МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Белорусский национальный технический университет

Кафедра «Информационные технологии в управлении»

ИЗБРАННЫЕ ГЛАВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Пособие

Минск
БНТУ
2021

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Белорусский национальный технический университет

Кафедра «Информационные технологии в управлении»

ИЗБРАННЫЕ ГЛАВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Пособие
для студентов специальностей
1-53 01 02 «Автоматизированные системы обработки информации», 1-40 01 01 «Программное обеспечение информационных технологий», 1-25 01 07 «Экономика и управление на предприятии», 1-26 02 02 «Менеджмент»

Рекомендовано учебно-методическим объединением высших учебных заведений Республики Беларусь по образованию в области автоматизации технологических процессов, производств и управления

Минск
БНТУ
2021

УДК 519.2 (075.8)
ББК 22.17
И32

С о с т а в и т е л и:
В. Ф. Голиков, *В. А. Казакевич*

Р е ц е н з е н т ы:
кафедра высшей математики Военной академии Республики Беларусь,
зав. кафедрой, д-р техн. наук, профессор *В. А. Липницкий*;
доцент кафедры информационных технологий БГУ,
канд. техн. наук *А. В. Овсянников*

И32 **Избранные** главы теории вероятностей и математической статистики : пособие для студентов специальностей 1-53 01 02 «Автоматизированные системы обработки информации», 1-40 01 01 «Программное обеспечение информационных технологий», 1-25 01 07 «Экономика и управление на предприятии», 1-26 02 02 «Менеджмент» / сост.: В. Ф. Голиков, В. А. Казакевич. – Минск : БНТУ, 2021. – 115 с.
ISBN 978-985-583-218-9.

Пособие предназначено для студентов высших учебных заведений, обучающихся по специальностям технического и экономического профиля и изучающих теорию вероятностей и математическую статистику как разделы высшей математики. Цель пособия – ознакомить студентов с основными понятиями, расчетными формулами, областью применения теории вероятностей и математической статистики. Наиболее сложные понятия и методика расчетов иллюстрируются примерами.

УДК 519.2 (075.8)
ББК 22.17

ISBN 978-985-583-218-9

ОГЛАВЛЕНИЕ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

1. СЛУЧАЙНЫЕ СОБЫТИЯ

1.1. Введение

Теория вероятностей (ТВ) изучает закономерности, которым подчиняются случайные явления в различных областях жизни.

ТВ развивалась из потребностей практики. Еще в начале 17 века Галилей, изучая ошибки, совершаемые при физических измерениях, столкнулся с необходимостью математического описания случайных явлений. Чуть позже ученые пытались создать математический аппарат для изучения таких массовых и в то же время случайных явлений, как заболеваемость, смертность, несчастные случаи и т. д. Однако эти явления были на то время для науки чересчур сложны, поэтому особых успехов в создании новой теории достигнуто не было.

Позднее ТВ как наука создавалась для анализа азартных игр (кости, карты, рулетка и т. д.). Слово «азарт» в переводе с французского означает случай. В азартных играх присутствует элемент случайности. Так говорят «везет», если случай помогает выиграть, и «не везет», если случай ведет к проигрышу. Для анализа азартных игр использовалась «схема урн», которая достаточно просто и наглядно позволяла делать некоторые предсказания.

Если вспомнить историю развития ТВ, то она создавалась трудами следующих известных ученых:

– в 17 веке – Паскаль, Ферма, Гюйгенс;

– в 18 веке – Бернулли, Лаплас, Гаусс, Пуассон;

– в 19 веке – Буняковский, Чебышев, Марков, Ляпунов;

– в 20 веке – Колмогоров, Хинчин, Феллер, Крамер, Фишер.

Методы ТВ находят широкое применение во многих областях жизни: в технике, экономике, медицине, социологии и т. д.

1.2. Событие. Вероятность случайного события

Каждая теория оперирует основными понятиями. В ТВ это случайное событие, вероятность случайного события.

Случайное событие – это событие, которое в результате опыта может произойти или нет. Например, появление герба при броске

монеты, отказ телевизора, получение отличной оценки на экзамене. События в ТВ принято обозначать прописными буквами латинского алфавита.

Рассматривая эти события, видим, что каждое из них обладает разной степенью возможности. Например, выпадение герба один раз более возможно, чем три раза подряд при трехкратном бросании монеты. Чтобы количественно сравнивать события по степени возможности, надо с каждым числом связать определенную величину, которая тем больше, чем больше возможность наступления события. Такую величину назвали вероятностью события.

Вероятность события – численная мера степени объективной возможности этого события.

Все события можно классифицировать по различным признакам.

По степени возможности события делятся на достоверные, случайные и невозможные. Событие называется достоверным, если в результате опыта оно обязательно произойдет. Событие называется невозможным, если в результате опыта оно обязательно не произойдет.

По взаимодействию между собой события делятся на совместные и несовместные. События называются несовместными, если в результате опыта появление одного события исключает появление другого. Например, получение положительной оценки на экзамене исключает получение неудовлетворительной оценки. События называются совместными, если в результате опыта появление одного события не исключает появление другого. Например, выпадение четной цифры при броске игральной кости, не исключает выпадения цифры 4.

Несколько событий образуют полную группу событий, если в результате опыта обязательно происходит одно из событий этой группы. Например, выпадение цифр 1, 2, 3, 4, 5, 6 при однократном бросании игральной кости образуют полную группу событий. Если полная группа событий состоит из двух событий, то такие события называются противоположными. Например, выпадение герба или решки при бросании монеты образуют полную группу событий.

1.3. Определение вероятности

Существует несколько определений понятия вероятность случайного события.

1.3.1. Классическое определение

Вероятностью случайного события называется отношение числа благоприятствующих этому событию исходов опыта к общему числу всех равновозможных несовместных исходов, образующих полную группу. Вероятность события A равна

$$P(A) = \frac{m}{n}, \tag{1.1}$$

где m – число благоприятствующих событию A исходов опыта;

n – общее число исходов опыта.

Например, в урне имеется 6 шаров: 2 красных, 3 синих, 1 белый. Какова вероятность вынуть красный шар?

Обозначим:

A – событие, заключающееся в том, что вынутый шар окажется красным;

A_1 – событие, заключающееся в том, что вынутый шар окажется первым красным;

A_2 – событие, заключающееся в том, что вынутый шар окажется вторым красным;

A_3 – событие, заключающееся в том, что вынутый шар окажется первым синим;

A_4, A_5 – событие, заключающееся в том, что вынутый шар окажется вторым, третьим синим;

A_6 – событие, заключающееся в том, что вынутый шар окажется белым.

Общее число возможных исходов (событий) равно $n = 6$. Число исходов, благоприятствующих извлечению красного шара $m = 2$ (события A_1, A_2). Итак,

$$P(A) = \frac{m}{n} = \frac{2}{6} = \frac{1}{3}.$$

Т. к. в выражении (1.1) всегда $m \le n$, то вероятность случайного события может изменяться от 0 до 1, т. е. $0 \le P(A) \le 1$. Вероятность

невозможного события равна 0, т. к. $m = 0$. Классическое определение вероятности может быть использовано при конечном числе исходов опыта. Причем все исходы опытов должны быть равновероятными. Такая модель случайных событий получила в ТВ название схема урн.

1.3.2. Статистическое определение вероятности

Если проводится n опытов в которых некоторое событие A наступило m раз, то для определения вероятности события A следует воспользоваться статистической формулой вероятности:

$$P(A) = \frac{m}{n}. \tag{1.2}$$

Например, требуется определить вероятность того, что наугад выбранная пара обуви из достаточно большой партии окажется бракованной. Выберем случайным образом из данной партии n пар и проверим их. Пусть среди них окажется m пар брака. Тогда за искомую вероятность можно выбрать частоту появления бракованной обуви (1.2).

Несмотря на внешнее сходство выражений (1.1) и (1.2) они по внутреннему содержанию существенно отличны. Классическое определение вероятности может быть применено для нахождения вероятности без проведения опыта. Достаточно путем теоретического анализа вычислить число благоприятствующих исходов и общее число исходов. Статистическое определение вероятности позволяет находить приближенное значение вероятности по результатам проведенного опыта (отсюда название статистическое). Как будет показано далее, чем больше объем опыта, тем меньше ошибка в определении вероятности.

1.3.3. Геометрическая интерпретация понятия вероятность

С геометрической точки зрения вероятность – это возможность попадания случайно брошенной точки в некоторую область (отрезок, фигуру, объем). Например, пусть A – событие, заключающееся

в попадании случайно брошенной точки на отрезок l, находящийся внутри отрезка L (рис. 1.1). Естественно, что

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

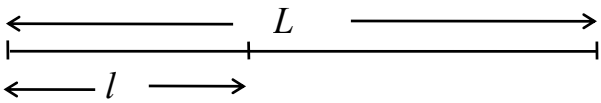


Рис. 1.1. Геометрическая интерпретация понятия вероятность (отношение отрезков)

Если речь идет о вероятности попадания в некоторую область, то $P(A) = \frac{s}{S}$, где s – площадь данной области, S – площадь всей фигуры (рис. 1.2).

Рис. 1.2. Геометрическая интерпретация понятия вероятность (отношение площадей)

1.4. Элементы комбинаторного анализа

Для расчета классической вероятности используются формулы комбинаторного анализа (комбинаторики). Комбинаторика позволяет определить количество комбинаций, составленных из конечного числа элементов, обладающих определенными свойствами. Чаще всего в ТВ используются такие комбинации, как перестановки и сочетания.

Перестановки – это комбинации элементов из n имеющихся, отличающихся только порядком расположения. Число возможных перестановок P_n равно

$$P_n = 1 \cdot 2 \cdot 3 \ldots n = n!. \qquad (1.3)$$

Пусть $n=3$: a_1, a_2, a_3. Найдем P_3. Из трех элементов можно составить следующие комбинации: a_1, a_2, a_3; a_1, a_3, a_2; a_2, a_1, a_3; a_2, a_3, a_1; a_3, a_1, a_2; a_3, a_2, a_1. Всего получилось 6 комбинаций. Если воспользоваться формулой (1.3), то получим $P_3 = 1 \cdot 2 \cdot 3 = 6$.

Сочетания – это комбинации элементов из n имеющихся по m выбранным, отличающихся хотя бы одним элементом. Из n элементов по m можно составить C_n^m сочетаний. Эта величина равна

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (1.4)$$

Пусть $n=4$: a_1, a_2, a_3, a_4; $m=2$. Найдем число возможных сочетаний C_4^2. Из 4 элементов по 2 можно составить следующие комбинации: a_1, a_2; a_1, a_3; a_1, a_4; a_2, a_3; a_2,a_4; a_3,a_4. Всего получилось 6 сочетаний. Если воспользоваться формулой (1.4), то получим

$$C_4^2 = \frac{4!}{2!(4-2)!} = 6.$$

2. ОСНОВНЫЕ ТЕОРЕМЫ СЛУЧАЙНЫХ СОБЫТИЙ

2.1. Теорема сложения вероятностей случайных событий

Суммой двух событий A и B называется событие C, состоящее в появлении события A или события B, или обоих этих событий. Условно это записывается как: $C = A + B$. Например, A – извлечение из колоды карт карты бубновой масти, B – извлечение из колоды карт карты червовой масти, C – извлечение из колоды карт карты красной масти.

Произведением двух событий A и B называется событие C, состоящее в появлении события A и события B. Условно это записывается как: $C = A \cdot B$. Например, A – попадание в мишень при первом выстреле, B – попадание в мишень при втором выстреле, C – двукратное попадание в мишень после двух выстрелов.

Теорема для несовместных событий

Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий. Т. е., если $C = A + B$, то

$$P(C) = P(A) + P(B). \tag{2.1}$$

Доказательство. Пусть n – общее число исходов опыта; m_1 – число исходов благоприятствующих событию A; m_2 – число исходов благоприятствующих событию B;

$m_1 + m_2$ – число исходов благоприятствующих событию $A + B$. Тогда

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

Теорема для совместных событий

Вероятность появления хотя бы одного из двух совместных событий равна

$$P(C) = P(A) + P(B) - P(AB). \tag{2.2}$$

Доказательство. Пусть n – общее число исходов опыта; m_1 – число исходов благоприятствующих событию A; m_2 – число исходов благоприятствующих событию B; k – число исходов благоприятствующих событию $A\,B$. Тогда

$$P(C) = \frac{m_1 + m_2 - k}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{k}{n} = P(A) + P(B) - P(AB).$$

Следствия

1. Если n несовместных событий A_i образуют полную группу событий, то вероятность появления одного из этих событий, безразлично какого, равна 1. Т. е.

$$\sum_{i=1}^{n} P(A_i) = 1.$$

2. Сумма вероятностей противоположных событий A и $\overline{A}$ равна 1, т. е.

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

2.2. Теорема умножения вероятностей

2.2.1. Зависимые и независимые события

Событие A называется независимым от события B, если $P(A)$ не зависит от того произошло событие B или нет.

Например, пусть A – выпадение герба в первом опыте; B – выпадение герба во втором опыте. События A и B независимы. Пусть в урне находится 3 шара: 2 – черных и 1 белый. Обозначим извлечение черного шара в первом опыте как событие A, извлечение черного шара во втором – B. События A и B зависимы, т. к.

$$P(A) = \frac{2}{3}, \quad P(B) = \begin{cases} \frac{1}{2}, & \text{если произошло } A, \\ 1, & \text{если произошло } \overline{A}. \end{cases}$$

Вероятность события B, вычисленная при условии, что событие A произошло, называется условной вероятностью события B.

Условная вероятность обозначается $P(B / A)$ или $P_A(B)$. Очевидно, что для независимых событий $P(B / A) = P(B)$ или $P(A / B) = P(A)$.

2.2.2. Теорема

Вероятность произведения двух событий равна произведению вероятности одного события на условную вероятность другого, т. е.

$$P(AB) = P(A)P(B / A) = P(B)P(A / B). \qquad (2.3)$$

Доказательство. Обозначим: n – общее число исходов опыта; m_1 – число исходов благоприятствующих событию A; m_2 – число исходов благоприятствующих событию B; k – число исходов благоприятствующих событию AB (рис. 2.1).

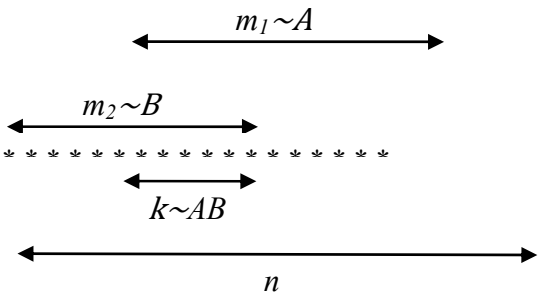


Рис. 2.1. Умножение вероятностей

Из рисунка видно, что

$$P(AB) = \frac{k}{n}, \quad P(A) = \frac{m_1}{n}, \quad P(B) = \frac{m_2}{n}, \quad P(B / A) = \frac{k}{m_1}.$$

Можно представить, что $\frac{k}{n} = \frac{m_1}{n}\frac{k}{m_1}$, тогда $P(AB) = P(A)P(B / A)$.

Следствие 1. Если A не зависит от B, то и B не зависит от A.

Следствие 2. Если A – произведение n независимых событий A_i, то

$$P(A_1 A_2 A_3 ... A_n) = P(A_1)P(A_2)P(A_3)...P(A_n).$$

2.3. Формула полной вероятности

Пусть требуется определить вероятность некоторого события A, которое может произойти вместе с одним из событий: $H_1, H_2, H_3, ..., H_n$, образующих полную группу несовместных событий. События, совместно с которыми может произойти интересующее нас событие, называются гипотезами. Вероятности гипотез, как правило, известны или могут быть определены до опыта.

Поскольку событие A может наступить с любой гипотезой H_i, то оно является суммой $A = AH_1 + AH_2 + AH_3 + ... + AH_n$. В соответствии с теоремой сложения вероятностей имеем

$$P(A) = P(AH_1) + P(AH_2) + P(AH_3) + ... + P(AH_n).$$

Т. к. события A и H_i зависимы, то $P(AH_i) = P(H_i)P(A / H_i)$. Окончательно имеем

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A / H_i). \qquad (2.4)$$

Формула (2.4) позволяет определять вероятность некоторого события с учетом знания доопытных (априорных) значений вероятностей гипотез.

Пример. Имеется две урны. В первой находятся 2 белых и 1 черный шар; во второй – 1 белый и 3 черных. Найти вероятность того, что вынутый наугад из одной из урн шар окажется белым.

Решение. Обозначим через A извлечение белого шара из наугад выбранной урны, через H_1 гипотезу – выбор первой урны, через H_2 гипотезу – выбор второй урны. Вероятность выбора гипотез $P(H_1) = P(H_2) = 0,5$. Условные вероятности равны

$$P(A/H_1)=\frac{2}{3},\ P(A/H_2)=\frac{1}{4},$$

$$P(A)=P(H_1)P(A/H_1)+P(H_2)P(A/H_2)=\frac{1}{2}\frac{2}{3}+\frac{1}{2}\frac{1}{4}=\frac{11}{24}.$$

2.4. Теорема гипотез (формула Байеса)

Пусть имеется группа несовместных гипотез $H_1, H_2, H_3,..., H_n$. Вероятности этих гипотез равны $P(H_1), P(H_2), P(H_3),..., P(H_n)$. Событие A может произойти совместно с одной из гипотез H_i. Произведен опыт, в результате которого наступило событие A совместно с гипотезой H_i. Найдем условную вероятность этой гипотезы.

В соответствии с теоремой умножения можно записать

$$P(AH_i)=P(A)P(H_i/A)=P(H_i)P(A/H_i).$$

Из последнего выражения получим

$$P(H_i/A)=\frac{P(H_i)P(A/H_i)}{P(A)}.$$

Или с учетом $P(A)=\sum_{i=1}^{n}P(H_i)P(A/H_i)$ окончательно имеем

$$P(H_i/A)=\frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n}P(H_i)P(A/H_i)}. \qquad (2.5)$$

Формула (2.5) называется формулой Байеса, она позволяет находить послеопытное (апостериорное) значение вероятности гипотез.

2.5. Частная теорема о повторении опытов (формула Бернулли)

Если производится n испытаний, в каждом из которых может наступить или не наступить событие A, а вероятность события A в каждом испытании одинакова и равна $P(A) = p$, результаты испытаний независимы, то часто возникает задача определения вероятности наступления заданного числа события A.

Пусть $n = 3$, событие A в серии из трех испытаний может наступить k раз (0, 1, 2, 3). Найдем вероятность того, что $k = 1$. Это событие заключается в появлении события A или в 1-ом или во 2-ом или в 3-ем испытании, т. е. $A\overline{A}\,\overline{A} + \overline{A}\,A\,\overline{A} + \overline{A}\,\overline{A}\,A$, где $\overline{A}$-событие, противоположное событию A. Вероятность того, что $k = 1$, равна

$$P(k=1) = P(A\overline{A}\,\overline{A} + \overline{A}\,A\,\overline{A} + \overline{A}\,\overline{A}\,A) =$$
$$= p(1-p)^2 + p(1-p)^2 + p(1-p)^2 = 3p(1-p)^2.$$

Аналогично

$$P(k=2) = P(AA\overline{A} + A\overline{A}A + \overline{A}AA) =$$
$$= p^2(1-p) + p^2(1-p) + p^2(1-p) = 3p^2(1-p),$$

$$P(k=3) = P(AAA) = p^3,$$

$$P(k=0) = P(\overline{A}\,\overline{A}\,\overline{A}) = (1-p)^3.$$

В общем случае вероятность того, что в n опытах событие A произойдет k раз, равна

$$P_n(k) = C_n^k\, p^k\,(1-p)^{n-k} = \frac{n!}{k!(n-k)!}\,p^k\,(1-p)^{n-k}.$$

3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

3.1. Понятие о случайной величине

Случайной величиной называется величина, которая в результате опыта может принять то или иное значение, причем неизвестно заранее, какое именно.

Например, число попаданий в мишень при трех выстрелах (0, 1, 2, 3); вес человека, случайно выбранного из толпы (от 40 кг до 120 кг); количество солнечных дней в году (0, 1, 2, ..., 365). Случайные величины, принимающие только отделенные друг от друга значения, которые можно заранее перечислить, называются прерывными или дискретными величинами (примеры 1 и 3). Случайные величины, значения которых непрерывно заполняют некоторый промежуток, называются непрерывными (пример 2).

Понятие случайная величина более богато по содержанию, чем случайное событие, и позволяет производить более сложные расчеты при изучении закономерностей случайных явлений. Кстати, всегда имеется возможность от случайного события перейти к случайной величине. Например, результат опыта, в котором событие A может наступить или нет, можно рассматривать как случайную величину, которая принимает два значения: 1 (когда событие A наступило) и 0 (когда событие A не наступило).

Принято случайные величины обозначать большой буквой, а ее возможные значения малыми буквами. Например, X – число попаданий в мишень при 3-х выстрелах; $x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ x_4 = 3.$

3.2. Закон распределения случайной величины

Пусть имеется случайная величина X, которая в результате опыта может принять одно из возможных значений: $x_1, x_2, ..., x_n$ с вероятностями: $P(X = x_1) = p_1,\ P(X = x_2) = p_2, ..., P(X = x_n) = p_n.$ Так как события $X = x_1, X = x_2, ..., X = x_n$ несовместны, то они образуют полную группу событий, и для них справедливо $\sum_{i=1}^{n} P(X = x_i) = 1,$ т. е. сумма вероятностей всех возможных значений случайной вели-

чины равна 1. Суммарная вероятность распределяется между отдельными значениями случайной величины. Это распределение вероятностей по значениям случайной величины называется законом распределения вероятностей случайной величины или для сокращения записи просто законом распределения. Таким образом, закон распределения случайной величины это соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Существуют различные формы задания закона распределения случайной величины.

3.2.1. Ряд распределения случайной величины

Ряд распределения – это таблица, содержащая возможные значения случайной величины и их вероятности (табл. 3.1). Применяется для дискретных случайных величин.

Таблица 3.1.

Ряд распределения случайной величины

Значения случайной величины x_i	x_1	x_2	x_3	. . .	x_n
Вероятности значений p_i	p_1	p_2	p_3	. . .	p_n

3.2.2. Функция распределения случайной величины

Интегральная Функция распределения случайной величины

Интегральной функцией распределения случайной величины X называется функция вида $F(x) = P(X < x)$, где x – некоторое зафиксированное значение случайной величины.

Эта форма закона распределения удобна как для дискретной, так и для непрерывной случайной величины. Интегральная функция распределения обладает рядом свойств.

1. $F(x)$ – неубывающая функция своего аргумента, т. е. если $x_2 > x_1$, то $F(x_2) \geq F(x_1)$.

Доказательство. Так как $F(x_1) = P(X < x_1)$, а $F(x_2) = P(X < x_2)$, но $P(X < x_2) > P(X < x_1)$, в силу того, что $x_2 > x_1$, значит $F(x_2) \geq F(x_1)$.

2. $F(-\infty) = 0.$

Доказательство. Так как $F(-\infty) = P(X < -\infty),$ а событие $X < -\infty$ невозможно, то $F(-\infty) = 0.$

3. $F(+\infty) = 1.$

Доказательство. Так как $F(+\infty) = P(X < +\infty),$ а событие $X < +\infty$ достоверно, то $F(+\infty) = 1.$

Таким образом, интегральная функция распределения представляет собой неубывающую функцию, изменяющуюся от 0 до 1, при изменении аргумента от $-\infty$ до $+\infty$ (рис. 3.1), причем в отдельных точках она может иметь скачки (разрывы). Для дискретной случайной величины справедливо

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i),$$

где $x_i < x$ под знаком суммы указывает, что суммирование распространяется на все те значения x_i, которые меньше *x*. Причем, когда текущая переменная проходит через какое-нибудь из возможных значений дискретных значений *X*, функция распределения меняется скачкообразно на величину вероятности этого значения.

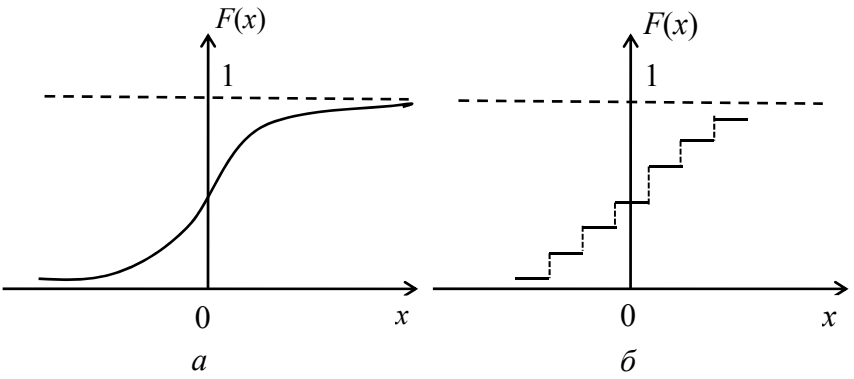


Рис. 3.1. Интегральная функция распределения:
а – для непрерывной случайной величины;
б – для дискретной случайной величины

Дифференциальная функция распределения (плотность распределения, плотность вероятности)

Для непрерывных случайных величин в качестве функции распределения наиболее часто используется дифференциальная функция распределения. Пусть имеется непрерывная случайная величина X с функцией распределения $F(x)$, непрерывной и дифференцируемой для всех значений x. Найдем вероятность попадания случайной величины X в интервал от x до $x+\delta x$, (обозначим это событие через A). Для этого рассмотрим еще два события: $X < x$ (событие B) и $X < x+\delta x$ (событие C). Из рис. 3.2 видно, что $C = A + B$.

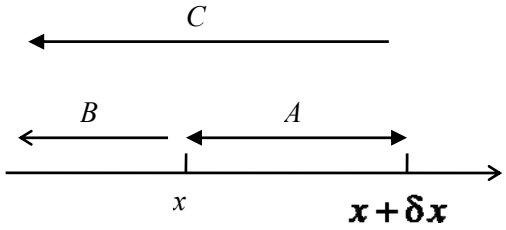


Рис. 3.2. Вычисление дифференциальной функции распределения

Согласно теореме о вероятности суммы несовместных событий, имеем $P(C) = P(A) + P(B)$ или $P(A) = P(C) - P(B)$. Т. к. $P(A) = P(x < X < x+\delta x)$, $P(B) = P(X < x) = F(x)$, $P(C) = P(X < x+\delta x) = F(x+\delta x)$, то искомая вероятность равна $P(x < X < x+\delta x) = F(x+\delta x) - F(x)$.

Последнее выражение имеет смысл приращения вероятности на интервале δx. Усредним это приращение, поделив его на δx. Если величину интервала взять очень малой, то можно считать, что полученный предел имеет смысл плотности вероятности в окрестности точки x, с математической точки зрения это производная от интегральной функции распределения в точке x

$$\lim_{\delta x \to 0} \frac{F(x+\delta x) - F(x)}{\delta x} = F'(x) = f(x).$$

Таким образом, дифференциальная функция распределения (плотность вероятности) есть производная от интегральной функции распределения

$$f(x) = F'(x). \tag{3.1}$$

Интегрируя (3.1) по t, получим

$$F(x) = \int_{-\infty}^{x} f(t)\,dt. \tag{3.2}$$

Дифференциальная функция распределения обладает следующими свойствами.

1. Дифференциальная функция распределения $f(x)$ – неотрицательная функция, т. е. $f(x) \geq 0$.

Доказательство. Т. к. $F(x)$ – неубывающая функция, а по определению (3.1) $f(x)$ есть производная от $F(x)$, то производная от неубывающей функции неотрицательна.

2. Площадь под кривой $f(x)$ и осью x равна 1, т. е.

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \tag{3.3}$$

Доказательство. В соответствии с (3.2) имеем, что если $x \to \infty$, то

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

Следует отметить, что интегральная функция распределения – величина безразмерная, а дифференциальная имеет размерность, обратную случайной величине. Типичный вид дифференциальной функции распределения изображен на рис. 3.3.

Зная $F(x)$, можно определить вероятность попадания случайной величины в некоторый интервал:

$$P(a < X < b) = F(b) - F(a). \tag{3.4}$$

Ранее было показано, что $P(x < X < x + \delta x) = F(x + \delta x) - F(x)$. Заменив x на a и $x + \delta x$ на $в$, т. е. $x = a$, $x + \delta x = в$, получим (3.4).

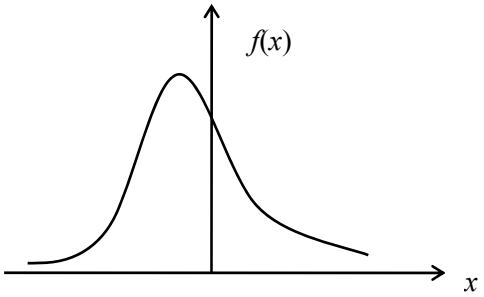


Рис. 3.3. Дифференциальная функция распределения

Зная $f(x)$, получим

$$P(a < X < b) = \int_a^b f(x)dx. \qquad (3.5)$$

Действительно,

$$P(a < X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_a^b f(x)dx.$$

Пример 3.1. В мишень производится два выстрела. Вероятность поадания при каждом выстреле равна p = 0,8. Рассматривается случайная величина – число попаданий в мишень при двух выстрелах. Необходимо построить ряд распределения и интегральную функцию распределения.

Решение. Обозначим случайную величину через X, попадание в мишень при первом выстреле – событие A_1, при втором – A_2. Возможные значения случайной величины равны:

а) $x_1 = 0$, если имели место два промаха, т. е. $\overline{A}_1\,\overline{A}_2$. Вероятность этого $P(X = x_1) = P(\overline{A}_1\,\overline{A}_2) = (1 - 0{,}8)(1 - 0{,}8) = 0{,}04$;

б) $x_2 = 1$, если имели место одно попадание и один промах, т. е. $A_1 \overline{A}_2 + \overline{A}_1 A_2$. Вероятность этого $P\left(X = x_2\right) = P(A_1 \overline{A}_2 + \overline{A}_1 A_2) =$ $= 0,8 \cdot 0,2 + 0,2 \cdot 0,8 = 0,32$;

в) $x_3 = 2$, если имели место два попадания, т. е. $A_1 A_2$. Вероятность этого $P\left(X = x_3\right) = P\left(A_1 A_2\right) = 0,8 \cdot 0,8 = 0,64$.

Зная значения случайной величины и их вероятности, составляем ряд распределения (табл. 3.2)

Значения интегральной функции распределения рассчитываем по формуле

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i)$$

и заносим их в табл. 3.2.

Таблица 3.2

Значения случайной величины x_i	0	1	2
Вероятности значений p_i	0,04	0,32	0,64
$F(x)$	0	0,04	0,36

Функция $F(x)$ изображена на рис. 3.4

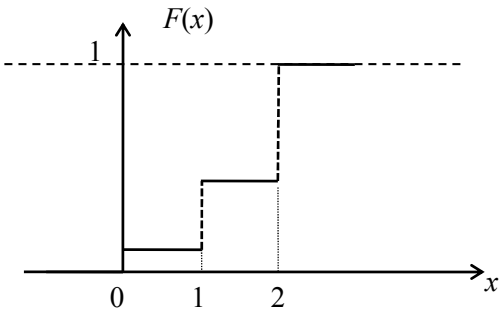


Рис. 3.4. Интегральная функция распределения

Пример 3.2. Случайная величина X имеет интегральную функцию распределения вида

$$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\le x\le 1\\ 1, & x>1.\end{cases}$$

Определить дифференциальную функцию распределения и вероятность попадания случайной величины X в интервал от 0,25 до 0,5.

В соответствии с (3.1) имеем

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

В соответствии с (3.4) имеем

$$P(0,25<X<0,5)=F(0,5)-F(0,25)=0,5^2-0,25^2=0,1875.$$

4. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

4.1. Общие сведения о числовых характеристиках случайных величин

Такие характеристики, как закон распределения в различных формах (ряд распределения, интегральная или дифференциальная функция распределения), полностью описывают поведение случайной величины с вероятностной точки зрения. Однако во многих практических задачах нет необходимости характеризовать случайную величину исчерпывающим образом. Достаточно знать некоторые параметры случайной величины. Например, среднее значение (значение, около которого группируются возможные значения случайной величины), рассеяние случайной величины относительно среднего значения и т. д. Использование таких характеристик бывает оправдано и тем, что часто результат расчета слабо зависит от вида закона распределения, поэтому и нет необходимости оперировать с законом распределения.

Таким образом, неслучайные характеристики, которые в числовой форме выражают наиболее существенные особенности случайной величины и ее закона распределения, называются числовыми характеристиками случайной величины. Наиболее употребляемыми числовыми характеристиками являются математическое ожидание, дисперсия, среднее квадратическое отклонение случайной величины.

4.2. Характеристики положения

Числовые характеристики положения случайной величины – это такие числовые характеристики, которые дают информацию о положении на числовой оси усредненного некоторым образом значения случайной величины.

Различают следующие характеристики положения: математическое ожидание, мода, медиана.

4.2.1. Математическое ожидание случайной величины (среднее значение)

Пусть X дискретная случайная величина с возможными значениями $x_1, x_2, ..., x_n$ и вероятностями этих значений $p_1, p_2, ..., p_n$. Математическим ожиданием случайной величины называется сумма произведений всех возможных значений случайной величины на вероятности этих значений

$$M[X] = m_x = \sum_{i=1}^{n} x_i p_i. \qquad (4.1)$$

Как видно из (4.1), математическое ожидание есть среднее взвешенное значение случайной величины. Причем весовыми коэффициентами являются вероятности каждого значения. Действительно, (4.1) можно представить в следующем виде:

$$M[X] = \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i \frac{p_i}{p_1 + p_2 + ... p_n}.$$

Таким образом, при расчете математического ожидания происходит усреднение не только по величинам возможных значений, но и по их вероятностям.

Пусть X – непрерывная величина с плотностью вероятности $f(x)$. По аналогии с дискретной случайной величиной имеем

$$M[X] = m_x = \int_{-\infty}^{+\infty} x f(x)\, dx. \qquad (4.2)$$

Действительно, интеграл (4.2) можно рассматривать как бесконечную сумму произведений возможных значений случайной величины на ее вероятность $f(x)dx$ (т. к. $P(x \approx x_i) = f(x_i)\, dx_i$).

Математическое ожидание является наиболее употребляемой характеристикой, поэтому полезно знать некоторые его свойства.

1. Математическое ожидание неслучайной величины C равно C, т. е. $M[C] = C$.

Доказательство. Неслучайная величина C может принимать только одно значение, равное C, с вероятностью 1. Поэтому, согласно (4.1), $m_x = C \cdot 1 = C$.

2. Математическое ожидание произведения неслучайной величины C на случайную величину X равно произведению C на m_x, т. е. $M[C \cdot X] = C \cdot m_x$.

Доказательство.

$$M[C \cdot X] = \sum_{i=1}^{n} C \cdot x_i \, p_i = C \sum_{i=1}^{n} x_i \, p_i = C \cdot m_x.$$

3. Если X и Y – независимые случайные величины, то математическое ожидание их суммы равно сумме их математических ожиданий, т. е. $M\left[X + Y\right] = M\left[X\right] + M\left[Y\right]$, а математическое ожидание их произведения равно произведению их математических ожиданий, т. е. $M[X \cdot Y] = M\left[X\right] \cdot M\left[Y\right]$.

4.2.2. Мода случайной величины

Модой дискретной случайной величины называется наиболее вероятное ее значение

$$M = x_m,$$

где $P(x_m) = \max\{P(x_j)\}$, $j = 1, ..., n$.

Аналогично для непрерывной случайной величины модой называется такое ее значение, для которого значение дифференциальной функции распределения максимально

$$M = x_m,$$

где $P(x_m) = \max\{f(x)\}$.

Понятие мода поясняется на рис. 4.1.

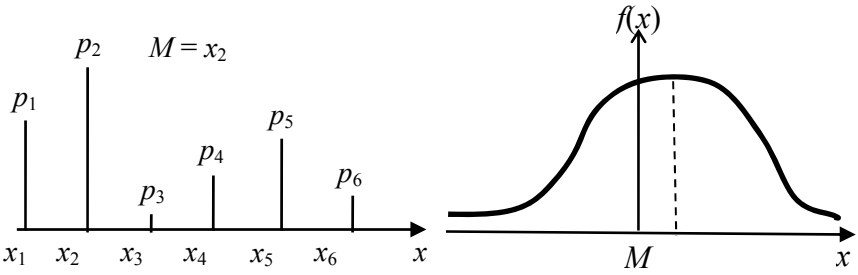


Рис. 4.1. Мода случайной величины

4.2.3. Медиана случайной величины

Медианой случайной величины называется такое ее значение μ, для которого выполняется $P(x < \mu) = P(x > \mu)$. Эти вероятности можно трактовать как площади под кривой плотности вероятности (рис. 4.2).

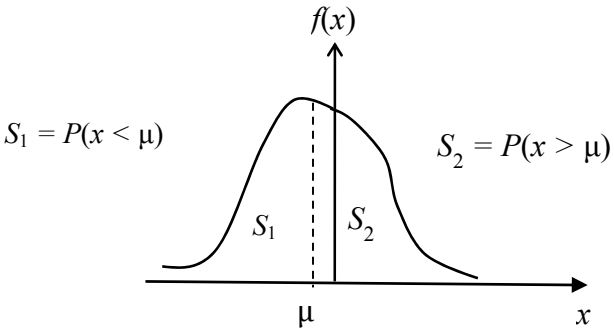


Рис. 4.2. Медиана случайной величины

4.3. Характеристики рассеяния

Числовые характеристики рассеяния случайной величины – это такие величины, которые характеризуют разброс возможных значений случайной величины относительно математического ожидания.

Различают следующие числовые характеристики рассеяния: дисперсию, среднее квадратическое отклонение.

4.3.1. Дисперсия

Пусть X – дискретная случайная величина с возможными значениями $x_1, x_2, ..., x_n$, вероятностями этих значений $p_1, p_2, ..., p_n$, математическим ожиданием m_x.

Дисперсией случайной величины X называется сумма произведений квадратов отклонений случайной величины от ее математического ожидания на вероятности этих отклонений:

$$D[X] = D_x = \sum_{i=1}^{n} (x_i - m_x)^2 p_i. \qquad (4.3)$$

Как видно из (4.3) дисперсия есть среднее взвешенное значение квадрата отклонения случайной величины относительно математического ожидания. Причем весовыми коэффициентами являются вероятности каждого значения квадрата отклонения. Необходимость использования в качестве меры рассеяния квадрата отклонения, а не просто отклонения, объясняется тем, что при суммировании всех возможных отклонений будет происходить компенсация положительных и отрицательных отклонений. В результате этого подобная характеристика не может характеризовать средний разброс случайной величины.

Для непрерывной случайной величины по аналогии с (4.2), (4.3) получим

$$D[X] = D_x = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x) dx. \qquad (4.4)$$

Размерность дисперсии равна квадрату случайной величины.

Дисперсия случайной величины обладает следующими свойствами.

1. Дисперсия случайной величины равна разности математического ожидания квадрата случайной величины и квадрата математического ожидания случайной величины

$$D[X] = M[X^2] - m_x^2. \tag{4.5}$$

Докажем справедливость (4.5) для непрерывной случайной величины с плотностью вероятности $f(x)$. Согласно (4.4) имеем

$$D[X] = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x)\,dx = \int_{-\infty}^{+\infty} (x^2 - 2x m_x + m_x^2) f(x)\,dx =$$

$$= \int_{-\infty}^{+\infty} x^2 f(x)\,dx - 2m_x \int_{-\infty}^{+\infty} x f(x)\,dx + m_x^2 \int_{-\infty}^{+\infty} f(x)\,dx =$$

$$= M[X^2] - 2m_x^2 + m_x^2 = M[X^2] - m_x^2.$$

2. Дисперсия неслучайной величины C равна 0.

Доказательство. Согласно (4.5) имеем

$$D[C] = M[C^2] - m_c^2 = C^2 - C^2 = 0.$$

3. Дисперсия произведения неслучайной величины C на случайную равна произведению квадрата неслучайной величины на дисперсию случайной величины

$$D[C\,X] = C^2 D[X]. \tag{4.6}$$

4. Дисперсия суммы (разности) двух независимых случайных величин равна сумме дисперсий этих величин

$$D[Y \pm X] = D[Y] + D[X]. \tag{4.7}$$

4.3.2. Среднее квадратическое отклонение

Среднее квадратическое отклонение есть корень квадратный из дисперсии

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

Среднее квадратическое отклонение имеет размерность случайной величины и позволяет оценивать среднюю величину отклонения случайной величины относительно ее математического ожидания. Поскольку среднее квадратическое отклонение связано с дисперсией монотонной зависимостью, все свойства дисперсии могут легко трансформироваться для среднего квадратического отклонения. Например, можно показать, что если $X = X_1 + X_2 + ... + X_n$, где X_i – независимые случайные величины, то справедливо

$$\sigma_x = \sqrt{D[X_1] + D[X_2] + ... + D[X_n]}. .$$

5. СИСТЕМЫ СЛУЧАЙНЫХ ВЕЛИЧИН

5.1. Функция распределения системы случайных величин

Часто результат опыта описывается не одной, а несколькими случайными величинами, образующими систему случайных величин. Например, результаты измерения роста и веса человека, координаты точки x, y, z, случайным образом помещенной в пространстве. Рассмотрим основные вероятностные характеристики системы случайных величин по аналогии с характеристиками одиночной случайной величины на примере системы, состоящей из двух величин.

5.1.1. Интегральная функция распределения системы случайных величин

Интегральной функцией распределения системы двух случайных величин (X, Y) называется вероятность совместного выполнения двух неравенств $X < x$, $Y < y$.

$$F(x, y) = P(X < x, Y < y). \tag{5.1}$$

Выражение (5.1) дает значение вероятности попадания точки в заштрихованную область (рис. 5.1).

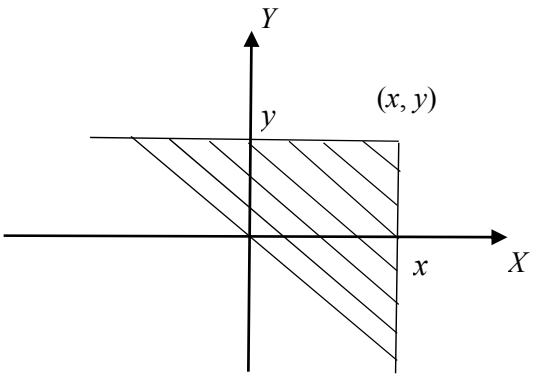


Рис. 5.1.

Рассмотрим свойства функции $F(x, y)$.

1. Функция $F(x, y)$ – неубывающая функция своих аргументов, т. е. если $x_1 < x_2,\ y_1 < y_2,$ то $F(x_2, y) \geq F(x_1, y)$ и $F(x, y_2) \geq F(x, y_1)$.

Доказательство. Так как $F(x, y) = P(X < x, Y < y),$ то меньшему значению x соответствует меньшее значение вероятности при фиксированном значении y. Аналогично меньшему значению y соответствует меньшее значение вероятности при фиксированном значении x.

2. $F(x, -\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0.$

Доказательство. Так как $F(x, -\infty) = P(X < x, Y < -\infty),$ а событие $Y < -\infty$ не возможно, то $F(x, -\infty) = 0.$ Аналогично событие $X < -\infty$ невозможно, поэтому $F(-\infty, y) = 0,\ F(-\infty, -\infty) = 0.$

3. $F(\infty, \infty) = 1.$

Доказательство. Так как $F(\infty, \infty) = P(X < \infty, Y < \infty),$ а события $X < \infty, Y < \infty$ достоверны, то $F(\infty, \infty) = 1.$

4. $F(x, \infty) = F_1(x),\ F(\infty, y) = F_2(y).$

Доказательство. Так как $F(x, \infty) = P(X < x, Y < \infty),$ а событие $Y < \infty$ достоверно, то $F(x, \infty)$ зависит только от вероятности $P(X < x),$ следовательно, $F(x, \infty) = F_1(x).$ Аналогично доказывается, что $F(\infty, y) = F_2(y).$

5.1.2. Дифференциальная функция распределения системы случайных величин

Дифференциальной функцией распределения системы случайных величин называется функция

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}. \qquad (5.2)$$

Функция (5.2) обладает аналогичными смыслом и свойствами, что и дифференциальная функция распределения одной случайной величины. Отметим, что

а) $f(x, y) \geq 0,$ т. к. $f(x, y)$ есть производная от неубывающей по x и y функции $F(x,y)$;

б) $F(x,y)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(x,y)dxdy;$

в) $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)dxdy=1.$

Типичный вид функции $f(x,y)$ представлен на рис. 5.2.

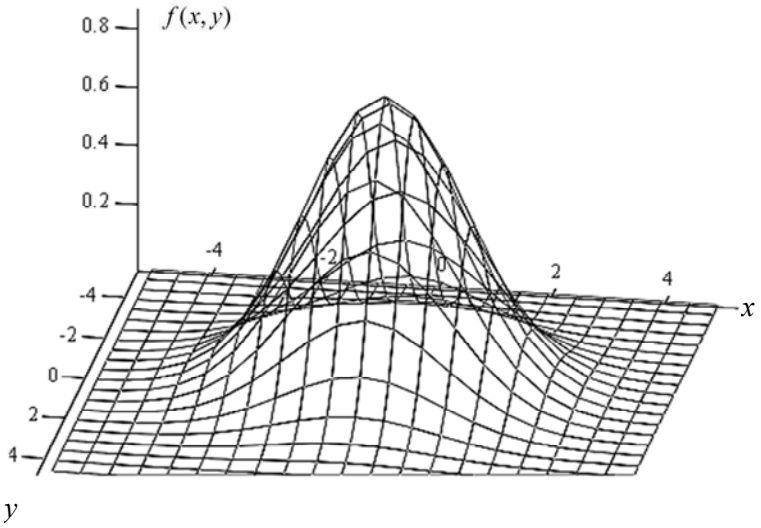


Рис. 5.2. Типичный вид функции $f(x,y)$

5.2. Законы распределения отдельных величин, условные законы распределения

Зная закон распределения системы случайных величин в виде интегральной или дифференциальной функции распределения, можно определить закон распределения отдельных величин, входящих в систему. Используя доказанные ранее свойства интегральной функции распределения, имеем

$$F_1(x)=F(x,\infty)\ ,F_2(y)=F(\infty,y).$$

Для дифференциальной функции распределения имеем

$$f_1(x) = \frac{dF_1(x)}{dx} = \frac{dF(x,\infty)}{dx} = \frac{d}{dx}[\int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(x,y)dxdy] = \int_{-\infty}^{+\infty} f(x,y)dy.$$

Аналогично для $f_2(y)$

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx.$$

5.3. Зависимые и независимые случайные величины

Как и случайные события, случайные величины могут быть зависимыми и независимыми.

Случайная величина Y называется независимой от случайной величины X, если ее закон распределения не зависит от того, какое значение приняла величина X. По аналогии с формулой умножения вероятностей можно показать, что для зависимых случайных величин справедливо

$$f(x,y) = f_1(x)\varphi_1(y/x), \tag{5.3}$$

$$f(x,y) = f_2(y)\varphi_2(x/y), \tag{5.4}$$

где $\varphi_1(y/x)$, $\varphi_2(x/y)$ – условные плотности вероятностей соответственно Y при фиксированном X и X при фиксированном Y. Для независимых случайных величин справедливо

$$f(x,y) = f_1(x)f_2(y), \tag{5.5}$$

т. к. для независимых величин плотность вероятности одной величины не зависит от того, какое значение принимает другая величина. Таким образом, условием независимости двух случайных величин является выполнение (5.5) или

$$\varphi_1(y/x) = f_2(y).$$

$$\varphi_2(x / y) = f_1(x).$$

Из (5.3), (5.4) получим

$$\varphi_1(y / x) = \frac{f(x,y)}{f(y)} = \frac{f(x,y)}{\int\limits_{-\infty}^{+\infty} f(x,y)dx},$$

$$\varphi_2(x / y) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int\limits_{-\infty}^{+\infty} f(x,y)dy}.$$

В отличии от функциональной зависимости, при которой две величины жестко связаны, вероятностная зависимость позволяет указать только закон распределения Y при известном X. Можно сказать, функциональная зависимость – это один полюс, независимость – второй, а между ними находится вероятностная зависимость. Примерами случайных величин, связанных такой зависимостью, являются вес человека и его рост, уровень образования и уровень доходов.

5.4. Корреляционный момент. Коэффициент корреляции

Система случайных величин, также как и отдельная случайная величина, может описываться числовыми характеристиками: математическим ожиданием, дисперсией, средним квадратическим отклонением величин, входящих в систему. Но кроме этих характеристик для описания зависимости между величинами, входящими в систему, вводятся корреляционные характеристики: корреляционный момент и коэффициент корреляции.

Корреляционным моментом двух случайных величин X и Y называется математическое ожидание произведения центрированных этих величин

$$K_{xy} = M[(X - m_x)(Y - m_y)]. \qquad (5.6)$$

Используя (5.6), можно получить расчетные формулы для дискретных и непрерывных случайных величин. Для дискретных величин

$$K_{xy} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_i - m_x)(y_j - m_y) p_{ij}. \qquad (5.7)$$

Для непрерывных величин

$$K_{xy} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - m_x)(y - m_y) f(x, y) dxdy. \qquad (5.8)$$

Величина корреляционного момента зависит от степени зависимости случайных величин и от величины их рассеяния. Так, для независимых случайных величин $K_{xy} = 0$. Покажем это, используя (5.8). Т. к. X и Y независимы и математическое ожидание центрированных величин равно 0, то (5.6) примет вид

$$K_{xy} = M[(X - m_x)(Y - m_y)] = M[(X - m_x)]M[Y - m_y)] = 0.$$

Пусть между X и Y существует функциональная связь типа $X = Y$, тогда

$$K_{xy} = M[(X - m_x)(Y - m_y)] = M[(X - m_x)(Y - m_x)] = D_x.$$

Таким образом, если корреляционный момент не равен 0, то это есть признак наличия зависимости между случайными величинами. С другой стороны на величину корреляционного момента влияет и рассеяние случайных величин: чем оно меньше, тем меньше корреляционный момент. Поэтому для оценки «чистой» силы связи используют нормированный корреляционный момент, называемый коэффициентом корреляции

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}. \qquad (5.9)$$

Коэффициент корреляции обращается в 0 одновременно с корреляционным моментом, следовательно, для независимых величин он равен 0 и принимает значение ± 1 для величин, связанных линейной функциональной зависимостью.

Случайные величины, коэффициент корреляции которых отличен от 0, называются коррелированными.

Независимые случайные величины всегда некоррелированные, некоррелированные не всегда независимые.

Если при возрастании одной случайной величины другая имеет тенденцию в среднем возрастать, то такая корреляция называется положительной, если убывать, то такая корреляция называется отрицательной.

Следует иметь в виду, что коэффициент корреляции характеризует не всякую зависимость, а только линейную.

Итак, коэффициент корреляции лежит в пределах $-1 \le r_{xy} \le 1,$ если $r_{xy} > 0,$ то корреляция положительна, если $r_{xy} < 0,$ то корреляция отрицательна.

6. ОСНОВНЫЕ ВИДЫ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

Изучение различных случайных величин, встречающихся в практических задачах, показало, что законы их распределения весьма разнообразны. Однако можно выделить небольшую группу законов, которые встречаются наиболее часто. Среди них нормальный, экспоненциальный, равномерный законы распределения для непрерывных случайных величин. Биномиальный, закон Пуассона – для дискретных. Рассмотрим эти законы распределения.

6.1. Закон равномерной плотности

Если возможные значения случайной величины лежат в некотором конечном интервале и все они равновероятны, то закон распределения такой величины равномерный. Например, при вращении вектора с силой, величина которой случайна, его отклонение θ относительно начального положения есть случайная величина, значения которой равномерно распределены в интервале от 0 до 2π.

6.1.1. Дифференциальная функция распределения

Случайная величина X имеет равномерный закон распределения, если ее дифференциальная функция распределения описывается следующим выражением

$$f(x)=\begin{cases}0, & \text{при } x<\alpha\\ c, & \text{при } \alpha\le x\le\beta\\ 0, & \text{при } x>\beta,\end{cases} \qquad (6.1)$$

где α, β – границы интервала;

c – некоторая постоянная величина ($c \ge 0$).

Функция (6.1) изображена на рис. 6.1. Величину c найдем с учетом выполнения для любой дифференциальной функции распределения условия

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

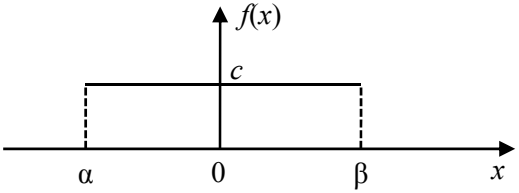


Рис. 6.1. $f(x)$ для равномерного закона распределения

Подставляя в это выражение функцию (6.1), получим $\int_{\alpha}^{\beta} c\,dx = c(\beta-\alpha)=1$, откуда $c=\frac{1}{\beta-\alpha}$. С учетом этого выражение для дифференциальной функции распределения примет вид

$$f(x)=\begin{cases}0, & \text{при } x<\alpha\\ \dfrac{1}{\beta-\alpha}, & \text{при } \alpha\le x\le\beta\\ 0, & \text{при } x>\beta.\end{cases} \qquad (6.2)$$

6.1.2. Интегральная функция распределения

Интегральная функция распределения равна

$$F(x)=\int_{-\infty}^{x} f(x)\,dx=\int_{\alpha}^{x}\frac{1}{\beta-\alpha}\,dx=\frac{x-\alpha}{\beta-\alpha}. \qquad (6.3)$$

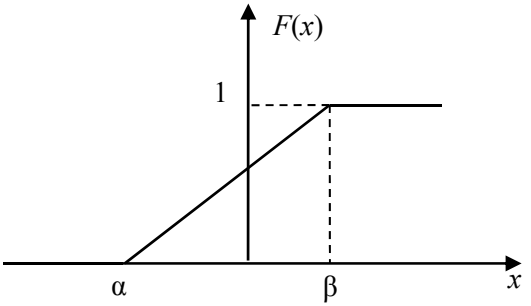


Рис. 6.2. $F(x)$ для равномерного закона распределения

6.1.3. Числовые характеристики

Определим основные числовые характеристики: математическое ожидание, дисперсию, среднее квадратическое отклонение

$$M[X] = \int_{-\infty}^{+\infty} x\, f(x)dx = \int_{\alpha}^{\beta} x \frac{1}{\beta - \alpha}\, dx = \frac{\beta + \alpha}{2}, \tag{6.4}$$

$$D[X] = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x)dx = \int_{\alpha}^{\beta} (x - m_x)^2 \frac{1}{\beta - \alpha}\, dx = \frac{(\beta - \alpha)^2}{12}, \tag{6.5}$$

$$\sigma[X] = \frac{\beta - \alpha}{2\sqrt{3}}.$$

6.1.4. Вероятность попадания в заданный интервал

Найдем вероятность попадания случайной величины X, имеющей равномерное распределение, в интервал $[a, b]$. Для этого воспользуемся формулой (6.4)

$$P(a < X < b) = F(b) - F(a) = \frac{b - \alpha}{\beta - \alpha} - \frac{a - \alpha}{\beta - \alpha} = \frac{b - a}{\beta - \alpha}. \tag{6.6}$$

6.2. Экспоненциальное (показательное) распределение

Экспоненциальное распределение широко используется в технике, экономике, медицине. Например, случайная величина – время между двумя соседними обращениями клиентов в страховую компанию, подчиняется экспоненциальному закону распределения, аналогично, время между соседними отказами сложной радиоэлектронной аппаратуры.

6.2.1. Дифференциальная функция распределения

Дифференциальная функция распределения случайной X, имеющей экспоненциальное распределение, имеет вид

$$f(x)=\begin{cases}0, & \text{при } x<0\\ \lambda e^{-\lambda x}, & \text{при } x\geq 0,\end{cases} \qquad (6.7)$$

где λ – параметр распределения ($\lambda > 0$).

Функция (6.7) изображена на рис. 6.3. Параметр λ определяет скорость уменьшения $f(x)$, чем больше λ, тем быстрее уменьшается $f(x)$.

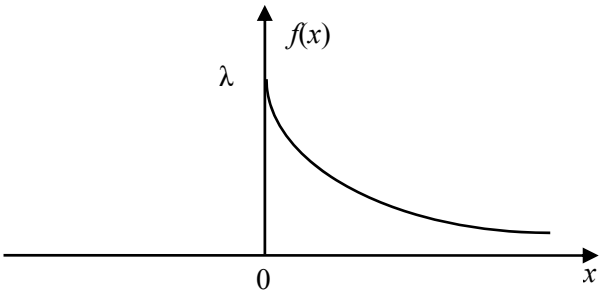


Рис. 6.3. $f(x)$ для экспоненциального закона распределения

6.2.2. Интегральная функция распределения

Проинтегрировав выражение (6.7), получим выражение для $F(x)$

$$F(x)=\int_{-\infty}^{x} f(x)dx=\begin{cases}0, & \text{при } x<0\\ \int_{0}^{x}\lambda e^{-\lambda x}dx=1-e^{-\lambda x}, & \text{при } x\geq 0.\end{cases} \qquad (6.8)$$

Функция (6.8) изображена на рис. 6.4.

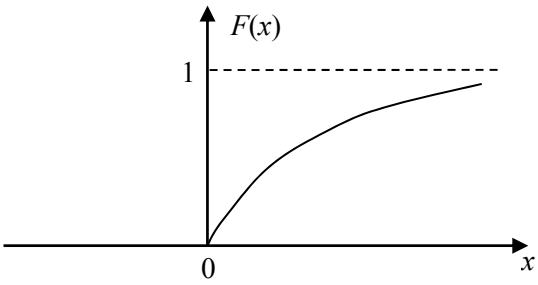


Рис. 6.4. $F(x)$ для экспоненциального закона распределения

6.2.3. Числовые характеристики

Найдем числовые характеристики случайной величины X, имеющей экспоненциальный закон распределения.

Математическое ожидание равно

$$M[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_{0}^{+\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}. \qquad (6.9)$$

Здесь и ниже при вычислении интегралов от показательной функции используется табличный интеграл $\int_{0}^{+\infty} x^n e^{-\lambda x} = \frac{n!}{\lambda^{n+1}}$, где $n = 1, 2, 3, ...$

Из (6.9) видно, что математическое ожидание зависит только от параметра λ, причем связь между ними обратно пропорциональная.

Дисперсия случайной величины X равна

$$D[X] = M[X^2] - m_x^2 =$$

$$= \int_{-\infty}^{+\infty} x^2 f(x) dx - \frac{1}{\lambda^2} = \int_{0}^{+\infty} x^2 \lambda e^{-\lambda x} dx - \frac{1}{\lambda^2} = \frac{2}{\lambda} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Из последнего выражения видно, что для экспоненциального распределения дисперсия равна квадрату математического ожидания. Очевидно, что среднее квадратическое отклонение равно математическому ожиданию. Действительно,

$$\sigma[X] = \sqrt{D[X]} = M[X] = \frac{1}{\lambda}. \qquad (6.10)$$

Вероятность попадания случайной величины X в интервал $[a, b]$ равна

$$P(a \le X \le b) = F(b) - F(a) = 1 - e^{-\lambda b} - 1 + e^{-\lambda a} = \\ = e^{-\lambda a} - e^{-\lambda b}. \qquad (6.11)$$

6.3. Нормальный закон распределения

Нормальный закон распределения (закон Гаусса) играет в ТВ исключительно большую роль. Это наиболее часто встречающийся на практике закон распределения. Он является предельным для многих других законов. Например, как будет показано далее, закон распределения суммы достаточно большого числа случайных величин с любыми законами распределений является нормальным.

6.3.1. Дифференциальная функция распределения

Случайная величина X имеет нормальный закон распределения, если ее дифференциальная функция распределения имеет вид

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad (6.12)$$

где m, σ – параметры распределения $(\sigma \geq 0)$. Зависимость (6.12) изображена на рис. 6.5.

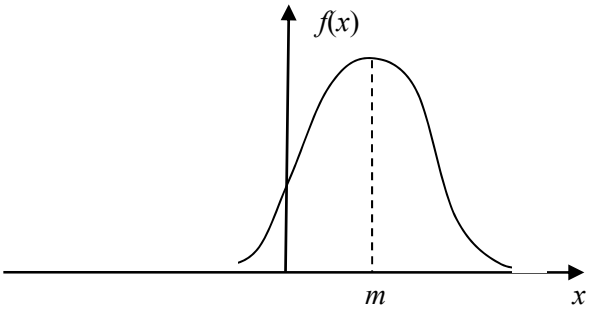


Рис. 6.5. $f(x)$ для нормального закона распределения.

6.3.2. Числовые характеристики

Найдем числовые характеристики случайной величины X, имеющей нормальный закон распределения.

Математическое ожидание равно

$$M[X]=\int\limits_{-\infty}^{+\infty} x\,f(x)dx=\int\limits_{-\infty}^{+\infty}\frac{x}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx.$$

Произведем замену переменной: $\frac{x-m}{\sqrt{2}\sigma}=t,\quad dx=dt\,\sqrt{2}\sigma,\quad x=$ $=\sqrt{2\sigma}t+m.$ Получим

$$M[X]=\frac{1}{\sqrt{2\pi}\,\sigma}\int\limits_{-\infty}^{+\infty}(\sqrt{2\sigma}\,t+m)e^{-t^2}\,\sqrt{2}\sigma\,dt=$$

$$=\frac{1}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty}\sigma\sqrt{2}\,te^{-t^2}dt+\frac{m}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty}e^{-t^2}dt.$$

В последнем выражении первый интеграл равен 0, как интеграл от нечетной функции в симметричных пределах, второй – интеграл Пуассона, равен $\sqrt{\pi}$. С учетом этого, получим

$$M[X]=m. \tag{6.13}$$

Таким образом, параметр m нормального закона распределения равен математическому ожиданию случайной величины и определяет положение максимума дифференциальной функции распределения.

Дисперсия случайной величины X равна

$$D[X]=M[X^2]-m_x^2=$$

$$=\int\limits_{-\infty}^{+\infty}x^2\,f(x)dx-m_x^2=\int\limits_{-\infty}^{+\infty}\frac{x^2}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx-m^2.$$

Можно показать, что после ряда преобразований

$$D[X]=\sigma^2. \tag{6.14}$$

Очевидно, что среднее квадратическое отклонение равно $\sigma[X] = \sigma$.

Таким образом, параметр σ нормального закона распределения равен среднему квадратическому отклонению случайной величины.

6.3.3. Влияние параметров распределения на положение и форму дифференциальной функции распределения

Как уже говорилось ранее, параметр m имеет смысл математического ожидания и определяет положение максимума функции вдоль оси x, не влияя при этом на форму кривой (рис. 6.6). Так при $m < 0$ максимум функции находится в области отрицательных значений x, при $m > 0$ – в области положительных значений.

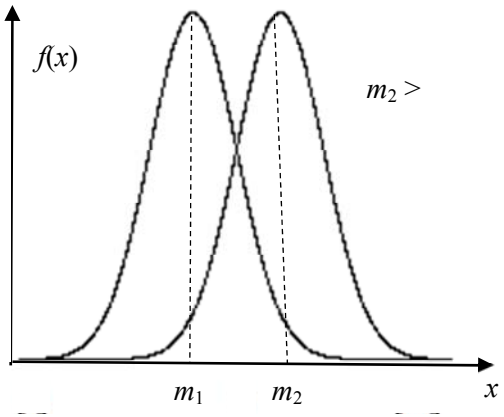


Рис. 6.6. Влияние параметра m на положение $f(x)$

Параметр σ имеет смысл среднего квадратического отклонения. Из (6.12) видно, что σ не влияет на положение максимума вдоль оси x, но влияет на форму кривой. Чем больше величина σ, тем меньше величина максимального значения функции и тем сильнее растягивается кривая вдоль оси x (рис. 6.7).

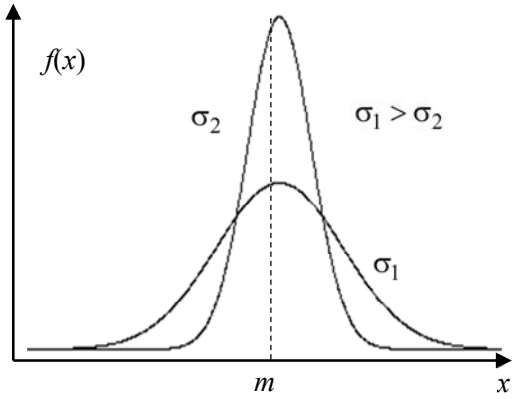


Рис. 6.7. Влияние параметра σ на форму $f(x)$

6.3.4. Интегральная функция распределения

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

Произведем замену переменой: $\frac{x-m}{\sqrt{2}\sigma} = t, \quad dx = dt\,\sqrt{2}\sigma, \quad x = \sqrt{2\sigma}\,t + m,$ тогда

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

Данный интеграл не вычисляется через элементарные функции, однако его можно выразить через табулированную функцию Лапласа (таблица значений этой функции приведена в табл. П1).

$$\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_{0}^{z} e^{-\frac{t^2}{2}}\,dt. \qquad (6.15)$$

Функция $\Phi(z)$ обладает следующими свойствами: $\Phi(z)$ – нечетная функция, т. е. $\Phi(-z) = -\Phi(z)$, $\Phi(-\infty) = -1$, $\Phi(0) = 0$, $\Phi(+\infty) = 1$. График $\Phi(z)$ имеет вид (рис. 6.8, *а*). С учетом (6.15) функция $F(x)$ равна

$$F(x) = 1/2[1 + \Phi(\frac{x-m}{\sigma})]. \qquad (6.16)$$

Функция $F(x)$ изображена на рис. 6.8, *б*.

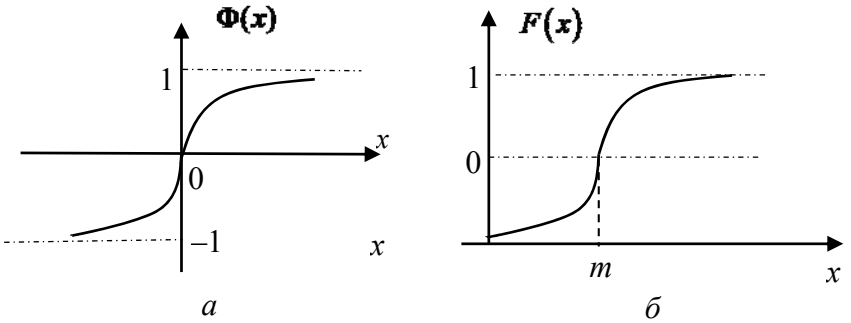


Рис. 6.8. Графики функций:
а – $\Phi(x)$; *б* – $F(x)$

Зная интегральную функцию распределения случайной величины, можно определить вероятность ее попадания в заданный интервал $[a, b]$

$$P(a < X < b) = F(b) - F(a) = \frac{1}{2}[\Phi(\frac{b-m}{\sigma}) - \Phi(\frac{a-m}{\sigma})]. \qquad (6.17)$$

Для симметричного относительно параметра m интервала (6.17) существенно упрощается. Пусть требуется найти вероятность

$$P(m-\delta < X < m+\delta) = \frac{1}{2}[\Phi(\frac{m+\delta-m}{\sigma}) - \Phi(\frac{m-\delta-m}{\sigma})] = \\ = \frac{1}{2}[\Phi(\frac{\delta}{\sigma}) - \Phi(\frac{-\delta}{\sigma})] = \frac{1}{2}[\Phi(\frac{\delta}{\sigma}) + \Phi(\frac{\delta}{\sigma})] = \Phi(\frac{\delta}{\sigma}). \qquad (6.18)$$

Расчеты показывают, что при $\delta = 3\sigma$ вероятность $P(m - 3\sigma < X < m + 3\sigma) = \Phi(3) = 0{,}997$. Т. е. вероятность отклонения случайной величины X от ее математического ожидания на величину больше, чем 3σ равна 0,003. Иными словами, отклонение случайной величины X от ее математического ожидания на величину больше, чем 3σ маловероятно. Этот факт получил в ТВ название «правило 3σ».

6.4. Биномиальное распределение

Биномиальный закон распределения имеет дискретная случайная величина, описанная в теореме Бернулли. Т. е. если производится n независимых испытаний, в каждом из которых некоторое событие A наступает или нет с вероятностями p и $q = 1 - p$, то случайная величина X – число появлений события A в этих испытаниях – имеет биномиальное распределение. В соответствии с теоремой Бернулли вероятность того, что случайная величина X после проведения n испытаний будет равна x, определяется по формуле

$$P(X = k) = P_n(k) = C_n^k p^k q^{n-k}, \qquad (6.19)$$

где $k = 0, 1, 2, ..., n$. Интегральная функция распределения равна

$$P(k < x) = \sum_{k=0}^{x-1} C_n^k p^k q^{n-k}. \qquad (6.20)$$

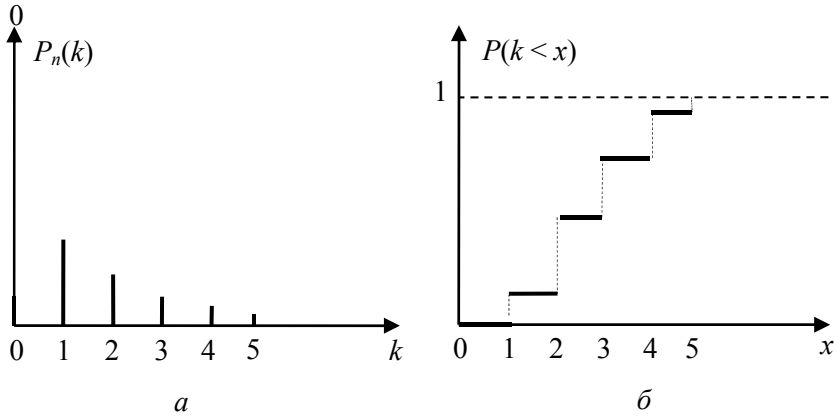
Распределение получило свое название ввиду того, что правая часть (6.19) представляет собой общий член разложения бинома Ньютона $(p + q)^n$. Примеры функций (6.19), (6.20) для $n = 5$ изображены на рис. 6.10.

Числовые характеристики случайной величины, имеющей биномиальное распределение, равны:

$$M[X] = \sum_{k=0}^{n} k\, C_n^k p^k q^{n-k} = np,$$

$$D[X] = \sum_{k=0}^{n} (k - np)^2 \, C_n^k p^k q^{n-k} = npq,$$

$$\sigma[X] = \sqrt{npq}.$$



а *б*

Рис. 6.10. Биномиальный закон распределения:
а – $P_n(k)$; *б* – $P(k < x)$

6.5. Закон распределения Пуассона

Дискретная случайная величина X, которая может принимать только целые значения 0, 1, 2, 3, ..., имеет распределение Пуассона, если вероятность того, что она примет значение m, равна

$$P(X = k) = P_k = \frac{a^k}{k!} e^{-a}, \qquad (6.21)$$

где a – параметр закона распределения Пуассона ($a > 0$).

Обычно этот закон распределения справедлив для случайных величин, представляющих собой сумму независимых случайных событий, последовательно попадающих в некоторый интервал, если расстояние между ними (время) подчиняется экспоненциальному закону распределения. Например, число вызовов машины «Скорой помощи» за 1 час, число отказов сложного оборудования за 1 месяц.

Интегральная функция распределения равна

$$P(k < x) = \sum_{k=0}^{x-1} \frac{a^k}{k!} e^{-a}. \qquad (6.22)$$

Примерный вид функций (6.21), (6.22) изображен на рис. 6.12.

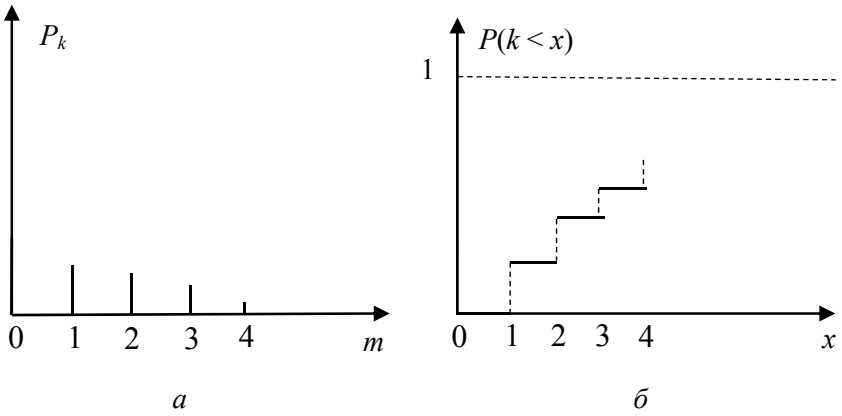


Рис. 6.12. Графики функций:
$а - P_k$; $б - P(k < x)$

Найдем числовые характеристики случайной величины, имеющей распределение Пуассона:

$$M[X] = \sum_{m=0}^{\infty} mP_m = \sum_{m=0}^{\infty} m \frac{a^m}{m!} e^{-a} = a,$$

$$D[X] = \sum_{m=0}^{\infty} (m-a)^2 P_m = \sum_{m=0}^{\infty} (m-a)^2 \frac{a^m}{m!} e^{-a} = a,$$

$$\sigma[X] = \sqrt{a}.$$

При выводе этих формул использовались табличные преобразования, которые из-за громоздкости здесь не приводятся.

7. ЗАКОН БОЛЬШИХ ЧИСЕЛ И ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

7.1. Общие сведения

Массовые случайные явления обладают определенной устойчивостью средних (числовых) характеристик: при большом числе опытов конкретные особенности отдельных случайных явлений почти не сказываются на среднем результате массы таких явлений. Именно устойчивость средних представляет собой физическое содержание закона больших чисел, понимаемого как: при большом числе случайных явлений средний их результат практически перестает быть случайным и может быть предсказан с большой степенью определенности.

В узком смысле под законом больших чисел понимается ряд теорем, в которых устанавливается факт приближения средних характеристик большого числа опытов к некоторым постоянным.

Кроме предсказания средних характеристик имеются возможности предсказания предельных законов распределения. Это охватывается группой теорем – центральной предельной теоремой.

7.2. Неравенство Чебышева

Для доказательства теорем закона больших чисел рассмотрим неравенство Чебышева.

Пусть имеется случайная величина X с числовыми характеристиками m_x, D_x. Неравенство Чебышева утверждает, что

$$P(|X - m_x| \geq \alpha) \leq \frac{D_x}{\alpha^2}, \qquad (7.1)$$

где α – некоторое положительное число.

Докажем приведенное неравенство. Преобразуем левую часть неравенства (7.1). Событие $|X - m_x| \geq \alpha$ можно представить как $X - m_x \geq \alpha$, при $X - m_x > 0$ и $-X + m_x \geq \alpha$, при $X - m_x < 0$.

Откуда $X \geq m_x + \alpha$, или $X \leq m_x - \alpha$.

Тогда левая часть (7.1) будет равна

$$P(|X - m_x| \geq \alpha) = P(X \geq m_x + \alpha) + P(X \leq m_x - \alpha) = \\ = \int_{-\infty}^{m_x - \alpha} f(x)dx + \int_{m_x + \alpha}^{+\infty} f(x)dx. \qquad (7.2)$$

Смысл последнего выражения поясняется на рис. 7.1. Затененные области – это области, в пределах которых ведется интегрирование.

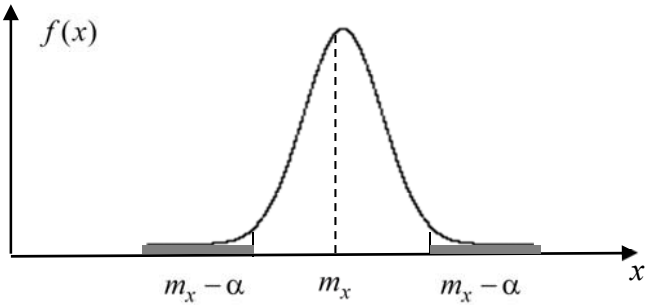


Рис. 7.1. Области интегрирования

С другой стороны дисперсия случайной величины X равна

$$D_x = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x)dx = \int_{-\infty}^{+\infty} |x - m_x|^2 f(x)dx \geq \int_{-\infty}^{m_x - \alpha} |x - m_x|^2 f(x)dx + \\ + \int_{m_x + \alpha}^{+\infty} |x - m_x|^2 f(x)dx.$$

Знак неравенства не изменится, если величину $|x - m_x|$ заменить на меньшую величину α, и с учетом (7.2) получим

$$D_x \geq \int_{-\infty}^{m_x - \alpha} \alpha^2 f(x)dx + \int_{m_x + \alpha}^{+\infty} \alpha^2 f(x)dx = \alpha^2 P(|x - m_x| \geq \alpha),$$

откуда $\dfrac{D_x}{\alpha^2} \geq P(|x - m_x| \geq \alpha)$, что эквивалентно (7.1).

7.3. Закон больших чисел (теорема Чебышева)

Теорема устанавливает связь между средним арифметическим наблюдаемых значений случайной величины и ее математическим ожиданием и формулируется следующим образом.

При достаточно большом числе независимых опытов среднее арифметическое наблюденных значений случайной величины сходится по вероятности к ее математическому ожиданию. Математически это означает

$$P(|X_n - m_x| < \varepsilon) > 1 - \delta, \tag{7.3}$$

где $X_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ – среднее арифметическое наблюденных значений X;

n – число наблюденных значений;

ε, δ – положительные произвольно малые числа.

Под сходимостью по вероятности понимается следующее. Говорят, что X_n сходится по вероятности к m_x, если с ростом n вероятность того, что X_n и m_x будут очень близки, стремится к 1.

Докажем сформулированную теорему. Вначале покажем, что математическое ожидание X_n равно математическому ожиданию X

$$M[X_n] = \frac{1}{n} M[\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \frac{1}{n}\sum_{i=1}^{n} m_x = m_x.$$

Дисперсия среднего арифметического X_n равна

$$D[X_n] = \frac{1}{n^2} D[\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} D_x = \frac{D_x}{n}.$$

Применим к X_n неравенство Чебышева, полагая $\alpha = \varepsilon$

$$P(|X_n - m_x| \geq \varepsilon) \leq \frac{D[X_n]}{\varepsilon^2} = \frac{D_x}{n\varepsilon^2}. \tag{7.4}$$

Как ни мало было бы ε, всегда можно найти такое значение n, чтобы выполнялось $\frac{D_x}{n\varepsilon^2} < \delta$. Тогда знак неравенства (7.4) $\leq$ изменится на $<$, если правую часть его заменить на большую величину

$$P(|X_n - m_x| \geq \varepsilon) < \delta. \qquad (7.5)$$

От события $|X_n - m_x| \geq \varepsilon$ перейдем к противоположному событию $|X_n - m_x| < \varepsilon$. Т. к. эти события составляют полную группу событий, то для их вероятностей справедливо $P(|X_n - m_x| \geq \varepsilon) = 1 - P(|X_n - m_x| < \varepsilon)$. С учетом этого из (7.5) получим $1 - P(|X_n - m_x| < \varepsilon) < \delta$, откуда $P(|X_n - m_x| < \varepsilon) > 1 - \delta$, что и требовалось доказать.

Теорема Чебышева может быть доказана и для случая, когда закон распределения случайной величины изменяется от опыта к опыту. И тем не менее при соблюдении некоторых условий среднее арифметическое является устойчивым и сходится по вероятности к определенной неслучайной величине (обобщенная теорема Чебышева). Обобщение закона больших чисел на зависимые случайные величины сделано в теореме Маркова.

7.4. Теорема Бернулли

Теорема Бернулли устанавливает связь между частотой появления случайного события в серии опытов и его вероятностью и доказывается как следствие закона больших чисел.

Пусть проводится n независимых опытов, в каждом из которых может появиться событие A с вероятностью p или не появиться с вероятностью $q = 1 - p$. Теорема Бернулли утверждает, что при неограниченном числе опытов n частота события A сходится по вероятности к вероятности события A:

$$P(|P^* - p| < \varepsilon) > 1 - \delta,$$

где ε, δ – малые положительные числа;

P^* – частота появления события A.

Доказательство. Теорема будет доказана, если доказать, что P^* есть среднее арифметическое некоторой случайной величины, математическое ожидание которой равно p. Рассмотрим независимые случайные величины X_i, каждая из которых принимает в каждом опыте значение 1, если событие A наступило, и принимает значение 0, если событие A не наступило. Следовательно, случайная величина X_i имеет следующий ряд распределения

x_i	0	1
p_i	q	p

Частоту появления события A в n опытах можно представить, как суммарное число наступления события A в n опытах, деленное на n

$$P^* = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad (7.6)$$

Из (7.6) видно, что частота события A есть среднее арифметическое случайной величины X_i. Найдем математическое ожидание P^*

$$M[P^*] = M[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} M[X_i].$$

Математическое ожидание случайной величины X равно

$$M[X_i] = 0 \cdot q + 1 \cdot p = p.$$

Следовательно, $M[P^*] = p$.

7.5. Центральная предельная теорема

Центральная предельная теорема устанавливает условия, при которых возникает нормальный закон распределения. Т. к. эти условия на практике очень часто соблюдаются, то нормальный закон является наиболее распространенным.

Ляпунов сформулировал и доказал центральную предельную теорему: если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на сумму ничтожно мало, то X имеет распределение, близкое к нормальному.

Пусть $X_1, X_2, ..., X_n$ – последовательность независимых случайных величин, каждая из которых имеет математическое ожидание и дисперсию:

$$M\left[X_i\right] = m_i, D\left[X_i\right] = D_i.$$

Для суммы случайных величин

$$S_n = \sum_{i=1}^{n} X_i.$$

Ляпунов доказал, что если каждое слагаемое, входящее в сумму, оказывает малое влияние, то S_n имеет нормальное распределение с параметрами

$$m = \sum_{i=1}^{n} m_i, \sigma^2 = \sum_{i=1}^{n} D_i.$$

Практически центральной предельной теоремой можно пользоваться и тогда, когда речь идет о сумме сравнительно небольшого числа случайных величин. Оказалось, что при суммировании независимых случайных величин, сравнимых по своему рассеянию, с увеличением суммы закон распределения суммы становится нормальным. Практически при 6–10 слагаемых закон распределения суммы близок к нормальному. Этот вывод объясняет тот факт, что очень многие случайные величины, встречающиеся на практике, подчиняются нормальному закону распределения. Дело в том, что по своей природе такие величины являются суммой других случайных величин, часто недоступных наблюдению.

8. СЛУЧАЙНЫЕ ФУНКЦИИ

8.1. Понятие о случайной функции. Характеристики случайных функций

Случайной функцией называется функция неслучайного аргумента, которая при каждом значении аргумента является случайной величиной. Обозначается случайная функция, как $X(t)$. Если аргумент случайной функции – непрерывная величина, то случайная функция непрерывна, если аргумент дискретен, то случайная функция дискретна и называется случайной последовательностью.

В результате опыта случайная функция при каждом значении аргумента принимает случайное значение. Совокупность этих значений называется реализацией случайной функции. Обозначается реализация – $x(t)$.

Случайная функция, аргументом которой является время, называется случайным процессом.

Примеры реализаций случайной функции изображены на рис. 8.1.

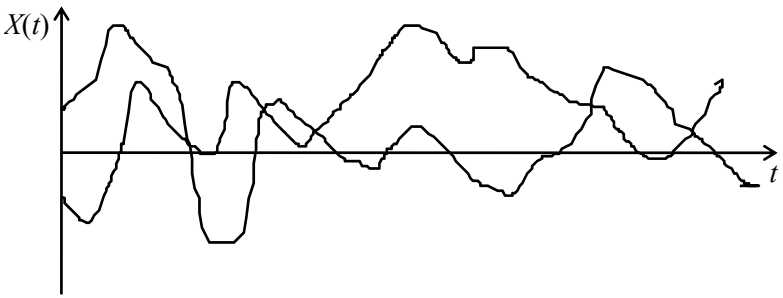


Рис. 8.1. Примеры реализаций случайной функции

Возможные значения случайной функции при фиксированном t называются сечением случайной функции. Для описания вероятностных свойств случайной функции используются характеристики, аналогичные вероятностным характеристикам случайных величин.

Одномерной интегральной функцией распределения случайной функции называется функция вида

$$F_1(x,t) = P(X(t) < x). \qquad (8.1)$$

Вероятность (8.1) равна вероятности того, что случайная функция будет меньше некоторого уровня x при всех значениях аргумента t.

Одномерной дифференциальной функцией распределения случайной функции называется функция вида

$$f_1(x,t) = \frac{dF_1(x,t)}{dx}. \tag{8.2}$$

Кроме того, часто используются двумерные функции распределения:

$$F_2(x_1, x_2, t_1, t_2) = P(X(t_1) < x_1; X(t_2) < x_2),$$

$$f_2(x_1, x_2, t_1, t_2) = \frac{\partial F^2(x_1, x_2, t_1, t_2)}{\partial x_1 \partial x_2}.$$

Чем выше порядок функций распределения, тем полнее характеризуют они вероятностные свойства случайной функции. Так одномерные функции распределения характеризуют случайную функцию как совокупность независимых случайных величин, в то время как двумерные позволяют учитывать корреляционную зависимость между отдельными сечениями. На практике зачастую вполне достаточно оказывается знания двумерных функций распределения.

Так же, как для случайных величин, для случайных функций используются понятия математическое ожидание, дисперсия, корреляционная функция и т. д.

8.1.1. Математическое ожидание

Математическим ожиданием случайной функции $X(t)$ называется неслучайная функция $m_x(t)$, значение которой при каждом значении t равно математическому ожиданию сечения при этом t, т. е. $m_x(t) = M[X(t)]$ или

$$m_x(t) = \int_{-\infty}^{+\infty} x f_1(x,t) dx. \tag{8.3}$$

Математическое ожидание случайной функции $X(t)$ обладает следующими основными свойствами:

1. Математическое ожидание неслучайной функции $W(t)$ равно этой функции, т. е.

$$M\left[W\left(t\right)\right] = W\left(t\right).$$

2. Математическое ожидание произведения случайной функции $X(t)$ на неслучайную функцию $W\left(t\right)$ равно произведению последней на математическое ожидание случайной функции, т. е.

$$M\left[X\left(t\right)W\left(t\right)\right] = W\left(t\right)m_x(t).$$

3. Математическое ожидание суммы (разности) независимых случайных функций $X(t)$ и $Y(t)$ равно сумме (разности) их математических ожиданий, т. е.

$$M[X\left(t\right)Y\left(t\right)] = m_x(t) \pm m_y(t).$$

8.1.2. Дисперсия и среднее квадратическое отклонение

Дисперсией случайной функции $X(t)$ называется неслучайная функция $D_x(t)$, значение которой при каждом значении t равно дисперсии сечения при этом t, т. е. $D_x(t) = D[X(t)]$ или

$$D_x(t) = \int_{-\infty}^{+\infty} (x - m_x)^2 f_1(x,t)dx. \qquad (8.4)$$

Дисперсия случайной функции $X(t)$ обладает следующими основными свойствами:

1. Дисперсия неслучайной функции $W\left(t\right)$ равна 0.

$$D\left[W\left(t\right)\right] = 0.$$

2. Дисперсия произведения случайной функции $X(t)$ на неслучайную функцию $W(t)$ равна произведению квадрата последней на дисперсию случайной функции, т. е.

$$D\left[X(t)W(t)\right] = W(t)^2 D_x(t).$$

3. Дисперсия суммы (разности) независимых случайных функций $X(t)$ и $Y(t)$ равна сумме их дисперсий, т. е.

$$D[X(t)Y(t)] = D_x(t) + D_y(t).$$

Среднее квадратическое отклонение - это корень квадратный из дисперсии

$$\sigma_x(t) = \sqrt{D_x(t)}. \qquad (8.5)$$

8.1.3. Корреляционная функция

Корреляционной функцией случайной функции $X(t)$ называется неслучайная функция $K_x(t_1, t_2)$, значение которой при каждом значении *t* равно корреляционному моменту двух сечений при этих значениях *t*, т. е.

$$K_x(t_1, t_2) = M[(X(t_1) - m_x(t_1))(X(t_2) - m_x(t_2))]]$$

или

$$K_x(t_1, t_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} ([x_1 - m_x(t_1))*(x_2 - m_x(t_2))] f_2(x_1, x_2, t_1, t_2) dx_1 dx_2.$$

Корреляционная функция $K_x(t_1, t_2)$ обладает следующими основными свойствами:

1. Значение корреляционной функции не изменится при перестановке аргументов, т. е. $K_x(t_1, t_2) = K(t_2, t_1)$.

2. Корреляционная функция, рассчитанная при одинаковых значениях аргументов, равна дисперсии случайной функции, т. е.

$$K_x(t_1,t_1) = D(t_1) \text{ или } K_x(t_2,t_2) = D(t_2).$$

3. Корреляционная функция случайной функции $X(t)$ не изменится, если к ней прибавить (отнять) неслучайную функцию $W(t)$, т. е.

$$K_x(t_1,t_2) = K_y(t_1,t_2),$$

где $K_y(t_1,t_2)$ – корреляционная функция суммы $Y(t) = X(t) + W(t)$.

4. Корреляционная функция произведения неслучайной функции на случайную равна произведению значений неслучайной функции при обоих значениях аргумента на корреляционную функцию случайной функции, т. е.

$$K_y(t_1,t_2) = W(t_1)W(t_2)K_x(t_1,t_2),$$

где $Y(t) = W(t)X(t)$.

Нормированной корреляционной функцией называется корреляционная функция, деленная на произведение средних квадратических отклонений, взятых при обоих значениях аргументов:

$$\rho_x(t_1,t_2) = \frac{K_x(t_1,t_2)}{\sigma_x(t_1)\sigma_x(t_2)}. \qquad (8.6)$$

Взаимной корреляционной функцией двух случайных функций $X(t)$ и $Y(t)$ называется неслучайная функция $R_{x,y}(t_1,t_2)$ двух аргументов t_1 и t_2, равная корреляционному моменту сечений обеих функций при этих значениях t:

$$R_{x,y}(t_1,t_2) = M[(X(t_1) - m_x(t_1))(Y(t_2) - m_y(t_2))]]$$

или

$$R_{x,y}(t_1,t_2) = \\ = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[(x-m_x(t_1))^*(y-m_y(t_2))]f_x(x,t_1)f_y(y,t_2)dxdy, \quad (8.7)$$

где $f_x(x,t_1)$, $f_y(y,t_2)$ – одномерные плотности вероятности случайных функций $X(t)$, $Y(t)$.

8.2. Стационарные случайные функции

Стационарной случайной функцией называется функция, вероятностные характеристики которой (функция распределения, математическое ожидание, дисперсия, корреляционная функция) не зависят от выбранного значения аргумента *t*.

Таким образом, для стационарных случайных функций справедливо:

$$m_x(t) = m_x, \quad D_x(t) = D_x, \quad F_1(x,t) = F(x), \quad K_x(t_1,t_2) = K_x(\tau),,$$

где $\tau = t_2 - t_1$.

Вероятностные характеристики стационарных случайных функций обладают теми же свойствами, что и характеристики нестационарных, причем многие характеристики существенно упрощаются. Например,

$$K(\tau) = K(-\tau), \quad K_x(0) = D_x.$$

8.3. Цепи Маркова

8.3.1. Общие сведения

Если аргумент случайной функции дискретен, то такая случайная функция называется случайной последовательностью. Математический аппарат для анализа произвольных случайных последовательностей достаточно сложен, поэтому ограничимся рассмотрением класса марковских последовательностей (цепей).

Цепью Маркова (марковской последовательностью) называется случайная последовательность, для которой вероятность того, что $X(t) = x_j(t_k)$ зависит только от предшествующего значения $X(t) = x_i(t_{k-1})$ и не зависит от значений в остальные предшествующие моменты времени.

Пусть $X(t)$ при каждом значении аргумента может принимать N дискретных значений, тогда марковская цепь имеет вид, представленный на рис. 8.2.

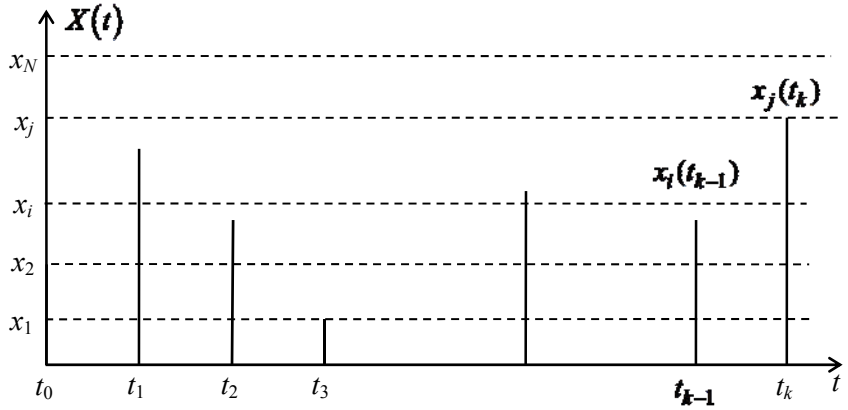


Рис. 8.2. Марковская цепь

Вероятность перехода процесса из состояния t_{k-1} в состояние t_{k-1} называется переходной вероятностью (вероятностью перехода). Обозначается эта вероятность $p_{ij}(k) = P(x_j, t_k / x_i, t_{k-1})$. Последовательность, у которой вероятности перехода одинаковы для всех значений t, называется однородной марковской цепью.

Чтобы нагляднее представить марковскую цепь, рассмотрим следующий пример. Пусть имеется марковская цепь, которая может принимать три значения x_1, x_2, x_3. Случайный процесс может в каждый момент времени скачком переходить из состояния в состояние. Обозначим вероятность перехода из состояния i в состояние j на k шаге через $p_{ij}(k)$. Диаграмма возможных переходов процесса изображена на рис. 8.3.

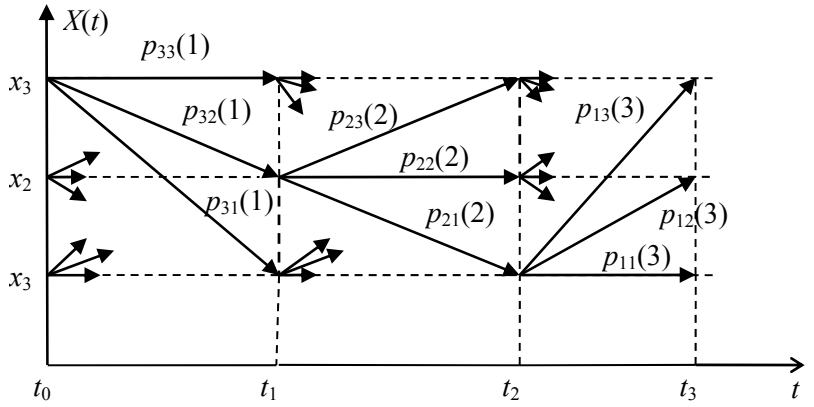


Рис. 8.3. Диаграмма возможных переходов марковской цепи

Все вероятности переходов образуют матрицу переходов, в которой i – номер предыдущего значения процесса X, j – номер текущего значения процесса X:

$$\left[p_{ij}(k)\right] = \begin{bmatrix} p_{11}(k) & p_{12}(k) & p_{13}(k) \\ P_{21}(k) & p_{22}(k) & p_{23}(k) \\ p_{31}(k) & p_{32}(k) & p_{33}(k) \end{bmatrix}. \qquad (8.9)$$

Для однородной марковской цепи справедливо $p_{ij}(k) = p_{ij}(k - -1) = p_{ij}$, а матрица переходов не зависит от номера шага:

$$\left[p_{ij}\right] = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ P_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}. \qquad (8.10)$$

Размер матрицы переходов в общем случае равен $N \times N$, N – число возможных значений $X(t)$. Кроме матрицы переходов необходимо знать начальное распределение $X(t)$, т. е. вероятности $p_i(0)$ в момент t_0. При этом соблюдается условие

$$\sum_{i=1}^{N} p_i(0) = 1.$$

Кстати, это условие соблюдается для каждой строки матрицы переходов, т. е.

$$\sum_{i=1}^{N} p_{ij}(0) = 1.$$

8.3.2. Равенство Маркова

В практических задачах часто необходимо определить $p_{ij}(n)$ – вероятность перехода случайной функции, представляющей собой однородную цепь Маркова, от *i*-го значения к *j*-му за *n* (интервалов) шагов. Для нахождения этой вероятности изобразим график перехода $X(t)$ от значения x_i к значению x_j за *n* шагов (рис. 8.4), считая, что на промежуточном шаге *m* значение $X(t)$ равно x_r.

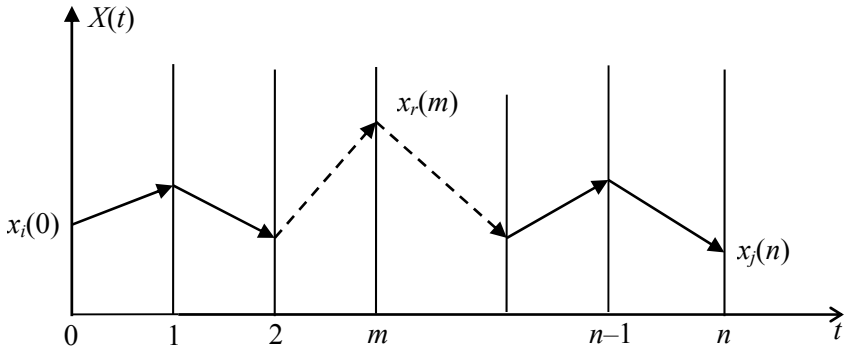


Рис. 8.4. График переходов случайной функции

Итак, будем считать, что за m интервалов случайная последовательность от x_i переходит к значению x_r с вероятностью $p_{ir}(m)$, за оставшиеся $n-m$ интервалов от x_r к x_J с вероятностью $p_{rj}(n-m)$. Тогда по формуле полной вероятности имеем

$$p_{ij}(n) = \sum_{r=1}^{N} p_{ir}(m)\, p_{rj}(n-m). \qquad (8.11)$$

Имея это выражение и матрицу вероятностей перехода $p_{ij}(1)$, можно найти $p_{ij}(n)$, где n = 1, 2, 3, ..:

$$p_{ij}(2) = \sum_{r=1}^{N} p_{ir}(1)\, p_{rj}(1) \qquad \text{для } m = 1,$$

$$p_{ij}(3) = \sum_{r=1}^{N} p_{ir}(2)\, p_{rj}(1) \qquad \text{для } m = 1,$$

$$p_{ij}(4) = \sum_{r=1}^{N} p_{ir}(3)\, p_{rj}(1) \qquad \text{для } m = 1.$$

Очень часто модель цепи Маркова используется для описания некоторой системы, которая может находится в N состояниях и переходить из одного состояния в другое с определенными вероятностями. Например, предприятие может находится в трех состояниях: 1 – работа; 2 – простой из-за отсутствия комплектующих элементов; 3 – простой из-за оборотных средств. Вероятности переходов – p_{ij}. Функционирование такого предприятия можно представить в виде графа (рис. 8.5).

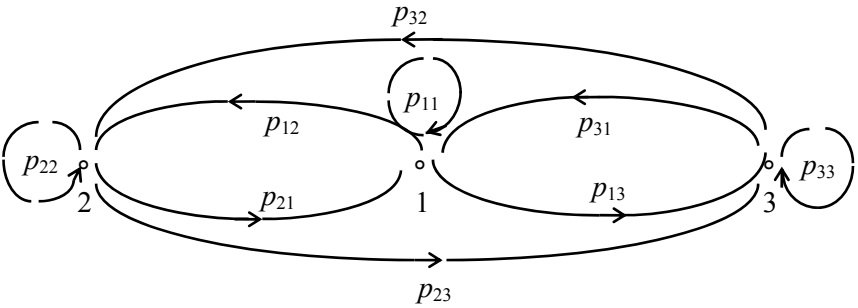


Рис. 8. 5. Граф состояний и переходов

Такое представление позволяет вычислять средние характеристики функционирования предприятия (среднее время простоя, среднее время работы).

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

9. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

9.1. Задачи математической статистики. Общие положения статистического оценивания

Теория вероятностей устанавливает закономерности, которым подчиняются массовые случайные явления, математическая статистика же определяет количественные характеристики этих закономерностей. В соответствии с этим основными задачами математической статистики являются:

– разработка способов сбора и группировки результатов наблюдения за случайными явлениями;

– обработка статистических данных.

Решая эти задачи, определяем виды законов распределения, параметры этих законов.

Пусть нас интересует некоторая случайная величина X. Для определения вида ее закона распределения или параметров этого закона проводятся наблюдения (эксперименты) за случайной величиной X. Результатом этих наблюдений является получение совокупности значений X: $x_1, x_2, x_3, ..., x_n$. Данная совокупность называется выборочной совокупностью или просто выборкой. Например, изучается качество товара. Из партии этого товара случайным образом выбирается некоторое количество образцов и определяется качество каждого образца. Совокупность результатов измерений и составляет выборочную совокупность.

Полная совокупность значений случайной величины называется генеральной совокупностью. Обычно из-за физических или экономических причин (партия объектов очень велика, процесс определения качества объектов разрушающий) нет возможности исследовать генеральную совокупность, а приходится оперировать лишь выборочной совокупностью. При этом свойства выборочной совокупности переносятся на генеральную совокупность. Правомочность такого переноса, как правило, справедлива при соблюдении некоторых правил при формировании выборочной совокупности. Выборка, которая обладает всеми свойствами генеральной совокупности называется представительной (репрезентативной).

9.2. Простой статистический ряд. Статистическая функция распределения

9.2.1. Простой статистический ряд

Пусть требуется определить вероятностные характеристики случайной величины X по результатам наблюдений ее значений в ряде опытов. Под вероятностными характеристиками будем понимать функцию распределения (интегральную или дифференциальную) и числовые характеристики.

Результаты наблюдения за случайной величиной X сводятся в таблицу, которая называется простая статистическая совокупность или простой статистический ряд (табл. 9.1).

Таблица 9.1

Простой статистический ряд

Номер наблюдения	1	2	3	…	n
Значение наблюдаемой величины	x_1	x_2	x_3	…	x_n

Если в простом статистическом ряду значения случайной величины разместить в порядке возрастания, то такой ряд называется вариационным. Первичный статистический ряд является начальной формой представления статистической информации. Из него могут быть получены более сложные характеристики случайной величины, такие как статистический ряд, гистограмма, статистическая функция распределения, числовые характеристики. Рассмотрим расчет этих характеристик.

9.2.2. Статистическая функция распределения

Статистической (эмпирической) функцией распределения случайной величины X называется частота события $X < x$ в данном статистическом материале:

$$F^*(x) = \frac{n_x}{n}, \tag{9.1}$$

где n – число наблюдений (объем выборки);

n_x – число значений X, оказавшихся меньше x.

Функция $F^*(x)$ имеет вид ступенчатой кривой (рис. 9.1) независимо от того непрерывная или дискретная случайная величина X.

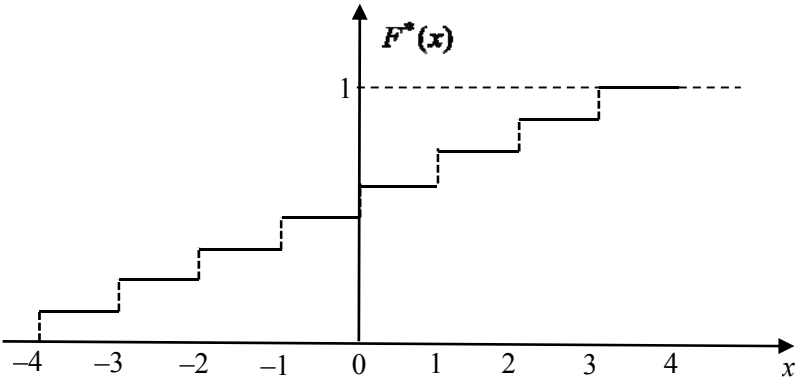


Рис. 9.1. Статистическая функция распределения

Эмпирическая функция распределения носит случайный характер. В отличие от нее функция распределения генеральной совокупности называется теоретической функцией распределения $F(x)$. Различие этих функций заключается в том, что теоретическая функция распределения $F(x)$ – есть вероятность $P(X < x)$, а эмпирическая $F^*(x)$ – есть частота события $X < x$. Согласно теореме Бернулли с ростом n частота $\frac{n_x}{n}$ сходится по вероятности к $P(X < x)$, т. е. $F^*(x)$ стремится к $F(x)$, или

$$\lim_{n\to\infty} P[|F(x) - F^*(x)| < \varepsilon] = 1,$$

где ε – сколь угодно малое положительное число.

Кстати, функция $F^*(x)$ обладает всеми свойствами функции $F(x)$:

$1.\, 0 \le F^*(x) \le 1.$

2. $F^*(x)$ – неубывающая функция аргумента x.

3. $F^*(x) = 1, \ F^*(x) = 0.$

9.3. Статистический ряд. Гистограмма

При большом числе наблюдений (более сотни) простой статистический ряд неудобен для использования. Для придания компактности первичному статистическому материалу строится статистический ряд (статистический интервальный ряд).

Пусть имеется n значений случайной величины X в виде простого статистического ряда. Расположим значения x_i в порядке возрастания. Разобьем диапазон значений x_i на k интервалов и подсчитаем количество значений x_i (обозначим его через m_i), попавших в i-й интервал, а также частоту попадания $P_i^* = \dfrac{m_i}{n}$. Статистический ряд будет иметь следующий вид

Таблица 9.2

Статистический ряд

Интервал значений					
Частота попадания					

Ширина каждого интервала, как правило, одинакова, количество выбирается из компромиссных соображений обеспечения компактности ряда и минимального загрубления первичной статистической информации за счет группировки.

Графически статистический ряд может быть оформлен в виде гистограммы. По горизонтальной оси откладываются значения случайной величины, разбитые на интервалы. На каждом интервале строится прямоугольник, высота которого равна эмпирической частоте, деленной на ширину интервала $H_i = \dfrac{p_i^*}{I_i}$. Площадь гистограммы равна 1. Гистограмма изображена на рис. 9.2.

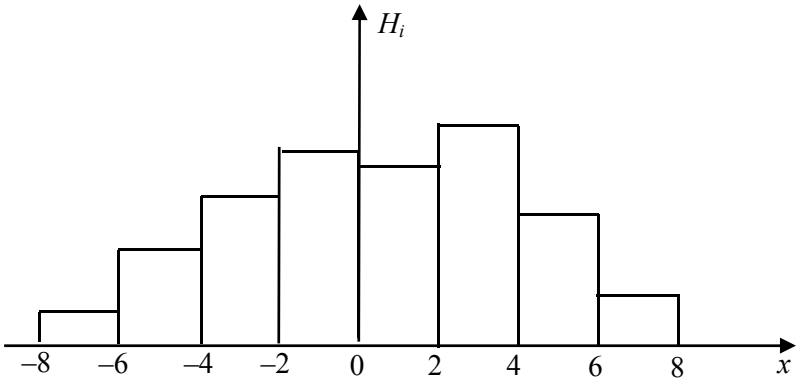


Рис. 9.2. Гистограмма

Пример 9.1. При обследовании птицефабрики получены следующие данные. Вес гусей равнялся:

– в интервале 3,0–3,5 кг – 8 шт.;

– 3,5–4,0 кг – 13 шт.;

– 4,0–4,5 кг – 7 шт.;

– 4,5–5,0 кг – 2 шт.

Необходимо построить статистический ряд и гистограмму распределения.

Решение

1. Рассчитаем частоты $P_i^* = \dfrac{m_i}{n}$

$$n = 8 + 13 + 7 + 2 = 30,$$

$$P_1^* = \frac{8}{30}, \quad P_2^* = \frac{13}{30}, \quad P_3^* = \frac{7}{30}, \quad P_4^* = \frac{2}{30}.$$

2. Составим таблицу статистического ряда

Интервал значений	3,0–3,5	3,5–4,0	4,0–4,5	4,5–5,0
Частота	8/30	13/30	7/30	2/30

3. Построим гистограмму
Рассчитаем значения

$$H_i = \frac{P_i^*}{I_i},$$

где I_i – ширина интервала, равная 0.5.

$$H_1 = \frac{8 \cdot 2}{30} = \frac{16}{30},\ H_2 = \frac{13 \cdot 2}{30} = \frac{26}{30},\ H_3 = \frac{7 \cdot 2}{30} = \frac{14}{30},\ H_4 = \frac{2 \cdot 2}{30} = \frac{4}{30}.$$

Гистограмма имеет вид

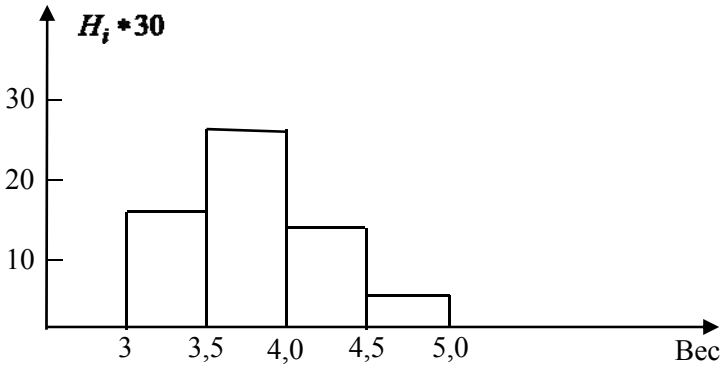


Рис. 9.3. Гистограмма

9.4. Статистическая точечная оценка параметров распределения случайной величины

Часто возникает задача определения значений числовых характеристик случайной величины, которые в свою очередь являются параметрами функции распределения, по статистическим данным. Т. е. необходимо по выборке $x_1, x_2, x_3, \ldots, x_n$ определить статистические значения числовых характеристик, которые в математической статистике получили название оценок.

9.4.1. Требования, предъявляемые к точечным оценкам

Естественно, что оценка, определяемая по статистическим данным, должна быть как можно ближе к истинному значению оцениваемого параметра (числовой характеристики). Однако понятие «ближе» не такое простое, как кажется на первый взгляд. Обозначим оцениваемый параметр θ, а его оценку θ^*. Если произвести оценивание k раз, то получим $\theta_1^*, \theta_2^*, \theta_3^*, ..., \theta_k^*$, где θ_i^* – оценка θ в i-й серии. Величина θ_i^* – случайная величина, определяемая по выборке конечного объема $x_{1i}, x_{2i}, x_{3i}, ..., x_{ni}$, где x_{ji} – j-е значение случайной величины X в i-й серии испытаний. Может получиться, что рассчитанная оценка $\theta_i^* = \varphi(x_{1i}, x_{2i}, x_{3i}, ..., x_{ni})$ будет иметь систематическое отклонение от θ (рис. 9.4).

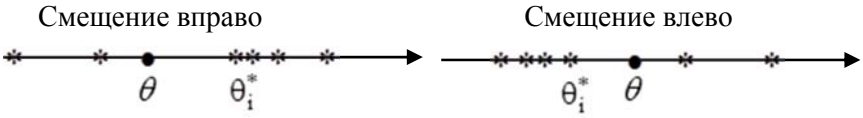


Рис. 9.4. Смещение оценки

Желательно, чтобы ожидаемое значение θ^* в среднем равнялось θ, т. е. $M[\theta^*] = \theta$. Оценки, у которых математическое ожидание равняется истинному значению оцениваемого параметра, называются несмещенными. Несмещенность характеризует отсутствие систематической ошибки оценивания.

Кроме несмещенности, требуется, чтобы оценка находилась в среднем как можно ближе к оцениваемому параметру. Это свойство характеризуется величиной разброса возможных ее значений (рис. 9.5).

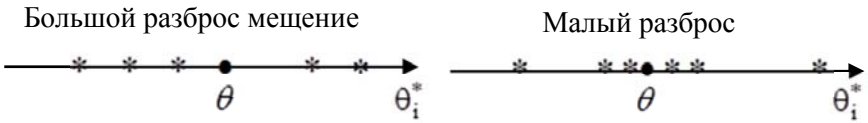


Рис. 9.5. Разброс оценок

Величина случайного разброса оценок зависит от их дисперсии. Следовательно, дисперсия оценки должна быть минимально возможной. Оценка, у которой $D[\theta^*] = D_{\min}$ при данном объеме выборки, называется эффективной. Эффективность оценки характеризует случайную составляющую ошибки оценивания. Кроме точности оценивания (несмещенности и эффективности), оценка с ростом объема выборки должна стремиться к истинному значению параметра. Это свойство называется состоятельностью. Оценки, которые при $n \to \infty$ стремятся по вероятности к оцениваемому параметру, называются состоятельными.

9.4.2. Оценка математического ожидания, дисперсии, корреляционного момента

Существует ряд методов построения точечных оценок числовых характеристик случайных величин. Воспользуемся наиболее распространенным – методом выборочных моментов. Согласно этому методу оценка начального момента *k*-го порядка равна

$$\mu_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k,$$

где n – объем выборки;

x_i – значение случайной величины X.

Т. е. оценка *k*-го момента есть среднее арифметическое *k*-х степеней выборочных значений.

Оценка математического ожидания (начального момента 1-го порядка) равна

$$\mu_1^* = m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Убедимся, что m_x^* является несмещенной, эффективной и состоятельной оценкой m_x:

1. $M[m_x^*] = M[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} M[x_i] = \frac{1}{n}\sum_{i=1}^{n} m_x = \frac{1}{n} n m_x = m_x.$

2. $D[m_x^*] = D[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n^2}\sum_{i=1}^{n} D[x_i] = \frac{1}{n^2} nD_x = \frac{1}{n} D_x.$

3. $\lim\limits_{n\to\infty} P(\left|m_x^* - m_x\right| < \varepsilon) = \lim\limits_{n\to\infty} P(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - m_x\right| < \varepsilon) = 1.$

Из приведенных выше расчетов следует, что оценка m_x^* – несмещенная ($M[m_x^*] = m_x$), эффективная ($D[m_x^*] = \frac{1}{n} D_x$ – минимально возможная), состоятельная.

Оценка дисперсии согласно методу выборочных моментов равна оценке центрального момента второго порядка

$$D_x^* = \alpha_2^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - m_x^*)^2. \qquad (9.2)$$

Исследование свойств D_x^* показывает, что эта оценка является смещенной. Действительно, можно показать, что

$$M[D_x^*] = \frac{1}{n}\sum_{i=1}^{n} M[x_i - m_x^*)^2] = \frac{n-1}{n} D_x.$$

Однако, если от D_x^* перейти к оценке $\hat{D}_x$, равной

$$\hat{D}_x = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} (x_i - m_x^*)^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_x^*)^2, \qquad (9.3)$$

то оценка $\hat{D}_x$ – несмещенная и поэтому при малых *n* целесообразно использовать эту оценку. Можно показать, что оценка (9.3) является асимптотически (т. е. при больших значениях *n*) эффективной и состоятельной.

В качестве оценки корреляционного момента используется выборочный смешанный момент

$$K_{xy}^* = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_x^*)(y_i - m_y^*). \qquad (9.4)$$

9.5. Интервальное оценивание числовых характеристик

Часто необходимо найти не только точечную оценку неизвестного параметра закона распределения, но и определить к каким ошибкам может привести замена θ на θ^*, определить вероятность того, что ошибка не превысит некоторую величину. Такая задача особенно актуальна для выборок малого объема, когда точечная оценка вследствие случайного разброса может в каждом конкретном случае сильно отличаться от истинного значения оцениваемого параметра.

Чтобы получить представление о точности и надежности оценки, используются понятия доверительный интервал, доверительная вероятность. Задача интервального оценивания формируется следующим образом.

Если имеется выборка $x_1, x_2, x_3, ..., x_n$, то необходимо найти такие границы интервала $\bar{\theta} = f(x_1, x_2, x_3, ..., x_n)$, $\underline{\theta} = f(x_1, x_2, x_3, ..., x_n)$, чтобы интервал $[\underline{\theta}, \bar{\theta}]$ накрывал истинное значение параметра θ с вероятностью не менее заданной. Последняя называется доверительной. Графически это представлено на рис. 9.6.

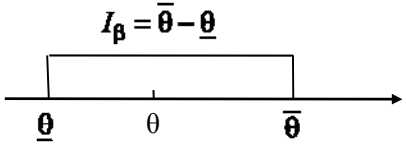


Рис. 9.6. Доверительный интервал:
$I_\beta = \bar{\theta} - \underline{\theta}$ – доверительный интервал; β – доверительная вероятность

Связь между доверительным интервалом и доверительной вероятностью имеет вид

$$P(\underline{\theta} \le \theta \le \bar{\theta}) = \beta.$$

Интервальная оценка должна строиться таким образом, чтобы удовлетворялись следующие требования: величина I_β должна быть

минимально возможной; с увеличением объема выборки величина I_β должна уменьшаться, границы интервала должны стремиться к 0; формулы для расчета $\underline{\theta}$, $\overline{\theta}$ должны быть простыми.

Существует ряд методов построения интервальных оценок для числовых характеристик. Рассмотрим построение интервальных оценок, основанное на неравенстве Чебышева.

9.5.1. Интервальная оценка для математического ожидания

Пусть имеется выборка $x_1, x_2, x_3, ..., x_n$ случайной величины X, математическое ожидание m_x которой неизвестно. Точечные оценки математического ожидания и дисперсии рассчитываются по формулам

$$m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad D_x^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2.$$

Будем считать, что m_x^* имеет нормальный закон распределения, т. к. это сумма независимых одинаково распределенных случайных величин (при $n > 10$ это предположение достаточно справедливо). Мы уже знаем, что

$$M[m_x^*] = m_x, \quad D[m_x^*] = \frac{D_x}{n}.$$

Найдем такую величину ε_β, чтобы $P(\left|m_x^* - m_x\right| < \varepsilon_\beta) \geq \beta.$

Неравенство $\left|m_x^* - m_x\right| < \varepsilon_\beta$ эквивалентно двойному неравенству $m_x - \varepsilon_\beta < m_x^* < m_x + \varepsilon_\beta$, тогда имеем

$$P(m_x - \varepsilon_\beta < m_x^* < m_x + \varepsilon_\beta) \geq \beta. \qquad (9.5)$$

Так как m_x^* имеет нормальный закон распределения, то

$$P(m_x - \varepsilon_\beta < m_x^* < m_x + \varepsilon) = F(m_x + \varepsilon_\beta) - F(m_x - \varepsilon_\beta) =$$
$$= \Phi(\frac{m_x + \varepsilon_\beta - m_x}{\sigma(m_x^*)}) - \Phi(\frac{m_x - \varepsilon_\beta - m_x}{\sigma(m_x^*)}) = \Phi(\frac{\varepsilon_\beta}{\sigma(m_x^*)}),$$

где $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$, $\sigma(m_x^*) = \frac{\sigma_x}{\sqrt{n}}$.

Переходя в (9.5) к равенству, имеем уравнение $\Phi(\frac{\varepsilon_\beta}{\sigma(m_x^*)}) = \beta$ или $\Phi(\frac{\varepsilon_\beta}{\sigma_x}\sqrt{n}) = \beta$. Обозначим $\frac{\varepsilon_\beta \sqrt{n}}{\sigma_x} = t_\beta$, тогда $\Phi\left(t_\beta\right) = \beta$. Откуда $t_\beta = \Phi^{-1}(\beta)$, где $\Phi^{-1}(\cdot)$ означает функцию, обратную $\Phi(\cdot)$. Поскольку $\varepsilon_\beta = t_\beta \frac{\sigma_x}{\sqrt{n}}$, то окончательно имеем

$$\underline{\theta} = m_x^* - \varepsilon_\beta = m_x^* - t_\beta \frac{\sigma_x}{\sqrt{n}}, \qquad \bar{\theta} = m_x^* + \varepsilon_\beta = m_x^* + t_\beta \frac{\sigma_x}{\sqrt{n}}, \qquad (9.6)$$

где t_β находится по таблицам функции $\Phi(\cdot)$ при заданном значении β.

Формулы (9.6) обеспечивают точный расчет, если известна дисперсия σ_x^2. При неизвестной дисперсии σ_x заменяется на ее точечную оценку

$$\sigma_x \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2} \qquad (9.7)$$

и расчеты являются приближенными.

9.5.2. Доверительный интервал для дисперсии

Точечная оценка $D_x^* = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2$ подчинена χ^2-распределению (хи-квадрат). Можно показать, что нижняя и верхняя границы для дисперсии равны

$$\underline{D}_x = \frac{D_x^*(n-1)}{\chi_1^2}, \qquad \bar{D}_x = \frac{D_x^*(n-1)}{\chi_2^2},$$

где χ^2 – квантиль хи-квадрат распределения с $n-1$ степенью свободы, который находится по соответствующим таблицам.

Квантиль χ_1^2 – соответствует вероятности $\dfrac{1-\beta}{2}$, χ_2^2 соответствует вероятности $\dfrac{1+\beta}{2}$.

Пример 9.2. При статистическом определении денежных доходов рабочих одного из предприятий было опрошено 10 человек. Необходимо составить простой статистический ряд, определить оценки математического ожидания и дисперсии заработной платы рабочих, а также получить интервальную оценку для математического ожидания с доверительной вероятностью $\beta = 0,9$.

Решение

1. Составим простой статистический ряд

Номер человека	1	2	3	4	5	6	7	8	9	10
Доход *x* (руб)	880	790	910	930	810	750	980	700	810	930

2. Рассчитаем оценки математического ожидания и дисперсии заработной платы в соответствии с (9.1), (9.3)

$$m_x^* = \frac{1}{10}(880 + 790 + 910 + 930 + 810 + 750 + 980 + 700 + $$
$$+810 + 930) = 849 \text{ руб.}$$

$$D_x^* = \frac{1}{9}(31^2 + 59^2 + 61^2 + 81^2 + 39^2 + 99^2 + 131^2 +$$

$$+149^2 + 29^2 + 81^2) = 8090 \text{ руб}^2.$$

3. Определим границы интервальной оценки

Среднеквадратическое отклонение случайной величины x в соответствии с (9.7) равно

$$\sigma_x \approx \sqrt{D_x^*} = 89,9 \text{ руб.}$$

По таблицам функции $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$ находим $t_\beta = \Phi^{-1}(0,9) = 1,64.$

Интервальные границы равны

$$\underline{m}_x = m_x^* - t_\beta \frac{\sigma_x}{\sqrt{n}} = 849 - 1,64 \frac{89,9}{\sqrt{10}} = 803,9 \text{ руб,}$$

$$\overline{m}_x = m_x^* + t_\beta \frac{\sigma_x}{\sqrt{n}} = 849 + 1,64 \frac{89,9}{\sqrt{10}} = 895,9 \text{ руб.}$$

10. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

10.1. Общие сведения о проверке статистических гипотез

Проверка предположения (гипотезы) о виде закона распределения случайной величины, о значениях ее числовых характеристик по имеющимся статистическим данным называется проверкой статистических гипотез.

Статистическая гипотеза H может быть простой или сложной. Простая гипотеза – это гипотеза однозначная, типа $m_x = 1$. Сложная – многозначная, типа $m_x < 1$, $m_x \neq 1$ и т. д. Проверяемая гипотеза называется нулевой и обозначается H_0. Альтернативная (конкурирующая) обозначается H_1. Условно гипотезы обозначаются следующим образом: H_0: $m_x = a$ или H_1: $m_x \neq a$.

Правило, по которому принимается решение принять или отклонить H_0, называется статистическим критерием. Проверка статистических гипотез осуществляется на основе выборочных значений случайной величины. Строится некоторая функция от выборочных значений, рассчитывается ее значение для конкретных наблюденных значений и согласно статистическому критерию сравнивается с некоторым порогом. В зависимости от результата сравнения гипотеза H_0 принимается или отвергается. Поскольку указанная процедура носит статистический характер, то при проверке гипотезы возможны ошибки двух видов: допускается ошибка первого рода, при которой гипотеза H_0 отвергается, хотя она верна; допускается ошибка второго рода, при которой гипотеза H_0 принимается, хотя она не верна. Ошибки первого и второго рода измеряются в вероятностях. Вероятность ошибки первого рода обозначается через α и называется уровнем значимости. Вероятность недопущения ошибки второго рода обозначается через β и называется мощностью критерия.

Обычно можно построить несколько статистических критериев, использование которых обеспечит требуемое значение α, но при этом вероятность ошибки второго рода $1-\beta$ может быть разной. Из этих критериев выбирают тот, который обеспечивает наименьшую

величину $1-\beta$ (наибольшее значение β), т. е. выбирается наиболее мощный критерий.

Если распределение случайной величины известно, а по выборке необходимо проверить гипотезу о значении числовой характеристики, то такая гипотеза называется параметрической, критерий – параметрическим. Критерий, не использующий знание закона распределения, называется непараметрическим.

10.2. Проверка гипотезы о среднем значении случайной величины при известной дисперсии

Пусть известно, что случайная величина X имеет нормальный закон распределения с неизвестным математическим ожиданием и известной дисперсией σ_x^2. В результате наблюдения за величиной X получена статистическая выборка $(x_1, x_2, \ldots, x_n)$. Требуется проверить гипотезу H_0: $m_x = a$, против H_1: $m_x = a_1$, если $a_1 > a$.

Перейдем от случайной величины X к некоторой нормированной случайной величине Z, имеющей: $M[Z]=0$, $D[Z]=1$, если H_0 верна. Такой величиной является величина вида

$$Z = \frac{m_x^* - a}{\sigma_x / \sqrt{n}},$$

где $m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Действительно, если гипотеза H_0: $m_x = a$ верна, то Z имеет следующие числовые характеристики:

$$M[Z] = \frac{M[m_x^*] - a}{\sigma_x / \sqrt{n}} = \frac{m_x - a}{\sigma_x / \sqrt{n}} = 0, \text{ т. к. } m_x = a;$$

$$D[Z] = \frac{D[m_x^*] - 0}{\sigma_x^2 / n} = \frac{D_x / n}{\sigma_x^2 / n} = 1, \text{ т. к. } D_x = \sigma_x^2.$$

Таким образом, при $m_x = a$ закон распределения Z нормальный с параметрами $m_z = 0,\ D_z = 1,$ т. е. $f_{H_0}(z) = N_0(0,1).$

Зададимся вероятностью ошибки первого рода α. Как правило, эта величина должна быть небольшой (0,01; 0,05; 0,1). Построим статистический критерий для проверки гипотезы, исходя следующих рассуждений. Если гипотеза H_0 справедлива, то вероятнее всего, что значения Z, будут недалеко от 0. Если значение Z оказывается далеко от 0, то возможно, что гипотеза H_0 неверна и ее следует отвергнуть. Таким образом, критерий сводится к назначению порогового уровня сравнения с ним наблюденного значения z_0. Если $Z \geq z_0$, то H_0 отвергается, если $Z < z_0,$ то H_0 принимается. Так как причиной того, что Z оказалась больше $z_0,$ может оказаться не тот факт, что H_0 не верна, а просто случайный разброс величины Z, то используя полученный критерий, мы допускаем ошибку первого рода: при $Z \geq z_0$ отвергаем гипотезу $H_0,$ хотя она верна. Вероятность этой ошибки равна $P_{H_0}(Z \geq z_0)$. Поскольку эта вероятность не должна превышать принятую величину $\alpha,$ то можно получить уравнение, решая которое определим z_0. Таким образом, имеем

$$P_{H_0}(Z \geq z_0) = \alpha.$$

или переходя к противоположному событию, получим

$$P_{H_0}(Z < z_0) = 1 - \alpha. \qquad (10.1)$$

Из последнего выражения видно, что z_0 есть квантиль нормального распределения уровня $1 - \alpha,$ т. е. $z_0 = z_{1-\alpha}$. Преобразуем уравнение (10.1)

$$P_{H_0}(Z < z_0) = F_{H_0}(z_{1-\alpha}) = 1/2[1 + \Phi(\frac{z_{1-\alpha} - m_z}{\sigma_z})\} =$$
$$= 1/2[1 + \Phi(z_{1-\alpha})] = 1 - \alpha$$

или $\Phi(z_{1-\alpha}) = 1 - 2\alpha,$ откуда

$$z_{1-\alpha} = \Phi^{-1}(1 - 2\alpha), \qquad (10.2)$$

где $\Phi^{-1}(*)$ – функция, обратная функции Φ(*).

Если верна гипотеза $H_1,$ то имеем

$$M[Z] = \frac{M[m_x^*] - a}{\sigma_x / \sqrt{n}} = \frac{a_1 - a}{\sigma_x / \sqrt{n}},$$

$$D[Z] = \frac{D[m_x^*] - 0}{\sigma_x^2 / n} = \frac{D_x / n}{\sigma_x^2 / n} = 1.$$

Таким образом, при $m_x = a_1$ закон распределения Z – нормальный с параметрами $m_x = \frac{a_1 - a}{\sigma_x / \sqrt{n}},$ $D_z = 1,$ т. е. $f_{H_1}(z) = N(\frac{a_1 - a}{\sigma_x / \sqrt{n}}, 1).$ Так как $a_1 > a,$ то $\frac{a_1 - a}{\sigma_x / \sqrt{n}} > 0.$ Поскольку мы выбрали в качестве порогового значения величину $z_0 = Z_{1-\alpha},$ то может оказаться, что наблюденное значение Z меньше $z_0,$ но верна гипотеза $H_1,$ а мы должны принять гипотезу $H_0.$ При этом совершается ошибка второго рода, вероятность которой равна

$$P_{H_1}(Z < z_{1-\alpha}) = F_{H_1}(z_{1-\alpha}) = 1/2[1 + \Phi(\frac{z_{1-\alpha} - m_z}{\sigma_z})] =$$

$$= 1/2[1 + \Phi(z_{1-\alpha} - \frac{a_1 - a}{\sigma_x / \sqrt{n}})] = 1 - \beta.$$

Откуда $\beta = 1/2[1 - \Phi(z_{1-\alpha} - \frac{a_1 - a}{\sigma_x / \sqrt{n}})].$

На рис. 10.1 изображены плотности вероятности f_{H_0}, f_{H_1}. Заштрихованные области показывают величины α и $1-\beta$ при выбранном значении $z_0 = z_{1-\alpha}$: чем больше $z_{1-\alpha}$, тем меньше α и тем больше $1-\beta$. Чем больше a_1 отличается от a, тем меньше ошибка второго рода $1-\beta$. Чем меньше дисперсия случайной величины X, тем меньшую величину $1-\beta$ можно обеспечить при выбранном значении α.

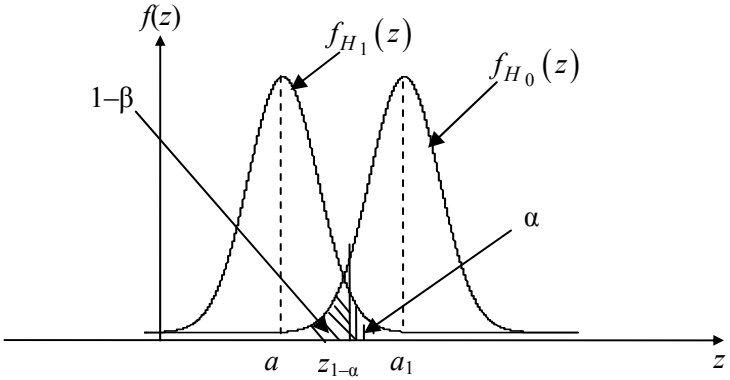


Рис. 10.1. Выбор порогового значения $z_{1-\alpha}$ для $a_1 > a$

Пусть альтернативная гипотеза H_1: $m_x = a_1$, где $a_1 < a$. В этом случае закон распределения величины Z при справедливости гипотезы H_0 остается таким же: $f_{H_0}(z) = N(0,1)$, но меняется статистический критерий. Гипотеза H_0 отвергается, если $Z \le z_0$, и принимается, если $Z > z_0$. Ошибка первого рода равна,

$$P_{H_0}\left(Z \le z_0\right) = F_{H_0}\left(z_\alpha\right) = \frac{1}{2}\left[1 + \Phi\left(\frac{z_\alpha - m_z}{\sigma_z}\right)\right] = \frac{1}{2}\left[1 + \Phi\left(z_\alpha\right)\right] = \alpha,$$

откуда $z_\alpha = \Phi^{-1}(2\alpha - 1)$.

Если верна гипотеза H_1, то имеем

$$M[Z] = -\frac{a - a_1}{\sigma_x / \sqrt{n}} < 0, \quad D[Z] = 1.$$

Ошибка второго рода при этом равна

$$P_{H_1}(Z > z_\alpha) = 1 - P_{H_1}(Z \le z_\alpha) = 1 - F_{H_1}(z_\alpha) =$$
$$= 1 - \frac{1}{2}[1 + \Phi(z_\alpha + \frac{a - a_1}{\sigma_x / \sqrt{n}})] = 1 - \beta,$$

откуда

$$\beta = \frac{1}{2}[1 + \Phi(z_\alpha + \frac{a - a_1}{\sigma_x / \sqrt{n}})].$$

Полученные выражения поясняются рис. 10.2.

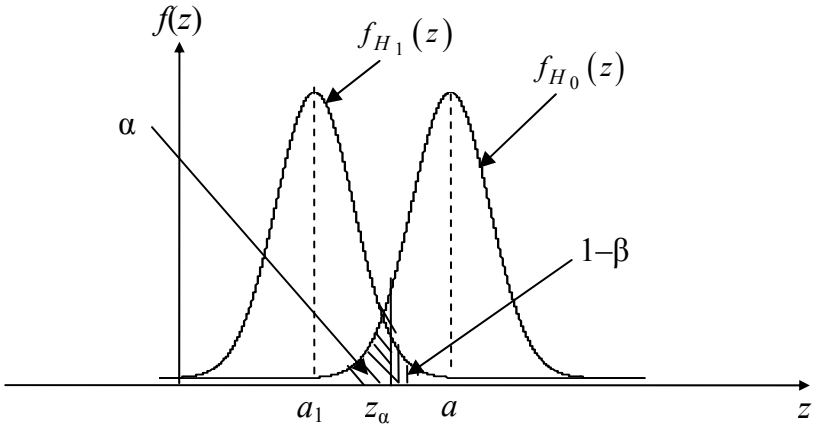


Рис. 10.2. Выбор порогового значения z_α для $a_1 < a$

10.3. Проверка гипотезы о виде закона распределения

Если в результате опытов над случайной величиной X построена эмпирическая функция ее распределения $F^*(x)$, то часто возникает задача подобрать такую функцию $F(x)$, которая наилучшим образом в некотором смысле соответствует $F^*(x)$.

Такая функция в дальнейшем будет называться теоретической функцией распределения, а задача ее отыскания формулируется следующим образом.

Имеется случайная величина X закон распределения $F(x)$ которой неизвестен. Получен ряд значений X, по которым построена эмпирическая функция распределения $F^*(x)$. Выдвигается гипотеза о виде закона распределения $F(x)$, следует ее проверить с использованием $F^*(x)$.

Наиболее простым и эффективным методом проверки такой гипотезы является метод, использующий критерий Колмогорова. При этом критерии за меру согласия эмпирического и теоретического распределений принимается величина

$$D_n = \max_x \left| F^*(x) - F(x) \right|,$$

где D_n – случайная величина, равная максимальному отклонению эмпирической функции от теоретической (рис. 10.3) и зависящая от количества n.

Колмогоров показал, что если перейти к случайной величине $\lambda_n = D_n\sqrt{n}$, то закон распределения последней с ростом n, независимо от вида закона распределения X, сходится к некоторой функции $K(\lambda)$. Эта функция получила название функции Колмогорова.

Если отклонение λ_n велико, то следует выдвинутую гипотезу H_0 отвергнуть, если мало – принять. При этом естественно существование вероятности ошибки первого рода. Пороговое значение λ_0 найдем из уравнения

$$P(\lambda_n \geq \lambda_0) = \alpha.$$

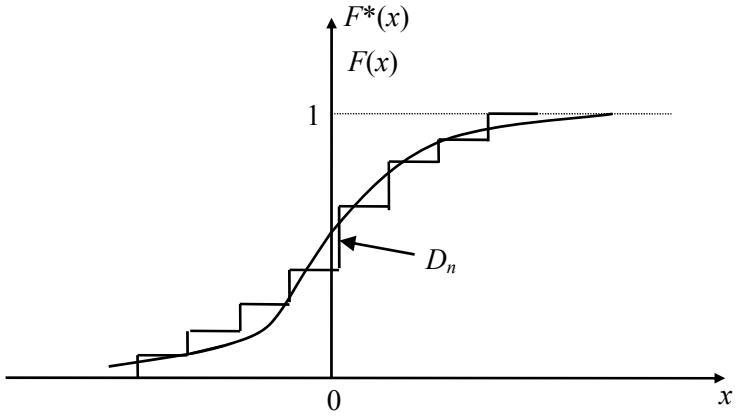


Рис 10.3. Определение D_n

Так как величина λ_n имеет распределение $K(\lambda)$, то

$$P(\lambda_n \geq \lambda_0) = 1 - K(\lambda_0) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{k^2 \lambda_0^2}.$$

Значения функции $P(\lambda_n \geq \lambda_0)$ сведены в таблицу, по которой на основании величины α можно определить значение $\lambda_0 = \lambda_\alpha$. В табл. 10.1 приведены значения квантилей распределения Колмогорова.

Таблица 10.1

Квантили распределения Колмогорова

Уровень значимости α	0,01	0,05	0,1
Квантиль распределения λ_0	1,627	1,358	1,224

Сравнивая λ_n^* (наблюденное значение λ_n) с λ_α, принимаем или отвергаем гипотезу H_0.

Пример 10.1. Для определения сорта автомобильных шин по критерию долговечности проводились выборочные испытания шин на износ. Для проведения испытаний было случайным образом отобрано 100 шин. В качестве показателя долговечности использовался пробег шины до наступления предельного износа. Норма пробега для шин первого сорта установлена равной 80 тыс. км, а для второго – 70 тыс. км. По результатам испытаний среднее значение долговечности шин оказалось равным $m_x^* = 75$ тыс. км, а среднее квадратическое отклонение $\sigma_x = 40$ тыс. км. Уровень значимости был выбран $\alpha = 0,05$.

Решение

В качестве нулевой гипотезы выдвинем гипотезу H_0: $m_x = a = 70$ тыс. км, а в качестве альтернативной – H_1: $m_x = a_1 = 80$ тыс. км. Перейдем к нормированной величине *Z*, равной

$$Z = \frac{m_x^* - a}{\sigma_x / \sqrt{n}} = \frac{75 - 70}{40 / \sqrt{100}} = 1,25.$$

Определим величину порогового значения z_0. Для этого необходимо использовать таблицу значений функции Ф(*) (табл. П1). При этом входом в таблицу является значение функции, выходом – значение аргумента

$$z_0 = z_{1-\alpha} = \Phi^{-1}(1 - 2\alpha) = \Phi^{-1}(0,9) = 1,645.$$

Сравнивая Z с z_0 видим, что $Z < z_0$. Следовательно, гипотеза H_0 принимается, т. е. принимается решение, что партия шин соответствует второму сорту. При этом ошибка второго рода равна

$$1 - \beta = \frac{1}{2}[1 + \Phi(z_{1-\alpha} - \frac{a_1 - a}{\sigma_x / \sqrt{n}})] = \frac{1}{2}[1 + \Phi(1,645 - \frac{80 - 70}{40 / \sqrt{100}})] =$$

$$= \frac{1}{2}[1 + \Phi(-0,855)] = \frac{1}{2}[1 - \Phi(0,855)] = 0,198.$$

Пример 10.2. Пользуясь критерием Колмогорова установить, согласуются ли данные по продолжительности телефонных разговоров с предположением о том, что длительность телефонного разговора является случайной величиной, распределенной по нормальному закону. Уровень значимости принять равным 0,01. Статистические данные о продолжительности 100 разговоров сведены в табл. 10.2.

Решение

Рассчитаем эмпирическую функцию распределения случайной величины X (продолжительности разговора).

$$F^*(x) = P^*(X < x_i) = \frac{n_x}{n}.$$

Результаты расчета запишем в табл. 10.2.

Таблица 10.2

Продолжительность разговора x_i (мин)	1	2	3	4	5	6	7	8	9	10
Количество разговоров n_i	4	10	20	36	14	9	3	3	0	1
$F^*(x)$	0	0,04	0,14	0,34	0,7	0,84	0,93	0,96	0,99	0,99
$F(x)$	0,01	0,06	0,18	0,42	0,68	0,88	0,97	0,99	0,999	0,999
$\lvert F^*(x) - F(x)\rvert$	0,01	0,02	0,04	0,08	0,02	0,04	0,04	0,03	0,009	0,009

Для построения теоретической функции распределения $F(x)$, рассчитаем ее параметры, согласующиеся со статистическими данными:

$$m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i = 4,3 \text{ мин};$$

$$D_x^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2 = 2,1 \text{ мин}^2.$$

Функцию $F(x)$ рассчитаем с использованием функции Ф(*)

$$F(x) = 1/2(1 + \Phi(\frac{x - m_x^*}{\sqrt{D_x^*}})).$$

Рассчитаем разность функций $\left|F(x) - F^*(x)\right|$ и занесем ее в табл. 10.2. Из последней строки таблицы находим наибольший модуль разности

$$D_n = \max_x \left|F(x) - F^*(x)\right| = 0,08.$$

Этой величине соответствует $\lambda_n = D_n\sqrt{n} = 0,08\sqrt{100} = 0,8.$

По табл. 10.1 находим для $\alpha = 0,01;\ \lambda_0 = 1,627.$

Так как 1,627 > 0,08 т. е. $(\lambda_0 > \lambda_n)$, то можно сделать вывод, что гипотеза о нормальном законе распределения X согласуется с опытными данными.

11. ДИСПЕРСИОННЫЙ АНАЛИЗ

11.1. Основные понятия дисперсионного анализа

Часто возникает задача установления факта влияния некоторых факторов на величину, носящую случайный характер. Раздел математической статистики, решающий эту задачу, называется дисперсионным анализом. В зависимости от количества учитываемых факторов различают однофакторный, двухфакторный и многофакторный дисперсионный анализ. Под влиянием факторов будем понимать изменение среднего значения случайной величины. Это изменение в значительной мере маскируется воздействием неизвестных факторов, которые в дальнейшем будем называть неучтенными. В итоге дисперсия случайной величины формируется как под воздействием учитываемых, так и не учитываемых факторов. Можно считать, что при одном учитываемом факторе A справедливо

$$\sigma^2 = \sigma_A^2 + \sigma_O^2,$$

где $\sigma^2, \sigma_A^2, \sigma_O^2$ – суммарная дисперсия случайной величины, доля дисперсии, вызванная воздействием фактора A, доля дисперсии, вызванная воздействием неучтенных факторов, соответственно.

Если в результате наблюдений возможно определить значения $\sigma_A^2, \sigma_O^2,$ то можно принять решение о влиянии фактора A на случайную величину X. Например, если окажется, что $\sigma_O^2 < \sigma_A^2,$ то делается вывод о значимом влиянии фактора A. Поскольку точные значения σ_O^2, σ_A^2 получить невозможно, то в дисперсионном анализе оперируют с их оценками (выборочными дисперсиями), которые будем в дальнейшем обозначать через S_O^2, S_A^2. Естественно, что результат сравнения выборочных дисперсий носит случайный характер. Действительно, если имеем две выборочные дисперсии

$$S_O^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2, \quad S_A^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - m_y^*)^2,$$

то величина $F = \dfrac{S_A^2}{S_O^2}$ имеет распределение Фишера с γ_1, γ_2 степенями свободы $(\gamma_1 = n-1, \gamma_2 = m-1)$. Выбирая уровень значимости гипотезы о равенстве выборочных дисперсий $(F = 1)$, равным α, будем считать, что при $F > F_{1-\alpha,\gamma_1,\gamma_2}$ фактор A влияет на X, при $F \le F_{1-\alpha,\gamma_1,\gamma_2}$ – не влияет. В этих неравенствах $F_{1-\alpha,\gamma_1,\gamma_2}$ – есть квантиль распределения Фишера порядка $1-\alpha$. Таблица значений распределения Фишера приведена в табл. П2.

11.2. Однофакторный дисперсионный анализ

Для практического использования аппарата дисперсионного анализа необходимо так проводить эксперименты со случайной величиной X, чтобы была возможность определить значения S_O^2, S_A^2. Это становится возможным, если удается наблюдать значения X при различных уровнях фактора A.

Пусть фактор A за время наблюдения за случайной величиной принимает значения $A_1, A_2, ..., A_m$. Для каждого уровня фактора A проводится n измерений случайной величины X. Результаты наблюдений сводятся в табл. 11.1.

Таблица 11.1

Таблица для однофакторного дисперсионного анализа

Номер испытания	Уровень фактора A					
	A_1	A_2	. . .	A_j	. . .	A_m
1	x_{11}	x_{12}	. . .	x_{1j}	. . .	x_{1m}
2	x_{21}	x_{22}	. . .	x_{2j}	. . .	x_{2m}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
i	x_{i1}	x_{i2}	. . .	x_{ij}	. . .	x_{im}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
n	x_{n1}	x_{n2}	. . .	x_{nj}	. . .	x_{nm}
Групповые средние	$\overline{x}_1$	$\overline{x}_2$	. . .	$\overline{x}_j$	. . .	$\overline{x}_m$

В этой таблице x_{ij} – i-е значение случайной величины X при $A = = A_j$. Произведя усреднение по всем значениям X при фиксированном A, получим значения групповых средних

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

т. е. $\overline{x}_j$ – есть среднее значение X при $A = A_j$.

Усредняя по всем значениям x при всех значениях A, получим общее среднее

$$\overline{X} = \frac{1}{n+m}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}.$$

Выборочная дисперсия случайной величины X равна

$$S^2 = \frac{1}{nm-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{X})^2.$$

Эта величина характеризует разброс случайной величины X относительно общего среднего. Выборочная дисперсия групповой средней равна

$$S_A^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\overline{x}_j - \overline{X})^2,$$

она характеризует разброс среднего значения X при фиксированном значении A относительно общего среднего. Если фактор A не изменяет среднее групповое (т. е. он не влияет на X), то эта дисперсия равна 0.

Выборочная дисперсия случайной величины X при фиксированном значении A относительно группового среднего равна

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2.$$

Усредняя эту дисперсию по всем группам, получим

$$S_o^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\right) = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2.$$

Дисперсия S_o^2 – это средняя по всем группам дисперсия X относительно групповых средних, она не зависит от A и характеризует разброс X за счет неучтенных факторов.

Таким образом, решение о влиянии фактора A принимается по критерию

$$\frac{S_A^2}{S_o^2} = \frac{\frac{1}{m-1}Q_A}{\frac{1}{m(n-1)}Q_o} > F_{1-\alpha,m-1,m(n-1)},$$

где

$$Q_A = n\sum_{j=1}^{m}(\overline{x_j}-\overline{X})^2, \quad Q_o = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2,$$

$F_{1-\alpha,m-1,m(n-1)}$ – квантиль распределения Фишера со степенями свободы: m–1, $m(n$–1). Если последнее неравенство выполняется, то фактор A влияет, в противном случае – не влияет.

При практических расчетах пользоваться приведенными формулами неудобно. Целесообразно пользоваться следующей методикой.

1. Находим сумму квадратов всех наблюдений

$$Q_1 = \sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}^2. \qquad (11.1)$$

2. Находим среднюю сумму квадратов по столбцам

$$Q_2 = \frac{1}{n}\sum_{j=1}^{m}x_j^2, \qquad (11.2)$$

где $x_j = \sum_{i=1}^{n}x_{ij}$.

3. Находим средний квадрат суммы всех наблюдений

$$Q_3 = \frac{1}{nm}\left(\sum_{j=1}^{m} x_j\right)^2. \qquad (11.3)$$

4. Вычисляем выборочные дисперсии

$$S_0^2 = \frac{Q_1 - Q_2}{m(n-1)}, \quad S_A^2 = \frac{Q_2 - Q_3}{m-1}.$$

11.3. Двухфакторный дисперсионный анализ

Дисперсионный анализ наиболее эффективен при одновременном изучении влияния нескольких факторов, хотя расчеты при этом существенно усложняются.

Рассмотрим дисперсионный анализ влияния двух независимых факторов A и B на случайную величину X. Пусть фактор A имеет m уровней: $A_1, A_2, A_3, \ldots, A_m$, а фактор B – k уровней: $B_1, B_2, B_3, \ldots, B_k$. Пусть при каждом сочетании уровней факторов A_j B_i может быть получено n значений величины X. Результаты наблюдений сведем в табл. 11.2.

Таблица 11.2

Таблица для двухфакторного дисперсионного анализа

Уровни фактора B	Уровни фактора A						
	A_1	A_2	. . .	A_j	. . .	A_m	Итого
B_1	A_1	x_{12}	. . .	x_{1j}	. . .	x_{1m}	x'_1
B_2	x_{21}	x_{22}	. . .	x_{2j}	. . .	x_{2m}	x'_2
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
B_i	x_{i1}	x_{i2}	. . .	x_{ij}	. . .	x_{im}	x'_i
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
B_k	x_{k1}	x_{k2}	. . .	x_{kj}	. . .	x_{km}	x'_k
Итого	$\overline{x}_1$	$\overline{x}_2$	. . .	$\overline{x}_j$	. . .	$\overline{x}_m$	

В этой таблице для простоты полагается, что $n = 1$, т. е. для каждого сочетания факторов получено только одно наблюдение. Через $\overline{x}_j \left(j = 1, m \right)$ обозначена сумма элементов по j-му столбцу. Через $x'_i \left(i = 1, k \right)$ – сумма элементов по i-й строке, т. е.:

$$\overline{x}_j = \sum_{i=1}^{k} x_{ij}, \quad x'_i = \sum_{j=1}^{m} x_{ij}.$$

Выборочные дисперсии S_0^2, S_A^2, S_B^2 рассчитываются по формулам:

$$S_0^2 = \frac{Q_1 - Q_2 + Q_4 - Q_3}{(k-1)(m-1)}, \tag{11.4}$$

$$S_A^2 = \frac{Q_2 - Q_4}{m-1}, \quad S_B^2 = \frac{Q_3 - Q_4}{k-1}, \tag{11.5}$$

где

$$Q_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} x_{ij}^2, \quad Q_2 = \frac{1}{k} \sum_{j=1}^{m} x_j^2,$$
$$Q_3 = \frac{1}{m} \sum_{i=1}^{k} (x_i^1)^2, \quad Q_4 = \frac{1}{mk} (\sum_{j=1}^{m} x_j)^2. \tag{11.6}$$

Решение о влиянии факторов принимается по критерию Фишера:

$$\frac{S_A^2}{S_0^2} > F_{1-\alpha,\gamma_1,\gamma_2}, \quad \frac{S_B^2}{S_0^2} > F_{1-\alpha,\gamma_1',\gamma_2'},$$

где α – уровень значимости; $\gamma_1 = m-1$, $\gamma_2 = (k-1)(m-1)$, $\gamma_1' = k - -1$, $\gamma_2' = (k-1)(m-1)$.

Пример 11.1. В течение шести лет использовались пять различных технологий по выращиванию сельскохозяйственной культуры. Необходимо установить влияние различных технологий на урожай-

ность культуры. Статистические данные об урожайности (в ц/га) сведены в табл. 11.3.

Таблица 11.3

Статистические данные об урожайности (в ц/га)

Номер наблюдения (год)	Уровень фактора A				
	A_1	A_2	A_3	A_4	A_5
1	1,2	0,6	0,9	1,7	1,0
2	1,1	1,1	0,6	1,4	1,4
3	1,0	0,8	0,8	1,3	1,1
4	1,3	0,7	1,0	1,5	0,9
5	1,1	0,7	1,0	1,3	1,5
6	0,8	0,9	1,1	1,3	1,5
Итого	6,5	4,8	5,4	8,4	7,1

Решение

По формулам (11.1–11.3) вычисляем: $Q_1 = 37$; $Q_2 = 35{,}9$; $Q_3 = 34$. Выборочные дисперсии равны

$$S_0^2 = \frac{37 - 35{,}9}{5 \cdot 5} = 0{,}044; \quad S_A^2 = \frac{35{,}9 - 34{,}56}{4} = 0{,}325.$$

Сравним выборочные дисперсии по критерию Фишера:

$$F = \frac{S_A^2}{S_0^2} = \frac{0{,}085}{0{,}044} = 1{,}93.$$

Задавая уровень значимости критерия $\alpha = 0{,}05$, по таблице находим $F_{0{,}95;\,4;\,25} = 2{,}8$. Так как $F < F_{0{,}95;\,4;\,25}$, то влияние технологии незначительно.

Пример 11.2. В четырех странах анализировался уровень эмиграции в зависимости от уровня экономического развития страны и среднего уровня образования населения. Необходимо установить

влияние данных факторов. Обозначим фактор уровень экономического развития через A, а фактор уровень образования через B. Результаты обследования (уровень эмиграции в % от всего населения страны) сведем в табл. 11.4.

Таблица 11.4

Результаты обследования

Уровни фактора B	Уровни фактора A				
	A_1	A_2	A_3	A_4	x_i'
B_1	1,2	0,6	0,8	0,4	3
B_2	1,0	1,1	0,6	0,4	3,1
B_3	0,9	1,0	0,7	0,6	3,2
$\overline{x}_j$	3,1	2,7	2,1	1,4	

Решение

По формулам (11.6) вычисляем:

$$Q_1 = 1,44 + 0,365 + 0,64 + 0,36 + 1,0 + 1,21 + 0,36 + 0,16 + \\ + 0,81 + 1,0 + 0,49 + 0,14 = 7,99.$$

$$Q_2 = \frac{1}{3}(9,61 + 7,29 + 4,41 + 2,56) = 7,757.$$

$$Q_3 = \frac{1}{4}\left(10,24 + 9,61 + 10,24\right) = 7,213.$$

$$Q_4 = \frac{9,5^2}{3 \cdot 4} = 7,208.$$

Выборочные дисперсии равны:

$$S_0^2 = \frac{1}{3 \cdot 2}\left(8,19 + 7,9 - 7,95 - 7,52\right) = 0,038.$$

$$S_A^2 = \frac{7,95 - 7,52}{3} = 0,183.$$

$$S_B^2 = \frac{7,525 - 7,52}{2} = 0,002.$$

Отношения дисперсий равны:

$$F_A = \frac{S_A^2}{S_0^2} = \frac{0,145}{0,038} = 4,82.$$

$$F_B = \frac{S_B^2}{S_0^2} = \frac{0,0025}{0,0386} = 0,05.$$

$$\gamma_1 = 3,\ \gamma_2 = 6,\ \gamma_1' = 2,\ \gamma_2' = 6.$$

По таблице значений распределения Фишера при $\alpha = 0,95$ находим $F_{0,95;\,3;\,6} = 4,76;\ F_{0,95;\,2;\,6} = 5,14$. Поскольку $F_A > F_{0,95;\,3;\,6}$, $F_B < F_{0,95;\,2;\,6}$, то фактор A влияет на уровень эмиграции, а фактор B не влияет.

12. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

12.1. Основные понятия корреляционного и регрессионного анализа

Между детерминированными величинами может существовать функциональная связь $y = f(x)$, при которой каждому значению аргумента соответствует одно или несколько известных значений функции. Между случайными величинами может существовать стохастическая (вероятностная) связь. Стохастическая связь это такая связь, при которой закон распределения одной случайной величины зависит от того какое значение приняла другая случайная величина. Выявление стохастической связи и ее оценка составляют суть корреляционного и регрессионного анализа.

Пусть имеется система двух случайных величин (X, Y). Проводятся эксперименты, в каждом из которых (X, Y) принимает некоторое значение (x_i, y_i). Результаты экспериментов сводятся в таблицу, которая называется корреляционной (табл. 12.1).

Таблица 12.1

Корреляционная таблица

(x, y)	y_1	y_2	. . .	y_s	y_i
x_1	m_{11}	m_{12}	. . .	m_{1s}	n_1
x_2	m_{21}	m_{22}	. . .	m_{2s}	n_2
. . .	. . .	. . .	. . .	. . .	. . .
x_k	m_{k1}	m_{k2}	. . .	m_{ks}	n_k
m_j	m_1	m_2	. . .	m_s	n

В этой таблице обозначено m_{ij} – количество появлений пары (x_i, y_i) в n экспериментах; $n_i = \sum_{j=1}^{S} m_{ij}$, $m_j = \sum_{i=1}^{k} m_{ij}$, $n = \sum_{i=1}^{k}\sum_{j=1}^{s} m_{ij}$.

Из табл. 12.1 видно, что каждому значению X соответствует некоторое распределение Y. Усредняя значения Y при фиксированных

X можно получить зависимость $\overline{y} = f(x)$, где $\overline{y}$ – среднее значение Y (табл. 12.2).

Таблица 12.2.

Зависимость $\overline{y} = f(x)$

X	x_1	x_2	x_3	. . .	x_k
$\overline{y}$	$\overline{y}_1$	$\overline{y}_2$	$\overline{y}_3$	. . .	$\overline{y}_4$

В этой таблице:

$$\overline{y}_1 = \frac{y_1 m_{11} + y_2 m_{12} + y_3 m_{13} + \ldots + y_s m_{1s}}{n_1},$$

$$\overline{y}_2 = \frac{y_1 m_{21} + y_2 m_{22} + y_3 m_{23} + \ldots + y_s m_{2s}}{n_2},$$

……..

$$\overline{y}_k = \frac{y_1 m_{k1} + y_2 m_{k2} + y_3 m_{k3} + \ldots + y_s m_{ks}}{n_k}.$$

Таким образом, между выборочными средними $\overline{y}$ и значениями x существует некоторая связь, обозначенная как $\overline{y} = f(x)$. Аналогично можно найти зависимость $\overline{x} = \Phi(y)$.

Уравнения $\overline{y} = f(x)$, $\overline{x} = \Phi(y)$ называются уравнениями регрессии соответственно Y на X и X на Y, а их графики – линиями регрессии.

Поскольку значения $\overline{y}$ случайны, то они рассеяны относительно некоторой средней линии (рис. 12.1).

В связи с этим возникает задача определения вида функции, описывающей регрессию, определения параметров этой функции и оценки силы корреляционной связи. Функцию, описывающую регрессию, называют моделью регрессии. Выбор модели регрессии

осуществляется либо на основе априорных сведений о характере зависимости Y и X (например, зависимость линейная, квадратичная и т. д.), либо, если таких сведений нет, путем подбора, идя от простой модели к более сложной, оценивая для каждой модели допускаемую ошибку. Чаще всего в качестве универсальной модели выбирается степенной полином вида $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_m x^m$, где a_i – неизвестные коэффициенты. Таким образом, уравнение регрессии заменяется его моделью.

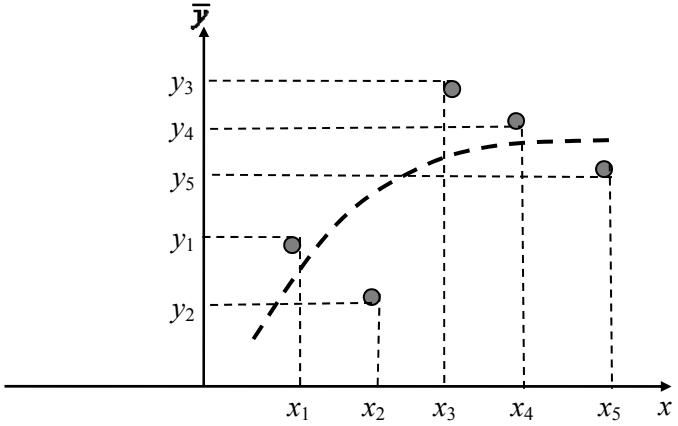


Рис. 12.1. График $\overline{y} = f(x)$

Неизвестные коэффициенты a_i определяются из условия минимизации суммы квадратов отклонения модели от опытных данных

$$S = \frac{1}{k}\sum_{i=1}^{k}[\overline{y}_i - f(x_i, a_0, a_1, \ldots, a_m)]^2 \to \min.$$

Различают простую регрессию (для системы двух случайных величин) и множественную (для системы трех и более случайных величин).

12.2. Линейная корреляционная зависимость и прямая регрессии

Пусть имеется совокупность k пар значений случайных величин $(x_1, \overline{y}_1), (x_2, \overline{y}_2), (x_k, \overline{y}_k)$, где $\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{s} y_j$ – рассчитывается по корреляционной таблице. Требуется найти уравнение регрессии Y на X, используя линейную модель.

Линейную модель регрессии запишем в виде

$$y = a_0 + a_1 x.$$

Неизвестные коэффициенты a_0, a_1 найдем из условия

$$S = \frac{1}{k}\sum_{i=1}^{k}[\overline{y}_i - (a_0 + a_1 x_i)]^2 \to \min_{a_0, a_1}. \qquad (12.1)$$

Условие (12.1) иллюстрируется рис. 12.2. Прямая $a_0 + a_1 x$ должна быть проведена таким образом, чтобы сумма квадратов отклонений ее от y_i была минимальной.

Для минимизации S возьмем ее производную по a_0 и a_1 и полученные выражения приравняем к нулю:

$$\left.\begin{aligned} \frac{\partial S}{\partial a_0} = 0 \\ \frac{\partial S}{\partial a_1} = 0 \end{aligned}\right\}.$$

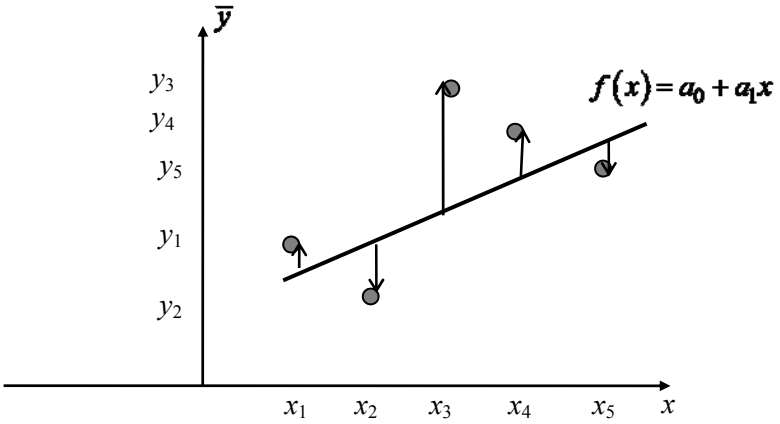


Рис. 12.2. Линейная модель регрессии

Можно показать, что при этом получается следующая система уравнений:

$$\left.\begin{aligned} a_1\bar{x}^2 + a_0\bar{x} &= \overline{xy} \\ a_1\bar{x} + a_0 &= \bar{y} \end{aligned}\right\}, \qquad (12.2)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$, $\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{s} y_i n_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{s} x_i y_j m_{ij}$.

Решая систему уравнений (12.2), получим

$$a_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}, \quad a_0 = \bar{y} - \frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}\bar{x},$$

где $\bar{K}_{xy}$, $\bar{\sigma}_x^2$ – выборочные значения корреляционного момента и дисперсии. Подставим полученные значения коэффициентов в исходное уравнение $\bar{y} = a_0 + a_1 x$ после подстановки получим

$$\bar{y}_x = \bar{y} - \frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}\bar{x} + \frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}x$$

или после преобразования

$$\bar{y}_x = \bar{y} - \frac{\bar{K}_{xy}}{\sigma_x^2}(\bar{x} - x).$$

Последнее уравнение называется эмпирическим уравнением прямой регрессии Y на X. Обозначив $\rho_{y/x} = \frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}$, где $\rho_{y/x}$ – выборочное значение коэффициента регрессии Y на X, получим окончательно

$$\bar{y}_x = \bar{y} - \rho_{y/x}(\bar{x} - x). \qquad (12.3)$$

Аналогично для регрессии X на Y имеем

$$\bar{x}_y = \bar{x} - \rho_{x/y}(\bar{y} - y),$$

где $\rho_{x/y}$ – выборочное значение коэффициента регрессии X на Y.

Из выражения (12.3) видно, при линейной модели регрессии линия регрессии имеет вид прямой, проходящей через точку с координатами $(\bar{x}, \bar{y})$, угол наклона этой прямой определяется коэффициентом регрессии. Чем он больше, тем больше угол наклона.

Известно, что сила корреляционной связи характеризуется коэффициентом корреляции $r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$. Установим связь между выборочными значениями коэффициентов корреляции и регрессии. Преобразуем выражение для r_{xy}.

$$\bar{r}_{xy} = \frac{\bar{K}_{xy}}{\sqrt{\bar{\sigma}_x^2 \bar{\sigma}_y^2}} = \sqrt{\frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}\frac{\bar{K}_{xy}}{\bar{\sigma}_y^2}} = \sqrt{\frac{\bar{K}_{xy}}{\bar{\sigma}_x^2}}\sqrt{\frac{\bar{K}_{xy}}{\bar{\sigma}_y^2}} = \sqrt{\rho_{x/y}}\sqrt{\rho_{y/x}} = \sqrt{\rho_{x/y}\rho_{y/x}}.$$

В свою очередь коэффициенты регрессии легко определяются через коэффициент корреляции:

$$\rho_{y/x} = \frac{\overline{K}_{xy}}{\overline{\sigma}_x^2} = \frac{\overline{K}_{xy}}{\sqrt{\overline{\sigma}_x^2 \overline{\sigma}_y^2}} \frac{\sqrt{\overline{\sigma}_y^2}}{\sqrt{\overline{\sigma}_x^2}} = \overline{r}_{xy} \frac{\sqrt{\overline{\sigma}_y^2}}{\sqrt{\overline{\sigma}_x^2}},$$

$$\rho_{x/y} = \frac{\overline{K}_{xy}}{\overline{\sigma}_y^2} = \frac{\overline{K}_{xy}}{\sqrt{\overline{\sigma}_x^2 \overline{\sigma}_y^2}} \frac{\sqrt{\overline{\sigma}_x^2}}{\sqrt{\overline{\sigma}_y^2}} = \overline{r}_{xy} \frac{\sqrt{\overline{\sigma}_x^2}}{\sqrt{\overline{\sigma}_y^2}}.$$

12.3. Оценка коэффициентов корреляции и регрессии по выборочным данным

В соответствие с приведенными ранее формулами оценки коэффициентов корреляции и регрессии равны:

$$\overline{r}_{xy} = \frac{\overline{K}_{xy}}{\sqrt{\overline{\sigma}_x^2 \overline{\sigma}_y^2}}, \quad \rho_{y/x} = \frac{\overline{K}_{xy}}{\overline{\sigma}_x^2}, \quad \rho_{x/y} = \frac{\overline{K}_{xy}}{\overline{\sigma}_y^2}.$$

Здесь $\overline{K}_{xy} = \overline{xy} - \overline{xy}$, $\overline{\sigma}_x^2 = \overline{x}^2 - x^{-2}$, $\overline{\sigma}_y^2 = \overline{y}^2 - y^{-2}$.

При $n > 50$ можно считать, что закон распределения $\overline{r}_{xy}$, $\rho_{y/x}(\rho_{x/y})$ – нормальный с дисперсиями:

$$\sigma_r = (1 - \overline{r}_{xy}^2)\,/\sqrt{n}, \quad \sigma_{y/x} = \frac{(1 - \overline{r}_{xy}^2)\sqrt{\overline{\sigma}_x^2}}{\sqrt{n\overline{\sigma}_y^2}}.$$

Тогда интервальные оценки равны

$$\overline{r}_{xy} - t_\gamma \sigma_r \le r_{xy} \le \overline{r}_{xy} + t_\gamma \sigma_r,$$

$$\rho_{y/x} - t_\gamma \sigma_{y/x} \le \rho_{y/x} \le \rho_{y/x} + t_\gamma \sigma_{y/x},$$

где t_γ – квантиль нормального распределения уровня γ;

γ – доверительная информация.

12.4. Проверка значимости модели и коэффициентов уравнения регрессии

После того, как по статистической выборке найдено уравнение приближенной регрессии, его следует подвергнуть регрессионному и корреляционному анализу:

– оценить ошибку от замены истинной регрессии приближенной;

– оценить долю регрессии в общем рассеянии значений $\overline{y}_i$.

Чтобы определить степень приближенности уравнения регрессии, необходимо найти оценки для всех коэффициентов уравнения. При достаточном объеме выборки закон распределения коэффициентов нормальный с дисперсией, вычисляемой через дисперсии *X* и *Y*.

Если уравнение регрессии находится путем последовательного уточнения, то рассчитав дисперсии коэффициентов для линейной модели, вводим поправку в модель и вновь рассчитываем дисперсии коэффициентов.

Например, пусть на первом этапе выбрана линейная модель регрессии $\overline{y}_x = 2 + 3x$. На втором этапе усложняем модель: $\overline{y}_x = a_0 + a_1 x + a_2 x^2$. Методом дисперсионного анализа сравниваем дисперсии коэффициентов регрессии. Если произошло значимое уменьшение дисперсии за счет усложнения уравнения, то следует отказаться от линейной модели.

ЛИТЕРАТУРА

1. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М.; Высшая школа, 1997.

2. Микулик, Н. А. Теория вероятностей и математическая статистика / Н. А. Микулик, А. В. Метельский. – Мн.: Пион, 2002.

3. Гусак, А. А. Высшая математика / А. А. Гусак : в 2 ч. – Ч. 2. – Мн.: ТетраСистемс, 2005.

4. Письменный, Д. Т. Конспект лекций по теории вероятностей и математической статистике / Д. Т. Письменный. – М.: Айрис-пресс, 2004.

5. Герасимович, А. И. Математическая статистика / А. И. Герасимович. – Мн.: Вышэйшая школа, 1983.

6. Вентцель, Е. С. Теория вероятностей / Е. С. Вентцель. – М.: Наука, 1969.

7. Чистяков, В. П. Курс теории вероятностей / В. П. Чистяков. – М.: Наука, 1987.

8. Пугачев, В. С. Теория вероятностей и математическая статистика / В. С. Пугачев. – М.: Наука, 1979.

9. Гмурман, В. Е. Руководство к решению задач по теории вероятностей и математической статистике / В. Е. Гмурман. – М.: Высшая школа, 1997.

10. Рябушко, А. П. Индивидуальные задания по высшей математике : в 4 ч. / А. П. Рябушко. – Ч. 4 – Мн.: Вышэйшая школа, 2007.

ПРИЛОЖЕНИЕ

Таблица П1

Таблица значений функции $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	3,0000	3,0080	3,0160	3,0239	3,0319	3,0399	3,0478	3,0558	0,0638	0,0717
0,1	3,0797	3,0876	3,0955	3,1034	3,1113	3,1192	3,1271	3,1350	0,1428	0,1507
0,2	3,1585	3,1663	3,1741	3,1819	3,1897	3,1974	0,2051	3,2128	0,2205	0,2282
0,3	3,2358	3,2434	3,2510	3,2586	3,2661	3,2737	3,2812	3,2886	0,2960	0,3035
0,4	3,3108	3,3182	3,3255	3,3328	3,3401	3,3473	3,3545	3,3616	0,3688	0,3759
0,5	3,3829	3,3899	3,3969	3,4039	3,4108	3,4177	3,4245	3,4313	0,4381	0,4448
0,6	3,4515	3,4581	3,4647	3,4713	3,4778	3,4843	3,4907	3,4971	0,5035	0,5098
0,7	3,5161	0,5223	3,5285	3,5346	3,5407	3,5467	0,5527	3,5587	0,5646	0,5705
0,8	3,5763	3,5821	3,5878	3,5935	3,5991	3,6047	3,6102	0,6157	0,6211	0,6265
0,9	3,6319	3,6372	3,6424	0,6476	3,6528	3,6579	0,6629	0,6679	0,6729	0,6778
1,0	0,6827	0,6875	3,6923	0,6970	3,7017	3,7063	3,7109	0,7154	0,7199	0,7243
1,1	3,7287	0,7330	0,7373	0,7415	0,7457	3,7499	0,7540	0,7580	0,7620	0,7660
1,2	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7994	0,8029
1,3	08064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,4	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,5	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,6	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,7	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,8	0,9281	0,9297	0,9312	0,9327	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,9	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	0,9643	0,9651	0,9660	0,9668	0,9666	0,9684	0,9692	0,9700	0,9707	0,9715
2,2	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,3	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0.9822	0,9827	0,9832
2,4	0,9836	0,9841	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,5	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904
2,6	0,9907	0,9910	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9928
2,7	0,9931	0,9933	0,9935	0,9937	0,9939	0,9940	0,9942	0,9944	0,9946	0,9947
2,8	0,9949	0,9951	0,9952	0,9953	0,9955	0,9956	0,9958	0,9959	0,9960	0,9961
2,9	0,9963	0,9964	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972

Окончание табл. П1

z	0	1	2	3	4	5	6	7	8	9
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986
3,2	0,9986	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,3	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,4	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995	0,9995
3,5	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997
3,6	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998	0,9998
3,7	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,0	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

Таблица П2

Таблица для значений функции Фишера для $1-\alpha=0,95$

γ_2 \ γ_1	1	2	3	4	5	6	8	12	16	24	50	∞
1	161,4	199	215	224	230	234	238	243	246	249	251	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,43	19,45	19,47	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,69	8,64	8,58	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,84	5,77	5,70	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,60	4,53	4,44	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,92	3,84	3,75	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,49	3,41	3,32	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,20	3,12	3,03	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,98	2,90	2,80	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,82	2,74	2,64	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,70	2,61	2,50	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,60	2,50	2,40	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,51	2,42	2,32	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,44	2,35	2,24	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,39	2,29	2,18	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,33	2,24	2,13	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,29	2,19	2,08	1,96

Окончание табл. П2

γ_2 \ γ_1	1	2	3	4	5	6	8	12	16	24	50	∞
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,25	2,15	2,04	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,21	2,11	2,00	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,18	2,08	1,96	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,15	2,05	1,93	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,13	2,03	1,91	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,11	2,00	1,88	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	2,09	1,98	1,86	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	2,07	1,96	1,84	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	2,05	1,95	1,82	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	2,03	1,93	1,80	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	2,02	1,91	1,78	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	2,00	1,90	1,77	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,99	1,89	1,76	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,94	1,83	1,70	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,90	1,70	1,66	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,87	1,76	1,63	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,85	1,74	1,60	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,81	1,70	1,56	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,79	1,67	1,53	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,77	1,65	1,51	1,32
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,76	1,64	1,49	1,30
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,75	1,63	1,48	1,28
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,72	1,60	1,45	1,25
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,71	1,59	1,44	1,22
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,69	1,57	1,42	1,19
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,68	1,55	1,39	1,15
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,67	1,54	1,38	1,13
500	3,86	3,01	2,62	2,39	2,23	2,1	1,96	1,77	1,66	1,54	1,38	1,11
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,65	1,53	1,36	1,08
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,64	1,52	1,35	1,00

Учебное издание

ИЗБРАННЫЕ ГЛАВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Пособие
для студентов специальностей
1-53 01 02 «Автоматизированные системы обработки информации», 1-40 01 01 «Программное обеспечение информационных технологий», 1-25 01 07 «Экономика и управление на предприятии», 1-26 02 02 «Менеджмент»

Составители:
ГОЛИКОВ Владимир Федорович
КАЗАКЕВИЧ Виктор Александрович

Редактор *Е. В. Герасименко*
Компьютерная верстка *Е. А. Беспанской*

Подписано в печать 07.09.2021. Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Ризография.
Усл. печ. л. 6,74. Уч.-изд. л. 5,27. Тираж 100. Заказ 383.

Издатель и полиграфическое исполнение: Белорусский национальный технический университет.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1/173 от 12.02.2014. Пр. Независимости, 65. 220013, г. Минск.