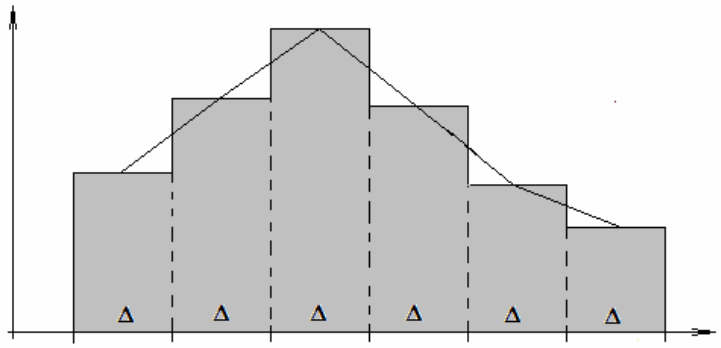
А.В. СОЛОПАХО

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

КРАТКИЙ КУРС ДЛЯ ЭКОНОМИСТОВ



♦ ИЗДАТЕЛЬСТВО ТГТУ ♦

Министерство образования и науки Российской Федерации
ГОУ ВПО «Тамбовский государственный технический университет»

А.В. СОЛОПАХО

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

КРАТКИЙ КУРС ДЛЯ ЭКОНОМИСТОВ

Утверждено Ученым советом университета
в качестве учебного пособия
для студентов специальностей 080105, 080109, 080500
всех форм обучения

Тамбов
Издательство ТГТУ
2007

УДК 519.2:33(075)
ББК В17я73
С606

Рецензенты:

Доктор физико-математических наук,
директор ИМФИ ТГУ им. Г.Р. Державина
Е.С. Жуковский

Доктор экономических наук, профессор
заведующий кафедрой «Бухгалтерский учет и аудит» ТГТУ
Л.В. Пархоменко

Солопахо, А.В.

С606 Теория вероятностей и математическая статистика: краткий курс для экономистов : учеб. пособие / А.В. Солопахо. – Тамбов : Изд-во Тамб. гос. техн. ун-та, 2007. – 108 с. – 120 экз. – ISBN 978-5-8265-0638-7.

Содержит материал, соответствующий программе дисциплины «Теория вероятностей и математическая статистика». Кратко, но достаточно полно излагаются сведения по разделам: «Вероятности случайных событий», «Теория случайных величин», «Оценка законов распределения», «Проверка статистических гипотез», «Регрессионный анализ». Рассматриваются примеры типовых задач и их решение.

Предназначено для студентов специальностей 080105, 080109, 080500 всех форм обучения.

УДК 519.2:33(075)
ББК В17я73

ISBN 978-5-8265-0638-7

Учебное издание

СОЛОПАХО Александр Владимирович

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

КРАТКИЙ КУРС ДЛЯ ЭКОНОМИСТОВ

Учебное пособие

Редактор Е.С. Мордасова
Инженер по компьютерному макетированию Т.Ю. Зотова

Подписано в печать 09.11.2007.
Формат 60 × 84 / 16. 6,28 усл. печ. л. Тираж 120 экз. Заказ № 715.

Издательско-полиграфический центр ТГТУ
392000, Тамбов, Советская, 106, к. 14

ВВЕДЕНИЕ

Теория вероятностей – математическая дисциплина, объектом изучения которой являются случайные события, т.е. события, происходящие в ходе эксперимента со случайным окончанием. На теории вероятностей основывается *математическая статистика*, которую иногда считают даже частью теории вероятностей. Задачей математической статистики является определение по имеющемуся набору экспериментальных данных некоторых общих характеристик случайных событий или явлений. За несколько десятилетий из теории вероятностей выделился целый ряд самостоятельных направлений, важнейшими из которых являются: теория случайных процессов; теория массового обслуживания; теория информации; эконометрическое моделирование. Большой вклад в развитие теории вероятностей внесли русские и советские ученые

Экономика и производственные процессы – одна из важнейших сфер применения теории вероятности и математической статистики. Исследование и прогнозирование экономических явлений трудно себе представить без использования методик статистического оценивания и проверки гипотез, регрессионного анализа, трендовых и сглаживающих эконометрических моделей и других методов, опирающихся на теорию вероятностей. С развитием общества экономика все более усложняется и, следовательно, по законам развития динамических систем должен усиливаться статистический характер законов, описывающих социально-экономические явления.

Все это предопределяет необходимость овладения методами теории вероятностей и математической статистики как важнейшим инструментом анализа и прогнозирования экономических явлений и процессов.

1. ИСЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ СЛУЧАЙНЫХ СОБЫТИЙ

1.1. Общие понятия о случайном событии и его вероятности. Действия над случайными событиями

Введем некоторые исходные понятия [1, 3, 7].

***Определение* 1.1.** *Случайным* называют событие, которое может как произойти, так и не произойти в ходе некоторого соответствующего наблюдения или эксперимента, называемого *вероятностным* или *стохастическим опытом*.

Важно заметить, что любое случайное событие всегда связано с конкретным вероятностным опытом, и не может рассматриваться само по себе, в отрыве от этого опыта. В определении также делается разница между экспериментом и наблюдением. Фактически при решении задач ее почти не существует. Однако, желательно для большей определенности понимать, к какому типу относится данный вероятностный опыт. Эксперимент носит активный характер, например бросается игральный кубик. Наблюдение – пассивный, например – мы просто наблюдаем, сколько покупателей приходит за день в магазин.

Обычно случайные события обозначаются заглавными латинскими буквами: A, B, Z и т.д.

***Определение* 1.2.** (*Неформальное определение вероятности*). *Вероятностью* случайного события называют меру возможности его осуществления в ходе соответствующего опыта.

Отметим, что это определение действительно носит *неформальный* характер, так как не позволяет рассчитывать эти самые вероятности. Вероятность случайного события A обозначается так

$$P(A).$$

***Определение* 1.3.** Если известно, что некоторое событие обязательно произойдет в ходе соответствующего опыта, то его называют *достоверным*. Достоверное событие обычно обозначают Ω, и при этом считают, что

$$P(\Omega) = 1.$$

***Определение* 1.4.** Если известно, что некоторое событие никогда не произойдет в ходе соответствующего опыта, то его называют *невозможным*. Невозможное событие обычно обозначают $\varnothing$, и при этом считают, что

$$P(\varnothing) = 0.$$

Таким образом, вероятность события – это число от нуля до единицы.

Над случайными событиями определяются следующие действия.

***Определение* 1.5.** Случайное событие C называется *суммой* событий A и B, если оно происходит $\Leftrightarrow$ происходит хотя бы одно из событий слагаемых. При этом пишут

$$C = A + B \quad \text{или} \quad C = A \cup B.$$

***Определение* 1.6.** Случайное событие C называется *произведением* событий A и B, если оно происходит $\Leftrightarrow$ происходят оба события сомножителя. При этом пишут

$$C = A \cdot B \quad \text{или} \quad C = A \cap B.$$

***Определение* 1.7.** Событие C называется *разностью* событий A и B, если оно происходит $\Leftrightarrow$ происходит A, но не происходит B. При этом пишут

$$C = A - B \quad \text{или} \quad C = A \setminus B .$$

Для иллюстрации самых разных определений и фактов теории вероятностей весьма удобным оказывается использование геометрических множеств (рис. 1.1.). При этом считают, что эксперимент состоит в случайном бросании точки на некоторое множество Ω. А случайные события состоят в попадании или непопадании данной точки на соответствующие подмножества этого множества. Вообще можно отметить с самого начала, что понятие случайного события в теории вероятностей *всегда* можно ассоциировать с понятием множества. Причем в различных случаях (при различных вероятностных экспериментах) вид этих множеств различен (дискретные конечные, дискретные бесконечные, непрерывные и т.д.).

***Определение* 1.8.** *Противоположным* событию A называется событие $\overline{A}$, которое происходит $\Leftrightarrow$ не происходит A.

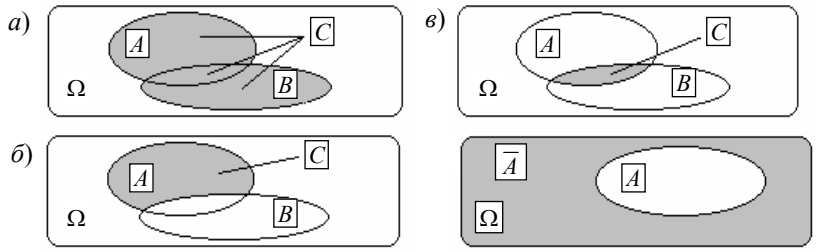


Рис. 1.1:
а – сложение; *б* – умножение; *в* – вычитание; *г* – противоположное событие

1.2. Схема с равновозможными исходами. Классическое определение вероятности

Несмотря на кажущуюся простоту понятия вероятности события, строгое формальное его определение является трудной проблемой. Об этом говорит хотя бы тот факт, что многие великие математики XVIII – XIX вв. так и не смогли этого сделать.

Проще всего формализовать понятие вероятности можно, если рассматриваемый вероятностный эксперимент соответствует следующей схеме, которая называется *схемой с равновозможными исходами*:

1) эксперимент может закончиться только появлением одного из n возможных исходов

$$w_1, w_2, \ldots w_n ,$$

называемых *элементарными исходами*;

2) эти исходы *равновозможны*, а это означает, что следует считать

$$P(w_i) = \frac{1}{n}, \quad i = \overline{1,n}.$$

Примеры вероятностных опытов соответствующих этой схеме, в изобилии имеются в сфере азартных игр, с которыми кстати связано само начало развития теории вероятностей, еще со времен средневековья. К сожалению в реальной жизни таких ситуаций не так много.

***Пример* 1.1.** Рассмотрим случайное бросание игрального кубика. Ясно, что этот эксперимент можно считать соответствующим схеме с равновозможными исходами, при $n = 6$. Обозначим w_i – элементарный исход, состоящий в выпадении на кубике i очков, $i = \overline{1,6}$.

***Определение* 1.9.** Некоторый элементарный исход опыта называется *благоприятствующим* для события A, если при его выпадении считают, что событие A произошло.

***Определение* 1.10.** Случайное событие A называется *сложным* или *составным*, если ему благоприятствует два и более исходов.

При этом, если, например, для A благоприятствующими являются исходы

$$w_{i1}, w_{i2}, \ldots, w_{im},$$

то так и пишут

$$A = \{w_{i1}, w_{i2}, \ldots, w_{im}\}.$$

***Пример* 1.2.** Пусть A – случайное событие, состоящее в выпадении на игральном кубике четного числа очков. Ясно, что это сложное событие, так как ему благоприятствуют три исхода, и

$$A = \{w_2, w_4, w_6\}.$$

Очень важным является следующее определение.

***Определение* 1.11.** (*Классическое определение вероятности*). Вероятностью случайного события A в схеме с равновозможными исходами называется число

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

где n – число всех возможных исходов, m – число исходов благоприятствующих A.

Пользуясь введенным определением, теперь можно формально строго находить вероятности событий, если известно количество благоприятствующих им исходов. А также вероятности событий являющихся результатом действий над другими событиями, если известны множества всех благоприятствующих исходов событий-операндов.

***Пример* 1.3.** Пусть, например, $n = 6$, $A = \{w_1, w_2, w_3\}$, $B = \{w_3, w_4\}$. Тогда, по классическому определению вероятностей, имеем

$$P(A) = 3/6 = 1/2; \quad P(B) = 2/6 = 1/3.$$

Далее получаем, что:

$$C = A + B = \{w_1, w_2, w_3, w_4\} \Rightarrow P(C) = 4/6 = 2/3,$$
$$D = AB = \{w_3\} \Rightarrow P(D) = 1/6,$$
$$Z = A - B = \{w_1, w_2\} \Rightarrow P(Z) = 1/3.$$

1.3. Использование комбинаторных формул

Комбинаторика – наука о перестановках. Ее возникновение так же связывают с азартными играми и относят к средневековью. Многие ее понятия широко используются в теории вероятностей [3].

***Определение* 1.12.** *Числом перестановок* из n называют количество всевозможных вариантов упорядочивания n имеющихся объектов. Это число обозначается и равно

$$P_n = n!.$$

***Определение* 1.13.** *Числом размещений* из n по k называют количество всевозможных вариантов выбора k объектов из n имеющихся, с учетом порядка их следования. Это число обозначается и равно

$$A_n^k = \frac{n!}{(n-k)!}.$$

***Определение* 1.14.** *Числом сочетаний без повторений* из n по k называется количество возможных способов выбора k объектов из n имеющихся (без учета порядка следования). Это число обозначается и равно

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

***Пример* 1.3.**

1. Сколькими способами можно вытащить три карты из колоды?

$$C_{36}^3 = \frac{36!}{3!(36-3)!} = \frac{34 \cdot 35 \cdot 36}{1 \cdot 2 \cdot 3} = 7140.$$

2. Какова вероятность, что все три наудачу выбранные из колоды карты окажутся королями?
По классическому определению вероятности, получаем

$$P(A) = \frac{C_4^3}{C_{36}^3} = \frac{4}{7140}.$$

Многие задачи на схему с равновозможными исходами являются частным случаем следующей. Пусть в закрытой урне имеется N шаров, из которых M белые. Какова вероятность, что из n наудачу выбранных шаров ровно m окажется белыми. Обозначим это событие через A, и построим формулу для его вероятности.

Вариантов выбора n из N шаров определяется числом C_N^n. Количество способов выбора m белых шаров из M имеющихся – C_M^m. Далее ясно, что способов выбора из оставшихся $N–M$ шаров равно $n–m$ не белых, определяется числом C_{N-M}^{n-m}, тогда искомая вероятность, по классическому определению, выразится формулой

$$P(A) = \frac{C_{N-M}^{n-m} C_M^m}{C_N^n}.$$

Отметим, что не было никаких ограничений на величины N, M, n и m. В частности, они могут быть равны нулю, или – друг другу. В этом и состоит причина большой универсальности данной формулы.

***Пример* 1.4.**

1. Для решения задачи предыдущего примера можно было бы использовать эту формулу, при

$$N = 36, \quad n = 4, \quad M = 3, \quad m = 3.$$

2. Пусть имеется 500 электрических лампочек. Завод-изготовитель гарантирует, что из них не более 2 бракованных. Оценить, какова вероятность, что из 5 выбранных лампочек нет ни одной бракованной. Используя рассмотренную формулу, при

$$N = 500,\ n = 5,\ M = 2,\ m = 0\ ,$$

получаем

$$P = \frac{C_{498}^{5} C_{2}^{0}}{C_{500}^{5}} \geq\ 0{,}98.$$

1.4. Схема с неравновозможными исходами. Статистическое определение вероятности

Как уже отмечалось, лишь небольшой спектр практических ситуаций соответствует схеме с равновозможными исходами. Например, стрельба по мишени. Нет никаких оснований полагать, что выбивание того или иного количества очков равновозможно. В таких ситуациях следует перейти к следующей схеме, называемой *схемой с неравновозможными исходами*:

1) эксперимент может закончиться только появлением одного из n возможных элементарных исходов

$$w_1, w_2, \ldots, w_n;$$

2) заданы вероятности этих исходов

$$P(w_i) = p_i,\ \ i = \overline{1,n}.$$

Из теоремы, которая будет рассмотрена ниже, следует, что должно выполняться условие

$$\sum_{i=1}^{n} p_i = 1\ .$$

Однако, его необходимость и без того достаточно очевидна. Фактически это условие означает, что в результате опыта обязательно произойдет один, и только один, элементарный исход из рассматриваемой совокупности.

***Определение* 1.15.** Вероятностью случайного события A в схеме с неравновозможными исходами называется число

$$P(A) = \sum_{w_i \in A} p_i\ .$$

Это определение дает возможность формально строго решать большинство соответствующих задач. Однако, для практических ситуаций остается неясным ответ на вопрос – откуда могут быть известны величины p_i? Частично этот ответ дает следующее определение.

***Определение* 1.16.** (*Статистическое определение вероятности*). Вероятностью события A называется предел

$$P(A) = \lim_{n\to\infty} \frac{k}{n},$$

где n – количество независимо проведенных одинаковых опытов, в результате которых рассматривается появление или непоявление A, k – количество появлений A.

Ясно, что на практике невозможно бесконечное количество экспериментов. Однако, если статистика наблюдений достаточно велика, то практически может оказаться допустимым считать

$$P(A) \approx \frac{k}{n}\ .$$

Величину $\mu_A = \dfrac{k}{n}$ называют *относительной частотой появления* события A.

***Пример* 1.5.** Имеется статистика объемов продаж некоторого товара случайному потоку покупателей за некоторый предшествующий период времени.

Сумма покупки, тыс. р.	До 1	1 – 2	2 – 4	4 – 10	Более 10
Количество покупателей	5	10	20	8	2

Оценить вероятность, что очередному покупателю потребуется товара на сумму более 4 тыс. р.

Рассмотрим ситуацию, как эксперимент с пятью неравновозможными элементарными исходами. Вероятности которых, в соответствии со статистическим определением и имеющимися данными, можно принять равными

$$p_1 = \frac{5}{45},\ \ p_2 = \frac{10}{45},\ \ p_3 = \frac{20}{45},\ \ p_4 = \frac{8}{45},\ \ p_5 = \frac{2}{45}.$$

Тогда для искомой вероятности имеем

$$P(A) = p_4 + p_5 = \frac{10}{45} = \frac{2}{9}.$$

***Определение* 1.16.** Говорят, что некоторое множество является *счетным*, или имеет *счетную мощность*, если оно состоит из бесконечного количества элементов, каждому из которых теоретически может быть приписан порядковый номер.

Так, например, счетными являются множества всех целых чисел на числовой оси R^1, или даже всех дробей (рациональных чисел).

Схема с неравновозможными исходами является обобщением классической схемы. Дальнейшим обобщением является схема с бесконечным, но счетным количеством исходов. Теоретические построения при этом настолько аналогичны рассмотренным в этом пункте, что мы не будем специально на них останавливаться. Тем более, что практические ситуации, соответствующие такой схеме, весьма редки. Стоит лишь отметить, что вероятности элементарных исходов должны убывать до нуля, так, чтобы выполнялось необходимое условие

$$\sum_{i=1}^{\infty} p_i = 1 .$$

1.5. Схема с несчетным множеством исходов. Геометрическое определение вероятности

***Определение* 1.17.** Говорят, что некоторое множество является *несчетным*, или имеет *более чем счетную мощность*, или является *континиумом*, если невозможно каждому его элементу приписать порядковый номер.

Оказывается, что любой сколь угодно малый непрерывный промежуток числовой оси имеет более чем счетную мощность. Также как и любые непрерывные множества на плоскости или в пространстве.

Рассмотрим эксперимент – бросание наугад точки на отрезок $[a,b] \in R^1$. Ясно, что множество элементарных исходов в нем является континиумом. А тем самым, этот эксперимент не может быть рассмотрен в рамках выше изложенных схем. Можно привести много других таких же примеров: траектория движения материального тела в пространстве, прогноз температуры воздуха на завтра, и т.д. Строгое построение теории для этих случаев возможно лишь на основе следующего понятия.

***Определение* 1.18.** *Мерой Лебега* множества $G \in R^n$ называется:

- сумма длин составляющих его интервалов, если $G \in R^1$;
- его площадь, если $G \in R^2$;
- его объем, если $G \in R^n, n \geq 3$.

***Определение* 1.19.** Множество $G \in R^n$ называется *измеримым*, если оно имеет меру Лебега.

***Теорема* 1.1.** Объединение и пересечение любого не более чем счетного количества количества измеримых множеств и их дополнений является измеримым множеством.

Оказывается, что существуют неизмеримые множества, т.е. множества, для которых нельзя указать ни длины, ни площади, ни объема.

***Пример* 1.6.** (*Неизмеримого множества*). Разобьем все точки отрезка [0,1] на классы, отнеся x и y к одному классу, тогда и только тогда, когда ($x - y$) есть число рациональное, т.е. класс К(x) состоит из точек, принадлежащих отрезку от 0 до 1 и таких, что $y = x + r$, где r – любое рациональное число. Выберем из каждого класса по одной точке, обозначим множество выбранных точек через A. Можно доказать, что A не является измеримым множеством.

Уже из этого случая можно понять, что примеры неизмеримых множеств достаточно сложны и искусственны. Обычно все множества измеримы.

Рассмотрим следующую схему проведения стохастического эксперимента, называемую *схемой с несчетным множеством исходов*:

1) опыт состоит в случайном бросании точки на некоторое измеримое множество $\Omega \in R^n$, т.е. возможными элементарными исходами являются все точки этого множества;

2) вероятность попадания точки на некоторое измеримое подмножество $A \subset \Omega$ зависит только от меры этого подмножества, и не зависит от его расположения на Ω.

Последнее условие можно сформулировать и иначе: вероятности попадания точки на подмножества одинаковой меры равны. Иногда, для простоты, используют и не совсем корректную формулировку – все точки Ω «одинаково возможны».

Очень важно отметить, что в «несчетной» схеме в качестве случайных событий можно рассматривать только попадание точки на измеримые подмножества множества Ω. Характерным является еще то, что вероятность любого элементарного исхода в этой схеме равна нулю.

***Определение* 1.20.** (*Геометрическое определение вероятности*). Вероятностью случайного события A в схеме с более чем счетным множеством исходов называется число

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

где Ω – множество всех исходов, $A \in \Omega$.

***Пример* 1.7.** (*Задача Бюффона*). Плоскость разграфлена параллельными прямыми, отстоящими друг от друга на расстоянии $2a$. На плоскость наугад бросают иглу длиной $2l$ ($l < a$). Найти вероятность того, что игла пересечет какую-нибудь прямую.

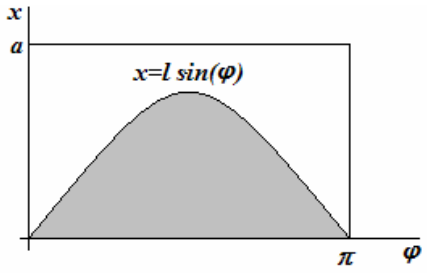


Рис. 1.2.

Решение.

Пусть x – расстояние от центра K иглы до ближайшей прямой (рис. 1.2.), φ – угол, составленный иглой с этой прямой. Множество всех возможных положений иглы, т.е. совокупность Ω элементарных исходов – это прямоугольник

$$\Omega = \{(x,\ \varphi) : 0 \le x \le a, \quad 0 \le \varphi \le \pi\}.$$

Совокупность благоприятствующих положений иглы образует множество

$$C = \{(x,\ \varphi) : x \le l \sin\varphi, \quad (x,\ \varphi) \in \Omega\}.$$

Тогда по геометрическому определению вероятности

$$P(A) = \frac{m(C)}{m(\Omega)} = \frac{\int_0^{\pi} l \sin\varphi\, d\varphi}{a\,\pi} = \frac{2l}{a\,\pi}.$$

Полученный результат интересен прежде всего тем, что позволяет экспериментально определить число π. Действительно, если принять

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

где $\frac{m}{n}$ – относительная частота пересечений, то

$$\pi \approx \frac{2l\,n}{a\,m}.$$

Такие эксперименты проводились многократно [10].

Являясь обобщением вышерассмотренных схем, схема с более чем счетным множеством исходов позволяет рассматривать самые разные эксперименты:

1. В случае равновозможных исходов можно считать, что опыт состоит в случайном бросании точки на отрезок, разбитый на n интервалов одинаковой длины.
2. В случае конечного числа неравновозможных исходов – на n интервалов, соответствующих их вероятностям длины.
3. В случае счетного числа исходов – на счетное число подынтервалов бесконечно убывающей длины, таких что их объединение составляет весь отрезок Ω.

Можно предложить и многомерные аналоги, но в этом нет необходимости.

Последние соображения делают, и без того достаточно понятную, аналогию между случайными событиями и геометрическими множествами полностью обоснованной.

1.6. Теоремы сложения и умножения вероятностей

Оказывается, часто удается находить вероятности событий, являющихся результатом действия над другими событиями и без знания конкретного множества исходов, благоприятствующих событиям-операндам.

***Определение* 1.21.** События A и B называют *несовместными*, если наступление одного из них исключает наступление другого.

Несовместность означает, что события не имеют общих благоприятствующих исходов, или, обращаясь к геометрическим иллюстрациям, – соответствующие им подмножества не пересекаются.

***Теорема* 1.2.** (*Теорема сложения вероятностей*). Если A и B являются несовместными, то

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

Доказательство. Если использовать геометрическое определение вероятности, то это и последующее утверждение выглядят очевидным.■

***Следствие* 1.1.** Если A и B любые случайные события, связанные с общим вероятностным экспериментом, то

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

Введем следующее весьма важное понятие.

***Определение* 1.22.** События A и B называются *независимыми*, если появление одного из них не влияет на вероятность появления другого.

Можно приводить многочисленные реальные примеры, как зависимых, так и независимых случайных событий.

Фактически целесообразнее считать, что независимыми могут быть лишь события, появление которых рассматривается в ходе различных, независимых опытов (хотя формально это и необязательно). Эти опыты, или испытания, могут быть как совершенно различной природы и содержания, так и в принципе одинаковыми, только независимо проводящимися. Например, два бросания игрального кубика – независимые опыты.

***Теорема* 1.3.** (*Теорема умножения вероятностей*). Пусть A и B независимы, тогда

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

Доказательство. Опять рассмотрим наиболее общую схему с несчетным множеством исходов. Для простоты геометрической иллюстрации будем считать, что эксперименты, с которыми связаны события A и B, соответствуют случайному бросанию точки на одномерные отрезки Ω_1 и Ω_2 соответственно. При этом A состоит в попадании точки на интервал $[a_1, a_2] \in \Omega_1$, а B – на интервал $[b_1, b_2] \in \Omega_2$ (рис. 1.3.). Тогда, очевидно, можно перейти от пары этих экспериментов к единому опыту, состоящему в случайном бросании точки на двухмерное множество Ω, подмножество C которого соответствует произведению A и B.

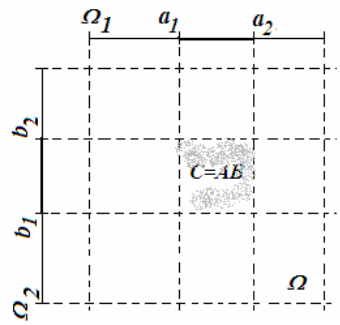


Рис. 1.3.

Далее, по геометрическому определению вероятности, имеем

$$P(AB) = \frac{m(C)}{m(\Omega)} = \frac{S(C)}{S(\Omega)} = \frac{l(a_1,a_2)\cdot l(a_1,a_2)}{l(\Omega_1)\cdot l(\Omega_2)} = \frac{l(a_1,a_2)}{l(\Omega_1)} \cdot \frac{l(a_1,a_2)}{l(\Omega_2)} = P(A)\cdot P(B),$$

что и требовалось доказать.■

1.7. Формулы условной вероятности, полной вероятности и формула Байеса

Пусть события A и B связаны с общим вероятностным опытом, и имеют общие благоприятствующие исходы. Тогда, если известно, что в результате опыта произошло событие B, то имеется вероятность, что произошло и A. Эта вероятность называется *условной вероятностью* события A при условии, что произошло B, и обычно обозначается

$$P_B(A) \text{ или } P(A/B).$$

Выведем формулу для расчета $P_B(A)$.

Итак, пусть известно, что событие B произошло. Тогда новой совокупностью возможных элементарных исходов являются лишь исходы, благоприятствующие B, т.е. B – достоверное событие. За исходы, благоприятствующие A, следует считать лишь те исходы, которые являются общими с B, т.е. составляют множество $A \cap B$.

Тогда, в соответствии с наиболее общей схемой с несчетным числом исходов и геометрическим определением вероятности получаем формулу

$$P(A/B) = P_B(A) = \frac{P(A \cap B)}{P(B)},$$

которая и называется *формулой условной вероятности*.

***Определение* 1.23.** Говорят, что случайные события A_1, A_2, …, A_n, связанные с общим вероятностным экспериментом, образуют полную группу событий, если

1. $A_i \cap A_j = \varnothing, \quad i \neq j$.

2. $\bigcup_{i=1}^{n} A_i = \Omega$.

***Теорема* 1.4.** (*Формула полной вероятности*). Если $A_1 \dots A_n$ образуют полную группу событий, причем $P(A_i) > 0$, $i = \overline{1,n}$, то для любого события B, связанного с тем же экспериментом, справедлива формула

$$P(B) = \sum_{i=1}^{n} P(B / A_i) P(A_i),$$

называемая *формулой полной вероятности*.

Доказательство. Из того, что

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} (B \cap A_i),$$

причем события $B \cap A_i$ несовместны, по теореме сложения вероятностей получаем

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i),$$

и используя формулу условной вероятности, окончательно

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

Формула полной вероятности, в определенном смысле, подобна разложению вектора по базису. Из нее легко выводится еще одна важная формула, называемая *формулой Байеса*

$$P(A_k / B) = \frac{P(A_k) P(B / A_k)}{P(B)}.$$

***Пример* 1.8.** На предприятии изготавливают некоторые изделия на трех поточных линиях. На первой производят 20 % всех изделий, на второй 30 %, на третьей – 50 %. Каждая линия характеризуется следующими процентами годности изделия: 95, 98, 97 %. Требуется определить вероятности того, что:

1. Взятое наугад изделие окажется бракованным.
2. Бракованное изделие изготовлено на 1, 2, 3 линии.

Решение. Обозначим A_1, A_2, A_3 – взятое наугад изделие изготовлено на 1, 2, 3 линиях. Согласно условию

$$P(A_1) = 0{,}2; \quad P(A_2) = 0{,}3; \quad P(A_3) = 0{,}5.$$

Обозначим за B – взятое наугад изделие оказалось бракованным, согласно условию

$$P(B/A_1) = 0{,}05; \quad P(B/A_2) = 0{,}02; \quad P(B/A_3) = 0{,}03.$$

Используя формулу полной вероятности находим

$$P(B) = P(B/A_1)\, P(A_1) + P(B/A_2)\, P(A_2) + P(B/A_3)\, P(A_3) =$$

$$= 0{,}2 \cdot 0{,}05 + 0{,}3 \cdot 0{,}02 + 0{,}5 \cdot 0{,}03 = 0{,}031.$$

Отметим, что последняя величина фактически выражает общий уровень брака по предприятию.

Вероятность того, что взятое наугад бракованное изделие изготовлено на той или иной линии, находим по формуле Байеса:

$$P(A_1/B) = \frac{P(A_1) P(B / A_1)}{P(B)} = \frac{0{,}02 \cdot 0{,}05}{0{,}031} = \frac{10}{31}; \; P(A_2/B) = \frac{P(A_2) P(B / A_2)}{P(B)} = \frac{0{,}006}{0{,}031} = \frac{6}{31};$$

$$P(A_3/B) = \frac{P(A_3) P(B / A_3)}{P(B)} = \frac{0{,}015}{0{,}031} = \frac{15}{31}.$$

Задача решена.

1.8. Аксиомы теории вероятности. Вероятностное пространство

Пусть имеется некоторый стохастический эксперимент, и известно множество всех его элементарных исходов Ω. Пусть в Ω выделена система подмножеств F, соответствующих условиям:

A_1) $\Omega \in F$;

A_2) если $A \in F$, то $\overline{A} = \Omega \setminus A \in F$;

A_3) если $A_i \in F$, $i = \overline{1,\infty}$, то $\bigcup_{i=1}^{\infty} A_i \in F$.

Такая система подмножеств называется σ-*алгеброй*. Элементы F будем называть *случайными событиями*.

Таким образом, строго говоря, случайными событиями в любом эксперименте допустимо считать не любое подмножество, принадлежащее Ω, а лишь элементы некоторой определенной нами σ-алгебры его подмножеств. Это существенное замечание.

Пусть на элементах F определена функция $P(\cdot)$, обладающая свойствами:

P_1) $P(A) \geq 0, \forall A \in F$;

P_2) $P(\Omega) = 1$;

P_3) если $A_i \cap A_j = \varnothing$, $i \neq j$, то $P(\quad A_i) = \sum^{\infty} P(A_i)$.

Такую функцию называют *вероятностной мерой*, или просто *вероятностью*.

Утверждение $A_1, A_2, A_3, P_1, P_2, P_3$ составляют, так называемую, *систему аксиом теории вероятности*, которая была предложена А.Н. Колмогоровым и оказалась исключительно плодотворной для развития этой науки.

Проиллюстрируем полезность введенного аксиоматического подхода на примере доказательства нескольких простых, но важных теорем.

***Теорема* 1.5.** Пусть A и B – случайные события, и $A \subset B$. Тогда

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

Доказательство. Поскольку $A \subset B$, то $B = A + (B - A)$, причем

$$A \cap (B - A) = \varnothing,$$

тогда по P_3 :

$$P(B) = P(A) + P(B - A) \Rightarrow P(B - A) = P(B) - P(A). \blacksquare$$

***Теорема* 1.6.** Пусть A и B случайные события, тогда

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

Доказательство. Поскольку $A + B = (A - (AB)) + (B - (AB)) + (AB)$, то по P_3 и предыдущей теореме

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

Последнюю формулу можно обобщить для любого числа событий.

Центральным понятием в теории вероятностей является следующее.

***Определение* 1.24.** Если для некоторого стохастического эксперимента заданы: множество всех его элементарных исходов Ω, σ-алгебра F подмножеств Ω, а на элементах F определена функция $P(\cdot)$, обладающая свойствами вероятностной меры, то говорят, что задана *вероятностная модель* этого эксперимента. Тройку (Ω, F, P) называют *вероятностным пространством*.

Для корректного решения практических задач исключительной важностью обладает правильное и четкое определение всех составляющих указанной тройки. Именно ошибки, возникающие на этом этапе, и являются основной причиной дальнейших неправильных расчетов и выводов. Рассматривая конкретную задачу или ситуацию, нужно очень четко выявить все множество элементарных исходов, установить их вероятности, выяснить, какие события следует считать случайными событиями данного эксперимента и определить для них функцию вероятности.

1.9. Последовательности испытаний. Схема Бернулли

Рассмотрим следующий вероятностный эксперимент:

1) последовательно проводится n независимых одинаковых опытов;
2) в каждом из которых имеется лишь два исхода, которые условно назовем «успех» – 1, и «неудача» – 0;
3) во всех опытах вероятность «успеха» – p, и, соответственно, «неудачи» – $q = 1 - p$, неизменны.

Такая последовательность испытаний называется *схемой Бернулли*, она соответствует весьма многим практическим ситуациям, а именно таким, которые характеризуются «массовостью» явления, что часто встречается в социологии, маркетинге и т.д.

Ясно, что элементарными исходами, в эксперименте по схеме Бернулли, являются всевозможные комбинации вида

$$\underbrace{100 \ldots 010}_{n-\text{позиций}}.$$

Множество всех таких исходов состоит из 2^n элементов.

По теореме умножения вероятностей несложно рассчитать вероятность любого такого исхода, например

$$P(100 \ldots 010) = p \cdot q \cdot q \ldots q \cdot p \cdot q.$$

Очевидно, что любой исход, состоящий из m «успехов» и $(n–m)$ «неудач» имеет вероятность $p^m q^{n-m}$.

А всего таких исходов C_n^m. Складывая их вероятности, получаем так называемую *формулу Бернулли*

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

которая выражает вероятность того, что при n испытаниях «успех» произойдет ровно m раз.

Вспоминая формулу бинома Ньютона устанавливаем справедливость необходимого условия

$$\sum_{m=0}^{n} P_n(m) = \sum_{m=0}^{n} C_n^m p^m q^{n-m} = (p+q)^n = 1^n = 1 .$$

Пример. Вероятность нормального приживления саженца плодового дерева $p = 0{,}8$. Найти вероятности, что:

1) из 10 саженцев приживется ровно 8;
2) не меньше 8.

Решение.

1. Ясно, что приживления отдельных саженцев можно считать независимыми, поэтому ситуация соответствует схеме Бернулли. Тогда по формуле Бернулли получаем

$$P(A) = C_{10}^8 \cdot 0{,}8^8 \cdot 0{,}2^{10-8} .$$

2. По теореме сложения

$$P(A) = \sum_{i=8}^{10} P_{10}(i) .$$

1.10. Локальная и интегральная теоремы Муавра–Лапласа

При большой величине m и n использование формулы Бернулли становится затруднительным. Это связано с очень большой скоростью роста функции $n!$, и даже с помощью ЭВМ при $n > 100$ факториал рассчитать невозможно. Однако существуют ориентированные на практические расчеты приближенные формулы.

***Теорема* 1.7.** (*Формула Пуассона*). Если при неограниченном увеличении числа испытаний n произведение np стремится к постоянному числу λ, то

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

Доказательство. По формуле Бернулли

$$P_n(m) = C_n^m p^m q^{n-m} = \frac{n(n-1)(n-2)\ldots(n-m+1)}{m!} p^m (1-p)^n (1-p)^{-m} = \ldots ,$$

или учитывая условие

$$\lim_{n\to\infty} n\,p = \lambda ,$$

при $n \to \infty$, используя второй замечательный предел, имеем

$$\ldots = 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)\frac{\lambda^m}{m!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m} \to \frac{\lambda^m e^{-\lambda}}{m!} . \blacksquare$$

В доказанном утверждении, строго говоря, не выполняется обязательное для схемы Бернулли условие постоянства величины p. Однако, на практике и n всегда конечно, поэтому полученное равенство (при достаточно больших n и условии $np \le 10$) используют как дающее приближенный результат. Формула Пуассона эффективна тогда, когда p достаточно мало, если же это не так, то можно использовать следующую важную формулу.

***Теорема* 1.8.** (*Локальная теорема Муавра–Лапласа*). Если вероятность p появления «успеха» в каждом из n независимых испытаний постоянна, и отлична от 0 и 1, то:

$$\lim_{\substack{n,m\to\infty \\ m\le n}} P_n(m) = \frac{1}{\sqrt{npq}}\,\varphi\left(\frac{m-np}{\sqrt{npq}}\right),$$

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

Функция $\varphi(x)$ называется *функцией Лапласа*. Таблица ее значений приводится во многих справочниках и учебниках. При их использовании следует помнить о четности этой функции. Считается, что локальная формула Лапласа дает достаточно хороший результат при $npq > 15 - 20$.

Вероятность, что из n испытаний по схеме Бернулли «успех» произойдет не менее m_1 раза, и не более m_2 раз, обозначается и равна

$$P_n(m_1, m_2) = \sum_{i=m_1}^{m_2} C_n^i\; p^i\; q^{n-i} .$$

И если слагаемых в этой формуле хотя бы несколько десятков, то ее непосредственное использование, очевидно, еще более затруднительно. Для этого случая имеется следующая теорема.

***Теорема* 1.9.** (*Интегральная теорема Муавра–Лапласа*). При условиях предыдущей теоремы справедливо равенство

$$\lim_{\substack{n\to\infty \\ 0\le m_1\le m_2\le n}} P_n(m_1,m_2)=\frac{1}{\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{z^2}{2}}\,dz\,,$$

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

Доказательства теорем 1.8 и 1.9 достаточно сложны, и поэтому мы не будем на них останавливаться. Эти доказательства можно прочитать, например, в [10, § 2.6].

Функция

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

называется *интегральной функцией Лапласа*, ее значения в виде таблиц представлены в справочной литературе. При их использовании следует помнить о нечетности этой функции.

Пример. Вероятность того, что деталь проходит проверку контроля качества равна 0,2. Найти вероятность того, что среди 400 случайно отобранных деталей проверенных окажется от 70 до 100.

Решение. По условию: $n = 400$; $m_1 = 70$; $m_2 = 100$; $q = 0{,}8$; $p = 0{,}2$. Тогда

$$x_1=\frac{70-400\cdot 0{,}2}{\sqrt{400\cdot 0{,}2\cdot 0{,}8}}=-1{,}25\,,\ x_2=2{,}5.$$

И используя таблицы получаем

$$P_{400}(70,\,100)\approx \Phi(2{,}5)-\Phi(-1{,}25)=\Phi(2{,}5)+\Phi(1{,}25)\approx$$
$$\approx 0{,}4938+0{,}3944=0{,}8882.$$

Иногда более удобной оказывается другая форма интегральной теоремы.

***Теорема* 1.10.** Если вероятность успеха p в каждом испытании постоянна, то при $n\to\infty$,

$$P\{\,a\le\frac{m-np}{\sqrt{npq}}\le b\,\}\to\frac{1}{\sqrt{2\pi}}\int_a^b e^{\frac{-x^2}{2}}\,dx\,,$$

где m – число «успехов», a и b – любые числа $\in R^1$.

Из последней теоремы вытекает следующая важная формула

$$\gamma=P\left\{\left|\frac{m}{n}-p\right|\le\varepsilon\right\}\approx 2\,\Phi\left(\varepsilon\sqrt{\frac{n}{p\,q}}\right),$$

которая позволяет решать целый спектр важных задач. Например:

1) сколько нужно провести испытаний n, чтобы с заданной вероятностью γ, можно было быть уверенным, что наблюдаемая частота $\dfrac{m}{n}$ интересующего нас события отклонится от его истинной вероятности p не больше чем на ε;

2) с какой вероятностью γ, при определенном количестве экспериментов n, частота $\dfrac{m}{n}$ отклоняется не больше чем на ε от p.

Вопросы для самопроверки

1. Какие схемы проведения вероятностного эксперимента вы знаете? Приведите примеры соответствующих ситуаций.
2. В чем ограниченность классического определения вероятности?
3. Что такое благоприятствующий исход?
4. Зачем нужно статистическое определение вероятности?
5. Что такое условная вероятность?
6. Приведите пример полной группы случайных событий.
7. Что такое вероятностное пространство?
8. Приведите пример вероятностного эксперимента, соответствующего схеме Бернулли? Схеме с несчетным множеством исходов?
9. Может ли локальная формула Лапласса заменить формулу Пуассона?
10. В чем состоит несоответствие условий теоремы Пуассона схеме Бернулли?

2. ОСНОВЫ ТЕОРИИ СЛУЧАЙНЫХ ВЕЛИЧИН

2.1. Определение случайной величины. Задание дискретной случайной величины

***Определение* 1.25.** *Случайной величиной* (с.в.) называется функция $X(w)$, определенная на некотором множестве элементарных событий Ω.

Если Ω конечно, или счетно, то на функцию $X(w)$ не накладывается никаких ограничений, т.е. это может быть любая функция. Если Ω континиум, то $X(w)$ должна быть такой, что для любого события вида $A = \{w: X(w) \le x_0\}$, где x_0 – любое число $\in R^1$, могла быть определена его вероятность $P(A) = P\{x < x_0\}$. Следует отметить, что при изучении с.в. и их свойств почти никогда не вспоминают об области их определения, т.е. о Ω [1, 3, 4].

Все встречающиеся в природе процессы и объекты так или иначе характеризуются численными значениями своих параметров. Причем по целому ряду причин эти значения должны рассматриваться нами как случайные. Поэтому понятие с.в. имеет очень большую практическую значимость. Примеров с.в. можно привести бесконечное количество – это и температура воздуха на завтра, и курс доллара на банковских торгах, и количество покупателей в магазине, и масса, даже стандартно упакованного, пакета некоторого продукта, и т.д. Фактически теория с.в. составляет бóльшую часть содержания теории вероятностей, и всю математическую статистику.

***Определение* 1.26.** Конкретные значения, которые принимала с.в. в ходе соответствующих экспериментов или наблюдений, называются *реализациями* данной с.в.

Например, X– курс доллара на торгах ММВБ, это с.в. А, $x_1 = 26{,}37$ – курс 10 августа, $x_2 = 26{,}64$ – 5 сентября, – это реализации данной с.в.

Чаще всего с.в. обозначаются заглавными латинскими буквами, например, X, Y, Z, аргумент при этом обычно опускают. А их реализации – соответствующими строчными буквами – x, y, z, и т.д.

***Определение* 1.27.** Говорят, что с.в. X *задана*, или задан ее *закон распределения*, если для любого измеримого множества $B \in R^1$ определена вероятность

$$P\{x \in B\},$$

т.е. вероятность того, что при очередном эксперименте реализация x окажется лежащей в этом множестве.

***Определение* 1.28.** С.в. X называется *дискретной*, если множество ее возможных значений конечно, или счетно.

Если с.в. X имеет конечное количество возможных значений, то задать ее можно простым перечислением этих значений x_i, $i = \overline{1,n}$, и их соответствующих вероятностей p_i,

$$P\{X = x_i\} = p_i .$$

Из теорем теории вероятностей ясно, что при этом должно выполняться условие

$$\sum_{i=1}^{n} p_i = 1 .$$

Такое перечисление обычно записывают в виде таблицы (табл. 1.1), которую называют *рядом распределения* данной с.в.

Таблица 1.1

x_i	x_1	x_2	x_3	…	x_n
p_i	p_1	p_2	p_3	…	p_n

Пример. Пусть X – с.в. числа очков, выпадающих на игральном кубике. Ясно, что это дискретная с.в., а ее ряд распределения имеет вид, указанный в табл. 1.2.

Таблица 1.2

x_i	1	2	3	…	6
p_i	1/6	1/6	1/6	…	1/6

Если с.в. имеет счетное количество возможных значений, то ее задают с помощью формульного описания этих значений и их вероятностей. При этом, формульные выражения зависят от i, т.е. номера значения

$$P\{X = x_i(i)\} = p_i(i), \;\; i = \overline{1,\infty}.$$

При этом должно выполняться

$$\sum_{i=1}^{\infty} p_i = 1 .$$

Пример. Пусть X – с.в. числа бросаний монеты до первого выпадания орла. Ясно, что эта с.в. имеет счетное количество возможных значений, и

$$P\{X = i\} = \frac{1}{2^i}, \quad i = \overline{1,\infty}.$$

Следует отметить, что реальных примеров счетных с.в. не очень много.

2.2. Непрерывная с.в. Функция распределения

***Определение* 1.29.** С.в. называется *непрерывной*, если область ее возможных значений имеет более чем счетную мощность, т.е. включает хотя бы один непрерывный интервал числовой оси.

На практике как правило встречаются непрерывные с.в. Непрерывными считают даже денежные суммы, человеческие ресурсы, и т.д. Тем более таковыми являются различные физические величины, или относительные экономические показатели, и т.д.

Ясно, что задать непрерывную с.в. простым перечислением ее значений и их вероятностей невозможно.

Оказывается, что любое измеримое множество числовой оси можно представить в виде не более чем счетного объединения непересекающихся интервалов вида

$$[a;\, b],\ (a;\, b],\ [a;\, b),\ (a;\, b),$$

или их дополнений. Такие интервалы называют *простыми множествами*. Поэтому достаточно определить вероятность попадания реализации некоторой с.в. в любой такой интервал, и тогда, на основании теоремы сложения вероятностей, эта с.в. будет задана. Рассмотрим, как это можно сделать.

***Определение* 1.30.** *Функцией распределения* с.в. X называется такая функция $F_X(t)$, что $\forall t \in R^1$ выполняется

$$P\{x \le t\} = F_X(t).$$

Покажем, что с помощью функции распределения можно определить вероятность попадания реализации с.в. в любое простое множество:

1) для $(a;\, b]$ это почти очевидно, так как по теореме 1.5

$$P\{a < x \le b\} = P\{x \le b\} - P\{x \le a\},$$

тогда

$$P\{a < x \le b\} = F_X(b) - F_X(a);$$

2) для $[a;\, b]$. Пусть $\{t_i\}$ – некоторая числовая последовательность, стремящаяся к a слева, т.е. это последовательность чисел меньших a, и таких что

$$\lim_{i\to\infty} t_i = a,$$

но тогда и для последовательности интервалов справедливо

$$(t_i, b] \to [a, b],$$

а тем самым

$$P\{a \le x \le b\} = F_X(b) - \lim_{i\to\infty} F_X(t_i),$$

если только соответствующий предел слева функции $F_X(t)$ существует. Но для большинства реальных с.в. это выполняется;

3) аналогично и для $[a;\, b)$, $(a;\, b)$.

Таким образом функция распределения полностью задает любую, в том числе и непрерывную, с.в.

Пример.

1. Для дискретной с.в., имеющей ряд распределения

x_i	x_1	x_2	x_3	…	x_n
p_i	p_1	p_2	p_3	…	p_n

Ее функция распределения имеет кусочно-постоянный, ступенчатый вид (рис. 2.1). Ясно, что задавать дискретную с.в. с помощью функции распределения смысла нет;

2. Для так называемой *равномерно распределенной*, на отрезке $[a,\, b] \in R^1$, с.в., функция распределения задается следующим кусочно-аналитическим выражением

$$F_X(t)=\begin{cases}0, & \text{при } t<a;\\ \dfrac{t-a}{b-a}, & \text{при } a\le t\le b;\\ 1, & \text{при } t>b.\end{cases}$$

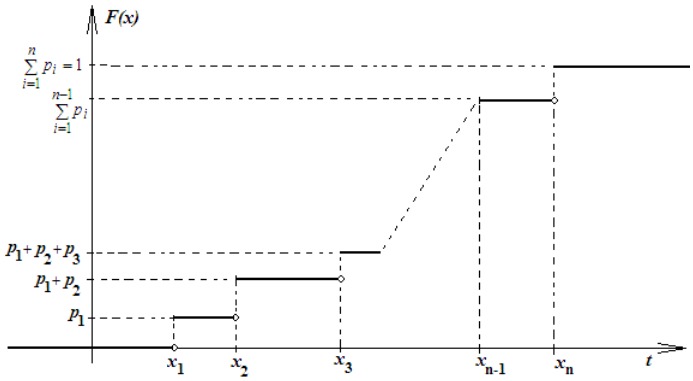


Рис. 2.1.

***Определение* 1.31.** С.в. называется *абсолютно непрерывной*, если ее функция распределения непрерывна на всей числовой оси.

Множеством возможных значений абсолютно непрерывной с.в., обычно является один единственный непрерывный интервал числовой оси, быть может с бесконечными границами. Большинство встречающихся на практике с.в. являются абсолютно непрерывными, поэтому в дальнейшем, говоря «непрерывная с.в.» , мы будем подразумевать абсолютно непрерывную с.в.

Достаточно очевидны следующие важные свойства функции распределения.

***Свойства функции распределения*:**

1) $0 \le F(x) \le 1$, для $\forall x \in R^1$;
2) $F(x)$ – неубывающая функция;
3) $F(x) \to 0$, при $x \to -\infty$; $F(x) \to 1$, при $x \to +\infty$.

2.3. Функция плотности распределения с.в.

Следующее понятие является чрезвычайно важным.

***Определение* 1.32.** Функция $P(x)$ называется *функцией плотности распределения* с.в. X (или просто *плотностью*), если для любого $(a, b] \in R^1$ выполняется

$$P\{x \in (a, b]\} = \int_a^b P(x)dx .$$

Ясно, что плотность тоже полностью задает с.в.

Из сказанного выше следует

$$\int_a^b P(x)dx = P\{x \in (a, b]\} = F_X(b) - F_X(a),$$

т.е. функция распределения является первообразной для плотности, т.е.

$$F'(x) = P(x), \quad F(x) = \int_{-\infty}^{x} P(t)dt .$$

Таким образом, зная плотность всегда можно найти функцию распределения, и наоборот. Однако, на практике для задания с.в. почти всегда используется плотность. Это связано со следующим обстоятельством.

Вспомним геометрический смысл определенного интеграла (рис. 2.2), а именно, что он численно равен площади криволинейной трапеции, ограниченной соответствующим участком графика функции на интервале интегрирования и самим этим интервалом.

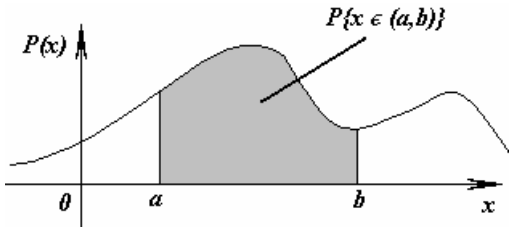


Рис. 2.2.

Таким образом, на тех участках числовой оси, где больше значения функции плотности, там и больше площадь под ее графиком, а тем самым и вероятность появления реализаций данной с.в. Поэтому график плотности весьма хорошо иллюстрирует закон распределения данной с.в. Этого нельзя сказать о графике функции распределения.

Пример.

1. *Равномерно распределенная*, на отрезке $[a, b] \in R^1$, с.в. Выражение для функции плотности распределения

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

А ее график – следующий вид (рис. 2.3).

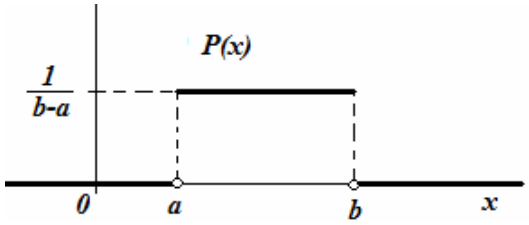


Рис. 2.3.

2. *Нормальное распределение*, называемое еще *гауссовским* распределением, играет важнейшую роль в теории вероятностей и математической статистике. Говорят, что с.в. X имеет нормальное распределение, если ее плотность задается выражением

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

где $\sigma > 0,\ a$ – заданные числа, параметры этого распределения. График нормальной плотности имеет вид (рис. 2.4).

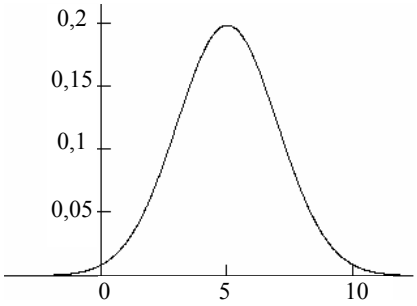


Рис. 2.4.

3. *Экспоненциальное* распределение

$$P(x) = \lambda e^{-\lambda x},$$

где λ – некоторый параметр. График этой плотности имеет вид (рис. 2.5).

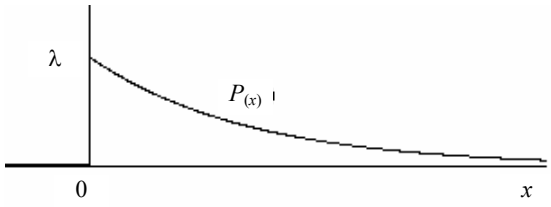


Рис. 2.5.

Считается, что экспоненциальное распределение имеет, например, случайное время обслуживания одного покупателя в магазине.

***Свойства функции плотности*:**

1. $P(x) \ge 0,$ при $\forall x \in R$;
2. $\int\limits_{-\infty}^{\infty} P(x)dx = 1$;
3. $P(x) \to 0$, при $x \to \pm\infty$.
4. $\int\limits_{-\infty}^{x} P(x)dx \to 0$, при $x \to -\infty$, $\int\limits_{x}^{+\infty} P(x)dx \to 0$, при $x \to +\infty$.

2.4. Математическое ожидание с.в.

***Определение* 1.33.** Любая числовая величина, так или иначе характеризующая закон распределения некоторой с.в., называется *параметром* этого распределения.

Из бесконечного количества всевозможных параметров, важнейшее значение имеют два: *математическое ожидание* и *дисперсия* с.в.

***Определение* 1.34.** *Математическим ожиданием* дискретной с.в. X, имеющей ряд распределения

x_i	x_1	x_2	x_3	…	x_n
p_i	p_1	p_2	p_3	…	p_n

называется число

$$M(x) = \sum_{i=1}^{n} x_i \, p_i \, .$$

Таким образом, математическое ожидание – это взвешенная сумма возможных значений с.в. с учетом их вероятностей. Математическое ожидание выражает, таким образом, некоторое «среднее» значение с.в. с учетом вероятностей всех ее возможных значений. Ясно, что это весьма важный параметр.

Часто математическое ожидание, так и называют – *среднее значение* с.в.

Пример. Найти математическое ожидание числа очков, выпадающих на кубике.

Решение. Ряд распределения этой с.в. приводился выше. Получаем

$$M(x) = \sum_{i=1}^{6} i \, \frac{1}{6} = \frac{1}{6} \cdot 21 = \frac{7}{2} = 3{,}5 \, .$$

Этот пример показывает, что математическое ожидание дискретной с.в. может не совпадать ни с одним из ее возможных значений.

***Определение* 1.35.** Математическим ожиданием непрерывной с.в., имеющей плотность распределения $P(x)$, называют число

$$M(x) = \int_{-\infty}^{\infty} x \, P(x) dx \, .$$

Пример.

1. Найдем математическое ожидание равномерно распределенной на отрезке $[a, \, b] \in R^1$ с.в.

В соответствии с определением, имеем

$$M(X) = \int_{-\infty}^{a} x \, 0 \, dx + \int_{a}^{b} x \frac{1}{b-a} dx + \int_{b}^{\infty} x \, 0 dx = \frac{x^2}{2(b-a)} \Big| \begin{matrix} b \\ a \end{matrix} = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2} \, .$$

Исходя из графика равномерной плотности этот результат можно считать вполне ожидаемым.

2. Для нормально распределенной с.в., можно доказать, что

$$M(X) = a \, ,$$

что также вполне согласуется с приведенным выше графиком нормальной плотности (рис. 2.4).

***Свойства математического ожидания*:**

1. $M(c) = c$, где c – любая константа;
2. $M(c \cdot X) = c \cdot M(X)$, где X – любая с.в.;
3. $M(X + Y) = M(X) + M(Y)$, где X, Y – любые с.в.;

Доказательство. Пусть имеется две дискретные с.в.

$$X : P\{ X = x_i \} = p_i, \; i = \overline{1, \, n}, \quad \text{и} \quad Y : P\{ Y = y_j \} = q_j, \; j = \overline{1, \, m} \, .$$

Тогда, используя формулы условной и полной вероятностей, имеем

$$M(X+Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j) P\{ X = x_i, Y = y_j \} =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j) P\{X = x_i\} P\{ Y = y_j \setminus X = x_i \} =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\, p_i\, P\{Y = y_j \setminus X = x_i\} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j\, P\{Y = y_j \setminus X = x_i\} =$$

$$= \sum_{i=1}^{n} x_i\, p_i \sum_{j=1}^{m} P\{Y = y_j \setminus X = x_i\} + \sum_{j=1}^{m} y_j \sum_{i=1}^{n} P\{Y = y_j \setminus X = x_i\} =$$

$$= \sum_{i=1}^{n} x_i\, p_i\, 1 + \sum_{j=1}^{m} y_j\, q_j = M(X) + M(Y).$$

Для непрерывных с.в. доказательство можно построить аналогично, используя переход от интегралов к пределам интегральных сумм, и наоборот. ■

4. $M(XY) = M(X)M(Y)$, где X, Y – любые независимые с.в. (см. п. 2.6).

2.5. Дисперсия случайных величин

***Определение* 1.36.** *Дисперсией* с.в. X называется число

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

т.е. математическое ожидание случайного квадрата отклонения значений с.в. от ее математического ожидания.

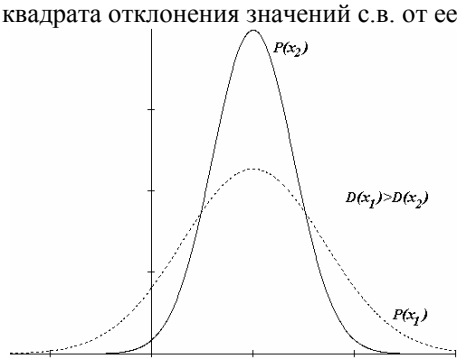


Рис. 2.6.

Таким образом, дисперсия характеризует величину разброса возможных значений с.в. вокруг ее среднего значения. Ясно, что это тоже очень важный параметр.

На рис. 2.6 приведены графики плотностей двух с.в., с одинаковым математическим ожиданием, но различными дисперсиями. Не сложно понять, у какой из с.в. дисперсия больше.

Используя свойства математического ожидания, легко доказать следующую, более удобную для практических расчетов, формулу дисперсии

$$D(x) = M(x^2 - 2xM(x) + M^2(x)) = M(x^2) - 2M(x)M(x) + M^2(x) = M(x^2) - M^2(x),$$

т.е.

$M(x^2) = \sum_{i=1}^{n} x_i^2 P_i$ – для дискретной с.в.,

$M(x^2) = \int_{-\infty}^{\infty} x^2 P(x)dx$ – для непрерывной с.в.

Дисперсия, по сути, является квадратическим показателем. Иногда более удобно использовать аналогичный линейный параметр, называемый *среднеквадратическим отклонением* с.в.

$$\sigma_X = \sqrt{D(X)}\,.$$

Пример.

1. Найти дисперсию равномерно распределенной на отрезке $[a, b] \in R^1$ с.в.

Находим сначала

$$M(X^2) = \int_a^b x^2 \frac{1}{b-a}dx = \frac{x^3}{3(b-a)}\Big|_a^b = \frac{a^2 + ab + b^2}{3}.$$

Окончательно получаем

$$D(X) = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}.$$

2. Для нормально распределенной с.в. можно доказать, что

$$D(X) = \sigma^2.$$

Таким образом важнейшее для практики нормальное распределение (как и многие другие) полностью задается двумя своими параметрами: математическим ожиданием и дисперсией. Тот факт, что с.в. X имеет нормальное распределение с параметрами

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

условно обозначается так

$$X \sim N\left(a, \sigma^2\right).$$

***Свойства дисперсии*:**

1. $D(X) \geq 0$, где X – любая с.в.;
2. $D(c) = 0$, где c – любая неслучайная константа;
3. $D(cX) = c^2 D(X)$;
4. $D(c + X) = D(X)$;
5. Если X и Y – независимые с.в., то $D(X+Y) = D(X) + D(Y)$.

2.6. Независимость с.в. и коэффициент корреляции

***Определение* 1.37.** С.в. X и Y называются *независимыми*, если закон распределения каждой из них не меняется от того, какие значения, в ходе соответствующих экспериментов, принимает вторая с.в.

Можно приводить множество примеров, как зависимых, так и независимых с.в. Курс доллара на торгах и счет в футбольном матче – очевидно независимые с.в. Курс доллара и объем продаж валюты – зависимые с.в., и т.д.

Выявление действительного наличия зависимости между какими-то показателями, и измерение силы этой зависимости, является важным инструментом самых разных исследований, в частности экономических. При этом используют следующие понятия.

***Определение* 1.38.** *Коэффициентом ковариации* между с.в. X и Y называют число

$$\operatorname{cov}(X, Y) = M((X - M(X))(Y - M(Y))).$$

Из свойств математического ожидания легко устанавливается, что для независимых с.в. ковариация равна нулю

$$\operatorname{cov}(X, Y) = 0.$$

Отсюда можно понять, что чем сильнее зависимы с.в., тем больше величина коэффициента ковариации. Важнейшим его недостатком как показателя силы зависимости с.в. является то, что его величина зависит и от абсолютных значений данных с.в. От этого недостатка свободен следующий показатель.

***Определение* 1.39.** *Коэффициентом корреляции* между с.в. X и Y называют число

$$r_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)D(Y)}}.$$

Как видим, коэффициент корреляции является нормированным показателем, т.е. он не зависит от абсолютных значений с.в. При независимости с.в. он также равен нулю. При наличии строго детерминированной (максимальной по силе) линейной зависимости, т.е. если

$$Y = a X + b,$$

можно доказать, что

$$r_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)D(Y)}} = \frac{\operatorname{cov}(X, aX+b)}{\sqrt{D(X)D(aX+b)}} = \frac{a\operatorname{cov}(X, X)}{\sqrt{D(X)a^2 D(X)}} = \frac{a}{|a|} = \pm 1.$$

Таким образом, коэффициент корреляции может принимать значения от –1 до 1. Важнейшим недостатком коэффициента корреляции является то, что он выражает лишь силу линейной составляющей в той или иной зависимости между с.в. Так, например, несложно доказать, что при наличии строго детерминированной квадратичной зависимости

$$Y = X^2$$

коэффициент корреляции окажется равным нулю. Тем не менее этот коэффициент является наиболее часто используемым на практике показателем, как по причине простоты, так и потому, что обычно в реальных зависимостях линейная составляющая весьма велика, или даже преобладает (рис. 2.7).

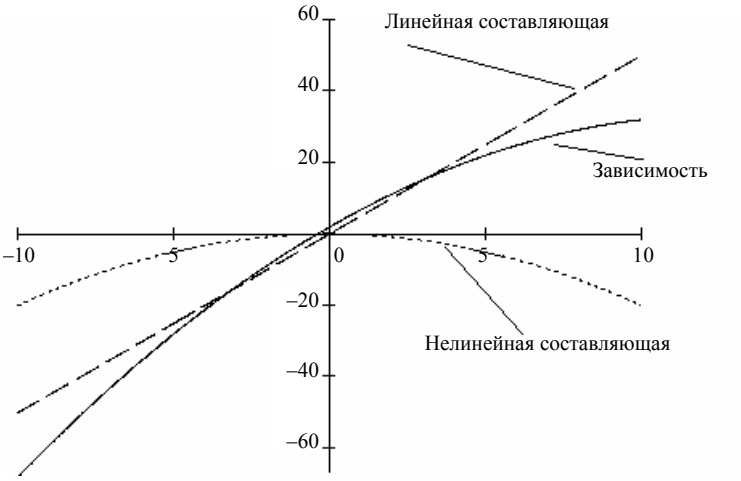


Рис. 2.7.

2.7. Неравенство Чебышева

Для решения различных задач и доказательства теорем необходимыми оказываются следующие утверждения.

Лемма (Маркова). Если с.в. X может принимать только неотрицательные значения, то для любого $t > 0$ выполняется

$$P\{x > t\} \le \frac{M(X)}{t}.$$

Доказательство. Имеем

$$P\{x > t\} = \int_t^\infty P(x)dx = \int_t^\infty \frac{x}{x} \cdot P(x)dx \le \int_t^\infty \frac{x}{t} \cdot P(x)dx \le \frac{1}{t} \cdot \int_0^\infty x \cdot P(x)dx = \frac{M(X)}{t}. \blacksquare$$

Следствие. В условиях предыдущего утверждения

$$P\{x \le t\} > 1 - \frac{M(X)}{t}.$$

***Теорема* 2.1.** (*Неравенство Чебышева*). Для любой с.в. X выполняется

$$P\{|x - M(X)| > \varepsilon\} \le \frac{D(X)}{\varepsilon^2},$$

где $\varepsilon > 0$ – любое число.

Доказательство. Достаточно применить лемму Маркова к неотрицательной с.в. $Z = (X - M(X))^2$. $\blacksquare$

Следствие. Для любой с.в. X выполняется

$$P\{|x - M(X)| \le \varepsilon\} > 1 - \frac{D(x)}{\varepsilon^2},$$

где $\varepsilon > 0$ – любое число.

2.7. Закон больших чисел

Из неравенства Чебышева вытекает следующее утверждение.

***Теорема* 2.2.** (*Закон больших чисел в форме Чебышева*).
Пусть X_i – независимые с.в., такие, что

$$M(X_i) = a_i, \quad D(X_i) \le C, \quad i = \overline{1,n}.$$

Тогда

$$\lim_{n\to\infty} P\left\{\left|\frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n a_i}{n}\right| \le \varepsilon\right\} = 1, \quad \forall\ \varepsilon > 0.$$

Доказательство. Рассмотрим с.в.

$$X = \frac{\sum_{i=1}^{n} X_i}{n},$$

из свойств математического ожидания и дисперсии

$$M(X) = \frac{\sum_{i=1}^{n} a_i}{n}, \quad D(X) = D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} D(X_i)}{n^2} \le \frac{C}{n}.$$

Применяя к X следствие неравенства Чебышева

$$P\{|X - M(X)| \le \varepsilon\} > 1 - \frac{C}{\varepsilon^2 n} \to 1,$$

при $n \to \infty$. Откуда и получаем утверждение теоремы. ■

Следствие. (*Закон больших чисел для одинаково распределенных с.в.*).

Пусть X_i – независимые с.в., такие, что

$$M(X_i) = a, \;\; D(X_i) = \sigma^2, \;\; i = \overline{1,n}.$$

Тогда

$$\lim_{n\to\infty} P\left\{\left|\frac{\sum_{i=1}^{n} x_i}{n} - a\right| \le \varepsilon\right\} = 1, \quad \forall\ \varepsilon > 0.$$

Рассмотренные утверждения являются весьма важными как теоретически, так и практически. Они устанавливают так называемый закон *статистической устойчивости* для значений с.в., а именно, хотя каждое из слагаемых является случайным, их среднее значение тем более неслучайно, чем больше этих слагаемых, т.е., при достаточно большом количестве наблюдений в случайности всегда будет прослеживаться закономерность. Тем самым, например, эти утверждения обосновывают возможность применения анализа статистик при принятии конкретных управленческих решений.

Также важным является следующий частный случай, который, помимо прочего, обосновывает статистическое определение вероятности.

***Теорема* 2.3.** (*Закон больших чисел в форме Бернулли*). Для схемы Бернулли

$$\lim_{n\to\infty} P\left\{\left|\frac{m}{n} - p\right| \le \varepsilon\right\} = 1, \;\; \forall\ \varepsilon > 0.$$

Данное утверждение вытекает и из формулы, рассматривавшейся в п. 1.10. Говорят, что теорема 2.3 устанавливает *статистическую устойчивость вероятности события*.

2.8. Центральная предельная теорема

Хорошо известно, что большинство встречающихся в природе с.в. имеют распределение, по крайней мере, очень близкое к нормальному. Теоретическое объяснение этому факту дает следующая очень важная теорема [7].

***Теорема* 2.4.** (*Центральная предельная или теорема Ляпунова*).

Пусть $X_1, X_2, ..., X_n$ – независимые с.в., имеющие какие угодно распределения, с параметрами

$$M(X_i) = a_i, \;\; D(X_i) = \sigma_i^2, \;\; M(X - M(X))^3 = C_i, \;\; i = \overline{1,n},$$

и если

$$\lim_{n\to\infty} \frac{\sqrt[3]{C_1 + C_2 + ... + C_n}}{\sqrt{\sigma_1^2 + ... + \sigma_n^2}} = 0,$$

тогда

$$P\left\{\frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} a_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} < t\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} dz,$$

т.е. закон распределения с.в.

$$Y_n = \sum_{i=1}^{n} X_i ,$$

при $n \to \infty$, асимптотически приближается к нормальному, с параметрами

$$M(Y_n) = \sum_{i=1}^{n} a_i ,\ D(Y_n) = \sum_{i=1}^{n} \sigma_i^2 .$$

Доказательство этой теоремы достаточно сложно, и поэтому мы его рассматривать не будем.

Проще говоря, утверждение теоремы состоит в том, что если на некоторую с.в. влияет большое количество других независимых с.в., то ее распределение будет близко к нормальному. Но в природе так и бывает, на большинство процессов и объектов оказывает влияние большое количество разнообразных случайных факторов, и поэтому параметры этих процессов и оказываются случайными, и имеющими почти нормальное распределение. Хотя необходимо отметить, что все же так бывает не всегда. И поэтому часто возникает необходимость проверки, можно ли считать ту или иную с.в. нормально распределенной или нет?

Забегая вперед укажем, что вся классическая теория математической статистики строится только для нормальных с.в., и при этом считается применимой почти всегда, именно в силу центральной предельной теоремы.

Пример. *В* кассе учреждения имеется сумма $d = 3500$ р., в очереди стоит $n = 20$ человек, сумма X_i, которую нужно выплатить отдельному человеку, является с.в. со средним значением 150 р. и $\sigma_X = 60$. Найти:

1. Вероятность того, что суммы d не хватит для выплаты всем людям из очереди.
2. Какую сумму d нужно иметь в кассе, чтобы с вероятностью $p = 0{,}995$ ее хватило всей очереди.

Решение. При $n = 20$ уже можно считать, что с.в. $Y = \sum_{i=1}^{20} X_i$ достаточно близка к нормальной, с параметрами

$$M(Y) = \sum_{i=1}^{20} M(X_i) = 150 \cdot 20 = 3000 ,\ D(Y) = \sum_{i=1}^{20} \sigma_i^2 = 20 \cdot 3600 = 72\,000 .$$

Тогда:

1. $P\{Y > d = 3500\} = 1 - P\{Y \le 3500\} = 1 - \Phi\left(\dfrac{3500 - 3000}{\sqrt{72000}}\right) \approx 0{,}032$;

2. $P\{Y < d\} \ge 0{,}995 \Rightarrow \Phi\left(\dfrac{d - 3000}{\sqrt{72000}}\right) \ge 0{,}995$, по таблицам находим,

что для этого должно быть $\left(\dfrac{d - 3000}{\sqrt{72000}}\right) \ge 2{,}58 \Rightarrow d \ge 3692$.

2.9. Многомерные случайные величины

Обычно, тот или иной процесс, или случайным образом отобранный объект, характеризуется не одним, а сразу несколькими своими параметрами (например, коммерческое предприятие и т.д.). Поэтому, оказывается необходимым следующее понятие.

***Определение* 1.40.** *n-мерной с.в. X*, называется вектор

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

все элементы которого являются случайными величинами.

Иногда многомерную с.в. называют *случайным вектором*. Основные факты для многомерных с.в. совершенно аналогичны случаю одномерных с.в.

***Определение* 1.41.** *Математическим ожиданием n-мерной с.в. X* называется вектор

$$M(X) = \begin{pmatrix} M(X_1) \\ M(X_2) \\ \vdots \\ M(X_n) \end{pmatrix}.$$

***Определение* 1.42.** Плотностью распределения n-мерной с.в. X называется функция $P_X(t_1, t_2, ..., t_n)$, определенная на некотором подмножестве R^n, такая, что

$$P\{x \in Q\} = \int_Q P_X(t_1, t_2, ..., t_n)\, dq\,,$$

для любого измеримого множества $Q \in R^n$.

Как и всегда для функции нескольких аргументов, в общем случае, график плотности распределения многомерной с.в., к сожалению, невозможно изобразить на плоскости. Что, естественно, является дополнительной сложностью.

***Теорема* 2.5.** Если элементы случайного вектора являются независимыми с.в., то для ее плотности распределения справедливо равенство

$$P_X(t_1, t_2, ..., t_n) = P_{X_1}(t_1) P_{X_2}(t_2) \cdot ... \cdot P_{X_n}(t_n)\,.$$

Для дискретной многомерной с.в. доказательство вполне очевидно. Для непрерывной оно строится на переходе от интеграла к пределу интегральной суммы и обратно.

***Определение* 1.43.** *Ковариационной матрицей* n-мерной с.в. X называется матрица

$$V(X) = \begin{pmatrix} \operatorname{cov}(x_1, x_1) & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_n) \\ \operatorname{cov}(x_2, x_1) & \operatorname{cov}(x_2, x_2) & \cdots & \operatorname{cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_n, x_1) & \operatorname{cov}(x_n, x_2) & \cdots & \operatorname{cov}(x_n, x_n) \end{pmatrix},$$

т.е. матрица, составленная из коэффициентов ковариации между элементами данной с.в.

Можно отметить, что ковариационная матрица является квадратной и симметричной, а на ее главной диагонали стоят дисперсии соответствующих компонент данного случайного вектора.

Ковариационная матрица многомерной с.в. фактически, играет роль дисперсии одномерной. Так, например, плотность, распределения
n-мерной нормально распределенной с.в. X описывается функцией

$$P_X(t) = \frac{1}{\sqrt{(2\pi)^n}\sqrt{|V(X)|}} e^{-\frac{(t-M(X))^T [V(X)]^{-1} (t-M(X))}{2}},$$

где t – вектор-аргумент соответствующей размерности.

2.10. Функции от случайных величин

Можно задаться вопросом: если задано распределение с.в. X – $P(x)$ и некоторая функция $g(\bullet)$, то каким будет распределение с.в. $Y = g(X)$?

Оказывается, в общем случае ответить на этот вопрос нельзя. Но существует несколько важных частных случаев. Рассмотрим два из них:

1. Пусть $g(\bullet)$ – монотонная функция (монотонно возрастающая или монотонно убывающая), определенная на некотором множестве $D(g) \in R$, тогда, как известно, существует функция $g^{-1}(\bullet)$, обратная к ней, и определенная на множестве значений функции $g(\bullet)$. Причем $g^{-1}(\bullet)$ также будет монотонной. Пусть для определенности $g(\bullet)$ монотонно возрастающая. Тогда

$$F_Y(t) = P\{Y \le t\} = P\{g(x) \le t\} = P\left\{x \le g^{-1}(t)\right\} = F_X\left(g^{-1}(t)\right).$$

Аналогично, для монотонно убывающей $g(\bullet)$ можно получить

$$F_Y(t) = 1 - F_X\left(g^{-1}(t)\right).$$

2. Пусть X_1 и X_2 независимые с.в. с заданными функциями плотности распределения $P(x_1)$ и $P(x_2)$. Рассмотрим распределение с.в. $Y = X_1 + X_2$.

Имеем

$$F_Y(t) = P\{X_1 + X_2 \le t\} = \iint_{A_X} P(x_1) P(x_2)\, dx_1 dx_2 = ...\,,$$

где область интегрирования A_X описывается неравенством $x_1 + x_2 \le t$, тогда

$$... = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t-x_1} P(x_1) P(x_2)\, dx_2 dx_1 \right) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t} P(x_1) P(t-x_1)\, dt\, dx_1 \right) =$$

$$= \int_{-\infty}^{t} \int_{-\infty}^{\infty} P(x_1) P(t-x_1)\, dx_1\, dt.$$

Последнее означает, что

$$P_Y(t) = \int_{-\infty}^{\infty} P(x_1)\,P(t-x_1)\,dx_1\,,$$

иногда формулы подобного вида называют *формулами свертки*.

Пример. Пусть $X \sim N(a_X, \sigma_X^2)$, $Y \sim N(a_Y, \sigma_Y^2)$ – независимые с.в. Тогда плотность распределения с.в. $Z = X + Y$ будет иметь вид

$$P_Z(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_X} e^{\frac{-(x-a_X)^2}{2\sigma_X^2}} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{\frac{-(t-x-a_Y)^2}{2\sigma_Y^2}} dx =$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}} e^{\frac{-(t-(a_X+a_Y))^2}{2(\sigma_X^2+\sigma_Y^2)}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\frac{\sigma_Y\sigma_X}{\sqrt{\sigma_X^2+\sigma_Y^2}}} e^{-\frac{(x-f(a_X,a_Y,\sigma_Y,\sigma_X,t))^2}{2\frac{\sigma_Y^2\sigma_X^2}{\sigma_X^2+\sigma_Y^2}}} dx,$$

где $f(a_X, a_Y, \sigma_Y, \sigma_X, t)$ – некоторое выражение от заданных значений. Видим, что последнее подынтегральное выражение является плотностью нормально распределенной с.в., с математическим ожиданием $f(a_X, a_Y, \sigma_Y, \sigma_X, t)$, а значит сам интеграл равен единице. Тем самым

$$X + Y \sim N(a_X + a_Y, \sigma_X^2 + \sigma_Y^2).$$

Таким образом, установлен следующий очень важный факт.

***Теорема* 2.6.** Сумма двух нормально распределенных с.в. имеет нормальное распределение.

Вопросы для самопроверки

1. Что значит задать с.в.?
2. Что такое дискретная с.в., и ее ряд распределения?
3. Почему теория с.в. так важна для практики?
4. Какие вы знаете основные дискретные распределения?
5. Что такое дробноминомиальное и гипергеометрическое распре-деления?
6. Какие вы знаете основные непрерывные распределения?
7. Почему так важно нормальное распределение?
8. Сколькими параметрами задается эспоненциальное распределение?
9. Что выражают матеиатическое ожидание с.в. и дисперсия?
10. Приведите пример многомерной с.в.
11. Приведите примеры независимых, слабо зависимых и сильно зависимых с.в.
12. Запишите выражение для плотности двумерной равномерно-распределенной с.в. Сколькими параметрами она задается? Найдите ее математическое ожидание и дисперсию.

3. Оценка параметров и закона распределения с.в.

3.1. Основные понятия выборочного метода

Математическая статистика является одним из наиболее широко используемых на практике разделов прикладной математики [1, 3].

***Определение* 3.1.** *Генеральной совокупностью* называется вероятностное пространство и определенная на этом пространстве с.в. X.

Проще говоря, под генеральной совокупностью подразумевают всю совокупность изучаемых объектов, или все множество возможных значений изучаемой с.в.

***Определение* 3.2.** *Случайной выборкой* или просто *выборкой* объема n значений с.в. X называется набор чисел

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

являющихся реализациями с.в. X, полученными в ходе n соответствующих независимых наблюдений или экспериментов.

Иногда выборку также называют *статистическими данными*, или *статистикой реализаций* с.в., или просто *статистикой*.

Различают *повторные* и *бесповторные* выборки. Говорят о *собственно-случайных*, *механических*, *серийных* и *типических* выборках.

Основной задачей математической статистики является оценка и анализ параметров распределения изучаемой с.в., или самого вида этого распределения (непараметрическое оценивание) по данным выборки ее значений. Часто отмечают, что основная задача математической статистики
является, в некотором смысле, обратной к задачам теории вероятностей.

Основными целями оценивания являются:

1) прогнозирование поведения изучаемой с.в. в будущем;
2) проверка соответствия значений полученных оценок некоторым регламентированным характеристикам.

И то и другое часто может служить обоснованием выбора наиболее оптимального варианта управленческих решений.

***Определение* 3.3.** Истинное значение того или иного параметра распределения изучаемой с.в. X называют его *теоретическим*, или *генеральным* значением.

***Определение* 3.4.** *Статистической оценкой* некоторого параметра γ распределения с.в. X называется функция, определенная на множестве выборок значений этой с.в.

$$\hat{\gamma} = \hat{\gamma}_n(x_1,\ x_2,\ ...,\ x_n)\ ,$$

значения которой, в некотором статистическом смысле, близки к теоретическому значению γ.

***Определение* 3.5.** Конкретное значение статистической оценки на данной конкретной выборке называют *выборочным значением* этой оценки, или *точечным* значением, или *выборочной оценкой*.

Важно четко понимать разницу между теоретическим значением параметра и его выборочными оценками.

Теоретическое значение того или иного параметра распределения с.в. в общем случае определяется только его плотностью распределения, т.е. бесконечной информацией. Данные же выборки всегда конечны. Поэтому никогда нет возможности найти точное теоретическое значение параметра.

Очень важно также, что выборочное значение параметра само является случайной величиной, поскольку рассчитывается по данным случайной выборки.

Иногда приходится говорить не о конкретной выборке, а вообще, абстрактно. Как, например, при выводе различных статистических формул и доказательствах теорем. Тогда приходится использовать следующее понятие.

***Определение* 3.6.** *Теоретической выборкой* объема n значений с.в. X будем называть совокупность независимых с.в.

$$X_1,\ X_2,....,\ X_n\ ,$$

каждая из которых имеет то же распределение, что и X, т.е.

$$X_i = X,\ \ i = \overline{1,n}\ .$$

3.2. Свойства статистических оценок

Для оценок, которые предполагается использовать на практике, очевидно, очень желательны следующие свойства.

***Определение* 3.7.** Оценка $\hat{\gamma}$ некоторого параметра γ называется *несмещенной*, если

$$M(\hat{\gamma}) = \gamma\ .$$

Несмещенность оценки означает, что она не дает какой-либо регулярной (постоянной) ошибки. Иногда на практике используют и смещенные оценки.

***Определение* 3.8.** Оценка $\hat{\gamma}_n$ называется *состоятельной*, если

$$P\{|\hat{\gamma}_n - \gamma| \le \varepsilon\} \to 1\ ,$$

при $n \to \infty$, для любого $\varepsilon > 0$.

Таким образом, состоятельность оценки означает, что чем больше объем выборки, тем относительно точнее выборочная оценка. Примеры несостоятельных оценок достаточно редки.

Справедлива теорема.

***Теорема* 3.1.** Если $\hat{\gamma}_n$ – несмещенная оценка параметра γ и

$$D(\hat{\gamma}_n) \to 0\ ,$$

при $n \to \infty$, то $\hat{\gamma}_n$ – состоятельна.

Доказательство этой теоремы легко вытекает из неравенства Чебышева.

***Определение* 3.9.** Оценка называется *эффективной*, если она имеет минимальную дисперсию среди всех других оценок данного параметра.

3.3. Оценка математического ожидания с.в.

Из теории вероятностей известно, что среди всех параметров распределения с.в. важнейшую роль играют математическое ожидание и дисперсия. Напомним, что нормальное распределение полностью задается этими двумя параметрами. Поэтому и в математической статистике их оценки занимают центральное место.

Поэтому, для решения основной задачи математической статистики зачастую бывает достаточно оценить именно эти два параметра.

Пусть X – некоторая с.в. Как всегда, будем обозначать

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

где a, σ – неизвестны. Пусть имеется выборка значений этой с.в.

$$x_1, x_2, \ldots, x_n.$$

Вспоминая определение и смысл математического ожидания в качестве его оценки естественно предложить следующую

$$\hat{a} = \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}\ .$$

Величину $\overline{X}$ называют *выборочным средним*, и действительно используют в качестве оценки математического ожидания. Изучим ее свойства как статистической оценки.

Оказывается

$$M\left(\overline{X}\right) = M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \ldots\ ,$$

переходя к теоретической выборке, и используя свойства математического ожидания, имеем

$$\ldots = M\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{\sum_{i=1}^{n} a}{n} = a\ .$$

Таким образом, выборочное среднее – несмещенная оценка дисперсии.

Рассмотрим состоятельность. Используя свойства дисперсии, имеем

$$D\left(\overline{X}\right) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sum_{i=1}^{n} D(X_i)}{n^2} = \frac{n\ \sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0\ ,\quad \text{при } n \to \infty.$$

Тогда по теореме 3.1, $\overline{X}$ – состоятельная оценка математического ожидания.

Рассмотрим эффективность. Все возможные *линейные* оценки математического ожидания имеют вид

$$\hat{a}(C_1\ ,\ \ldots,\ C_n) = \sum_{i=1}^{n} C_i\ x_i\ ,$$

где C_i – любые весовые коэффициента, т.е.

$$\sum_{i=1}^{n} C_i = 1\ .$$

Отметим, что все такие оценки и несмещены, и состоятельны.

Найдем, при каких значениях коэффициентов C_i дисперсия соответствующей взвешенной суммы будет наименьшей, т.е. имеем задачу

$$D(\hat{a}(C_1, \ldots, C_n)) \to \min\ ,$$

при ограничении

$$\sum^{n} C_i = 1\ .$$

Это задача на условный экстремум, которую мы решим методом множителей Лагранжа. Составляем функцию Лагранжа:

$$L = D(\hat{a}(C_1, \ldots, C_n)) + \lambda\left[\sum^{n} C_i - 1\right],$$

где

$$D(\hat{a}(C_1, \ldots, C_n)) = D\left(\sum C_i\ x_i\right) = \sum^{n} C_i^2\ D(X_i) = \sigma^2 \sum C_i^2\ .$$

Тогда

$$L = \sigma^2 \sum C_i^2 + \lambda\left[\sum^{n} C_i - 1\right].$$

Система необходимых условий минимума функции Лагранжа

$$\begin{cases} \dfrac{\partial L}{\partial C_i} = 2C_i\,\sigma^2 + \lambda = 0, \;\; i = \overline{1,n}, \\ \dfrac{\partial L}{\partial \lambda} = \sum\limits_{i=1}^{n} C_i - 1 = 0, \end{cases} \Rightarrow$$

$$\begin{cases} C_i = -\lambda / 2\sigma^2 \\ \sum C_i = 1 \end{cases} \Rightarrow \sum^{n} \frac{\lambda}{2\sigma^2} = \frac{n\lambda}{2\sigma^2} = -1 \Rightarrow \lambda = -\frac{2\sigma^2}{n} \Rightarrow C_i = \frac{1}{n}.$$

Таким образом, минимум дисперсии достигается при тех самых коэффициентах, которые входят в $\overline{X}$, т.е. $\overline{X}$ – эффективная оценка.

3.4. Оценки дисперсии с.в.

Возможны два случая:

1. a – известно. Тогда в качестве оценки дисперсии естественно предложить следующую функцию

$$S_a^2 = \frac{1}{n}\sum^{n}(x_i - a)^2 .$$

Изучаем несмещенность

$$M(S_a^2) = M\left[\frac{1}{n}\sum^{n}(x_i - a)^2\right] = \left[\frac{1}{n}\sum M(X_i - a)^2\right] = \frac{1}{n}\sum^{n} D(X_i) = D(X) = \sigma^2 ,$$

т.е., рассматриваемая оценка является несмещенной.

Рассмотрим состоятельность этой оценки

$$\lim_{n\to\infty} D\left(S_a^2\right) = \lim_{n\to\infty} D\left(\frac{1}{n}\sum^{n}(x_i - a)^2\right) = \lim_{n\to\infty} \frac{D\left(\sum X_i^2 - 2a X_i + a^2\right)}{n^2} =$$

$$= \lim_{n\to\infty} \frac{n D\left(X^2 - 2aX\right)}{n^2} = 0,$$

если только

$$D\left(X^2 - 2aX\right) < \infty ,$$

т.е. дисперсия с.в. $\left(X^2 - 2aX\right)$ конечна, что является весьма не обременительным предположением.

Эффективность доказывается по схеме, аналогичной той, что была применена для математического ожидания.

2. a – неизвестно. При этом, естественно предложить оценку

$$\widetilde{S}^2 = \frac{1}{n}\sum^{n}\left(x_i - \overline{X}\right)^2 .$$

Изучаем ее несмещенность

$$M(\widetilde{S}^2) = M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] =$$

$$= M\left[\frac{1}{n}\sum (X_i - a - (\overline{X} - a))^2\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i - a)^2 - 2(\overline{X} - a)(X_i - a) + (\overline{X} - a)^2\right)\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - \frac{2}{n}(\overline{X} - a)\sum_{i=1}^{n}(X_i - a) + (\overline{X} - a)^2\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - 2(\overline{X} - a)(\overline{X} - a) + (\overline{X} - a)^2\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - (\overline{X} - a)^2\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - a)\right)^2\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - \frac{1}{n^2}\sum_{i=1}^{n}(X_i - a)\sum_{j=1}^{n}(X_j - a)\right] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - a)(X_j - a)\right] = \ldots ,$$

здесь при $i \neq j$,

$$M\left[(X_i - a)(X_j - a)\right] = 0 ,$$

так как X_i – независимые с.в.

$$\ldots = \frac{1}{n}\sum_{i=1}^{n}M(X_i - a)^2 - \frac{1}{n^2}M\sum_{i=1}^{n}(X_i - a) = \frac{1}{n}n\sigma^2 - \frac{1}{n^2}n\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\frac{n-1}{n} \neq \sigma^2 .$$

Таким образом, рассматриваемая оценка является смещенной. Однако видим, что это легко исправить перейдя к оценке

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1},$$

которую так и называют *несмещенной* или *исправленной* оценкой дисперсии.

Аналогично рассмотренному выше, не сложно доказать состоятельность как $\widetilde{S}^2$, так и S^2.

Очевидно, дисперсия оценки S^2 больше, чем у $\widetilde{S}^2$, и поэтому S^2 не является эффективной. Тем не менее, свойство несмещенности считается практически более значимым, и для оценки дисперсии обычно используется именно S^2. Однако и $\widetilde{S}^2$ находит практическое применение, в частности, величина $\widetilde{S}$ называется *стандартной ошибкой выборки*.

3.5. Оценка доли признака

Для весьма широкого спектра задач важным является следующий параметр.

Определение 3.10. *Генеральной долей* признака называется величина

$$p = \frac{M}{N},$$

где N – общее количество объектов, составляющих некоторую генеральную совокупность, M – количество объектов из этой совокупности, обладающих некоторым соответствующим признаком.

Например, N – количество всех работников на предприятии, M – число женщин из них.

Во многих маркетинговых, социологических и прочих исследованиях требуется оценить генеральную долю признака. Например, долю потенциальных покупателей, предпочитающих некоторый определенный товар. При этом невозможно протестировать всю совокупность N, а следует сделать выводы только на основании некоторой выборки объема n.

Определение 3.11. *Выборочной долей* признака называется величина

$$w = \frac{m}{n},$$

где n – общее количество протестированных объектов из соответствующей генеральной совокупности, m – количество объектов из n, обладающих интересующим нас признаком.

Следует ли считать, что w можно использовать, как оценку p?

Во-первых, вспомним, что выборки бывают повторные и бесповторные. При оценке доли признака, тип выборки имеет большое значение. В случае повторной выборки, тестируемый объект возвращается в генеральную совокупность, и может оказаться выбранным еще, и еще раз. Например, на бензозаправке решили выяснить, какова доля иномарок среди всех автомобилей. Ясно, что при этом за исследуемый день возможно неоднократное посещение одной и той же иномаркой данной бензозаправки. В случае бесповторной выборки тестируемый объект помечается и в дальнейшем не учитывается, т.е. исключается из генеральной совокупности. Чаще имеют место именно такие ситуации.

Не сложно понять, что при бесконечном количестве объектов в генеральной совокупности разницы между повторной и бесповторной выборками нет.

Рассмотрим сначала более простой случай повторной выборки.

Тогда каждый раз вероятность выбора объекта, обладающего признаком, по классическому определению вероятности, составляет

$$p = \frac{M}{N},$$

а вероятность, что из n отборов объектов с признаком попадется ровно m, выражается формулой Бернулли, т.е. в этом случае $\frac{m}{n}$ – с.в. имеющая, так называемое, *дробно биномиальное* распределение, для которого, как известно,

$$M\left(\frac{m}{n}\right) = p\,,$$

т.е. w – несмещенная оценка генеральной доли. Дисперсия дробно биномиальной с.в.

$$D\left(\frac{m}{n}\right) = \frac{p\,q}{n} \to 0\,,$$

при $n \to \infty$, т.е. w – состоятельная оценка.

В случае бесповторной выборки эксперимент соответствует «урновой» схеме, и m имеет, так называемое, *гипергеометрическое распределение*. Тогда

$$M\left(\frac{m}{n}\right) = \frac{1}{n}\sum_{i=0}^{n} i\,\frac{C_M^i\,C_{N-M}^{n-i}}{C_N^n}\,,$$

и можно доказать, что

$$M\left(\frac{m}{n}\right) = \frac{M}{N} = p\,,$$

т.е. опять w – несмещенная оценка. При этом

$$D\left(\frac{m}{n}\right) = \frac{p\,q}{n}\,\frac{N-n}{N-1} \to 0\,,$$

при $n \to \infty$. А, точнее говоря, предельно возможное значение для n это N, и при $n = N$ эта дисперсия равна нулю, т.е. опять w – состоятельная оценка.

3.6. Стандартные статистические распределения и их критические границы

Во многих формулах математической статистики используются, так называемые, *стандартные статистические* распределения. К ним, прежде всего, относят: *стандартное гауссовское* распределение, а также распределения Пирсона, Стьюдента и Фишера.

***Определение* 3.11.** Говорят, что с.в. X имеет *стандартное гауссовское* распределение, если

$$X \sim N(0,1)\,.$$

***Определение* 3.12.** Пусть $\varepsilon_1, \ldots, \varepsilon_n$ – независимые стандартные гауссовские с.в., тогда распределение с.в.

$$z = \sum_{i=1}^{n} \varepsilon_i^2$$

называют распределением Пирсона с n степенями свободы и пишут

$$z \sim \chi^2(n)\,.$$

График плотности χ^2-распределения зависит от числа степеней свободы, однако схематически он имеет вид, представленный на рис. 3.1.

***Определение* 3.13.** Пусть $\varepsilon_0, \ldots, \varepsilon_n$ – независимые стандартные гауссовские с.в., тогда распределение с.в.

$$z = \frac{\varepsilon_0}{\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\varepsilon_i^2}}$$

называют распределением Стьюдента с n степенями свободы, и пишут

$$z \sim t(n)\,.$$

График плотности t-распределения зависит от числа степеней свободы, однако в целом он подобен стандартному гауссовскому.

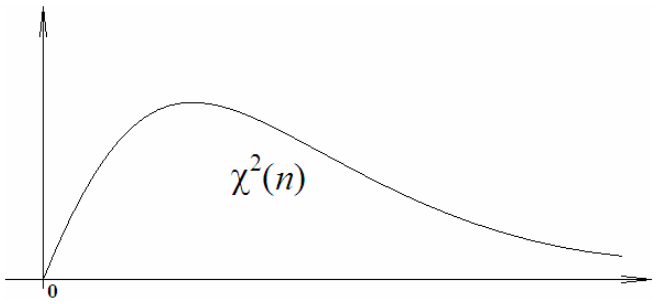


Рис. 3.1.

Замечание. Распределение Стьюдента с n степенями свободы можно было бы определить и так

$$t = \frac{\varepsilon_0}{\sqrt{\frac{1}{n}\,\chi^2(n)}}.$$

***Определение* 3.14.** Пусть $z_1 \sim \chi^2(n)$, $z_2 \sim \chi^2(m)$ – независимые с.в., тогда распределение с.в.

$$z = \frac{\frac{1}{n}\,z_1}{\frac{1}{m}\,z_2}$$

называют распределением Фишера с n степенями свободы числителя и m знаменателя, и пишут

$$z \sim F\left(\frac{n}{m}\right).$$

График плотности F-распределения зависит от числа степеней свободы, однако в целом он подобен распределению Пирсона.

***Определение* 3.15.** Будем называть распределение *симметричным*, если график его плотности симметричен относительно оси ординат.

Так, например, стандартное гауссовское и распределение Стьюдента – симметричные распределения, а Пирсона и Фишера – нет.

Огромную важность имеют следующие понятия.

***Определение* 3.16.** *Односторонней критической границей* уровня значимости α некоторого распределения $P(x)$ с.в. X (рис. 3.2), называется такое число z_α, что

$$P\{x \le z_\alpha\} = 1 - \alpha.$$

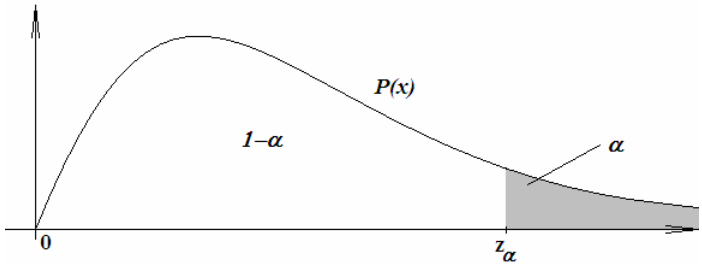


Рис 3.2.

***Определение* 3.17.** *Двухсторонней критической границей* уровня значимости α некоторого симметричного распределения $P(x)$ с.в. X (рис. 3.3) называется такое число $z_{\alpha/2}$, что

$$P\{-z_{\alpha/2} \le x \le z_{\alpha/2}\} = 1 - \alpha.$$

Отметим, что односторонняя критическая граница уровня значимости $\alpha/2$ некоторого симметричного распределения одновременно является его двухсторонней границей уровня значимости α.

Критические границы различных уровней значимости, стандартных статистических распределений приводятся в виде справочных таблиц, в соответствующей литературе (например, в учебниках и задачниках по математической статистике). Умение пользоваться этими таблицами является необходимым навыком для решения самых разных задач математической статистики.

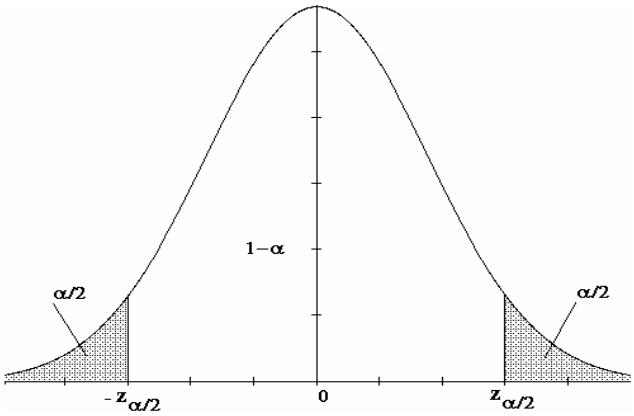


Рис 3.3.

3.7. Понятие доверительного интервала

Как уже говорилось, конкретное значение оценки некоторого параметра, рассчитанное по числовым данным конкретной выборки, называют *точечной оценкой* данного параметра.

Поскольку у нас нет информации о точности этих оценок, то на практике конкретные управленческие решения, выработанные на основе сведений о значениях точечных оценок, не являются достаточно обоснованными. То есть, пока мы не установили *меры близости* между теоретическим значением интересующего нас параметра и его точечной оценкой, невозможно говорить о допустимости практического использования последней.

Поэтому большую важность в теории статистического оценивания имеет следующее понятие.

***Определение* 3.18.** *Доверительным интервалом* уровня значимости α некоторого параметра γ называется любой интервал $[a,\ b] \in R^1$, такой, что

$$P\{\gamma \in [a,\ b]\} = 1 - \alpha\ .$$

Величину $p = 1 - \alpha$ при этом называют *доверительной вероятностью* или *уровнем доверия.*

Иначе говоря, доверительный интервал – это интервал, в котором с заданной (желаемой) вероятностью содержится теоретическое значение оцениваемого параметра. Обычно строят доверительные интервалы при $\alpha = 0{,}1$ или, чаще всего, при $\alpha = 0{,}05$, или, в особо ответственных случаях, при $\alpha = 0{,}01$.

Доверительные интервалы часто называют *интервальными оценками* данного параметра.

Ясно, что доверительный интервал уже несет информацию о мере точности наших знаний об истинном значении параметра. Именно эта информация и может действительно служить основанием для тех или иных конкретных практических выводов.

3.8. Доверительный интервал для математического ожидания нормальной с.в.

Рассмотрим задачу построения интервальной оценки для математического ожидания нормально распределенной с.в. $X \sim N\left(a,\ \sigma^2\right)$.

Во-первых, отметим, что для выборочного среднего

$$M\left(\overline{X}\right) = M\left(\frac{1}{n}\sum^{n} X_i\right) = \frac{n\,M(X)}{n} = a,$$

$$D\left(\overline{X}\right) = D\left(\frac{1}{n}\sum^{n} X_i\right) = \frac{1}{n^2}\sum^{n} D(X_i) = \frac{\sigma^2}{n}\ ,$$

т.е.

$$\overline{X} \sim N\left(a, \frac{\sigma^2}{n}\right).$$

Тогда, из свойств математического ожидания и дисперсии, получаем

$$\frac{\sqrt{n}(\overline{X} - a)}{\sigma} \sim N(0,1)\ \text{– стандартная гауссовская с.в.}$$

И если $u_{\alpha/2}$ – двухсторонняя критическая граница уровня значимости α стандартного гауссовского распределения, то с вероятностью $p = 1 - \alpha$ выполняется неравенство

$$-u_{\alpha/2} \le \frac{\sqrt{n}(\overline{X} - a)}{\sigma} \le u_{\alpha/2}\ .$$

Отсюда с той же вероятностью выполняется неравенство

$$\overline{X} - \frac{\sigma}{\sqrt{n}}\, u_{\alpha/2} \le a \le \overline{X} + \frac{\sigma}{\sqrt{n}}\, u_{\alpha/2},$$

которое и определяет искомый доверительный интервал для математического ожидания нормально распределенной с.в. при известной дисперсии σ^2.

Рассмотрим гораздо более важный для практики случай, когда дисперсия неизвестна.

Справедлива теорема.

***Теорема* 3.2.** Пусть с.в. $X \sim N\left(a, \sigma^2\right)$, и задана выборка ее значений $x_1, x_2, \ldots, x_n$, тогда

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1),$$

где S^2 – несмещенная оценка дисперсии с.в. X по этой выборке.

Поскольку

$$(n-1)\frac{S^2}{\sigma^2} = \frac{\sum\left(x_i - \overline{X}\right)^2}{\sigma^2},$$

то для с.в. q имеем

$$q = \frac{\sqrt{n}(\overline{X} - a)}{S} = \frac{\sqrt{n}(\overline{X} - a)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}} = \frac{\frac{\sqrt{n}}{\sigma}(\overline{X} - a)}{\sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{\sigma^2}}} = \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi^2(n-1)}},$$

т.е.

$$q \sim t(n-1).$$

Отсюда получаем, что с вероятностью $p = 1 - \alpha$ выполняется неравенство

$$-t(n-1) \le \frac{\sqrt{n}(\overline{X} - a)}{S} \le t(n-1),$$

где $t_{\alpha/2}(n-1)$ – табличное значение двухсторонней критической границы уровня значимости α распределения Стьюдента с (n-1) степенями свободы. Или, с той же вероятностью, неравенство

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le a \le \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

которое и определяет искомый доверительный интервал для математического ожидания нормально распределенной с.в. при неизвестной дисперсии.

Отметим, что тот же интервал можно выразить через формулу

$$\overline{X} - t_{\alpha/2}(n-1)\frac{\widetilde{S}}{\sqrt{n-1}} \le a \le \overline{X} + t_{\alpha/2}(n-1)\frac{\widetilde{S}}{\sqrt{n-1}}.$$

Все полученные доверительные интервалы строились в предположении, что X имеет нормальное распределение. В этом случае $\overline{X}$ имеет в точности нормальное распределение, что и является, в действительности, единственным необходимым условием корректности всех выкладок. Однако и при любом распределении X при $n \to \infty$, $\overline{X}$ будет асимптотически нормальной с.в., в соответствии с центральной предельной теоремой. Эта близость, разумеется, зависит от объема выборки n, но можно сказать, что уже при $n > 5 \div 10$ ее можно считать достаточной.

Пример. Производитель шин заинтересован в получении оценки средней износоустойчивости шин одной особой модели. Над 10 случайно выбранными шинами произвели специальные испытания, оказалось, что средняя длина пробега $\overline{X}$ = 22 500 миль, при стандартном отклонении $\widetilde{S}$ = 3000 миль. Найдем 95 и 99 %-ные доверительные интервалы для $M(x) = a$, X – длина пробега.

Решение. По таблицам находим $t_{0,025}(10 - 1) = 2{,}26$; $t_{0,005}(9) = 3{,}25$; тогда с вероятностью:

– $p = 0{,}95$ износоустойчивости будет лежать в пределах

$$22\,500 \pm 2{,}26\,\frac{3000}{\sqrt{9}} = 22\,500 \pm 2260 \text{ миль},$$

отклонение от $\overline{X}$ при этом составляет ± 10 %;

– $p = 0{,}99$ % в пределах 22 500 ± 3250, отклонение ± 14 %.

Практически считается, что чем меньше размах доверительного интервала, тем оценка достовернее. В разных случаях считается по-разному, но обычно исходят из того, что при размахе до 10 % от значения $\overline{X}$ полученная информация достаточно точно отражает реальную ситуацию. При размахе более 20 – 30 % достоверность прогноза мала. Очевидно, уменьшить размах интервала можно увеличив объем выборки.

3.9. Доверительный интервал для дисперсии нормального распределения

Пусть опять с.в. $X \sim N(a, \sigma^2)$ и имеется выборка ее значений

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

Рассмотрим вопрос построения доверительного интервала для ее дисперсии. Опять разберем два случая:

1. a – известно, тогда

$$n\frac{S_a^2}{\sigma^2} = \sum^{n}\left(\frac{x_i - a}{\sigma}\right)^2 \sim \chi^2(n)\,,$$

поскольку

$$\frac{X_i - a}{\sigma} \sim N(0,1)\,.$$

Тогда с вероятностью $p = 1-\alpha$ выполняется неравенство

$$\chi^2_{1-\alpha/2}(n) \le \frac{n\,S_a^2}{\sigma^2} \le \chi^2_{\alpha/2}(n)\,,$$

где $\chi^2_{\alpha/2}(n),\ \chi^2_{1-\alpha/2}(n)$ – соответствующие односторонние критические границы распределения Пирсона. Отсюда и получаем требуемый доверительный интервал уровня и значимости α, для истинного значения σ^2

$$\frac{nS_a^2}{\chi^2_{\alpha/2}(n)} \le \sigma^2 \le \frac{nS_a^2}{\chi^2_{1-\alpha/2}(n)}\,.$$

2. a – неизвестно, тогда, по теореме 3.2

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)\,.$$

И, аналогично, доверительный интервал с уровнем доверия $p = 1 - \alpha$ имеет вид

$$\frac{(n-1)\,S^2}{\chi^2_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2}(n-1)}\,.$$

3.10. Доверительный интервал для генеральной доли признака

При достаточно большом количестве наблюдений (считается, что при $n \ge 20$) на основании центральной предельной теоремы, можно считать, что выборочная доля

$$w = \frac{m}{n}$$

имеет распределение, достаточно близкое к нормальному, параметры которого были указаны в п. 3.5.

Тогда, например, для повторной выборки, с вероятностью $1-\alpha$, выполняются неравенства

$$-u_{\alpha/2} \le \frac{w-p}{\sqrt{p\,q/n}} \le u_{\alpha/2} \Rightarrow\ w - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le w + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\ .$$

Последнее из них дает границы искомого доверительного интервала в неявном виде. Для их явного выражения решаем соответствующее квадратное уравнение

$$(w-p)^2 = (u_{\alpha/2})^2\,\frac{p(1-p)}{n}\,.$$

Откуда получаем, что эти границы задаются выражениями

$$p_{1,2} = \frac{1}{1+\dfrac{u^2_{\alpha/2}}{n}}\left[w + \frac{u^2_{\alpha/2}}{2n} \pm u_{\alpha/2}\sqrt{\frac{w(1-w)}{n} + \left(\frac{u_{\alpha/2}}{2n}\right)^2}\right].$$

Для бесповторной выборки аналогично получаем следующие границы интервальной оценки

$$p_{1,2} = \frac{1}{1+\dfrac{u^2_{\alpha/2}}{n}\,\theta}\left[w + \frac{u^2_{\alpha/2}}{2n}\,\theta \pm u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}\,\theta + \left(\frac{u_{\alpha/2}}{2n}\,\theta\right)^2}\right],$$

где $\theta = \dfrac{N-n}{N-1}$.

Если объем данных выборки не достаточно велик ($n < 20$), то для построения границ p_1 и p_2 искомого интервала строят специальные уравнения на основе конкретных выражений для вероятностей значений биномиального или гипергеометрического распределений.

3.11. Определение необходимого объема выборки

Формулы доверительных интервалов рассмотренных параметров позволяют ответить на следующий весьма важный вопрос: выборку какого объема мы должны иметь, чтобы оценить интересующий нас параметр с заданной точностью?

Для оценки:

1. Математического ожидания. Во-первых, отметим, что размах интервала

$$\lim_{n\to\infty}\left(t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)\to 0,$$

поскольку

$$\lim_{n\to\infty} t_\alpha(n)\to u_\alpha, \quad \lim_{n\to\infty} S\to\sigma.$$

Итак, с ростом n доверительный интервал сжимается. Для определения достаточного объема выборки:

а) сначала находят S по некоторой пробной выборке;

б) для объема пробной выборки (или большего) принимают соответствующую критическую границу *t*-распределения;

в) далее корректируют объем.

Пример. Пусть было отобрано n = 25 пакетов некоторого стандартно расфасованного продукта, средний вес которых оказался $\overline{X}$ = 1020 г при стандартном отклонении $\tilde{S}$ = 12 г. Каким должен быть объем выборки, чтобы установить 99 %-ный интервал с размахом не более ±5 г.

Решение. Принимаем, например, $t_{0,005}\,(n-1) = t_{0,005}(40) = 2{,}797$.

Получаем

$$2{,}797\,\frac{12}{\sqrt{n}} \le 5 \Rightarrow \sqrt{n} \ge 6{,}71 \Rightarrow n \ge 45{,}06 \Rightarrow n = 46.$$

Далее можно скорректировать $t_{0,005}(45) = 2{,}41$, и т.д.

При известной генеральной дисперсии задача решается еще проще.

2. Дисперсии. Как известно

$$\lim_{n\to\infty}\chi^2_\alpha(n)\to n\left(1-\frac{2}{9n}+u_\alpha\sqrt{\frac{2}{9n}}\right),$$

т.е. и для дисперсии размах интервала с ростом n уменьшается. С помощью несложных численных процедур можно подобрать необходимое n. Опять следует использовать данные пробной выборки.

3. Генеральной доли. Задача решается непосредственным выбором достаточно большого n, если только этот объем выборки может быть практически реализован.

3.12. Оценка функции распределения

Определение 3.19. **Выборка значений с.в., упорядоченная по возрастанию, называется *вариационным рядом*.**

В качестве оценки $\hat{F}_X(t)$ функции распределения $F_X(t)$ с.в. X естественно предложить следующую кусочно-постоянную функцию

$$\hat{F}_X(t) = \begin{cases} 0, & t < x_1 \\ \dfrac{k-1}{n}, & x_{k-1} < t \le x_k, \quad k = \overline{1,n}, \\ 1, & t > x_n \end{cases}$$

где x_k – элементы вариационного ряда.

Функция $\hat{F}_X(t)$ называется *эмпирической функцией распределения*. Заметим, она соответствует всем необходимым свойствам функции распределения.

Необходимо определить точность этой оценки. В качестве меры уклонения $\hat{F}_X(t)$ от $F_X(t)$ обычно рассматривают величину

$$\hat{F}_X(t)\ D_n = \max_{1\le k\le n} d_k,$$

где

$$d_k = \max\left\{\frac{k}{n} - F_X(x_k),\ F_X(x_k) - \frac{k-1}{n}\right\}.$$

т.е. D_n – это наибольшее отклонение $\hat{F}_X(t)$ от $F_X(t)$ в точках вариационного ряда.

Справедлива следующая теорема.

Теорема 3.3. Если $F_X(t)$ непрерывна, то при $z>0$ и $n\to\infty$

$$P\left\{\sqrt{n}\,D_n<z\right\}\to K(z).$$

В этой теореме $K(z)$ – так называемое распределение Колмогорова, которое имеет свойство

$$\lim_{z\to\infty}K(z)=1.$$

Из теоремы 3.3 получаем, что с вероятностью $p=1-\alpha$ выполняется неравенство

$$\left|\hat{F}_X(t)-F_X(t)\right|\le\frac{z_\alpha}{\sqrt{n}},\ \ \forall t\in R^1,$$

где z_α – соответствующая критическая граница распределения Колмогорова. Последнее неравенство, очевидно, может выполнять роль доверительного интервала для $F_X(t)$. Например, при $\alpha=0{,}05$

$$\hat{F}_X(t)-\frac{1{,}358}{\sqrt{n}}\le F_X(t)\le\hat{F}_X(t)+\frac{1{,}358}{\sqrt{n}}.$$

Существуют и другие меры уклонения $\hat{F}_X(t)$ от $F_X(t)$, например, мера Мизеса.

3.13. Оценка функции плотности распределения

Плотность распределения также может быть оценена, причем несколькими способами:

1. *Гистограмма.* Находят $x_{\min}$ и $x_{\max}$ – минимальный и максимальный элементы выборки. Интервал $[x_{\min},x_{\max}]$, длина которого называется *размахом выборки*, делят на k подынтервалов одинаковой длины

$$\Delta=\frac{x_{\max}-x_{\min}}{k}.$$

Рекомендуется, согласно известной формуле Стерджеса, принимать

$$k=1+3{,}322\cdot\ln(n),$$

так как здесь имеется ввиду ближайшее целое число.

Далее, находят величины ν_i – количество элементов выборки, попадающих в *i*-й подынтервал

$$[x_{\min}+\Delta i,x_{\min}+\Delta(i+1)),\ \ i=\overline{0,k-1}.$$

Затем над каждым таким интервалом строят прямоугольник высотой

$$h_i=\frac{\nu_i}{n},$$

где n – объем выборки (рис. 3.4).

2. *Полигон частот* – это ломаная, соединяющая середины соседних ступеней гистограммы. Если с.в. X непрерывна, то полигон частот лучше оценивает плотность распределения.

Известна теорема, дающая оценку точности аппроксимации плотности посредством полигона частот.

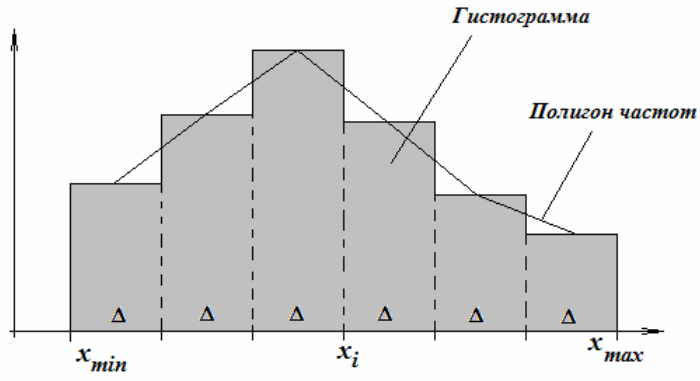


Рис. 3.4.

Теорема 3.4. (*Смирнова*). Для любого $(a,\ b)\in R$

$$\lim_{n\to\infty} P\left\{ \max_{a\le x\le b} \frac{\left|\hat{P}(x)-P(x)\right|}{\sqrt{P(x)}} < \frac{\lambda+\frac{z}{\lambda}}{\sqrt{n\Delta}} \right\} = e^{-2e^{-z}},$$

где λ – решение уравнения $\Phi(\lambda)=\frac{1}{2}-\frac{1}{2\sqrt[4]{n}}$, $P(x)$ – теоретическая плотность распределения.

3.14. Метод моментов

Рассмотренные выше точечные оценки параметров строились эвристически и лишь затем доказывалось, что они действительно являются оценками интересующих нас параметров. Оказывается имеются регулярные методы построения статистических оценок. Одним из них является метод моментов.

Определение 3.20. Пусть с.в. X имеет плотность распределения $P(x)$, тогда величины

$$m_k = \int_{-\infty}^{\infty} x^k P(x)dx, \quad d_k = \int_{-\infty}^{\infty} (x-M(X))^k P(x)dx$$

называются, соответственно, *теоретическими начальными* и *центральными моментами* k-го порядка.

Отметим, что математическое ожидание является начальным моментом первого порядка, а дисперсия – *центральным* второго.

Определение 3.21. Пусть имеется выборка $x_1, x_2, ..., x_n$ значений с.в. X. Тогда величины

$$\hat{m}_k = \frac{\sum_{i=1}^{n} x_i^k}{n}, \quad \hat{d}_k = \frac{\sum_{i=1}^{n} \left(x_i-\overline{X}\right)^k}{n}$$

называются, соответственно, *выборочными начальными* и *центральными моментами* k-го порядка.

Пусть $P(x,\gamma)$ – плотность распределения с.в. X, γ – неизвестный параметр этой плотности. Достаточно часто теоретические моменты могут быть выражены явно через неизвестный параметр γ. Идея метода моментов состоит в приравнивании теоретических и выборочных моментов. Решая соответствующие уравнения и получают искомую оценку неизвестного параметра. Если неизвестных параметров несколько, то приравнивая несколько моментов получают целую систему уравнений, которую также достаточно часто удается разрешить относительно оцениваемых параметров. Полученные таким образом статистические оценки называют *оценками метода моментов*.

Пример.

1. Пусть X равномерно распределена на отрезке $[c, d]\in R$. Ясно, что c, d являются параметрами данного распределения. Построим для них оценки, используя метод моментов.

Ранее мы видели, что для равномерного распределения

$$M(X)=\frac{c+d}{2}, \quad D(X)=\frac{(d-c)^2}{12}.$$

Приравнивая $M(X)$ и $D(X)$ с соответствующими выборочными моментами получаем систему уравнений

$$\begin{cases} \dfrac{c+d}{2} = \overline{X}, \\ \dfrac{(d-c)^2}{12} = \widetilde{S}^2, \end{cases}$$

решая которую находим

$$\begin{cases} c = \overline{X} - \sqrt{3}\,\widetilde{S}, \\ d = \overline{X} + \sqrt{3}\,\widetilde{S}. \end{cases}$$

Полученные равенства и являются оценками метода моментов искомых параметров.

2. Аналогично для нормального распределения получаем систему

$$\begin{cases} a = \overline{X}, \\ \sigma^2 = \widetilde{S}^2, \end{cases}$$

которая в целом соответствует полученным ранее оценкам. Однако видим, что оценка дисперсии получилась смещенной. Это происходит часто и при использовании других методов построения оценок и не считается серьезным недостатком, так как смещенные оценки обычно удается «исправить».

Достоинством метода моментов является его относительно простая вычислительная реализация. Важнейшим недостатком – неоднозначность получения оценок. Приравнивая различные моменты мы, вообще говоря, можем получить различные оценки для одних и тех же параметров. Кроме того, оценки метода моментов часто менее эффективны, чем некоторых других методов, например, следующего.

3.15. Метод максимального правдоподобия

Пусть $P(x,\gamma)$ – плотность распределения с.в. X, где γ – неизвестный, подлежащий оценке параметр. И пусть имеется выборка $x_1, x_2, ..., x_n$ значений этой с.в. Идея метода максимального правдоподобия состоит в том, что в качестве оценки γ следует принять такое значение, при котором вероятность появления именно этой имеющейся выборки максимальна. А эта вероятность, условно говоря, пропорциональна функции

$$L(\gamma) = P(x_1, \gamma)\,P(x_2, \gamma)\cdot ...\cdot P(x_n, \gamma)\,,$$

которую называют *функцией правдоподобия* Фишера. Таким образом мы приходим к задаче на нахождение экстремума функции $L(\gamma) \to \max$.

Необходимое условие экстремума дает уравнение

$$\frac{\partial L(\gamma)}{\partial \gamma} = 0\,,$$

которое, будучи разрешенным относительно γ, и определяет оценку метода максимального правдоподобия. Если неизвестных параметров несколько $\gamma_1, ..., \gamma_k$, то, аналогично, получаем систему необходимых условий экстремума

$$\begin{cases} \dfrac{\partial L(\gamma_1, ..., \gamma_k)}{\partial \gamma_1} = 0, \\ \vdots \\ \dfrac{\partial L(\gamma_1, ..., \gamma_k)}{\partial \gamma_k} = 0, \end{cases}$$

из которой выражаются необходимые оценки.

Часто оказывается удобнее использовать не $L(\gamma)$, а

$$\widetilde{L}(\gamma) = \ln(L(\gamma)) = \ln(P(x_1, \gamma)\,P(x_2, \gamma)\cdot ...\cdot P(x_n, \gamma)) = \sum_{i=1}^{n} \ln(P(x_i, \gamma))\,,$$

которая называется *логарифмической функцией правдоподобия*. В силу свойств логарифма, и $L(\gamma)$, и $\widetilde{L}(\gamma)$ достигают максимума при одном и том же значении.

Пример.

1. Найдем оценки максимального правдоподобия для параметров нормального распределения. Функция правдоподобия имеет вид

$$L(a, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum\limits_{i=1}^{n} \frac{(x_i - a)^2}{2\sigma^2}}\,,$$

а логарифмическая функция правдоподобия

$$\widetilde{L}(a, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\,.$$

Ясно, что удобнее максимизировать последнюю. Система необходимых условий экстремума

$$\begin{cases} \dfrac{\partial \widetilde{L}(a, \sigma^2)}{\partial a} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i - a) = 0, \\ \dfrac{\partial \widetilde{L}(a, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2}\dfrac{1}{\sigma^2} + \dfrac{1}{2(\sigma^2)^2}\sum\limits_{i=1}^{n}(x_i - a)^2 = 0. \end{cases}$$

Выражая из этой системы a и σ^2 получаем искомые оценки

$$\begin{cases} \hat{a} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i, \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(x_i - \overline{X}\right)^2 = 0, \end{cases}$$

т.е. получили хорошо знакомый результат. В данном случае, оценки максимального правдоподобия совпадают с оценками метода моментов, что на самом деле случается редко.

2. Найдем оценки максимального правдоподобия границ интервала распределения равномерно распределенной с.в. (см. пример п. 3.14).

Подлежащая максимизации функция правдоподобия в данном случае имеет вид

$$L(c,\ d) = \frac{1}{(d-c)^n} \to \max$$

и, очевидно, что должны выполняться ограничения

$$c \le x_{\min},\ \ x_{\max} \le d\ ,$$

где $x_{\min}, x_{\max}$ – соответственно, минимальный и максимальный элементы выборки. Решением задачи, а значит искомыми оценками являются значения

$$\hat{c} = x_{\min},\ \ \hat{d} = x_{\max}\ .$$

Как видим, они не совпадают с полученными ранее оценками метода моментов.

Вопросы для самопроверки

1. В чем заключается основная задача математической статистики?
2. Что такое выборочное среднее?
3. Какие вы знаете оценки дисперсии?
4. Что такое генеральная доля?
5. В чем недостаток точечных оценок?
6. Какие стандартные статистические распределения вы знаете?
7. Постройте оценку максимального правдоподобия для параметра экспоненциального распределения.
8. Постройте доверительный интервал для математического ожидания экспоненциального распределения.
9. Что такое распределение Колмогорова, и для чего оно может быть использовано?
10. Зачем нужны методы построения статистических оценок?

4. проверка статистических гипотез

4.1. Общая схема проверки статистических гипотез

***Определение* 4.1.** *Статистической гипотезой* называется любое высказывание о конкретных значениях параметров распределения некоторой с.в. или о виде этого распределения.

В первом случае гипотеза называется *параметрической*, во втором – *непараметрической* [1, 3, 4].

Нередко конкретные управленческие решения могут или должны быть приняты на основе анализа статистической информации о рассматриваемом процессе или явлении. Причем эти решения напрямую определяются тем, каковы конкретные параметры этого процесса. Например:

1. Станок производит фасовку некоторого продукта в стандартные упаковки весом по 1 кг. Средний вес 50-ти, случайным образом отобранных упаковок оказался равным $X = 0{,}98$ кг, при $S = 0{,}05$ кг. Можно ли считать, что имеющееся отклонение является результатом случайности и что станок настроен правильно? Или следует останавливать производство для переналадки станка?
2. В с/х районе опробуются два новых сорта пшеницы. Средняя урожайность первого составила 22 ц/га, второго – 22,5 ц/га. Следует ли считать, что второй сорт действительно урожайнее первого? Или имеющуюся разницу можно объяснить всегда присутствующими случайными причинами.
3. Имеется статистика страховых случаев по поводу угонов автомобилей различных марок. Следует ли считать, что частота случаев зависит от марки автомобилей? Очевидно, что если да, то страховая компания должна учитывать это в условиях страхования.

Оказывается, в этих и во многих других подобных случаях, ответ на поставленный вопрос можно свести к проверке соответствующей статистической гипотезы. Именно имеющиеся в математической статистике методики проверки таких гипотез делают ее одним из самых практически значимых разделов прикладной математики.

Общую схему проверки статистических гипотез кратко сформулировать непросто. Но можно выделить в ней следующие моменты:

1. Во-первых, формулируют, так называемую, *основную* или *нулевую* гипотезу. Ее обычно обозначают H_0. Различают простые и сложные гипотезы.

***Определение* 4.2.** Гипотеза называется *простой*, если она состоит в равенстве одного или нескольких параметров заданным числам. Если множество допустимых, для справедливости гипотезы, значений параметров состоит более, чем из одного элемента, то ее называют *сложной*.

Например

$H_0 : a = 5,$
$H_0 : a = 5, \sigma^2 = 10$ } – простые гипотезы,

$H_0 : 4 \le a \le 5$ – сложная гипотеза.

2. Одновременно рассматривается некоторая гипотеза называемая *альтернативной* или *конкурирующей*, ее обычно обозначают H_1. Различают односторонние и двухсторонние конкурирующие гипотезы. Например,

$H_1 : a > 5,$
$H_1 : a < 5$ } – односторонние гипотезы,

$H_1 : a \ne 5$ – двухсторонняя гипотеза.

Выбор вида альтернативной гипотезы определяется смыслом задачи.

3. Всегда имеется некоторая величина γ, которая рассчитывается из данных выборки, и поэтому ее называют *выборочной статистикой* или *критерием* проверки гипотезы. Причем из теории бывает известно, какое распределение $P(\gamma)$ будет иметь данная величина, если верна нулевая гипотеза.

4. Из данных конкретной выборки находится расчетное значение $\gamma_{\text{расч}}$, и если оно плохо соответствует теоретическому распределению $P(\gamma)$ (рис. 4.1), то отсюда делается вывод, что в действительности с.в. γ имеет другое распределение, а значит, и нулевая гипотеза H_0 не верна. Она отклоняется, и принимается альтернативная гипотеза H_0.

5. При этом всегда имеется вероятность сделать неправильный вывод. Ошибкой I-го рода называют отклонение, на самом деле истинной, нулевой гипотезы. Вероятность этой ошибки называют *уровнем значимости* проверки гипотезы, и обычно обозначают α. Ошибкой II-го рода называют принятие на самом деле ложной, нулевой гипотезы. Вероятность этой ошибки обычно обозначают β, а величину $1-\beta$ называют *мощностью критерия*. В общем случае α и β не связаны каким-либо однозначным соотношением, хотя для конкретных критериев это возможно. Считается, что хороший критерий должен обладать свойством $\alpha < \beta$.

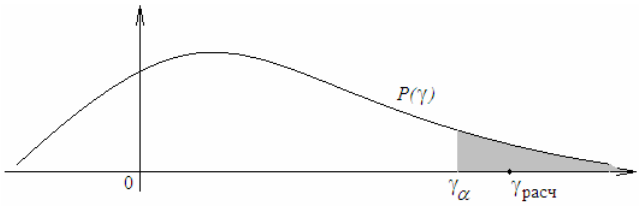


Рис. 4.1.

***Определение* 4.2.** Область Q, при попадании в которую выборочной статистики $\gamma_{\text{расч}}$ отвергается основная гипотеза, называется *критической областью*.

Говорят об уровне значимости критической области. В соответствии с видом альтернативной гипотезы различают *односторонние* и *двухсторонние* критические области.

Иногда говорят, что критерием проверки статистической гипотезы называется правило построения критической области. Отметим еще, что иногда критерием называют теоретическое распределение $P(\gamma)$.

4.2. Проверка простых гипотез с помощью доверительных интервалов

Наиболее просто общая схема проверки статистических гипотез может быть проиллюстрирована на примере проверки простых гипотез, что, фактически, делается с помощью доверительных интервалов для соответствующих параметров.

Например, рассмотрим гипотезу

$$H_0 : a = a_0,$$

где a – теоретическое математическое ожидание с.в. X, a_0 – некоторое заданное число.

Как известно с вероятностью $p = 1 - \alpha$ выполняется неравенство

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le a \le \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

где $t_{\alpha/2}(n-1)$ – табличное значение двухсторонней критической границы, уровня значимости α, распределения Стьюдента с $(n-1)$ степенями свободы. И если H_0 верна, то с той же вероятностью должно выполняться неравенство

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le a_0 \le \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

а, значит, и неравенство

$$a_0 - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le \overline{X} \le a_0 + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}.$$

Если выборочное среднее $\overline{X}$ не удовлетворяет этому неравенству, то произошло событие, вероятность которого меньше α (напомним, что это в предположении истинности H_0), т.е. событие достаточно маловероятное, чтобы можно было предположить, что на самом деле истинное значение a не равно a_0, т.е. H_0 – не верна, она отвергается. При этом имеется вероятность, не больше α, что все же H_0 верна. Однако мы сами выбираем уровень значимости в зависимости от ответственности принимаемого
решения.

Таким образом, в данном случае выборочной статистикой является величина

$$\gamma = \frac{\sqrt{n}(\overline{X} - a_0)}{S},$$

а ее теоретическим распределением $t(n-1)$. Односторонняя критическая область Q уровня значимости α, соответствующая односторонней альтернативе $H_1 : a > a_0$, определяется неравенством

$$Q : \frac{\sqrt{n}(\overline{X} - a_0)}{S} > t_\alpha(n-1),$$

а двухсторонняя критическая область уровня α, соответствующая двухсторонней альтернативе $H_1 : a \neq a_0$, – двумя неравенствами

$$Q : \frac{\sqrt{n}(\overline{X} - a_0)}{S} < -t_{\alpha/2}(n-1), \quad t_{\alpha/2}(n-1) < \frac{\sqrt{n}(\overline{X} - a_0)}{S}.$$

Фактически для проверки гипотезы $H_0 : a = a_0$ следует проверить выполнение неравенства

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \le a_0 \le \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}.$$

Рассмотрим гипотезу о равенстве дисперсии с.в. заданному числу

$$H_0 : \ \sigma^2 = \sigma_0^2,$$

где σ_0^2 – заданное число. И если она справедлива, то с вероятностью $p = 1 - \alpha$ будут выполняться неравенства

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \le \sigma_0^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \ \Rightarrow \ \frac{\sigma_0^2 \chi^2_{1-\alpha/2}(n-1)}{n-1} \le S^2 \le \frac{\sigma_0^2 \chi^2_{\alpha/2}(n-1)}{n-1}.$$

Если эти неравенства не выполняются, т.е. выполняется, например, одно из неравенств, определяющих двустороннюю критическую область с уровнем значимости α

$$Q : \ S^2 > \frac{\sigma_0^2 \chi^2_{\alpha/2}(n-1)}{n-1}, \quad \text{или} \quad S^2 < \frac{\sigma_0^2 \chi^2_{1-\alpha/2}(n-1)}{n-1},$$

то значит произошло очень маловероятное событие, и что, по-видимому, просто не верна нулевая гипотеза и ее нужно отвергнуть.

Пример. Компания выпускает электролампочки, установленный нормативный срок использования которых составляет $a_0 = 1500$ ч. Из новой партии было выбрано $n = 10$ ламп, оказалось

$$\overline{X} = 1410 \text{ ч},$$

при

$$\tilde{S} = \sqrt{\frac{1}{n}\sum^{n}(x_i - \overline{X})^2} = 90 \text{ ч}.$$

Свидетельствует ли это о том, что срок использования отклонился от нормативного?

Решение. Рассматриваем гипотезу $H_0 : a = 1500$, при альтернативе $H_1 : a < 1500$.

Строим одностороннюю критическую область при $\alpha = 0{,}95$

$$Q : \overline{X} < a_0 - t_\alpha(n-1)\frac{\tilde{S}}{\sqrt{n-1}} = 1500 - t_{0,95}(10-1)\frac{\tilde{S}}{\sqrt{10-1}} = 1445{,}1.$$

Таким образом, нулевая гипотеза отклоняется, срок действительно отклонился от нормативного.

4.3. Проверка гипотезы о равенстве дисперсий двух выборок

Пусть $X \sim N(a_X, \sigma_X^2)$, $Y \sim N(a_Y, \sigma_Y^2)$ – независимые с.в., и имеются выборки их значений $X: x_1, \ldots, x_n$; $Y: y_1, \ldots, y_m$.

Рассмотрим гипотезу $H_0: \sigma_X^2 = \sigma_Y^2$.

Следует отметить, что это сложная гипотеза, однако, ее легко можно привести к эквивалентной простой

$$H_0: z = 1,$$

где $z = \dfrac{\sigma_X^2}{\sigma_Y^2}$ – можно считать некоторым параметром распределения этих с.в.

Рассмотрим опять два случая:

1. a_X, a_Y – известны. Тогда наилучшими оценками дисперсий являются

$$S_X^2 = \frac{1}{n}\sum (x_i - a_X)^2 \quad \text{и} \quad S_Y^2 = \frac{1}{m}\sum (y_i - a_Y)^2 .$$

И если верна нулевая гипотеза, т.е. если $\sigma_X^2 = \sigma_Y^2 = \sigma^2$,то выполняется

$$F = \frac{S_X^2}{S_Y^2} = \frac{\frac{1}{n} n \frac{S_X^2}{\sigma^2}}{\frac{1}{m} m \frac{S_Y^2}{\sigma^2}} = \frac{\frac{1}{n}\chi^2(n)}{\frac{1}{m}\chi^2(m)} \sim F\left(\frac{n}{m}\right).$$

Таким образом с.в. F имеет распределение Фишера с указанным числом степеней свободы. Именно F является критерием при проверке этой гипотезы. Если при расчетах всегда брать

$$F = \frac{\text{большая дисперсия}}{\text{меньшая дисперсия}},$$

то можно ограничиться рассмотрением лишь односторонней альтернативной гипотезы $H_1: \sigma_X^2 > \sigma_Y^2$, для которой критическая область уровня значимости α определяется неравенством

$$Q: F_{\text{расч}} > F_\alpha\left(\frac{n}{m}\right).$$

2. a_X, a_Y – неизвестны, рассуждения совершенно аналогичны. Наилучшими оценками дисперсий являются

$$S_X^2 = \frac{1}{n-1}\sum (x_i - \overline{X})^2 \quad \text{и} \quad S_Y^2 = \frac{1}{m-1}\sum (y_i - \overline{Y})^2 .$$

И если верна нулевая гипотеза, в силу теоремы 3.2, выполняется

$$F = \frac{S_X^2}{S_Y^2} = \frac{\frac{1}{(n-1)}(n-1)\frac{S_X^2}{\sigma^2}}{\frac{1}{(m-1)}(m-1)\frac{S_Y^2}{\sigma^2}} = \frac{\frac{1}{n-1}\chi^2(n-1)}{\frac{1}{m-1}\chi^2(m-1)} \sim F\left(\frac{n-1}{m-1}\right).$$

При расчетах берут

$$F = \frac{\text{большая дисперсия}}{\text{меньшая дисперсия}},$$

тогда односторонняя критическая область уровня значимости α определяется неравенством

$$Q: F_{\text{расч}} > F_\alpha\left(\frac{n-1}{m-1}\right).$$

Если выборки не из нормальных совокупностей, то при достаточно больших m, n ($m, n \geq 20 \div 30$) рассматриваемый критерий также считается применимым.

Пример. В университете проведен анализ успеваемости среди студентов и студенток за последние 25 лет. С.в. X и Y – соответственно их суммарный балл за время учебы в эти годы. Получены следующие результаты:

$$\overline{X} = 400; \quad S_X^2 = 300; \quad \overline{Y} = 420; \quad S_Y^2 = 150.$$

Определить, есть ли основания считать, что разброс оценок у студентов больше, чем у студенток. Принять уровень значимости проверки гипотезы $\alpha = 0{,}05$.

Решение. Находим расчетное значение критерия

$$F_{\text{расч}} = \frac{300}{150} = 2 .$$

Критическая точка распределения Фишера $F_{\text{крит}} = F_{0,05}(24,24) = 1{,}98$. Поскольку $F_{\text{расч}} > F_{\text{крит}}$, нулевую гипотезу следует отклонить, т.е. действительно, разброс оценок у студентов в данном университете больше.

4.4. Проверка гипотезы о равенстве средних

Пусть $X \sim N(a_X, \sigma_X^2)$, $Y \sim N(a_Y, \sigma_Y^2)$ – независимые с.в., и имеются выборки их значений X: $x_1, \ldots, x_n$; Y: $y_1, \ldots, y_m$. Рассмотрим гипотезу $H_0 : a_X = a_Y$.

Формально это сложная гипотеза, так как ей соответствует целое множество точек

$$a_X = a_Y = a, \forall\, a \in R^1,$$

однако, ее легко можно привести к эквивалентной простой

$$H_0 : \; z = 0,$$

где $z = a_X - a_Y$.

Рассмотрим два случая:

1. σ_X^2, σ_Y^2 – известны. Тогда, как мы знаем, имеет место

$$\overline{X} \sim N\left(a_X, \frac{\sigma_X^2}{n}\right), \qquad \overline{Y} \sim N\left(a_Y, \frac{\sigma_Y^2}{m}\right),$$

и, следовательно,

$$\overline{X} - \overline{Y} \sim N\left(a_X - a_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

Если справедлива нулевая гипотеза, то

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}} \sim N(0,1).$$

Именно величина U, таким образом, является критерием при проверке этой гипотезы, а его теоретическим распределением – стандартное гауссовское распределение. Тогда двухсторонняя критическая область с уровнем значимости α имеет вид

$$Q : U < -u_{\alpha/2}, \quad U > u_{\alpha/2}.$$

На практике рассматриваемый критерий считается применимым не только тогда, когда теоретические дисперсии действительно известны, но и если объем выборок достаточно велик (порядка 50 и более наблюдений), тогда в качестве теоретических значений просто принимают их точечные оценки, так как они будут уже достаточно точны;

2. σ_X^2, σ_Y^2 – неизвестны, но равны. Если дисперсии действительно равны, то ясно, что наилучшая оценка соответствующей величины σ^2 должна рассчитываться исходя из совокупной выборки по формуле

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}.$$

Из теоремы 3.2 следует, что

$$(n+m-2)\frac{S^2}{\sigma^2} \sim \chi^2(n+m-2).$$

А тогда

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\left((n-1)S_x^2 + (m-1)S_y^2\right)/(m+n-2)}} \sqrt{\frac{nm}{n+m}} =$$

$$= \frac{\dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m}}}}{\sqrt{\dfrac{1}{(m+n-2)}(m+n-2)\dfrac{S^2}{\sigma^2}}} = \frac{N(0,1)}{\sqrt{\dfrac{1}{(m+n-2)}\chi^2(m+n-2)}} \sim t(n+m-2).$$

Расчетной статистикой проверки гипотезы в данном случае является t, а ее теоретическим распределением – распределение Стьюдента с $(n+m-2)$ степенями свободы.

Отсюда получаем, например, одностороннюю критическую область с уровнем значимости α

$$Q : t > t_\alpha(m+n-2),$$

где $t_\alpha(m+n-2)$ – табличное значение соответствующей односторонней критической границы рапределения Стьюдента. Аналогично предыдущему пункту все формулы для выборок не из нормальных генеральных совокупностей считаются применимыми при $m, n \geq 20 \div 30$.

Остается обсудить предположение о равенстве дисперсий. На практике вначале проверяют гипотезу о равенстве дисперсий X и Y, и если нет оснований отклонить ее, то рассматриваемый критерий сравнения средних считают применимым. Отметим, что две выборочные дисперсии должны отличаться довольно сильно (в 3 – 5 раз), чтобы были основания отклонить возможность равенства их генеральных значений, поэтому рассматриваемый критерий оказывается применимым весьма часто.

Пример. Исходя из данных предыдущего примера можно ли утверждать, что девушки в среднем учатся лучше ребят? Принять уровень значимости проверки гипотезы $\alpha = 0{,}05$.

Решение. Рассматриваем нулевую гипотезу $H_0 : a_X = a_Y$. Против альтернативной односторонней гипотезы $H_1 : a_X < a_Y$.

Находим, что

$$t = \frac{420-400}{\sqrt{24\cdot 300 + 24\cdot 150}}\sqrt{\frac{25\cdot 25\cdot(25+25-2)}{25+25}} = 4{,}71 .$$

По таблицам находим критическое значение $t_{\text{кр}} = t_{0,05}(25+25-2) = 1{,}68$. Рассчитанное значение больше его, следовательно действительно результаты девушек значимо превосходят успехи ребят.

4.5. Однофакторный дисперсионный анализ

Важное место в методах статистического анализа реальных данных и принятия соответствующих практических решений занимают приемы дисперсионного анализа. Суть этого анализа сводится к расчленению общей дисперсии некоторого показателя на компоненты, обусловленные влиянием отдельных факторов, и проверке гипотез о значимости этого влияния. Существует многофакторный дисперсионный анализ, но мы пока разберем простейший случай.

Пусть имеется выборка объема n значений некоторой интересующей нас величины X (например, некоторого экономического показателя). Задача состоит в выявлении ее зависимости от некоторого фактора Z. При этом фактор Z обычно характеризуется не численно, а некоторыми своими уровнями (категориями признака). При числовом выражении уровня Z для решения задачи более пригодным может оказаться коэффициент корреляции.

Разобьем имеющуюся статистику на, так называемые, *частичные выборки* X_{i1}, X_{i2}, ..., X_{ik_m} $(i = \overline{1,m})$, каждая из которых соответствует своему уровню фактора Z. Здесь k_i – количество элементов выборки, которые соответствуют объектам.

Если значение фактора непрерывно, тогда разбиение производится в смысле соответствия реализаций некоторому интервалу из общего диапазона значений фактора.

Эти выборки и соответствующие им промежуточные величины записывают в виде таблицы, которую называют *таблицей дисперсионного анализа* (табл. 4.1.).

Общую сумму квадратов отклонений от $\overline{X}$, т.е. общую вариацию данных можно разбить на два слагаемых:

$$Q_\Sigma = \sum_i^m \sum_j^{k_i} (X_{ij} - \overline{X})^2 = \sum_i \sum_j (X_{ij} - \overline{X}_j)^2 + \sum_{i=1}^m k_i\,(\overline{X}_i - \overline{X})^2 = Q_1 + Q_2 ,$$

где Q_1 – называется *внутригрупповой вариацией*, Q_2 – *межгрупповой вариацией*. Ясно, что чем сильнее влияние фактора Z на величину X, тем относительно большей будет составляющая Q_2.

Таблица 4.1.

Номер выборки	Наблюденное значение	Объем выборки	Групповая средняя
1	$X_{11}, X_{12}, ..., X_{1k_1}$	k_1	$\overline{X_1} = \dfrac{\sum_{j=1}^{k_1} X_{1j}}{k_1}$
::::	::::	::::	
::::	::::	::::	::::::::::
::::	::::	::::	
::::	::::	::::	
m	$X_{m1}, X_{m2}, ..., X_{mk_m}$	k_m	$\overline{X_m} = \dfrac{\sum_{j=1}^{k_m} X_{mj}}{k_m}$
		$n = \sum_j^m k_j$	$\overline{X} = \dfrac{\sum_{i=1}^m \sum_{j=1}^{k_i} X_{ij}}{n}$

Величины

$$S_1^2 = \frac{Q_1}{n-m}, \quad S_2^2 = \frac{Q_2}{m-1}$$

являются несмещенными оценками, дисперсий, соответственно:

- σ_1^2 – внутригрупповой (остаточная дисперсия);
- σ_2^2 – межгрупповой (регулярная дисперсия).

Схема дальнейших рассуждений такова. Если межгрупповая дисперсия значимо превосходит внутригрупповую, т.е. разброс данных, порождаемый влиянием фактора Z, значимо выше разброса, порождаемого другими случайными причинами, то и влияние фактора Z на величину X следует признать значимым.

Для сравнений же самих дисперсий σ_1^2 и σ_2^2 можно использовать обычный критерий Фишера. Рассмотрим пример.

Пример. На предприятии опробывались 4 технологии производства деталей. Выработка шт./день указана в таблице (см. табл. 4.2.). Можно ли утверждать, что фактор технологии существенно влияет на объем производства?

Решение. Находим средние значения объема выпуска по отдельным технологиям, и записываем их в той же таблице. Объем всей выборки

$$n_1 + n_2 + n_3 + n_4 = n = 23.$$

Находим общее среднее

$$\overline{X} = 1782/23 = 77{,}48\,.$$

Таблица 4.2.

Объем	Технология			
	1	2	3	4
1	60	75	60	95
2	80	66	80	85
3	75	85	65	100
4	80	80	60	80
5	85	70	86	–
6	70	80	75	–
7	–	90	–	–
$\overline{X}_i$, $i = 1, ..., 4$	450/6 = 75	78	71	90

Далее рассчитаем следующие величины

$$S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m} n_i(\overline{X}_i - \overline{X})^2 =$$

$$= \frac{1}{4-1}\left(6(75-77{,}48)^2 + 7(78-77{,}48)^2 + ... + 4(90-77{,}48)^2\right) =$$

$$= \frac{917{,}7}{3} = 305{,}9 \text{ – оценка межгрупповой дисперсии;}$$

$S_1^2 = \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 = \frac{1}{23-4}\left((60-75)^2 + ... + (80-90)^2\right) = \frac{1497}{19} = 78{,}8$ –оценка внутригрупповой дисперсии. Теперь находим расчетное значение критерия сравнения дисперсий

$$F_{\text{расч}} = \frac{S_2^2}{S_1^2} = \frac{305{,}9}{78{,}8} = 3{,}88$$

и сравниваем ее с табличным значением критической точки

$$F_{0{,}05}\left(\frac{3}{19}\right) = 3{,}13\,.$$

Видим, что расчетное значение превышает критическое. Теперь есть все основания предпочесть четвертую технологию.

4.6. Проверка непараметрических гипотез. Критерий согласия Пирсона

Пусть имеется некоторая с.в. X и выборка ее значений $x_1, \ldots, x_n$.

Требуется выяснить, может ли быть, что с.в. X имеет заданное распределение $P(x)$. Иначе говоря, что генеральная совокупность имеет это распределение. Такую гипотезу называют *непараметрической*, поскольку в ней идет речь не о параметрах, а о самом виде распределения.

Проверку строят следующим образом. Разбивают предполагаемый диапазон возможных значений изучаемой с.в. X на m подынтервалов. Это можно делать различным образом. Но, обычно, рекомендуют

- диапазон принимать равным $\Delta = x_{\max} - x_{\min}$,
- **количество интервалов разбиения, в соответствии с формулой Стерджеса,** $m = 1 + 3{,}322 \ln(n)$**.**

Далее находят, какое количество m_i элементов выборки попадают в i-й подынтервал, и рассчитывают наблюдаемую относительную частоту

$$\nu_i = \frac{m_i}{n}.$$

Исходя из $P(x)$ рассчитывают теоретические вероятности p_i попадания значений с.в X в соответствующие подынтервалы.

Рассматриваемую гипотезу формально записывают так

$$H_0 : \nu_1 = p_1,\ \nu_2 = p_2, \ldots,\ \nu_m = p_m.$$

Таким образом, следует сравнить наблюдаемые относительные частоты и теоретические вероятности. Ясно, что возможно использование различных мер их близости. Эти меры называют *критериями согласия* (иногда этот термин понимают более широко [4]). Известны и находят практическое применение критерии согласия Колмогорова, Романовского, Ястремского, ω-критерий и др. Но чаще всего, используется *критерий согласия Пирсона*

$$T_n = \sum_{i=1}^{m} \frac{(m_i - n \cdot p_i)^2}{n \cdot p_i},$$

который в случае истинности нулевой гипотезы и при $n > 20$ имеет распределение, приближающееся к

$$T \sim \chi^2(m - k - 1),$$

где k – число параметров теоретического распределения $P(x)$, уже оцененных по данным имеющейся выборки.

При использовании этого и других критериев согласия должны иметься статистики достаточно большого объема. Считается что должно быть $m_i \geq 5 \div 10$, для $\forall\ i = \overline{1, m}$. В частности поэтому не рекомендуют сильно дробить диапазон возможных значений с.в.

Критерий Пирсона используется для решения различных задач, в частности, такой важной, как проверка, что имеющаяся выборка является выборкой значений нормально распределенной с.в. (взята из нормальной генеральной совокупности). Отметим, что для этой задачи проще использовать другой подход, состоящий в расчете следующих эмпирических коэффициентов

$$A_c = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^3}{\left(\sqrt{\sum_{i=1}^{n} (x_i - \overline{X})^2}\right)^3} - 3, \qquad Э_k = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^4}{\left(\sum_{i=1}^{n} (x_i - \overline{X})^2\right)^2},$$

называемых коэффициентами асимметрии и эксцесса, соответственно. Для нормальной выборки их значения близки к нулю. Имеются таблицы [21] критических значений этих коэффициентов для проверки соответствующей гипотезы при различных уровнях значимости.

Критерий Пирсона используется также для такой важной и часто встречающейся на практике задачи как установление связи между признаками некоторых объектов, когда эти признаки не выражаются численно (если они выражаются численно, как бывает обычно, то можно использовать коэффициент корреляции). Рассмотрим суть этой задачи и методику расчетов на примере.

Пример. На некотором предприятии решили выяснить, имеется ли зависимость между процентом выполнения плана рабочими и тем, учатся ли они в настоящее время на курсах повышения квалификации или на заочном отделении. Собранные статистические данные были сведены в табл. 5.3.

Таблица 5.3.

Отношение к учебе	Выполнение нормы выработки, %				Всего
	85 – 99	100 – 105	106 – 116	117 – 130	
	1	2	3	4	
Учатся	2	5	10	20	37

Не учатся	10	5	2	2	19
Всего	12	10	12	22	56

Таблица 5.4.

Отношение к учебе	Выполнение нормы выработки, %			
	85 – 99	100 – 105	106 – 116	117 – 130
	1	2	3	4
Учатся	[37 × 12/56 = 7,93]	[6,6]	[7,93]	[14,54]
Не учатся	[4,07]	[3,39]	[4,07]	[7,46]

Ответить на вопрос: имеется ли зависимость между данными признаками?

Решение. Сравним имеющееся распределение рабочих с теоретическим распределением, соответствующим ситуации, когда никакой связи между признаками нет. Но сначала это распределение еще нужно составить. Им не будет являться то, которое получится, если в каждую клетку этой таблицы записать равные доли от общего количества рабочих, т.е. число 56/8 = 7. Действительно, нужно учитывать, сколько всего рабочих соответствует тому или иному признаку по отдельности без связи между ними. Этому условию удовлетворяют величины, посчитанные по формуле

$$w_{ij} = k_i q_j / N ,$$

где k_i – количество рабочих попавших в i -ю категорию первого признака, q_j – количество рабочих попавших в j -ю категорию второго признака, N – общий объем выборки. Запишем эти величины, для каждой из клеток, в квадратных скобках в табл. 5.4.

Отметим, что входящие в формулу критерия Пирсона величины $n \cdot p_j$, и есть эти теоретические частоты. Теперь находим расчетную величину

$$T_n = \frac{(7{,}93-2)^2}{7{,}93} + \frac{(4{,}07-10)^2}{4{,}07} + \frac{(6{,}6-5)^2}{6{,}6} + \ldots + \frac{(7{,}46-2)^2}{7{,}46} = 21{,}86 .$$

Находим критическую границу при $\alpha = 0{,}05$ и числе степеней свободы $(k-1)(q-1) = 3 \cdot 1 = 3$, где k, q – количества категорий соответствующих признаков

$$\chi^2_{0{,}05}(3) = 7{,}815 .$$

Видим, что расчетная величина весьма сильно превосходит критическую, поэтому нулевая гипотеза об отсутствии связи между признаками решительно отклоняется.

В заключении отметим, что в задачах аналогичных только что описанной нередко ошибочно применяют дисперсионный анализ. Это недопустимо, поскольку здесь данные не являются независимыми, а заранее сгруппированы.

Вопросы для самопроверки

1. Какие вы можете привести примеры статистических гипотез? Какие из них простые гипотезы?
2. Что такое статистический критерий и его теоретическое распределение?
3. Что такое мощность критерия?
4. Что такое критическая область?
5. В каком случае нельзя проверить гипотезу о равенстве математических ожиданий?
6. Постройте критерий сравнения дисперсий двух выборок, когда математическое ожидание одной из с.в. известно, а другой нет.
7. Сформулируйте своими словами основную идею однофакторного дисперсионного анализа.
8. Перечислите гипотезы, в которых теоретическим распределением является распределение Пирсона.
9. Можно ли задачу дисперсионного анализа решить на основе критерия Пирсона?

5. ОСНОВЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА

5.1. Детерминированные, статистические и регрессионные зависимости

Как известно, основной задачей прикладной математики является построение так называемых *математических моделей* различных процессов и явлений. Целью изучения любого процесса или явления является его *прогноз* и *оптимизация*. Математические модели часто позволяют избежать труднореализуемых, а иногда и практически невозможных реальных экспериментов над процессом, и ограничиться численными расчетами.

Чаще всего, математические модели представляют собой уравнения, описывающие взаимозависимости между параметрами процесса. Иногда построение этих моделей возможно на основе количественно известных законов природы, таких, как законы механики, законы сохранения энергии, массы, импульса и т.д., химические законы. Такой подход нередко оказывается возможным при моделировании технических и технологических процессов. В таких случаях модель может иметь вид

$$y = f(x_1, ..., x_m) ,$$

где y – какой-то интересующий нас показатель качества протекания процесса, или, как говорят, *объясняемый*, *эндогенный* его параметр, $x_1, ..., x_m$ – независимые, *экзогенные*, объясняющие управляемые параметры. Такая модель представляет собой строго *детерминированную*, *функциональную* зависимость.

В таких отраслях человеческих знаний как экономика, социология, психология, медицина и т.д., не существует количественно известных законов природы, а можно говорить лишь о качественно известных закономерностях. Например, уровень производительности труда на предприятии в среднем тем выше, чем больше его электровооруженность. Однако нет никаких оснований утверждать об однозначности такой зависимости, т.е. между параметрами явления существуют зависимости, но не жесткие, неоднозначные. Тогда математическая модель будет иметь вид

$$y = f(x_1, \ ..., \ x_m, \ \varepsilon) ,$$

здесь ε – некоторая, быть может многомерная, с.в. Такие зависимости называют *статистическими*, или *стохастическими*. При этом каждому набору значений объясняющих факторов $x_1, \ ..., \ x_m$, соответствует целая плотность распределения $P(y / x_1, \ ..., \ x_m)$ возможных значений зависимого показателя y. Ее называют *условной* плотностью распределения. Отсутствие возможности установить точную величину показателя y объясняется прежде всего тем, что y заведомо испытывает влияние не только факторов $x_1, ..., x_m$, но и других факторов, которые либо не известны, либо не учитываются в целях упрощения модели, и поэтому должны рассматриваться нами как случайные.

Построение статистических зависимостей возможно лишь на основе обработки статистического материала наблюдений за поведением процесса при различных значениях его управляемых параметров. На самом деле такой подход нередко используется и при моделировании сложных технологических процессов.

Зависимость вида

$$M(y / x_1, ..., x_m) = f(x_1, ..., x_m) ,$$

где $M(y / x_1, ..., x_m)$ – условное математическое ожидание параметра y, называется *корреляционной* зависимостью. Если, кроме того, факторы $x_1, ..., x_m$ – неслучайны и независимы между собой, то ту же зависимость называют *регрессионной*. А функцию $f(x_1, ..., x_m)$ – *функцией регрессии*. Следует отметить, что на практике далеко не всегда можно провести четкое разграничение между этими двумя случаями, особенно при работе с экономическими данными.

Говорят, что основной задачей так называемого *корреляционного анализа* является выявление связи между случайными переменными и оценка ее тесноты. Основной задачей *регрессионного анализа* является установление формы и изучение зависимости между переменными, т.е., в частности, оценка неизвестной функции регрессии.

Применение методов корреляционного и регрессионного анализа для анализа экономических данных положило начало новой науке – *эконометрике*.

5.2. Регрессионная модель и предпосылки регрессионного анализа

Центральная задача регрессионного анализа, как уже говорилось, состоит в конструировании неизвестной функции регрессии $f(x_1, ..., x_m)$ по исходным статистическим данным. Из-за ограниченности этих данных как и всегда возможно построение лишь функции, приближающейся к истинной функции регрессии, т.е. ее оценки $\hat{f}(x_1, ..., x_m)$, или, как говорят, *аппроксимации*. Соответствующую зависимость

$$\hat{y} = \hat{f}(x_1, ..., x_m)$$

будем называть *регрессионным уравнением*. Здесь $\hat{y}$ – точечная оценка значения фактора y при заданных $x_1, ..., x_m$, которую называют еще *расчетным* или *прогнозным* значением фактора y. Для выражения наличия случайности необходимо еще ввести случайную составляющую, что почти всегда делают следующим образом

$$y = \hat{f}(x_1, ..., x_m) + \varepsilon .$$

Если в последнем соотношении определен общий вид функции $\hat{f}(x_1, ..., x_m)$, то его называют *регрессионной моделью*. Выбор общего вида регрессионной модели является очень важным и очень непростым этапом регрессионного анализа.

Классическая теория многомерного регрессионного анализа строится для, так называемой, *линейной модели множественной регрессии*

$$y = a_1 x_1 + ... + a_m x_m + \varepsilon ,$$

где $a_i, \ i = \overline{1, m}$ – требующие своей оценки коэффициенты регрессии. Заметим, что важнейшим ограничением при этом является линейность правой части модели по коэффициентам, а не по экзогенным переменным, так как последние могут являться просто обозначениями сколь угодно сложных выражений от исходных объясняющих факторов, поэтому линейная модель достаточно универсальна.

Пусть имеется многомерная выборка объема n

$$\begin{array}{c|cccc} y_1 & x_{11} & x_{12} & \cdots & x_{1m}, \\ y_2 & x_{21} & x_{22} & \cdots & x_{2m}, \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nm}, \end{array}$$

где x_{ij} – наблюдаемое значение *j*-го объясняющего фактора в *i*-м наблюдении. Если принята линейная модель множественной регрессии, то это означает, что считается выполненной следующая система соотношений

$$y_i = a_1 x_{i1} + ... + a_m x_{im} + \varepsilon_i, \ i = \overline{1, n},$$

где $\varepsilon_i, \ i = \overline{1, n}$ – случайные возмущения в отдельных наблюдениях. Эту систему соотношений, в свою очередь, называют *классической нормальной линейной моделью множественной регрессии*, если выполняются условия:

1) ε_i – являются реализациями независимых с.в. с распределением $N(0, \sigma^2)$;

2) $x_1, ..., x_m$ – неслучайные величины;

3) столбцы матрицы

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

линейнонезависимы. Это означает, что объясняющие факторы взаимно независимы, т.е. не выражаются друг через друга, и что матрица $(X^T X)$ невырождена.

Матрицу *X* будем называть *матрицей значений объясняющих факторов*.

Предположения (1) – (3) называют *основными предпосылками регрессионного анализа*, или условиями Гаусса–Маркова.

Если ввести обозначения

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

то нормальную модель можно записать в гораздо более компактной матричной форме

$$Y = X a + \varepsilon.$$

5.3. Метод наименьших квадратов (МНК)

Метод наименьших квадратов – это метод нахождения оценок коэффициентов регрессии.

Уравнение регрессии, соответствующее многомерной линейной модели, имеет вид

$$\hat{y} = \alpha_1 x_1 + ... + \alpha_m x_m,$$

где α_i – оценки неизвестных коэффициентов линейной регрессии $a_i, \ i = \overline{1, m}$. Вектор

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = Y - \hat{Y}$$

называют *вектором ошибок регрессии* или вектором *остатков регрессии*. Идея МНК состоит в том, что в качестве искомых оценок следует принять такие значения, при которых вектор ошибок оказывается наименьшим. При этом используется следующая мера величины этого вектора

$$Q(\alpha_1, ..., \alpha_m) = e^T e = \sum_{i=1}^{n} e_i^2,$$

которая соответствует обычной формуле длины вектора в n-мерном евклидовом пространстве. Ясно, что величину $Q(\alpha_1, ..., \alpha_m)$, как это и указано, следует рассматривать как функцию искомых оценок. Эту функцию называют *квадратической функцией невязки*. В принципе возможно использование и других мер величины вектора ошибок, например, сумму модулей элементов этого вектора и т.п. Однако, квадратическая функция невязки имеет те преимущества, что она дифференцируема, и система необходимых условий минимума этой функции является линейной, а значит легко решается.

Имеем,

$$Q(\alpha_1, ..., \alpha_m) = e^T e = \left(Y - \hat{Y}\right)^T \left(Y - \hat{Y}\right) = (Y - X\alpha)^T (Y - X\alpha) =$$
$$= YY - \alpha^T X^T Y - Y^T X\alpha + \alpha^T X^T X\alpha = YY - 2\alpha^T X^T Y + \alpha^T X^T X\alpha,$$

где α – вектор искомых оценок. Дифференцируя векторно, получаем систему необходимых условий минимума

$$\frac{\partial Q}{\partial \alpha} = -2X^T Y + 2X^T X\alpha = 0\,.$$

Или после простых преобразований – следующую систему уравнений

$$\left(X^T X\right)\alpha = X^{T\cdot} Y\,,$$

называемую *системой нормальных уравнений*. В алгебраической форме она имеет вид

$$\begin{cases} \alpha_1 \sum x_{i1}^2 + \alpha_2 \sum x_{i1}\, x_{i2} + \ldots + \alpha_m \sum x_{i1}\, x_{im} = \sum x_{i1}\, y_i, \\ \alpha_1 \sum x_{i1}\, x_{i2} + \alpha_2 \sum x_{i2}^2 + \ldots + \alpha_m \sum x_{i2}\, x_{im} = \sum x_{i2}\, y_i, \\ \vdots \\ \alpha_1 \sum x_{i1}\, x_{im} + \alpha_2 \sum x_{i2}\, x_{im} + \ldots + \alpha_m \sum x_{im}^2 = \sum x_{im}\, y_i, \end{cases}$$

где суммирование везде ведется по i, от 1 до n. Видим, что у этой системы симметричная матрица. И, вообще, структура системы такова, что ее несложно запомнить.

Используя условие (3) из предыдущего пункта, получаем

$$\alpha = \left(X^T X\right)^{-1} X^T Y\,,$$

т.е. явное выражение для искомого вектора оценок МНК.

5.4. Модели парной регрессии

Простейшим и достаточно часто используемым видом регрессии является, так называемая, *парная регрессия*, т.е. регрессия y только на один фактор x. Она, конечно же, является частным случаем множественной. Но благодаря простоте на ее примере легче проиллюстрировать некоторые аспекты регрессионного анализа в целом.

В случае парной регрессии, выбор общего вида регрессионной модели можно осуществить на основе анализа так называемого *корреляционного поля*. Пусть, имеется парная выборка

$$\begin{array}{c|c|c|c} y_1 & y_2 & \cdots & y_n \\ x_1 & x_2 & \cdots & x_n \end{array}\,.$$

Корреляционным полем называется чертеж, на котором изображены точки с координатами, соответствующими этой выборке (рис. 5.1). Ясно, что для многомерной выборки построить корреляционное поле, к сожалению, нельзя.

По форме корреляционного поля можно сделать вывод о том, в каком классе зависимостей следует искать аппроксимацию функции регрессии. Пусть, например, выбрана, наиболее часто используемая параболическая модель

$$y = a_0 + a_1\, x + a_2\, x^2 + \varepsilon\,.$$

Она может быть рассмотрена в рамках общей теории многомерных линейных моделей, если считать, что матрица значений объясняющих факторов имеет вид

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}.$$

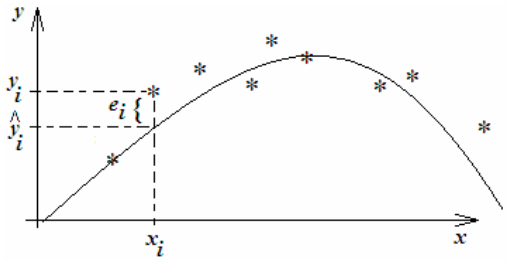


Рис. 5.1.

Система нормальных уравнений в этом случае запишется так

$$\begin{cases} \sum\limits_i y_i = n\,\alpha_0 + \alpha_1 \sum\limits_i x_i + \alpha_2 \sum\limits_i x_i^2; \\ \sum\limits_i x_i\, y_i = \alpha_0 \sum\limits_i x_i + \alpha_1 \sum\limits_i x_i^2 + \alpha_2 \sum\limits_i x_i^3; \\ \sum\limits_i x_i^2\, y_i = \alpha_0 \sum\limits_i x_i^2 + \alpha_1 \sum\limits_i x_i^3 + \alpha_2 \sum\limits_i x_i^4. \end{cases}$$

Если же, например, принята гиперболическая модель

$$y = a_0 + \frac{a_1}{x} + \varepsilon ,$$

то матрица значений объясняющих факторов будет иметь вид

$$X = \begin{pmatrix} 1 & \frac{1}{x_1} \\ 1 & \frac{1}{x_2} \\ \cdots & \cdots \\ 1 & \frac{1}{x_n} \end{pmatrix}.$$

Кроме рассмотренных используются еще, например, кубическая, экспоненциальная, логарифмическая и показательная парные регрессии.

Пример. Пусть зарегистрированы следующие данные, характеризующие интенсивность орошения (x) и урожайность (y) некоторой зерновой культуры (табл. 5.5) [16].

Таблица 5.5.

y, ц/га	6	7	13	16	20	24	22	20
x, дм	0,9	1,0	1,8	2,4	4,0	5,8	7,6	8,5

Найти коэффициенты параболического регрессионного уравнения.

Решение. Рассчитываем сначала коэффициенты системы нормальных уравнений

$$n = 8;\ \sum_{i=1}^{n} x_i = 32;\ \sum_{i=1}^{n} y_i = 128;\ \sum_{i=1}^{n} x_i\, y_i = 630{,}6;\ \sum_{i=1}^{n} x_i^2 = 190{,}48,$$

$$\sum_{i=1}^{n} y_i\, x_i^2 = 3989{,}22;\ \sum_{i=1}^{n} x_i^3 = 1333{,}6;\ \sum_{i=1}^{n} x_i^4 = 9989{,}3.$$

Решая соответствующую систему, например, методом Гаусса, находим искомые оценки коэффициентов регрессии

$$\alpha_0 = 0{,}9127;\ \alpha_1 = 7{,}7231;\ \alpha_2 = -0{,}6638 .$$

Итак, искомое уравнение параболической регрессии имеет вид

$$\hat{y} = 0{,}9127 + 7{,}7231x - 0{,}6638\, x^2.$$

5.5. Статистические свойства вектора оценок МНК

Легко видеть, что

$$M(\alpha) = M\left(\left(X^T X\right)^{-1} X^T Y\right) = \left(X^T X\right)^{-1} X^T M(Y) =$$

$$= \left(X^T X\right)^{-1} X^T M(X a + \varepsilon) = \left(X^T X\right)^{-1} X^T X a + \left(X^T X\right)^{-1} X^T M(\varepsilon) = a ,$$

т.е. МНК оценка оказывается несмещенной.

Из той же цепочки равенств следует, что

$$\alpha = a + \left(X^T X\right)^{-1} X^T \varepsilon ,$$

т.е. точечные значения МНК оценок будут содержать случайные ошибки.

Изучим дисперсию этих оценок. Поскольку в данном случае мы имеем случайный вектор, то следует говорить о ковариационной матрице этого вектора. Имеем

$$V(\alpha) = V\left(a + \left(X^T X\right)^{-1} X^T \varepsilon\right) = V\left(\left(X^T X\right)^{-1} X^T \varepsilon\right) =$$

$$= \left(X^T X\right)^{-1} X^T V(\varepsilon) X \left(\left(X^T X\right)^{-1}\right)^T = \ldots ,$$

и поскольку в силу условия (1) Гаусса–Маркова

$$V(\varepsilon) = \sigma^2 I_n ,$$

а матрица $\left(X^T X\right)^{-1}$ симметрична, получаем

$$V(\alpha) = \sigma^2 \left(X^T X\right)^{-1}.$$

Напомним, что на диагонали этой матрицы стоят дисперсии отдельных элементов вектора оценок, т.е. дисперсии оценок отдельных коэффициентов. Заметим, что, по крайней мере, диагональные элементы матрицы $\left(X^T X\right)$ с ростом *n* неограниченно возрастают, а это означает, что диагональные элементы обратной матрицы $\left(X^T X\right)^{-1}$ будут стремиться к нулю при $n \to \infty$. Отсюда следует состоятельность оценок МНК.

В известной теореме Гаусса–Маркова [26] устанавливается эффективность МНК оценок.

Для некоторых ниже рассматривающихся фактов необходимо иметь оценку ковариационной матрицы $V(\alpha)$. Ясно, что в качестве нее следует использовать

$$\hat{V}(\alpha) = S^2 \left(X^T X\right)^{-1},$$

где S^2 – оценка неизвестной дисперсии σ^2 случайной составляющей ε. Оказывается, что несмещенной, состоятельной и асимптотически эффективной оценкой σ^2 является

$$S^2 = \frac{e^T e}{n-m}.$$

Справедлива теорема, аналогичная теореме 3.2.

***Теорема* 5.1.** $(n-m)\dfrac{S^2}{\sigma^2} \sim \chi^2(n-m)$.

5.6. Проверка статистической значимости отдельных коэффициентов регрессии

Любая математическая модель должна выражать реальные связи между параметрами процесса. Это необходимо для того, чтобы анализ процесса на основе этой модели был бы адекватен. В частности, модель должна учитывать факторы действительно влияющие на результирующий показатель *y*, и не содержать факторов несущественно влияющих на процесс. Ясно, что отбор таких действительно значимых факторов – важная и непростая задача. Оказывается, в регрессионном анализе имеются инструменты, позволяющие решать эту задачу.

Предположим, что в число объясняющих факторов линейной модели включен параметр x_i, который в действительности не влияет на y, тогда теоретическое значение соответствующего коэффициента a_i равно нулю. Но мы имеем лишь статистическую оценку α_i этого коэффициента, которая, как и всегда, не равна в точности теоретическому значению, а значит не равна нулю, и возникает возможность сделать ошибочный вывод о наличии влияния x_i на y.

Таким образом, следовало бы проверить статистическую гипотезу

$$H_0 : a_i = 0.$$

Оказывается, это можно сделать. Действительно, мы знаем, что

$$\alpha_i \sim N\left(a_i, \sigma^2_{\alpha_i}\right) \Rightarrow \frac{\alpha_i - a_i}{\sigma_{\alpha_i}} \sim N(0,1),$$

где $\sigma^2_{\alpha_i}$ – соответствующий диагональный элемент матрицы

$$V(\alpha) = \sigma^2 \left(X^T X\right)^{-1}.$$

Из теоремы 5.1

$$(n-m)\frac{S^2_{\alpha_i}}{\sigma^2_{\alpha_i}} \sim \chi^2(n-m),$$

и несложно доказать, что

$$\frac{\alpha_i - a_i}{S_{\alpha_i}} \sim t(n-m).$$

Из этого распределения, в частности, легко строятся доверительные интервалы для теоретических значений коэффициентов регрессии.

Тогда двухсторонняя критическая область уровня значимости γ для рассматриваемой нулевой гипотезы H_0 имеет вид

$$Q : \frac{\alpha_i}{S_{\alpha_i}} < -t_{\gamma/2}(n-m),\ t_{\gamma/2}(n-1) < \frac{\alpha_i}{S_{\alpha_i}}.$$

Описанная методика проверки статистической значимости коэффициентов регрессии является важным инструментом в регрессионном анализе. На ее основе строятся, так называемые, *пошаговые процедуры отбора значимых переменных*, кото-

рые позволяют выявить факторы, действительно влияющие на интересующий нас эндогенный показатель y, и, тем самым, получить не только количественную, но и качественную информацию о внутреннем содержании изучаемого процесса.

Эти пошаговые процедуры состоят из следующих этапов:

1) составляют список всех возможных объясняющих факторов;
2) оценивают коэффициенты соответствующей модели;
3) проверяют значимость полученных коэффициентов и исключают из списка наименее значащий фактор;
4) повторяют второй и третий этапы до тех пор, пока в модели не останутся только значимые факторы;
5) обычно в завершении проверяют, нельзя ли включить в модель, значимым образом, какие-то ранее исключенные факторы.

Ясно, что подобные процедуры почти всегда требуют использования компьютеров.

5.7. Проверка гипотез о нескольких коэффициентах регрессии

Гораздо более общей, чем рассмотренная в предыдущем пункте, является гипотеза

$$H_0 : K\,a = b\,,$$

здесь матрица $K_{r\times m}$ и вектор b – заданы. Эта гипотеза позволяет проверять любые линейные соотношения между коэффициентами. Так, например, для гипотезы

$$H_0 : \begin{cases} a_1 - 2a_3 = 4, \\ a_2 = \dfrac{a_3 + a_4}{2}, \\ a_4 = 5, \end{cases}$$

матрица K и вектор b имеют вид

$$K = \begin{pmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & -\dfrac{1}{2} & -\dfrac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 0 \\ 5 \end{pmatrix}.$$

Оказывается, при справедливости рассматриваемой нулевой гипотезы

$$F = \frac{(K\,\alpha - b)^T \left(K\left(X^T X\right)K^T\right)^{-1} (K\,\alpha - b)/r}{S^2} \sim F\left(\frac{r}{n-m}\right),$$

из чего обычным образом строятся критические области. Из этого же факта вытекает методика построения эллипсоида, в котором с заданной доверительной вероятностью лежит истинный вектор коэффициентов регрессии.

Особенный интерес представляет гипотеза

$$H_0 : K\,a = 0\,,$$

где матрица K имеет вид

$$K_{(m-1)\times m} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Эта гипотеза состоит в незначимости всех коэффициентов, кроме первого. И если первый столбец матрицы X состоит из единиц, т.е. модель содержит свободное слагаемое, то расчетная статистика для проверки гипотезы имеет вид

$$F = \frac{\left(\hat{Y} - \overline{Y}\right)^T \left(\hat{Y} - \overline{Y}\right)/m - 1}{S^2} \sim F\left(\frac{m-1}{n-m}\right).$$

При таких матрицах K и X гипотеза H_0 состоит в том, что построенное регрессионное уравнение дает прогноз не лучше, чем простое среднее значение объясняемого фактора y, поэтому такую гипотезу называют гипотезой о *значимости* всего *регрессионного уравнения* в целом. Если эту гипотезу нельзя отклонить, то, значит, построенная регрессия дает такие большие ошибки, что нет никакого смысла использовать ее ни для прогноза, ни вообще для анализа процесса.

5.8. Коэффициент детерминации

Рассмотрим вопрос: как вообще оценить качество построенной регрессии? Введем следующее понятие.

***Определение* 5.1.** *Коэффициентом детерминации* с.в. Y по X называется число

$$K_d(Y,\ X) = 1 - \frac{D(Y/X)}{D(Y)}\,.$$

В этом определении:

- $D(Y/X)$ – условная дисперсия Y, при заданном значении X, т.е. величина возможного разброса значений Y при заданном X;
- $D(Y)$ – общая дисперсия Y.

Если Y полностью определяется значением X, т.е. между ними строго функциональная детерминированная зависимость, то условная дисперсия будет равна нулю, и

$$K_d(Y,\ X) = 1\,.$$

Если знание значения X не уменьшает возможной дисперсии Y, то

$$D(Y/X) = D(Y), \quad \text{и} \quad K_d(Y,\ X) = 0\,.$$

Таким образом, $K_d(Y, X)$ может принимать значения от 0 до 1 и является универсальным измерителем тесноты статистической связи между различными величинами.

В регрессионном анализе рассматривают выборочный коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2},$$

который выражает отношение так называемой *необъясненной вариации* признака к его общей вариации. Если например, $R^2 = 0{,}9$, то говорят, что изменение y на 90 % объясняется изменением соответствующих экзогенных признаков, и на 10 % какими-то другими причинами. Величина R^2 имеет важную геометрическую интерпретацию [26].

Для проверки значимости уравнения регрессии можно использовать статистику

$$F = \frac{R^2}{1-R^2}\frac{n-m}{m-1} \sim F\left(\frac{m-1}{n-m}\right),$$

которая, как не сложно выяснить, совпадает с рассмотренной выше.

5.9. Прогнозирование с помощью регрессионной зависимости

Ясно, что величина

$$\widetilde{y} = \alpha_1 \widetilde{x}_1 + ... + \alpha_m \widetilde{x}_m$$

является точечным прогнозом случайного значения фактора y при заданных значениях объясняющих факторов $\widetilde{x}_1, ..., \widetilde{x}_m$.

Будем обозначать это значение $y|_{\widetilde{x}}$.

Далее видим, что

$$M(\widetilde{y}) = M(\alpha_1 \widetilde{x}_1 + ... + \alpha_m \widetilde{x}_m) = a_1 \widetilde{x}_1 + ... + a_m \widetilde{x}_m + M(\varepsilon) = M(y|\widetilde{x}_1, ..., \widetilde{x}_m)\,,$$

т.е. $\widetilde{y}$ – является также несмещенной оценкой условного математического ожидания с.в. $y|_{\widetilde{x}}$, а тогда

$$(\widetilde{y} - y|_{\widetilde{x}}) \sim N\left(0, \sigma^2_{\widetilde{y}-y|_{\widetilde{x}}}\right).$$

Находим

$$D(\widetilde{y} - y|_{\widetilde{x}}) = D(\alpha_1 \widetilde{x}_1 + ... + \alpha_m \widetilde{x}_m - (a_1 \widetilde{x}_1 + ... + a_m \widetilde{x}_m + \varepsilon)) = ...$$

по свойствам дисперсии

$$... = \sum_{i=1}^{m} \widetilde{x}_i^2\, D(\alpha_i) + \sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} \widetilde{x}_i\, \widetilde{x}_j\, \mathrm{cov}(\alpha_i, \alpha_j) + D(\varepsilon) =$$

$$= \widetilde{x}^T V(\alpha)\widetilde{x} + \sigma^2 = \sigma^2\, \widetilde{x}^T \left(X^T X\right)\widetilde{x} + \sigma^2 = \sigma^2\left(1 + \widetilde{x}^T \left(X^T X\right)\widetilde{x}\right).$$

Таким образом, оценкой этой дисперсии является величина

$$S^2_{\widetilde{y}-y|_{\widetilde{x}}} = S^2\left(1 + \widetilde{x}^T \left(X^T X\right)\widetilde{x}\right).$$

Отсюда, аналогично вышесказанному, получаем

$$\frac{\widetilde{y} - y|_{\widetilde{x}}}{S_{\widetilde{y}-y|_{\widetilde{x}}}} \sim t(n-m)\,.$$

Тогда, с вероятностью $p = 1-\gamma$ выполняется

$$\widetilde{y} - t_{\gamma/2}(n-m)\, S_{\widetilde{y}-y|_{\widetilde{x}}} \le y|_{\widetilde{x}} \le \widetilde{y} + t_{\gamma/2}(n-m)\, S_{\widetilde{y}-y|_{\widetilde{x}}}\,.$$

Это неравенство определяет интервальный прогноз значения $y|_{\widetilde{x}}$.

В случае парной регрессии выражения для оценки дисперсии прогноза можно записать в виде:

- для прямой $S^2_{\tilde{y}-y|\tilde{x}} = \dfrac{e^T e}{n-2}\left(1+\dfrac{1}{n}+\dfrac{(\tilde{x}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$;
- для параболы $S^2_{\tilde{y}-y|\tilde{x}} = \dfrac{e^T e}{n-3}\times$

$$\times\left(1+\frac{(\tilde{x}-\bar{x})^2}{\sum(x_i-\bar{x})^2}+\frac{\sum(x_i-\bar{x})^4-2\left(\sum(x_i-\bar{x})^2\right)\left((\tilde{x}-\bar{x})^2\right)+n(\tilde{x}-\bar{x})^4}{n\sum(x_i-\bar{x})^4-\left(\sum(x_i-\bar{x})^2\right)^2}\right).$$

5.10. Выборочный коэффициент корреляции и проверка его статистической значимости

Пусть имеется парная выборка значений с.в. X и Y. Тогда, в качестве оценки коэффициента корреляции r_{XY} между ними, используют, так называемый, *выборочный коэффициент корреляции*

$$\rho_{XY} = \frac{n\sum x_i\, y_i - \sum x_i \sum y_i}{\sqrt{[n\sum {x_i}^2 - \left(\sum x_i\right)^2]\,[n\sum {y_i}^2 - \left(\sum y_i\right)^2]}},$$

здесь суммирование везде ведется от 1 до *n*. Как и всегда, даже если теоретическое значение $r_{XY}=0$, т.е. с.в. X и Y – независимы, его выборочная оценка ρ_{XY} не будет равна нулю. Отсюда появляется необходимость проверки гипотезы $H_0: r_{XY}=0$, для чего используют статистику

$$t = \frac{\rho_{XY}\sqrt{n-2}}{\sqrt{1-\rho^2_{XY}}},$$

которая, при истинности H_0 имеет распределение $t(n-2)$, т.е. $t \sim t(n-2)$.

Построение критических областей проводят стандартным образом.

5.11. Коэффициенты ранговой корреляции

Как известно, признаки могут измеряться не в интервальной, а в порядковой шкале. Оказывается, что и в этом случае можно изучать степень тесноты их статистической связи (с.т.с.с.). Для этого используют так называемые *коэффициенты ранговой корреляции*.

***Определение* 5.2**. *Ранжировкой n* объектов называется приписывание каждому из этих объектов порядкового номера от 1 до n, в соответствии с уровнем некоторого их количественного или качественного признака.

Таким образом ранг 1 приписывается наиболее важному или крупному объекту, ранг 2 – следующему, и т.д. При этом за начало может быть взят как наименьший, так и наибольший уровень признака.

Ранжировки могут строиться как на основе экспертных оценок, так и объективной информации о численных значениях параметров объектов параметров. Например, совокупность некоторых объектов (товаров и т.д.) может быть проранжирована по их стоимости.

Если два или более объектов поставлены экспертом на один уровень, то их ранги считаются равными среднему арифметическому номеров позиций, которые они заняли. Так могут возникать *дробные* ранги. Например, два объекта считаются одинаковыми, а их качество следует за 20-м объектом, тем самым они занимают 21 и 22 позиции. Тогда их рангами будет 21,5.

Если объекты ранжированы по двум признакам, то может возникнуть необходимость измерения степени согласованности этих двух ранжировок. Такая задача возникает, в частности, при:

- необходимости установить силу связи между качественными признаками (если бы они были количественные, то можно было бы использовать обычный коэффициент корреляции);
- желании сопоставить ранжировки одних и тех же объектов, предложенные двумя разными экспертами, чтобы выяснить степень согласованности их мнений, а, тем самым, – достоверность этих ранжировок.

Чаще всего используют два коэффициента ранговой корреляции. Первым из них обычно рассматривают *коэффициент ранговой корреляции Спирмэна*

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)},$$

где d_i – разность значений рангов, приписанных разными экспертами одному и тому же объекту.

Коэффициент r_S является частным случаем коэффициента парной корреляции. В самом деле его можно получить, преобразовав соответствующее выражение путем введения в него рангов вместо значений *x* и *y*.

Если два ряда рангов полностью совпадают, то

$$\sum_{i=1}^{n} d_i^2 = 0,$$

и, следовательно, $r_S = 1$.

Можно доказать, что при полной обратной связи, т.е. когда ранги двух рядов расположены в обратном порядке, будет иметь место $r_S = -1$.

Поскольку r_S определяется на основе выборки, возникает необходимость в проверке гипотезы

$$H_0 : \rho_S = 0,$$

где ρ_S – генеральный коэффициент ранговой корреляции. Оказывается, при отсутствии связи между рангами и $n \to \infty$

$$\frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \sim t(n-2).$$

Вторым показателем с.т.с.с. для ранжировок является *коэффициент ранговой корреляции Кендалла*.

Пусть выборка объема *n* содержит независимые объекты и имеется две их ранжировки. Упорядочим объекты в соответствии с одной их них. Обозначим y_i – ранг объекта по второй ранжировке, имеющего *i*-й ранг, в соответствии с первой (т.е. стоящего на i-м месте после упорядочивания). Обозначим, R_i – количество объектов, стоящих справа от i-го, и имеющих ниже ранг (большее число) по второй ранжировке. Находим

$$R = R_1 + R_2 + \ldots + R_{n-1}.$$

Выборочный коэффициент ранговой корреляции Кендалла определяется формулой

$$\tau_e = \frac{4R}{n(n-1)} - 1.$$

Величину R_i будем называть количеством соответствий. Можно наоборот считать количество, так называемых, *инверсий*, тогда

$$\tau_e = 1 - \frac{4R}{n(n-1)}.$$

Оказывается, при отсутствии связи между рангами и $n \to \infty$

$$T = \frac{\tau_e}{\sqrt{2(2n+5)/(9n^2-9n)}} \sim N(0, 1).$$

Пример. Допустим, что при ранжировании отметок на вступительных экзаменах и средних баллов за первую экзаменационную сессию одних и тех же лиц получены ранги, указанные в табл. 5.6. Проверить согласованность этих ранжировок.

Таблица 5.6.

Студент	А	Б	В	Г	Д	У	Ж	З	И	К
Вступительные экзамены	2	5	6	1	4	10	7	8	3	9
Экзаменационная сессия	3	6	4	1	2	7	8	10	5	9
d_i	–1	–1	2	0	2	3	–1	–2	–2	0

Решение. **Находим коэффициент Спирмэна**

$$\sum d_i^2 = 28 \Rightarrow r_S = 1 - \frac{6 \cdot 28}{10 \cdot (10^2 - 1)} = 0{,}83.$$

Такая величина коэффициента ранговой корреляции говорит о достаточно высокой связи между результатами двух видов экзаменов. Однако сделаем проверку его значимости, например, при α=0,05. Если нулевая гипотеза верна, то должно быть

$$\left|\frac{0{,}83 \cdot \sqrt{10-2}}{\sqrt{1-0{,}83^2}}\right| = 4{,}2 \le t_{0{,}05/2}(8) = 2{,}31.$$

Видим, что неравенство не выполняется – нулевая гипотеза отклоняется. Связь между результатами на вступительных и экзаменах в сессии есть.

Рассмотрим на этом же примере использование коэффициента Кендалла. Запишем список студентов, упорядочив их по результату вступительных экзаменов. Такой список и величины R_i указаны в табл. 5.7.

Таблица 5.7.

Студент	Г	А	И	Д	Б	В	Ж	З	К	Е
Вступительные экзамены	1	2	3	4	5	6	7	8	9	10
Экзаменационная сессия	1	3	5	2	6	4	8	10	9	7
R_i	9	7	5	6	4	4	2	0	0	0

Находим

$$R = 9 + 7 + 5 + 6 + 4 + 4 + 2 = 37 \Rightarrow \tau_e = \frac{4 \cdot 37}{10 \cdot 9} - 1 \approx 0{,}64.$$

При той же значимости $\alpha = 0{,}05$ проверим нулевую гипотезу

$$T_{\text{расч}} = \frac{0{,}64}{\sqrt{2 \cdot (2 \cdot 10 + 5)/(9 \cdot 10 \cdot (10 - 1))}} = 2{,}576 .$$

По таблицам нормального распределения $u_{0{,}05/2} = 1{,}96$. Тогда, поскольку

$$T_{\text{расч}} = 2{,}576 > u_{0{,}05/2} = 1{,}96 ,$$

то и по коэффициенту Кендалла нулевую гипотезу следует отвергнуть. Связь между оценками по двум экзаменам значима.

Вопросы для самопроверки

1. В чем разница между корреляционной и регрессионной зависимостями?
2. Что такое регрессионная модель? В чем сложность ее выбора?
3. В чем состоят цели построения регрессионных зависимостей?
4. Приведите пример бытовой ситуации, когда было бы желательно иметь регрессионную зависимость.
5. Какие бы вы использовали объясняющие факторы в модели цены на квартиру?
6. В чем основная идея МНК? В чем его преимущества перед другими методами оценки коэффициентов регрессии?
7. Какие еще вы можете предложить методы оценки коэффициентов регрессии?
8. Зачем нужны интервальные прогнозы?
9. Зачем нужно проверять значимость коэффициентов регрессии и корреляции?
10. Придумайте пример использования коэффициента Спирмена.

ЗАКЛЮЧЕНИЕ

Данное пособие содержит лишь основы методов математической статистики. Его цель – ввести нематематиков в область соответствующих задач и дать представление об огромном потенциале математической статистики для анализа данных. Упрощения, которые иногда встречаются в тексте, связаны именно с этим. Для более углубленного изучения соответствующих методов следует прибегнуть к многочисленной и богатой специальной литературе, которой всегда издавалось, а в последнее время особенно, весьма много [6, 9, 11]. В частности, в последние годы выходят серьезные издания, содержащие зарубежный опыт практического использования этих методов, особенно регрессионного анализа [18].

Следует помнить также, что серьезное использование методов математической статистики немыслимо без применения компьютеров, чему посвящено также немало литературы последних лет [12, 13]. Современная математическая статистика все в большей мере становится многомерной [19], чему также способствует распространение компьютерной техники.

Появление мощных и удобных статистических пакетов для персональных компьютеров позволяет использовать их не только как специальный инструмент научных исследований, но и как общеупотребительный инструмент плановых, аналитических, маркетинговых отделов производственных и торговых корпораций, банков и страховых компаний, правительственных и медицинских учреждений и даже представителей мелкого бизнеса. Статистические программные пакеты сделали соответствующие методы более доступными и наглядными, так как трудоемкую работу по расчету различных статистик, параметров, характеристик, построению таблиц и графиков в основном стал выполнять компьютер, а исследователю осталась главным образом творческая работа: постановка задачи, выбор методов ее решения и интерпретация результатов.

Среди множества используемых для этих целей пакетов прикладных программ выделим популярные в России универсальные и специализированные статистические пакеты: отечественные STADIA, Эвриста, Статистик-консультант, Олимп: СтатЭксперт и американские STATGRAPHICS, SPSS, SYSTAT, STATISTICA/w и др.

СПИСОК ЛИТЕРАТУРЫ

1. Гмурман, В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М. : Высшая школа, 1977.
2. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике / В.Е. Гмурман. – М. : Высшая школа, 1979.
3. Кремер, Н.Ш. Теория вероятностей и математическая статистика / Н.Ш. Кремер. – М. : ЮНИТИ, 2000.
4. Колемаев, В.А. Теория вероятностей и математическая статистика / В.А. Колемаев, В.Н. Калинина. – М. : Инфра-М, 1997.
5. Хили, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хили ; пер. с англ. – СПб. : Питер, 2005.
6. Айвазян, С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998.
7. Вентцель, Е.С. Теория вероятностей и ее инженерные приложения / Е.С. Вентцель, Л.А. Овчаров – М. : Наука, 1988.
8. Войтенко, М.А. Руководство к решению задач по теории вероятностей / М.А. Войтенко. – М. : Изд. ВЗФЭИ, 1988.
9. Дубров, А.М. Многомерные статистические методы / А.М. Дубров, В.С. Мхитарян, Трошин Л.И. – М. : Финансы и статистика, 1998.
10. Гихман, И.И. Теория вероятностей и математическая статистика / И.И. Гихман, А.В. Скороход, М.И. Ядренко. – К. : Выща школа, 1988.
11. Справочник по прикладной статистике / под ред. Э. Ллойда, У. Лидермана ; пер. с англ. – М. : Финансы и статистика, 1989.
12. Тюрин, Ю.Н. Анализ данных на компьютерах / Ю.Н. Тюрин, А.А. Макаров ; под ред. В.Э. Фигурнова. – М. : Инфра-М, 1998.
13. Боровиков, В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В. Боровиков. – СПБ. : Питер, 2001.
14. Уотшем, Т. Дж. Количественные методы в финансах / Т. Уотшем, К. Паррамоу ; пер. с англ. – М. : ЮНИТИ, Финансы, 1999.
15. Ферстер, Э. Методы корреляционного и регрессионного анализа / Э. Ферстер, Б. Ренц ; пер. с нем. – М. : Финансы и статистика, 1983.
16. Четыркин, Е.М. Вероятность и статистика / Е.М. Четыркин, И.Л. Калихман. – М. : Финансы и статистика, 1982.
17. Шарп, У. Инвестиции / У. Шарп, А. Гордон Док., Д. Бейли ; пер. с англ. – М. : Инфра-М, 1997.
18. Берндт, Э.Р. Практика эконометрики: классика и современность : учебник / Э.Р. Берндт ; под ред. проф. С.А. Айвазяна ; пер. с англ. – М. : ЮНИТИ-ДАНА, 2005.
19. Многомерный статистический анализ в экономике : учеб. пособие для вузов / под ред. проф. В.Н. Тамашевича. – М. : ЮНИТИ-ДАНА, 1999.
20. Болч, Б. Многомерные статистические методы для экономики / Б. Болч, К. Дж.Хуань. – М. : Финансы и статистика, 1979.
21. Большев, Л.Н. Таблицы математической статистики / Л.Н. Большев, К.В.Смирнов. – М. : Наука, 1983.
22. Бро, Г.Г. Математические методы экономического анализа на предприятии / Г.Г. Бро, Л.М. Шнайдман. – М. : Экономика, 1976.
23. Джонсон, Дж. Эконометрические методы / Дж. Джонсон. – М. : Статистка, 1980.
24. Дрейпер, К. Прикладной регрессионный анализ. Кн. 1, 2 / К. Дрейпер, Г. Смит. – М. : Финансы и статистика, 1986.
25. Дубровский, С.А. Прикладной многомерный статистический анализ / С.А. Дубровский. – М. : Финансы и статистика, 1982.
26. Магнус, Я.Р. Эконометрика / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – М. : Дело, 2001.
27. Дюк, В.А. Компьютерная психодиагностика / В.А. Дюк. – СПб. : Братство, 1994.
28. Елисеева, И.И. Прикладной статистический анализ / И.И. Елисеева, В.О. Рукавишников. – М. : Финансы и статистика, 1982.
29. Енюков, И.С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА / И.С. Енюков. – М. : Финансы и статистика, 1986.
30. Кади, Дж. Количественные методы в экономике / Дж. Кади. – М. : Прогресс, 1977.
31. Кейн, Э. Экономическая статистика и эконометрия / Э. Кейн – Вып. 2. – М. : Статистика, 1977.
32. Кендалл, М. Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт. – М. : Наука, 1976.
33. Кендалл, М. Статистические выводы и связи \ М. Кендалл, А. Стьюарт. – М. : Наука, 1973.

ВВЕДЕНИЕ …… 3

1. **ИСЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ СЛУЧАЙНЫХ СОБЫТИЙ** ……. 4

1.1. Общие понятия о случайном событии и его вероятности. Действия над случайными событиями …….. 4

1.2. Схема с равновозможными исходами. Классическое определение вероятности …….. 6

1.3. Использование комбинаторных формул ……... 7

1.4. Схема с неравновозможными исходами. Статистическое определение вероятности ……... 9

1.5. Схема с несчетным множеством исходов. Геометрическое определение вероятности …….. 11

1.6. Теоремы сложения и умножения вероятностей ……... 14

1.7. Формулы условной вероятности, полной вероятности и формула Байеса ……... 15

1.8. Аксиомы теории вероятности. Вероятностное пространство … 18

1.9. Последовательности испытаний. Схема Бернулли …….. 19

1.10. Локальная и интегральная теоремы Муавра-Лапласа ……. 20

2. ОСНОВЫ ТЕОРИИ СЛУЧАЙНЫХ ВЕЛИЧИН ……. 24

2.1. Определение случайной величины. Задание дискретной случайной величины ……... 24

2.2. Непрерывная с.в. Функция распределения ……... 26

2.3. Функция плотности распределения с.в. …… 28

2.4. Математическое ожидание с.в. …….. 30

2.5. Дисперсия случайных величин …….. 32

2.6. Независимость с.в. и коэффициент корреляции …….. 34

2.7. Закон больших чисел …….. 37

2.8. Центральная предельная теорема …….. 38

2.9. Многомерные случайные величины …… 40

2.10. Функции от случайных величин …… 42

3. ОЦЕНКА ПАРАМЕТРОВ И ЗАКОНА РАСПРЕДЕЛЕНИЯ С.В. ... 44

3.1. Основные понятия выборочного метода …….. 44

3.2. Свойства статистических оценок ……... 45

3.3. Оценка математического ожидания с.в. ……... 46

3.4. Оценки дисперсии с.в. ……..... 48

3.5. Оценка доли признака …… 50

3.6. Стандартные статистические распределения и их критические границы …… 52

3.7. Понятие доверительного интервала …….. 55

3.8. Доверительный интервал для математического ожидания нормальной с.в. …….. 55

3.9. Доверительный интервал для дисперсии нормального распределения …….. 58

3.10. Доверительный интервал для генеральной доли признака …. 59

3.11. Определение необходимого объема выборки ……... 60

3.12. Оценка функции распределения …… 61

3.13. Оценка функции плотности распределения …… 62
3.14. Метод моментов …….. 63
3.15. Метод максимального правдоподобия ……. 65
4. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ ……..……... 68
4.1. Общая схема проверки статистических гипотез …….. 68
4.2. Проверка простых гипотез с помощью доверительных интервалов …….. 70
4.3. Проверка гипотезы о равенстве дисперсий двух выборок ….. 72
4.4. Проверка гипотезы о равенстве средних …….. 74
4.5. Однофакторный дисперсионный анализ …….. 76
4.6. Проверка непараметрических гипотез. Критерий согласия Пирсона …….. 79
5. ОСНОВЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА …… 84
5.1. Детерминированные, статистические и регрессионные зависимости …….. 83
5.2. Регрессионная модель и предпосылки регрессионного анализа …….. 84
5.3. Метод наименьших квадратов (МНК) …….. 86
5.4. Модели парной регрессии …….. 88
5.5. Статистические свойства вектора оценок МНК …….. 90
5.6. Проверка статистической значимости отдельных коэффициентов регрессии ……. 91
5.7. Проверка гипотез о нескольких коэффициентах регрессии ... 93
5.8. Коэффициент детерминации …….. 94
5.9. Прогнозирование с помощью регрессионной зависимости … 95
5.10. Выборочный коэффициент корреляции и проверка его статистической значимости ……. 96
5.11. Коэффициенты ранговой корреляции ……... 97
ЗАКЛЮЧЕНИЕ ……. 101
СПИСОК ЛИТЕРАТУРЫ ……... 102

ДЛЯ ЗАМЕТОК