac{m_1}{N} x_1 + \frac{m_2}{N} x_2 + \ldots + \frac{m_n}{N} x_n = \sum_{i=1}^{n} \frac{m_i}{N} x_i = \sum_{i=1}^{n} \omega_i x_i .$$

Поскольку относительная частота события при увеличении числа испытаний стремится к вероятности p_i появления значения x_i , то есть $\omega_i \approx p_i$ при больших N, то $\overline{X} \approx \sum_{i=1}^{n} x_i p_i = m_x$.

Замечание 1. Произведением постоянной величины C на дискретную случайную величину X называется дискретная случайная величина X, возможные значения которой равны произведению постоянной C на возможные значения X, а вероятность возможных значений $C \cdot X$ равна вероятностям соответствующих

$$\overline{X} \approx \sum_{i=1}^{n} x_i p_i = m_x \ .$$

Замечание 2. Случайные величины называются независимыми, если закон распределения каждой из них не зависит от того, какое значение приняла другая случайная величина (в противном случае случайные величины называются зависимыми).

Для независимых случайных величин X и Y для любых x и y события ($X<x$) и ($Y< y$) являются независимыми, поэтому

$$P\{(X < x) \cdot (Y < y)\} = P(X < x) \cdot P(Y < y) = F_1(x) \cdot F_2(y).$$

Аналогично случайные величины $X_1, X_2 \ldots X_n$ называют взаимно независимыми, если для любой группы этих случайных величин $X_{i1}, X_{i2} \ldots X_{ik}$ $(k \le n)$ и любых $x_{i1}, x_{i2} \ldots x_{ik}$ события $(X_{i1} < x_{i1})$, $(X_{i2} < x_{i2})$ … $(X_{ik} < x_{ik})$ — независимые.

Замечание 3. Произведением независимых случайных величин X и Y называется случайная величина $X \cdot Y$, возможные значения которой равны произведениям каждого возможного значения X на каждое возможное значение Y, а вероятности возможных значений произведений $X \cdot Y$ равны произведениям вероятностей возможных значений сомножителей.

Если некоторые произведения $x_i y_i$ оказываются равными между собой, то вероятность возможного значения произведения равна сумме соответствующих вероятностей.

Замечание 4. Сумма двух случайных величин X и Y есть случайная величина $Z=X+Y$, возможные значения которой равны суммам каждого возможного значения X с каждым возможным значением Y, а вероятности возможных значений $X+Y$ равны произведению вероятности одного слагаемого на условную вероятность второго.

Если некоторые суммы $x_i + y_i$ оказываются равными между собой, то вероятность возможного значения такой суммы равна сумме соответствующих вероятностей.

Свойства математического ожидания

1. Математическое ожидание постоянной величины $C = const$ равно этой постоянной:

$$M[C]=C. \qquad (3.9)$$

Доказательство. Постоянную можно рассматривать как дискретную случайную величину, принимающую лишь одно значение C с вероятностью $p=1$, поэтому $M[C]=C \cdot 1=C$.

2. Постоянный множитель C можно выносить за знак математического ожидания:

$$M[C \cdot X]=C \cdot M[X]. \qquad (3.10)$$

Доказательство. По формуле (3.8) для $M[X]$ находим:

$$M[C \cdot X]=Cx_1p_1+Cx_2p_2+\ldots+Cx_np_n=C(x_1p_1+x_2p_2+\ldots+x_np_n)=C \cdot M[X].$$

3. Математическое ожидание алгебраической суммы случайных величин равно алгебраической сумме их математических ожиданий:

$$M[X \pm Y]=M[X] \pm M[Y]. \qquad (3.11)$$

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

$$M[X \cdot Y]=M[X] \cdot M[Y]. \qquad (3.12)$$

5. Математическое ожидание отклонения $(x-m_x)$ значения случайной величины от ее математического ожидания равно нулю:

$$M[(X-m_x)]=0. \qquad (3.13)$$

Модой M_0 дискретной случайной величины X называется ее наиболее вероятное значение, то есть такое значение x_m, что предшествующее и последующее за ним значения имеют вероятности, меньшие $P(x_m)$. Распределения могут иметь одну или несколько мод или не одной моды.

Медиана M_e определяется для непрерывной случайной величины, это такое ее значение M_e, для которого

$$P(X < M_e) = P(X > M_e) = 0,5 \text{ или } F(M_e) = 0,5.$$

Кроме характеристик положения, для описания свойств распределения используются и характеристики рассеяния (дисперсия, средне-
квадратичное отклонение, различные моменты распределения порядка выше первого и др.).

Начальным моментом S-го порядка распределения дискретной случайной величины X называется математическое ожидание s-й степени этой случайной величины:

$$\alpha_s[X] = M[X^s] = \sum_{i=1}^{n} x_i^s \cdot p_i \,. \qquad (3.14)$$

Очевидно, что первый начальный момент случайной величины представляет собой ее математическое ожидание.

Центрированной случайной величиной $\overset{0}{X}$, соответствующей величине X, называется отклонение случайной величины X от ее математического ожидания:

$$\overset{0}{X} = X - m_x \,. \qquad (3.15)$$

Математическое ожидание центрированной случайной величины согласно свойству **5** математического ожидания равно нулю. Центрирование случайной величины равносильно переносу начала координат в среднюю, центральную точку, абсцисса которой равна математическому ожиданию.

Моменты центрированной случайной величины называются центральными моментами.

Центральным моментом порядка S распределения случайной величины X называется математическое ожидание S-й степени соответствующей центрированной случайной величины:

$$\mu_s(X) = M[\overset{o}{(X)}^S] = M[(X - m_x)^s] \,. \qquad (3.16)$$

Для дискретной случайной величины центральный момент равен:

$$\mu_s = \sum_{i=1}^{n} (x_i - m_x)^s \cdot p_i . \qquad (3.17)$$

Второй центральный момент называется дисперсией случайной величины, то есть дисперсия дискретной случайной величины X равна математическому ожиданию квадрата отклонения случайной величины от ее математического ожидания:

$$D[X] = \sigma_x^2 = M[(X - m_x)^2] . \qquad (3.18)$$

Более удобна для вычисления дисперсии следующая формула:

$$D[X] = M[X^2] - m_x^2 . \qquad (3.19)$$

Для ДСВ X дисперсия вычисляется по формуле:

$$D[X] = \sum_{i=1}^{n} (x_i - m_x)^2 \cdot p_i \qquad (3.20)$$

или

$$D[X] = \sum_{i=1}^{n} x_i^2 \cdot p_i - m_x^2 . \qquad (3.21)$$

Свойства дисперсии

- Дисперсия постоянной величины равна нулю: $D[C]=0$, где $C=const$.
- Постоянный множитель можно выносить за знак дисперсии, предварительно возведя его в квадрат: $D[CX]=C^2 \cdot D[X]$.
- $D[X \pm Y]=D[X]+D[Y]$, если X и Y — независимые случайные величины.

Средним квадратическим отклонением σ_x дискретной случайной величины X называется корень квадратный из дисперсии:

$$\sigma_x = \sqrt{D[X]} .$$

Дисперсия и среднее квадратическое отклонение характеризуют степень разбросанности значений СВ относительно математического ожидания (величину σ_x иногда называют стандартным отклонением).

Третий центральный момент служит для характеристики асимметрии распределения, так как для симметричных распределений он равен нулю.

Замечание. Основные числовые характеристики дискретной случайной величины, распределенной по а) биномиальному закону: $M[X]=np$, $D[X]=npq$; б) закону Пуассона: $M[X]=D[X]=\lambda$.

Пример 2. По заданному закону распределения дискретной случайной величины X (число вызовов «скорой помощи» в час) найти: а) ее математическое ожидание $M[X]$, дисперсию $D[X]$ (двумя способами) и среднее квадратичное отклонение; б) вероятность события $X \le M[X]$.

x_i	10	15	20	25
p_i	0,4	0,3	0,1	0,2

Решение

1. Математическое ожидание ДСВ X равно:

$$M[X] = \sum_{i=1}^{4} x_i p_i = 10 \cdot 0,4 + 15 \cdot 0,3 + 20 \cdot 0,1 + 25 \cdot 0,2 = 4 + 4,5 + 2 + 5 = 15,5.$$

Дисперсию $D[x]$ вычислим двумя способами:

— по формуле:

$$D[X] = M[(X - m_x)^2] = \sum_{i=1}^{4} (x_i - m_x)^2 \cdot p_i = (10 - 15,5)^2 \cdot 0,4 + (15 - 15,5)^2 \cdot 0,3 +$$

$$+ (20 - 15,5)^2 \cdot 0,1 + (25 - 15,5)^2 \cdot 0,2 = 12,1 + 0,075 + 2,025 + 28,05 = 32,25;$$

— по формуле:

$$D[X] = M[X^2] - m_x^2 = \sum_{i=1}^{4} x_i^2 p_i - m_x^2 = (10^2 \cdot 0,4 + 15^2 \cdot 0,3 + 20^2 \cdot 0,1 +$$

$$+ 25^2 \cdot 0,2) - 15,5^2 = (40 + 6,5 + 40 + 125) - 240,25 = 275,5 - 240,25 = 32,25.$$

Среднее квадратичное отклонение $\sigma_x = \sqrt{D[X]} = \sqrt{32,25} \approx 5,68$.

2. Событие ($X \le M[X]$), то есть в нашем случае ($X \le 15,5$) — сложное, равное сумме несовместимых событий ($X = 10$) ={случайная величина X приняла значение 10} и ($X = 15$), поэтому:

$$P(X \le 15,5) = P\{(X = 10) + (X = 15)\} = P(X = 10) + P(X = 15) = 0,4 + 0,3 = 0,7.$$

Пример 3. В некотором городе, по оценкам, происходит в среднем одно рождение в час. Какова вероятность того, что за данный час: а) не произойдет ни одного рождения; б) произойдет более двух рождений?

Решение. Дискретная случайная величина X — число рождений в час распределена по закону Пуассона с $m_x=\lambda=1=const$. Следовательно, по формуле (3.5) будем иметь:

а) $P(X=0)=\dfrac{1^0\cdot e^{-1}}{0!}=e^{-1}\approx 0{,}368;$

б) $P(X>2)=1-P(X\le 2)=1-[P(X=0)+P(X=1)+P(X=2)]=$

$$=1-\left(0{,}368+\frac{1^1\cdot e^{-1}}{1!}+\frac{1^2\cdot e^{-1}}{2!}\right)=1-(0{,}368+0{,}368+0{,}184)=0{,}08.$$

Задачи для самостоятельного решения

1. Вероятность полного выздоровления при лечении некоторым препаратом равна ***0,8.*** Составить закон распределения числа полностью выздоровевших из ***3*** больных, составить многоугольник распределения, построить график функции распределения вероятностей ***F(x).***
2. Некоторый эксперимент завершается успехом в ***75 %*** всех попыток, причем в день можно провести только ***4*** таких эксперимента. Найти закон распределения случайной величины: а) ***X*** — числа успешных экспериментов в день; б) ***Y*** — числа проведенных экспериментов в день, если они проводятся только до первого успешного.
3. Вероятность порчи при хранении в течение года одной единицы продукции равна ***0,001***. Составить закон распределения числа испорченных единиц продукции после года хранения среди ***5000*** имеющихся и найти вероятности того, что повреждено: а) менее ***3*** единиц; б) более ***2*** единиц; в) хотя бы одна единица продукции.
4. В течение часа станция «скорой помощи» получает в среднем ***40*** вызовов. Какова вероятность того, что не будет вызова в течение: а) ***30*** сек.; б) ***2*** мин.?
5. Из ***25*** проверенных налоговой полицией предприятий у ***6*** имеются нарушения в оплате налогов. Через ***3*** месяца для повторной проверки случайным образом выбираются ***3*** предприятия. Составить закон распределения числа предприятий из этих трех, у которых ранее были нарушения в оплате налогов, построить график функции распределения вероятностей ***F(x)***.
6. В рекламной лотерее на каждые ***100*** билетов приходится ***15*** денежных выигрышей: ***1*** выигрыш — по ***50*** руб., ***4*** — по ***10*** руб. и ***10*** —

   по ***5*** руб. Составить закон распределения дискретной случайной величины ***X*** — размер выигрыша; найти средний размер выигрыша и его среднее квадратичное отклонение.
7. По заданному закону распределения дискретной случайной величины ***X*** (число вызовов «скорой помощи» в час) найти: 1) ее математиче-

ское ожидание ***M [X]***, дисперсию ***D [X]*** (двумя способами) и среднее квадратичное отклонение; 2) вероятность события ($X \le M[X]$) .

x_i	10	15	20	25
p_i	0,1	0,3	0,4	0,2

8. По известному закону распределения дискретной случайной величины ***X*** (число стандартных деталей в партии деталей, изготовленных станком-автоматом) найти математическое ожидание **M[X]**, дисперсию **D[X],** среднее квадратичное отклонение σ_x и вероятность события ($|X - M[X]| \le \sigma_x$).

x_i	10	15	20	25
p_i	0,1	0,1	0,2	0,6

а

x_i	2	5	8	10
p_i	0,5	0,2	0,2	0,1

б

9. Сравнить математические ожидания и дисперсии двух дискретных случайных величин ***X*** и ***Y***, заданных законами распределения:

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

y_i	-1	0	1
p_i	0,4	0,5	0,1

4. НЕПРЕРЫВНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ (НСВ)

4.1. Функция распределения и плотность распределения НСВ

Непрерывной называют случайную величину, если она может принимать все значения из некоторого конечного или бесконечного интервала, или если ее функция распределения ***F(x)*** непрерывна на всей числовой оси.

Закон распределения непрерывной случайной величины невозможно описать с помощью таблицы, в которой были бы перечислены все возможные значения этой величины и их вероятности. Поэтому для количественной характеристики распределения вероятностей НСВ пользуются функцией распределения случайной величины, равной (см. 3.2): $F(x)=P(X<x)=P(-\infty<X<x)$.

Геометрически равенство $F(x)=P(X<x)$ можно представить как вероятность того, что случайная величина X (случайная точка на оси OX)
в результате опыта попадет левее точки x (см. рис. 4.1).

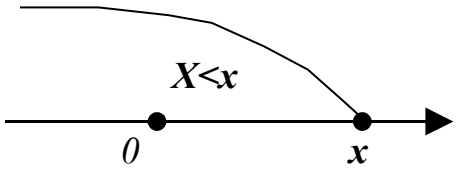


Рис. 4.1

Свойства функции распределения F(x):

• $0\le F(x)\le 1$ — это свойство следует из определения функции распределения, как вероятности $(0\le P\le 1)$.

• $F(x_2)\ge F(x_1)$, если $x_2>x_1$; то есть $F(x)$ — неубывающая функция.

Доказательство. Пусть $x_2>x_1$, тогда событие $X<x_2$ можно представить в виде суммы двух несовместимых событий $X<x_1$ и $x_1\le X<x_2$. По теореме сложения несовместимых событий получим:

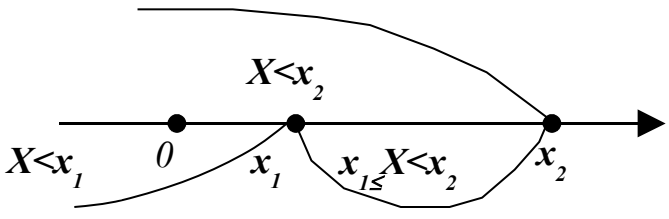


Рис. 4.2

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

или

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

Так как вероятность любого события есть число неотрицательное, то есть $P(x_1 \le X < x_2) \ge 0$, то, следовательно, $F(x_2) \ge F(x_1)$.

Следствие 1. Вероятность того, что непрерывная случайная величина примет какое-либо значение внутри интервала (x_1, x_2), равна разности значений функции распределения для границ этого интервала

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1). \qquad (4.1)$$

Следствие 2. Вероятность того, что непрерывная случайная величина примет какое-либо заранее заданное значение, равна *нулю*: из формулы (4.1) для вероятности попадания в интервал имеем при $x_1 = x_2 = x$:

$$P(X=x) = F(x_2) - F(x_1) = F(x) - F(x) = 0. \qquad (4.2)$$

Следствие 3. $\lim\limits_{x \to -\infty} F(x) = F(-\infty) = 0; \quad \lim\limits_{x \to +\infty} F(x) = F(+\infty) = 1.$

График функции распределения, как следует из ее свойств, расположен в полосе, ограниченной прямыми $y=0$ и $y=1$. График функции произвольного распределения представлен на рис. 4.3.

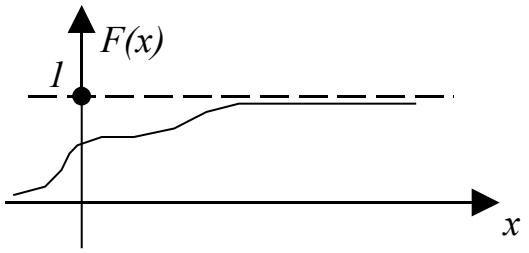


Рис. 4.3

Непрерывную случайную величину можно задать и с помощью другой функции, называемой плотностью распределения.

Плотностью распределения вероятностей $f(x)$ непрерывной случайной величины X называется функция, определяемая как первая производная от функции распределения:

$$f(x) = F'(x). \qquad (4.3)$$

Свойства плотности распределения f(x):

- $f(x) \ge 0$, как производная неубывающей функции.
- Функция распределения может быть найдена по известной плотности распределения по формуле:

$$F(x) = \int_{-\infty}^{x} f(x)dx. \qquad (4.4)$$

• Вероятность того, что непрерывная случайная величина ***X*** примет некоторое значение в интервале (***a, b***), равна определенному интегралу от плотности распределения, взятому в пределах от ***a*** до ***b***:

$$P(a < x < b) = \int_a^b f(x)dx. \qquad (4.5)$$

• Интеграл от плотности распределения в интервале от -∞ до +∞ равен 1 (условие нормировки):

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \qquad (4.6)$$

• Если все значения случайной величины ***X*** принадлежат интегралу (***a; b***), то вероятность ее попадания в интервал (***a; b***) равна единице, то есть

$$\int_a^b f(x)dx = P(a < x < b) = 1. \qquad (4.7)$$

• Вероятностный смысл плотности распределения

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

Следовательно, плотность распределения приближенно равна вероятности попадания НСВ в единичный интервал при данном значении величины ***x***.

$$f(x) \approx \frac{P(X+\Delta x < X < x)}{\Delta x}. \qquad (4.8)$$

• График плотности функции распределения в соответствии со свойством 1 расположен над осью OX.

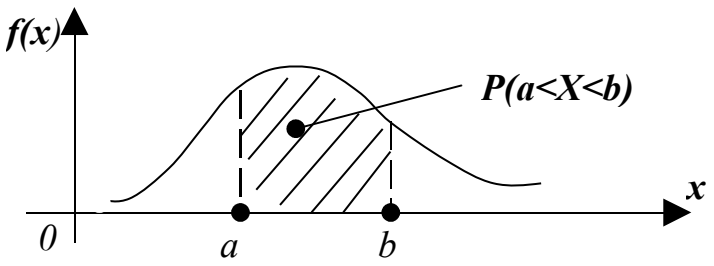


Рис. 4.4

Геометрический смысл плотности вероятности: вероятность попадания в интервал (***a; b***) равна площади криволинейной трапеции под кривой плотности распределения в пределах от ***a*** до ***b*** (рис. 4.4).

Числовые характеристики непрерывной случайной величины X с плотностью распределения $f(x)$ имеет смысл, аналогичный числовым характеристикам дискретных случайных величин.

Математическое ожидание непрерывной случайной величины равно:

$$M[X] = m_x = \int_{-\infty}^{+\infty} x \cdot f(x)dx . \qquad (4.9)$$

Дисперсия непрерывной случайной величины равна:

$$D[X] = \sigma_x^2 = \int_{-\infty}^{+\infty} (X - m_x)^2 \cdot f(x)dx \text{ или } D[X] = \sigma_x^2 = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - m_x^2 . \qquad (4.10)$$

Среднее квадратическое отклонение непрерывной случайной величины равно:

$$\sigma_x = \sqrt{D[X]} . \qquad (4.11)$$

Свойства числовых характеристик НСВ такие же, как и для ДСВ.

Модой M_0 дискретной случайной величины X называется ее наиболее вероятное значение, то есть точка максимума плотности распределения $f(x)$.

Медиана M_e — это такое значение M_e непрерывной случайной величины, для которого

$$P(X < M_e) = P(X > M_e) = 0,5 \text{ или } \int_{-\infty}^{M_e} f(x)dx = F(M_e) = 0,5 .$$

Пример 1. Найти математическое ожидание и дисперсию случайной величины X, заданной функцией распределения $F(x)$:

$$F(x) = \begin{cases} 0 & \text{при } x \le 0; \\ 1/4x^2 & \text{при } 0 < x \le 2; \\ 1 & \text{при } x > 2. \end{cases}$$

Решение. Найдем плотность распределения:

$$f(x) = F'(x) = \begin{cases} 0 & \text{при } x \le 0; \\ 1/2x & \text{при } x < x \le 2; \\ 0 & \text{при } x > 2. \end{cases}$$

Найдем математическое ожидание по формуле (4.9):

$$m_x = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_0^2 x \cdot \frac{x}{2} dx = \left. \frac{x^3}{6} \right|_0^2 = \frac{4}{3}.$$

Найдем дисперсию по формуле (4.10):

$$D[x] = \int_0^2 x^2 \cdot \frac{x}{2} \cdot dx - \left(\frac{3}{4}\right)^2 = \frac{x^4}{8}\Big|_0^2 - \frac{16}{9} = 2 - \frac{16}{9} = \frac{2}{9}.$$

4.2. Основные виды распределений непрерывных случайных величин

4.2.1. Равномерное распределение

Закон распределения называется равномерным, если на отрезке [***a; b***], которому принадлежат все возможные значения случайной величины, плотность распределения сохраняет постоянное значение. Равномерное распределение имеет, например, ошибка округления при измерении тех или иных физических величин или вычислениях.

Из условия (4.7) имеем: $\int_a^b c \cdot dx = c(b-a) = 1$. Отсюда $c = \frac{1}{b-a}$ на отрезке [***a, b***]. Поэтому равномерная плотность распределения имеет вид

$$f(x) = \begin{cases} 0; & x < a; \\ \dfrac{1}{b-a}; & a \le x \le b; \\ 0; & x > b. \end{cases} \qquad (4.12)$$

График плотности распределения равномерно распределенной случайной величины приведен на рис. 4.5а.

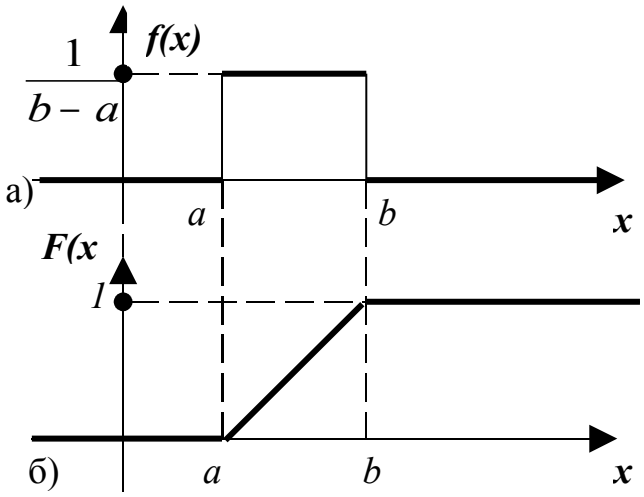


Рис. 4.5

Функция распределения для равномерного закона при $a \le x \le b$ равна:

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

В общем виде:

$$F(x) = \begin{cases} 0; & x < a; \\ \dfrac{x-a}{b-a}; & a \le x \le b; \\ 1; & x > b. \end{cases} \qquad (4.13)$$

График функции $F(x)$ приведен на рис. 4.5б. Вероятность попадания в интервал $(x_1; x_2) \in [a;b]$ равна:

$$P(x_1 < x < x_2) = \frac{x_2 - x_1}{b-a}. \qquad (4.14)$$

Математическое ожидание равно:

$$m_x = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a} \cdot dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}. \qquad (4.15)$$

Дисперсия равномерного распределения равна:

$$D[X] = \int_a^b \frac{1}{b-a} \cdot (x - m_x)^2 dx = \frac{(b-a)^2}{12}. \qquad (4.16)$$

Среднеквадратическое отклонение имеет вид:

$$\sigma_x = \frac{b-a}{2\sqrt{3}}. \qquad (4.17)$$

Пример 2. Шкала рычажных весов, установленных в лаборатории, имеет цену деления ***1*** г. При измерении массы вещества отсчет делается с точностью до целого деления с округлением в ближайшую сторону. Поэтому случайная ошибка округления X будет распределена по равномерному закону в интервале ***[-0,5г; 0,5г].*** Какова вероятность того, что ошибка округления по абсолютной величине не превысит ***0,2*** г?

Решение. Плотность распределения данной случайной величины X-ошибки округления согласно формуле (4.12) будет иметь вид:

$$f(x) = \begin{cases} 0, \text{ если } & x \le -0,5; \\ \dfrac{1}{0,5-(-0,5)} = 1, \text{ если } & -0,5 < x \le 0,5; \\ 0, \text{ если } & x > 0,5. \end{cases}$$

Тогда искомая вероятность будет равна:

$$P(|X| \le 0,2) = P(-0,2 < X < 0,2) = \int_{-0,2}^{0,2} f(x)dx = \int_{-0,2}^{0,2} 1 \cdot dx = x\Big|_{-0,2}^{0,2} = 0,4.$$

4.2.2. Экспоненциальное распределение

Случайная величина называется распределенной по экспоненциальному закону, если ее плотность распределения имеет вид:

$f(x) = \lambda \cdot e^{-\lambda x}$ для $x \ge 0$ (рис. 4.6а),

где $\lambda = const > 0$ — параметр.

Функция распределения $F(x)$ (рис. 4.6б), имеет вид:

$$F(x) = 1 - e^{-\lambda x};\ x \ge 0.$$

Экспоненциальное распределение часто встречается в теории массового обслуживания (например, X — время ожидания обслуживания или длительность обслуживания) и в теории надежности (например,
X — срок службы прибора, микросхемы).

Для экспоненциального распределения математическое ожидание и среднеквадратическое отклонение равны:

$$m_x = \sigma_x = 1/\lambda\ , \qquad (4.18)$$

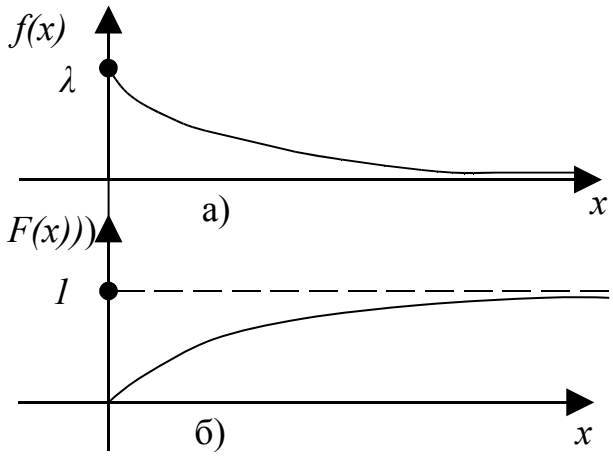


Рис. 4.6

так как: $m_x = \lambda \int_0^{\infty} x \cdot e^{-\lambda x} \cdot dx = \frac{1}{\lambda};$

$$\sigma_x = \sqrt{\lambda \int_0^{\infty} (x - m_x)^2 e^{-\lambda x} \cdot dx} = \frac{1}{\lambda}.$$

Пример 2. Продолжительность жизни растений данного вида в определенной среде представляет собой НСВ ***X*** с плотностью распределения

$$f(x) = \frac{1}{120} \cdot e^{-x/120}, \quad (x \geq 0).$$

- Найти функцию распределения ***F(x).***
- Вычислить вероятность того, что:

— данное растение проживет не более ***120*** дней;

— растение, прожившее ***120*** дней, умрет в течение ***60*** следующих дней;

— данное растение проживет более ***120*** дней.

Решение

1. Найдем функцию распределения ***F(x)*** (при ***x≥0***):

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_0^x \frac{1}{120} \cdot e^{-x/120} dx = -e^{-x/120}\Big|_0^x = 1 - e^{-x/120}.$$

2. а) $P(X \leq 120) = F(120) = 1 - e^{-120/120} = 1 - e^{-1} \approx 0{,}6321.$

б) Вероятность попадания НСВ ***X*** в интервал (***120;180***) можно вычислить двумя способами:

— по известной плотности распределения ***f(x)*** по формуле (4.5):

$$P(120 < X < 180) = \int_{120}^{180} \frac{1}{120} \cdot e^{-x/120} \cdot dx = -e^{-x/120}\Big|_{120}^{180} =$$

$$= -e^{-120/120} - e^{-180/120} = e^{-1} - e^{-1,5} \approx 0{,}3679 - 0{,}2231 = 0{,}1448;$$

— по известной функции распределения ***F(x)*** по формуле (4.1):

$$P(120 < X < 180) = F(180) - F(120) = (1 - e^{-180/120}) - (1 - e^{-120/120}) = e^{-1} - e^{-1,5} = 0{,}1448.$$

Результат, как видно, один и тот же.

в) Так как событие (**x>120**) является противоположным событию (**x≤120**), то

$$P(X > 120) = 1 - P(x \leq 120) = 1 - F(120) = 1 - 1 - e^{-120/120} = e^{-1} \approx 0{,}3679.$$

4.2.3. Нормальный закон распределения (закон Гаусса)

Нормально распределенные случайные величины широко распространены на практике. Объясняется это тем, что если а) на формирование значения величины оказывает влияние очень большое число случайных факторов и б) влияние каждого из этих факторов

незначительно по сравнению с влиянием всех факторов вместе, то, как показывает опыт, такая величина будет иметь приближенно нормальное распределение.

Например, на результат измерения физической величины влияет огромное количество случайных факторов: колебание атмосферных условий, сотрясения измерительного прибора, усталость наблюдателя и т.п. Каждый из этих факторов, взятый в отдельности, порождает ничтожную x_i ошибку в результате измерения. Общая ошибка X будет, следовательно, суммой большого числа малых случайных величин x_i, поэтому можно уверенно заключить, что случайная ошибка X будет иметь закон распределения, близкий к нормальному.

Другой пример: при массовом производстве каких-либо изделий вследствие малых случайных отклонений от нормы при проведении большого числа операций характеристики однотипных изделий (размеры, масса, объем и др.) будут иметь приближенно нормальное распределение. Поэтому в теории вероятностей нормальный закон распределения имеет фундаментальное значение. Плотность нормального распределения имеет вид:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \cdot e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} = \frac{1}{\sigma_x \sqrt{2\pi}} \cdot exp(-\frac{(x-m)^2}{2\sigma^2}), \qquad (4.19)$$

где параметры m и σ совпадают с характеристиками распределения:

$$m = m_x,\ \sigma = \sigma_x.$$

Свойства плотности нормального распределения:

- Функция $f(x)$ определена на всей оси, то есть — $-\infty < x < +\infty$.
- $f(x) > 0$, то есть нормальная кривая расположена над осью OX.
- Предел функции при $|x| \to \infty$ равен 0:

$$\lim_{|x| \to \infty} f(x) = 0,$$

то есть ось OX служит горизонтальной асимптотой графика.

- Исследуем $f(x)$ на экстремум.

$$f'(x) = -\frac{x - m_x}{\sigma_x^3 \sqrt{2\pi}} e^{-(x-m_x)^2/2\sigma_x^2} = 0.$$

Максимум имеет место при $x=m_x$; так как при $x<m_x$ производная $f'(x)>0$ и при $x>m_x$ $f'(x) <0$.

Величина максимума равна: $\frac{1}{\sigma_x \sqrt{2\pi}}$.

- Разность $(x-m_x)$ в формуле для $f(x)$ — в квадрате, поэтому график $f(x)$ симметричен относительно прямой $x= m_x$.

• Точки перегиба определим из условия $f''=0$.

$$f''(x) = -\frac{1}{\sigma_x^3\sqrt{2\pi}} e^{-(x-m_x)^2/2\sigma_x^2} \cdot \left[1 - \frac{(x-m_x)^2}{\sigma_x^2}\right].$$

При $x=m_x+\sigma_x$ и $x=m_x-\sigma_x$ вторая производная равна нулю и при переходе через эти точки она меняет знак, следовательно, эти точки — точки перегиба. В обеих этих точках значение функции равно $1/(\sigma_x\sqrt{2\pi \cdot e})$.

Графики плотности нормального распределения (нормальные кривые) при различных значениях σ_x представлены на рис. 4.7. Среднеквадратическое отклонение σ_x определяет форму кривой, при его уменьшении величина максимума увеличивается, при увеличении — уменьшается. Математическое ожидание m_x определяет положение нормальной кривой на числовой оси x. При уменьшении математического ожидания график нормальной плотности сдвигается влево, при увеличении математического ожидания график сдвигается вправо.

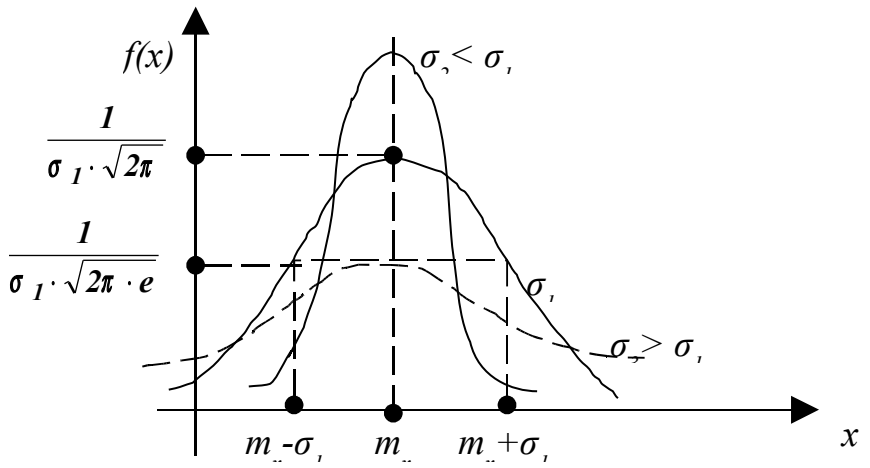


Рис. 4.7. Кривые нормального распределения

Вероятность попадания нормально распределенной случайной величины в интервал от x_1 до x_2 равна:

$$P(x_1 < X < x_2) = \frac{1}{\sigma_x\sqrt{2\pi}} \int_{x_1}^{x_2} \cdot exp(-\frac{(x-m)^2}{2\sigma_x^2}) \cdot dx. \qquad (4.22)$$

Последнюю формулу преобразуем, чтобы можно было воспользоваться готовыми таблицами. Для этого введем новую переменную $t = \frac{x-m_x}{\sigma_x}$ в определенном интеграле. Тогда $dt = dx / \sigma_x$, а пределы

интегрирования для переменной t будут: нижний — $t_1 = \frac{x_1 - m_x}{\sigma_x}$, верхний — $t_2 = \frac{x_2 - m_x}{\sigma_x}$. В результате получим:

$$P(x_1 < X < x_2) = \int_{t_1}^{t_2} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}} \int_{t_1}^{0} exp\left(-\frac{t^2}{2}\right) dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{t_2} exp\left(-\frac{t^2}{2}\right) dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{t_2} exp\left(-\frac{t^2}{2}\right) dt - \frac{1}{\sqrt{2\pi}} \int_{0}^{t_1} exp\left(-\frac{t^2}{2}\right) dt = \Phi(t_2) - \Phi(t_1),$$

где $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} exp\left(-\frac{t^2}{2}\right) dt$ — функция Лапласа (см. 2.10), значения которой сведены в таблицу. Таким образом,

$$P(x_1 < X < x_2) = \Phi(t_2) - \Phi(t_1) = \Phi\left(\frac{x_2 - m_x}{\sigma_x}\right) - \Phi\left(\frac{x_1 - m_x}{\sigma_x}\right). \qquad (4.23)$$

Замечание 1. Случайная величина T называется *нормированной*, если ее математическое ожидание равно нулю, а дисперсия — единице:

$$M[T] = 0, \; D[T] = 1.$$

От любой случайной величины X можно перейти к нормированной случайной величине T с помощью линейного преобразования:

$$T = \frac{X - m_x}{\sigma_x}.$$

Для нормированной нормальной случайной величины ($m_t=0$, $\sigma_t=1$) плотность распределения $f(t)$ имеет вид:

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}},$$

а функция распределения равна интегралу: $F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-t^2/2} \cdot dt$.

Очевидно, что $F(t)=\Phi(t)+0,5$, где $\Phi(t)$ — функция Лапласа.

Пользуясь формулой (4.23), вычислим вероятность попадания нормально распределенной величины X в следующие симметричные интервалы с центром в m_x: а) ($m_x-\sigma_x$; $m_x+\sigma_x$); б) ($m_x-2\sigma_x$; $m_x+2\sigma_x$); в) ($m_x-3\sigma_x$; $m_x+3\sigma_x$):

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

$$= \Phi(1) - \Phi(-1) = \Phi(1) + \Phi(1) = 2\Phi(1) = 2 \cdot 0{,}3413 \approx 0{,}683 = 68{,}3\,\%.$$

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

$= \Phi(2) - \Phi(-2) = 2\Phi(2) = 2 \cdot 0{,}4772 \approx 0{,}954 = 95{,}4\,\%.$

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

$= \Phi(3) - \Phi(-3) = 2 \cdot \Phi(3) = 2 \cdot 0{,}49865 \approx 0{,}997 = 99{,}7\,\%.$

Вероятность нахождения случайной величины, распределенной по нормальному закону, в интервале (m_x-3σ_x; m_x+3σ_x) весьма близка к единице, поэтому «трехсигмовые» границы $m_x \pm 3\sigma_x$ принимаются за границы практически возможных значений нормально распределенной случайной величины.

4.2.4. Распределение «хи квадрат»

Пусть X_i $(i=1,2,\ldots, n)$ — нормированные нормальные независимые случайные величины, то есть математическое ожидание каждой из них равно нулю, $m_x=0$ и $\sigma_x=1$. Тогда величина $\chi^2 = \sum_{i=1}^{n} x_i^2$ распределена по закону χ_k^2 с n степенями свободы. Если же эти величины связаны одним линейным соотношением, например, $\sum_{i=1}^{n} x_i = n \cdot \bar{x}$, то есть $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, то $k=n-1$.

Плотность этого распределения

$$f(x) = \begin{cases} 0; & x \le 0 \\ \dfrac{1}{2^{k/2} \cdot \Gamma(k/2)} \cdot e^{-x/2} \cdot x^{(k/2)-1}; & x > 0 \end{cases}, \qquad (4.24)$$

где $\Gamma(x) = \int_0^{\infty} t^{x-1} \cdot e^{-t} \cdot dt$ — гамма функция; в частности $\Gamma(n+1)=n!$

Отсюда видно, что распределение χ^2 определяется одним параметром — числом степеней свободы k. С увеличением k распределение медленно приближается к нормальному (рис. 4.8).

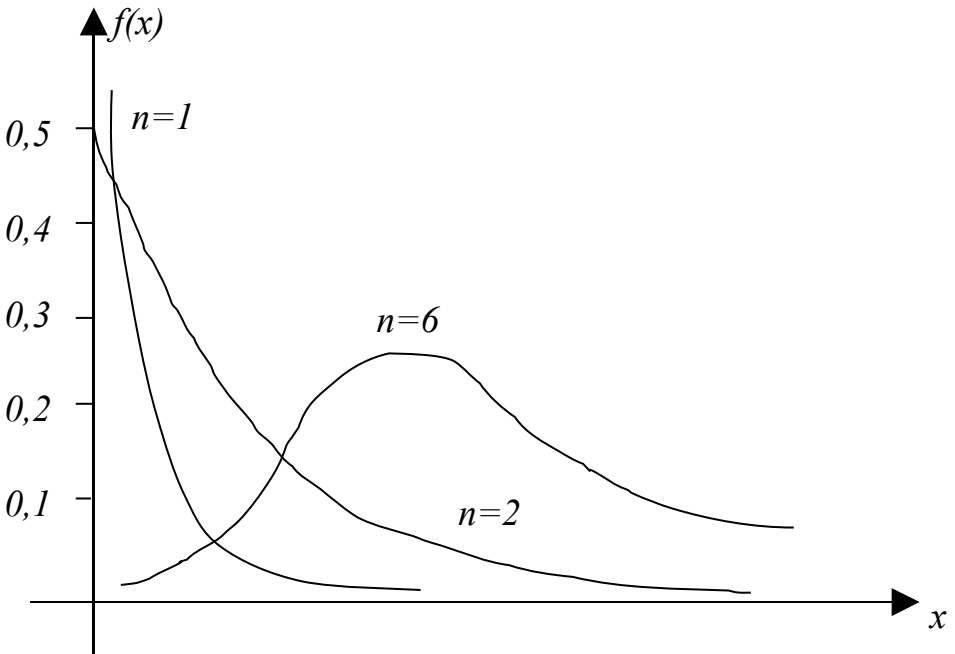


Рис. 4.8. Распределение χ^2

4.2.5. Распределение Стьюдента

Пусть ***Z*** — нормированная нормальная случайная величина, то есть ***M[Z]=0, σ[Z]=1,*** а ***V*** — независимая от ***Z*** величина, которая распределена по закону χ^2 с ***k*** степенями свободы. Тогда величина

$$T = \frac{Z}{\sqrt{V/k}} \tag{4.25}$$

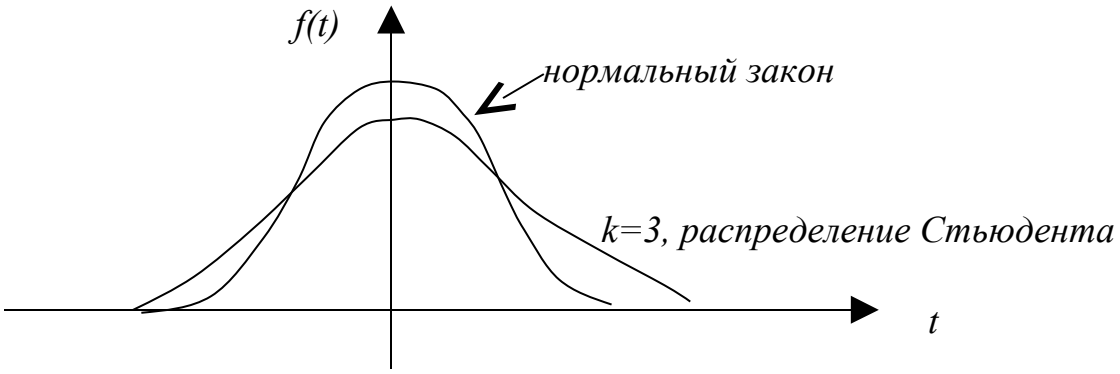


Рис. 4.9. Распределение Стьюдента

имеет распределение, которое называют ***t***-распределением или распределением Стьюдента с ***k*** степенями свободы. С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормированному нормальному распределению. При ***k*>30** они практически совпадают (рис. 4.9).

4.2.6. Распределение Фишера

Если ***U*** и **V** — независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2, то величина

$$F = \frac{U/\kappa_1}{V/\kappa_2}. \tag{4.26}$$

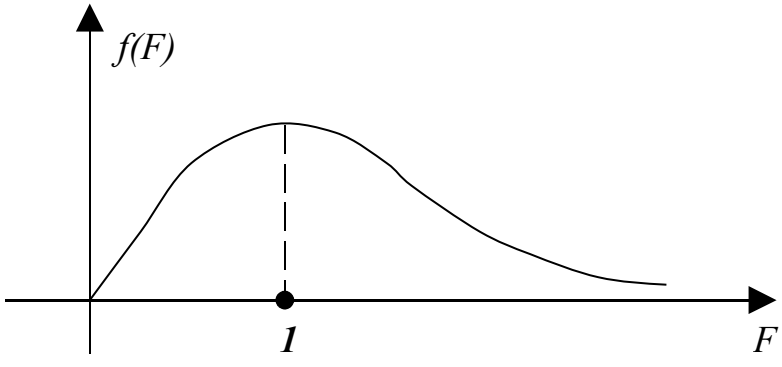


Рис. 4.10. Распределением Фишера

имеет распределение, которое называют распределением Фишера со степенями свободы k_1 и k_2 (рис. 4.10). Максимум плотности распределения соответствует $F=1$.

4.3. Центральная предельная теорема

В теории вероятностей в 1901 году русским математиком А.М. Ляпуновым были получены условия, при которых справедлива центральная предельная теорема: сумма очень большого числа взаимно независимых случайных величин имеет распределение, близкое к нормальному.

Сформулируем эти условия. Пусть $X_1, X_2, \ldots, X_n, \ldots$ — последовательность взаимно независимых случайных величин, каждая из которых имеет конечные математическое ожидание и дисперсию:

$$M[X_k]=a_k;\ D[X_k]=b_k^2. \tag{4.27}$$

Составим «частичные» суммы S_n этих величин:

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{k=1}^{n} X_k . \tag{4.28}$$

Тогда математическое ожидание и дисперсия этих сумм будут равны:

$$M[S_n] = A_n = \sum_{k=1}^{n} ak, \text{ а } D[S_n] = B_n^2 = \sum_{k=1}^{n} b_k^2 . \tag{4.29}$$

Перейдем от суммы S_n случайных величин к их нормированной сумме $\dfrac{S_n - M[S_n]}{\sqrt{D[S_n]}} = \dfrac{S_n - A_n}{B_n} = \dfrac{1}{B_n}\sum_{i=1}^{n}(X_i - a_i)$, имеющей нулевое математическое ожидание и дисперсию, равную единице.

Функцию распределения нормированной суммы будет равна

$$F_n(x) = P\left(\frac{S_n - A_n}{B_n} < x\right).$$

Говорят, что к последовательности случайных величин $X_1, X_2, \ldots$ применима центральная предельная теорема, если при любом x функция

распределения нормированной суммы при $n \to \infty$ стремится к нормированной нормальной функции распределения:

$$\lim_{n \to \infty} P\left[\frac{S_n - A_n}{B_n} < x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \cdot dt \,. \qquad (4.30)$$

А.М. Ляпунов доказал, что если для $\delta > 0$ при $n \to \infty$ отношение

$$L_n = C_n / B_n^{2+\delta}, \quad \text{где} \quad C_n = \sum_{k=1}^{n} M[X_k - a_k]^{2+\delta} \qquad (4.31)$$

стремится к нулю, то к последовательности $X_1, X_2, \ldots$ применима центральная предельная теорема. Сущность условия А.М. Ляпунова состоит в требовании, чтобы каждое слагаемое нормированной суммы $(S_n - A_n)/B_n$ оказывало на сумму ничтожное влияние. В частности, если все случайные величины $X_1, X_2 \ldots$ одинаково распределены, то к ним применима центральная предельная теорема, если дисперсии всех величин X_i $(i = 1, 2 \ldots n)$ конечны и отличны от нуля.

4.4. Оценка отклонения теоретического распределения от нормального, асимметрия и эксцесс

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 симметрия и эксцесс имеют небольшие значения, то можно предположить близость этого распределения к нормальному. Наоборот, большие значения асимметрии и эксцесса указывают на значительные отклонения от нормального закона. Для симметричных распределений каждый центральный момент нечетного порядка, то есть вида

$$M[(x - m_x)^{2k+1}], \quad k = 0, 1, 2, \ldots$$

равен нулю. Простейший из них момент третьего порядка μ_3 (момент первого порядка μ_1 равен нулю для любого распределения) удобно выбрать для характеристики степени асимметрии распределения:

$$\mu_3 = M[(x - m_x)^3].$$

Чтобы его величина не зависела от единиц, в которых измеряется случайная величина, его делят на σ^3. Итак, асимметрия равна:

$$A_3=\mu_3/\sigma^3. \qquad (4.32)$$

Асимметрия положительна, если «длинная часть» кривой плотности распределения расположена справа от математического ожидания; асимметрия отрицательна, если длинная часть кривой расположена слева от математического ожидания (рис. 4.11).

а б

Рис. 4.11

Для оценки «крутости», то есть большего или меньшего подъема кривой теоретического распределения по сравнению с нормальной кривой, пользуются характеристикой — эксцессом. Эксцессом теоретического распределения называют величину, равную

$$E_k = (\mu_4/\sigma^4) - 3. \qquad (4.33)$$

Для нормального распределения $\mu_4/\sigma^4 = 3$, следовательно, эксцесс равен нулю. Если эксцесс положительный, то кривая имеет более высокую и «острую» вершину, чем нормальная кривая, если эксцесс отрицательный, то сравниваемая кривая имеет более низкую и «плоскую» вершину, чем нормальная кривая (рис. 4.12). При этом предполагается, что нормальное и данное распределения имеют одинаковые математические ожидания и дисперсии.

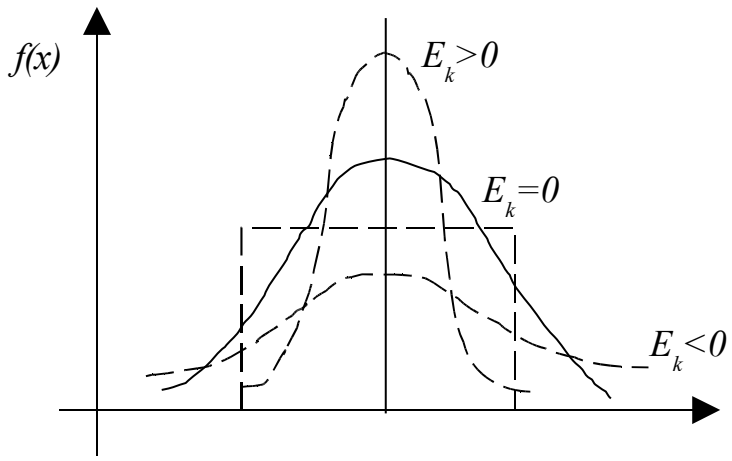


Рис. 4.12. Эксцесс теоретического распределения

Пример 7. Длина деталей, изготовленных на станке, является случайной величиной X, распределенной по нормальному закону с математическим ожиданием $m_x = 50$ *мм* и средним квадратичным отклонением $\sigma_x = 0{,}8$ *мм*. Наугад выбирают одну деталь. Найти вероятность того, что длина детали: а) окажется в интервале $(49;\ 50{,}4)$; б) отклонится от m_x по абсолютной величине не более чем на $1{,}2$ мм.

Решение

а) Для нормально распределенной величины X вероятность того, что X примет значение в интервале $(x_1,\ x_2)$, вычисляется по формуле (4.34):

$$P(x_1 < X < x_2) = \Phi(t_2) - \Phi(t_1) = \Phi\left(\frac{x_2 - m_x}{\sigma_x}\right) - \Phi\left(\frac{x_1 - m_x}{\sigma_x}\right). \qquad (4.34)$$

Подставим числовые данные и, найдя значения $\Phi(t)$ по таблице функции Лапласа, получим:

$$P(49 < X < 50{,}4) = \Phi\left(\frac{50{,}4 - 50}{0{,}8}\right) - \Phi\left(\frac{49 - 50}{0{,}8}\right) = \Phi(0{,}5) - \Phi(-1{,}25) =$$

$$= \Phi(0{,}5) + \Phi(1{,}25) = 0{,}19146 + 0{,}39435 = 0{,}58581.$$

б) Согласно условию задачи, необходимо найти вероятность следующего события $\{|X - m_x| < \delta\}$, где $\delta = 1{,}2$ мм. Преобразуем это событие, использовав определение модуля и добавив затем ко всем частям двойного неравенства m_x:

$$\{|X - m_x| < \delta\} = \{-\delta < (X - m_x) < \delta\} = \{m_x - \delta < X < m_x + \delta\}.$$

Тогда $P\{|X - m_x| < \delta\} = P(m_x - \delta < X < m_x + \delta) = \Phi\left(\frac{(m_x + \delta) - m_x}{\delta_x}\right) -$

$$\Phi\left(\frac{(m_x - \delta) - m_x}{\delta_x}\right) = \Phi\left(\frac{\delta}{\sigma_x}\right) - \Phi\left(-\frac{\delta}{\sigma_x}\right) = 2\Phi\left(\frac{\delta}{\sigma_x}\right).$$

Подставив численные значения, получим:

$$P\{|X - m_x| < 1{,}2\} = 2\Phi\left(\frac{1{,}2}{0{,}8}\right) = 2\Phi(1{,}5) = 2 \cdot 0{,}43319 = 0{,}86638.$$

Следовательно, вероятность того, что длина детали будет отличатся от m_x по абсолютной величине не более чем на $1{,}2$ мм, равна $0{,}86638$.

Задачи для самостоятельного решения

1. Случайная величина X задана плотностью распределения $f(x)$:

$$f(x) = \begin{cases} 0 & \text{при } x \le 0; \\ 2x & \text{при } 0 < x \le 1; \\ 0 & \text{при } x > 1. \end{cases}$$

Найти:

— функцию распределения $F(x)$;

— вероятность $P(0 < X < \pi/4)$;

— m_x и σ_x.

2. Все значения случайной величины принадлежат интервалу (**0; 2**), причем плотность распределения

$$f(x) = \begin{cases} 3/4 & \text{при } 1 < x < 2; \\ 1/4 & \text{при } 0 < x \le 1. \end{cases}$$

Найти функцию распределения $F(x)$, математическое ожидание и вероятности $P(0 < X < 1{,}5)$ и $P(0 X > 1{,}4)$.

3. Случайная величина X задана функцией распределения $F(x)$:

$$F(x) = \begin{cases} 0 & \text{при } x \le -\pi / 2; \\ a \cdot \sin(x) & \text{при } -\pi//2 < x \le \pi / 2; \\ 1 & \text{при } x > \pi . \end{cases}$$

Найти коэффициент a и вероятность $P(|X| < \pi/4)$.

4. Время ожидания приема у врача является случайной величиной, распределенной по показательному закону с $m_x = 20$ минут. Найти вероятность того, что пациенту придется ждать приема: а) не более 30 минут; б) более 1 часа.

5. Функция распределения случайного времени безотказной работы радиоаппаратуры имеет вид: $F(t) = 1 - e^{-t/T}, (t \ge 0)$. Найти вероятность безотказной работы в течение времени T и вероятность отказа радиоаппаратуры в интервале времени $(0; 2 \cdot T)$.

6. Шкала секундомера имеет цену деления $0{,}2$ секунды. Какова вероятность сделать по этому секундомеру отсчет времени с погрешностью более $0{,}05$ сек., если отсчет делается с точностью до целого деления с округлением в ближайшую сторону.

7. Автобусы ходят с интервалом в 20 минут. Предполагая, что время X ожидания автобуса на остановке имеет равномерное распределение, найти: а) плотность распределения $f(x)$; б) среднюю продолжительность ожидания; в) вероятности того, что время ожидания будет меньше 5 мин.; больше 15 мин.; более 1 часа.

8. Высотомер имеет случайные погрешности, распределенные по нормальному закону. Какую среднеквадратичную погрешность должен иметь прибор, чтобы с вероятностью $0{,}9$ погрешность измерения высоты была меньше $0{,}1$ м?

9. Детали, изготовленные автоматом, считаются стандартными, если отклонения их диаметра от проектного размера не превышают ***0,2*** мм. Случайные отклонения диаметра подчиняются нормальному закону с дисперсией ***0,0256*** мм^2 и математическим ожиданием $m_x=0$. Сколько процентов стандартных деталей изготовляет автомат?
10. Известно, что для человека pH крови является нормально распределенной случайной величиной со средним значением ***7,4*** и средним квадратичным отклонением ***0,2***. Какова вероятность того, что уровень pH крови: а) превышает ***7,3***; б) находится между ***7,35*** и ***7,45***?

5. ОСНОВНЫЕ ПРЕДЕЛЬНЫЕ ЗАКОНЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

5.1. Среднее арифметическое *n* одинаково распределенных взаимно независимых случайных величин

Случайные величины $X_1, X_2, \ldots, X_n$ называют одинаково распределенными, если они имеют одинаковые законы распределения. Следовательно, при любом x

$$F(x) = P(X_1 < x) = P(X_2 < x) = \ldots = P(X_i < x) = \ldots = P(X_n < x). \qquad (5.1)$$

Пусть имеется n взаимно независимых (см. разд. 3.2) случайных величин $X_1, X_2, \ldots, X_n$, которые имеют одинаковые распределения, а следовательно, одинаковые ***M[x]*** и ***Д[x]***. Существует связь между числовыми характеристиками каждой из отдельно взятых случайных величин и их средним арифметическим.

Среднее арифметическое рассматриваемых случайных величин равно:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (5.2)$$

Найдем для него математическое ожидание:

$$M[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = M[X], \qquad (5.3)$$

где $M[X_i] = M[X]$. Это означает, что математическое ожидание среднего арифметического n одинаково распределенных взаимно независимых случайных величин равно математическому ожиданию каждой из этих величин.

Дисперсия среднего арифметического n взаимно независимых одинаково распределенных случайных величин в n раз меньше дисперсии каждой из этих величин:

$$D[\overline{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{D[X]}{n}, \qquad (5.4)$$

где $D[X_i] = D[X]$.

Для среднеквадратичного отклонения аналогично имеем:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}. \qquad (5.5)$$

Так как дисперсия и среднеквадратичное отклонение характеризуют разброс, рассеивание значений случайной величины относительно m_x, то отсюда следует, что рассеивание среднего арифметического случайных величин значительно меньше, чем для отдельно взятой случайной величины.

5.2. Закон больших чисел

Ранее отмечалось, что при проведении серии одинаковых испытаний относительная частота появления случайного события при большом числе испытаний обнаруживает свойство статистической устойчивости, что позволяет ввести меру объективной возможности появления случайного события в отдельном испытании, то есть его вероятности. Аналогично отдельные значения случайной величины могут значительно отклоняться от ее среднего значения, но среднее арифметическое большого числа отдельных значений случайной величины уже незначительно отклоняется от ее математического ожидания. Мы наблюдаем единичные явления вместе со всеми их индивидуальными особенностями, мешающими проявлению тех закономерностей, которые имеют место при наблюдении большого числа аналогичных явлений. Закономерности случайных явлений и величин проявляются при большем их количестве. Случайные отклонения от закономерностей, имеющиеся при отдельных наблюдениях, взаимно погашаются по мере возрастания числа наблюдений за случайными явлениями или величинами. Это формулировка закона больших чисел в широком смысле.

Для практики очень важно знание условий, при выполнении которых совокупное действие очень многих случайных причин приводит к результатам, почти не зависящим от случая, то есть позволяет предвидеть ход явлений. Эти условия и указываются в теоремах, носящих общее название закона больших чисел. В узком смысле слова под законом больших чисел понимают ряд математических теорем: Чебышева, Бернулли, Ляпунова и др. Для доказательства этих теорем используется неравенство Чебышева.

Если X — случайная величина с конечными математическим ожиданием $M[X]$ и дисперсией $D[X]$, то для любого положительного ε имеет место неравенство:

$$P(|X - M[X]| < \varepsilon) \geq 1 - \frac{D[X]}{\varepsilon^2}. \qquad (5.6)$$

С помощью этого неравенства, например, можно оценить сверху вероятность того, что случайная величина отклонится от m_x не более чем на $k\sigma_x$:

$$P(|X - M[X]| < k\sigma_x) \geq 1 - \frac{\sigma_x^2}{(k\sigma_x)^2} = 1 - \frac{1}{k^2}.$$

В частности, при $k=3$ ($\varepsilon = 3\sigma_x$) получим: $P(|X - M[X]| < 3\sigma_x) \geq 1 - \frac{1}{3^2} = \frac{8}{9}$.

Значит, при любом распределении случайной величины X не менее $^8/_9$ **89 %** всех ее значений находится в интервале $(m_x - 3\sigma_x; m_x + 3\sigma_x)$.

Теорема Чебышева. Пусть имеется совокупность n попарно независимых случайных величин $X_1, X_2 \ldots X_n$ с какими угодно распределениями вероятностей. Пусть все эти случайные величины имеют определенные математические ожидания и дисперсии:

$$M[X_i] = m_i; \qquad D[X_i] = \sigma_i^2.$$

Образуем среднюю арифметическую этих случайных величин:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{5.7}$$

Математическое ожидание среднего арифметического $\overline{X}$ будет равно среднему арифметическому их математических ожиданий:

$$M[\overline{X}] = M\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \overline{m}. \tag{5.8}$$

Дисперсия среднего арифметического равна:

$$D[\overline{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2. \tag{5.9}$$

Пусть все дисперсии ограничены сверху величиной C: $\sigma_i^2 \leq C$, тогда $D[\overline{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \leq \frac{C \cdot n}{n^2} = \frac{C}{n}$ и справедлива следующая теорема Чебышева:

Вероятность того, что абсолютное отклонение среднего арифметического $\overline{X}$ случайных величин, дисперсии которых ограничены, от среднего арифметического их математических ожиданий $\overline{m}$ меньше, чем $\varepsilon > 0$, с возрастанием количества случайных величин становится сколь угодно близкой к единице, то есть

$$P\left\{|\overline{X} - \overline{m}| < \varepsilon\right\} \geq 1 - \frac{C}{n \cdot \varepsilon^2}, \tag{5.10}$$

где ε — сколь угодно малое положительное число.

Отсюда можно сделать вывод, что среднее арифметическое достаточно большого числа независимых случайных величин, дисперсии которых ограничены, утрачивают характер случайной величины, то есть

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots X_i + \ldots + X_n}{n} - \overline{m}\right| < \varepsilon\right) = 1. \qquad (5.11)$$

На теореме Чебышева основан широко применяемый в статистике выборочный метод, суть которого состоит в том, что по сравнительно небольшой случайной части совокупности (выборке) судят обо всей совокупности исследуемых объектов.

Частным случаем теоремы Чебышева является *теорема Бернулли*, устанавливающая связь между относительной частотой события и его вероятностью:

Вероятность того, что абсолютное отклонение относительной частоты события в n независимых испытаниях от вероятности его появления в одном испытании меньше, чем $\varepsilon > 0$, с возрастанием числа испытаний становиться сколь угодно близкой к единице, то есть

$$P\left\{\left|\frac{m}{n} - p\right| \le \varepsilon\right\} \ge 1 - \frac{pq}{n\varepsilon^2}, \qquad (5.12)$$

где ε — сколь угодно малое положительное число, p и q — вероятности появления и непоявления события в одном испытании.

Теоремы Чебышева и Бернулли формулируют те общие условия, выполнение которых влечет за собой статистическую устойчивость средних результатов серии опытов или испытаний.

Пример 1. При проверке ампул необходимо, чтобы вероятность того, что абсолютное отклонение среднего значения массы ампул от их математического ожидания меньше $0,2$ г, превышала $0,95$. Дисперсия массы ампул не превышает $0,04$ г. Сколько ампул надо взять для удовлетворения указанных условий?

Решение. По условию задачи случайные величины X_1, X_2 … X_n — массы отдельных ампул — имеют одинаковые математические ожидания:

$M[X_i] = m_i = m$, тогда $\overline{m}_n = \frac{1}{n}\sum_{i=1}^{n} m_i = m$ и $P(|\overline{X} - m| < 0,2) > 0,95.$

По теореме Чебышева имеем: $P(|\overline{X} - m| < 0,2) \ge 1 - \frac{0,04}{n \cdot 0,04}$, откуда:

$$0,95 \le 1 - 0,04/(n \cdot 0,04).$$

Для n получим: $1/n \le 0,05$ и $n \ge 20$ — для проверки достаточно взять 20 ампул.

Пример 2. Найти вероятность того, что абсолютная разность между относительной частотой и вероятностью появления таблетки со стан-

дартной массой будет меньше ***0,2***, если вероятность того, что одна таблетка стандартна, равна ***0,9*** и для проверки взято ***60*** таблеток.

Решение. Согласно условию имеем: ***p=0,9; q=1- p=0,1; n=60; ε=0,2.***

По теореме Бернулли получим:

$$P(|w - 0{,}8| < 0{,}2) \geq 1 - pq/(n \cdot \varepsilon^2) = 1 - \frac{0{,}9 \cdot 0{,}1}{60 \cdot 0{,}2^2} = 1 - \frac{0{,}09}{2{,}4} = 1 - 0{,}0375 \approx 0{,}962.$$

Задачи для самостоятельного решения

1. Статистическая вероятность рождения девочки равна ***0,485***. Оцените вероятность того, что доля девочек среди ***3000*** новорожденных будет отличаться по абсолютной величине от вероятности рождения девочки не более чем на ***0,02.***
2. Выборочным путем требуется определить среднюю массу зерен пшеницы. Сколько нужно обследовать зерен, чтобы с вероятностью, большей ***0,9,*** можно было утверждать, что средняя масса отобранных зерен будет отличаться от средней массы всех зерен не более чем на ***0,001*** *г* ? Установлено, что дисперсия массы зерен не превышает ***0,002*** $г^2$.
3. При взвешивании на аналитических весах дисперсия случайных погрешностей равна $\sigma_x^2 = 4(мг^2)$. Найти вероятность того, что погрешность при десятикратном взвешивании не превзойдет по абсолютной величине ***1*** мг.
4. Вероятность брака при выпуске одноразовых шприцов составляет ***0,02***. Какое количество шприцов надо взять для контроля, чтобы отклонение относительной частоты годных от вероятности ***0,98*** не превысило ***0,05*** с вероятностью ***0,99***.
5. Вероятность того, что посетитель потребует некоторый препарат, равна ***0,2.*** Найти вероятность того, что при ***1000*** посетителях отклонение относительной частоты от вероятности спроса будет меньше ***0,02.***
6. Дисперсия роста студента не превышает ***40*** $см^2$. Сколько студентов надо взять для контроля их роста, чтобы абсолютное отклонение среднего арифметического значения их роста от среднего роста (то есть погрешность) не превышало ***1,5*** см при вероятности не менее ***0,95***.

6. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

6.1. Задачи математической статистики

Математическая статистика — раздел математики, посвященный методам сбора, анализа и обработки статистических данных для получения научных и практических выводов.

Статистические данные — данные, полученные в результате обследования большого числа объектов или явлений, то есть математическая статистика имеет дело с массовыми явлениями.

Теория вероятностей является основной, фундаментом для математической статистики. Основное отличие математической статистики от теории вероятностей состоит в том, что в математической статистике рассматриваются не действия над законами распределения и числовыми характеристиками случайных величин, а приближенные методы отыскания этих законов и характеристик по результатам ограниченного числа экспериментов (наблюдений).

Любое значение искомой характеристики, вычисленное на основе ограниченного числа опытов, всегда содержит влияние фактора случайности в виде погрешности. Такое приближенное, случайное значение называется оценкой характеристики. Из всех возможных оценок желательно выбирать такую, чтобы эти погрешности были минимальны.

Первая задача математической статистики — описательная: указать способы сбора, представления и группировки статистических данных.

Вторая задача математической статистики — аналитическая: разработка методов анализа статистических данных в зависимости от целей исследования. Сюда относятся:

— оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения и т.д.;

— проверка статистических гипотез о виде распределения, параметрах распределений, статистической связи между случайными величинами, влиянии определенных факторов на изучаемый параметр;

— разработка методов планирования эксперимента с целью получения оценок параметров с заданными точностью и надежностью.

6.2. Генеральная и выборочная совокупности, способы отбора

Генеральной совокупностью называется вся совокупность однородных объектов, подлежащих изучению относительно некоторого общего качественного или количественного признака $\boldsymbol{X}$. Например, если имеется большая партия деталей, то качественным признаком может служить стандартность детали, а количественным — контролируемый размер детали.

Распределение признака в генеральной совокупности определяется функцией $F(x)$ или плотностью $f(x)$ распределения, или же характеризуется числовыми характеристиками $M[X], D[X], \sigma_x$ и др. Если генеральная совокупность содержит очень большое число объектов, то провести сплошное обследование физически невозможно. Проводить сплошное обследование также не имеет смысла, если исследование объекта связанно с его уничтожением или требует больших материальных затрат, например, определение химического состава лекарства, проверка качества продуктов и т.д. Поэтому на практике обычно из генеральной совокупности случайно выбирают часть объектов, чтобы на основе их изучения оценить неизвестное распределение признака X в генеральной совокупности.

Выборочной совокупностью (или просто выборкой) называется отобранная для исследования часть генеральной совокупности.

Объемом совокупности (генеральной или выборочной) называется число объектов совокупности. Например, если из 1000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности $N = 1000$ деталей, а объем выборки $n = 100$.

На практике часто генеральная совокупность содержит конечное число N объектов, в математической статистике для облегчения теоретических выводов полагают, что генеральная совокупность — это бесконечно большая (или приближающаяся к ней) совокупность.

Для того чтобы свойства выборки достаточно правильно отражали свойства генеральной совокупности, выборка должна быть представительной (репрезентативной).

Согласно закону больших чисел, можно утверждать, что выборка будет репрезентативной, если ее осуществить случайно, то есть если каждый объект генеральной совокупности имеет одинаковую вероятность попасть в выборку.

Применяемые на практике способы отбора разделяются на два вида:

• Отбор, не требующий расчленения генеральной совокупности на части. Сюда относятся: простой случайный повторный отбор; простой случайный бесповторный отбор.

Повторной называют выборку, при которой отобранный объект после исследования возвращают в генеральную совокупность перед отбором следующего, то есть численность единиц генеральной совокупности в процессе выборки остается неизменной.

Бесповторной называют выборку, при которой отобранный объект обратно в генеральную совокупность не возвращается и в дальнейшем в выборке не участвует.

На практике более распространена бесповторная выборка. Если объем генеральной совокупности достаточно велик, а выборка состав-

ляет незначительную часть этой совокупности, то различие между повторной и бесповторной выборками исчезает.

• Отбор, при котором генеральная совокупность разбивается на части. Сюда относятся: типический отбор; механический отбор; серийный отбор.

Типическим называется отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой типической ее части. Типическим отбором пользуются, если признак заметно колеблется в типических частях. Например, если продукция изготавливается на нескольких станках, среди которых есть более и менее изношенные, то целесообразен типический отбор.

Механическим называют отбор, при котором генеральную совокупность «механически» делят на столько групп, сколько объектов должно войти в выборку, а из каждой группы отбирают один объект. Иногда механический отбор не обеспечивает репрезентативной выборки.

Серийным называют отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Серийным отбором пользуются тогда, когда признак в разных сериях колеблется незначительно. Например, проверяется полностью продукция только нескольких станков из большой группы станков-автоматов.

6.3. Статистическое распределение выборки

Пусть из генеральной совокупности объемом N извлечена выборка объемом n. Наблюдаемые различающиеся значения признака $x_1, x_2, \ldots, x_к$ называются вариантами (k — число вариант), а последовательность вариант, записанная в упорядоченном виде (чаще всего в порядке возрастания), — вариационным рядом.

Пусть варианта x_1 появилась в выборке m_1 раз; x_2: m_2, …, $x_к$: $m_к$ раз, где $\sum_{i=1}^{к} m_i = n$ – объем выборки. Число m_i появлений варианты x_i называют частотой, а отношение $w_i = \frac{m_i}{n}$ — относительной частотой этой варианты. Относительные частоты являются оценками соответствующих вероятностей

$$w_i \approx p_i = P(X = x_i), \tag{6.1}$$

причем соблюдается условие нормировки: $\sum_{i=1}^{k} w_i = 1$. (6.2)

Статистическим дискретным рядом распределения называется таблица 6.1, содержащая расположенные обычно в порядке возрастания варианты x_i признака X и их частоты m_i, или относительные частоты w_i:

Таблица 6.1

x_i	x_1	x_2	…	x_i	…	x_k
w_i	w_1	w_2	…	w_i	…	w_k

Дискретный ряд распределения обычно служит для описания малых выработок ($n < 30$).

Если же имеется выборка большого объема ($n \geq 30$), то ее предварительно группируют — распределяют варианты выборки по частичным интервалам, каждый из которых содержит некоторый диапазон значений изучаемого признака. В общем случае ширины интервалов могут быть различными, но обычно выбирают интервалы одинаковой длины.

Число k и длины Δx_i ($i = 1, 2, \ldots, k$) частичных интервалов группировки выбирают исходя из целей исследования, объема n выборки и степени варьирования признака в выборке. В случае равных интервалов их число k можно приближенно оценить исходя только из объема выборки n по формуле Стерджеса:

$$k = 1 + 3{,}32 \cdot \lg(n), \qquad (6.3)$$

или в соответствии с таблицей 6.2.

Таблица 6.2

Объем выборки n	25—40	40—60	60—100	100—200	>200
Число интервалов k	5—6	6—8	7—10	8—12	10—15

Статистическим интервальным рядом распределения называется таблица 6.3, содержащая частичные интервалы, расположенные в порядке возрастания, и их частоты m_i (частота интервала равна числу вариант, попавших в интервал) или относительные частоты w_i:

Таблица 6.3

Интервал (m_i)	(x_0, x_1)	(x_1, x_2)	…	(x_{i-1}, x_i)	…	(x_{k-1}, x_k)
w_i	w_1	w_2	…	w_i	…	w_k

Для графического изображения статистического дискретного ряда распределения используется полигон. Для построения полигона на оси Ox откладывают значения вариант x_i, на оси Oy — значения

относительных частот ω_i. Построенную таким образом ломаную, отрезки которой соединяют точки с координатами $(x_i;\omega_i)$, называют полигоном относительных частот (рис. 6.1).

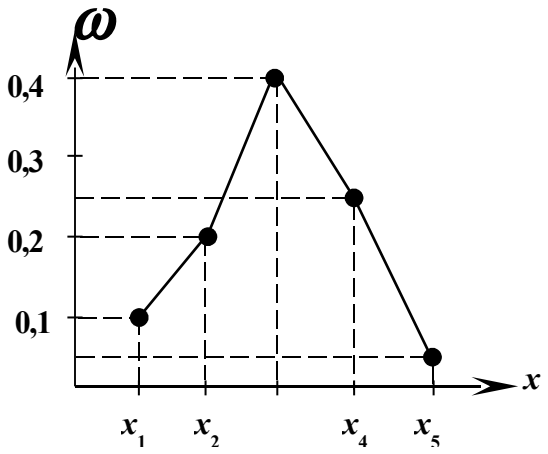


Рис. 6.1. Полигон относительных частот

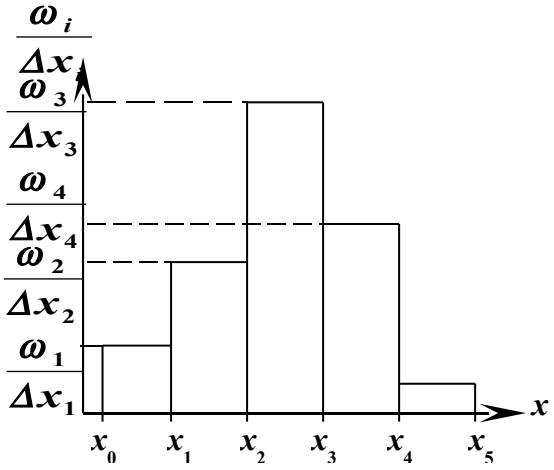


Рис. 6.2. Гистограмма

Для графического изображения статистического интервального ряда распределения строится гистограмма. Для этого на оси ***Ox*** откладывают частичные интервалы длиной $\Delta x_i = x_i - x_{i-1}$, которые в общем случае могут быть различными по величине. Далее на каждом частичном интервале строят прямоугольник с основанием Δx_i и высотой, равной частному от деления относительной частоты интервала на длину интервала $\omega_i/\Delta x_i$. Площадь ***i***-го прямоугольника равна $\Delta x_i \cdot \frac{\omega_i}{\Delta x_i} = \omega_i$, а суммарная площадь всех прямоугольников равна сумме всех относительных частот, то есть единице. Полученную таким образом ступенчатую фигуру, состоящую из прямоугольников, называют гистограммой (рис. 6.2).

Высота прямоугольников на гистограмме равна плотности частоты $\omega_i/\Delta x_i$, поэтому гистограмма является оценкой плотности распределения вероятности $f(x) \approx p_i/\Delta x_i$ для непрерывных распределений. При неограниченном увеличении числа наблюдений n и уменьшении длин Δx_i частичных интервалов ступенчатая верхняя линия гистограммы будет стремиться к плавной кривой плотности распределения $f(x)$ генеральной совокупности. Сравнивая кривую, огибающую гистограмму сверху, с графиком плотности вероятности типичного распределения, можно отнести изучаемое распределение к тому или иному типу.

Эмпирическая (статистическая) функция распределения $F(x)$ является оценкой неизвестной функции распределения $F(x)$ генеральной

совокупности и для каждого значения x равна относительной частоте события ($X < x$):

$$F(x) = w(X < x) = \frac{n_x}{n}, \qquad (6.4)$$

где n_x — число вариант в выборке, меньших x.

Эмпирическая функция распределения обладает всеми свойствами функции распределения:

• Значения эмпирической функции принадлежат отрезку *[0,1]* — $0 \le F(x) \le 1$.

• $F(x)$ — неубывающая функция.

• Если x_1 — наименьшая варианта, то $F(x) = 0$ при $x \le x_1$; если x_k — наибольшая варианта, то $F(x) = 1$ при $x > x_k$.

График $F(x)$ имеет ступенчатый вид (рис. 6.3).

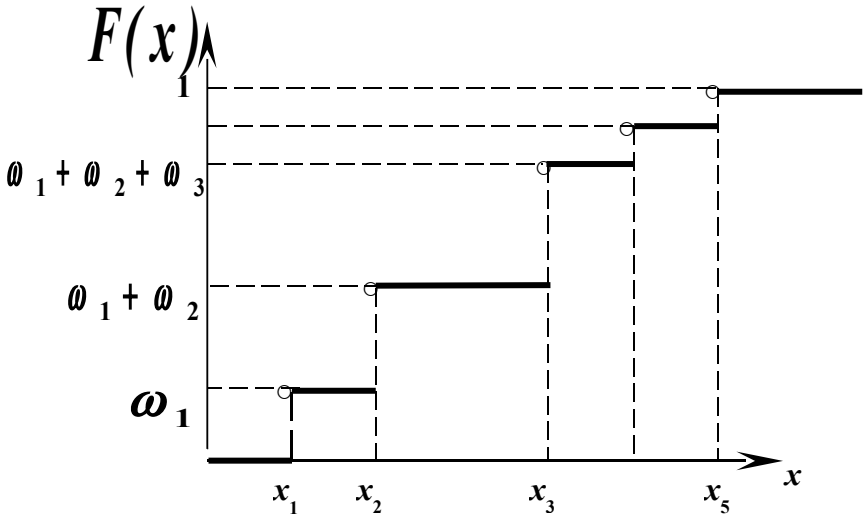


Рис. 6.3. График $F(x)$

Пример 1. В результате $n = 15$ независимых наблюдений некоторой дискретной случайной величины X получена выборка: ***5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4***. Требуется составить:

— вариационный ряд;

— статистический дискретный ряд распределения.

Решение

1. Выбирая варианты из выборки и располагая их в порядке возрастания, получим вариационный ряд: ***2, 3, 4, 5, 7, 10.***

2. Для нахождения относительных частот $w_i = m_i / 15$ предварительно подсчитаем для каждой варианты частоту m_i ее повторения в выборке:

$m_i = 3, 1, 2, 3, 4, 2$, причем $\sum m_i = 3 + 1 + 2 + 3 + 4 + 2 = 15$.

Статистический дискретный ряд распределения:

x_i	2	3	4	5	7	10
ω_i	3/15	1/15	2/15	3/15	4/15	2/15

Пример 2. Дан статистический интервальный ряд распределения:

Время реакции, мин.	120—160	160—180	180—200	200—220	220—240	240—280
ω_i	6/50	10/50	14/50	12/50	6/50	2/50

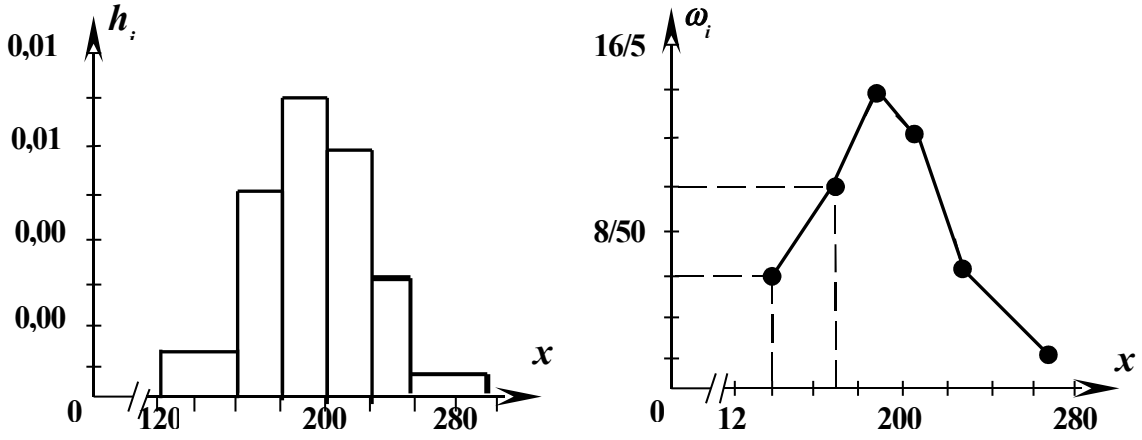
Построить: гистограмму; полигон относительных частот.

Решение

1. Для построения гистограммы на отрезках оси абсцисс длиной Δx_i, соответствующих данным интервалам, строятся прямоугольники с высотой $h_i = \frac{\omega_i}{\Delta x_i}$. Вычислим Δx_i и h_i, например, ширина первого интервала $\Delta x_1 = 160 - 120 = 40$; высота первого прямоугольника $h_1 = (6/50)/40 = 0{,}003$ и т.д.

Δx_i	40	20	20	20	20	40
h_i	0,003	0,010	0,014	0,012	0,006	0,001

Выбираем соответствующий масштаб и строим гистограмму (см. рис. 6.4а):



а
Гистограмма

б
Полигон относительных частот

Рис. 6.4

2. Чтобы построить полигон, данный интервальный статистический ряд преобразуем в дискретный. Для этого каждый интервал заменяется дискретным значением признака, равного середине интервала:

Время реакции, мин.	*140*	*170*	*190*	*210*	*230*	*260*
ω_i	***6/50***	*10/50*	*14/50*	*12/50*	*6/50*	*2/50*

Теперь в прямоугольной системе координат строим точки, координатами которых являются пары чисел из полученного дискретного ряда распределения: (***140***; ***6/50***), (***170***; ***10/50***) и т.д. Последовательно соединяя отрезками прямых построенные точки, получим полигон относительных частот (см. рис. 6.4б).

6.4. Точечные оценки параметров распределения

Точечной называют оценку параметра, которая определяется одним числом — его приближенным значением.

Пусть θ — неизвестный параметр (m_x, ***D(X)***, σ_x и т.д.) распределения изучаемого количественного признака ***X***, а θ^* — его точечная оценка по выборке, зависящая от вариант $x_1, x_2 \ldots x_k$ выборки и приближенно равная θ, то есть $\theta^* \approx \theta$. Оценка θ^* является случайной величиной, так как она является функцией случайных вариант, и поэтому ее значение будет меняться от одной выборки к другой.

Несмещенной оценкой параметра θ называют такую оценку θ^*, математическое ожидание которой равно оцениваемому параметру θ, то есть $M[\theta^*] = \theta$. В противном случае оценка называется смещенной.

Оценка называется состоятельной, если при $n \to \infty$ она стремится по вероятности к оцениваемому параметру, например, является несмещенной и ее дисперсия стремится к нулю при неограниченном увеличении объема выборки.

Эффективной называется оценка, имеющая дисперсию, наименьшую среди дисперсий несмещенных оценок данного параметра.

Из отмеченных требований, предъявленных к оценке, наиболее важными являются требования несмещенности и состоятельности.

Рассмотрим оценки числовых характеристик генеральной совокупности по выборочным данным.

Оценкой неизвестного математического ожидания $\mathbf{т_x}$ генеральной совокупности является выборочное среднее $\overline{x}_в$, равное:

$$\hat{m}_x = \overline{x_в} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot x_i \,, \qquad (6.5)$$

где m_i — частота варианты x_i; $\sum_{i=1}^{k} m_i = n$.

Если все значения x_1, x_2 … x_n выборки объема n различны, то $\bar{x}_в$ можно вычислить по формуле:

$$\overline{x_в} = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (6.6)$$

Покажем, что $\bar{x}_в$ является несмещенной и состоятельной оценкой m_x. Для этого найдем математическое ожидание выборочного среднего:

$$M[\bar{x}_в] = M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[x_i]. \qquad (6.7)$$

Выборочные значения x_1, x_2,…, x_n случайной величины X также можно рассматривать как независимые одинаково распределенные случайные величины X_1, X_2, …, X_n, имеющие одинаковые числовые характеристики m_x, $D[X]$, σ_x. Тогда M [X] = m_x и из (6.6) получаем: $M[\bar{x}_в] = \frac{1}{n}\sum_{i=1}^{n} m_x = m_x$ то есть $\bar{x}_в$ – несмещенная оценка математического ожидания m_x.

Аналогично получим, что дисперсия выборочной средней $\bar{x}_в$ равна:

$$D[\bar{x}_в] = \frac{1}{n^2}\sum_{i=1}^{n} D[x_i] = \frac{1}{n^2}\cdot n \cdot D[X] = \frac{D(X)}{n}. \qquad (6.8)$$

Из формулы (6.7) следует, что при увеличении объема выборки ($n \to \infty$) дисперсия среднего арифметического стремится к нулю, то есть среднее выборочное является состоятельной оценкой математического ожидания.

Из понятия состоятельности следует, 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 проявляется свойство устойчивости выборочных средних. Выборочное среднее обладает следующими свойствами:

- Сумма отклонений вариант от выборочного среднего равна нулю: $\sum_{i=1}^{n} (x_i - \bar{x}_в) = 0$.

- Сумма квадратов отклонений вариант от выборочного среднего меньше, чем от любого другого числа a: $\sum_{i=1}^{n} (x_i - \bar{x}_в)^2 < \sum_{i=1}^{n} (x_i - a)^2$; $a \neq \bar{x}_в$.

Оценкой неизвестной дисперсии $D[X]$ генеральной совокупности может служить выборочная дисперсия $D_в[X]$, равная:

$$\hat{D}[X] \approx D_в[X] = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot (x_i - \bar{x}_в)^2 \qquad (6.9)$$

или

$$D_в[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}_в^2\right). \qquad (6.10)$$

Однако выборочная дисперсия $D_в[X]$ является смещенной оценкой дисперсии $D[X]$, так как $M[D_в(X)] = \frac{n-1}{n}\cdot D[X] \neq D[X]$. Поэтому $D_в[X]$ «исправляют», умножая ее на дробь $n/(n-1)$, и получают *исправленную выборочную дисперсию* S_x^2, являющуюся несмещенной ($M[S_x^2] = D[X]$) оценкой неизвестной дисперсии $D[X]$, равную:

$$\hat{D}[X] \approx S_x^2[X] = \frac{1}{n-1}\cdot\sum_{i=1}^{k} m_i\cdot(x_i - \bar{x}_в)^2 \qquad (6.11)$$

или $$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}_в^2\right). \qquad (6.12)$$

При малом объеме выборки ($n \leq 30$) для оценки $D[X]$ используют S_x^2; при больших же n ($n > 30$) практически безразлично, какую из двух оценок (S_x^2 или $D_в[X]$) использовать.

В качестве оценки неизвестного среднего квадратичного отклонения σ_x используют выборочное среднее квадратичное отклонение $\sigma_в$, равное:

$$\sigma_в = \sqrt{D_в[X]} = \sigma_x \qquad (6.13)$$

или исправленное среднее квадратичное отклонение:

$$S_x = \sqrt{S_x^2}\,. \qquad (6.14)$$

Если статистическое распределение выборки задано статистическим интервальным рядом распределения, то для применения формул (6.4) — (6.13) для оценки параметров в качестве x_i обычно берут середину интервала $(x_{i-1}; x_i)$.

Согласно (6.7) дисперсия и среднее квадратичное отклонение выборочного среднего $\bar{x}_в$ соответственно в n и в $\sqrt{n}$ раз меньше $D[X]$ и σ_x:

$$D[\bar{x}_в] = \frac{D(X)}{n}\,, \text{ откуда } \sigma_{\bar{x}_в} = \frac{\sigma_x}{\sqrt{n}} = \sqrt{\frac{D[X]}{n}}\,. \qquad (6.15)$$

Аналогичные формулы существуют и для оценок числовых характеристик выборочного среднего:

$$\hat{D}[\bar{x}_в] = S_{\bar{x}}^2 = \frac{S_x^2}{n}\,, \qquad (6.16)$$

$$\sigma_{\bar{x}} = S_{\bar{x}} = \frac{S_x}{\sqrt{n}}. \qquad (6.17)$$

Пример 3. Для оценки качества поставляемого сырья случайным образом были отобраны ***20*** образцов, в каждом из которых измерялась концентрация ***X*** (в %) некоторого вещества. Результаты измерений приведены в таблице:

x_i, %	4	5	6	8
Число образцов m_i	6	10	3	1

Найти: выборочное среднее; выборочную и исправленную выборочную дисперсии; в) исправленное среднее квадратичное отклонение.

Решение

1. По формуле (6.4) среднее выборочное равно

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot x_i = \frac{1}{20}(6\cdot 4 + 10\cdot 5 + 3\cdot 6 + 1\cdot 8) = \frac{100}{20} = 5.$$

2. По формуле (6.8) $D_в[X] = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot (x_i - \bar{x}_в)^2 =$

$$= \frac{1}{20}(6\cdot(4-5)^2 + 10\cdot(5-5)^2 + 3\cdot(6-5)^2 + 1\cdot(8-5)^2) = \frac{18}{20} = 0{,}9.$$

По формуле (6.10) $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{k} m_i \cdot (x_i - \bar{x}_в)^2 = \frac{18}{20-1} \approx 0{,}95.$

3. $S_x = \sqrt{S_x^2} \approx \sqrt{0{,}95} \approx 0{,}97.$

Таким образом, по результатам измерений можно считать, что среднее содержание вещества в поставляемом сырье, то есть математическое ожидание m_x, приближенно равно ***5*** %, а среднее квадратичное отклонение содержания вещества в образцах

$$\sigma_x \approx S_x \approx 0{,}97 \approx 1\%.$$

6.5. Точность и надежность оценки, доверительный интервал

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, поэтому при малых ***n*** следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами — концами доверительного интервала.

Доверительным интервалом для оценки параметра θ называется интервал $(\hat{\theta} - \delta ; \hat{\theta} + \delta)$, который заключает в себе (накрывает) неизвестный параметр θ с заданной надежностью (доверительной вероятностью) γ или в $100 \cdot \gamma\%$ случаев:

$$P(\hat{\theta} - \delta < m_x < \hat{\theta} + \delta) = P(|\hat{\theta} - \theta| < \delta) = \gamma , \qquad (6.18)$$

где положительное число δ — полуширина доверительного интервала — называется точностью оценки.

Надежность оценки γ обычно задают заранее, причем в качестве доверительной вероятности берут числа, достаточно близкие к единице (*0,95; 0,99; 0,999* и т.п.). При исследовании в экономике, экологии, биологии доверительную вероятность принимают равной $\gamma = 0{,}95$, при разработке стандартов $\gamma = 0{,}99$.

Иногда в литературе для задания надежности оценки используют не γ, а противоположную ей по смыслу величину — уровень значимости $\alpha = 1 - \gamma$, равный вероятности того, что оцениваемый параметр θ окажется за пределами доверительного интервала:

$$\alpha = P(|\hat{\theta} - \theta| > \delta) = 1 - P(|\hat{\theta} - \theta| < \delta) = 1 - \gamma .$$

Пусть количественный признак X генеральной совокупности распределен нормально с неизвестным математическим ожиданием m_x. При этом возможны два случая — дисперсия σ^2_x генеральной совокупности: а) известна; б) неизвестна.

Доверительным интервалом для оценки неизвестного математического ожидания m_x нормально распределенной случайной величины X будет интервал $(\bar{x}_в - \delta ; \bar{x}_в + \delta)$, который заключает в себе математическое ожидание с заданной доверительной вероятностью γ:

$$P(\bar{X}_в - \delta < m_x < \bar{X}_в + \delta) = \gamma ,$$

причем полуширина δ доверительного интервала находится:

— при известной дисперсии σ^2_x по формуле:

$$\delta = t_\gamma \cdot \frac{\sigma_x}{\sqrt{n}} = t_\gamma \cdot \sigma_{\bar{x}} , \qquad (6.19)$$

где n — объем выборки, t_γ — такое значение аргумента функции Лапласа $\Phi(t)$, при котором $2\Phi(t_\gamma) = \gamma$;

— при неизвестной дисперсии σ^2_x по формуле:

$$\delta = t_\gamma (k) \cdot \frac{S_x}{\sqrt{n}} = t_\gamma (k) \cdot S_{\bar{x}} , \qquad (6.20)$$

где $t_\gamma(k)$ — коэффициент Стьюдента, который находится по заданным доверительной вероятности γ и числу степеней свободы $k = n-1$ по таблице распределения Стьюдента.

Примечание. По определению, коэффициент Стьюдента $t_\gamma(k)$ — это такое постоянное число, зависящее от γ и k, при котором случайная величина $t = \dfrac{\bar{x}_в - m_x}{S_x/\sqrt{n}} = \dfrac{\bar{x}_в - m_x}{S_{\bar{x}}}$, распределенная по закону Стьюдента, удовлетворяет условию $P\{|t| < t_\gamma(k)\} = P\{-t_\gamma(k) < t < t_\gamma(k)\} = \gamma$.

Пример 4. В результате выборочного статистического исследования 5 предприятий отрасли получены следующие данные о стоимости их основных производственных фондов X (млн. руб.): ***3; 5; 5; 7*** и ***10***. Найти:

— точечные оценки средней стоимости фондов, дисперсии, среднего квадратичного отклонения;

— доверительный интервал для средней стоимости при доверительной вероятности $\gamma = 0{,}95$, считая, что выборка взята из нормально распределенной генеральной совокупности для двух случаев: среднее квадратичное отклонение стоимости фондов известно и равно ***2,5*** (млн руб.); дисперсия стоимости фондов неизвестна.

Решение

1. Вычислим сначала сумму значений x_i и x_i^2 (вместо $\sum_{i=1}^{n}$ условимся писать просто Σ):

$$\sum x_i = 30 \text{ и } \sum x_i^2 = 9+25+25+49+100=208.$$

Находим выборочное среднее: $\bar{x}_в = \dfrac{1}{n}\sum x_i = \dfrac{1}{5}\cdot 30 = 6$ *(млн руб.)*

Вычисляем исправленную выборочную дисперсию S_x^2:

$$S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x}_в)^2 = \frac{1}{5-1}\left[(3-6)^2 + (5-6)^2 + (5-6)^2 + (7-6)^2 + (10-6)^2\right] = \frac{28}{4} = 7.$$

Удобно вычислять S_x^2 по другой формуле:

$$S_x^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\cdot \bar{x}_в^2\right) = \frac{1}{4}(208 - 5\cdot 6^2) = \frac{1}{4}(208-180) = \frac{28}{4} = 7.$$

Находим оценку среднего квадратичного отклонения σ_x:

$$S_x = \sqrt{S_x^2} = \sqrt{7} \approx 2{,}65 \text{ (млн руб.)}.$$

2. Находим доверительный интервал для средней стоимости фондов:

• Найдем $t_\gamma = t_{0,95}$. Из соотношения $2\Phi(t_\gamma) = \gamma$ или $2\Phi(t_{0,95}) = 0{,}95$ получим $\Phi(t_{0,95}) = 0{,}95/2 = 0{,}475$. По таблице для функции Лапласа $\Phi(t)$ нахо-

дим $t_{0,95}=1,96$. По формуле (6.18) вычислим полуширину δ доверительного интервала: $\delta = t_\gamma \cdot \frac{\sigma_x}{\sqrt{n}} = 1,96 \cdot \frac{2,5}{\sqrt{5}} \approx 2,2.$

Найдем границы доверительного интервала:

$$\bar{x}_в - \delta = 6 - 2,2 = 3,8; \quad \bar{x}_в + \delta = 6 + 2,2 = 8,2.$$

Итак, при доверительной вероятности $\gamma = 0,95$ средняя стоимость основных производственных фондов всех предприятий отрасли заключено в доверительном интервале (*3,8; 8,2*) (в млн руб.).

• По таблице критических точек *t*-распределения Стьюдента при доверительной вероятности $\gamma = 0,95$ и числе степеней свободы $k = n-1 = 5-1 = 4$ находим $t_\gamma(k)$: $t_{0,95}(4) = 2,78$.

Вычисляем полуширину доверительного интервала:

$$\delta = t_{\gamma,k} \cdot \frac{S_x}{\sqrt{n}} = 2,78 \cdot \frac{2,65}{\sqrt{5}} \approx 3,3.$$

Запишем доверительный интервал для m_x при надежности $\gamma = 0,95$:

$$6 - 3,3 < m_x < 6 + 3,3, \text{ или } 2,7 < m_x < 9,3 \text{ (в млн руб).}$$

Доверительным интервалом для оценки неизвестного среднего квадратичного отклонения σ_x нормально распределенной случайной величины X с заданной надежностью γ являются интервалы:

$$S_x(1-q) < \sigma_x < S_x(1+q) \text{ (при } q<1) \text{ и} \qquad (6.21)$$

$$0 < \sigma_x < S_x(1+q) \text{ (при } q>1), \qquad (6.22)$$

где S_x — исправленное среднее квадратичное отклонение, а величина q находится по специальной таблице для $q_\gamma(n)$.

Пример 5. По выборке объема $=20$ найдено исправленное среднее квадратичное отклонение $S_x=4$. Найти доверительный интервал, накрывающий среднее квадратичное отклонение σ_x с надежностью *0,95*.

Решение. По таблице для $q_\gamma(n)$ по данным $\gamma=0,95$ и $n=20$ найдем $q_{0,95}(20)=0,37$. Так как $q<1$, то искомый доверительный интервал таков:

$$4(1-0,37) < \sigma_x < 4(1+0,37), \text{ или } 2,52 < \sigma_x < 5,48.$$

6.6. Доверительный интервал для оценки вероятности по относительной частоте

Пусть проводятся n независимых испытаний с целью оценки неизвестной постоянной вероятности p появления события A в каждом испытании. Например, определяют долю (удельный вес) единиц в генераль-

ной совокупности, обладающих (событие A) некоторым *качественным* признаком (годность детали, признак X больше или меньше заданного значения, семейное положение обследуемого и т. п.).

В качестве точечной оценки неизвестной вероятности p принимают относительную частоту (выборочную долю) появления события A $\omega=m/n$, где m — число появлений события в n испытаниях. Например, если из 100 деталей выборки ($n=100$) 95 деталей оказались стандартными ($m=95$), то выборочная доля равна $\omega=95/100=0,95$, то есть вероятность того, что случайно выбранная деталь окажется стандартной, приближенно равна $0,95$.

Относительная частота ω — несмещенная оценка p, то есть ее математическое ожидание равно оцениваемой вероятности. Действительно, учитывая, что случайная величина m распределена по биномиальному закону (3.4) с математическим ожиданием $M[m]=n\cdot p$, получим

$$M(\omega)=M[m/n]=M[m]/n=np/n=p. \qquad (6.23)$$

Дисперсия относительной частоты как оценки вероятности равна:

$$D[\omega]=D[m/n]=D[m]/n^2=npq/n^2=pq/n, \qquad (6.24)$$

так как $D[m]=np(1-p)= npq.$

Среднеквадратическое отклонение оценки равно:

$$\sigma_\omega = \sqrt{D[\omega]} = \sqrt{pq/n}\,. \qquad (6.25)$$

Доверительный интервал для оценки вероятности p находится исходя из предположения, что относительная частота ω распределена приближенно по нормальному закону с $M(\omega)=p$ и $D[\omega]= pq/n$ (что хорошо выполняется, если n достаточно велико и вероятность p не очень близка к нулю и единице). Тогда согласно формуле для вероятности того, что абсолютная величина отклонения нормальной случайной величины X от m_x не превысит положительного числа δ, будем иметь приближенное равенство:

$$P(|\omega - p| < \delta) \approx 2\Phi(\delta/\sigma_\omega) = \gamma\,,$$

где γ — доверительная вероятность.

Учитывая, что $\sigma_\omega = \sqrt{pq/n}$, получим

$$p(|\omega - p| < \delta) \approx 2\Phi(\delta\sqrt{n}/\sqrt{pq}) = 2\Phi(t) = \gamma\,,$$

где $t = \delta\sqrt{n}/\sqrt{pq}$. Отсюда $\delta = t\sqrt{pq/n}$.

Таким образом, с надежностью γ выполняется неравенство

$$|\omega - p| < t\sqrt{pq/n}\,. \qquad (6.26)$$

Поскольку вероятность p неизвестна, решим это неравенство относительно p. Для этого возведем его в квадрат и решим полученное равносильное квадратное неравенство $p^2(1+t^2/n)-2(\varpi+t^2/n)p+\varpi^2<0$.

Решением последнего неравенства будет искомый доверительный интервал $(p_1; p_2)$ с приближенными концами

$$p_1=\frac{n}{t^2+n}(\alpha+\beta) \quad \text{и} \quad p_2=\frac{n}{t^2+n}(\alpha-\beta),$$

где $\alpha=\omega+\frac{t^2}{2n}$; $\beta=t\sqrt{\frac{\omega(1-\omega)}{n}+\left(\frac{t}{2n}\right)^2}$, а t — такое значение аргумента функции Лапласа $\Phi(t)$, при котором $2\Phi(t)=\gamma$.

При больших $n(n>10^2)$ слагаемые $\frac{t^2}{2n}$ и $\left(\frac{t}{2n}\right)^2$ очень малы, а множитель $n/(t^2+n)\approx 1$, поэтому в качестве приближенных границ доверительного интервала можно взять

$$p_1\approx\omega-t\sqrt{\omega(1-\omega)/n} \quad \text{и} \quad p_2\approx\omega+t\sqrt{\omega(1-\omega)/n}.$$

Пример 6. Среди выборочно обследованных ***500*** семей города по уровню душевого дохода малообеспеченных оказалось ***150*** семей. Требуется с надежностью ***0,99*** определить долю малообеспеченных семей во всем городе.

Решение. Выборочная доля малообеспеченных семей равна:

$$\omega=150/500=0,3.$$

По таблице для функции Лапласа $\Phi(t)$ для доверительной вероятности $\gamma=0,99$ находим $t=2,58$. Вычисляем границы доверительного интервала $(p_1; p_2)$:

$$p_1\approx 0,3-2,58\sqrt{0,3(1-0,3)/500}=0,247;$$

$$p_2\approx 0,3+2,58\sqrt{0.3(1-0,3)/500}=0,353.$$

Таким образом, с вероятностью ***0,99*** можно утверждать, что доля p малообеспеченных семей среди всех семей города находится в интервале

$$0,247<p<0,353 \text{ или } 24,7\%<p<35,3\ \%.$$

6.7. Другие выборочные характеристики

Модой m_0 называют варианту, которая имеет наибольшую частоту.

Медианой m_e называют такое значение признака X, когда одна половина значений меньше ее, а вторая половина — больше. Для определения медианы сначала ранжируют выборку, то есть располагают дан-

ные в порядке возрастания или убывания. Если объем выборки нечетен, то есть $n=2k+1$ $(k=0, 1, 2, \ldots)$, то медиана равна $(k+1)$ члену ранжированной выборки $m_e=x_{k+1}$; при четном числе $n=2k$ медиана равна $m_e=(x_k+x_{k+1})/2$.

Размахом вариации R называют разность между наибольшей и наименьшей вариантами:

$$R=x_{max}-x_{min}.$$

Средним абсолютным отклонением d называют среднее арифметическое абсолютных отклонений отдельных вариант от среднего выборочного:

$$d=\left(\sum_{i=1}^{k} n_i\left|x_i-\bar{x}_в\right|\right)/\sum_{i=1}^{k} n_i\ ;$$

оно служит для характеристики рассеивания вариант относительно $\bar{x}_в$.

Коэффициентом вариации ν называют выраженное в процентах отношение исправленного выборочного среднего квадратичного отклонения S_x к среднему выборочному $\bar{x}_в$

$$\nu = S_x / \bar{x}_в \cdot 100\,\%.$$

Коэффициент вариации служит для сравнения величин рассеяния по отношению к выборочной средней двух вариационных рядов. Коэффициент вариации — безразмерная величина, поэтому он пригоден для сравнения рассеяний вариационных рядов, варианты которых имеют различную размерность, например, если варианты одного ряда выражены в сантиметрах, а другого — в граммах.

Коэффициент вариации используют и для характеристики однородности выборочной совокупности. Совокупность считается количественно однородной, если коэффициент вариации не превышает ***33 %.***

Задачи для самостоятельного решения

1. Имеются следующие данные о размерах основных фондов (в млн руб.) ***30*** предприятий:
   4,2; 2,4; 4,9; 6,7; 4,5; 2,7; 3,9; 2,1; 5,8; 4,0;
   2,8; 7,3; 4,4; 6,6; 2,0; 6,2; 7,0; 8,1; 0,7; 6,8;
   9,4; 7,6; 6,3; 8,8; 6,5; 1,4; 4,6; 2,0; 7,2; 9,1.
   Составить интервальный статистический ряд с шириной интервалов ***2*** (млн руб.); построить гистограмму.
2. Для изучения распределения новорожденных по весу были собраны данные для ***100*** детей и составлен интервальный статистический ряд:

Интервалы, кг	***1,0—2,0***	***2,0—2,5***	***2,5—3,0***	***3,0—3,5***	***3,5—4,0***	***4,0—5,0***
Относительная частота,	***0,03***	***0,05***	***0,15***	***0,35***	***0,28***	***0,14***

ω_i						

Построить гистограмму, найти выборочное среднее и выборочное среднее квадратичное отклонение веса новорожденных.

Указание. Найти середины интервалов и принять их в качестве значений x_i для расчета числовых характеристик.

3. Возраст ***20*** случайно отобранных студентов представляется следующими данными: ***17, 20, 18, 19, 18, 17, 20, 21, 24, 22, 20, 21, 20, 19, 18, 18, 18, 20, 21, 19***. Составить вариационный ряд, статистический дискретный ряд распределения. Построить график эмпирической функции распределения и полигон относительных частот.
4. Подбросить игральную кость ***50*** раз.
   - Составить дискретный статистический ряд распределения числа очков на верхней грани очков и сравнить его с законом распределения числа очков при одном подбрасывании кости.
   - Построить полигон относительных частот и сравнить его с многоугольником распределения вероятностей.
   - Найти выборочное среднее $\bar{x}_в$ числа очков и сравнить его с математическим ожиданием m_x числа очков, равного ***3,5***.
5. По данным задачи ***1*** вычислить выборочную среднюю размера основных фондов ***30*** предприятий двумя способами: по исходным данным; по интервальному статистическому ряду, приняв за значение размера основных фондов середины интервалов. Вычислить по второму способу выборочную и исправленную выборочную дисперсии.
6. Имеются данные распределения предприятий области по росту производительности труда (в % к предыдущему году):

Рост производительности труда, %	***80—90***	***90—100***	***100—110***	***110—120***	***120—130***
Число предприятий	***2***	***14***	***60***	***20***	***4***

   - Составить интервальный статистический ряд.
   - Преобразовать его в дискретный статистический ряд, взяв середины интервалов за значения роста производительности труда.
   - Найти несмещенные оценки математического ожидания и дисперсии роста производительности труда.
7. Найти доверительный интервал с доверительной вероятностью γ для неизвестного математического ожидания m_x нормально распределенной случайной величины X, если известна дисперсия σ_x^2 и найдено выборочное среднее $\bar{x}_в$ по выборке объема n:
   1) $\gamma=0,95;\ \sigma_x^2=4;\ \bar{x}_в=10;\ n=25.$
   2) $\gamma=0,95;\ \sigma_x^2=4;\ \bar{x}_в=10;\ n=16.$
   3) $\gamma=0,99;\ \sigma_x^2=4;\ \bar{x}_в=10;\ n=25.$

4) $\gamma=0,95;\ \sigma_x^2=25;\ \bar{x}_в=14;\ n=25.$

8. Известен объем n выборки для нормально распределенной случайной величины X, выборочное среднее $\bar{x}_в$ и исправленное среднее квадратичное отклонение S_x. Найти доверительный интервал для математического ожидания m_x с доверительной вероятностью γ:
   1) $n=16;\ \bar{x}_в=4,20;\ S_x=0,40;\ \gamma=0,95.$
   2) $n=25;\ \bar{x}_в=4,20;\ S_x=0,40;\ \gamma=0,95.$
   3) $n=16;\ \bar{x}_в=4,20;\ S_x=0,40;\ \gamma=0,99.$
   4) $n=16;\ \bar{x}_в=4,20;\ S_x=0,40;\ \gamma=0,999.$
9. Найти минимальный объем выборки, при котором с вероятностью $0,95$ точность оценки математического ожидания нормально распределенного признака по выборочной средней будет $\delta=0,2$, если среднее квадратичное отклонение равно $\sigma_x=2$.
10. Считая Вашу учебную группу малой выборкой из совокупности всех студентов университета, оценить с доверительной вероятностью $0,95$:
    — средний рост;
    — среднюю массу тела студентов университета (отдельно для юношей и девушек).
11. Произведено пять независимых измерений концентрации сахара (в %) в растворе. Получены следующие результаты: $2,15;\ 2,18;\ 2,14;\ 2,16;\ 2,17$. Предполагая, что результаты измерений распределены по нормальному закону, найти $95\ \%$ доверительные интервалы для математического ожидания и среднего квадратичного отклонения концентрации сахара.
12. Для оценки качества закупаемого сырья случайным образом были отобраны 30 образцов, в каждом из которых измерялась концентрация X (в %) некоторого вещества. Результаты измерений приведены в таблице.

x_i, %	3	5	6	7
Число образцов m_i	5	14	10	1

Найти: выборочное среднее; выборочную и исправленную выборочную дисперсии; исправленное среднее квадратичное отклонение, коэффициент вариации; моду, медиану, среднее абсолютное отклонение.

13. Найти $99\ \%$ доверительный интервал для оценки неизвестной вероятности p биномиального распределения, если в 60 испытаниях событие появилось 24 раза.
14. Среди выборочно обследованных 200 семей города по числу детей семей с тремя и более детьми оказалось 50. Требуется с надежно-

стью ***0,95*** определить долю семей с тремя и более детьми во всем городе.

7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

7.1. Основные понятия и определения

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений. Например:

— генеральная совокупность распределена по нормальному закону;

— дисперсии двух нормальных генеральных совокупностей равны между собой и т.д.

Нулевой (основной) гипотезой H_0 называется выдвинутая гипотеза. Это всегда гипотеза об отсутствии различия (нулевое различие) между законами или параметрами распределения случайных величин.

Конкурирующей гипотезой H_1 называется гипотеза, противоположная нулевой, то есть гипотеза о том, что различие есть.

Для проверки статистической гипотезы используют статистические (вероятностные) методы, поэтому решение, принимаемое в итоге проверки, может оказаться ошибочным. Ошибка, совершаемая при отклонении правильной нулевой гипотезы H_0, называется ошибкой первого рода. Вероятность ошибки первого рода обозначается α. Обычно α берут достаточно малым: $\alpha=0,05;\ 0,01;\ 0,001$. Ошибка второго рода — это принятие неправильной нулевой гипотезы, а вероятность этой ошибки обозначается β.

Проверка статистической гипотезы основывается на принципе, в соответствии с которым маловероятные события $(p<<1)$ считаются практически невозможными.

Предельно большое значение вероятности события, при котором оно еще может считаться практически невозможным, называется уровнем значимости и обозначается α (по смыслу то же, что и вероятность ошибки первого рода).

Для того чтобы проверить выдвинутую нулевую гипотезу по экспериментальным данным, необходимо вычислить вероятность того, что наблюдаемое различие между тем, что должно получиться при условии справедливости H_0 и экспериментом, случайно. Если эта вероятность больше уровня значимости α, то нулевая гипотеза не отвергается.

Если же эта вероятность окажется очень малой, меньше уровня значимости α, то нулевая гипотеза отвергается. Это означает, что наблюдаемое различие практически невозможно при справедливости нулевой гипотезы H_0, то есть различие вызвано реальными причинами.

Необходимо отметить, что принятие нулевой гипотезы не равносильно утверждению, что отсутствие различия доказано. Это означает, что данные опыта не противоречат предположению об отсутствии различия.

Статистический критерий проверки нулевой гипотезы. Вычисление вероятности того или иного различия (в законах или параметрах распределений) довольно сложно. Поэтому практически проверку H_0 осу-

ществляют с помощью различных статистических критериев. В качестве критерия используют некоторую случайную величину (например, t, F и т.п.).

Практическим (наблюдаемым) значением критерия (например, $t_{пр}$) называется его значение, вычисленное по экспериментальным данным.

Критической областью называется множество значений критерия, при которых нулевую гипотезу H_0 отвергают. При этом вероятность попаданий значения критерия в критическую область при условии истинности нулевой гипотезы H_0 не больше уровня значимости α.

Областью допустимых значений называется множество значений критерия, при которых нулевую гипотезу принимают.

Критическими точками (границами) называются значения критерия, отделяющие критическую область от области допустимых значений (например, $t_{кр}$ и т.п.).

Для каждого статистического критерия составлены заранее таблицы (см. приложения), в которых приведены значения критических точек, соответствующих различным уровням значимости α и степеням свободы k (k равно числу экспериментальных данных минус число параметров, вычисленных по ним и используемых при расчетах).

Правило проверки H_0 с помощью статистического критерия: если вычисленное практическое значение критерия будет принадлежать критической области, то нулевую гипотезу отвергают; в противном случае нулевая гипотеза принимается.

Между статистическими критериями и доверительными интервалами существует тесная связь: если принимается гипотеза о том, что значение проверяемого параметра, например m_x равно определенному значению (m_0) с уровнем значимости α, то это эквивалентно заданию $100(1-\alpha)=100\gamma$-ного доверительного интервала для данного параметра с центром в точке m_0. И наоборот, если доверительный интервал для m_x с надежностью γ включает в себя значение m_0, это эквивалентно принятию гипотезы о равенстве $m_x = m_0$ с уровнем значимости $\alpha=1-\gamma$.

7.2. Сравнение двух дисперсий нормально распределенных генеральных совокупностей

На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, инструментов, методов измерений и т.д. Очевидно, предпочтительнее тот прибор, инструмент или метод, который обеспечивает наименьшее рассеивание результатов измерений, то есть наименьшую дисперсию.

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам с объемами n_x и n_y, находят $\bar{x}$, $\bar{y}$, S_x^2, S_y^2.

Требуется по S_x^2 и S_y^2 при заданном уровне значимости α проверить нулевую гипотезу H_0 о равенстве генеральных дисперсий: $\sigma_x^2 = \sigma_y^2$ или $M[S_x^2] = M[S_y^2]$ при конкурирующей двусторонней гипотезе H_1: $\sigma_x^2 \neq \sigma_y^2$, то есть определить, значимо или незначимо отличаются S_x^2 и S_y^2.

Примечание. Гипотеза H_1: $\sigma_x^2 \neq \sigma_y^2$ называется двусторонней, так как она предполагает, что различие между величинами может быть любого знака, то есть $\sigma_x^2 - \sigma_y^2 > 0$ или $\sigma_x^2 - \sigma_y^2 < 0$.

В качестве критерия значимости различия исправленных дисперсий принимается их отношение F — критерий Фишера:

$$F_{пр} = \frac{S_1^2}{S_2^2}, \qquad (7.1)$$

где S_1^2 — бóльшая из исправленных дисперсий S_x^2 и S_y^2, S_2^2 — меньшая из них.

Случайная величина F при условии нормальности распределения совокупностей X и Y и справедливости нулевой гипотезы H_0: $\sigma_x^2 = \sigma_y^2$ имеет распределение Фишера с числами степеней свободы числителя $k_1=n_1-1$ и знаменателя $k_2=n_2-1$.

Критическое значение $F_{кр}$, зависящее от уровня значимости α и чисел степеней свободы k_1 и k_2 и удовлетворяющее условию

$$P\{F > F_{кр}\} = P\{F > F_\alpha(k_1; k_2)\} = \alpha,$$

находят по таблице критических точек распределения Фишера при уровне значимости $\alpha/2$ (вдвое меньше заданного): $F_{кр}=F_{\alpha/2}(k_1; k_2)$.

Примечание: При односторонних конкурирующих гипотезах H_1: $\sigma_x^2 > \sigma_y^2$ или H_1: $\sigma_x^2 < \sigma_y^2$ $F_{кр}= F_\alpha(k_1; k_2)$.

Правило проверки нулевой гипотезы с помощью критерия Фишера следующее: если $F_{пр}>F_{кр}$, то нулевая гипотеза отвергается; если $F_{пр}\leq F_{кр}$, то нет оснований отвергнуть нулевую гипотезу, то есть в этом случае, очевидно, $\sigma_x^2 = \sigma_y^2$, а S_x^2 и S_y^2 отличаются незначимо.

Пример 1. При измерении концентрации (в %) одного и того же вещества разными методами были получены следующие оценки дисперсий: при первом методе — $S_x^2 = 6{,}52$ (число измерений $n_x=15$), при втором методе — $S_y^2 = 11{,}41$ (число измерений $n_y=12$). Предполагая независимость и нормальность распределения результатов измерения концентрации, при уровне значимости $\alpha=0{,}02$ проверить нулевую гипотезу H_0: $\sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе H_1: $\sigma_x^2 \neq \sigma_y^2$, то есть определить, можно ли считать, что точности измерений двух методов существенно не отличаются.

Решение

1. Найдем отношение $F_{пр}$ бóльшей исправленной дисперсии (в нашем случае это S_y^2) к меньшей: $F_{пр}=S_y^2/S_x^2=11{,}41/6{,}52=1{,}75$.

2. По таблице критических точек распределения Фишера (см. приложение 6) по $\alpha/2=0{,}01$ и числам степеней свободы числителя $k_1=n_y-1=12-1=11$ и знаменателя $k_2=n_x-1=15-1=14$ находим критическую точку: $F_{кр}=F_{0{,}01}(11;\ 14)=3{,}86$.

3. Сравним $F_{пр}$ и $F_{кр}$: $F_{пр}<F_{кр}$, следовательно, нет оснований отвергнуть нулевую гипотезу, то есть по точности измерения оба метода существенно не различаются.

7.3. Сравнение предполагаемого математического ожидания m_0 нормальной генеральной совокупности с выборочным средним $\bar{x}_в$

Пусть из некоторой нормально распределенной генеральной совокупности X взята выборка объемом n: $x_1, x_2, \ldots, x_n$. Будем считать, что дисперсия генеральной совокупности неизвестна.

Требуется при заданном уровне значимости α по вычисленной выборочной средней $\bar{x}_в$ проверить нулевую гипотезу H_0: $m_x=m_0$ (или $M[\bar{x}_в]=m_0$) при двусторонней конкурирующей гипотезе H_1: $m_x \neq m_0$, где m_0 — неизвестное, но предполагаемое математическое ожидание генеральной совокупности, из которой взята выборка, то есть требуется выяснить, значимо или незначимо различаются предполагаемое математическое ожидание и выборочное среднее.

В качестве статистического критерия проверки принимают случайную величину t — нормированное отклонение выборочного среднего:

$$t = \frac{\bar{x}_в - m_0}{S_{\bar{x}}}, \qquad (7.2)$$

где $S_{\bar{x}} = \dfrac{S_x}{\sqrt{n}}$ — исправленное среднеквадратичное отклонение выборочной средней. Величина t при условии справедливости нулевой гипотезы H_0: $m_x=m_0$ имеет распределение Стьюдента с $k=n-1$ степенями свободы.

Правило проверки H_0: $m_x=m_0$ при H_1: $m_x \neq m_0$ по t — критерию Стьюдента.

1. Вычислим практическое значение $t_{пр}$ критерия по формуле (7.2).

2. По таблице распределения Стьюдента (см. приложение 5) при уровне значимости α, помещенному в верхней строке таблицы, и числу степеней свободы $k=n-1$ найдем значение критической точки $t_{кр}=t_\alpha(k)$.

3. Сравним $|t_{пр}|$ и $t_{кр}$:

— если $|t_{пр}| \le t_{кр}$, то нет оснований отвергнуть нулевую гипотезу;

— если $|t_{пр}| > t_{кр}$, то нулевую гипотезу отвергаем, то есть $\bar{x}_в$ и предполагаемое m_0 значимо различаются и практически можно считать, что у генеральной совокупности, из которой взята выборка, другое математическое ожидание, чем мы предполагали.

При односторонних конкурирующих гипотезах $H_1: m_x > m_0$ или $H_1: m_x < m_0$ правило проверки остается таким же, только из таблицы критических точек t-распределения Стьюдента берут критические значения для уровня значимости α, помещенного в нижней строке таблицы.

Пример 2. Для проверки исправности работы станка-автомата, который должен прессовать таблетки массой $m_0=0,5$ *г*, наугад отобрано $n=25$ таблеток. Выборочное среднее $\bar{x}_в$ при этом оказалось равным $0,48$ *г* и исправленное среднеквадратичное отклонение $S_x=0,05$ *г*. Можно ли при уровне значимости $\alpha=0,01$ считать, что станок работает нормально, то есть справедлива ли нулевая гипотеза H_0: $M[\bar{x}_в]=0,5$ *г* при конкурирующей гипотезе H_1: $M[\bar{x}_в] \neq 0,5$?

Решение

1. Вычислим $t_{пр}$: $t_{пр} = \dfrac{\bar{x}_в - m_0}{S_x/\sqrt{n}} = \dfrac{0,48 - 0,5}{0,05/\sqrt{25}} = \dfrac{-0,02}{0,01} = -2$.

2. По таблице распределения Стьюдента найдем критическую точку $t_{кр}=t_\alpha(k)$:

$$t_{кр}=t_{0,01}(24)=2,80.$$

3. Так как $|t_{пр}| = |-2| = 2 < 2,80$, то есть $|t_{пр}| < t_{кр}$, то нет оснований считать, что станок неисправен — $\bar{x}_в$ и нормальная масса таблетки, равная $0,5$ *г*, несущественно отличаются друг от друга.

7.4. Сравнение двух выборочных средних произвольно распределенных генеральных совокупностей (большие независимые выборки)

Необходимо проверить при заданном уровне значимости α нулевую гипотезу $H_0: m_x=m_y$ или $M[\bar{x}_в] = M[\bar{y}_в]$ при двусторонней конкурирующей гипотезе $H_1: m_x \neq m_y$. Сами случайные величины X и Y хотя и могут быть распределены по любому закону, их выборочные средние $\bar{x}_в$ и $\bar{y}_в$ при больших объемах выборок ($n_x>30$, $n_y>30$) будут распределены приближенно нормально, а исправленные дисперсии S_x^2 и S_y^2 являются достаточно надежными оценками генеральных дисперсий. Поэтому нормирован-

ное отклонение разности выборочных средних $t = \dfrac{(\bar{x}_в - \bar{y}_в) - (m_x - m_y)}{S_{\bar{x}-\bar{y}}}$, при условии справедливости нулевой гипотезы H_0: $m_x=m_y$ равное $t = \dfrac{\bar{x}_в - \bar{y}_в}{S_{\bar{x}-\bar{y}}}$, приближенно имеет распределение Стьюдента с $k=n_x+n_y-2$ степенями свободы, так что и в этом случае для проверки по выборочным средним нулевой гипотезы H_0: $m_x=m_y$ используется t –критерий.

Правило проверки

1. Найдем $\bar{x}_в$, S_x^2, $\bar{y}_в$, S_y^2 (n_x и n_y — объемы выборок). Затем вычислим исправленное среднеквадратичное отклонение $S_{\bar{x}-\bar{y}}$ разности выборочных средних ($\bar{x}_в - \bar{y}_в$) по формуле:

$$S_{\bar{x}-\bar{y}} = \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}, \qquad (7.3)$$

согласно свойству дисперсии разности случайных величин:

$$D[X-Y]=D[X]+D[Y].$$

Далее вычислим $t_{пр}$:

$$t_{пр} = \frac{\bar{x}_в - \bar{y}_в}{S_{\bar{x}-\bar{y}}}. \qquad (7.4)$$

2. Найдем по уровню значимости α и числу степеней свободы $k=n_x+n_y-2$ из таблицы распределения Стьюдента критическое значение $t_{кр}=t_\alpha(k)$.

3. Нулевая гипотеза отвергается, если $|t_{пр}| > t_{кр}$; если же $|t_{пр}| \le t_{кр}$, то нулевая гипотеза не отвергается.

Пример 3. При исследовании влияния времени приема некоторого лекарственного препарата (до еды или после еды) на его концентрацию (в мг/л) в крови через 6 часов после приема получены следующие данные:

Время приема	Количество подопытных животных	Средняя концентрация, мг/л	Исправленное среднее квадратичное отклонение концентрации, мг/л
до еды	$n_x=40$	$\bar{x}_в=42$	$S_x=2,0$
после еды	$n_y=35$	$\bar{y}_в=38$	$S_y=2,3$

Можно ли на основании этих данных считать, что время приема лекарства существенно не влияет на концентрацию лекарства в крови (принять $\alpha=0,01$)? Другими словами, требуется при уровне значимости $\alpha=0,01$ проверить нулевую гипотезу H_0: $m_x=m_y$ при конкурирующей гипотезе H_1: $m_x \ne m_y$.

Решение

1. Вычислим $t_{пр}$:

$$t_{пр} = \frac{\bar{x}_в - \bar{y}_в}{S_{\bar{x}-\bar{y}}} = \frac{\bar{x}_в - \bar{y}_в}{\sqrt{S_x^2/n_x + S_y^2/n_y}} = \frac{42 - 38}{\sqrt{\frac{(2,0)^2}{40} + \frac{(2,3)^2}{35}}} \approx \frac{4}{\sqrt{0,25}} = 8.$$

2. По таблице распределения Стьюдента при $\alpha=0,01$ и $k=40+35-2=73$ найдем критическую точку $t_{кр}=t_{0,01}(73)=2,65$.

3. Сравним $|t_{пр}| = 8$ и $t_{кр}$: $|t_{пр}| > t_{кр}$. Следовательно, нулевая гипотеза H_0: $m_x=m_y$ отвергается, то есть время приема лекарства существенно влияет на его концентрацию в крови через 6 часов после его приема.

7.5. Сравнение выборочных средних двух нормально распределенных генеральных совокупностей, дисперсии которых неизвестны, но одинаковы (малые независимые выборки)

Требуется проверить при заданном уровне значимости α нулевую гипотезу о равенстве средних двух нормально распределенных генеральных совокупностей H_0: $m_x=m_y$ при двусторонней конкурирующей гипотезе H_1: $m_x \neq m_y$. Так как дисперсии считаются одинаковыми $\sigma_x^2 = \sigma_y^2 = \sigma^2$, то S_x^2 и S_y^2 являются двумя оценками одной и той же одинаковой дисперсии σ^2. При малых объемах выборок ($n_x<30$, $n_y<30$) оценки дисперсий генеральных совокупностей содержат большую статистическую погрешность. Поэтому для получения более точной оценки σ^2 по исправленным выборочным дисперсиям S_x^2 и S_y^2 находят их среднее арифметическое S^2:

$$S^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = \frac{\sum_{i=1}^{n_x}(x_i - \bar{x}_в)^2 + \sum_{i=1}^{n_y}(y_i - \bar{y}_в)^2}{n_x + n_y - 2}. \qquad (7.5)$$

Так как генеральные совокупности X и Y распределены по нормальному закону, то для проверки нулевой гипотезы о равенстве средних при конкурирующей гипотезе, заключающейся в их неравенстве, используют также критерий Стьюдента. Для этого вычислим $t_{пр}$ по формуле:

$$t_{пр} = \frac{\bar{x}_в - \bar{y}_в}{S_{\bar{x}-\bar{y}}}, \qquad (7.6)$$

где

$$S_{\bar{x}-\bar{y}} = \sqrt{\frac{S^2}{n_x} + \frac{S^2}{n_y}} = \sqrt{S^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}. \qquad (7.7)$$

Затем по таблице критических значений критерия Стьюдента по уровню значимости α и числу степеней свободы $k=n_x+n_y-2$ находим $t_{кр}=t_\alpha(k)$.

Если конкурирующая гипотеза H_1 состоит в том, что $m_x>m_y$ или $m_y>m_x$ (односторонняя гипотеза), то критическое значение из таблицы Стьюдента берется при уровне значимости α, помещенном в нижней строке таблицы.

Если $|t_{пр}| \le t_{кр}$, то нулевая гипотеза H_0 не отвергается и считают, что $m_x=m_y$. Если $|t_{пр}|> t_{кр}$, то нулевая гипотеза отвергается.

Пример 4. Завод, работая по обычной технологии в течение двух рабочих недель ($n_x=10$ дней), выпускал в сутки в среднем $\bar{x}_в = 16$ *кг* некоторого лекарственного препарата с $S_x=0,6$ *кг*. При работе по новой технологии завод за следующую неделю ($n_y=5$ дней) стал выпускать в сутки в среднем $\bar{y}_в=17,2$ *кг* этого же препарата с $S_y=0,7$ *кг*. При уровне значимости $\alpha=0,05$ проверить нулевую гипотезу $H_0: m_x=m_y$ при конкурирующей гипотезе $H_1: m_x<m_y$, то есть выпуск продукции по новой технологии существенно превышает выпуск при старой, обычной технологии. Предполагается, что X и Y распределены нормально и их дисперсий равны $(\sigma_x^2 = \sigma_y^2)$.

Решение

1. Вычислим сначала S^2:

$$S^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x+n_y-2} = \frac{9\cdot 0,6^2 + 4\cdot 0,7^2}{10+5-2} = \frac{3,24+1,96}{13} = 0,4.$$

Далее находим $S_{\bar{x}-\bar{y}}$: $S_{\bar{x}-\bar{y}} = \sqrt{0,4\left(\frac{1}{10}+\frac{1}{5}\right)} = \sqrt{0,12} \approx 0,35.$

Вычислим $t_{пр}$: $t_{пр} = \frac{\bar{y}_в - \bar{x}_в}{S_{\bar{x}-\bar{y}}} = \frac{17,2-16}{0,35} \approx 3,4$.

2. Гипотеза $H_1: m_x<m_y$ — односторонняя, поэтому берем значение $t_{кр}$ для $k=n_x+n_y-2=13$ из таблицы приложения 5 для $\alpha=0,05$, помещенного в нижней строке: $t_{кр}=t_{0,05}(13)=1,77$.

3. Так как $|t_{пр}| = 3,4 > t_{кр}$, то H_0 отвергается, то есть новая технология существенно отличается от обычной по количеству выпускаемого лекарственного препарата, причем $m_y>m_x$.

7.6. Проверка соответствия экспериментального распределения определенному теоретическому виду по критерию Пирсона (χ^2)

Для качественной проверки соответствия экспериментального распределения заданному теоретическому можно построить гистограмму и сравнить кривую, огибающую ее сверху, с теоретически рассчитанной кривой. Если отклонения велики, то распределение нельзя считать соответствующим теоретическому. Если же различие между огибающей кривой и графиком плотности распределения невелико, то необходим более строгий количественный анализ.

Проверка соответствия экспериментального распределения теоретическому распределению является задачей статистической проверки нулевой гипотезы H_0 о принадлежности эмпирического распределения к данному теоретическому виду.

Для сравнения эмпирического и теоретического распределения необходимо выбрать критерий соответствия. Рассмотрим часто применяемый для решения этой задачи критерий χ^2 (*хи-квадрат*). Он основан на сравнении эмпирических и теоретически ожидаемых частот.

Пусть $x_1, x_2, \ldots, x_n$ — выборка независимых наблюдений случайной величины X объемом n. Проверяется гипотеза H_0, утверждающая, что X имеет закон распределения $F(x)$. Процедура применения критерия χ^2 для проверки гипотезы H_0 состоит из следующих этапов:

1. По выборочным данным находят s оценок неизвестных параметров предполагаемого закона распределения $F(x)$ (если все параметры известны, то s равно нулю). Пуассоновский и экспоненциальный законы распределения характеризуются одним вычисляемым по выборке параметром $\bar{\lambda}\ (\bar{\mu})$, поэтому для таких случаев число $s=1$, для нормального распределения $s=2$, если m_x и σ_x неизвестны (или меньше 2 в случае, когда один или два параметра известны).

2. Если X — дискретная случайная величина, а k — число вариант или групп вариант, то определяют частоты m_i ($i = 1, 2, \ldots, k$), с которыми каждая варианта или группа вариант встречается в выборке.

Если X — непрерывная случайная величина, то разбивают область ее значений на k непересекающихся частичных интервалов $\Delta x_1, \Delta x_2, \ldots, \Delta x_k$, и определяют их частоты m_i ($i = 1, 2, \ldots, k$) — число значений выборки, попавших в каждый интервал. Очевидно, что в обоих случаях

$$\sum_{i=1}^{k} m_i = n$$

.

3. Используя предполагаемый закон распределения $F(x)$, вычисляют вероятности p_i ($i = 1, 2, \ldots, k$):

— появления вариант или группы вариант в случае, если X — дискретная случайная величина;

— попадания в каждый интервал Δx_i в случае, если X — непрерывная случайная величина.

Очевидно, что в обоих случаях $\sum_{i=1}^{k} p_i = 1$.

4. Умножая найденные значения вероятностей p_i на объем выборки, находят теоретически ожидаемые частоты $m_{iтеор}=np_i$.

5. Различие эмпирических m_i и теоретически вычисленных $n \cdot p_i$ частот характеризуется суммой квадратов их отклонений — наблюдаемым значением $\chi^2_{пр}$ критерия χ^2:

$$\chi^2_{пр} = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}. \qquad (7.8)$$

6. По таблице критических точек распределения χ^2 (см. приложение 4) по заданному уровню значимости α и числу степеней свободы

$$f=k-s-1 \qquad (7.9)$$

(s — число неизвестных параметров распределения, оцениваемых по выборке, k — число вариант (групп вариант) или интервалов) находят критическое значение $\chi^2_{кр} = \chi^2_{\alpha}(f)$.

7. Проверка нулевой гипотезы сводится к сравнению практически полученной величины $\chi^2_{пр}$ с критическим значением $\chi^2_{кр}$. Если $\chi^2_{пр} < \chi^2_{кр}$, то нулевая гипотеза о виде распределения не отвергается. Если $\chi^2_{пр} > \chi^2_{кр}$, то нулевая гипотеза отвергается — эмпирические и теоретические частоты различаются значимо.

Примечание. Применения критерия χ^2 ограничивается рядом допущений. Одно из самых важных — это то, что при больших объемах n выборок частота m_i приближенно распределена по нормальному закону с математическим ожиданием $n \cdot p_i$ и среднеквадратическим отклонением $\sigma_i = \sqrt{np_i}$. Тогда при условии справедливости нулевой гипотезы величина нормированного отклонения:

$$t_i = \frac{m_i - np_i}{\sqrt{np_i}} \qquad (7.10)$$

будет иметь распределение, близкое к нормированному нормальному с $m_x=0$ и $\sigma_x=1$. Чтобы эти утверждения были достаточно точными, необходимо, чтобы для всех частот выполнялось условие $m_i \geq 5$ ($np_i \geq 5$). Если для некоторых вариант (интервалов) это условие не выполняется, то их следует объединить с соседними, сложив соответствующие им частоты. Тогда при определении числа степеней свободы по формуле $f=k-s-1$ следует в качестве k принять число вариант (интервалов), оставшихся после объединения.

Пример 5. Число пациентов, посещающих аптеку за ***3*** минуты, распределено в течение ***7*** часов (всего ***n=140*** трехминутных интервалов) в соответствии с таблицей 1.

Таблица 7.1

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	≥13
m_i	0	3	8	9	20	24	19	21	12	11	7	5	1	0

При уровне значимости α**=0,05** проверить гипотезу о том, что число пациентов имеет распределение Пуассона.

Решение

1. В таблице 1 объединяем три первых значения в один интервал (***i ≤ 2***) и последние четыре тоже (***i ≥ 10***), чтобы m_i для них были больше 10. Количество ***k*** интервалов значений x_i стало равным ***9***:

x_i	i≤2	3	4	5	6	7	8	9	i ≥ 10
m_i	11	9	20	24	19	21	12	11	13

Вычислим общее число пациентов ***N*** и среднее число пациентов в трехминутный интервал (интенсивность потока пациентов) $\overline{\lambda} = N / n$:

$$N = \sum x_i \cdot m_i = 839 \text{ и } \overline{\lambda} = 839/140 \approx 6 \text{ (пациентов/мин).}$$

2. Для полученной интенсивности $\lambda = 6$ по формуле распределения Пуассона $P_i = P(X = i) = \dfrac{\lambda^i \cdot e^{-\lambda}}{i!}$, находим величины вероятностей P_i поступления ***i*** пациентов ***(i = 0, 1, 2, 3, …)*** за 3 минуты:

$$P_0 = \frac{6^i \cdot e^{-6}}{0!} \approx 0{,}00248 \approx 0{,}002;\ P_1=0{,}01488\approx 0{,}015;\ P_2=0{,}04464\approx 0{,}045,$$

тогда вероятность попадания в ***1***-й интервал (для ***i≤2***) будет равна:

$$P(i\le 2) = 0{,}002+0{,}015+0{,}045=0{,}062.$$

Далее имеем: $P_3=0{,}089;\ P_4=0{,}134;\ P_5=0{,}161;\ P_6=0{,}161;\ P_7=0{,}138;\ P_8=0{,}103;\ P_9=0{,}069.$ Вероятность поступления не менее ***10*** пациентов вычислим исходя из того, что $\sum_{i=0}^{\infty} P_i = 1$, откуда $P(i \ge 10) = 1 - \sum_{i=0}^{9} P_i = 1 - 0{,}917 = 0{,}083.$

3. Обозначая вероятность попадания в ***k***-й (***k=1, 2, 3, …, 9***) интервал малой буквой p_i (причем $p_1 = P(i \le 2)$, $p_2 = P_3$, $p_3 = P_4$, …, $p_9 = P(i \ge 10)$), вычисляем теоретически ожидаемые частоты $n \cdot p_k$ попадания в каждый интервал:

$n \cdot p_1 = 140 \cdot 0{,}062 = 8{,}68;$ $n \cdot p_2 = n \cdot P_3 = 140 \cdot 0{,}089 = 12{,}46;$ $n \cdot p_3 = n \cdot P_4 = 140 \cdot 0{,}134 = 18{,}76;$

$n \cdot p_4 = 22{,}54;$ $n \cdot p_5 = 22{,}54;$ $n \cdot p_6 = 19{,}32;$ $n \cdot p_7 = 14{,}42;$ $n \cdot p_8 = 9{,}66;$ $n \cdot p_9 = 11{,}62.$

Проверка: $\sum_{k=1}^{9} np_k =$

$8{,}68+12{,}46+18{,}76+22{,}54+22{,}54+19{,}32+14{,}42+9{,}66+11{,}62 = 140{,}00 = n.$

4. Вычисляем экспериментальное значение критерия χ^2:

$$\chi^2_{пр} = \sum_{i=1}^{9} \frac{(m_i - np_i)^2}{np_i} = \frac{(11-8{,}68)^2}{8{,}4} + \frac{(9-12{,}46)^2}{12{,}46} + \frac{(20-18{,}76)^2}{18{,}76} + \ldots + \frac{(13-11{,}62)^2}{10{,}648} = 3{,}318.$$

5. По таблице приложения 4 для числа степеней свободы $f=l-1-1=9-2=7$ и уровня значимости $\alpha=0{,}05$ находим $\chi^2_{кр} = 14{,}1$. Поскольку $\chi^2_{пр} < \chi^2_{кр}$, делаем вывод о соответствии экспериментального распределения пуассоновскому.

7.7. Непараметрические критерии

Непараметрическими являются критерии, свободные (независимые) от вида распределения. В них используют не сами численные значения элементов выборки, а структурные свойства выборки (например, отношения порядка между ее элементами). В связи с этим теряется часть информации, содержащаяся в выборке, поэтому, например, мощность (чувствительность) непараметрических критериев меньше, чем мощность соответствующих им параметрических критериев. Однако непараметрические критерии более просты с точки зрения вычислений.

Критерий знаков

Этот непараметрический метод используется для проверки нулевой гипотезы о том, что две выборочные совокупности взяты из одной и той же генеральной совокупности. При этом на вид распределения не накладывается никаких ограничений, предполагается только непрерывность распределения генеральной совокупности.

Критерий знаков не позволяет судить, по каким именно характеристикам различаются выборочные совокупности, поэтому он обладает меньшей чувствительностью (мощностью), чем критерий Стьюдента и Фишера, но в то же время проще в вычислениях, чем они.

Критерий знаков используется для сравнения зависимых, попарно связанных выборок, когда результаты эксперимента представлены в виде так называемых сопряженных пар. Таковы, например, результаты:

—анализа одних и тех же объектов до и после воздействия в разное время (суток, года, лет);

—измерения одних и тех же величин двумя разными методами (приборами) и других подобных исследований.

Нулевая гипотеза при этом состоит в том, что при всех значениях аргумента (обозначим его через x) функции распределения двух генеральных совокупностей отличаются незначимо $H_0: F_1(x)=F_2(x)$ (такие генеральные совокупности называются *однородными*). Конкурирующими могут являться гипотезы: а) $F_1(x)\neq F_2(x)$; б) $F_1(x)<F_2(x)$ то есть $X>Y$ или в) $F_1(x)>F_2(x)$, то есть $X<Y$.

Замечание. Принятие конкурирующей гипотезы $F_1(x)<F_2(x)$ означает, что $X>Y$. Действительно, неравенство $F_1(x)<F_2(x)$ равносильно неравенству $P(X<x)<P(Y<x)$, откуда $P(X>x)>P(Y>x)$. Другими словами, вероятность того, что случайная величина X превзойдет определенное число x больше, чем вероятность случайной величине Y оказаться большей x — в этом смысл $X>Y$. Аналогично, если справедлива гипотеза $F_1(x)>F_2(x)$, то $X<Y$.

Пусть, например, $x_1, x_2,\ldots, x_n$ и $y_1, y_2,\ldots, y_n$ — данные анализа крови одних и тех же больных до и после применения некоторого лечебного препарата. Если сравниваемые выборки взяты из одной и той же генеральной совокупности (то есть воздействие препарата на кровь незначимо), то значения x_i и y_i ($i=1, 2, 3, \ldots, n$) взаимозаменяемы и, следовательно, вероятности появления положительных и отрицательных разностей (x_i-y_i) равны:

$$P(x_i - y_i > 0) = P(x_i - y_i < 0) = \frac{1}{2}. \qquad (7.11)$$

Кроме того, в силу предполагаемой непрерывности распределения X и Y, $x_i\neq y_i$, то есть вероятность появления нулевой разности равна нулю. Нулевые разности на практике могут появиться из-за случайных погрешностей или ошибок округления, такие пары наблюдений исключаются из дальнейшего рассмотрения и объем n парной выборки при этом уменьшается.

Если число знаков «+» близко к числу знаков «–», то, очевидно, различие между выборками незначимо; если же преобладает один из знаков, то различие может оказаться существенным.

Статистикой критерия знаков является m — число знаков «+» или «–», встречающихся менее часто (отсюда название критерия). При условии, что проверяемая гипотеза H_0 верна, число знаков m имеет биномиальное распределение с параметрами $p = P("+") = P("-") = \frac{1}{2}$ и n — чис-

ло ненулевых разностей. Вероятность того, что из n знаков встретится ровно m знаков («+» или «–»), вычисляется по формуле Бернулли:

$$P_n(m) = C_n^m \cdot p^m \cdot (1-p)^{n-m} = C_n^m \cdot \left(\frac{1}{2}\right)^n = \frac{n!}{m!(n-m)!} \cdot \left(\frac{1}{2}\right)^n . \qquad (7.12)$$

С помощью этой формулы можно найти вероятность того, что количество знаков «+» или «–», менее часто встречающихся, меньше или равно получившемуся (практическому) числу $m_{пр}$:

$$2 \cdot P(0 \le m \le m_{пр}) = 2(P_n(0) + P_n(1) + \ldots + P_n(m_{пр})) = 2\sum_{m=0}^{m_{пр}} P_n(m) \qquad (7.13)$$

(цифра «**2**» стоит потому, что $m_{пр}$ может быть числом любых знаков; если $m_{пр}$ — число определенных знаков, то без «двойки»).

Нулевая гипотеза не отклоняется, если эта вероятность будет больше уровня значимости:

$$2 \cdot P(0 \le m \le m_{пр}) > \alpha \quad \text{или} \quad P(0 \le m \le m_{пр}) > \frac{\alpha}{2}, \qquad (7.14)$$

и отклоняется, если меньше:

$$2 \cdot P(0 \le m \le m_{пр}) < \alpha \quad \text{или} \quad P(0 \le m \le m_{пр}) < \frac{\alpha}{2}. \qquad (7.15)$$

Практически эти вероятности не вычисляют, а сравнивают $m_{пр}$ с критическими значениями $m_{кр} = m_{\alpha}(n)$ числа менее часто встречающихся знаков — это наименьшее целое число, удовлетворяющее неравенству $2 \cdot P(0 \le m \le m_{кр}) \ge \alpha$ (если m — число определенных знаков, то без «двойки» в левой части неравенства). Значения $m_{кр}$ для разных α и n заранее рассчитаны и сведены в таблицу.

Правила проверки H_0 по критерию знаков:

1. Определить $m_{пр}$ — число знаков, менее встречающихся.

2. Подсчитать n — общее число знаков (нулевые разности отбрасываются).

3. По таблице критических значений числа менее часто встречающихся знаков для заданного уровня значимости α и найденного n найти $m_{кр} = m_{\alpha}(n)$.

4. Сравнить $m_{пр}$ и $m_{кр}$. Если $m_{пр} \ge m_{кр}$, то нулевая гипотеза не отвергается, если $m_{пр} < m_{кр}$, то нулевая гипотеза отвергается, то есть выборки не принадлежат одной генеральной совокупности.

Пример 6. При уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу о том, что прием больными некоторого гормонального препарата

не приводит к существенному изменению их массы тела по следующим данным:

Масса до начала приема x_i, кг	66	72	72	77	83	56	81	67	75	93	89	74	68	69	58	76
Масса через месяц после начала приема y_i, кг	68	75	73	79	84	59	82	67	76	89	92	75	69	67	63	78
Знаки ($x_i - y_i$)	–	–	–	–	–	–	–	0	–	+	–	–	–	+	–	–

Решение

1. Найдем знаки разностей парных наблюдений и подпишем их под таблицей. Всего знаков $n=15$ (одна нулевая разность), из них менее встречающихся знаков «+» $m_{пр}=2$.

2. По таблице критических значений по $\alpha=0,01$ и $n=15$ находим $m_{кр}=3$.

3. Так как $m_{пр}<m_{кр}$, то нулевая гипотеза отвергается, то есть прием данного препарата приводит к значимому изменению массы тела больного.

Можно в данном примере прямо вычислить вероятность полученного различия в числе знаков «+» и «–» и сравнить ее с уровнем значимости α:

$$P(0 \le m \le 2) = P_{15}(0) + P_{15}(1) + P_{15}(2) = (C_{15}^0 + C_{15}^1 + C_{15}^2)\cdot\left(\frac{1}{2}\right)^{15} = = 0.00369 < \frac{0,01}{2} = 0.005,$$

то есть H_0 отвергается.

Критерий Вилкоксона

Критерий Вилкоксона служит для проверки однородности двух независимых выборок $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$. Он применим к случайным величинам, распределения которых неизвестны, но обязательно непрерывны. Нулевая гипотеза состоит в том, что генеральные совокупности имеют одинаковые непрерывные функции распределения — H_0: $F_1(x)=F_2(y)$ (при $x=y$) или, обозначив аргумент через x, $F_1(x)=F_2(x)$. Конкурирующими могут быть следующие гипотезы: а)$F_1(x)\neq F_2(x)$, б) $F_1(x)<F_2(x)$, то есть $X>Y$ или в) $F_1(x)>F_2(x)$, то есть $X<Y$ (см. замечание в подразделе «Критерий знаков»).

Предполагается, что объем *первой* выборки не больше объема второй: $n_1 \le n_2$; если это не так, то выборки можно поменять местами.

1. Проверка нулевой гипотезы для малых выборок ($n_{1,2}<25$). Чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: F_1(x)=F_2(x)$ об однородности двух независимых выборок объемов n_1 и n_2 при конкурирующей гипотезе $H_1: F_1(x)\neq F_2(x)$, необходимо располо-

жить варианты обеих выборок в возрастающем порядке, то есть в виде одного вариационного ряда, и найти в этом ряду наблюдаемое значение критерия $W_{набл}$ — сумму порядковых номеров первой выборки. Затем найти по таблице критических точек критерия Вилкоксона нижнюю критическую точку $W_{н.кр}(\alpha/2; n_1; n_2)$ и верхнюю критическую точку по формуле:

$$W_{в.кр}=(n_1+n_2+1)\cdot n_1 - W_{н.кр}.$$

Если $W_{набл}<W_{н.кр}$ или $W_{набл}>W_{в.кр}$, то нулевую гипотезу отвергают.

Если $W_{н.кр}< W_{набл}< W_{в.кр}$, то нет оснований отвергнуть нулевую гипотезу.

При конкурирующей гипотезе $F_1(x)>F_2(x)$ находят по таблице нижнюю критическую точку $W_{н.кр}(\alpha; n_1; n_2)$. Если $W_{набл}>W_{н.кр}$, то нет оснований отвергнуть нулевую гипотезу. Если $W_{набл}<W_{н.кр}$, то нулевую гипотезу отвергают.

При конкурирующей гипотезе $H_1: F_1(x)<F_2(x)$ надо найти верхнюю критическую точку: $W_{в.кр}(\alpha; n_1; n_2)=(n_1+n_2+1)\cdot n_1-W_{н.кр}(\alpha; n_1; n_2)$. Если $W_{набл}<W_{в.кр}$, то нет оснований отвергнуть нулевую гипотезу, если $W_{набл}>W_{в.кр}$, то нулевую гипотезу отвергают.

Если несколько вариант одной выборки одинаковы, то в общем ряду им приписываются обычные порядковые номера, как если бы они были различными числами. Если совпадают варианты разных выборок, то всем им приписывают один и тот же порядковый номер, равный среднему арифметическому порядковых номеров, которые имели бы эти варианты до совпадения.

2. Проверка нулевой гипотезы для больших выборок — n_1 и (или) n_2 больше **25**. При конкурирующей гипотезе $F_1(x)\neq F_2(x)$ вычисляется нижняя критическая точка.

$$W_{н.кр}\left(\frac{\alpha}{2}; n_1; n_2\right) = int\left[\frac{(n_1+n_2+1)\cdot n_1 - 1}{2} - t_{кр.}\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}\right], \qquad (7.15)$$

где $t_{кр}$ находят по таблице функции Лапласа из равенства $\Phi(t_{кр})=(1-\alpha)/2$; $int\,[a]$ означает целую часть числа a.

В остальном правило проверки гипотезы соответствует п. 1.

При конкурирующих гипотезах $F_1(x)>F_2(x)$ и $F_1(x)<F_2(x)$ нижнюю критическую точку $W_{н.кр}(\alpha; n_1; n_2)$ находят по формуле (7.15), а соответствующий $t_{кр}$ — из равенства $\Phi(t_{кр})=(1-2\alpha)/2$. Остальные правила, приведенные в п. 1, сохраняются.

Пример 7. При уровне значимости **0,05** проверить нулевую гипотезу об однородности двух выборок объемом $n_1=6$ и $n_2=8$:

x_i	15	23	25	26	28	29	
y_i	12	14	18	20	22	24	27	30

при конкурирующей гипотезе $H_1: F_1(x) \neq F_2(x)$.

Решение. Расположим варианты обеих выборок в виде одного вариационного ряда и перенумеруем их:

— порядковые номера ***1 2 3 4 5 6 7 8 9 10 11 12 13 14;***

— варианты ***12 14 15 18 20 22 23 24 25 26 27 28 29 30.***

Найдем наблюдаемое значение критерия Вилкоксона — сумму порядковых номеров (они набраны курсивом) вариант первой выборки:
$W_{набл} = 3+7+9+10+12+13 = 54$.

Найдем по таблице критических точек критерия Вилкоксона нижнюю критическую точку $W_{н.кр}(0{,}025;\ 6,\ 8) = 29$.

Найдем верхнюю критическую точку:

$$W_{в.кр} = (n_1 + n_2 + 1) \cdot n_1 - W_{н.кр} = (6+8+1) \cdot 6 - 29 = 61.$$

Так как $29 < 54 < 61$, то есть $W_{н.кр} < W_{набл} < W_{в.кр}$, то нет оснований отвергнуть нулевую гипотезу об однородности выборок.

Пример 8. При уровне значимости ***0,01*** проверить нулевую гипотезу об однородности двух выборок объемом $n_1 = 30$ и $n_2 = 50$ при конкурирующей гипотезе $H_1: F_1(x) \neq F_2(x)$, если известно, что в общем вариационном ряду, составленном из вариант обеих выборок, сумма порядковых номеров вариант первой выборки $W_{набл} = 1600$.

Решение. По таблице функции Лапласа по равенству

$$\Phi(t_{кр}) = (1-\alpha)/2 = (1-0{,}01)/2 = 0{,}495$$

находим $t_{кр} = 2{,}58$.

Подставив $n_1 = 30$, $n_2 = 50$ и $t_{кр} = 2{,}58$ в формулу (7.15), получим:

$$W_{н.кр}(0{,}005;\ 30;\ 50) = 954.$$

Найдем верхнюю критическую точку:

$$W_{в.кр} = (n_1 + n_2 + 1) \cdot n_1 - W_{н.кр} = 2430 - 954 = 1476.$$

Так как $1600 > 1476$, то есть $W_{набл} > W_{в.кр}$, то нулевая гипотеза отвергается.

Задачи для самостоятельного решения

В задачах 1—11 предполагается, что выборки получены из нормально распределенных генеральных совокупностей.

1. Завод-изготовитель гарантирует среднее время m_x безотказной работы своего прибора не менее ***1000*** часов. Однако выборочное среднее время $\overline{X}$ безотказной работы ***25*** приборов, полученных предприятием, оказалось равным ***954*** часам с $S_x = 100$ часов. Можно ли считать, что

полученная партия из ***25*** приборов не удовлетворяет условиям гарантии, если уровень значимости α равен: а) ***0,01***; б) ***0,05*** ?

2. Контролер, в задачу которого входит наблюдение за правильностью фасовки некоторого химпрепарата, отобрал случайным образом десять упаковок и взвесил их. Он получил следующие результаты (в г): ***0,94; 0,95; 0,92; 1,02; 0,97; 0,95; 1,02; 0,96; 0,92; 0,97*** (масса упаковки по норме должна быть ***1,00*** г). Может ли контролер с уровнем значимости α=***0,01*** считать, что полученные данные не противоречат гипотезе о правильной фасовке химпрепарата?
3. Из многолетнего опыта лечения некоторого заболевания известно, что среднее число дней с повышенной температурой у больных с этим заболеванием равно ***12***, то есть m_x=***12*** дням. При клинической проверки действия нового лекарственного препарата для лечения этого же заболевания у случайно отобранных ***20*** больных были получены следующие значения числа дней с повышенной температурой: ***10, 7, 9, 14, 12, 9, 7, 13, 8, 15, 14, 10, 11, 16, 9, 10, 12, 11, 7, 6***. Можно ли с α=***0,01*** считать, что применение нового лекарственного препарата существенно уменьшает среднее число дней с повышенной температурой?
4. В эксперименте с крысами изучалось сравнительное влияние двух лекарств на количество соляной кислоты (HCl), выделяемой в желудке крысы. Первое лекарство давалось случайной выборке из n_x=***8*** крыс, второе — выборке из n_y=***10*** крыс. Получены следующие данные (количество HCl в соответствующих единицах):

1 лекарство, x_i	***5***	***6***	***8***	***2***	***3***	***1***	***1***	***4***	—	—
2 лекарство, y_i	***8***	***5***	***7***	***8***	***3***	***7***	***5***	***3***	***2***	***6***

Можно ли с α=***0,01*** утверждать, что проверяемые лекарства не различаются по своему влиянию на секрецию HCl в желудке крысы (считать, что $\sigma_x^2 = \sigma_y^2$)?

5. При исследовании зависимости времени химической реакции (в сек.) от концентрации ***C*** катализатора (в %) были получены следующие данные:

C, %	время	№ эксперимента							
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
2	x_i, с	***6,1***	***7,2***	***5,0***	***5,4***	***6,3***	***5,2***	***6,4***	***7,2***
4	y_i, с	***8,2***	***6,8***	***7,0***	***8,1***	***7,5***	***7,9***	***8,0***	***8,1***

При α=***0,01*** проверить нулевую гипотезу о том, что содержание катализатора не влияет на время химической реакции (предварительно проверить гипотезу о равенстве дисперсий H_0: $\sigma_x^2 = \sigma_y^2$).

6. Данные товарооборота за полугодие в двух аптеках составили следующую сводку (в млн руб.):

	Месяцы					
	1	***2***	***3***	***4***	***5***	***6***
Аптека *1*	***20***	***20***	***32***	***27***	***30***	***21***
Аптека ***2***	***20***	***18***	***23***	***18***	***17***	***18***

Можно ли с α=***0,01*** считать, что среднемесячные товарообороты двух аптек существенно не отличаются (предварительно проверить гипотезу о равенстве дисперсий)?

7. Используя данные из задачи 6, сравнить среднемесячный товарооборот каждой из двух аптек за первое и второе полугодия, если сводка за второе полугодие имеет вид:

	Месяцы					
	7	***8***	***9***	***10***	***11***	***12***
Аптека ***1***	***19***	***23***	***26***	***18***	***20***	***26***
Аптека ***2***	***16***	***15***	***18***	***26***	***17***	***16***

Считать, что дисперсии выборок равны и принять α=***0,01***.

8. При α=***0,01*** проверить:
   — нулевую гипотезу H_0: $\sigma_x^2 = \sigma_y^2$;
   — нулевую гипотезу H_0: $m_x=m_y$ при условии, что $\sigma_x^2 = \sigma_y^2$.

x_i	***3,0***	***3,5***	***3,4***	***3,1***	***3,4***	***3,1***	***2,9***	***3,2***
y_i	***2,6***	***2,8***	***2,7***	***2,5***	***2,6***	***2,4***	-	-

9. До наладки станка-автомата, прессующего таблетки, была проверена точность прессовки ***10*** таблеток и найдено значение S_x=***4,5*** *мг*. После наладки подвергли контролю еще ***12*** таблеток и получили следующие данные (в *мг*): ***1026, 1028, 1026, 1025, 1027, 1024, 1029, 1025, 1030, 1025, 1027, 1024.*** Можно ли с уровнем значимости α=***0,01*** считать, что в результате наладки станка-автомата точность изготовления таблеток увеличилась?
10. Чтобы определить, какое влияние оказывает температура окружающей среды на значение измеряемого показателя преломления вещества, было проведено n_x=***6*** измерений при ***t=18°C*** и n_y=***8*** измерений при t=***25°C***. Результаты измерений следующие:

t=18,	***1,382***	***1,384***	***1,387***	***1,381***	***1,38***	***1,388***	—	—

°C					***3***			
t=25, °C	***1,395***	***1,397***	***1,395***	***1,396***	***1,392***	***1,394***	***1,393***	***1,392***

Можно ли считать, что температура окружающей среды не влияет на средние значения измеряемого показателя преломления? Принять $\alpha=\boldsymbol{0,05}$.

11. Для производства некоторого препарата должна использоваться вода с малой жесткостью. Ожидается, что добавление специальных веществ уменьшает жесткость воды. Измерения жесткости до и после добавления этих веществ дали следующие результаты (в градусах жесткости):

Без добавки	***2,9***	***5,5***	***8,8***	***5,9***	***4,5***	***5,2***	***3,9***	***5,8***
После добавки	***3,5***	***5,5***	***3,0***	***2,6***	***2,2***	***3,1***	***2,3***	***4,3***

Подтверждают ли эти результаты ожидаемый эффект (принять $\alpha=\boldsymbol{0,05}$)?

В задачах 12—14 на вид распределения генеральных совокупностей никаких ограничений не накладывается, кроме непрерывности распределения.

12. Значения концентрации лекарственного вещества в крови (в *мг/л*) через ***6*** часов после приема в двух случаях, когда его запивали водой ($\boldsymbol{n_x=50}$) и молоком ($\boldsymbol{m_y=30}$), представлены в таблице:

Вода	$\boldsymbol{x_i}$, *мг/л*	***11***	***12***	***13***	***14***
	$\boldsymbol{n_i}$	***20***	***15***	***10***	***5***
Молоко	$\boldsymbol{y_i}$, *мг/л*	***12***	***14***	***15***	***16***
	$\boldsymbol{m_i}$	***8***	***9***	***10***	***3***

Можно ли при $\alpha=\boldsymbol{0,01}$ считать, что в среднем концентрация лекарственного вещества в крови не зависит от того, чем его запивать после приема?

13. При работе по старой технологии среднее содержание некоторого химического вещества (в %) в выпускаемой продукции, полученное по выборке $\boldsymbol{n_x=40}$ образцов, было равно $\boldsymbol{\overline{X}=18,4}$ % с $\boldsymbol{S_x=1,8}$ %. Предложена новая технология, которая должна увеличить содержание этого вещества в продукции. По выборке из $\boldsymbol{n_y=50}$ образцов, полученных при работе по новой технологии, получены следующие результаты:

$\boldsymbol{y}$, %	***18***	***20***	***22***	***24***

n_i	20	10	15	5

Можно ли считать, что новая технология оправдывает ожидание, если α=0,05?

14. В результате выборочного социологического исследования у $n_x=60$ провизоров, имеющих стаж работы до **5** лет, средняя зарплата оказалась равной $\bar{X}=1250$ руб. с $S_x=90$ руб., а у $n_y=50$ провизоров, имеющих стаж работы от **5** до **10** лет, средняя зарплата оказалась равной $\bar{Y}=1520$ руб. с $S_y=130$ руб. Существенно ли различается заработок провизора в зависимости от стажа (принять α=0,01)?

Для решения задач 15—17 использовать критерий Пирсона.

15. Число вызовов скорой помощи за смену распределилось по количеству вызовов за 5 минут следующим образом:

Число вызовов x_i	3	4	5	6	7	8	9	10	11	12	13
Число пятиминутных интервалов m_i	4	9	10	21	25	18	22	13	10	8	4

Проверить гипотезу о соответствии эмпирического распределения пуассоновскому с $\alpha=0,05$.

16. Число пациентов, обслуживаемых поликлиникой, распределено по длительности обслуживания следующим образом:

Интервал (t_i-t_{i+1}), мин.	0—5	5—10	10—15	15—20
Число пациентов m_i, чел.	64	21	8	7

Проверить гипотезу об экспоненциальном законе распределения времени обслуживания t при $\alpha=0,05$.

17. При уровне значимости **0,05** проверить гипотезу о нормальном распределении генеральной совокупности X по данным выборки объема $n = 100$:

Интервал (x_i-x_{i+1})	3–8	8–13	13–18	18–23	23–28	28–33	33–38
Частота m_i	6	8	15	40	16	8	7

Для решения задач 18—20 использовать критерий знаков.

18. У ***18*** больных одновременно измеряли температуру тела двумя термометрами — обычным ручным и электротермометром. Получены следующие данные:

Ртутный термометр	***38,2***	***37,4***	***39,1***	***38,3***	***36,7***	***37,6***	***36,9***	***36,2***	***39,3***	***38,8***	***38,4***	***38,2***	***37,3***	***37,4***	***36,2***	***36,9***	***36,2***	***38,5***
Электро-термометр	***38,1***	***37,2***	***39,0***	***38,4***	***36,8***	***37,5***	***37,1***	***36,0***	***39,4***	***38,6***	***38,3***	***38,3***	***37,2***	***37,3***	***36,1***	***37,1***	***36,0***	***38,5***

Можно ли считать, что различие в показаниях термометров существенно (принять $\alpha=0,01$, для него $m_{кр}=m_{0,01}(17)=4$)?

19. Предполагается, что один из pH-метром имеет систематическую ошибку. Для проверки этого предположения определили pH ***10*** растворов одновременно двумя приборами. В результате получены следующие данные:

I прибор	***7,00***	***8,52***	***6,32***	***5,43***	***6,57***	***7,48***	***7,26***	***6,83***	***5,87***	***7,04***
II прибор	***6,97***	***8,55***	***6,29***	***5,42***	***6,58***	***7,46***	***7,23***	***6,79***	***5,86***	***7,05***

Позволяют ли эти результаты утверждать, что один из приборов действительно имеет систематическую ошибку? Принять $\alpha=0,05$.

20. У ***90*** почти здоровых мужчин в возрасте ***30*** лет в ходе диспансеризации измерялось артериальное давление крови. Затем измерения повторили через ***10*** лет (у ***40***-летних) и через ***20*** лет (у **50**-летних). Оказалось, что через ***10*** лет повышение артериального давления было зафиксированно у ***50*** мужчин, а через ***20*** лет — у **60**. Можно ли с $\alpha=0,05$ утверждать, что давление крови существенно увеличилось: а) через *10* лет; б) через ***20*** лет?

Для решения задач 21—22 использовать критерий Вилкоксона.

21. Известны результаты измерения (в мм) длины изделий двух выборок, объемы которых соответственно равны $n_1=6$ и $n_2=6$:

x_i	***112***	***110***	***108***	***115***	***114***	***111***
y_i	***113***	***109***	***116***	***117***	***107***	***118***

При уровне значимости ***0,05*** проверить нулевую гипотезу $F_1(x)=F_2(x)$ об однородности выборок при конкурирующей гипотезе $H_1: F_1(x)\neq F_2(x)$.

22. При уровне значимости ***0,05*** проверить нулевую гипотезу об однородности двух выборок, объемы которых соответственно равны $n_1=30$ и $n_2=50$, при конкурирующей гипотезе $F_1(x)>F_2(x)$, если известно, что сумма порядковых номеров вариант первой выборки в общем вариационном ряду $W_{набл}=1150$.

8. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

8.1. Статистическая и корреляционная связь

Статистической называется такая связь, когда каждому значению одной случайной величины соответствует множество значений другой случайной величины, или, если одна случайная величина реагирует на изменение другой случайной величины изменением своей функции (или плотности) распределения вероятностей. Например, связь между производительностью труда и зарплатой, дозой лекарственного препарата и лечебным эффектом, между ростом и весом животного одного вида и т.д.

Корреляционной связью называется частный случай статистической связи, когда одна случайная величина реагирует на изменение другой случайной величины изменением своего математического ожидания то есть

$$M[Y/X=x]=f(x) \text{ или } M[X/Y=y]=f(y), \qquad (8.1)$$

где $M[Y/X=x]=\bar{Y}(x)=m_y(x)$ — математическое ожидание случайной величины Y при условии, что случайная величина X приняла конкретное значение x, то есть условное математическое ожидание величины Y, $m_x(y)=M[X/Y=y]$ — условное математическое ожидание величины X.

Функция $f(x)$ называется функцией регрессии (или просто регрессией) Y на X, а ее график — линией регрессии Y на X, соответственно $f(y)$ — регрессия X на Y.

Функция регрессии может быть линейной $M[Y/X=x]=\bar{Y}(x)=a+bx$ или нелинейной. Соответственно различают линейную и нелинейную корреляционную связь.

Коэффициент корреляции K_{xy}, равный

$$K_{xy}=M[(X-m_x)(Y-m_y)]=M[X\cdot Y]-m_x\cdot m_y, \qquad (8.2)$$

является числовой характеристикой линейной корреляционной связи двух случайных величин X и Y.

Если случайные величины X и Y независимы, то $K_{xy}=0$. Следовательно, если $K_{xy}\neq 0$, то X и Y — зависимые случайные величины.

Знак K_{xy} указывает на направление корреляционной связи. Если $K_{xy}>0$, то с возрастанием x возрастает в среднем и y — положительная корреляция между величинами X и Y; если $K_{xy}<0$, то при возрастании x величина y в среднем убывает — отрицательная корреляция.

Численное значение K_{xy} зависит от выбора единиц измерения случайных величин X и Y, что затрудняет сравнение коэффициентов корреляции различных пар случайных величин. Вследствие этого вводят другую числовую характеристику линейной связи между X и Y, лишенную этого недостатка, — нормированный коэффициент корреляции ρ_{xy}:

$$\rho_{xy}=\frac{K_{xy}}{\sigma_x\cdot\sigma_y} \qquad (8.3)$$

Свойства нормированного коэффициента корреляции:

1. Нормированный коэффициент корреляции принимает значения в интервале от -1 до 1: $-1 \le \rho_{xy} \le 1$
2. Если случайные величины X и Y независимы, то $\rho_{xy}=0$. Следовательно, если $\rho_{xy} \ne 0$, то X и Y — зависимые величины.
3. Если $\rho_{xy}=0$, то X и Y не связаны линейной корреляционной зависимостью, то есть некоррелированны, но могут быть зависимы, даже связаны функционально, но только нелинейной связью.
4. Если X и Y — нормально распределенные случайные величины, то из $\rho_{xy}=0$ следует, что X и Y — независимы.
5. Если $|\rho_{xy}| = 1$, то связь линейная функциональная.

Если $|\rho_{xy}| \ne 1$, но близок к 1, то корреляционная зависимость будет близка к линейной.

На основании свойств 3 и 5 абсолютное значение $|\rho_{xy}|$ используется как характеристика силы линейной связи между двумя случайными величинами X и Y, то есть степени близости корреляционной зависимости к линейной. Ориентировочно можно считать, что если $|\rho_{xy}| \ge 0{,}7$, то линейная связь сильная, если $0{,}4 \le \rho_{xy} < 0{,}7$ — средняя, если $0 < |\rho_{xy}| < 0{,}4$ — слабая.

Если линии регрессии не являются прямыми, то ρ_{xy} может лишь с некоторым приближением рассматриваться как показатель силы связи между X и Y.

Оценка нормированного коэффициента корреляции

По выборке из n пар экспериментальных значений $(x_i;y_i)$ в качестве оценки неизвестного ρ_{xy} берется выборочный нормированный коэффициент корреляции ρ_{xy}, вычисляемый по формуле:

$$\rho_{xy} = \frac{K_{xy}}{\delta_x \cdot \delta_y} = \frac{K_{xy}}{S_x \cdot S_y}, \tag{8.4}$$

где

$$K_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i \cdot Y_i - n\overline{X}\cdot\overline{Y}\right) \tag{8.5}$$

Проверка значимости ρ_{xy}

Если выборочный ρ_{xy} отличен от нуля, то, учитывая случайность выборки, еще нельзя сделать вывод о том, что и ρ_{xy} генеральной совокупности тоже отличен от нуля. Для проверки существенности (значимости) корреляционной связи при заданном уровне значимости α, то

есть для проверки нулевой гипотезы $H_0: \rho_{xy}=0$ при альтернативной гипотезе $H_1: \rho_{xy} \neq 0$ используют критерий Стьюдента:

$$t_{пр} = \frac{\rho_{xy} - 0}{S\rho_{xy}} = \rho_{xy} \cdot \sqrt{\frac{n-2}{1-\rho^2_{xy}}}. \qquad (8.6)$$

По таблице критических точек Стьюдента по заданному уровню значимости α и числу степеней свободы $K=n-2$ находят $t_{крит} = t_{\alpha/2}(k)$. Если $|t_{пр}| \leq t_{кр}$, то $H_0: \rho_{xy}=0$ принимается, то есть X и Y некоррелированны (а в случае нормального распределения X и Y — независимы), если же $|t_{пр}| > t_{крит}$, то H_0 отвергается, то есть ρ_{xy} значимо отличается от нуля и X и Y коррелированны, а значит и зависимы.

8.2. Метод наименьших квадратов

Метод наименьших квадратов служит для выявления связей между величинами, характеризующими изучаемый объект или процесс, например, между видом образования и заработком, величиной пульса и артериального давления у больных, между температурой и выходом конечного продукта при химической реакции и т.д.

В общем случае имеется n результатов совместных измерений величин X и Y: $(x_1, y_1); (x_2, y_2)\ldots(x_n, y_n)$. Необходимо по этим данным наилучшим (оптимальным) образом составить уравнение связи $y=f(x)$ и вычислить коэффициенты этого уравнения. В случае нормального распределения величин X и Y оптимальное решение получают методом наименьших квадратов (МНК).

Его сущность состоит в определении коэффициентов предполагаемого уравнения связи таким образом, чтобы сумма квадратов отклонений экспериментальных значений y_i от вычисленных по уравнению связи $y_{i\,теор.}=f(x_i)$ являлась бы минимальной, то есть

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 = min. \qquad (8.7)$$

Если зависимость между Y и X близка к линейной, то уравнение связи имеет вид:

$$Y=a+bx. \qquad (8.8)$$

В этом случае необходимо найти такие значения коэффициентов a и b, при которых сумма

$$Q = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = Q(a,b) \qquad (8.9)$$

принимала бы минимальное значение.

В высшей математике доказывается, что сумма квадратов отклонений $Q(a,b)$, как функция двух аргументов, будет иметь минимум, если частные производные от нее по a и по b будут равны нулю. Для линейного уравнения (8.8) и соответствующего ему $Q(a,b)$ эти условия имеют вид:

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}\left[y_i - (a + bx_i)\right] = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}\left[y_i - (a + bx_i)\right]x_i = 0. \end{cases} \tag{8.10}$$

Решая совместно эти уравнения, получаем следующие выражения для коэффициентов:

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i) - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}, \tag{8.11}$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad и \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Уравнение связи при этом удобно записывать в виде:

$$y = \bar{y} + b(x - \bar{x}), \tag{8.12}$$

или

$$y = \bar{y} - b\bar{x} + bx = (\bar{y} - b\bar{x}) + bx = a + bx . \tag{8.13}$$

Для уравнений связи второго и высших порядков составляется система уравнений типа (8.10). Для уравнений связи второго порядка

$$y=a+bx+cx^2 \tag{8.14}$$

будем иметь:

$$Q = Q(a,b,c) = \sum_{i=1}^{n}\left[y_i - (a + bx_i + cx_i^2)\right]^2 = min . \tag{8.15}$$

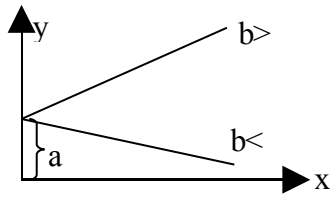
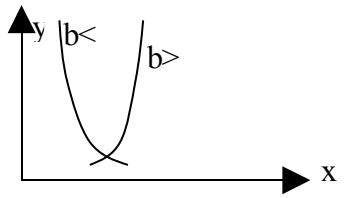
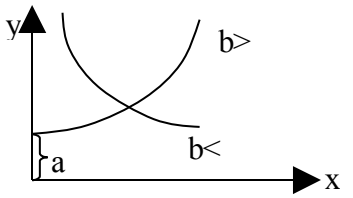
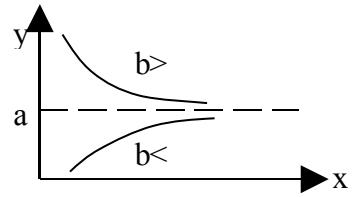
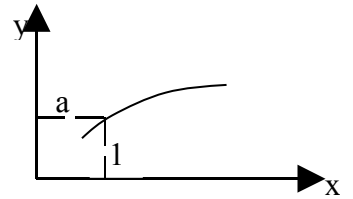
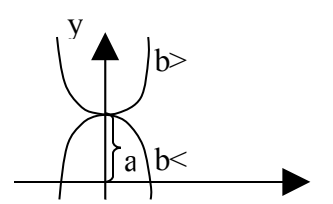
Берем частные производные по a, b, c, получим:

$$\left.\begin{aligned} \frac{\partial Q}{\partial a} &= \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} x_i^2 = 0 \\ \frac{\partial Q}{\partial b} &= \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 - c\sum_{i=1}^{n} x_i^3 = 0 \\ \frac{\partial Q}{\partial c} &= \sum_{i=1}^{n} y_i x_i^2 - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i^3 - c\sum_{i=1}^{n} x_i^4 = 0 \end{aligned}\right\} . \tag{8.16}$$

Эта система уравнений решается обычными алгебраическими методами, так как суммы при коэффициентах a, b и c имеют конкретные числовые значения.

Вычисление коэффициентов уравнений более чем третьего порядка довольно сложно и приводит к большим погрешностям в значениях коэффициентов, поэтому уравнения четвертого и высших порядков обычно не используют для расчетов. В ряде случаев нелинейные уравнения можно путем преобразований свести к линейным, коэффициенты ***a*** и ***b*** которых находят по МНК из уравнений (8.11). Примеры таких преобразований приведены в таблице 8.1.

Таблица 8.1

Название связи	*Уравнение связи*	*График уравнения*	*Преобразование*	*Преобразованное уравнение связи*
Линейная	$y=a+bx$		—	$y=a+bx$
Степенная	$y=ax^b$		$y'=\ln y;$ $x'=\ln x;$ $a'=\ln a$	$y'=a'+bx'$
Экспоненциальная	$y=ae^{bx}$		$y'=\ln y;$ $a'=\ln a$	$y'=a'+bx$
Гиперболическая	$y=a+\frac{b}{x}$		$x'=\frac{1}{x}$	$y=a+bx'$
Логарифмическая	$y=a+b\ln x$		$x'=\ln x$	$y=a+bx'$
Параболическая	$y=a+bx^2$		$x'=x^2$	$y=a+bx'$

8.3. Оценка параметров уравнения регрессии

Если корреляционная зависимость между X и Y близка к линейной (ρ_{xy} близок к 1), то естественно составить вопрос о нахождении такого уравнения регрессии $\bar{Y}(x) = a + bx$ или $\bar{X}(y) = a + by$, которое наилучшим образом выражало бы эту зависимость. Оценки параметров a и b линейной регрессии Y на X по выборке из n пар значений $(x_i;y_i)$, где $i=1, 2, 3, \ldots, n$ получают, используя рассмотренный в предыдущем разделе метод наименьших квадратов.

Коэффициенты линейного уравнения регрессии равны:

$$b = \frac{K_{xy}}{S_x^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}. \quad (8.17)$$

$$a = \bar{y} - b\bar{x}. \quad (8.18)$$

Оценку b можно также найти по следующей формуле:

$$b = \rho_{xy}\cdot\frac{S_y}{S_x}. \quad (8.19)$$

Так как прямая регрессии $\bar{y}(x) = a + bx$ проходит через точку ($\bar{x};\bar{y}$), то уравнение линейной регрессии удобно записывать в виде:

$$\bar{y}(x) - \bar{y} = b(x - \bar{x})$$

или

$$\bar{y}(x) = \bar{y} + b(x - \bar{x}). \quad (8.20)$$

Оценки параметров a и b линейной регрессии X на Y вычисляют по аналогичной формуле:

$$b = \frac{K_{xy}}{S_y^2}, \quad (8.21)$$

$$a = \bar{x} - b\bar{y}, \quad (8.22)$$

$$b = \rho_{xy}\cdot\frac{S_x}{S_y}, \quad (8.23)$$

$$\bar{x}(y) = \bar{x} + b(y - \bar{y}). \quad (8.24)$$

Полученные выборочные уравнения линейной регрессии можно использовать для предсказания среднего значения одной случайной величины по заданному значению другой случайной величины.

Пример 1. В результате выборочного статистического исследования ***5*** предприятий отрасли получены следующие данные о стоимости основных производственных фондов ***X*** (млн руб.) и объеме товарной продукции ***Y*** (млн руб.):

x_i, млн руб.	3	5	5	7	10
y_i, млн руб.	20	31	37	52	70

1. Для каждой выборки найти точечные оценки среднего значения, дисперсии, среднего квадратичного отклонения и доверительный интервал для среднего значения при уровне значимости $\alpha=0,05$, считая, что выборки взяты из нормально распределенных генеральных совокупностей.
2. Вычислить оценку нормированного коэффициента линейной корреляции ρ_{xy} и оценить тесноту линейной связи между ***X*** и ***Y***.
3. Найти выборочные сравнения линейной регрессии ***Y*** на ***X*** и ***X*** на ***Y***. Построить их графики совместно с корреляционным полем.

Решение

1. Для удобства вычислений составим расчетную таблицу (вместо $\sum_{i=1}^{n}$ условимся писать просто $\sum$):

x_i	3	5	5	7	10	$\sum x_i = 30$
y_i	20	31	37	52	70	$\sum y_i = 210$
x_i^2	9	25	25	49	100	$\sum x_i^2 = 208$
y_i^2	400	961	1369	2704	4900	$\sum y_i^2 = 10334$
$x_i y_i$	60	155	185	364	700	$\sum x_i y_i = 1464$

Находим выборочные средние:

$$\bar{x}_B = \frac{1}{n}\sum x_i = \frac{1}{5} * 30 = 6,$$

$$\bar{y}_B = \frac{1}{n}\sum y_i = \frac{1}{5} * 210 = 42.$$

Вычисляем исправленные выборочные дисперсии S_x^2 и S_y^2:

$$S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x}_B)^2 = \frac{1}{5-1}\left[(3-6)^2 + (5-6)^2 + (5-6)^2 + (7-6)^2 + (10-6)^2\right] = \frac{28}{4} = 7;$$

$$S_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y}_B)^2 = \frac{1}{4}\left[(20-42)^2 + (31-42)^2 + (37-42)^2 + (52-42)^2 + (70-42)^2\right] = 378.5.$$

Удобно вычислять $S_x^2 (S_y^2)$ по другой формуле, например, для S_y^2 имеем:

$$S_y^2 = \frac{1}{n-1}\left(\sum y_i^2 - n \cdot \bar{y}_B^2\right) = \frac{1}{4}(10334 - 5 \cdot 42^2) = \frac{1}{4}(10334 - 8820) = \frac{1514}{4} = 378{,}5.$$

Находим оценки средних квадратичных отклонений σ_x и σ_y:

$$S_x = \sqrt{S_x^2} = \sqrt{7} \approx 2{,}65\,; \quad S_y = \sqrt{S_y^2} = \sqrt{378{,}5} \approx 19{,}46\,.$$

По таблице критических точек t-распределения Стьюдента при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = n\text{-}1 = 5\text{-}1 = 4$ находим $t\,(\alpha, k)$:

$$t\,(0{,}95;4) = 2{,}78.$$

Вычисляем полуширину доверительного интервала:

$$\delta_x = t(\alpha, k) \cdot \frac{S_x}{\sqrt{n}} = 2{,}78\,\frac{2{,}65}{\sqrt{5}} \approx 3{,}3\,; \quad \delta_y = t(\alpha, k) \cdot \frac{S_y}{\sqrt{n}} = 2{,}78\,\frac{19{,}46}{\sqrt{5}} \approx 24{,}19 \approx 24\,.$$

Запишем доверительные интервалы для m_x и m_y при уровне значимости $\alpha = 0{,}05$:

$$7 - 3{,}3 < m_x < 7 + 3{,}3 \text{ и } \quad 42 - 24 < m_y < 42 + 24$$

или

$$3{,}7 < m_x < 10{,}3, \text{ и } \quad 18 < m_y < 66.$$

Итак, при уровне значимости $\alpha = 0{,}05$ среднее значение X заключено в доверительном интервале $(3{,}7;\ 10{,}3)$, а среднее значение Y — в доверительном интервале $(18;66)$.

2. Вычисляем оценку нормированного коэффициента ρ_{xy} линейной корреляции по формуле:

$$\rho_{xy} = \frac{\frac{1}{n-1}\left(\sum x_i y_i - n \cdot \bar{x}_B \cdot \bar{y}_B\right)}{S_x \cdot S_y} \approx \frac{\frac{1}{4}\left(1464 - 5 \cdot 6 \cdot 42\right)}{2{,}65 \cdot 19{,}46} \approx \frac{51}{51{,}57} \approx 0{,}982.$$

Так как ρ_{xy} близок к единице, то можно считать, что признаки X (основные фонды) и Y (объем продукции) связаны тесной (близкой к функциональной) линейной корреляционной связью.

3. Находим сначала выборочное уравнение линейной регрессии Y на X: $\bar{y}_x = a + bx$. Оценки параметров a и b получим методом наименьших квадратов по формулам (8.17 — 8.18):

$$b = \frac{1464 - 5 \cdot 6 \cdot 42}{208 - 5 \cdot 6^2} = \frac{204}{28} \approx 7{,}286.$$

Тогда $a = 42 - 7{,}286 \cdot 6 = -1{,}714.$

Записываем выборочное уравнение линейной регрессии Y на X:

$$\bar{y}_x = -1{,}714 + 7{,}286 \cdot x.$$

Аналогично получим выборочное уравнение регрессии X на Y: $\bar{x}_y = c + d \cdot y$.

$$d = \frac{\sum x_i \cdot y_i - n \cdot \bar{x}_B \cdot \bar{y}_B}{\sum y_i^2 - n \cdot \bar{y}_B^2} = \frac{1464 - 5 \cdot 6 \cdot 42}{10334 - 5 \cdot 42^2} = \frac{204}{1514} \approx 0{,}135,$$

$$\hat{c} = 6 - 0{,}135 \cdot 42 = 0{,}33.$$

Тогда $\bar{x}_y = 0{,}33 + 0{,}135 \cdot y$.

Строим графики (рис. 8.1) выборочных уравнений линейной регрессии совместно с корреляционным полем (совокупностью точек с координатами $(x_i; y_i)$).

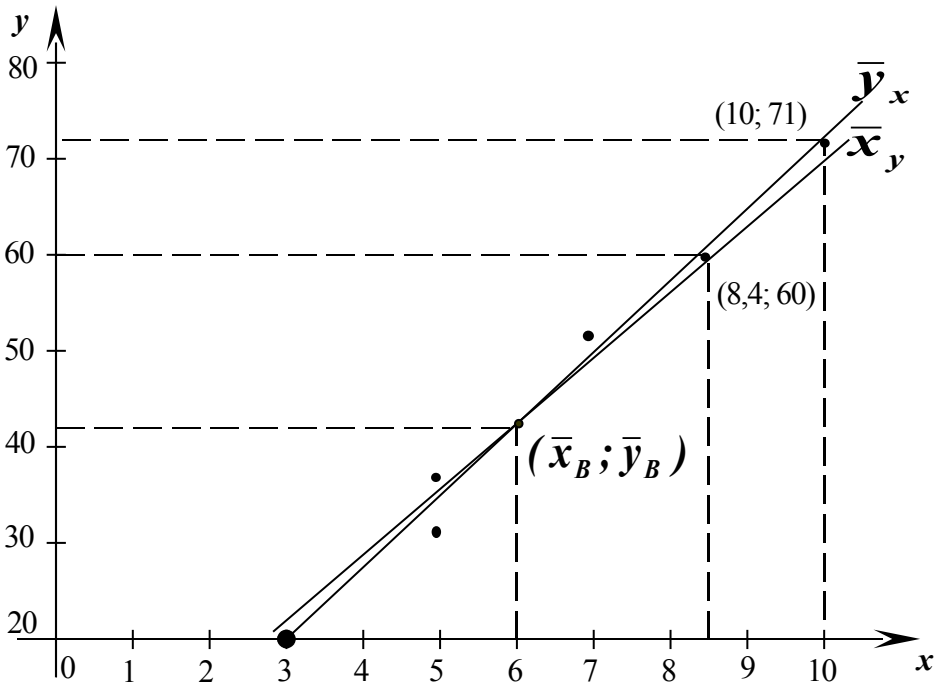


Рис. 8.1

Оба графика выборочных уравнений регрессии должны проходить через точку $(\bar{x}_B; \bar{y}_B) = (6;42)$. Найдем вторые точки для построения прямых, например:

$$\bar{y}_{10} = -1{,}714 + 7{,}2863 \cdot 10 = 71{,}146 \approx 71;$$
$$\bar{x}_{60} = 9{,}33 + 0{,}135 \cdot 60 = 8{,}43 \approx 8{,}4.$$

Проводим прямые через рассчитанные точки и точку $(\bar{x}_B ; \bar{y}_B)$.

8.4. Корреляционное отношение

Если регрессия является нелинейной, то удобным показателем силы корреляционной связи, характеризующим тесноту распределения около кривых $m_y(x)$ или $m_x(y)$ регрессии, служит корреляционное отношение $\eta_{y/x}$ или $\eta_{x/y}$:

$$\eta_{y/x} = \frac{\sigma_{m_y(x)}}{\sigma_y} = \sqrt{\frac{M[(m_y(x) - m_y)^2]}{\sigma_y^2}} \quad \text{или} \quad \eta_{x/y} = \frac{\sigma_{m_x(y)}}{\sigma_x} = \sqrt{\frac{M[(m_x(y) - m_x)^2]}{\sigma_x^2}}, \quad (8.25)$$

где $\sigma^2_{m_y}(x)$ и $\sigma^2_{m_x}(y)$ — условные дисперсии.

Свойства корреляционного отношения $\eta_{y/x}$ (аналогичные свойства и у $\eta_{x/y}$):

1. Корреляционное отношение принимает значение в интервале от ***0*** до ***1***:

$$0 \le \eta_{y/x} \le 1.$$

2. Если $m_y(x)=m_y=const$, то есть X и Y некоррелированны, то $\eta_{y/x}=0$. Верно и обратное: если $\eta_{y/x}=0$, то X и Y некоррелированны.

3. Если существует функциональная зависимость Y от X: $m_y(x)=y=f(x)$, то есть все значения Y лежат на линии регрессии, то $\eta_{y/x}=1$. Верно и обратное: если $\eta_{x/y}=1$, то Y является функцией X.

4. Между $\eta_{y/x}$ и $\eta_{x/y}$ нет простой связи — одно из них может равняться 1, а другое — нулю и наоборот. Если $\eta_{y/x}= \eta_{x/y}=1$, то между величинами X и Y существует обратимая функциональная связь. Верно и обратное предположение.

5. Корреляционное отношение не меньше абсолютного значения нормированного коэффициента корреляции ρ_{xy}:

$$\eta_{y/x} \ge |\rho_{xy}| \text{ и } \eta_{x/y} \ge |\rho_{xy}|.$$

Следовательно, если $\eta_{y/x}=0$, то и $\rho_{xy}=0$.

Корреляционная таблица. При большом числе пар наблюдений $(x_i; y_j)$ случайных величин X и Y одно и то же значение x_i может встретиться n_{Xi} раз, одно и то же значение y_j — n_{Yj} раз, одна и та же пара значений $(x_i; y_j)$ может наблюдаться n_{XiYj} раз ($i=1,2,\ldots, k$; $j=1,2,\ldots, l$, где k и l — число различных значений величин X и Y соответственно). Для удобства обработки результатов парных наблюдений их группируют, подсчитывают частоты n_{Xi}, n_{Yj} и n_{XiYj} и записывают в виде корреляционной таблицы.

Y \ X	x_1	x_2	…	X_K	n_{Yj}
y_1	$n_{X1\ Y1}$	$n_{X2\ Y1}$	…	$n_{Xk\ Y1}$	n_{Y1}
y_2	$n_{X1\ Y2}$	$n_{X2\ Y2}$	…	$n_{Xk\ Y2}$	n_{Y2}
…	…	…	…	…	…
Y_l	$n_{X1\ Yl}$	$n_{X2\ Yl}$	…	$n_{Xk\ Yl}$	n_{Yl}
n_{Xi}	n_{X1}	n_{X2}	…	n_{Xk}	n

В первой строке указаны все наблюдаемые значения случайной величины X, в первом столбце — случайной величины Y. На пересечении строк и столбцов находятся частоты $n_{Xi\ Yj}$ пар $(x_i; y_j)$ значений величин. В последнем столбце (последней строке) записаны суммы частот соответствующих строк (столбцов), причем n — объем парной выборки:

$$n = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{l} n_{y_j} .$$

Оценка корреляционного отношения $\eta_{y/x}$ по выборке общим объемом n пар значений величин X и Y, представленной в виде корреляционной таблицы, вычисляется по формуле:

$$\eta_{y/x} = S_{\bar{y}(x)} \Big/ S_y , \tag{8.26}$$

где

$$S_{\bar{y}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_{x_i}[\bar{y}(x_i) - \bar{y}]^2} ; \tag{8.27}$$

$$\bar{y}(x_i) = \frac{1}{n_{x_i}} \cdot \sum_{j=1}^{n_{x_i}} y_j(x_i),$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_{x_i} \cdot \bar{y}(x_i) = \frac{1}{n}\sum_{j=1}^{n} y_j \quad ; \tag{8.28}$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{l} n_{y_j}(y_j - \bar{y})^2} . \tag{8.29}$$

В этих формулах $\bar{y}(x_i)$ — условное среднее значение величины Y при $X=x_i$, n_{xi}– число значений x_i.

По аналогичным формулам находятся $\eta_{x/y}$:

$$\eta_{x/y} = \frac{S_{\bar{x}(y)}}{S_x} , \tag{8.30}$$

где

$$S_{\bar{x}(y)} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{e} n_{y_j}\left[\bar{x}(y_j) - \bar{x}\right]^2}\ ; \qquad (8.31)$$

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}\ ; \qquad (8.32)$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{l} n_{y_j}\bar{x}(y_j) = \frac{1}{n}\sum_{i=1}^{n} x_i\ . \qquad (8.33)$$

Пример 2. Для изучения корреляционной связи между температурой ***X*** (°С) и временем химической реакции ***Y***(с) были выполнены ***n=20*** измерений:

X	***40***	***50***	***40***	***40***	***60***	***60***	***50***	***50***	***60***	***40***	***50***	***50***	***50***	***60***	***60***	***50***	***60***	***60***	***40***	***50***
Y	***14***	***16***	***14***	***14***	***15***	***16***	***15***	***14***	***16***	***14***	***15***	***15***	***15***	***18***	***18***	***15***	***18***	***18***	***14***	***15***

• Построить корреляционное поле и корреляционную таблицу.

• Найти условные средние $\bar{y}(x_i)$ и построить по ним эмпирическую линию регрессии ***y*** на ***x***.

• Вычислить и сравнить между собой оценки корреляционных отношений $\eta_{y/x}$ и $\eta_{x/y}$ и коэффициента линейной корреляции ρ_{xy}.

Решение

1. Составить корреляционную таблицу и построить поле корреляции (рис. 8.2).

X / ***Y***	***40***	***50***	***60***	n_{Yj}
14	***5***	***1***	***-***	***6***
15	***-***	***6***	***1***	***7***
16	***-***	***1***	***2***	***3***
18	***-***	***-***	***4***	***4***
n_{Xi}	***5***	***8***	***7***	***20***

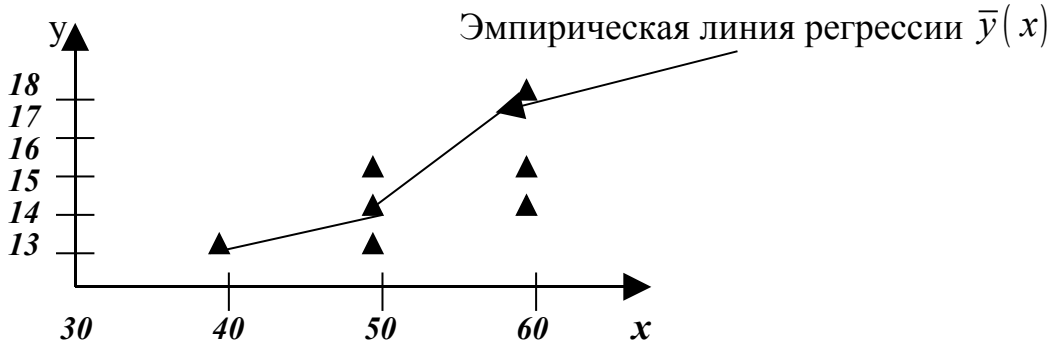


Рис. 8.2

2. Найдем условные средние $\overline{y}(x_i)$ как средние арифметические значения величины ***Y***, соответствующие x_i, по формуле (8.28а):

$$\overline{y}(x_1) = \overline{y}(40) = \frac{14 \cdot 5}{5} = 14;$$

$$\overline{y}(x_2) = \overline{y}(50) = \frac{14 \cdot 1 + 15 \cdot 6 + 16 \cdot 1}{8} = \frac{120}{8} = 15;$$

$$\overline{y}(x_3) = \overline{y}(60) = \frac{15 \cdot 1 + 16 \cdot 2 + 18 \cdot 4}{7} = \frac{119}{7} = 17.$$

По условным средним $\overline{y}(x_i)$ строим эмпирическую линию регрессии ***y*** на ***x*** (рис. 8.2).

3. Найдем общую среднюю $\overline{y}$:

$$\overline{y} = \frac{1}{20}(6 \cdot 14 + 7 \cdot 15 + 3 \cdot 16 + 4 \cdot 18) = \frac{309}{20} = 15{,}45 \approx 15{,}5.$$

Вычислим оценку межгруппового среднего квадратического отклонения ***Y*** по формуле (8.27):

$$S_{\overline{y}(x)} = \sqrt{\frac{1}{19}\left[5 \cdot (14 - 15{,}5)^2 + 8 \cdot (15 - 15{,}5)^2 + 7 \cdot (17 - 15{,}5)^2\right]} =$$

$$= \sqrt{\frac{1}{19}(11{,}25 + 2 + 15{,}75)} = \sqrt{\frac{29}{19}} \approx 1{,}235$$

и общего среднего квадратического отклонения y по формуле (8.29):

$$S_y = \sqrt{\frac{1}{19}\left[6 \cdot (14 - 15{,}5)^2 + 7 \cdot (15 - 15{,}5)^2 + 3 \cdot (16 - 15{,}5)^2 + 4 \cdot (18 - 15{,}5)^2\right]} =$$

$$= \sqrt{\frac{1}{19}(13{,}5 + 1{,}75 + 0{,}75 + 25)} = \sqrt{\frac{41}{9}} \approx 1{,}469.$$

Тогда выборочное корреляционное отношение $\eta_{x/y}$, согласно формуле (8.26), будет равно:

$$\eta_{x/y} = \frac{S_{\overline{y}(x)}}{S_y} = \frac{1{,}235}{1{,}469} \approx 0{,}841.$$

Аналогично по формулам находим $\eta_{y/x}$:

$$\overline{x}(y_1) = \overline{x}(14) = \frac{40 \cdot 5 + 50 \cdot 1}{6} = \frac{250}{6} \approx 41{,}7;$$

$$\overline{x}(y_2) = \overline{x}(15) = \frac{50 \cdot 6 + 60 \cdot 1}{7} = \frac{360}{7} \approx 51{,}4;$$

$$\overline{x}(y_3) = \overline{x}(16) = \frac{50 \cdot 1 + 60 \cdot 2}{3} = \frac{170}{3} \approx 56{,}7;$$

$$\bar{x}(y_4) = \bar{x}(18) = \frac{60 \cdot 4}{4} = 60;$$

$$\bar{x} = \frac{5 \cdot 40 + 8 \cdot 50 + 7 \cdot 60}{20} = \frac{1020}{20} = 51;$$

$$S_{\bar{x}(y)} = \sqrt{\frac{1}{19}\left[6(41,7-51)^2 + 7 \cdot (51,4-51)^2 + 3 \cdot (56,7-51)^2 + 4 \cdot (60-51)^2\right]} =$$

$$= \sqrt{\frac{1}{19}(518,94 + 1,12 + 97,47 + 324)} = \sqrt{\frac{941,53}{19}} \approx 7,04;$$

$$S_x = \sqrt{\frac{1}{19}\left[5 \cdot (40-51)^2 + 8 \cdot (50-51)^2 + 7 \cdot (60-51)^2\right]} =$$

$$= \sqrt{\frac{1}{19}(605 + 8 + 567)} = \sqrt{\frac{1180}{19}} \approx 7,88;$$

$$\eta_{x/y} = \frac{S_{\bar{x}(y)}}{S_x} = \frac{7,04}{7,88} \approx 0,893.$$

Вычислим теперь ρ_{xy} по формуле: $\rho_{xy} = \dfrac{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}\right)}{S_x \cdot S_y}$.

$$\rho_{xy} = \frac{\{[40 \cdot 14 \cdot 5 + 50 \cdot (14 + 15 \cdot 6 + 16) + 60 \cdot (15 + 16 \cdot 2 + 18 \cdot 4)] - 20 \cdot 15,45 \cdot 51\} / 19}{7,88 \cdot 1,469} =$$

$$= \frac{[(2800 + 6000 + 7140) - 15759] / 19}{11,576} \approx 0,823.$$

Видно, что

$$\eta_{y/x} > |\rho_{xy}| \text{ и } \eta_{x/y} > |\rho_{xy}|,$$

причем разница между выборочными корреляционными отношениями и ρ_{xy} не превышает 0,1; следовательно, корреляционную зависимость между X и Y можно считать практически линейной.

Задачи для самостоятельного решения

В задачах 1, 2, 3 вычислить ρ_{xy}, проверить его значимость при $\alpha = 0,05$, найти выборочные уравнения линейной регрессии ***Y*** на ***X*** и ***X*** на ***Y.***

1.

X	***8***	***5***	***10***	***8***	***9***	***14***
Y	***1***	***1***	***3***	***2***	***3***	***5***

2.

X	***5***	***9***	***10***	***12***	***14***	***16***

Y	***3***	***6***	***4***	***7***	***10***	***6***

3.

X	***10***	***2***	***7***	***5***	***4***	***1***
Y	***8***	***2***	***6***	***4***	***4***	***3***

4. Результаты анализа ***n=10*** образцов сырья, содержащего два полезных вещества ***A*** и ***B***, приведены в таблице, где ***X*** и ***Y*** — соответственно содержание веществ ***A*** и ***B*** в мг на кг сырья.

X	***66***	***70***	***75***	***80***	***82***	***85***	***90***	***92***	***95***	***98***
Y	***60***	***78***	***65***	***87***	***74***	***70***	***78***	***95***	***88***	***90***

Построить корреляционное поле, вычислить ρ_{xy} и проверить его значимость при $\alpha = 0{,}01$, найти выборочные уравнения линейной регрессии ***Y*** на ***X*** и ***X*** на ***Y***.

5. Определить при уровне значимости $\alpha = 0{,}05$ существенность линейной корреляционной связи между частотой пульса ***X*** и температурой тела ***Y***°С у инфекционных больных. Найти выборочное уравнение линейной регрессии ***Y*** на ***X*** и по нему оцените температуру больного при частоте пульса ***75***.
Результаты измерения следующие:

X, уд/мин.	***60***	***60***	***65***	***70***	***80***	***90***	***100***
Y°C	***36***	***37***	***38***	***37***	***38***	***40***	***40***

6. В клинике исследовалось влияние дозы лекарственного препарата ***X*** (мкг/кг) на время выздоравливания ***Y*** (в сутках). Получены следующие результаты:

X, мкг/кг	***5***	***10***	***20***	***30***	***40***	***40***	***50***
Y, сут	***24***	***20***	***14***	***16***	***16***	***14***	***14***

Построить корреляционное поле, при уровне значимости $\alpha = 0{,}05$ оценить существенность линейной корреляционной связи между дозой лекарств и временем выздоравливания и найти уравнение линейной регрессии ***Y*** на ***X***.

7. Вычислить выборочный нормированный коэффициент корреляции ρ_{xy} между температурой ***X***°С и временем некоторой химической реакции ***Y***(с) по результатам $n = 10$ измерений.

X° C	***55***	***71***	***53***	***67***	***81***	***75***	***59***	***89***	***65***	***81***
Y, c	***206***	***116***	***221***	***113***	***32***	***128***	***248***	***113***	***284***	***215***

8. Некоторый химический технологический процесс характеризуется двумя параметрами: давлением смеси ***X*** (мм рт. ст.) и температурой протекаемой реакции ***Y***°С.

X	***51***	***67***	***84***	***81***	***101***	***109***	***71***	***97***	***109***	***51***	***105***	***89***
Y	***25***	***30***	***43***	***44***	***57***	***58***	***43***	***46***	***62***	***45***	***55***	***45***

Найти выборочное уравнение линейной регрессии ***Y*** на ***X*** и ***X*** на ***Y*** и вычислить ρ_{xy} между этими характеристиками. Существенна ли корреляционная связь между ними при $\alpha = 0{,}01$?

9. По выборке объемом $n = 11$ получен $\rho_{xy} = -0{,}7$. Можно ли считать, что случайные величины X и Y отрицательно коррелированны, если уровень значимости $\alpha = 0{,}05$? А если $\alpha = 0{,}01$? При каком объеме n выборки с уровнем значимости $\alpha = 0{,}01$ можно было бы считать, что ρ_{xy} существенно отличается от нуля?

10. Найти $\eta_{y/x}$ и ρ_{xy} по следующим данным:

а

X	***10***	***20***	***30***	***40***
Y	***4,8***	***4,4***	***3,3***	***4,1***
	4,7	***4,2***	***3,5***	***3,9***
	4,8	***4,0***		***3,8***
	4,9			***4,2***

б

X	***18***	***19***	***23***	***26***
Y	***25***	***27***	***27***	
	23	***25***	***28***	
	27	***28***	***29***	***22***
		24		

11. Вычислить и сравнить между собой оценки корреляционных отношений $\eta_{y/x}$ и $\eta_{x/y}$ и коэффициента линейной корреляции ρ_{xy} по данным, приведенным в следующих корреляционных таблицах:

а

Y	*X*					
	2	***5***	***6***	***8***	***10***	n_y
5	***19***	***1***	***1***			***21***
8	***2***	***14***				***16***
10		***3***	***22***	***2***		***27***
15				***15***		***15***
20					***21***	***21***
n_x	***21***	***18***	***23***	***17***	***21***	***n=100***

б

Y	*X*					
	1	***4***	***6***	***15***	***20***	n_y
2	***9***	***1***				***10***
4	***1***	***4***				***5***
5		***3***	***15***	***2***		***20***
8			***2***	***8***	***1***	***11***
10					***4***	***4***
n_x	***10***	***8***	***17***	***10***	***5***	***n=50***

9. ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

9.1. Задачи дисперсионного анализа

Дисперсионный анализ является одним из методов определения влияния основных факторов на результаты наблюдений (например, различных доз лекарственного вещества на состав крови). Основной фактор изменяется заданным или случайным образом и его изменение может
в различной степени влиять на результаты эксперимента. Основные факторы могут иметь количественный характер (например, влияние концентрации катализатора на скорость химической реакции) или качественный характер (например, влияние способа лечения на длительность болезни, вида удобрений на урожай). Существенность влияния основного фактора характеризуется его вкладом в общую дисперсию результатов. Р. Фишер, выдающийся математик-статистик, разработавший основы дисперсионного анализа, выразил его сущность так: «Отделение дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам».

Таким образом, основная идея дисперсионного анализа состоит в сравнении дисперсии, отражающей действие изменения основного фактора с дисперсией, характеризующей случайность в результатах наблюдений.

Дисперсионный анализ можно использовать при планировании эксперимента и для выявления наиболее важных действующих факторов при составлении регрессионных уравнений связи. Иногда дисперсионный анализ пригоден также для проверки однородности нескольких совокупностей с целью их последующего объединения, что позволит получить более точные статистические оценки параметров и видов распределения.

В зависимости от количества факторов различают однофакторный и многофакторный дисперсионный анализ.

9.2. Однофакторный анализ

Однофакторный дисперсионный анализ используется для выявления существенности влияния изменения одного фактора A на изучаемый показатель X. В общем случае у основного фактора A имеется n значений (уровней): A_1, A_2, ..., A_i, …, A_n. Изучаемый показатель X считается распределенным по нормальному закону, дисперсии генеральных совокупностей X_i, соответствующих каждому уровню фактора A_i, будем считать одинаковыми:

$$D[X_1]=D[X_2]=\ldots=D[X_i]=\ldots=D[X_n]. \quad (9.1)$$

Если это неизвестно, то предварительно проверяют гипотезы об их равенстве.

Для каждого уровня A_i фактора A образуют выборочную совокупность $X_{i1}, X_{i2},..., X_{ij},\ldots, X_{im}$ объемом «m» из генеральной совокупности X_i. Выборочные средние $\overline{X}_i$, соответствующие различным уровням фактора, практически всегда различаются между собой. Это различие может получиться или вследствие случайности выборки (тогда оно несущественно), или вследствие действия изменения фактора A (и тогда различие существенно).

Требуется при заданном уровне значимости по полученным данным проверить нулевую гипотезу H_o о равенстве математических ожиданий:

$$M[X_1]=M[X_2]=\ldots=M[X_i]=\ldots=M[X_n]. \quad (9.2)$$

Если эта гипотеза отвергается, то, следовательно, различие между математическими ожиданиями существенно — среди них имеются хотя бы два неодинаковых. Проведение такого сравнения для каждой пары уровней фактора нецелесообразно, так как требует большого объема вычислений и не всегда дает правильный результат. В самом деле, соседние уровни фактора могут вызывать эффекты, существенно не различающиеся. В то же время различие в воздействиях крайних уровней фактора на изучаемый показатель может быть существенным.

Сведем все данные о результатах наблюдений в таблицу 9.1, где каждый результат наблюдения X_{ij} имеет два индекса: i — номер уровня фактора и j — номер наблюдения для данного уровня фактора.

В таблице каждая строка представляет собой самостоятельную выборочную совокупность (группу), соответствующую определенному уровню фактора, для которой можно найти выборочное групповое среднее:

$$\overline{X}_i = \frac{1}{m}\sum_{j=1}^{m} X_{ij} . \quad (9.3)$$

Таблица 9.1

Изучаемый фактор Ai	*j — номер наблюдения*						$\sum_{j=1}^{m} X_{ij}$	$(\sum_{j=1}^{m} X_{ij})^2$	$\sum_{j=1}^{m} X_{ij}^2$
	1	*2*	*…*	*j*	*…*	*m*			
A_1	X_{11}	X_{12}	...	X_{1j}	...	X_{im}			
A_2	X_{12}	X_{22}	...	X_{2j}	...	X_{2m}			
...	...	...	...	...	...	...	...	...	...
A_i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{im}			
...	...	...	...	...	...	...	...	...	...
A_n	X_{n1}	X_{n2}	...	X_{n3}	...	X_{nm}			

$\sum_{i=1}^{n}$									

Исправленная дисперсия для каждой группы равна:

$$S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 . \qquad (9.4)$$

Среднее по всем группам наблюдений находится как среднее арифметическое групповых средних:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\bar{X}_i = \frac{1}{n \cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{ij} . \qquad (9.5)$$

Для всей совокупности наблюдений можно найти общую исправленную дисперсию:

$$S^2 = \frac{1}{(nm-1)} = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 , \qquad (9.6)$$

где $(nm-1)$ — число степеней свободы.

Результаты эксперимента X_{ij} можно представить в виде суммы составляющих:

$$X_{ij} = \bar{X} + \alpha_i + \varepsilon_{ij} , \qquad (9.7)$$

где $\bar{X}$ — среднее значение (постоянная составляющая);

α_i — эффект фактора на i-м уровне;

ε_{ij} — случайная составляющая.

В случае, когда влияние фактора A несущественно, вследствие действия случайностей ε_{ij} выборочные средние для различных уровней фактора будут отличаться. Это различие характеризуется дисперсией, называемой остаточной.

Если влияние уровней фактора A существенно, то выборочные средние будут различаться вследствие случайности выборки (ε_{ij}) и вследствие действия уровней фактора (α_i). Степень этого различия характеризуется дисперсией, которая называется факторной.

Итак, выборочные средние $\bar{X}_i$, соответствующие различным уровням A_i фактора A, всегда различаются между собой. Требуется выяснить, является это различие случайным или оно вызвано действием изучаемого фактора. Это можно определить, проверив нулевую гипотезу о равенстве между собой факторной и остаточной дисперсий по критерию Фишера.

Для проверки нулевой гипотезы H_o найдем общую сумму квадратов отклонений значений X_{ij} от общей средней $\bar{X}$:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 . \qquad (9.8)$$

Представим эту сумму квадратов отклонений в виде суммы двух слагаемых. Для этого прибавим и вычтем в скобках величину группового среднего $\overline{X}_i$:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \overline{X}_i + \overline{X}_i - \overline{X})^2 . \qquad (9.9)$$

Разложим выражение в скобках как квадрат суммы:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}[(X_{ij} - \overline{X}_i) + (\overline{X}_i - \overline{X})]^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \overline{X}_i)^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \overline{X}_i)(\overline{X}_i - \overline{X}) + \sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{X}_i - \overline{X})^2 . \qquad (9.10)$$

В этой формуле второе слагаемое можно представить в виде:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \overline{X}_i)(\overline{X}_i - \overline{X}) = \sum_{i=1}^{n}(\overline{X}_i - \overline{X})\sum_{j=1}^{m}(X_{ij} - \overline{X}_i) .$$

Сумма $\sum_{j=1}^{m}(X_{ij} - \overline{X}_i) = 0$, так как является суммой отклонений вариант от своего выборочного (группового) среднего.

Таким образом, выражение для общей суммы квадратов примет вид:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \overline{X}_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{X}_i - \overline{X})^2 = Q_0 + Q_\phi . \qquad (9.11)$$

Итак, общая сумма квадратов отклонений состоит из двух сумм. Одна из них — Q_ϕ, называемая факторной суммой квадратов отклонений, характеризует рассеивание групповых средних $\overline{X}_i$ относительно общей средней $\overline{X}$ вследствие случайности и действия изучаемого фактора. Факторную сумму удобно преобразовать.

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{X}_i - \overline{X})^2 = m\sum_{i=1}^{n}(\overline{X}_i - \overline{X})^2 . \qquad (9.12)$$

Другое слагаемое — Q_o, называемое остаточной суммой квадратов отклонений, характеризует рассеяние значений X_{ij} относительно групповых средних $\overline{X}_i$ вследствие только случайных причин. Q_o не зависит от уровня фактора, так как его действие на X_{ij} компенсируется вычитанием группового среднего $\overline{X}_i$.

Разделив суммы квадратов отклонений Q_ϕ и Q_o на соответствующее число степеней свободы, получим факторную S_ϕ^2 и остаточную S_0^2 дисперсии:

$$S_{\phi}^2 = \frac{Q_{\phi}}{n-1}; \qquad (9.13)$$

$$S_0^2 = \frac{Q_0}{n(m-1)}. \qquad (9.14)$$

Число степеней свободы для остаточной дисперсии определяется как разность между числом степеней свободы общей и факторной дисперсии:

$$(n\,m - 1) - (n - 1) = n\,m - n = n\,(m - 1).$$

Если нулевая гипотеза H_o о равенстве средних справедлива, то есть эффекты уровней факторов $\alpha_i=0$, то эти дисперсии различаются незначительно. Поэтому общая, факторная и остаточная дисперсии являются оценкой дисперсии генеральной совокупности при справедливости нулевой гипотезы.

Сравним по критерию Фишера две дисперсии — факторную S_{ϕ}^2, обусловленную действием фактора и случайностью, и остаточную S_0^2, вызванную только случайными причинами. Если различие между ними окажется несущественным, то нулевая гипотеза H_o не отвергается и, следовательно, различие между групповыми средними незначимо. Если нулевая гипотеза не верна, то есть различие между математическими ожиданиями генеральных совокупностей существенно, то групповые средние $\overline{X}_i$, как оценки существенно различных математических ожиданий, будут сильно отличаться между собой и от их общей средней $\overline{X}$. В результате значительно увеличивается факторная дисперсия, в то время как остаточная дисперсия практически не изменится. В итоге различие между S_{ϕ}^2 и S_0^2 окажется значимым. Обратно, если S_{ϕ}^2 окажется существенно больше S_0^2, при заданном уровне значимости, то и групповые средние $\overline{X}_i$ существенно различны, нулевая гипотеза отвергается и действие изменения фактора A на показатель X значимо. Поэтому проверку нулевой гипотезы можно проводить по критерию Фишера:

$$F = S_{\phi}^2 / S_0^2. \qquad (9.15)$$

Итак, чтобы проверить нулевую гипотезу H_o о равенстве математических ожиданий нормально распределенных генеральных совокупностей с одинаковыми дисперсиями, достаточно проверить по критерию Фишера нулевую гипотезу о равенстве факторной S_{ϕ}^2 и остаточной S_0^2 дисперсий, причем в числителе их отношения всегда стоит факторная дисперсия S_{ϕ}^2.

Правило проверки всегда выглядит так: нулевая гипотеза не отвергается, если практически вычисленное значение критерия $F_{пр}$ не больше его критического значения:

$$F_{пр} \le F_{кр}(\alpha, n-1, n(m-1)).$$

Если H_o принимается, это не означает, что математические ожидания равны. Единственное, что можно сказать, это то, что они существенно не различаются. Если $S_{ф}^2$ окажется меньше S_0^2, то отсюда сразу следует справедливость нулевой гипотезы о равенстве средних и нет надобности прибегать к F-критерию.

Для упрощения расчетов остаточную дисперсию S_0^2 можно находить как среднее арифметическое всех групповых исправленных дисперсий S_i^2, действительно:

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n} S_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2 / n(m-1). \qquad (9.16)$$

Изложенная выше последовательность расчетов в принципе проста, но требует значительного количество вычислений. При больших объемах выборок удобнее преобразовать указанные выше формулы, разложив суммы квадратов отклонений. Для общей суммы квадратов получается следующая формула:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2 - \frac{1}{n \cdot m}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}\right)^2. \qquad (9.17)$$

Остаточная сумма квадратов равна:

$$Q_0 = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2 - \frac{1}{m}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij}\right)^2; \qquad (9.18)$$

$$Q_{ф} = \frac{1}{m}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij}\right)^2 - \frac{1}{n \cdot m}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}\right)^2. \qquad (9.19)$$

Для проверки правильности расчетов можно использовать равенство (9.11). Далее производится расчет полученного практически значения критерия Фишера:

$$F_{пр} = \frac{Q_{ф}}{Q_0} \cdot \frac{n(m-1)}{(n-1)}. \qquad (9.20)$$

Из формул (9.17) — (9.20) видно, что процесс анализа производится с вычислением промежуточных величин:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}; \qquad \sum_{i=1}^{n}\sum_{j=1}^{m} X^2{}_{ij}; \qquad \sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij}\right)^2,$$

которые помещаются в таблицу 9.1 в три последних столбца и последнюю строку.

Пример 1. Данные о времени химической реакции X (сек.) при различном содержании катализатора (фактор A) представлены в таблице 9.2:

Таблица 9.2

Содержание катализатора A	*Номер эксперимента*			
	1	2	3	4
A_1=5 %	7	6	8	7
A_2=10 %	7	4	7	6
A_3=15 %	11	10	10	13

Методом дисперсионного анализа при уровне значимости $\alpha=0,05$ проверить, оказывает ли существенное влияние изменение содержания катализатора (в пределах ***5—15 %***) на время химической реакции. Предполагается, что выборки извлечены из нормальных генеральных совокупностей с одинаковыми дисперсиями.

Решение. С математической точки зрения требуется проверить нулевую гипотезу H_0: $M[X_1] = M[X_2] = M[X_3]$, где X_1, X_2, X_3 — генеральные совокупности, соответствующие ***5***, ***10*** и ***15 %*** содержанию катализатора.

1. Находим групповые средние $\overline{X}_i$ и общее среднее $\overline{X}$:

$$\overline{X}_1 = \frac{1}{4}\sum_{j=1}^{4} X_{ij} = \frac{1}{4}(7 + 6 + 8 + 7) = 7;$$

$$\overline{X}_2 = \frac{1}{4}(7 + 4 + 7 + 6) = 6;$$

$$\overline{X}_3 = \frac{1}{4}(11 + 10 + 10 + 13) = 11;$$

$$\overline{X} = \frac{7 + 6 + 11}{3} = 8.$$

2. Находим групповые исправленные дисперсии S_i^2 по формуле (9.7):

$$S_1^2 = \frac{1}{4 - 1}[(7 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (7 - 7)^2] = 2 / 3 \approx 0,67;$$

$$S_2^2 = \frac{1}{4-1}[(7-6)^2 + (4-6)^2 + (7-6)^2 + (6-6)^2] = 6/3 = 2;$$

$$S_3^2 = \frac{1}{4-1}[(11-11)^2 + (10-11)^2 + (10-11)^2 + (13-11)^2] = 6/3 = 2.$$

3. Вычисляем общую сумму квадратов отклонений Q:

$$Q = \sum_{i=1}^{3}\sum_{j=1}^{4}(X_{ij} - \overline{X})^2 = [(7-8)^2 + (6-8)^2 + (8-8)^2 + ... + (13-8)^2] = 70.$$

4. Находим факторную сумму квадратов отклонения Q:

$$Q_ф = \sum_{i=1}^{3}(\overline{X}_i - \overline{X})^2 = 4[(7-8)^2 + (6-8)^2 + (11-8)^2] = 4 \cdot 14 = 56.$$

5. Остаточную сумму квадратов отклонений можно найти двумя способами:

— непосредственно по формуле для Q_o:

$$Q_0 = \sum_{i=1}^{3}\sum_{j=1}^{4}(X_{ij} - \overline{X}_i)^2 = [(7-7)^2 + (6-7)^2 + (8-7)^2 + ... + (13-11)^2] = 14;$$

— как разность между Q и $Q_ф$: $Q_o = Q - Q_ф = 70 - 56 = 14$.

Лучше воспользоваться первым способом, так как по вычисленным значениям Q, $Q_ф$ и Q_o можно затем проверить правильность расчетов: $Q = Q_ф + Q_o$. В нашем случае $70 = 14 + 56$.

6. Вычисляем факторную $S_ф^2$ и остаточную S_o^2 дисперсии:

$$S_ф^2 = 56/(3-1) = 28; \qquad S_o^2 = 14/3(4-1) = 14/9 \approx 1{,}56.$$

Можно сильно сократить расчеты, не вычисляя Q и Q_o, а сразу найти S_o^2 как среднеарифметическое групповых исправленных дисперсий по формуле (9.14):

$$S_o^2 = 1/3(0{,}67 + 2 + 2) \approx 1{,}56.$$

7. Определяем $F_{пр}$:

$$F_{пр} = S_ф^2 / S_o^2 = 28/1{,}56 = 18.$$

8. По таблице распределения Фишера для заданного $\alpha = 0{,}05$ и числу степеней свободы числителя $k_1 = 2$ и знаменателя $k_2 = 9$ находим критическое значение $F_{кр}$:

$$F_{кр} = F_{0.05}(2; 9) = 8{,}02.$$

9. Так как $F_{пр} > F_{кр}$, то нулевая гипотеза отвергается, таким образом изменение содержания катализатора в пределах **5—15 %** значимо влияет на время химической реакции.

9.3. Схема однофакторного дисперсионного анализа при различных объемах выборки на разных уровнях

Если объемы выборок m_i $(i=1,2,...,n)$ на разных уровнях фактора A окажутся разными, то последовательность дисперсионного анализа останется прежней, лишь некоторые расчетные формулы немного усложнятся. Средние значения и исправленные дисперсии по каждому уровню определяются так:

$$\bar{X}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}; \qquad S_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_i)^2 . \tag{9.21}$$

Общее среднее по всей совокупности равно:

$$\bar{X} = \frac{1}{l}\sum_{i=1}^{n}\sum_{j=1}^{m_i} X_{ij} = \frac{1}{l}\sum_{i=1}^{n} m_i \bar{X}_i; \qquad l = \sum_{i=1}^{n} m_i ,$$

где l — общее число наблюдений. Общая сумма квадратов отклонений, факторная и остаточная суммы определяются соответственно по формулам:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m_i} (X_{ij} - \bar{X})^2 ; \tag{9.22}$$

$$Q_\phi = \sum_{i=1}^{n} m_i \cdot (\bar{X}_i - \bar{X})^2 ; \tag{9.23}$$

$$Q_o = \sum_{i=1}^{n}\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_2)^2 . \tag{9.24}$$

Факторная и остаточная дисперсия равны:

$$S_\phi^2 = \frac{Q_\phi}{n - 1}; \qquad S_o^2 = \frac{Q_o}{l - n}. \tag{9.25}$$

Для упрощенных расчетов формула (9.14) примет вид:

$$S_o^2 = \frac{1}{l - n}\sum_{i=1}^{n} (m_i - 1) S_i^2 . \tag{9.26}$$

Правильность расчетов Q, Q_ϕ и Q_o можно проверить по равенству (9.12).

Пример 2. Данные об урожайности пшеницы в ц/га по ***13*** участкам с различными комплексами удобрений (фактор A) представлены в таблице 9.3.

Таблица 9.3

Комплексы удобрений	*Номер наблюдения (участка)*					
	1	*2*	*3*	*4*	*5*	*6*
I комплекс	***32***	***36***	***34***	–	–	–
II комплекс	***37***	***38***	***40***	***33***	***35***	***33***
III комплекс	***34***	***41***	***52***	***49***	–	–

Методом дисперсионного анализа при $\alpha=0,01$ проверить, есть ли существенное различие в средней урожайности в зависимости от вида комплекса удобрений. Предположить, что выборки взяты из нормально распределенных генеральных совокупностей с одинаковыми дисперсиями.

Решение. По условию задачи требуется проверить нулевую гипотезу о равенстве групповых средних в случае, когда объемы групп (выборок) различны: $m_1=3,\ m_2=6,\ m_3=4.$ Всего наблюдений (участков) $l=m_1+m_2+m_3=13$.

1. Находим групповые средние $\overline{X}_i$ и общее среднее $\overline{X}$:

$$\overline{X}_1 = (32+36+34)/3 = 34;$$

$$\overline{X}_2 = 36;$$

$$\overline{X}_3 = 44;$$

$$\overline{X} = \frac{1}{13}(3\,34+6\,36+4\,44) = 494/13 = 38.$$

2. Находим групповые исправленные дисперсии S_i^2:

$$S_1^2 = \frac{1}{3-1}[(32-34)^2+(36-34)^2(34-34)^2] = 8/2 = 4;$$

$$S_2^2 = \frac{1}{6-1}[(37-36)^2+(38-36)^2+\ldots+(33-36)^2] = 40/5 = 8;$$

$$S_3^2 = \frac{1}{4-1}[(34-44)^2+(41-44)^2+(52-44)^2+(49-44)] = 198/3 = 66.$$

Так как различие между ними достаточно велико, например $S_3^2/S_1^2 = 66/4 = 16,5$, то возникает сомнение в справедливости предположения о равенстве дисперсий генеральных совокупностей, соответствующих каждому уровню фактора. Поэтому предварительно проверим гипотезу о равенстве дисперсий: $M[S_1^2]=M[S_2^2]=M[S_3^2]$, то есть о том, что существующее различие между ними: 4; 8 и 66 —

несущественно. Для сравнения дисперсии используем критерий Фишера. Для этого попарно найдем их F — отношения ($F_{пр}$= S^2большая/ S^2меньшая) и сравним с критическими значениями, взятыми при $\alpha=0,01$ и при соответствующих числах степеней свободы числителя k_1 и знаменателя k_2 из таблицы распределения Фишера:

$$Fnp_{31} = S_3^2/S_1^2 = 66/4 = 16,5 < F(0,01; 3; 2) = 99,25;$$

$$Fnp_{32} = S_3^2/S_2^2 = 66/8 = 8,75 < F(0,01; 3; 5) = 12,06;$$

$$Fnp_{33} = S_2^2/S_1^2 = 8/4 = 2 < F(0,01; 5; 2) = 99,30.$$

Так как все три $F_{пр}$ меньше критических значений, то, следовательно, нулевые гипотезы о попарном равенстве генеральных дисперсий не отвергаются, то есть можно действительно считать, что дисперсии трех генеральных совокупностей, соответствующих трем разным уровням фактора A, одинаковы. Продолжаем дальше дисперсионный анализ.

3. Находим общую, факторную и остаточную сумму квадратов отклонений по формулам (9.22), (9.23) и (9.24):

$$Q = [(32 — 38) + (36 — 38) + ... + (49 — 38)] = 462;$$

$$Q_ф = 3(34 — 38) + 6(36 — 38) + 4(44 — 38) = 216;$$

$$Q_o = [(32 — 34) + (36 — 34) + ... + (49 — 34] = 246.$$

Для контроля правильности расчетов воспользуемся тождеством:
$Q = Q_ф + Q_o = 462; 216 + 246 = 462$ — расчеты верны.

4. Находим факторную и остаточную дисперсии по формулам (9.25):

$$S_ф^2 = \frac{Q_ф}{n - 1} = 216 / 2 = 108; \quad S_o^2 = \frac{Q_o}{l - n} = 246 /(13 - 3) = 24,6.$$

Для быстроты расчетов S_0^2 можно найти сразу, не вычисляя Q и Q_o, по формуле (9.26) и исправленным дисперсиям S_i^2 :

$$S_o^2 = \frac{1}{13 - 3}[(3 - 1)4 + (6 - 1)8 + (4 - 1)66] = 246 / 10 = 24,6.$$

5. Определяем практическое значение критерия Фишера:

$$F_{пр} = S_ф^2/S_o^2 = 108/24,6 = 4,39.$$

6. По таблице распределения Фишера для $\alpha=0,01$, $k_1= 2$ и $k_2= 10$ находим критическое значение $F_{кр}$:

$$F_{кр} = F_{0,01}(2; 10) = 7,56.$$

Так как $F_{пр}<F_{кр}$, то нулевая гипотеза о равенстве средних не отвергается. Другими словами, зависимость среднего урожая от вида комплекса удобрений не установлена (несущественна).

9.4. Двухфакторный дисперсионный анализ

Общие понятия и определения

Дисперсионный анализ, при котором изучается действие изменения совокупности двух или более факторов на результат эксперимента, называется двух- или многофакторным.

При двухфакторном анализе на количественный признак X действуют два фактора; обозначим их A и B (например, влияние процентного содержания катализатора и температуры на скорость химической реакции). Будем считать, что фактор A имеет, как и ранее, n уровней, то есть по фактору A результаты измерения признака X делятся на n различных групп (выборок), а по фактору B — на m групп. Всего будет $n \cdot m$ групп. Предполагается нормальность распределения генеральных совокупностей, соответствующих каждому сочетанию уровней факторов и равенство их дисперсий. Для простоты ограничимся случаем, когда в каждой группе имеется лишь одно значение X, то есть всего значений будет ($n \cdot m$). Обозначим через X_{ij} значение показателя X, соответствующее сочетанию уровней фактора A_i и B_j.

Величину X_{ij} можно представить в виде суммы:

$$X_{ij} = \overline{X} + \alpha_i + \beta_i + \varepsilon_{ij},$$

где $\overline{X}$ — постоянная составляющая;

α_i — эффект от действия i-го уровня фактора A;

β_i — эффект от действия j-го уровня фактора B;

ε_{ij} — случайная составляющая.

При двухфакторном анализе удобно все результаты записать в виде таблицы 9.4, где в строках записаны группы по фактору A, а в столбцах — группы по фактору B. Для каждой из групп находят групповые средние $\overline{X}_{i\cdot}$ и $\overline{X}_{\cdot j}$:

$$\overline{X}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} X_{ij}; \qquad \overline{X}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \tag{9.27}$$

и общее среднее $\overline{X}$:

$$\overline{X} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}. \tag{9.28}$$

Таблица 9.4

A_i	B_j	J						X_i
		B_1	B_2	...	B_j	...	B_m	
i	A_1	X_{11}	X_{12}	...	X_{1j}	...	X_{1m}	$\overline{X}_{1.}$

	A_2	X_{21}	X_{22}	...	X_{2j}	...	X_{2m}	$\overline{X}_{2.}$
	...	...	...	...	...	...	...	...
	A_n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{nm}	$\overline{X}_{n.}$
		$\overline{X}_{.1}$	$\overline{X}_{.2}$	...	$\overline{X}_{.j}$	...	$\overline{X}_{.m}$	$\overline{X}$

Полная сумма квадратов отклонений Q от общего среднего может быть представлена в виде сумм квадратов Q_A, Q_B и Q_O:

$$Q = Q_A + Q_B + Q_O, \tag{9.29}$$

где

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - \overline{X})^2 . \tag{9.30}$$

— общая сумма квадратов отклонений;

$$Q_A = m\sum_{i=1}^{n} (\overline{X}_{i\bullet} - \overline{X})^2 \tag{9.31}$$

— сумма квадратов отклонений по фактору A;

$$Q_B = n\sum_{j=1}^{m} (\overline{X}_{\bullet j} - \overline{X})^2 \tag{9.32}$$

— сумма квадратов по фактору B;

$$Q_o = \sum_{j=1}^{m} (X_{ij} - \overline{X}_{i\bullet} - X_{\bullet j} + \overline{X})^2 \tag{9.33}$$

— остаточная сумма квадратов.

Разделив суммы квадратов отклонений Q, Q_A, Q_B и Q_O на соответствующее число степеней свободы, получим общую S^2, факторные S_A^2 и S_B^2 и остаточную S_o^2 дисперсии:

$$S^2 = \frac{Q}{nm-1}; \qquad S_A^2 = \frac{Q_A}{n-1}; \qquad S_B^2 = \frac{Q_B}{m-1}; \qquad S_0^2 = \frac{Q_0}{(n-1)(m-1)}. \tag{9.34}$$

Нулевая гипотеза об отсутствии существенного различия между групповыми средними здесь проверяется по критерию Фишера отдельно для фактора A и отдельно для фактора B. Для этого находят отношения факторных дисперсий к остаточной:

$$F_{Anp} = S_A^2 / S_o^2; \qquad F_{Bnp} = S_B^2 / S_o^2 \tag{9.35}$$

и сравнивают их с критическими значениями $F_{Akp}=F_\alpha(n-1, (n-1)(m-1)$ и $F_{Bkp}=F_\alpha(m-1, (n-1)(m-1)$. По результатам сравнения делают вывод о том, принимается или отвергается нулевая гипотеза H_o (отдельно по фактору A и по фактору B).

Пример 3. В ***n=3*** аптеках данные товарооборота за ***m=4*** месяца работы (в млн руб.) составили таблицу 9.5.

При уровне значимости ***α=0,05*** оценить существенность влияния на товарооборот фактора ***A*** — разные аптеки и фактора ***B*** — различные месяцы года.

Таблица 9.5

A_i \ B_j		месяц 1	2	3	4	$\overline{X}_{i\bullet}$
Аптеки	1	18	15	10	9	13
	2	11	11	8	6	9
	3	10	7	6	9	8
$\overline{X}_{\bullet j}$		13	11	8	8	10

Решение

1. Находим среднее по строкам $\overline{X}_{i\bullet} = 1/4\sum_{j=1}^{4} X_{ij}$ и среднее по столбцам $\overline{X}_{\bullet j} = 1/3\sum_{i=1}^{3} X_{ij}$ и значения их записываем в крайний столбец и нижнюю строку соответственно.

2. На пересечении крайнего столбца и нижней строки записываем вычисленное общее среднее $\overline{X}$:

$$\overline{X} = \frac{1}{3\,4}\sum_{i=1}^{3}\sum_{j=1}^{4} X_{ij} = 10.$$

3. Находим факторные суммы квадратов:

$$Q_A = 4\sum_{i=1}^{3} (\overline{X}_{i\bullet} - \overline{X})^2 = 4[(13-10)^2 + (9-10)^2 + (8-10)^2] = 56;$$

$$Q_B = 3\sum_{j=1}^{4} (X_{\bullet j} - X)^2 = 3[(13-10)^2 + (11-10)^2 + (8-10)^2] = 54.$$

4. Находим остаточную сумму квадратов Q_o. Ее можно получить непосредственно по формуле (9.33):

$$Q_o=(18-13-13+10)^2+(15-13-11+10)^2+(10-13-8+10)^2+(9-13-8+10)^2+$$
$$+(11-9-13+10)^2 + \ldots + (9-8-8+10)^2=28.$$

Можно сначала найти полную сумму квадратов:

$$Q=(18-10)^2+(15-10)^2+(10-10)^2+\ldots+(9-10)^2=138.$$

Тогда Q_o вычисляется как разность:

$$Q_o = Q - Q_a - Q_b = 138-56-54 = 28.$$

5. Найдем факторные дисперсии:

$$S_A^2 = \frac{Q_A}{n-1} = 56/2 = 28;$$

$$S_B^2 = \frac{Q_B}{m-1} = 54/3 = 18;$$

$$S_o^2 = \frac{Q_o}{(n-1)(m-1)} = 28/6 \approx 4{,}67.$$

6. Проверяем нулевую гипотезу по фактору A:

$$F_{Anp} = S_A^2 / S_o^2 = 28/4{,}67 = 6.$$

Из таблицы распределения Фишера находим соответствующее критическое значение $F_{Akp}=F_{0,05}(2; 6)=5{,}14$.

Так как $F_{Anp}>F_{Akp}$, то влияние фактора A на средний товарооборот существенно, или, другими словами, средние товарообороты (за 4 месяца) у разных аптек различаются существенно.

7. Проверяем нулевую гипотезу по фактору B:

$$F_{Bnp} = S_B^2 / S_o^2 = 18/4{,}67 \approx 3{,}9;$$

из таблицы находим $F_{Bkp} = F_{0,05}(3; 6) = 4{,}76$.

Так как $F_{Bnp}<F_{Bkp}$, то влияние фактора B несущественно, то есть средние товарообороты аптек в разных месяцах различаются несущественно.

9.5. Понятие о многофакторном дисперсионном анализе и планировании эксперимента

При обработке результатов эксперимента могут встретиться случаи, когда надо провести трех-, четырех- и т.д. факторный анализ. Во всех случаях многофакторный анализ сводится к последовательному применению двухфакторного анализа. Например, при трехфакторном анализе выборки по факторам A, B, C расположатся в виде трехмерной таблицы в кубических ячейках параллелепипеда со сторонами n_A, n_B, n_C. Сначала это трехмерное распределение проектируется на одну из граней, например AB, затем проводится двухфакторный анализ. После этого распределение проектируется, например, на грань BC ит.д.

Многофакторный анализ значительно упрощается, если каждый фактор имеет только два уровня. Причина в том, что для этого специаль-

ного случая используется отличная от обычной, гораздо более простая схема расчетов. Более подробно с многофакторным анализом можно ознакомиться в специальной литературе.

На основе методов двухфакторного и многофакторного дисперсионного анализа разработаны методы научного планирования статистических экспериментов. Основная идея этих методов состоит в том, что эксперименты могут быть оптимизированы без привлечения дополнительных затрат. Это можно сделать путем сокращения числа опытов без потери точности и достоверности результатов за счет сокращения избыточной информации, что позволяет получить экономию материальных ресурсов и высвободить время для дополнительных исследований. Можно также при сохранении затрат на исследования намного увеличить объем новой получаемой информации. При этом сокращается вероятность ошибочных выводов по экспериментальным данным. Наиболее известны среди них метод случайных блоков и метод латинских квадратов. Метод случайных блоков позволяет при планировании эксперимента исключить влияние разнородности блоков, вывести это влияние из оценки дисперсии случайности, что повышает точность анализа.

Метод латинских квадратов может быть рекомендован для организации эксперимента при различных значениях исходных параметров при воздействии различными уровнями факторов.

9.6. Дисперсионный анализ в Microsoft Excel

В состав электронной таблицы Microsoft Excel входит набор средств анализа данных (называемый пакет анализа), предназначенный для решения сложных статистических и инженерных задач. Для проведения анализа данных с помощью этих инструментов следует указать входные данные и выбрать параметры вывода; анализ будет проведен с помощью подходящей статистической или инженерной макрофункции, а результат будет помещен в выходной диапазон. Другие инструменты позволяют представить результаты анализа в графическом виде.

Установка и применение статистического пакета анализа данных. Для работы с инструментами анализа данные следует представить в виде строк или столбцов листа Excel. Совокупность ячеек, содержащих анализируемые данные, называется входным диапазоном.

Чтобы запустить пакет анализа:

- В меню *Сервис* выберите команду *Анализ данных*.
- В списке *Инструменты анализа* выберите нужную строку.
- Введите *входной* и *выходной диапазоны*, затем выберите необходимые параметры.

Примечание. Если команда *Анализ данных* отсутствует в меню *Сервис*, то необходимо запустить программу установки пакета анализа:

• В меню *Сервис* выберите команду *Надстройки*.

• Установите флажок *Пакет анализа*.

Примечание. Дополнительно в Microsoft Excel содержится большое число статистических функций. Для того чтобы вывести список доступных функций листа, выберите команду *Функция* в меню *Вставка*.

Для демонстрации возможностей статистического пакета решим, например, с его помощью задачу на однофакторный дисперсионный анализ, условие и решение которой приведены на с. 131—132:

«***Пример 2.*** Данные об урожайности пшеницы в ц/га по ***13*** участкам с различными комплексами удобрений (фактор ***A***) представлены в таблице 9.6.

Таблица 9.6

Комплексы удобрений	*Номер наблюдения (участка)*					
	1	***2***	***3***	***4***	***5***	***6***
I комплекс	***32***	***36***	***34***	—	—	—
II комплекс	***37***	***38***	***40***	***33***	***35***	***33***
III комплекс	***34***	***41***	***52***	***49***	—	—

Методом дисперсионного анализа при ***α=0,01*** проверить, есть ли существенное различие в средней урожайности в зависимости от вида комплекса удобрений.».

Сначала введем в ячейки A1–C1, A2–F2, A3–D3 исходные данные (см. рис. 9.1). Далее выбираем в меню *Сервис* команду *Анализ данных* и в списке *Инструменты анализа* — строку «*Однофакторный дисперсионный анализ*».

В открывшемся диалогового окне заполняем *Входные данные* и *Параметры вывода*, как показано на рисунке 9.1.

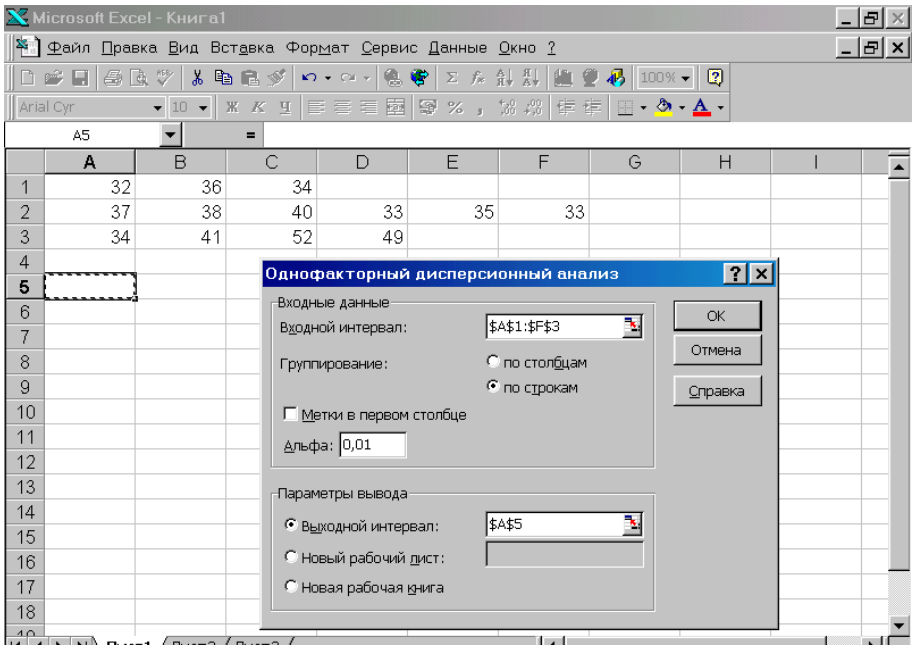
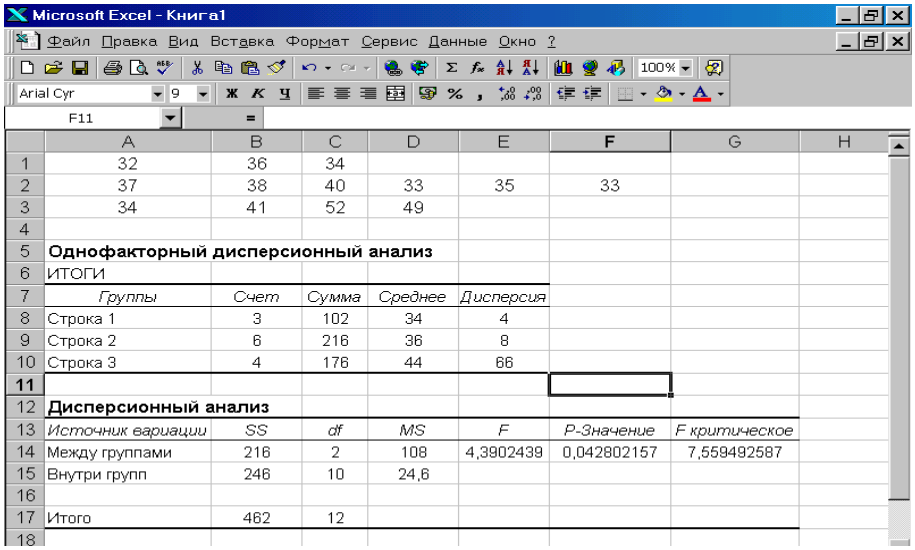


Рис. 9.1

Результаты решения представлены на рисунке 9.2.



Microsoft Excel - Книга1

Файл Правка Вид Вставка Формат Сервис Данные Окно ?

Arial Cyr 9 100%

F11 =

	A	B	C	D	E	F	G	H
1	32	36	34					
2	37	38	40	33	35	33		
3	34	41	52	49				
4								
5	**Однофакторный дисперсионный анализ**							
6	ИТОГИ							
7	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*			
8	Строка 1	3	102	34	4			
9	Строка 2	6	216	36	8			
10	Строка 3	4	176	44	66			
11								
12	**Дисперсионный анализ**							
13	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*	
14	Между группами	216	2	108	4,3902439	0,042802157	7,559492587	
15	Внутри групп	246	10	24,6				
16								
17	Итого	462	12					
18								

Рис. 9.2

Как видно, результаты «ручного» решения и решения с помощью Microsoft Excel полностью совпадают.

Более подробные сведения о других инструментах анализа статистического пакета Microsoft Excel и о параметрах соответствующих диалоговых окон можно найти в его справке (см. рис. 9.3).

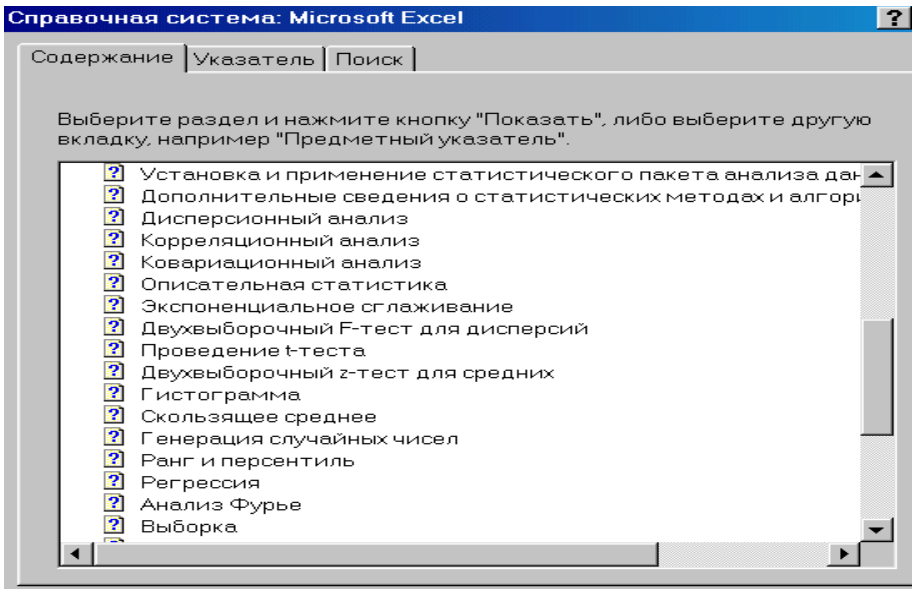


Рис. 9.3

Например, выбрав пункт «Описательная статистика», можно получить следующие выходные статистические данные:

Среднее, Стандартная ошибка (среднего), Медиана, Мода, Стандартное отклонение, Дисперсия выборки, Эксцесс, Асимметричность, Интервал, Минимум, Максимум, Сумма, Счет, Наибольшее (#), Наименьшее (#), Уровень надежности.

Задачи для самостоятельного решения

В задачах 1—8 предполагается, что выборки получены из нормально распределенных генеральных совокупностей с равными дисперсиями.

1. На химико-фармацевтическом заводе разработаны два новых варианта технологического процесса. Чтобы оценить, как изменится дневная производительность труда при переходе на работу по новым технологиям, завод в течение недели (5 рабочих дней) работает по каждому варианту, включая и существующий ранее. Результаты работы в течение трех недель представлены в таблице 9.6. Можно ли с α**=0,05** считать, что средние результаты работы по новым и старому вариантам технологического процесса существенно различаются?

Таблица 9.6

Вариант технологического процесса ***A***	*Дни недели*				
	1	***2***	***3***	***4***	***5***
Существующий A_1	***42***	***44***	***52***	***50***	***47***
I новый вариант A_2	***51***	***52***	***56***	***57***	***54***
II новый вариант A_3	***54***	***55***	***56***	***58***	***52***

2. Результаты эксперимента по изучению влияния температуры (фактор ***A***) на выход продукции при некотором химическом процессе представлены в таблице 9.7.

Таблица 9.7

Температура, ^{0}***C***	*Номер эксперимента*			
	1	***2***	***3***	***4***
40–50	***40***	***44***	***48***	***36***
50–60	***62***	***80***	***71***	***91***
60–70	***92***	***76***	**–**	**–**

Методом дисперсионного анализа при ***α=0,01*** оценить существенность влияния температуры на выход продукции.

3. Сравнение часовой производительности (в литрах) четырех дистилляторов в течение четырех часов дало следующие результаты (табл. 9.8).

Таблица 9.8

Номер дистиллятора	*Часы работы*			
	1	***2***	***3***	***4***
1	***6***	***7***	***8***	***11***
2	***6***	***7***	***11***	***12***
3	***9***	***12***	***14***	***13***
4	***7***	***9***	***10***	***10***

Методом дисперсионного анализа при ***α=0,05*** проверьте, значимо ли различается часовая производительность (в литрах) этих дистилляторов.

4. При ***α=0,05*** методом дисперсионного анализа проверить гипотезу о несущественном влиянии возрастного состава населения города (фактор ***A***) на годовое потребление лекарств (в условных единицах) на душу населения по результатам опроса 20 граждан (по 5 человек в каждом возрастном интервале) (табл. 9.9).

Таблица 9.9

Возраст в годах, ***A***	*Номер наблюдения*				
	1	***2***	***3***	***4***	***5***
20—30	***10***	***8***	***9***	***12***	***13***
30—45	***16***	***17***	***13***	***15***	***14***
45—55	***19***	***21***	***18***	***22***	***30***
Более 55	***15***	***14***	***16***	***19***	***21***

5. В четырех аптеках данные товарооборота (в тыс. руб.) за первые шесть месяцев года сведены в таблицу 9.10.

Таблица 9.10

Аптеки	*Номер месяца*					
	1	***2***	***3***	***4***	***5***	***6***
1	***19***	***23***	***26***	***18***	***20***	***26***

2	***20***	***20***	***32***	***27***	***30***	***21***
3	***16***	***15***	***18***	***26***	***17***	***16***
4	***20***	***18***	***23***	***18***	***17***	***18***

При ***α=0,01*** методом дисперсионного анализа проверить гипотезу о равенстве среднемесячных товарооборотов этих аптек.

6. При ***α=0,01*** оценить существенность влияния фактора ***A*** (температура, ^{0}C) и фактора ***B*** (содержание катализатора, %) на время химической реакции (в сек.). Экспериментальные данные представлены в таблицах 9.11 и 9.12.

Таблица 9.11

A_i \ B_j	***5 %***	***10 %***	***30 %***
20^0C	***2***	***1***	***3***
25^0C	***4***	***6***	***8***
30^0C	***3***	***8***	***10***
35^0C	***11***	***13***	***15***

Таблица 9.12

A_i \ B_j	***2 %***	***4 %***	***8 %***	***10 %***	***6 %***
40^0C	***16***	***17***	***15***	***14***	***13***
43^0C	***10***	***8***	***12***	***11***	***9***
46^0C	***15***	***14***	***19***	***21***	***16***
50^0C	***19***	***21***	***22***	***30***	***18***

7. При анализе работы городской поликлиники исследовалась зависимость числа ее посетителей от их возраста (фактор ***B***) и дня недели (фактор ***A***). Полученные данные представлены в таблице 9.13.

Таблица 9.13

A_i \ B_j	***До 20 лет***	***25—25 лет***	***25—30 лет***	***30—40 лет***	***40—50 лет***	***После 50 лет***
Понедельник						
Вторник						
Среда						
Четверг						
Пятница						

При уровне значимости ***α=0,01*** проверьте, значима ли разница в числе посетителей поликлиники:

— в зависимости от дня недели;

— в зависимости от их возраста.

8. Акционерное общество, специализирующееся на производстве лекарственного растения, провело эксперимент с целью увеличения урожайности. Использовалось пять различных видов удобрений (фактор ***A***) и четыре способа химической обработки семян, почвы и

посевов (фактор ***B***). Результаты эксперимента представлены в таблицах 9.14 и 9.15 в пересчете ц/га.

При ***α=0,05*** оцените значимость влияния вида удобрения и способа химической обработки на урожайность лекарственного растения.

Таблица 9.14

B_j / A_i	Удобрения				
	I	***II***	***III***	***IV***	***V***
1	***16***	***18***	***14***	***12***	***8***
2	***20***	***24***	***32***	***44***	***30***
3	***28***	***32***	***48***	***64***	***50***
4	***24***	***22***	***38***	***50***	***36***

Таблица 9.15

B_i / A_i	Удобрения				
	I	***II***	***III***	***IV***	***V***
1	***10***	***8***	***9***	***12***	***11***
2	***20***	***21***	***17***	***19***	***18***
3	***19***	***21***	***18***	***22***	***30***
4	***11***	***10***	***12***	***15***	***17***

10. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

10.1. Основные понятия и определения

1. Временным рядом (ВР) называется множество наблюдений $x(t_1)$, $x(t_2)$ … $x(t_i)$ … $x(t_n)$ изучаемого признака некоторого процесса или явления, получаемых последовательно во времени.

Временной ряд является непрерывным, если его значение может быть определено в произвольный момент времени, например, изменение температуры химической реакции в технологическом процессе; изменение биопотенциалов, сопровождающих изменение жизнедеятельности клетки или органоида, и т.д.

Дискретными называются временные ряды, значения которых определены в отдельные изолированные моменты времени.

Моментные дискретные временные ряды получают путем измерения значений непрерывного временного ряда в определенные моменты времени, например, измеряя температуру химической реакции через час в течение дня, мы получим моментный временной ряд.

Интервальный дискретный временной ряд получают при усреднении значений временного ряда (непрерывного или дискретного) за некоторый интервал времени, например, отчетные показатели товарооборота медикаментов за месяц, квартал, год и т.д.

2. Временной ряд называется детерминированным, если его будущие значения могут быть заранее точно вычислены.

Временной ряд называется случайным, если его будущие значения нельзя точно предсказать (то есть они описываются с помощью плотности распределения).

Практически временной ряд $x(t)$ всегда содержит как детерминированную $\bar{x}(t)$, так и случайную $\xi(t)$ составляющие:

$$x(t) = \bar{x}(t) + \xi(t). \qquad (10.1)$$

3. Временной ряд называется стационарным, если все его характеристики (например, плотность распределения, m_x, D_x) не изменяются во времени. На практике временные ряды можно считать стационарными на некоторых коротких промежутках времени. Для стационарных рядов их характеристики можно находить усреднением по времени:

$$\hat{m}_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x(t_i), \qquad (10.2)$$

$$\hat{D}_x = S_x^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left[x(t_i) - \bar{x}\right]^2. \qquad (10.3)$$

4. Определение тренда (основной тенденции изменения) временного ряда.

В формуле (1) слагаемое $\bar{x}(t)$ отражает зависимость детерминированной составляющей ВР от времени, то есть основную тенденцию изменения ВР. Чтобы ее найти, необходимо уменьшить случайную составляющую $\xi(t)$ — выровнять (сгладить) временной ряд. Существует ряд методов, позволяющих определить тренд достаточно точно.

10.2. Аналитическое выравнивание методом наименьших квадратов (МНК)

Пусть имеется последовательность значений временного ряда, взятых в моменты времени $t_1, t_2 \ldots t_n$. По этим данным необходимо подобрать вид уравнения, описывающего тренд, и вычислить его коэффициенты.

Наиболее распространены полиномиальные модели тренда, их которых самым простым является линейное уравнение:

$$\bar{x}(t) = a + bt, \tag{10.4}$$

где коэффициенты находятся по методу наименьших квадратов:

$$a = \bar{x} - b\bar{t}, \tag{10.5}$$

$$b = \frac{\sum_{i=1}^{n} x_i t_i - n\bar{x}\cdot\bar{t}}{\sum_{i=1}^{n}(t_i - \bar{t})^2} = \frac{n\sum_{i=1}^{n} x_i t_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} t_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2}. \tag{10.6}$$

Для квадратичной зависимости:

$$\bar{x}(t) = a + b(t - \bar{t}) + c(t - \bar{t})^2 \tag{10.7}$$

коэффициенты равны:

$$b = \frac{\sum_{i=1}^{n} x_i t_i - n\bar{x}\bar{t}}{\sum_{i=1}^{n}(t_i - \bar{t})^2}, \tag{10.8}$$

$$c = \frac{\left(\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n}(t_i - \bar{t})^2 - n\sum_{i=1}^{n} x_i(t_i - \bar{t})^2}{\left[\sum_{i=1}^{n}(t_i - \bar{t})^2\right]^2 - n\sum_{i=1}^{n}(t_i - \bar{t})^4}, \tag{10.9}$$

$$a = \frac{1}{n}\left[\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n}(t_i - \bar{t})^2\right]. \tag{10.10}$$

Вид формулы, описывающей тренд, практически удобно выбирать исходя из минимума дисперсии:

$$S_0^2 = \frac{1}{n - p - 1}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2, \qquad (10.11)$$

где $\bar{x}_i$ — вычисленное по уравнениям значения временного ряда;

P — количество вычисляемых параметров.

Для каждого уравнения вычисляется дисперсия S_0^2 и выбирается то уравнение, для которого она наименьшая.

Использование уравнений третьего и высших порядков для тренда приводит к усложнению расчетов и увеличению погрешностей при оценке коэффициентов ***a, b, c, …*** . Поэтому без хорошей вычислительной базы их практически не применяют, а приводят путем преобразований уравнение тренда к линейному виду.

Точный вид уравнения тренда, как правило, неизвестен, поэтому оптимальное сглаживание не достигается. Если тренд имеет сложный характер или его уравнение неизвестно, то рациональней на первом этапе использование других методов сглаживания.

Пример 1. Имеются данные о поставке предприятием товара (тыс. руб.) за ***5*** месяцев.

Месяц, ***t***	***1***	***2***	***3***	***4***	***5***
Поставки, ***x***	***24,0***	***26,0***	***26,0***	***30,0***	***34,0***

Произвести аналитическое выравнивание ряда:

— по прямой $\bar{x}(t) = a + bt$;

— по параболе $\bar{x}(t) = a + bt + ct^2$.

Определить, какие из этих двух уравнений лучше описывают тренд.

Решение. Результаты расчета промежуточных величин оформим в виде таблицы:

						Сумма	Среднее
Месяц, t_i	***1***	***2***	***3***	***4***	***5***	***15***	$\bar{t} = 15/5 = 3$
x_i, тыс. руб.	***24***	***26***	***26***	***30***	***34***	***140***	$\bar{t} = 140/5 = 28$
$(t_i - \bar{t})$	***-2***	***-1***	***0***	***1***	***2***	***0***	
$x_i \cdot t_i$	***24***	***52***	***78***	***120***	***170***	***444***	
$(t_i - \bar{t})^2$	***4***	***1***	***0***	***1***	***4***	***10***	
$x_i(t_i - \bar{t})^2$	***96***	***26***	***0***	***30***	***136***	***288***	
$(t_i - \bar{t})^4$	***16***	***1***	***0***	***1***	***16***	***34***	

- Вычисляем коэффициент ***a*** и ***b*** по формулам (10.5) и (10.6):

$$b = \frac{444 - 5 \cdot 28 \cdot 3}{10} = \frac{24}{10} = 2{,}4; \quad a = 28 - 2{,}4 \cdot 3 = 20{,}8,$$

отсюда $\bar{x}(t) = 20{,}8 + 2{,}4 \cdot t$.

• Для уравнения параболы по формулам (10.8) — (10.10) находим:

$$b = 2{,}4 \quad ;$$

$$c = \frac{140 \cdot 10 - 5 \cdot 288}{10^2 - 5 \cdot 34} = \frac{1400 - 1440}{100 - 170} = \frac{40}{70} \approx 0{,}57 \; ;$$

$$a \approx \frac{1}{5}[140 - 0{,}57 \cdot 10] \approx \frac{1}{5} \cdot 134{,}3 \approx 26{,}86 \; .$$

Итак, искомое уравнение: $\bar{x}(t) = 26{,}86 + 2{,}4(t-3) + 0{,}57(t-3)^2$. Теперь рассчитаем дисперсии S_0^2 для каждого уравнения тренда, для чего вычислим выровненные по ним значения $\bar{x}_i$.

Для линейного тренда:

$$\bar{x}_1 = \bar{x}(1) = 20{,}8 + 2{,}4 \cdot 1 = 23{,}2 \; ;$$

$$\bar{x}_2 = \bar{x}(2) = 20{,}8 + 2{,}4 \cdot 2 = 25{,}6 \text{ и т.д.}$$

Полученные значения $\bar{x}_i$ поместим в таблицу:

x_i	24	26	26	30	34	$\sum x_i = 140$
$\bar{x}_i$	23,2	25,6	28	30,4	32,8	$\sum \bar{x}_i = 140$
$(x_i - \bar{x}_i)$	0,8	0,4	-2,0	-0,4	1,2	$\sum (x_i - \bar{x}_i) = 0$
$(x_i - \bar{x}_i)^2$	0,64	0,16	4,0	0,16	1,44	$\sum (x_i - \bar{x}_i)^2 = 6{,}40$

Отсюда $S_0^2 = \frac{1}{5-2-1} \cdot 6{,}40 = \frac{6{,}40}{2} = 3{,}2$.

Аналогичные вычисления производим для квадратичного тренда:

$$\bar{x}_1 = \bar{x}(1) = 26{,}86 + 2{,}4(1-3) + 0{,}57(1-3)^2 = 26{,}86 - 4{,}8 + 2{,}28 \approx 24{,}3;$$

$$\bar{x}_2 = \bar{x}(2) = 26{,}86 + 2{,}4(2-3) + 0{,}57(2-3)^2 = 26{,}86 - 2{,}4 + 0{,}5 \approx 25{,}0 \text{ и т.д.}$$

Составим таблицу:

x_i	24	26	26	30	34	$\sum x_i = 140$
$\bar{x}_i$	23,3	25,0	26,9	29,8	33,9	$\sum \bar{x}_i = 139{,}9$
$(x_i - \bar{x}_i)$	-0,3	1,0	-0,9	0,2	0,1	$\sum (x_i - \bar{x}_i) = 0{,}1$
$(x_i - \bar{x}_i)^2$	0,09	1,0	0,81	0,04	0,01	$\sum (x_i - \bar{x}_i)^2 = 1{,}95$

Отсюда $S_0^2 = \frac{1}{5-3-1} \cdot 1{,}95 = 1{,}95$.

Видно, что дисперсия для квадратичного тренда меньше, то есть уравнение параболы лучше описывает основную тенденцию изменения поставок во времени, чем уравнение прямой.

10.3. Метод скользящего среднего

Суть метода состоит в замене значений $x(t_i)$ ВР их среднеарифметическими за определенный интервал из K значений ряда. При нечетном числе (что более удобно) $K=2m+1$ среднее равно:

$$\bar{x}_i = \frac{x_{i-m} + x_{i-m+1} + \ldots + x_i + \ldots + x_{i+m}}{2m+1}. \qquad (10.12)$$

При четном значении временного ряда среднее относят к моменту времени, находящемуся в середине между t_i и t_{i+1} моментом времени:

$$\bar{x}_i = \frac{1/2 \cdot x_{i-m} + x_{i-m+1} + \ldots + x_i + \ldots + 1/2 \cdot x_{i+m}}{2m}. \qquad (10.13)$$

Увеличивая последовательно номер i сглаживаемого значения x_i и вычисляя по формулам (10.13) или (10.14), получают последовательность сглаженных значений временного ряда $\bar{x}_i$.

Такой метод соответствует сглаживанию ряда по МНК отрезками прямых. Ряд сглаженных значений имеет дисперсию в n раз меньшую, чем дисперсия значений исходного ВР:

$$S_{\bar{x}}^2 = \frac{S_x^2}{n}. \qquad (10.14)$$

Поэтому сглаженный ряд более точно отражает тренд.

Как видно из формул (10.13) и (10.14), первые m и последние m значений временного ряда остаются не сглаженными. Для их сглаживания выведены по МНК специальные формулы. Так, при сглаживании по трем значениям для первого и последнего сглаженных значений имеем:

$$\bar{x}_1 = \frac{1}{9}(7x_1 + 4x_2 - 2x_3). \qquad (10.15)$$

Для последнего сглаживания значения порядок нумераций берется обратный.

Длина интервала сглаживания выбирается с учетом скорости изменения ряда и величины разброса его значений. При большой скорости изменения нужно сглаживать по малому числу: 3—5 значений. При малой скорости изменения тренда можно сглаживать по большому интервалу.

Пример 2. Имеются следующие данные о количестве амбулаторных рецептов, поступивших в аптеку за первые ***8*** дней отчетного месяца.

День, t_i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Количество рецептов, x_i	*201*	*143*	*235*	*303*	*266*	*340*	*318*	*344*

Перевести сглаживание ряда методом скользящего среднего по трем отчетам.

Решение. Вычисляем сглаженные значения:

$$\bar{x}_2 = \frac{x_1 + x_2 + x_3}{3} = \frac{201 + 143 + 235}{3} = \frac{579}{3} = 193;$$

$$\bar{x}_3 = \frac{x_2 + x_3 + x_4}{3} = \frac{143 + 235 + 303}{3} = \frac{681}{3} = 227.$$

Аналогично находим: $\bar{x}_4 = 268$; $\bar{x}_5 = 303$; $\bar{x}_6 = 308$; $\bar{x}_7 = 334$.
Крайние значения $\bar{x}_1$ и $\bar{x}_8$ вычисляем по формуле (10.15):

$$\bar{x}_1 = \frac{1}{9}(7 \cdot 201 + 4 \cdot 143 + 2 \cdot 235) = \frac{1509}{9} \approx 168;$$

$$\bar{x}_8 = \frac{1}{9}(7x_8 + 4x_7 + 2x_6) = \frac{1}{9}(7 \cdot 344 + 4 \cdot 318 - 2 \cdot 340) = \frac{1}{9} \cdot 3900 \approx 333.$$

Нанесем скользящие средние и фактические данные на графике (рис. 10.1).

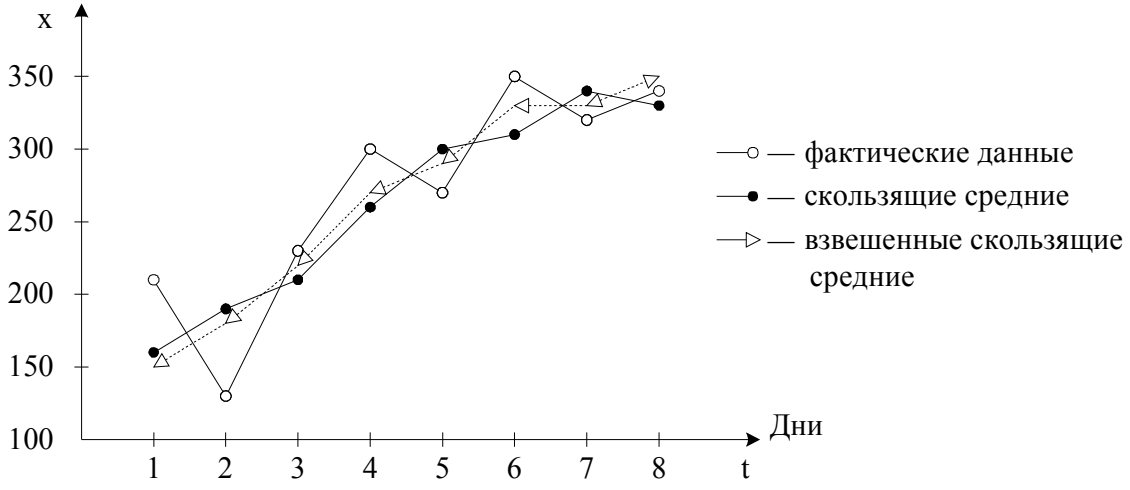


Рис. 10.1

10.4. Метод взвешенного скользящего среднего

Если сглаженный ряд имеет большой разброс, то его можно сгладить еще раз. Повторное сглаживание приводит к тому, что сглаженные значения получаются не просто как средние арифметические, значения временного ряда входят в них уже с некоторыми коэффициентами — «весами»:

$$\bar{x}_i = \sum_{v=i-m}^{i+m} x_v c_v \ , \qquad (10.16)$$

где c_v — весовые коэффициенты, причем $\sum_{v=i-m}^{i+m} c_v = 1$.

Дисперсия сглаженного значения равна:

$$D[x] = \sum_{v=i-m}^{i+m} c_v^2 \ . \qquad (10.17)$$

Разработаны наборы весовых коэффициентов для сглаживания по полиному до пятого порядка. Они представлены в таблице.

Таблица 10.1

Порядок уравнения	*Коэффициенты*
P=2	***(1/12, 4/12, 6/12, 4/12, -1/12)*** или ***(0,1 0,2 0,4 0,2 0,1)*** или ***(1/16, 4/16, 6/16, 4/16, 1/16)***
P=3	***1/22(-2, 1, 4, 5, 6, 5, 4, 1, -2)***
P=4 ***Или*** ***P=5***	***1/32(1, -4, 2, 10, 14, 10, 14, 10, 2, -4, 1)*** ***1/24(1, -3, 2, 7, 10, 7, 2, -3, 1)*** ***1/40(-5, 3, 13, 16, 13, 3, -5, 1)***

Первый набор коэффициентов при ***P=2*** является оптимальным, он получен по МНК. Второй набор при ***P=2*** удобен для ручных расчетов, он не дает оптимального сглаживания. Третий набор коэффициентов при ***P=2*** удобен для реализации на ЭВМ, он также не оптимален. Наборы коэффициентов при ***P=3*** и ***P=4***, ***P=5*** округлены относительно точных зна-

чений для удобства расчетов. При сглаживании по взвешенному скользящему среднему первые и последние значения ***m*** остаются не сглаженными. В некоторых случаях получены по МНК формулы для их сглаживания. Например, при ***P=2*** и ***n=5***:

$$\bar{x}_1 = \frac{1}{10}(7x_1 + 5x_2 - x_3 - x_4), \qquad (10.18)$$

$$\bar{x}_2 = \frac{1}{10}(3x_1 + 5x_2 + x_3 + x_4).$$

Для сглаживания последних значений порядок нумерации в (10.18) меняется на обратный.

Пример 3. Сгладить временной ряд из примера 2 методом взвешенного скользящего среднего по ***n=5*** отсчетам отрезками параболы ***(P=2)***.

Решение. Для удобства расчетов будем сглаживать с набором весов $_1/_{10}$ ***{1; 2; 4; 2; 1}***. Получим:

$$\bar{x}_3 = \frac{x_1 + 2x_2 + 4x_3 + 2x_4 + x_5}{10} = \frac{201 + 2\cdot 143 + 4\cdot 235 + 2\cdot 303 + 266}{10} = \frac{2299}{10} \approx 230;$$

$$\bar{x}_4 = \frac{x_2 + 2x_3 + 4x_4 + 2x_5 + x_6}{10} = \frac{143 + 2\cdot 235 + 4\cdot 303 + 2\cdot 266 + 340}{10} = \frac{2697}{10} \approx 270;$$

$$\bar{x}_5 = \frac{x_3 + 2x_4 + 4x_5 + 2x_6 + x_7}{10} = \frac{235 + 2\cdot 303 + 4\cdot 266 + 2\cdot 340 + 318}{10} = \frac{2903}{10} \approx 290;$$

$$\bar{x}_6 = \frac{x_4 + 2x_5 + 4x_6 + 2x_7 + x_8}{10} = \frac{303 + 2\cdot 266 + 4\cdot 340 + 2\cdot 318 + 344}{10} = \frac{3175}{10} = 317{,}5 \approx 318.$$

Два первых и два последних сглаженных значения находим по формулам (10.18):

$$\bar{x}_1 = \frac{7x_1 + 5x_2 - x_3 - x_4}{10} = \frac{7\cdot 201 + 5\cdot 143 - 235 - 303}{10} = \frac{1584}{10} \approx 158;$$

$$\bar{x}_8 = \frac{7x_8 + 5x_7 - x_6 - x_5}{10} = \frac{7\cdot 344 + 5\cdot 318 - 340 - 266}{10} = \frac{3392}{10} \approx 339;$$

$$\bar{x}_2 = \frac{3x_1 + 5x_2 + x_3 + x_4}{10} = \frac{3\cdot 201 + 5\cdot 143 + 235 + 303}{10} = \frac{1856}{10} \approx 186;$$

$$\bar{x}_7 = \frac{3x_8 + 5x_7 + x_6 + x_5}{10} = \frac{3\cdot 344 + 5\cdot 318 + 340 + 266}{10} = \frac{3228}{10} \approx 323.$$

Полученные сглаженные значения нанесем на график (см. рис. 10.1).

10.5. Экспоненциальное сглаживание

Весовые коэффициенты при экспоненциальном сглаживании уменьшаются по экспоненциальному закону по мере удаления в «прошлое», то есть весовой коэффициент сглаживаемого значения ряда максимален, а первого значения ряда — минимален. Экспоненциальные средние вычисляются по рекуррентной формуле:

$$\bar{x}_i^{(n)} = \alpha \cdot x_i^{(k-1)} + (1-\alpha)\cdot \bar{x}_{i-1}^{(k)}, \qquad (10.19)$$

где ***(к)*** — порядок сглаживания;

$\alpha = \dfrac{2}{n+1}$ — коэффициент сглаживания;

n — длина интервала сглаживания, которая выбирается по тем же принципам, что и при сглаживании по взвешенному скользящему среднему, дисперсия сглаживания по этому методу значения примерно в два раза больше, чем при усреднении. При вычислениях первое сглаженное значение первого порядка удобно считать равным первому значению временного ряда: $\bar{x}_1^{(1)} = x_1$.

Пример 4. Вычислить экспоненциальные средние первого и второго порядка для временного ряда из примера 1.

Решение. Выберем длину интервала сглаживания ***n***, равную ***2***, тогда

$$\alpha = \frac{2}{n+1} = \frac{2}{2+1} = \frac{2}{3}.$$

Исходные данные и вычисляемые средние разместим в таблице:

t_i	1	2	3	4	5
x_i	24	26	26	30	34
$\bar{x}_i^{(1)}$	24	25,3	25,8	28,6	32,2
$\bar{x}_i^{(2)}$	24	24,9	25,5	27,6	30,7

Примем, что $\bar{x}_1^{(1)} = \bar{x}_1^{(2)} = x_1 = 24$. Остальные экспоненциальные средние рассчитываются:

$$\bar{x}_2^{(1)} = \alpha \cdot x_2 + (1-\alpha)\bar{x}_1^{(1)} = \frac{2}{3}\cdot 26 + \left(1-\frac{2}{3}\right)\cdot 24 = \frac{2\cdot 26+24}{3} = \frac{76}{3} \approx 25,3;$$

$$\bar{x}_3^{(1)} = \alpha \cdot x_3 + (1-\alpha)\bar{x}_2^{(1)} = \frac{2\cdot 26+25,3}{3} = \frac{77,3}{3} \approx 25,8;$$

$$\bar{x}_4^{(1)} = \frac{2\cdot 30+25,8}{3} = \frac{85,8}{3} \approx 28,6;$$

$$\bar{x}_5^{(1)} = \frac{1}{3}(2 \cdot 34 + 28{,}6) = \frac{96{,}6}{3} = 32{,}2 .$$

Теперь вычисляем $\bar{x}_i^{(2)}$:

$$\bar{x}_2^{(2)} = \alpha \cdot \bar{x}_2^{(1)} + (1-\alpha)\bar{x}_1^{(2)} = \frac{2}{3} \cdot 25{,}3 + \frac{1}{3} \cdot 24 = \frac{2 \cdot 25{,}3 + 24}{3} = \frac{74{,}6}{3} \approx 24{,}9 ;$$

$$\bar{x}_3^{(2)} = \frac{1}{3}(2 \cdot 25{,}8 + 24{,}9) = \frac{76{,}5}{3} = 25{,}5 ;$$

$$\bar{x}_4^{(2)} = \frac{1}{3}(2 \cdot 28{,}6 + 25{,}5) = \frac{82{,}7}{3} \approx 27{,}6 ;$$

$$\bar{x}_5^{(2)} = \frac{1}{3}(2 \cdot 32{,}2 + 27{,}6) = \frac{92{,}0}{3} \approx 30{,}7 .$$

10.6. Прогнозирование временных рядов

Для решения задачи прогнозирования необходимо по экспериментальным данным о временном ряде составить уравнение тренда и, предлагая, что выявленная тенденция изменения сохранится в будущем, рассчитать по этому уравнению значения ряда в предстоящие моменты времени. Чаще всего в медицине для прогнозирования используются метод наименьших квадратов и экспоненциальное сглаживание, которые наиболее просты и наглядны.

Прогноз обычно осуществляется по полиномиальной модели первого, второго и так далее порядков. Для метода наименьших квадратов используются уравнения (10.4) — (10.10).

При экспоненциальном сглаживании прогноз осуществляется по линейной модели:

$$\bar{x}(t + \Delta t) = \bar{a}_0 + \bar{a}_1 \Delta t , \qquad (10.20)$$

где

$$\bar{a}_0 = 2\bar{x}^{(1)}(t) - \bar{x}^{(2)}(t) ,$$

$$\bar{a}_1 = \frac{\alpha}{1-\alpha}[\bar{x}^{(1)}(t) - \bar{x}^{(2)}(t)] ;$$

или по квадратичной зависимости:

$$\bar{x}(t + \Delta t) = \bar{a}_0 + \bar{a}_1 \Delta t + \bar{a}_2 \Delta t^2 ,$$

где

$$\bar{a}_0 = 3[\bar{x}^{(1)}(t) - \bar{x}^{(2)}(t)] + \bar{x}^{(3)}(t) ;$$

$$\bar{a}_1 = \frac{\alpha}{2(1-\alpha)^2}[(6-5\alpha)\bar{x}^{(1)}(t) - 2(5-4\alpha)\bar{x}^{(2)}(t) + (4-3\alpha)\bar{x}^{(3)}(t)] ;$$

$$\bar{a}_2 = \frac{\alpha}{(1-\alpha)^2}\left[\bar{x}^{(1)}(t) - 2\bar{x}^{(2)}(t) + \bar{x}^{(3)}(t)\right]. \qquad (10.21)$$

Как видно из уравнений (10.20) и (10.21), коэффициенты $\bar{a}_0, \bar{a}_1, \bar{a}_2$ вычисляются по экспоненциальным сглаженным $\bar{x}(t)$, полученным к моменту времени t, то есть отражающим тенденцию изменения временного ряда. Поэтому прогноз по методу экспоненциального сглаживания, как правило, точней, чем по МНК.

Пример 5. Вычислить прогнозируемые значения поставок предприятием товара в ***6*** и *7* месяцы (условия примера 1) по линейной модели, коэффициенты которой найти: а) по МНК, б) с помощью экспоненциального сглаживания.

Решение

1. При решении примера 1 было найдено МНК линейное уравнение тренда: $\bar{x}(t) = 20{,}8 + 2{,}4t$. Предполагая, что выявленная тенденция изменения сохранится на ***6 (t=6)*** и *7 (t=7)* месяцы, получим прогноз:

$$\bar{x}(6) = 20{,}8 + 2{,}4 \cdot 6 = 20{,}8 + 14{,}4 = 35{,}2 \text{ (тыс. руб.);}$$

$$\bar{x}(7) = 20{,}8 + 2{,}4 \cdot 7 = 20{,}8 + 16{,}8 = 37{,}6 \text{ (тыс. руб.).}$$

2. Для нахождения коэффициентов $\bar{a}_0$ и $\bar{a}_1$ линейной модели прогноза: $\bar{x}(t + \Delta t) = \bar{a}_0 + \bar{a}_1 \Delta t$ по формулам (10.20) воспользуемся при результатами решения примера 4, а именно:

$$\bar{x}_5^{(1)} = 32{,}2 \text{ и } \bar{x}_5^{(2)} = 30{,}7.$$

Тогда

$$\bar{a}_0 = 2 \cdot \bar{x}_5^{(1)} - \bar{x}_5^{(2)} = 2 \cdot 32{,}2 - 30{,}7 = 33{,}7;$$

$$\bar{a}_1 = \frac{\alpha}{1-\alpha}\left(\bar{x}_5^{(1)} - \bar{x}_5^{(2)}\right) = \frac{2/3}{1-2/3}\left(32{,}2 - 30{,}7\right) = 2 \cdot 1{,}5 = 3{,}0.$$

Искомое уравнение для прогноза будет иметь вид:

$$\bar{x}(5 + \Delta t) = 33{,}7 + 3{,}0 \cdot \Delta t.$$

Отсюда:

$$\bar{x}(6) = \bar{x}(5+1) = 33{,}7 + 3{,}0 \cdot 1 = 36{,}7 \text{ (тыс. руб.);}$$

$$\bar{x}(7) = \bar{x}(5+2) = 33{,}7 + 3{,}0 \cdot 2 = 39{,}7 \text{ (тыс. руб.).}$$

Во втором случае прогнозируемые значения получились больше, чем по МНК (а). Это и ожидалось, так как скорость изменения последних значений временного ряда $(\Delta x = x_5 - x_4 = 34 - 30 = 4)$, которым больше «доверяет» метод экспоненциального сглаживания, выше, чем в среднем по всему ряду $\left(\overline{\Delta x} = \frac{34-24}{4} = 2{,}5\right)$.

10.7. Корреляционный анализ временных рядов по наборам условных средних

Корреляционный анализ временных рядов используется в медицине и фармации для решения разнообразных задач: выявление скрытых периодичностей, взаимосвязи между значениями временного ряда в различные моменты времени, анализ ЭКГ, потребления лекарств и т.п.

Корреляционной функцией (КФ) случайных стационарных процессов ***X(t)*** и ***Y(t)*** называется математическим ожидание произведения отклонения значений процессов, взятых в моменты времени t и $t+\tau$, от их математических ожиданий:

$$K_{xy}(\tau) = M[(x(t) - m_x)(y(t+\tau) - m_y)]. \qquad (10.22)$$

В качестве аргумента здесь используется величина τ, называемая временем задержки. Таким образом, корреляционная функция — это зависимость коэффициента корреляции от времени. Оценка нормированной корреляционной функции (НКФ) для каждого значения производится аналогично оценке нормированного коэффициента корреляции методом умножения. Однако при оперативном анализе использование этого метода приводит к громоздким и сложным расчетам. Одним из упрощенных и достаточно точных методов вычисления НКФ является метод набора условных средних, в котором операции умножения заменены весовым суммированием. Оценка НКФ при этом получается в виде:

$$\hat{\rho}_{xy}(\tau) = \frac{\sigma_x}{\sigma_y} \cdot \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{\ell} y(t_{ij} + \tau) \cdot c_i , \qquad (10.23)$$

где $c_i = \dfrac{a_i - \bar{a}}{\sum_{i=1}^{\ell} (a_i - \bar{a})^2}$ — весовые коэффициенты; ℓ — количество уровней;

$\bar{a} = \dfrac{1}{\ell} \sum_{i=1}^{\ell} a_i$ — среднее значение назначенных уровней a_i; $y(t_{ij} + \tau)$ — значение процесса ***Y(t)***, взятые через интервал времени τ после пересечения процессом ***X(t)*** уровня i; n — объем выборки; $j=1, 2, \ldots, n$.

Как видно из формулы (10.23), эта оценка НКФ инвариантна по отношению к математическому ожиданию, то есть не требует вычисления m_x и m_y и центрирования. По сравнению с методом умножения точность оценки (10.23) в два раза выше при равных объемах выборки, что позволяет сократить объем вычислительных операций в два раза. Формула (10.23) применима в любом законе распределения величины ***X*** и ***Y***.

Пример 6. Вычислить оценку НКФ для $\tau=1$ *с* по данным рисунке 10.2.

Решение

1. Находим моменты пересечения кривой $x(t)$ с уровнями $a_1=20$ $a_2=60$ и через интервал τ измеряем значения $x(t_{11}+\tau)$, $x(t_{21}+\tau)$ и т.д.

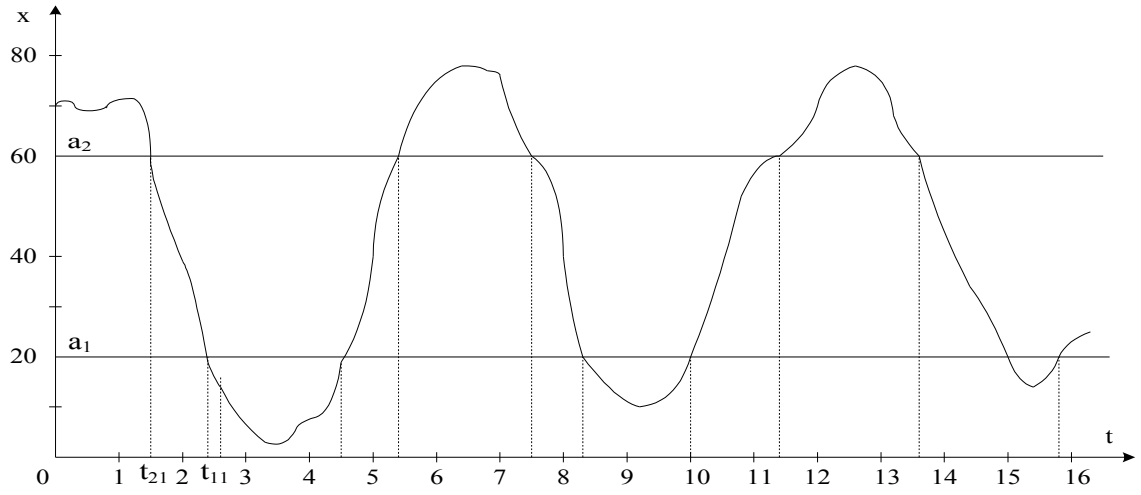


Рис. 10.2. Альфа-ритм энцефалограммы

2. Измеренные значения заносим в таблицу:

$x(t_{1j}+\tau)$	2	66	14	56	27	n=5
$x(t_{2j}+\tau)$	15	78	17	74	39	n=5
$x(t_{2j}+\tau)-x(t_{rj}+\tau)$	13	12	3	18	12	

3. Вычисляем весовые коэффициенты:

$$\bar{a}=\frac{1}{2}(a_1+a_2)=\frac{20+60}{2}=40\ ;\quad c_i=\frac{a_i-\bar{a}}{\sum_{i=1}^{\ell}(a_i-\bar{a})^2}=\frac{20-40}{800}=-\frac{1}{40}\ ;$$

$$c_2=\frac{60-40}{800}=\frac{1}{40}\ .$$

4. Находим оценку НКФ:

$$\hat{\rho}_{xy}(\tau)=\frac{1}{5}\sum_{j=1}^{2}\sum_{i=1}^{5}x(t_{ij}+\tau)\cdot c_i=\frac{1}{5}\left[-\frac{1}{40}(2+66+14+56+27)+\frac{1}{40}(15+78+17+74+39)\right]=$$

$$=\frac{1}{200}[(15-2)+(78-66)+(17-14)+(74-56)+(39-27)]=0{,}29\ .$$

Аналогичным образом находятся оценки НКФ при $\tau=2c, 3c, \ldots, 10c$. Полученные данные наносят на график и определяют основную частоту альфа-ритма.

Задачи для самостоятельного решения

В задачах 1—5 определить вид уравнения тренда, найти по МНК его коэффициенты и вычислить прогнозируемые значения для $\Delta t=1$ *и* $\Delta t=2$.

1. Розничный товарооборот аптеки (в тыс. руб.) за ***10*** месяцев отчетного года:

Месяц, t_i	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
x_i, *тыс. руб.*	***7,4***	***7,9***	***8,7***	***8,2***	***7,9***	***8,2***	***8,7***	***8,1***	***8,3***	***8,4***

2. Поставка товаров заводом за первое полугодие:

Месяц, t_i	***1***	***2***	***3***	***4***	***5***	***6***
x_i, *тыс. руб.*	***240***	***280***	***260***	***270***	***340***	***410***

3. Уровень готовых лекарственных средств в общей рецептуре аптеки за год:

Год	***1990***	***1991***	***1992***	***1993***	***1994***	***1995***	***1996***	***1997***	***1998***	***1999***	***2000***
x_i *%*	***51,6***	***52,4***	***53,6***	***54,2***	***54,4***	***55,5***	***56,0***	***56,9***	***58,0***	***59,4***	***60,2***

4. Количество стационарных рецептов в год:

Год	***1994***	***1995***	***1996***	***1997***	***1998***	***1999***	***2000***
x_i, *тыс.*	***167,4***	***168,3***	***170,7***	***171,0***	***174,2***	***175,1***	***178,6***

5. Фонд заработной платы за год:

Год	***1991***	***1992***	***1993***	***1994***	***1995***	***1996***	***1997***	***1998***	***1999***	***2000***
x_i, тыс. руб.	***25,4***	***25,8***	***26,3***	***26,9***	***26,7***	***27,8***	***29,4***	***30,2***	***30,4***	***31,5***

6. Имеются следующие данные о количестве лекарств индивидуального изготовления за первые ***10*** дней месяца:

День, t_i	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
x_i	***183***	***102***	***164***	***153***	***128***	***171***	***113***	***144***	***163***	***157***

Произвести сглаживание ряда: а) методом скользящего среднего по ***n=3*** значениям; б) методом взвешенного скользящего среднего по ***n=5*** значениям; в) методом аналитического выравнивания по прямой.

Вычислить прогнозируемые значения на ***t=11*** и ***t=12*** дни с помощью экспоненциального сглаживания.

7. Имеются данные о ежегодных затратах на закупку лекарственных трав по аптеке (руб.) в 1990—2000 годах:

Год	***1990***	***1991***	***1992***	***1993***	***1994***	***1995***	***1996***	***1997***	***1998***	***1999***	***2000***
x_i, *руб.*	***560***	***608***	***985***	***807***	***839***	***914***	***1100***	***1196***	***1490***	***1574***	***1513***

- Произведите аналитическое выравнивание ряда по следующим уравнениям тренда:

а) линейному $\bar{x}(t) = a + bt$ и криволинейным, преобразовав их в линейные:

б) $\bar{x}(t) = ae^{bt}$;

в) $\bar{x}(t) = at^b$;

г) $\bar{x}(t) = a + b/t$;

д) $\bar{x}(t) = \dfrac{1}{a + bt}$;

е) $\bar{x}(t) = a + b\ln(t)$.

- Вычислите по этим уравнениям прогнозируемое значение на 2001 г. ***(t=12)***.

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

	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	3652	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	3064	3011	2989	2966	2943	2920
0,8	2897	2874	2850	.2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2168	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	1589	]518
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	0449
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	0043
3,0	0,0044	0043	0042	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 $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$

x	Ф (x)	x	Ф(x)	x	Ф (x)	x	Ф (x)	x	Ф (x)
0,00	0,0000	0,37	0,1443	0,74	0,2703	1,11	0,3665	1,48	0,4306
0,01	0,0040	0,38	0,1480	0,75	0,2734	.1,12	0,3686	1,49	0,4319
0,02	0,0080	0,39	0,1517	0,76	0,2764	1,13	0,3708	1,50	0,4332
0,03	0,0120	0,40	0,1554	0,77	0,2794	1,14	0,3729	1,51	0,4345
0,04	0,0160	0,41	0,1591	0,78	0,2823	1,15	0,3749	1,52	0,4357
0,05	0,0199	0,42	0,1628	0,79	0,2852	1,16	0,3770	1,53	0,4370
0,06	0,0239	0,43	0,1664	0,80	0,2881	1,17	0,3790	1,54	0,4382
0,07	0,0279	0,44	0,1700	0,81	0,2910	1,18	0,3810	1,55	0,4394
0,08	0,0319	0,45	0,1736	0,82	0,2939	1,19	0,3830	1,56	0,4406
0,09	0,0359	0.46	0,1772	0,83	0,2967	1,20	0,3849	1,57	0,4418
0,10	0,0398	0,47	0,1808	0,84	0,2995	1,21	0,3869	1,58	0,4429
0,11	0,0438	0,48	0,1844	0,85	0,3023	1,22	0,3883	1,59	0,4441
0,12	0,0478	0,49	0,1879	0,86	0,3051	1,23	0,3907	1,60	0,4452
0,13	0,0517	0,50	0,1915	0,87	0,3078	1,24	0,3925	1,61	0,4463
0,14	0,0557	0,51	0,1950	0,88	0,3106	1,25	0,3944	1,62	0,4474
0,15	0,0596	0,52	0,1985	0,89	0,3133	1,26	0,3962	1,63	0,4484
0,16	0,0636	0,53	0,2019	0,90	0,3159	1,27	0,3980	1,64	0,4495
0,17	0,0675	0,54	0,2054	0,91	0,3186	1,28	0,3997	1,65	0,4505
0,18	0,0714	0,55	0,2088	0,92	0,3212	1,29	0,4015	1,66	0,4515
0,19	0,0753	0,56	0,2123	0,93	0,3238	1,30	0,4032	1,67	0,4525
0,20	0,0793	0,57	0,2157	0,94	0,3264	1,31	0,4049	1,68	0,4535
0,21	0,0832	0,58	0,2190	0,95	0,3289	1,32	0,4066	1,69	0,4545
0,22	0,0871	0,59	0,2224	0,96	0,3315	1,33	0,4082	1,70	0,4554
0,23	0,0910	0,60	0,2257	0,97	0,3340	1,34	0,4099	1,71	0,4564
0,24	0,0948	0,61	0,2291	0,98	0,3365	1,35	0,4115	1,72	0,4573
0,25	0,0987	0,62	0,2324	*0,*99	0,3389	1,36	0,4131	1,73	0,4582
0,26	0,1026	0,63	0,2357	***1,00***	***0,341***	1,37	0,4147	1,74	0,4591
0,27	0,1064	0,64	0,2389	1,01	0,3438	1,38	0,4162	1,75	0,4599
0,28	0,1103	0,65	0,2422	1,02	0,3461	1,39	0,4177	1,76	0,4608
0,29	0,1141	0,66	0,2454	1,03	0,3485	1,40	0,4192	1,77	0,4616
0,30	0,1179	0,67	0,2486	1,04	0,3508	1,41	0,4207	1,78	0,4625
0,31	0,1217	0,68	0,2517	1,05	0,3531	1,42	0,4222	1,79	0,4633
0,32	0,1255	0,69	0,2549	1,06	0,3554	1,43	0,4236	1,80	0,4641
0,33	0,1293	0,70	0,2580	.1,07	0,3577	1,44	0,4251	1,81	0,4649
0,34	0,1331	0,71	0,2611	1,08	0,3599	1,45	0,4265	1,82	0,4656
0,35	0,1368	0,72	0,2642	1,09	0,3621	1,46	0,4279	1,83	0,4664
0,36	0,1406	0,73	0,2673	1,10	0,3643	1,47	0,4292	1,84	0,4671

x	**Ф (x)**	**x**	**Ф(x)**	**x**	**Ф (x)**	**x**	**Ф (x)**
1,85	0,4678	2,08	0,4812	2,46	0,4931	2,84	0,4977
1,86	0,4686	2,10	0,4821	2,48	0,4934	2,86	0,4979
1,87	0,4693	2,12	0,4830	2,50	0,4938	2,88	0,4980
1,88	0,4699	2,14	0,4838	2,52	0,4941	2,90	0,4981
1,89	0,4706	2,16	0,4846	2,54	0,4945	2,92	0,4982
1,90	0,4713	2,18	0,4854	2,56	0,4948	2,94	0,4984
1,91	0,4719	2,20	0,4861	2,58	0,4951	2,96	0,4985
1,92	0,4726	2,22	0,4868	2,60	0,4953	2,98	0,4986
1,93	0,4732	2,24	0,4875	2,62	0,4956	***3,00***	***0,49865***
1,94	0,4738	2,26	0,4881	2,64	0,4959	3,20	0,49931
1,95	0,4744	2,28	0,4887	2,66	0,4961	3,40	0,49966
1,96	0,4750	2,30	0,4893	2,68	0,4963	3,60	0,499841
1,97	0,4756	2,32	0,4898	2,70	0,4965	3,80	0,499928
1,98	0,4761	2,34	0,4904	2,72	0,4967	4,00	0,499968
1,99	0,4767	2,36	0,4909	2,74	0,4969	4,50	0,499997
2,00	***0,4772***	2,38	0,4913	2,76	0,4971	5,00	0,499997
2,02	0,4783	2,40	0,4918	2,78	0,4973		
2,04	0,4793	2,42	0,4922	2,80	0,4974		
2,06	0,4803	2,44	0,4927	2,82	0,4976		

Приложение 3

Таблица значений $t_\gamma = t(\gamma, n)$

n	*γ*			*n*	*γ*		
	0,95	***0,99***	***0,999***		***0,95***	***0,99***	***0,999***
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,991	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

Критические точки распределения χ^2

Уровень значимости α						
Число степеней свободы k	***0,001***	***0,005***	***0,01***	***0,025***	***0,05***	***0,1***
1	10,8	7,88	6,6	5,0	3,8	2,71
2	13,8	10,6	9,2	7,4	6,0	4,61
3	16,3	12,8	11,3	9,4	7,8	6,25
4	18,5	14,9	13,3	11,1	9,5	7,78
5	20,5	16,7	15,1	12,8	11,1	9,24
6	22,5	18,5	16,8	14,4	12,6	10,6
7	24,3	20,3	18,5	16,0	14,1	12,0
8	26,1	22,0	20,1	17,5	15,5	13,4
9	27,9	23,6	21,7	19,0	16,9	14,7
10	29,6	25,2	23,2	20,5	18,3	16,0
11	31,3	26,8	24,7	21,9	19,7	17,3
12	32,9	28,3	26,2	23,3	21,0	18,5
13	34,5	29,8	27,7	24,7	22,4	19,8
14	36,1	31,3	29,1	26,1	23,7	21,1
15	37,7	32,8	30,6	27,5	25,0	22,3
16	39,3	34,3	32,0	28,8	26,3	23,5
17	40,8	35,7	33,4	30,2	27,6	24,8
18	42,3	37,2	34,8	31,5	28,9	26,0
19	43,8	38,6	36,2	32,9	30,1	27,2
20	45,3	40,0	37,6	34,2	31,4	28,4
21	46,8	41,4	38,9	35,5	32,7	29,6
22	48,3	42,8	40,3	36,8	33,9	30,8
23	49,7	44,2	41,6	38,1	35,2	32,0
24	51,2	45,6	43,0	39,4	36,4	33,2
25	52,6	46,9	44,3	40,6	37,7	34,4
26	54,1	48,3	45,6	41,9	38,9	35,6
27	55,5	49,6	47,0	43,2	40,1	36,7
28	56,9	51,0	48,3	44,5	41,3	37,9
29	58,3	52,3	49,6	45,7	42,6	39,1
30	59,7	53,7	50,9	47,0	43,8	40,3
40	73,4	66,8	63,7	59,3	55,8	51,8
50	86,7	79,5	76,2	71,4	67,5	63,2
75	118,6	110,3	106,4	100,8	96,2	91,1
100	149,4	140,2	135,6	129,6	124,3	118,5

Приложение 5

Критические точки распределения Стьюдента

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	3,95
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	***0,025***	***0,01***	***0,005***	***0,001***	***0,0005***
	Уровень значимости α (односторонняя критическая область)					

Приложение 6

Критические точки распределения *F* Фишера

(k_1 — число степеней свободы большей дисперсии,
k_2 — число степеней свободы меньшей дисперсии)

Уровень значимости $\alpha = 0,01$												
k_1 / k_2	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	99,30	99,33	99,36	99,37	99,39	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	9,86	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

Уровень значимости $\alpha = 0,05$												
k_1 / k_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	210	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3.59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

ОТВЕТЫ

1

1. а) $^1/_3$; б) $^1/_5$; в) $^{59}/_{90}$; г) $^{41}/_{90}$; д) $^1/_{10}$. **2.** а) $^1/_{30}$; б) $^1/_{60}$. **3.** а) $^{24}/_{91}$; б) $^2/_{91}$; в) $^{45}/_{91}$. **4.** а) $^1/_{20}$; б) $^1/_{49}$. **5.** а) $\frac{C_5^1 \cdot C_{15}^2}{C_{20}^3}$; б) $\frac{C_5^2 \cdot C_{15}^1}{C_{20}^3}$; в) $\frac{C_5^3}{C_{20}^3}$. **6.** а) $^{28}/_{435}$; б) $^{231}/_{435}$; в) $^{60}/_{435}$. **7.** $^{24}/_{105}$. **8.** а) (2·8!)/10!=1/45; б) (2·9!)/10!= =1/5. **9.** а) (6·7!)/10!=1/120; б) 1/120. **10.** а) $\frac{C_{12}^6 \cdot C_8^4}{C_{20}^{10}}$; б) $\frac{C_8^4 \cdot C_{12}^6}{C_{20}^{10}}$. **11.** N≈M·n/m. **14.** $\frac{3\sqrt{3}}{4\pi}$. **15.** $^1/_8$. **16.** а) 1; б) 2/π. **17.** (a-r)/a. **18.** 1/6. **19.** 0,25. **20.** 0,75. **21.** $^5/_9$. **22.** а) 2/π; б) $\frac{3\sqrt{3}}{4\pi}$. **23.** $\frac{\left(a-2\sqrt{2}r\right)^2}{a^2}$. **24.** а) 0,33; б) 0,67; в) $^{790}/_{2937}$. **25.** а) $^7/_{22}$; б) $^5/_{33}$; в) $^{70}/_{132}$; г) $^{31}/_6$. **26.** а) 0,0494; б) ≈0,0684. **27.** а) 0,648; б) 0,954. **28.** а) ≈0,922; б) $31{,}36\cdot10^{-6}$; в) ≈0,0753. **29.** а) ≈0,5577; б) ≈0,411; в) ≈0,440. **30.** а) 0,1543; б) 0,15355; в) 0,9922. **31.** а) $^4/_9$; б) $^5/_9$; в) $^2/_9$. **32.** а) ≈0,986; б) ≈0,1855; в) ≈0,9963. **33.** ≈0,9963. **34.** а) $^5/_{18}$; б) $^2/_5$. **35.** $^7/_{16}$. **36.** а) 0,9878; б) ≈0,3968. **37.** 0,9989. **38.** а) $^3/_{29}$; б) $^{10}/_{29}$; в) $^{16}/_{29}$. **39.** а) 0,56; б) 0,32; в) 0,12. **40.** а) $^4/_7$; б) $^3/_7$. **41.** а) ≈0,931; б) ≈0,069. **42.** ≈0,1284. **43.** а) ≈0,4598; б) ≈0,04598.

2

1. а) 0,0486; б) 0,9963. **2.** а) 0,0183; б) 0,9817. **3.** 0,9776. **4.** а) 0,322; б) 0,0113. **5.** а) 0,156; б) 0,762. **6.** $1-e^{-10x}>0{,}9$, откуда $x>0{,}230$ м^3. **7.** а) 0,0419; б) 0,419. **8.** а) 0,168; б) 0,483. **9.** а) 0,554; б) 0,664. **10.** а) 0,3164; б) 0,9694.

3

1.

x_i	0	1	2	3
p_i	0,008	0,096	0,384	0,512

$$F(x)=\begin{cases}0, & \text{если } x\le 0,\\ 0{,}008, & \text{если } 0<x\le 1,\\ 0{,}104, & \text{если } 1<x\le 2,\\ 0{,}488, & \text{если } 2<x\le 3,\\ 1, & \text{если } x>3.\end{cases}$$

2. а)

x_i	0	1	2	3	4

p_i	1/256	12/256	54/256	108/256	81/256

б)

y_i	1	2	3	4
p_i	3/4	3/16	3/64	1/64

3.

x_i	0	1	2	3	…	5000
p_i	e^{-5}	$5\,e^{-5}$	$\frac{25}{2}e^{-5}$	$\frac{125}{6}e^{-5}$		$\frac{5^{5000}}{5000!}e^{-5}$

а) $P(x<3)=18{,}5\cdot e^{-5}\approx 0{,}125$; б) $P(x>2)=1-P(x\le 2)\approx 0{,}875$; в) $P(x\ge 1)= 1-P(x=0)= =1- e^{-5}\approx 0{,}993$.

4. а) $e^{-1/3}=\frac{1}{\sqrt[3]{e}}\approx 0{,}716$; б) $e^{-4/3}\approx 0{,}263$.

5.

x_i	0	1	2	3
p_i	≈0,421	0,446	≈0,124	≈0,009

6.

x_i	0	5	10	50
p_i	0,85	0,10	0,04	0,01

m_x=1,4 руб., $\sigma_x\approx$5,44 руб.

7. 1) m_x=18,5, D[x]=20,25; σ_x=4,5; 2) $P(x\le 18{,}5)=0{,}4$. **8.** а) m_x=21,5; D[x]=25,25; $\sigma_x\approx$5,02; 0,8; б) m_x=4,6; D[x]=8,64; $\sigma_x\approx$2,94; 0,7. **9.** m_x=0,9; D[x]=0,49; m_y=0,3; D[y] ≈0,41.

4

1. а) $F(x)=x^2$; ($0\le x\le 1$); б) 0,84; в) $m_x=2/3$; $D_x=1/18$.

2. а) $F(x)=\begin{cases}0; & x\le 0,\\ 1/4x; & 0<x\le 1,\\ 1/4(3x-2); & 1<x\le 2,\\ 1; & x>2.\end{cases}$, m_x=1,25; D_x=13/48; $P(0<x<1{,}5)=5/8$; $P(x>1{,}5)=3/8$. **3.** a=0,5; $P(|x|<\pi/4)=\sqrt{2}/2$. **4.** а) $1-e^{-1,5}$; б) e^{-3}. **5.** а) $1-e^{-1}$; б) e^{-2}. **6.** 0,25. **7.** б) 10 мин.; в) 0,25; 0,25; 0. **8.** $P(|x|<0{,}1)=2\Phi(0{,}1/\sigma_x)-1=0{,}9;\ \sigma_x\approx 6{,}8$ *см*. **9.** ≈ *78,9%*. **10.** а) 0,4404; б) 0,1974.

5

1. 0,792. **2.** n≥50. **3.** 0,6. **4.** n≥784. **5.** 0,6. **6.** n≥366.

6

1.

Интервалы	2—4	4—6	6—8	8—10
Частоты m_i	9	7	10	4

2. $\bar{x}_в = 3{,}37$ кг; $\sigma_в \approx 0{,}63$ кг. **5.** а) $m_x = 6{,}47$; б) $m_x = 6{,}3$; $D_в[x] \approx 11{,}0$; $S_x^2 \approx 11{,}4$. **6.** $\bar{x}_в = 106\,\%$; $S_x^2 \approx 57{,}6$. **7.** 1) (9,2; 10,8); 2) (9,0; 11,0); 3) (9,0; 11,0); 4) (12,0; 16,0). **8.** 1) 4,20±0,21; 2) 4,20±0,17; 3) 4,20±0,30; 4) 4,20±0,41. **9.** $n \geq t_\gamma^2 \cdot \frac{\sigma_x^2}{\delta^2} = 384{,}16$; n≥385. **11.** $2{,}14<m_x<2{,}18$; $0<\sigma_x<0{,}037$; $q_{0{,}95}(5) = 1{,}37$. **12.** а) $\bar{x}_в \approx 5{,}07$ кг; б) $D_в[x] \approx \approx 1{,}129$; $S_x^2 \approx 1{,}168$; в) $S_x \approx 1{,}08$; ν=21,3%; г) $m_0=5$; $m_l=5$; d=0,752. **13.** $0{,}255<p<0{,}564$. **14.** $0{,}19<p<0{,}31$.

7

1. а) нет; б) да. **2.** $\bar{x} = 0{,}962$; $S_{\bar{x}} = 0{,}011$; **нет.**

3. $\bar{x} = 10{,}5$ дней; $S_{\bar{x}} = 0{,}643$; $t_{пр} = 2{,}33$; **нет.**

4. $\bar{x} = 3{,}75$; $S_x^2 = 6{,}21$; $\bar{y} = 5{,}4$; $S_y^2 = 4{,}71$; $S_{\bar{x}-\bar{y}} = 1{,}1$; $t_{пр} = 1{,}5$; **да.**

5. $\bar{x} = 6{,}1$; $S_x^2 = 0{,}723$; $\bar{y} = 7{,}7$; $S_y^2 = 0{,}291$; $t_{пр} = 4{,}49$; H_0 : −; $F_{пр} = 2{,}48$; H_0 : +.

6. $\bar{x} = 25$; $S_x^2 = 28{,}8$; $\bar{y} = 19$; $S_y^2 = 4{,}8$; H_0 : −; $F_{пр} = 6$; H_0 : +; $t_{пр} = 2{,}54$; H_0 : +.

7. $\bar{x} = 22$; $S_x^2 = 12{,}4$; $\bar{y} = 18$; $S_y^2 = 16{,}4$; H_0 : +.

8. $\bar{x} = 3{,}2$; $S_x^2 = 0{,}046$; $\bar{y} = 2{,}6$; $S_y^2 = 0{,}02$; H_0 : −; а) $F_{пр} = 2{,}48$; H_0 : +; б) $t_{пр} = 5{,}9$; H_0 : −.

9. $F_{пр} = 5{,}47$; **да.** **10.** Да. **11.** $t_{пр}=2{,}74$; да. **12.** $t_{пр}=6{,}7$; нет. **13.** $t_{пр}=4{,}37$; да. **14.** $t_{пр}=3{,}5$; да. **15.** H_0:+. **16.** H_0:+. **17.** $\chi^2_{пр}=13{,}22$; $\chi^2_{кр}(0{,}05;\ 4)=9{,}5$; H_0 отвергается. **18.** Нет. **19.** Нет. **20.** а) Нет; б) Да. **21.** $W_{н.кр}(0{,}025;\ 6;\ 6) = 26$; $W_{в.кр}=52$; $W_{набл}=70$; H_0 отвергается. **22.** $W_{н.кр}(0{,}05;\ 30;\ 50) = 1048$; $W_{в.кр}=1382$; H_0 не отвергается.

8

1. $\rho_{xy} \approx 0{,}934$; $\bar{y}(x) = -1{,}795 + 0{,}477x$; $\bar{x}(y) = 4{,}435 + 1{,}826y$.

2. $\rho_{xy} \approx 0{,}691$; $\bar{y}(x) = 1{,}224 + 0{,}434x$; $\bar{x}(y) = 4{,}4 + 1{,}1y$.

3. $\rho_{xy} \approx 0{,}961;\ \bar{y}(x) = 1{,}459 + 0{,}629x;\ \bar{x}(y) = -1{,}773 + 1{,}468y.$
4. $\rho_{xy} \approx 0{,}743;\ \bar{y}(x) = 12{,}245 + 0{,}795x;\ \bar{x}(y) = 28{,}791 + 0{,}694y.$
5. $\rho_{xy} \approx 0{,}912;\ \bar{y}(x) = 31{,}28 + 0{,}0897x;\ \bar{y}(75) = 38.$
6. $\rho_{xy} \approx -0{,}801;\ \bar{y}(x) = 21{,}91 - 0{,}181x.$
7. $\rho_{xy} = -0{,}596.$
8. $\rho_{xy} \approx 0{,}837;\ \bar{y}(x) = 8{,}96 + 0{,}44x;\ \bar{x}(y) = 10{,}96 + 1{,}6y.$
9. Да; нет; n≥13.
10. а) $\eta_{y/x} \approx 0{,}964;\ \rho_{xy} = -0{,}725;$ б) $\eta_{y/x} \approx 0{,}779;\ \rho_{xy} = 0{,}00727.$
11. а) $\eta_{y/x} \approx 0{,}981;\ \eta_{x/y} \approx 0{,}967;\ \rho_{xy} \approx 0{,}952;$ б) $\eta_{y/x} \approx 0{,}920;\ \eta_{x/y} \approx 0{,}902;\ \rho_{xy} \approx 0{,}892.$

9

1. $F_{пр}$=10; H_0: — . **2.** H_0: -. **3.** $F_{пр}$=2,4; H_0: +. **4.** $F_{пр}$=13,7; H_0: -. **5.** $F_{пр}$=3,85; H_0: +. **6.** а) $F_{Aпр}$=31; H_{A0}:+; $F_{Bпр}$=8; H_{B0}:-; б) $F_{Aпр}$=18,5; H_{A0}:-; $F_{Bпр}$=2,4; H_{B0}:+. **7.** а) $F_{Aпр}$=3,36; H_{A0}:+; б) $F_{Bпр}$=8,41; H_{B0}:-. **8.** а) $F_{Aпр}$=13,91; H_{A0}:-; $F_{Bпр}$=4,50; H_{B0}:-; б) $F_{Aпр}$=22,5; H_{A0}:-; $F_{Bпр}$=2,40; H_{B0}:+.

10

1. 7,83+0,0642·t. **2.** 281-34,0·t+9,1·t^2; 196+29,7·t. **3.** 50,7+0,83·t. **4.** 164,9+1,81·t. **5.** 24,2+0,698·t. **6.** в) 147+0,0970·t. **7.** а) 366+110·t; б) 496·$e^{0,111t}$; в) 458·$t^{0,469}$; г) 1300,5-997/t; д) $(0{,}001816-0{,}000119\cdot t)^{-1}$; е) 315,6+447·lnt.

ЛИТЕРАТУРА

1. Андрухаев, Х.М. Сборник задач по теории вероятностей / под ред. А.С. Солодовникова. — М. : Просвещение, 1985. — 160 с.
2. Баврин, И.И. Высшая математика. — М. : Просвещение, 1980. — 383 с.
3. Батунер, Л.М. Математические методы в химической технике / Л.М. Батунер, М.Е. Позин. — Л. : Химия, Ленингр. отделение, 1971. — 823 с.
4. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике. — М. : Высшая школа, 1999. — 400 с.
5. Гмурман, В.Е. Теория вероятностей и математическая статистика. — М. : Высшая школа, 1999. — 479.
6. Гроссман, С. Математика для биологов / С. Гроссман, Дж. Тернер. — М. : Высшая школа, 1983. — 383 с.
7. Дунаев, А.А. Высшая математика / А.А. Дунаев, В.А. Тупицын. — Ч. II : Основы математической статистики : метод. указания. — Рязань : РГМУ, 1994.
8. Дунаев, А.А. Высшая математика / А.А. Дунаев, В.А. Тупицын. — Ч. I : Дифференциальные уравнения. Основы теории вероятностей : метод. указания. — Рязань : РГМУ, 1995.
9. Мелник, М. Основы прикладной статистики. — М. : Энергоатомиздат, 1983. — 414 с.
10. Основы математической статистики / под ред. В.С. Иванова. — М. : Физкультура и спорт, 1990. — 176 с.
11. Смирнов, Н.В. Курс теории вероятностей и математической статистики / Н.В. Смирнов, И.В. Дунин-Барковский. — М. : Наука, 1969. — 511 с.
12. Солодовников, А.С. Теория вероятностей. — М. : Просвещение, 1983. — 207 с.

ОГЛАВЛЕНИЕ

Для заметок

Учебное издание

Дунаев Александр Анатольевич

Основы
статистических методов
компьютерной обработки
результатов наблюдений

Учебное пособие

Редактор *В.Л. Рубайлова*
Технический редактор *О.С. Верещагина*

Подписано в печать 8.11.07. Поз. № 056. Бумага офсетная. Формат 60x84$^{1}/_{16}$.
Гарнитура Times New Roman. Печать трафаретная.
Усл. печ. л. 10,46. Уч.-изд. л. 10,2. Тираж 150 экз. Заказ №

Государственное образовательное учреждение высшего профессионального образования
«Рязанский государственный университет имени С.А. Есенина»
390000, г. Рязань, ул. Свободы, 46

Редакционно-издательский центр РГУ
390023, г. Рязань, ул. Урицкого, 22