МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»

А.Г. Трофимов

ОСНОВЫ

МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Рекомендовано к изданию
УМО «Ядерные физика и технологии»

Москва 2016

УДК 519.22(075.8)
ББК 22.172я7
Т76

Трофимов А.Г. **Основы математической статистики**: Учебное пособие. М.: НИЯУ МИФИ, 2016. – 256 с.

Изложены базовые понятия описательной статистики, теории статистического оценивания, проверки статистических гипотез, основы дисперсионного, корреляционного и регрессионного анализов данных. Теоретический материал сопровождается многочисленными примерами, которые иллюстрируют основные приемы решения задач по математической статистике и проведения статистических исследований. По каждой теме курса предложены вопросы и задачи для самоконтроля.

Предназначено для студентов, специализирующихся по прикладной математике и информатике. Книга будет полезна для аспирантов и инженеров как справочное пособие при проведении прикладного статистического анализа данных.

Подготовлено в рамках Программы создания и развития НИЯУ МИФИ.

Рецензент:
д-р физ.-мат. наук, проф. А.П. Карпенко (МГТУ им. Н.Э. Баумана)

ISBN 978-5-7262-2262-2

ОГЛАВЛЕНИЕ

Предисловие 6

Глава 1. Методы статистического описания результатов наблюдений 8

§ 1. Понятие выборки 8

§ 2. Способы представления выборки 11

§ 3. Эмпирическая функция распределения. Числовые характеристики выборки 18

§ 4. Диаграмма Box-and-Whisker 30

§ 5. Выборочные характеристики двумерного случайного вектора 31

Глава 2. Точечные оценки 37

§ 6. Свойства точечных оценок 37

§ 7. Методы получения точечных оценок 48

§ 8. Точечные оценки математического ожидания и дисперсии 53

Глава 3. Интервальные оценки 56

§ 9. Понятие доверительного интервала 56

§ 10. Метод построения доверительных интервалов 58

§ 11. Законы распределения некоторых статистик нормальной выборки 62

§ 12. Примеры построения интервальных оценок параметров нормального распределения 68

§ 13. Интервальная оценка вероятности «успеха» в схеме Бернулли 76

Глава 4. Проверка статистических гипотез 81

§ 14. Основные понятия и определения 81

§ 15. Алгоритм проверки статистических гипотез 94

§ 16. Проверка гипотез о параметрах нормально распределённой генеральной совокупности 100

§ 17. Проверка гипотез о вероятности «успеха» в схеме Бернулли 111

Глава 5. Критерии согласия 115

§ 18. Проверка гипотез о виде распределения. Критерий Колмогорова 115

§ 19. Критерий «омега-квадрат» 119

§ 20. Критерий Пирсона 122

§ 21. Проверка гипотезы о нормальности распределения 127

§ 22. Проверка гипотез об однородности выборок. Критерий знаков 128

§ 23. Критерий Манна–Уитни 133

§ 24. Модификации критериев Колмогорова, «омега-квадрат» и Пирсона для проверки гипотез об однородности выборок 137

Глава 6. Анализ статистических взаимосвязей 144

§ 25. Виды связей между величинами 144

§ 26. Анализ статистической связи между номинальными величинами. Таблицы сопряжённости 147

§ 27. Виды дисперсий в совокупности, разделённой на части 152

§ 28. Однофакторный дисперсионный анализ 158

§ 29. Статистическая связь между компонентами нормально распределённого случайного вектора 165

§ 30. Корреляционное отношение 171

§ 31. Оценивание коэффициента корреляции по выборочным данным 177

§ 32. Оценивание коэффициента детерминации и корреляционного отношения по выборочным данным 183

§ 33. Ранговый коэффициент корреляции по Спирмену 194

§ 34. Ранговый коэффициент корреляции по Кендаллу 202

Глава 7. Регрессионный анализ 205

§ 35. Статистические модели 205

§ 36. Задачи регрессионного анализа 207

§ 37. Оценивание параметров уравнения регрессии. Метод наименьших квадратов 211

§ 38. Простейшая линейная регрессионная модель 215

§ 39. Линейная регрессионная модель общего вида 225

§ 40. Множественная линейная регрессия 234

§ 41. Некоторые регрессионные модели, сводящиеся к линейным 239

Приложение. Таблицы квантилей распределений 243

Список рекомендуемой литературы 254

ПРЕДИСЛОВИЕ

Установление закономерностей, которым подчинены массовые случайные явления, – одна из важнейших задач математической статистики, встречающаяся в природе. Опираясь на теорию вероятностей, математическая статистика изучает методы описания и анализа статистических данных наблюдений и экспериментов, предлагая вероятностную оценку результатов.

В связи с огромным разнообразием областей практического применения математической статистики и важностью получаемых статистических выводов, в настоящее время сложно представить специалиста по прикладной математике и информатике, не владеющего методами статистического анализа данных.

Анализ данных входит в область исследований всех научно-исследовательских отделов крупных компаний, многие компании имеют специализированные подразделения анализа данных, спрос на специалистов в этой области постоянно растёт, а математическая статистика – это основа анализа данных.

Владение методами математической статистики необходимо при изучении ряда прикладных математических дисциплин – теории исследования операций, статистического анализа временных рядов, машинного обучения, статистического моделирования систем и пр.

Предлагаемое учебное пособие соответствует программе преподавания математической статистики студентам факультета «Кибернетика и информационная безопасность» НИЯУ МИФИ. Учебное пособие включает семь глав, посвященных методам статистического описания результатов наблюдений, точечному и интервальному оцениванию параметров распределений, проверке параметрических и непараметрических статистических гипотез, основам дисперсионного, корреляционного и регрессионного анализов данных. В пособии излагается теория по всем темам курса в объёме, необходимом для решения прикладных задач, проведения статистических исследований и интерпретации их результатов.

Теоретический материал сопровождается многочисленными примерами, демонстрирующими применение на практике основных приемов статистического анализа данных. По каждой теме курса приводятся вопросы и задачи для самоконтроля.

Для приобретения навыков решения задач рекомендуется использовать книги: «Сборник задач по математике для втузов» под ред. А.В. Ефимова (ч. 4 «Теория вероятностей и математическая статистика») и «Математическая статистика» под ред. В.С. Зарубина и А.П. Крищенко, которые наиболее соответствуют содержанию и стилю настоящего учебного пособия.

Глава 1. МЕТОДЫ СТАТИСТИЧЕСКОГО ОПИСАНИЯ РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

§ 1. Понятие выборки

Математическая статистика – наука о математических методах, позволяющих по статистическим данным сформулировать выводы о свойствах изучаемого массового явления.

На практике редко доступна полная информация о модели изучаемого явления, описываемого в терминах некоторой случайной величины X. Чаще о законе распределения X имеется лишь частичная информация или эта информация полностью отсутствует. В таком случае возникают задачи восстановления параметров или вида неизвестного распределения F_X или определения его свойств.

Задачи математической статистики являются, в некотором смысле, обратными к задачам теории вероятностей. Если теория вероятностей позволяет при заданной вероятностной модели вычислить вероятности тех или иных случайных событий, то математическая статистика по результатам проводимых наблюдений (по исходам эксперимента) уточняет структуру вероятностной модели изучаемого явления.

Математическая статистика решает следующие задачи:

1) систематизация полученного статистического материала (этап описания массового явления);

2) выявление свойств и закономерностей изучаемого явления (этап анализа и прогноза).

Первой задачей занимается раздел математической статистики, называемый *описательной* (*дескриптивной*) *статистикой*. Описательная статистика предоставляет методы первичной обработки эмпирических данных, их наглядного представления в форме графиков и таблиц, а также их количественного описания с использованием статистических показателей. Методы описательной статистики, как правило, не требуют предположений о вероятностной природе данных.

Решению второй задачи посвящены теория оценивания и теория проверки статистических гипотез. В основе этих теорий лежат методы построения математических моделей наблюдений и выявления статистических закономерностей.

Точечное оценивание – вычисление приближённых значений характеристик статистических закономерностей по результатам наблюдений.

Интервальное оценивание – построение случайных множеств, называемых доверительными, которые с заданной вероятностью содержат оцениваемые характеристики.

Проверка статистических гипотез – принятие или отклонение по реализации наблюдений априорного предположения о неизвестных характеристиках статистических закономерностей.

С особенностями различных постановок задач оценивания связаны и различия в соответствующих методах статистических исследований.

Центральным понятием математической статистики является *выборка*. Выборка понимается следующим образом. Пусть случайная величина X наблюдается в эксперименте с комплексом условий G. Результатом этого эксперимента будет некоторое случайное число x – реализация случайной величины X. Повторим эксперимент n раз с неизменным комплексом условий. Результатом такого эксперимента будет случайный вектор $(x_1, \ldots, x_j, \ldots, x_n)$, где x_j – реализация случайной величины X в j-м эксперименте, $j = \overline{1,n}$. С другой стороны, вектор $(x_1, \ldots, x_n)$ можно рассматривать как единственную реализацию случайного вектора $(X_1, \ldots, X_n)$, где случайные величины $X_1, \ldots, X_n$ независимы в совокупности и каждая из которых имеет тот же закон распределения, что и случайная величина X.

Совокупность всех наблюдений случайной величины X, которые могли бы быть сделаны при данном комплексе условий, называется генеральной совокупностью случайной величины X, или просто *генеральной совокупностью* X. Распределение случайной величины X называется *распределением генеральной совокупности*. Число элементов, входящих в генеральную совокупность, называется

объёмом генеральной совокупности. Объём генеральной совокупности может быть как конечным, так и бесконечным.

Совокупность независимых случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и наблюдаемая случайная величина X, называется *случайной выборкой* из генеральной совокупности X. При этом число n называют *объёмом* случайной выборки, а случайные величины $X_1, \ldots, X_n$ – *элементами* случайной выборки. Любую реализацию $x_1, \ldots, x_n$ случайной выборки $X_1, \ldots, X_n$ будем называть *выборкой* из генеральной совокупности X, или *выборочной совокупностью*. Выборка из генеральной совокупности X представляет собой некоторое подмножество этой совокупности.

Пример 1.1. Эксперимент состоит в подбрасывании правильной игральной кости. Случайная величина X – число очков, выпавшее на верхней грани, возможные значения случайной величины X: 1, …, 6. В результате эксперимента получаем случайное число x – реализацию случайной величины X, $x \in \{1, ..., 6\}$. При повторении эксперимента n раз получаем выборку $x_1, \ldots, x_n$ наблюдений случайной величины X, $x_i \in \{1, ..., 6\}$, $i = \overline{1,n}$, или, что то же самое, единственное наблюдение $(x_1, \ldots, x_n)$ случайной выборки $X_1, \ldots, X_n$ объёма n. Генеральная совокупность случайной величины X содержит бесконечное число значений 1, …, 6 в равных пропорциях.

Пример 1.2. Исследуется качество партии выпущенных предприятием изделий. Случайная величина X – индикатор брака в изделии – принимает значение 1, если изделие оказалось бракованным, и 0 – в противном случае. В результате наблюдения случайной величины X (выбирая случайным образом изделие) получаем её реализацию x (0 или 1). Обследуя n изделий, получаем выборку наблюдений $x_1, \ldots, x_n$, $x_i \in \{0,1\}$, $i = \overline{1,n}$. Объём генеральной совокупности определяется объёмом партии выпущенных изделий. Объём выборки n не может превышать объём генеральной совокупности.

Понятие выборки может быть обобщено на случай, когда в результате эксперимента с некоторым комплексом условий G наблюдается несколько случайных величин. Например, пусть (x, y) – наблюдение двумерного случайного вектора (X, Y). Тогда случайная выборка объёма n представляет собой последовательность (X_1, Y_1), …, (X_n, Y_n) случайных векторов, а её реализация – последовательность векторов (x_1, y_1), …, (x_n, y_n).

Контрольные вопросы и задачи

1. В чём состоят основные задачи математической статистики?
2. Назовите основные задачи описательной статистики.
3. Дайте определение генеральной совокупности, случайной выборки и выборочной совокупности.
4. В каких случаях объём генеральной совокупности бесконечен?
5. Приведите пример эксперимента и опишите генеральную и выборочную совокупности для него.

§ 2. Способы представления выборки

Результаты наблюдений x_1, …, x_n генеральной совокупности X, записанные в порядке их регистрации, обычно труднообозримы и неудобны для дальнейшего анализа. Основная задача описательной статистики – получение такого представления выборки, которое позволит выявить характерные особенности совокупности исходных данных.

Одним из самых простых преобразований статистических данных является их упорядочивание по величине. *Вариационным рядом* выборки x_1, …, x_n называется способ её записи, при котором элементы упорядочиваются по возрастанию, т.е. вариационный ряд выборки – это последовательность выборочных значений

$$x_{(1)}, ..., x_{(i)}, ..., x_{(n)},$$

удовлетворяющих условию $x_{(1)} \le ... \le x_{(i)} \le ... \le x_{(n)}$.

Вариационный ряд $x_{(1)}, ..., x_{(i)}, ..., x_{(n)}$ выборки x_1, …, x_n можно рассматривать как реализацию вариационного ряда

$X_{(1)}, ..., X_{(i)}, ..., X_{(n)}$ случайной выборки $X_1, \ldots, X_n$. Случайную величину $X_{(i)}$ называют *i-й порядковой статистикой* (*i-th order statistic*). Число $x_{(i)}$ называют *i-м членом вариационного ряда*, или реализацией *i*-й порядковой статистики. Крайние члены $X_{(1)}$ и $X_{(n)}$ вариационного ряда называются *экстремальными порядковыми статистиками*. Для любой выборки реализации экстремальных порядковых статистик – это её минимальное и максимальное значения.

Можно показать, что функции распределения экстремальных порядковых статистик имеют вид

$$F_{(1)}(x) = P(X_{(1)} < x) = 1 - \left(1 - F_X(x)\right)^n, \qquad (1.1)$$

$$F_{(n)}(x) = P(X_{(n)} < x) = F_X^n(x). \qquad (1.2)$$

Эти соотношения позволяют оценить неизвестную функцию распределения $F_X(x)$ генеральной совокупности X, имея в эксперименте лишь минимальные и максимальные значения выборок.

Разность между максимальным и минимальным элементами выборки $x_{(n)} - x_{(1)}$ называется *размахом выборки* (*range of a sample*).

Различные значения случайной величины X называются *вариантами*.

Пусть выборка $x_1, \ldots, x_n$ случайной величины X содержит k вариантов $z_1, \ldots, z_k$, причём вариант z_i встречается n_i раз ($i = 1, ..., k$). Число n_i называется *частотой* варианта z_i. Очевидно, что сумма частот всех вариантов равна объёму выборки, $\sum_{i=1}^{k} n_i = n$.

Статистическим рядом выборки называется последовательность пар (z_i, n_i), $i = 1, ..., k$. Обычно статистический ряд записывают в виде таблицы, первая строка которой содержит варианты z_i, а вторая – частоты n_i (табл. 1.1), при этом варианты располагаются в порядке возрастания.

Таблица 1.1

Статистический ряд выборки

Варианты z_i	z_1	…	z_i	…	z_k
Частоты n_i	n_1	…	n_i	…	n_k

В частном случае, если все элементы выборки различны, то $k = n$, а частоты всех вариантов равны единице.

Пример 1.3. Игральная кость подбрасывается $n = 10$ раз. Случайная величина X – число очков, выпавшее на верхней грани. В результате эксперимента получены следующие наблюдения:

2; 5; 3; 3; 6; 5; 2; 1; 5; 2.

Записать вариационный и статистический ряды выборки. Рассчитать размах выборки.

Упорядочив элементы выборки по возрастанию, получим вариационный ряд:

1; 2; 2; 2; 3; 3; 5; 5; 5; 6.

Размах выборки: $6-1=5$. В выборке представлены $k = 5$ вариантов случайной величины X, т.е. 1, 2, 3, 5, 6, с частотами 1, 3, 2, 3, 1 соответственно. Таким образом, статистический ряд имеет вид:

z_i	1	2	3	5	6
n_i	1	3	2	3	1

При большом числе вариантов (например, при наблюдении случайной величины непрерывного типа с высокой точностью измерений) выборка может быть представлена в виде группированного статистического ряда. Для этого отрезок $[x_{(1)}; x_{(n)}]$, содержащий все элементы выборки, разбивается на k непересекающихся интервалов

$$J_1 = [\alpha_0 = x_{(1)}; \alpha_1),\ J_2 = [\alpha_1; \alpha_2),\ \ldots,\ J_k = [\alpha_{k-1}; \alpha_k = x_{(n)}],$$

как правило, одинаковой ширины h. Правые границы всех интервалов, за исключением последнего, задаются открытыми, чтобы исключить попадание граничных точек в соседний интервал.

Число интервалов k выбирается, как правило, в зависимости от объёма выборки. Для ориентировочной оценки числа k можно воспользоваться *формулой Стерджесса* (*Herbert Sturges*, 1926):

$$k \approx \left[1 + \log_2 n\right], \qquad (1.3)$$

где оператор $[\cdot]$ означает взятие целой части.

Например, при $n = 100$ оценка числа интервалов по формуле Стерджесса даёт $k = 7$, при $n = 1000$ получаем $k = 10$.

Ширина группировочных интервалов и число групп связаны формулой

$$h = \frac{x_{(n)} - x_{(1)}}{k}. \qquad (1.4)$$

Более теоретически обоснованный подход к выбору ширины группировочных интервалов дают *формула Скотта* (*David Scott*, 1979):

$$h \approx 3{,}5 s n^{-1/3}, \qquad (1.5)$$

и *формула Фридмана* (*David Freedman*, 1981):

$$h \approx 2\Delta n^{-1/3}, \qquad (1.6)$$

где s – среднеквадратичное отклонение выборки; Δ – интерквартильный размах выборки (см. § 3). Число группировочных интервалов k определяется из формулы (1.4).

В случае если распределение генеральной совокупности существенно отличается от нормального, число интервалов может быть увеличено. С уменьшением числа интервалов k происходит потеря статистической информации, содержащейся в исходной выборке.

Группированным статистическим рядом называется последовательность пар (J_i, n_i), $i = 1, ..., k$. Группированный статистический ряд записывается в виде таблицы, первая строка которой содержит интервалы J_i, а вторая – частоты n_i. Иногда в группированном статистическом ряду в первой строке таблицы вместо интервалов $J_1, \ldots, J_k$ записывают середины интервалов $c_1, \ldots, c_k$, где $c_i = \frac{\alpha_{i-1} + \alpha_i}{2}$ – середина i-го интервала, $i = 1, ..., k$.

Наряду с частотами n_i, $i = 1, ..., k$, попадания выборочных значений в группировочные интервалы рассматриваются также:

- относительные частоты n_i / n;
- накопленные (*cumulative*) частоты $m_i = \sum_{j=1}^{i} n_j$;
- относительные накопленные частоты m_i / n.

Полученные результаты сводятся в таблицу, называемую *таблицей частот группированной выборки* (табл. 1.2).

Таблица 1.2

Таблица частот группированной выборки

Номер интервала i	Границы интервала	Середина интервала c_i	Частота n_i	Накопленная частота m_i	Относительная частота n_i / n	Накопленная относительная частота m_i / n
1	$[\alpha_0; \alpha_1)$	c_1	n_1	m_1	n_1 / n	m_1 / n
...	...	...	...	...	...	...
k	$[\alpha_{k-1}; \alpha_k]$	c_k	n_k	$m_k = n$	n_k / n	$m_k / n = 1$

Визуально таблица частот может быть представлена с помощью гистограмм и полигонов частот. Выделяют четыре типа гистограмм (полигонов) частот:

1) гистограмма (полигон) абсолютных частот;
2) гистограмма (полигон) относительных частот;
3) гистограмма (полигон) накопленных частот;
4) гистограмма (полигон) относительных накопленных частот.

Гистограмма частот представляет собой кусочно-постоянную функцию, принимающую постоянные значения внутри интервалов группировки. В зависимости от типа гистограммы это значение может быть абсолютной, относительной, накопленной или относительной накопленной частотой.

Полигоны абсолютных и относительных частот строятся следующим образом: если построена гистограмма частот, то ординаты, соответствующие средним точкам интервалов, последовательно соединяются отрезками прямых.

Полигоны накопленных частот и относительных накопленных частот строятся так: если построена гистограмма частот, то ординаты, соответствующие правым точкам интервалов, последовательно соединяются отрезками прямых.

Пример 1.4. По результатам выборочного обследования 100 случайно отобранных из партии электрических лампочек проведена группировка наблюдений продолжительностей горения, и построена следующая таблица частот группированной выборки:

Номер интервала i	Границы интервала	Середина интервала c_i	Частота n_i	Накопленная частота m_i	Относительная частота n_i / n	Накопленная относительная частота m_i / n
1	[900; 920)	910	8	8	0,08	0,08
2	[920; 940)	930	15	23	0,15	0,23
3	[940; 960)	950	22	45	0,22	0,45
4	[960; 980)	970	36	81	0,36	0,81
5	[980; 1000)	990	12	93	0,12	0,93
6	[1000; 1020]	1010	7	100	0,07	1

Гистограммы и полигоны частот представлены на рис. 1.1.

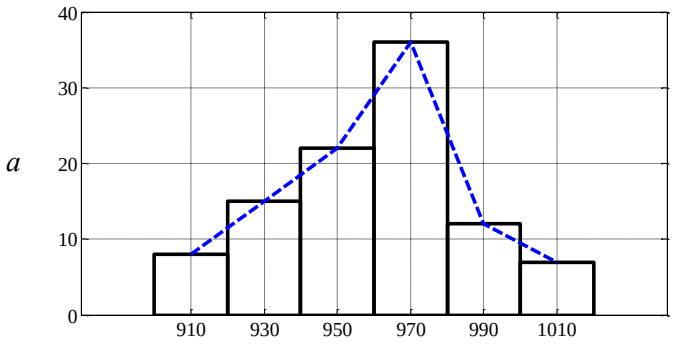

Рис. 1.1. Гистограммы и полигоны частот: *а* – абсолютных частот; *б* – относительных частот; *в* – накопленных частот; *г* – накопленных относительных частот (см. также с. 17)

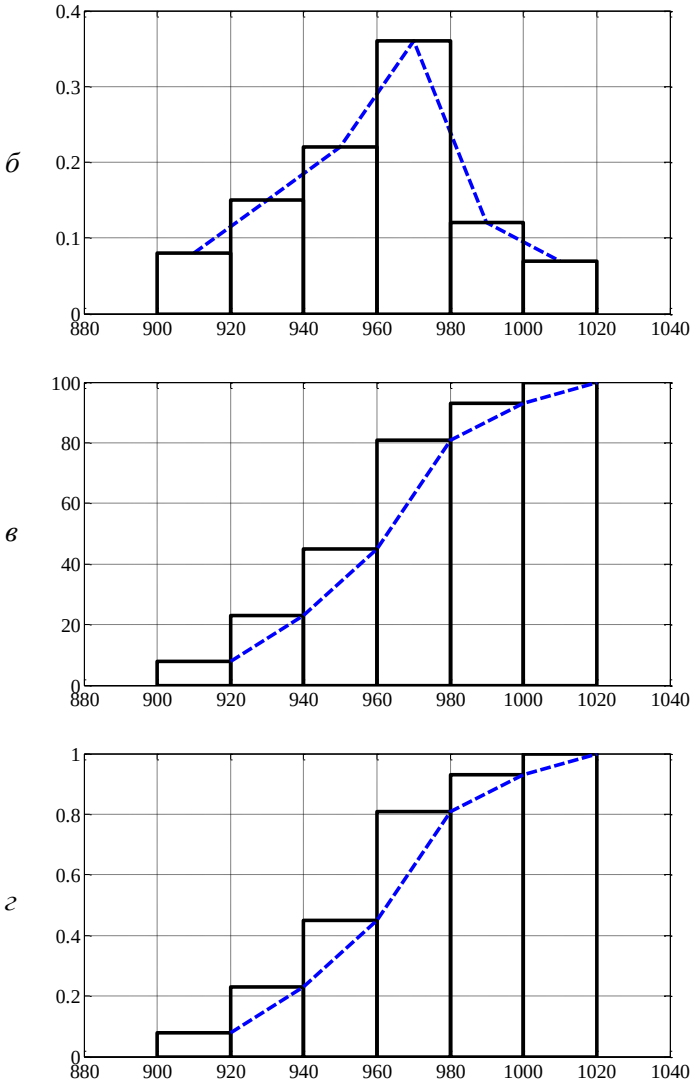

Рис. 1.1. Окончание

Контрольные вопросы и задачи

1. Дайте определение вариационного ряда случайной выборки и вариационного ряда выборки.
2. Что называют статистическим рядом выборки?
3. Докажите выражения (1.1) и (1.2).
4. В каких случаях целесообразно проводить группировку выборочных наблюдений?
5. Как выбираются число групп и ширина интервалов при группировке?
6. Какие негативные эффекты могут возникать при выборе слишком большого или слишком малого числа интервалов группировки?
7. Что называют группированным статистическим рядом выборки?
8. Чему равна сумма всех относительных частот в таблице частот группированной выборки?
9. Что показывает накопленная частота попадания выборочных значений в группировочный интервал?
10. Объясните принцип построения гистограмм и полигонов частот группированной выборки.

§ 3. Эмпирическая функция распределения. Числовые характеристики выборки

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей распределение $F_X(x)$. Пусть выборка содержит k вариантов $z_1, \ldots, z_k$, причём вариант z_i встречается с частотой n_i, $i = \overline{1,k}$.

Введём случайную величину дискретного типа X_n^*, принимающую значения $z_1, \ldots, z_k$ с вероятностями, равными соответствующим относительным частотам $\frac{n_1}{n}, \ldots, \frac{n_k}{n}$, т.е.

$$P(X_n^* = x_i) = \frac{n_i}{n}, \quad i = \overline{1,k}.$$

Относительные частоты принадлежат отрезку [0; 1], причём их сумма равна единице, т.е. для них выполнены все требования, предъявляемые к вероятности распределения. Распределение слу-

чайной величины X_n^* называется *распределением выборки* $x_1, \ldots, x_n$ (табл. 1.3).

Таблица 1.3

Распределение выборки

Варианты z_i	z_1	...	z_i	...	z_k
Вероятности p_i	n_1 / n	...	n_i / n	...	n_k / n

В связи с тем, что в выборке может присутствовать лишь конечное (или счётное) число вариантов наблюдаемой случайной величины X, распределение случайной величины X_n^* всегда является дискретным.

Функция распределения случайной величины X_n^* называется *эмпирической* (*выборочной*) *функцией распределения* (*ЭФР*) и обозначается $F_n^*(x)$:

$$F_n^*(x) = F_{X_n^*}(x) = P(X_n^* < x) = \sum_{z_i < x} \frac{n_i}{n} = \frac{1}{n} \sum_{z_i < x} n_i \,. \qquad (1.7)$$

Как известно, функция распределения случайной величины дискретного типа кусочно-постоянна. График ЭФР для выборки $x_1, \ldots, x_n$ с вариантами $z_1, \ldots, z_k$ приведён на рис. 1.2. Несложно показать, что ЭФР может принимать лишь значения, равные накопленным относительным частотам вариантов $z_1, \ldots, z_k$, или равняться нулю:

$$F_n^*(x) = \begin{cases} 0, & x \le z_1; \\ n_1 / n = m_1 / n, & z_1 < x \le z_2; \\ (n_1 + n_2) / n = m_2 / n, & z_2 < x \le z_3; \\ \ldots & \\ 1 = m_k / n, & x > z_k. \end{cases} \qquad (1.8)$$

В точках $z_1, \ldots, z_k$ ЭФР претерпевает разрыв и является, как и любая функция распределения, непрерывной слева.

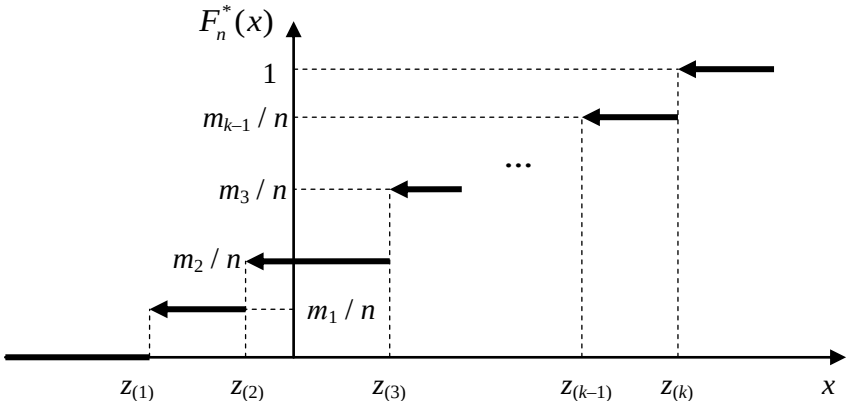

Рис. 1.2. Эмпирическая функция распределения

Поскольку ЭФР – это функция распределения дискретной случайной величины X_n^*, то для неё справедливы все свойства функции распределения дискретной случайной величины.

Эмпирическую функцию распределения $F_n^*(x)$ выборки $x_1, \ldots, x_n$ можно рассматривать как реализацию случайной эмпирической функции распределения $\mathcal{F}_n^*(x)$ соответствующей случайной выборки $X_1, \ldots, X_n$. При каждой конкретной реализации случайной выборки получаем соответствующую ей реализацию случайной ЭФР.

Пример 1.5. Для выборки из примера 1.3 построить эмпирическую функцию распределения.

По построенному статистическому ряду (см. решение примера 1.3) запишем распределение выборки:

z_i	1	2	3	5	6
p_i	0,1	0,3	0,2	0,3	0,1

ЭФР претерпевает разрыв в точках 1, 2, 3, 5, 6, а величина скачка равна соответствующей вероятности. Таким образом, ЭФР имеет вид

$$F^*(x)=\begin{cases}0, & x\le 1;\\ 0+0,1=0,1, & 1<x\le 2;\\ 0,1+0,3=0,4, & 2<x\le 3;\\ 0,4+0,2=0,6, & 3<x\le 5;\\ 0,6+0,3=0,9, & 5<x\le 6;\\ 0,9+0,1=1, & x>6.\end{cases}$$

Выборочными (*эмпирическими*) *числовыми характеристиками* называются числовые характеристики случайной величины X_n^*. К таким характеристикам относятся, например, моменты случайной величины. Напомним, что зная функцию плотности распределения $f_X(x)$ случайной величины *X* непрерывного типа (или распределение вероятностей $p_1, \ldots, p_k$ для случайной величины дискретного типа), математическое ожидание элементарной действительной функции $\xi(X)$ случайной величины *X* рассчитывается по формулам

$$\mathbf{M}[\xi(X)]=\int_{-\infty}^{\infty}\xi(x)\,f(x)dx \qquad (1.9)$$

и

$$\mathbf{M}[\xi(X)]=\sum_{i=1}^{k}\xi(x_i)\,p_i \qquad (1.10)$$

для непрерывного и дискретного случаев соответственно.

Учитывая (1.10), запишем выражение для расчёта *выборочного начального момента r-го порядка*. Все выборочные числовые характеристики будем обозначать с верхним знаком «звёздочка»:

$$\alpha_r^*=\mathbf{M}\left[(X_n^*)^r\right]=\sum_{i=1}^{k}z_i^r\,p_i=\sum_{i=1}^{k}z_i^r\,\frac{n_i}{n}=\frac{1}{n}\sum_{i=1}^{k}z_i^r n_i\,. \qquad (1.11)$$

В связи с тем, что каждый вариант z_i встречается в выборке $x_1, \ldots, x_n$ с соответствующей частотой n_i, $i=\overline{1,k}$, каждое произведение $z_i^r n_i$ может быть записано как сумма n_i одинаковых элементов

выборки, равных варианту z_i, возведённых в r-ю степень. Таким образом, выражение (1.11) примет вид:

$$\alpha_r^* = \frac{1}{n}\sum_{i=1}^{n} x_i^r \ . \qquad (1.12)$$

Выражение (1.11) называется *взвешенной* формой записи выборочного начального момента r-го порядка, а выражение (1.12) – *невзвешенной*.

Взвешенная форма записи выборочного начального момента r-го порядка представляет собой среднее арифметическое различных элементов (вариантов) выборки, возведённых в r-ю степень и взвешенных их частотами. Из невзвешенной формы записи видно, что выборочный начальный момент r-го порядка представляет собой простое среднее арифметическое элементов выборки, возведённых в r-ю степень. В связи с этим выборочный начальный момент r-го порядка также обозначается через $\overline{x^r}$.

Выборочный начальный момент первого порядка α_1^* называется *выборочным математическим ожиданием* и представляет собой простое среднее арифметическое элементов выборки, обозначаемое через $\overline{x}$:

$$m_X^* = \alpha_1^* = \mathbf{M}[X_n^*] = \frac{1}{n}\sum_{i=1}^{k} z_i n_i = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \ . \qquad (1.13)$$

Нижний индекс «X» в обозначении выборочного математического ожидания и других выборочных характеристик определяется случайной величиной, наблюдениями которой являются рассматриваемые выборочные значения $x_1, \ldots, x_n$.

Операция *центрирования* выборки состоит в смещении её значений на $\overline{x}$:

$$\dot{x}_i = x_i - \overline{x}, \quad i = \overline{1,n} \ . \qquad (1.14)$$

Несложно показать, что выборочное математическое ожидание (среднее) центрированной выборки $\dot{x}_1, ..., \dot{x}_n$ равно нулю:

$$\overline{\dot{x}} = \frac{1}{n}\sum_{i=1}^{n} \dot{x}_i = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x} = 0 \ . \qquad (1.15)$$

Учитывая определение центрального момента *r*-го порядка случайной величины дискретного типа, запишем выражение для расчёта *выборочного центрального момента r-го порядка*:

$$\mu_r^* = \mathbf{M}\left[(X_n^* - \overline{x})^r \right] = \sum_{i=1}^{k} (z_i - \overline{x})^r \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} (z_i - \overline{x})^r n_i \,. \qquad (1.16)$$

Выражение (1.16) представляет собой взвешенную форму записи. Невзвешенная форма получается из взвешенной заменой произведений $(z_i - \overline{x})^r n_i$, $i = \overline{1,k}$, на сумму n_i одинаковых слагаемых:

$$\mu_r^* = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^r \,. \qquad (1.17)$$

Выборочный центральный момент второго порядка μ_2^* называется *выборочной дисперсией*:

$$d_X^* = \mu_2^* = \frac{1}{n} \sum_{i=1}^{k} (z_i - \overline{x})^2 n_i = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \,. \qquad (1.18)$$

Выборочная дисперсия характеризует меру рассеяния значений $x_1, \ldots, x_n$ относительно их среднего арифметического $\overline{x}$.

Выборочное среднеквадратичное отклонение (*с.к.о.*) σ_X^* выборки $x_1, \ldots, x_n$ определяется как квадратный корень из выборочной дисперсии d_X^*:

$$\sigma_X^* = \sqrt{d_X^*} \,. \qquad (1.19)$$

Для выборочных начального и центрального моментов применимы все тождества, справедливые для начального и центрального моментов случайной величины дискретного типа. В частности, полезное на практике соотношение между выборочной дисперсией и выборочным начальным моментом второго порядка:

$$d_X^* = \overline{x^2} - \overline{x}^2 \,. \qquad (1.20)$$

Это равенство следует читать как «выборочная дисперсия равна разности между средним квадратом и квадратом среднего».

Выборочный коэффициент асимметрии γ_X^* (*skewness*) и *выборочный эксцесс* ε_X^* (*kurtosis*) – это коэффициент асимметрии и эксцесс случайной величины X_n^*:

$$\gamma_X^* = \frac{\mu_3^*}{(\sigma_X^*)^3}, \qquad (1.21)$$

$$\varepsilon_X^* = \frac{\mu_4^*}{(\sigma_X^*)^4} - 3. \qquad (1.22)$$

Выборочный коэффициент асимметрии характеризует степень асимметрии, а эксцесс – степень «плосковершинности» распределения выборки.

Пример 1.6. Для выборки из примера 1.3 рассчитать выборочные математическое ожидание и среднеквадратичное отклонение.

Варианты $z_1, \ldots, z_k$ данной выборки – числа 1, 2, 3, 5, 6, встречающиеся с частотами 1, 3, 2, 3, 1 соответственно. Число вариантов $k = 5$. Рассчитаем выборочные математическое ожидание и дисперсию на основе построенного статистического ряда выборки, используя взвешенные формы записи (1.13) и (1.18):

$$m_X^* = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} z_i n_i = \frac{1}{10}(1\cdot 1 + 2\cdot 3 + 3\cdot 2 + 5\cdot 3 + 6\cdot 1) = 3,4;$$

$$d_X^* = \frac{1}{n}\sum_{i=1}^{k} (z_i - \bar{x})^2 n_i = \frac{1}{10}\Big(2,4^2\cdot 1 + 1,4^2\cdot 3 + 0,4^2\cdot 2 + 1,6^2\cdot 3 + 2,6^2\cdot 1\Big) = 2,64;$$

$$\sigma_X^* = \sqrt{d_X^*} = \sqrt{2,64} \approx 1,63.$$

Выборочные характеристики также могут быть рассчитаны для группированной выборки. Пусть проведена группировка выборочных данных $x_1, \ldots, x_n$ на k интервалов $[\alpha_0; \alpha_1)$, $[\alpha_1; \alpha_2), \ldots, [\alpha_{k-1}; \alpha_k]$; n_i – частота попадания выборочных значений в i-й интервал; $c_i = \dfrac{\alpha_{i-1} + \alpha_i}{2}$ – середина i-го интервала, $i = \overline{1,k}$.

При расчёте выборочных характеристик группированной выборки предполагается, что все элементы выборки, попавшие в i-й интервал, находятся в его середине. Таким образом, выборочный начальный момент r-го порядка рассчитывается как среднее ариф-

метическое взвешенное середин интервалов, возведённых в *r*-ю степень, а выборочный центральный момент *r*-го порядка – как среднее арифметическое взвешенное центрированных середин интервалов, возведённых в *r*-ю степень. В обоих случаях взвешивание проводится частотами попадания выборочных значений в интервалы:

$$\alpha_r^* = \frac{1}{n}\sum_{i=1}^{k} c_i^r n_i \ . \qquad (1.23)$$

$$\mu_r^* = \frac{1}{n}\sum_{i=1}^{k} (c_i - \overline{x})^r n_i \ . \qquad (1.24)$$

Пример 1.7. Для выборки из примера 1.4 рассчитать выборочные математическое ожидание и среднеквадратичное отклонение.

Запишем группированный статистический ряд выборки:

Середины интервалов c_i	910	930	950	970	990	1010
Частоты n_i	8	15	22	36	12	7

Для расчёта выборочных математического ожидания и дисперсии используем формулы (1.23) при $r = 1$ и (1.24) при $r = 2$ соответственно:

$$m_X^* = \overline{x} = \frac{1}{n}\sum_{i=1}^{k} c_i n_i = \frac{910\cdot 8 + 930\cdot 15 + ... + 1010\cdot 7}{100} = 960;$$

$$d_X^* = \frac{1}{n}\sum_{i=1}^{k} (c_i - \overline{x})^2 n_i = \frac{50^2\cdot 8 + 30^2\cdot 15 + ... + 50^2\cdot 7}{100} = 676;$$

$$\sigma_X^* = \sqrt{d_X^*} = \sqrt{676} = 26 .$$

Выборочной квантилью на уровне вероятности p (или *порядка p*) выборки $x_1, \ldots, x_n$ называется квантиль случайной величины X_n^* на уровне вероятности p. Напомним, квантилью случайной величины X на уровне вероятности p называется точная верхняя граница x_p множества значений x, для которых выполнено условие:

$$F_X(x) = P(X < x) = p\,.$$

Для дискретной случайной величины, в частности для случайной величины X_n^*, точная верхняя граница этого множества не может быть определена однозначно. В связи с этим при расчёте выборочной квантили x_p^* на практике используются следующие правила.

1. Значение i-го элемента вариационного ряда $x_{(i)}$ является выборочной квантилью на уровне $p_i = \dfrac{i-0{,}5}{n}$. Таким образом, соответствие между элементами вариационного ряда и уровнями квантилей устанавливается таблицей:

Выборочная квантиль x_p^*	$x_{(1)}$	…	$x_{(i)}$	…	$x_{(n)}$
Уровень p	$\dfrac{0{,}5}{n}$	…	$\dfrac{i-0{,}5}{n}$	…	$\dfrac{n-0{,}5}{n}$

2. Для расчёта квантили произвольного уровня p, $0 \le p \le 1$, используется линейная интерполяция значений, приведённых в таблице выше.

Выборочной медианой выборки x_1, …, x_n называется выборочная квантиль $x_{0{,}5}^*$ на уровне $p = 0{,}5$. Из правил расчёта выборочных квантилей следуют правила расчёта выборочной медианы.

1. Если объём выборки n – нечётный, то, решая уравнение $\dfrac{i-0{,}5}{n} = 0{,}5$ относительно i, получаем номер $i = \dfrac{n+1}{2}$ элемента вариационного ряда, являющегося медианой, т.е.

$$x_{0{,}5}^* = x_{((n+1)/2)}\,. \qquad (1.25)$$

2. Если объём выборки n – чётный, то выборочная медиана определяется путём линейной интерполяции элементов вариационного ряда с номерами $\dfrac{n}{2}$ и $\dfrac{n}{2}+1$, являющихся квантилями на уров-

нях $0{,}5-\frac{1}{2n}$ и $0{,}5+\frac{1}{2n}$ соответственно. Результатом этой интерполяции будет среднее значение

$$x^*_{0,5}=\frac{x_{(n/2)}+x_{(n/2+1)}}{2}. \qquad (1.26)$$

Выборочные квантили $x^*_{0,25}$ и $x^*_{0,75}$ на уровнях 0,25 и 0,75 называются *выборочными нижней* и *верхней квартилями* соответственно. Разность Δ между верхней и нижней квартилями называется *интерквартильным интервалом*:

$$\Delta=x^*_{0,75}-x^*_{0,25}. \qquad (1.27)$$

Интерквартильный интервал характеризует разброс выборочных значений и, в некотором смысле, представляет собой аналог среднеквадратичного отклонения.

Выборочные квантили $x^*_{0,1}$, …, $x^*_{0,9}$ на уровнях, кратных 0,1, называются *выборочными децилями*, а выборочные квантили $x^*_{0,01}$, …, $x^*_{0,99}$ на уровнях, кратных 0,01, – *выборочными процентилями*.

Выборочной модой выборки x_1, …, x_n с вариантами z_1, …, z_k называется вариант z_i, $i\in\{1,...,k\}$, частота которого максимальна.

Пример 1.8. Для выборки из примера 1.3 рассчитать выборочные медиану, верхнюю и нижнюю квартили, интерквартильный интервал и квантиль на уровне 0,88.

Так как объём выборки $n = 10$ – чётный, то медиану рассчитываем по формуле (1.26):

$$x^*_{0,5}=\frac{x_{(5)}+x_{(6)}}{2}=\frac{3+3}{2}=3.$$

Элементы вариационного ряда 1; 2; **2**; 2; 3; 3; 5; **5**; 5; 6 являются выборочными квантилями на уровнях 0,05; 0,15; …; 0,95 соответственно. Таким образом, выборочные нижняя и верхняя квартили равны $x^*_{0,25}=2$ и $x^*_{0,75}=5$ (выделены в вариационном ряду). Интерквартильный интервал $\Delta=5-2=3$.

Для расчёта $x^*_{0,88}$ используем интерполяционную формулу. Ближайшие к 0,88 уровни, которым соответствуют квантили из вариационного ряда, – 0,85 и 0,95. Эти квантили – 5 и 6. Абсциссу точки, лежащей на прямой, проходящей через две точки (5; 0,85) и (6; 0,95), с ординатой 0,88 находим из пропорции:

$$\frac{x^*_{0,88}-5}{6-5}=\frac{0,88-0,85}{0,95-0,85},$$

получаем искомую квантиль

$$x^*_{0,88}=5,3.$$

Контрольные вопросы и задачи

1. Что называется распределением выборки?
2. Дайте определение эмпирической функции распределения.
3. Перечислите свойства эмпирической функции распределения.
4. Что называется выборочными числовыми характеристиками?
5. Что называется выборочным начальным моментом *r*-го порядка? В каких единицах измеряется выборочный начальный момент α^*_r?
6. Что называется выборочным центральным моментом *r*-го порядка? В каких единицах измеряется выборочный центральный момент μ^*_r?
7. Чему равно значение α^*_0?
8. Чему равно значение μ^*_1?
9. Как называется выборочный начальный момент первого порядка и что он характеризует?
10. Как называется выборочный центральный момент второго порядка и что он характеризует?
11. Чему равны размерности выборочного математического ожидания, выборочной дисперсии, выборочного среднеквадратичного отклонения?

12. В чём состоит операция центрирования выборки? Чему равны выборочные математическое ожидание и дисперсия центрированной выборки?

13. Докажите тождество (1.20), определяющее связь между выборочной дисперсией и выборочным начальным моментом второго порядка.

14. Покажите, что для любой выборки $x_1, \ldots, x_n$ и любого $a \in \mathbb{R}$ справедливо тождество: $\sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - a)^2$.

15. Как изменятся выборочные математическое ожидание, дисперсия и среднеквадратичное отклонение, если к каждому элементу выборки прибавить константу $x_0 \in \mathbb{R}$?

16. Как изменятся выборочные математическое ожидание, дисперсия и среднеквадратичное отклонение, если каждый элемент выборки умножить на константу $a \in \mathbb{R}$?

17. Дайте определение выборочного коэффициента асимметрии. Что показывает этот коэффициент?

18. Дайте определение выборочного эксцесса. Что показывает этот коэффициент?

19. Докажите, что центрирование выборки не изменяет значение выборочного коэффициента асимметрии.

20. Докажите, что центрирование выборки не изменяет значение выборочного эксцесса.

21. Что называется выборочной квантилью на уровне вероятности p? Что показывает эта характеристика?

22. Приведите правила расчёта выборочной квантили на уровне вероятности p.

23. Дайте определение понятий: выборочная медиана, верхняя и нижняя квартили, дециль, процентиль.

24. Что называется интерквартильным интервалом выборки? Что показывает эта характеристика?

25. Рассчитайте выборочные медиану и квартили для выборки: 4,2; 3,7; 1,1; 2,9; 4,4; 3,5; 3,2; 1,5.

26. Рассчитайте 35-ю и 65-ю процентили для выборки: 5; 3; 8; 2; 6; 1; 3; 9; 3; 9.

§ 4. Диаграмма Box-and-Whisker

Диаграмма *Box-and-Whisker* является удобной на практике формой визуального представления выборочного распределения, использующей выборочные медиану и квартили. Эта диаграмма позволяет также определить так называемые *выбросные значения выборки* (*outliers*), т.е. значения, существенно отличающиеся от большинства остальных значений выборки (например, возникшие в результате сбоя регистрирующей аппаратуры).

Схема диаграммы *Box-and-Whisker* приведена на рис. 1.3.

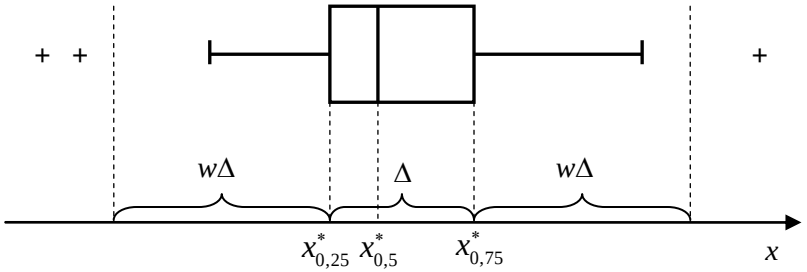

Рис. 1.3. Схема диаграммы *Box-and-Whisker*

Принцип построения диаграммы состоит в следующем.

1. Границы «коробочки» откладываются на уровнях нижней $x^*_{0,25}$ и верхней $x^*_{0,75}$ выборочных квартилей.

2. Вертикальная линия внутри «коробочки» проводится на уровне медианы $x^*_{0,5}$.

3. От правой границы «коробочки» вправо и от левой границы влево рисуются «усики», длина каждого из которых не превосходит $w\Delta$, где w – некоторая константа (обычно, $w = 1{,}5$); Δ – интерквартильный интервал. Левый «усик» заканчивается в точке, которой соответствует выборочное наблюдение, ближайшее к $x^*_{0,25} - w\Delta$ справа, а правый – в выборочной точке, ближайшей к $x^*_{0,75} + w\Delta$ слева.

4. Выборочные точки, не попавшие в отрезок $[x^*_{0,25} - w\Delta; x^*_{0,75} + w\Delta]$, обозначаются знаком «+» и считаются выбросными наблюдениями.

С помощью диаграмм *Box-and-Whisker* можно оценить степень разброса и асимметрию выборочных наблюдений, а также провести визуальное сравнение нескольких распределений, не требующее сложных вычислений. Границы «усиков» показывают минимальное и максимальное значения выборки, исключая выбросные наблюдения.

Контрольные вопросы и задачи

1. Объясните принцип построения диаграммы *Box-and-Whisker*.
2. Что называется выбросными значениями выборки?
3. Что можно сказать о выборочном распределении, если правая часть «коробочки» на диаграмме много длиннее, чем левая?
4. Что можно сказать о выборочном распределении, если правый «усик» диаграммы много длиннее, чем левый?

§ 5. Выборочные характеристики двумерного случайного вектора

Выборочные характеристики можно ввести и для выборок из многомерных генеральных совокупностей. Пусть $(x_1, y_1), \ldots, (x_n, y_n)$ – выборка наблюдений двумерного случайного вектора (X, Y), имеющего распределение $F_{XY}(x, y)$. Пусть выборка содержит k различных пар наблюдений (вариантов) $z_1, \ldots, z_k$, $z_i = (x_i, y_i)$, причём вариант z_i встречается с частотой n_i, $i = \overline{1,k}$.

По аналогии с одномерным случаем введём случайный вектор дискретного типа (X^*_n, Y^*_n), принимающий варианты $z_1, \ldots, z_k$ с вероятностями, равными соответствующим относительным частотам $\frac{n_1}{n}, \ldots, \frac{n_k}{n}$, т.е. $P\left((X^*_n = x_i) \bigcap (Y^*_n = y_i)\right) = \frac{n_i}{n}$, $i = \overline{1,k}$.

Распределение случайного вектора (X^*_n, Y^*_n) называется *распределением двумерной выборки* $(x_1, y_1), \ldots, (x_n, y_n)$. Предварительное

представление о распределении выборки можно получить, изображая элементы выборки точками на плоскости координат *xOy* (рис. 1.4). Это представление выборки называется *диаграммой рассеяния* (*scatter plot*).

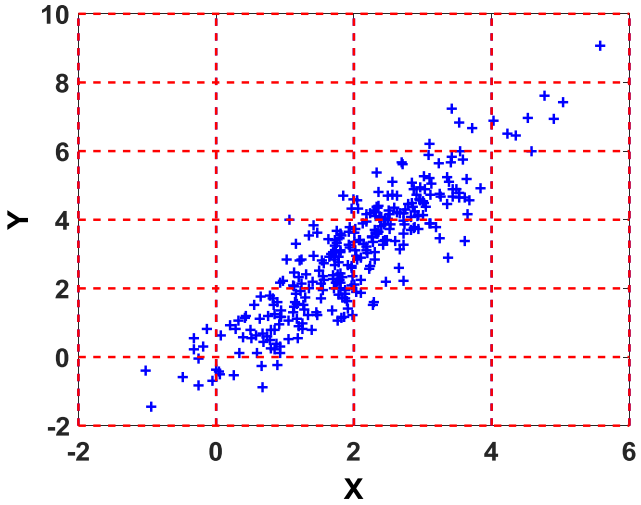

Рис. 1.4. Пример диаграммы рассеяния двумерной выборки

Выборочными числовыми характеристиками двумерной выборки $(x_1, y_1), \ldots, (x_n, y_n)$ называются числовые характеристики случайного вектора (X_n^*, Y_n^*). К таким характеристикам относятся, например, моменты случайного вектора.

Выборочный смешанный начальный момент порядка $(q+r)$ равен:

$$\alpha_{q,r}^* = \mathbf{M}\left[(X_n^*)^q (Y_n^*)^r\right] = \sum_{i=1}^{k} x_i^q y_i^r p_i \,, \qquad (1.28)$$

где $p_i = P\left((X_n^* = x_i) \bigcap (Y_n^* = y_i)\right)$, а суммирование проводится по всем вариантам случайного вектора (X_n^*, Y_n^*).

Учитывая, что случайный вектор (X_n^*, Y_n^*) принимает вариант (x_i, y_i) с вероятностью, равной относительной частоте $\frac{n_i}{n}$ этого наблюдения в выборке, и представляя произведения $x_i^q y_i^r n_i$ как суммы n_i одинаковых слагаемых $x_i^q y_i^r$, $i = \overline{1,k}$, формула (1.28) может быть записана в невзвешенном виде:

$$\alpha_{q,r}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^q y_i^r \ . \qquad (1.29)$$

Аналогично, *выборочный смешанный центральный момент* порядка $(q+r)$ определяется формулой

$$\mu_{q,r}^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^q (y_i - \bar{y})^r \ . \qquad (1.30)$$

Наиболее часто используемая числовая характеристика двумерного вектора – коэффициент корреляции. Напомним, что для случайного вектора дискретного типа (X, Y) коэффициент корреляции r_{XY} равен:

$$r_{XY} = \frac{k_{XY}}{\sigma_X \sigma_Y} \ , \qquad (1.31)$$

где k_{XY} – ковариационный момент, по определению $k_{XY} = \mu_{1,1}^{(X,Y)}$.

Учитывая (1.31), определим выражение для выборочного коэффициента корреляции ρ_{XY}^*:

$$\rho_{XY}^* = \frac{k_{XY}^*}{\sigma_X^* \sigma_Y^*} \ , \qquad (1.32)$$

где k_{XY}^* – выборочный ковариационный момент:

$$k_{XY}^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \ . \qquad (1.33)$$

Для выборочных ковариационного момента и коэффициента корреляции применимы все тождества, справедливые для ковариационного момента и коэффициента корреляции случайного вектора дискретного типа. В частности, полезное на практике соотношение между выборочным ковариационным моментом и выборочным смешанным начальным моментом второго порядка:

$$k^*_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y} = \overline{xy} - \overline{x} \cdot \overline{y} \,. \qquad (1.34)$$

Это равенство следует читать как «выборочный ковариационный момент равен разности между средним произведением и произведением средних».

Пример 1.9. У восьми учащихся колледжа зафиксировано следующее количество баллов, полученных за самостоятельные работы по математике (x) и по гуманитарным предметам (y):

Студент п/п	x	y
1	90	75
2	60	69
3	46	45
4	68	49
5	82	58
6	71	54
7	66	59
8	78	70

Рассчитать выборочный коэффициент корреляции между x и y.

Для расчёта коэффициента корреляции заполним таблицу:

Студент п/п	x	y	xy	x^2	y^2
1	90	75	6750	8100	5625
2	60	69	4140	3600	4761
3	46	45	2070	2116	2025
4	68	49	3332	4624	2401
5	82	58	4756	6724	3364
6	71	54	3834	5041	2916
7	66	59	3894	4356	3481
8	78	70	5460	6084	4900
Σ	561	479	34236	40645	29473
Σ / n	70	60	4280	5081	3684

Рассчитаем выборочные с.к.о.:

$$\sigma_X^* = \sqrt{\overline{x^2} - \overline{x}^2} \approx \sqrt{5081 - 70^2} \approx 12{,}8;$$

$$\sigma_Y^* = \sqrt{\overline{y^2} - \overline{y}^2} \approx \sqrt{3684 - 60^2} \approx 9{,}9;$$

и выборочные ковариационный момент и коэффициент корреляции:

$$k_{XY}^* = \overline{xy} - \overline{x} \cdot \overline{y} \approx 4280 - 70 \cdot 60 \approx 80;$$

$$\rho_{XY}^* = \frac{80}{12{,}8 \cdot 9{,}9} \approx 0{,}63.$$

Двумерная выборка может быть представлена в виде *корреляционной таблицы* – аналога группированного статистического ряда для одномерной выборки (табл. 1.4).

Таблица 1.4

Корреляционная таблица

Интервал по X	Интервал по Y				
	$[\beta_0; \beta_1)$	…	$[\beta_{j-1}; \beta_j)$	…	$[\beta_{m-1}; \beta_m]$
$[\alpha_0; \alpha_1)$	n_{11}	…	n_{1j}	…	n_{1m}
…	…	…	…	…	…
$[\alpha_{i-1}; \alpha_i)$	n_{i1}	…	n_{ij}	…	n_{im}
…	…	…	…	…	…
$[\alpha_{l-1}; \alpha_l]$	n_{l1}	…	n_{lj}	…	n_{lm}

Для построения корреляционной таблицы отрезок $[x_{(1)}; x_{(n)}]$, содержащий все наблюдения случайной величины X, разбивается на l непересекающихся интервалов

$$[\alpha_0 = x_{(1)}; \alpha_1), [\alpha_1; \alpha_2), \ldots, [\alpha_{l-1}; \alpha_l = x_{(n)}],$$

как правило, одинаковой ширины h_X. Аналогично отрезок $[y_{(1)}; y_{(n)}]$, содержащий все наблюдения случайной величины Y, разбивается на m непересекающихся интервалов

$$[\beta_0 = y_{(1)}; \beta_1), [\beta_1; \beta_2), \ldots, [\beta_{m-1}; \beta_m = y_{(n)}],$$

как правило, одинаковой ширины h_Y. Правые границы всех интервалов, за исключением последнего, задаются открытыми, чтобы исключить попадание граничных точек в соседний интервал.

Процедуру группировки двумерных выборочных наблюдений можно выполнить непосредственно по диаграмме рассеяния (см. рис. 1.4), нанеся на неё сетку горизонтальных и вертикальных прямых, взятых с постоянными шагами h_X и h_Y и рассчитав частоты n_{ij} попадания выборочных точек в каждую ячейку сетки, $i=\overline{1,l}$, $j=\overline{1,m}$. Очевидно, что сумма всех частот в корреляционной таблице равна объёму выборки, $\sum_{i=1}^{l}\sum_{j=1}^{m} n_{ij} = n$.

Контрольные вопросы и задачи

1. Что называется распределением двумерной выборки?
2. Что называется выборочными числовыми характеристиками двумерной выборки?
3. Что называется выборочным смешанным начальным моментом $(q+r)$-го порядка?
4. Что называется выборочным смешанным центральным моментом $(q+r)$-го порядка?
5. Что называется выборочным ковариационным моментом двумерной выборки $(x_1, y_1), \ldots, (x_n, y_n)$?
6. Докажите тождество (1.34), определяющее связь между выборочным ковариационным моментом и средним произведением выборочных данных.
7. В каких единицах измеряется выборочный ковариационный момент?
8. Что называется выборочным коэффициентом корреляции?
9. В каких единицах измеряется выборочный коэффициент корреляции?
10. Как изменятся выборочные ковариационный момент и коэффициент корреляции, если к каждому элементу выборки прибавить вектор констант $(x_0, y_0) \in \mathbb{R}^2$? Если каждый элемент выборки умножить на константу $a \in \mathbb{R}$?

Глава 2. ТОЧЕЧНЫЕ ОЦЕНКИ

§ 6. Свойства точечных оценок

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей распределение $F_X(x)$. При проведении ряда статистических исследований вид функции распределения наблюдаемой случайной величины зачастую предполагается известным (например, случайная величина имеет нормальное или биномиальное распределение). Неизвестными же являются параметры этого распределения.

Одной из задач математической статистики является оценка неизвестных параметров распределения наблюдаемой случайной величины X по выборке $x_1, \ldots, x_n$ её наблюдений.

Параметром $\theta \in \Theta$ распределения $F_X(x)$ случайной величины X называется любая числовая характеристика этой случайной величины (математическое ожидание, дисперсия и т.д.) или любая константа, явно входящая в выражение для функции распределения $F_X(x)$.

В общем случае будем считать, что распределение $F_X(x)$ характеризуется вектором параметров $\theta = (\theta_1, ..., \theta_k)$ размерности k.

Например, пусть масса деталей, изготавливаемых станком, в силу присутствия неточности работы станка является случайной величиной X, имеющей нормальное распределение, но его параметры $\theta_1 = m_X$ и $\theta_2 = \sigma_X$ неизвестны. Требуется найти приближённое значение этих параметров по выборке наблюдений $x_1, \ldots, x_n$ масс n изготовленных станком деталей.

Напомним, что любая выборка наблюдений $x_1, \ldots, x_n$ является реализацией случайной выборки $X_1, \ldots, X_n$ (см. § 1). *Статистикой* Z в математической статистике называется произвольная функция случайной выборки, не зависящая от неизвестных параметров распределения:

$$Z = \varphi(X_1, ..., X_n). \qquad (2.1)$$

В связи с тем, что статистика Z – функция случайных аргументов, то Z также является случайной величиной. Для каждой реализации $x_1, \ldots, x_n$ случайной выборки $X_1, \ldots, X_n$ получим соответствующую ей реализацию z статистики Z:

$$z = \varphi(x_1, ..., x_n), \tag{2.2}$$

называемую *выборочным значением статистики Z.*

Точечной оценкой $\tilde{\theta}_n$ неизвестного параметра $\theta \in \Theta$ (или вектора параметров) распределения $F_X(x)$ называется произвольная статистика $\tilde{\theta}_n$, построенная по случайной выборке $X_1, \ldots, X_n$ из генеральной совокупности X и принимающая значения из множества Θ:

$$\tilde{\theta}_n = \tilde{\theta}(X_1, ..., X_n). \tag{2.3}$$

Точечная оценка $\tilde{\theta}_n$ – случайная величина. Для выборки $x_1, \ldots, x_n$ может быть рассчитана *реализация точечной оценки*, или *выборочное значение точечной оценки*, неизвестного параметра $\theta \in \Theta$. Далее точечную оценку и её выборочное значение будем обозначать одинаково через $\tilde{\theta}_n$, при необходимости дополнительно оговаривая, является ли $\tilde{\theta}_n$ случайной величиной или её реализацией.

В соответствии с определением (2.3) существует бесконечно много точечных оценок неизвестного параметра θ. Формально точечная оценка $\tilde{\theta}_n$ может не иметь ничего общего с интересующим нас параметром θ. Её полезность для получения практически приемлемых оценок вытекает из статистических свойств, которыми она обладает.

Ниже рассмотрены три основных свойства точечных оценок.

1. *Состоятельность* (*Consistency*). Точечная оценка $\tilde{\theta}_n = \tilde{\theta}(X_1, ..., X_n)$ называется *состоятельной* оценкой параметра $\theta \in \Theta$, если последовательность случайных величин $\tilde{\theta}_1, \tilde{\theta}_2, ..., \tilde{\theta}_n, ...$ сходится по вероятности к оцениваемому параметру θ при $n \to \infty$, т.е.

$$\forall \varepsilon > 0 \quad P\left(\left|\tilde{\theta}_n - \theta\right| < \varepsilon\right) \to 1. \tag{2.4}$$

Иными словами, для состоятельной оценки вероятность её отклонения от оцениваемого параметра θ на любую малую величину ε при увеличении объёма выборки стремится к нулю. Это свойство оценки является очень важным, ибо несостоятельная оценка практически бесполезна. Для несостоятельной оценки её значение, рассчитанное даже для выборки очень большого объёма, может существенно отличаться от значения параметра θ, а увеличение объёма выборки не улучшает её качество.

Состоятельность оценки может быть проверена, используя *достаточное условие состоятельности*: если $\mathbf{M}[\tilde{\theta}_n] \to \theta$ и $\mathbf{D}[\tilde{\theta}_n] \to 0$ при $n \to \infty$, то оценка $\tilde{\theta}_n$ состоятельна.

Доказательство этого утверждения следует из второго неравенства Чебышева, согласно которому

$$\forall \varepsilon > 0 \quad P\left(\left|\tilde{\theta}_n - \mathbf{M}[\tilde{\theta}_n]\right| \geq \varepsilon\right) \leq \frac{\mathbf{D}[\tilde{\theta}_n]}{\varepsilon^2}.$$

Переходя к пределу при $n \to \infty$ получаем

$$\forall \varepsilon > 0 \quad P\left(\left|\tilde{\theta}_n - \theta\right| \geq \varepsilon\right) \to 0,$$

из чего следует состоятельность оценки $\tilde{\theta}_n$.

2. *Несмещённость* (*Unbiasedness*). Точечная оценка $\tilde{\theta}_n = \tilde{\theta}(X_1, ..., X_n)$ называется *несмещённой* оценкой параметра $\theta \in \Theta$, если её математическое ожидание равно оцениваемому параметру θ, т.е.

$$\mathbf{M}[\tilde{\theta}_n] = \theta. \qquad (2.5)$$

Разность $b_n(\theta) = \mathbf{M}[\tilde{\theta}_n] - \theta$ называется *смещением* (*bias*) точечной оценки $\tilde{\theta}_n$.

Статистика $\tilde{\theta}$ называется несмещённой оценкой параметра $\theta \in \Theta$, если условие (2.5) выполнено для любого фиксированного объёма выборки *n*.

Статистика $\tilde{\theta}$ называется *асимптотически несмещённой оценкой* параметра $\theta \in \Theta$, если числовая последовательность математи-

ческих ожиданий $\mathbf{M}[\tilde{\theta}_1], \mathbf{M}[\tilde{\theta}_2], ..., \mathbf{M}[\tilde{\theta}_n], ...$ сходится к оцениваемому параметру θ при $n \to \infty$, т.е.

$$\mathbf{M}[\tilde{\theta}_n] \to \theta. \qquad (2.6)$$

Несмещённость оценки $\tilde{\theta}_n$ означает, что реализации этой оценки, рассчитанные для различных реализаций случайной выборки $X_1, \ldots, X_n$ объёма n, будут группироваться в среднем около оцениваемого параметра θ. Реализации смещённой оценки $\hat{\theta}_n$ группируются около величины $\theta + b_n(\theta)$, где $b_n(\theta)$ – смещение оценки (рис. 2.1).

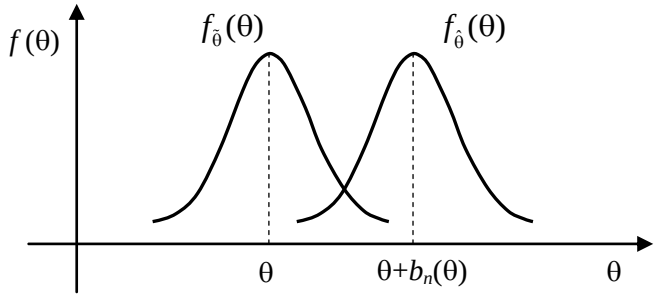

Рис. 2.1. Иллюстрация понятия несмещённости точечной оценки

3. *Эффективность* (*Efficiency*). Для оценивания параметра θ может быть предложено множество несмещённых оценок. Вследствие несмещённости различные реализации этих оценок будут группироваться относительно их математического ожидания, равного θ, однако разброс этих значений может быть различным. Как известно, мерой разброса значений случайной величины относительно математического ожидания является её дисперсия.

Пусть $\tilde{\theta}_n = \tilde{\theta}(X_1, ..., X_n)$ и $\hat{\theta}_n = \hat{\theta}(X_1, ..., X_n)$ – две несмещённые оценки параметра θ по выборке объёма n. Оценка $\tilde{\theta}_n$ называется *более эффективной*, чем оценка $\hat{\theta}_n$, если её дисперсия меньше, т.е.

$$\mathbf{D}[\tilde{\theta}_n] < \mathbf{D}[\hat{\theta}_n]. \qquad (2.7)$$

Статистика $\tilde{\theta}$ называется более эффективной оценкой параметра $\theta \in \Theta$, чем статистика $\hat{\theta}$, если условие (2.7) выполнено для любого фиксированного объёма выборки *n*.

Если оценка $\tilde{\theta}_n$ более эффективна, чем оценка $\hat{\theta}_n$, то это означает, что реализации оценки $\tilde{\theta}_n$, рассчитанные для различных реализаций случайной выборки $X_1, \ldots, X_n$ объёма *n*, будут иметь меньший разброс около оцениваемого параметра θ, чем реализации менее эффективной оценки $\hat{\theta}_n$ (рис. 2.2).

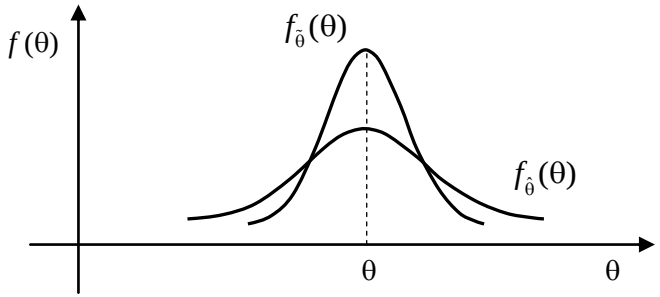

Рис. 2.2. Иллюстрация понятия эффективности точечной оценки

Оценка параметра θ, имеющая минимально возможную дисперсию среди всех оценок, называется *эффективной оценкой* параметра θ. В математической статистике наряду с термином «эффективная оценка» используют и другие: «несмещённая оценка с минимальной дисперсией», «оптимальная оценка».

Для того чтобы ответить на вопрос, является ли статистика $\tilde{\theta}$ эффективной оценкой параметра θ, используется *неравенство Рао–Крамера* (*Calyampudi Radhakrishna Rao*, *Harald Cramer*, 1945):

$$\mathbf{D}[\tilde{\theta}] \geq \frac{1}{I_n(\theta)}, \qquad (2.8)$$

согласно которому любая оценка $\tilde{\theta}$ параметра θ ограничена снизу величиной $\frac{1}{I_n(\theta)}$ при выполнении некоторых условий регулярно-

сти (выполнены практически для всех используемых на практике оценок), где $I_n(\theta)$ – количество информации по Фишеру о параметре θ, содержащейся в выборке объёма *n*.

Таким образом, *критерием эффективности* оценки $\tilde{\theta}$ является обращение для неё в равенство неравенства Рао–Крамера.

Эффективностью оценки $\tilde{\theta}$ параметра θ называется величина

$$e\left(\tilde{\theta}\right) = \frac{I_n^{-1}(\theta)}{\mathbf{D}[\tilde{\theta}]}. \qquad (2.9)$$

Согласно неравенству Рао–Крамера эффективность любой точечной оценки ограничена сверху единицей, а для эффективных оценок $e\left(\tilde{\theta}\right) = 1$.

При выполнении условий регулярности каждый элемент независимой случайной выборки $X_1, \ldots, X_n$ вносит равный вклад в информацию по Фишеру $I_n(\theta)$, т.е.

$$I_n(\theta) = nI(\theta), \qquad (2.10)$$

где $I(\theta)$ – количество информации по Фишеру о параметре θ, содержащейся в одном выборочном наблюдении.

Величина информации по Фишеру зависит от вида распределения генеральной совокупности *X*. Так, выборки, полученные из генеральных совокупностей с разными распределениями (например, нормальным и биномиальным) будут содержать различное количество информации о неизвестных математическом ожидании или дисперсии.

Чем больше информации по Фишеру о параметре θ содержится в выборочных наблюдениях, тем меньший разброс имеют реализации эффективной оценки этого параметра, а следовательно, являются более точными.

Формально информация по Фишеру о параметре θ, содержащаяся в одном выборочном наблюдении из генеральной совокупности с функцией плотности распределения $f_X(x, \theta)$, рассчитывается по формуле

$$I(\theta) = \mathbf{M}\left[U^2(X, \theta)\right], \qquad (2.11)$$

где функция

$$U(x,\theta)=\frac{\partial}{\partial\theta}\ln f_X(x,\theta) \tag{2.12}$$

называется *вкладом выборки*.

В случае дискретной генеральной совокупности с распределением вероятностей $P(x, \theta)$, $\sum_x P(x,\theta)=1$, вклад выборки определяется как

$$U(x,\theta)=\frac{\partial}{\partial\theta}\ln P(x,\theta)\,. \tag{2.13}$$

Отметим, что в выражении (2.11) одним из аргументов функции вклада является случайная величина X.

Статистика $\tilde{\theta}$ называется *асимптотически эффективной* оценкой параметра θ, если последовательность дисперсий $\mathbf{D}[\tilde{\theta}_1], \mathbf{D}[\tilde{\theta}_2], ..., \mathbf{D}[\tilde{\theta}_n], ...$ сходится к величине, обратной информации Фишера, при $n\to\infty$, т.е.

$$\mathbf{D}[\tilde{\theta}_n]\to\frac{1}{I_n(\theta)}\,. \tag{2.14}$$

Пример 2.1. Покажем, что среднее арифметическое $\bar{x}$ выборки $x_1, \ldots, x_n$ наблюдений случайной величины X является состоятельной и несмещённой оценкой математического ожидания m случайной величины X.

Выборка $x_1, \ldots, x_n$ является реализацией случайной выборки $X_1, \ldots, X_n$, а среднее арифметическое $\bar{x}$ – реализацией статистики $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$.

Найдём математическое ожидание и дисперсию статистики $\bar{X}$:

$$\mathbf{M}[\bar{X}]=\mathbf{M}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbf{M}[X_i]=\frac{1}{n}nm=m\,,$$

$$\mathbf{D}[\bar{X}]=\mathbf{D}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}[X_i]=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}\,.$$

Поскольку математическое ожидание точечной оценки $\bar{X}$ равно оцениваемому параметру m, то оценка является несмещённой. Оценка состоятельна, так как для неё выполнено достаточное условие состоятельности: $\mathbf{M}[\bar{X}] = m$ и $\mathbf{D}[\bar{X}] \to 0$ при $n \to \infty$.

Пример 2.2. Покажем, что выборочная дисперсия d_X^* выборки $x_1, \ldots, x_n$ наблюдений случайной величины X является смещённой оценкой дисперсии σ^2 случайной величины X.

Найдём математическое ожидание и дисперсию статистики $D_X^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, реализацией которой является выборочная дисперсия d_X^*. Предварительно преобразуем выражение для выборочной дисперсии:

$$D_X^* = \frac{1}{n}\sum_{i=1}^{n}\Big((X_i - m) - (\bar{X} - m)\Big)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - 2(\bar{X} - m)\frac{1}{n}\sum_{i=1}^{n}(X_i - m) + \frac{1}{n}\sum_{i=1}^{n}(\bar{X} - m)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - 2(\bar{X} - m)^2 + \frac{1}{n}n(\bar{X} - m)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\bar{X} - m)^2.$$

Используя свойства математического ожидания и дисперсии, запишем:

$$\mathbf{M}[D_X^*] = \frac{1}{n}\mathbf{M}\left[\sum_{i=1}^{n}(X_i - m)^2\right] - \mathbf{M}\left[(\bar{X} - m)^2\right] =$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}[X] - \mathbf{D}[\bar{X}] = \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Из полученного выражения, в частности, следует, что:

1) выборочная дисперсия является асимптотически несмещённой оценкой дисперсии, а её смещение равно $b_n(\sigma^2) = -\dfrac{\sigma^2}{n}$;

2) несмещённой оценкой выборочной дисперсии является статистика $S_X^2 = \dfrac{n}{n-1} D_X^* = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, называемая *исправленной выборочной дисперсией*.

Можно показать, что выборочная и исправленная дисперсии являются состоятельными оценками. Для доказательства состоятельности следует воспользоваться достаточным условием состоятельности или законом больших чисел.

Пример 2.3. Покажем, что среднее арифметическое $\bar{x}$ выборки $x_1, \ldots, x_n$ наблюдений случайной величины X из нормально распределённой генеральной совокупности $N(m, \sigma)$ является эффективной оценкой математического ожидания m.

Проверим критерий эффективности, а именно: обращается ли неравенство Рао–Крамера в равенство для оценки $\bar{X}$. Для этого вначале рассчитаем вклад выборки и информацию Фишера:

$$U(x, m) = \frac{\partial}{\partial m}\ln f_X(x, m) = \frac{\partial}{\partial m}\ln\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)\right] =$$

$$= \frac{\partial}{\partial m}\ln\left[\frac{1}{\sigma\sqrt{2\pi}}\right] + \frac{\partial}{\partial m}\left(-\frac{(x-m)^2}{2\sigma^2}\right) =$$

$$= -\frac{1}{2\sigma^2}\frac{\partial}{\partial m}(x-m)^2 = \frac{1}{2\sigma^2}2(x-m) = \frac{x-m}{\sigma^2};$$

$$I(m) = \mathbf{M}\left[U(X,m)^2\right] = \mathbf{M}\left[\left(\frac{X-m}{\sigma^2}\right)^2\right] =$$

$$= \frac{1}{\sigma^4}\mathbf{M}\left[(X-m)^2\right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Таким образом, $I_n(m) = \frac{n}{\sigma^2}$, а поскольку $\mathbf{D}[\bar{X}] = \frac{\sigma^2}{n}$ (см. пример 2.1), то статистика $\bar{X}$ является эффективной оценкой математического ожидания m.

Пример 2.4. Покажем, что относительная частота x события A в серии из n независимых испытаний, вероятность события A в каждом из которых равна p, является несмещённой, состоятельной и эффективной оценкой вероятности p.

Случайная величина K – число появлений события A в серии из n независимых испытаний – распределена по биномиальному закону $B(n, p)$. Как известно, числовые характеристики биномиального распределения

$$\mathbf{M}[K] = np,$$
$$\mathbf{D}[K] = np(1-p).$$

Случайная величина X – частота появления события A в серии из n независимых испытаний – связана со случайной величиной K соотношением $K = \frac{X}{n}$. Учитывая свойства математического ожидания и дисперсии, запишем выражения для расчёта числовых характеристик случайной величины X:

$$\mathbf{M}[X] = \mathbf{M}\left[\frac{K}{n}\right] = \frac{1}{n}np = p,$$
$$\mathbf{D}[X] = \mathbf{D}\left[\frac{K}{n}\right] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}.$$

Таким образом, статистика X является состоятельной (в силу выполнения достаточного условия состоятельности) и несмещённой (по определению) оценкой параметра p.

Для проверки эффективности используем неравенство Рао–Крамера. В результате эксперимента у нас имеется только одно наблюдение x случайной величины X, т.е. объём выборки в формуле расчёта информации Фишера (2.10) полагаем равным 1.

По формулам (2.11) и (2.13) рассчитаем вклад выборки и информацию Фишера о параметре p, содержащуюся в единственном наблюдении x:

$$U(x,\,p)=\frac{\partial}{\partial p}\ln P(x,\,p)=\frac{\partial}{\partial p}\ln\left(C_n^{nx}p^{nx}(1-p)^{n-nx}\right)=$$

$$=\frac{\partial}{\partial p}\left(\ln C_n^{nx}+nx\ln p+(n-nx)\ln(1-p)\right)=\frac{nx}{p}-\frac{n-nx}{1-p}=$$

$$=\frac{nx-npx-np+npx}{p(1-p)}=\frac{nx-np}{p(1-p)},\qquad nx=0,1,\ldots,n;$$

$$I(p)=\mathbf{M}\left[U(X,\,p)^2\right]=\mathbf{M}\left[\left(\frac{nX-np}{p(1-p)}\right)^2\right]=$$

$$=\frac{n^2\mathbf{M}\left[\left(X-p\right)^2\right]}{p^2(1-p)^2}=\frac{n^2\mathbf{D}[X]}{p^2(1-p)^2}=\frac{n^2p(1-p)}{np^2(1-p)^2}=\frac{n}{p(1-p)}.$$

Таким образом, $\mathbf{D}[X]=\dfrac{1}{I(p)}$, и согласно неравенству Рао–Крамера статистика X имеет минимально возможную дисперсию, т.е. является эффективной.

Контрольные вопросы и задачи

1. Что в математической статистике называется статистикой?
2. Дайте определения точечной оценки и реализации точечной оценки неизвестного параметра θ.
3. Какую точечную оценку называют состоятельной?
4. Сформулируйте достаточное условие состоятельности.
5. Какую точечную оценку называют несмещённой?
6. Какую точечную оценку называют асимптотически несмещённой?
7. Приведите пример несмещённой, но несостоятельной оценки. Приведите пример состоятельной, но смещённой оценки.
8. В каком случае точечная оценка $\tilde{\theta}$ более эффективна, чем точечная оценка $\hat{\theta}$ неизвестного параметра θ?
9. Сформулируйте неравенство Рао–Крамера.

10. Что называется эффективностью точечной оценки? Какую точечную оценку называют эффективной?

11. Сформулируйте критерий эффективности точечной оценки.

12. Что называется вкладом выборки?

13. Что называется информацией Фишера о неизвестном параметре θ, содержащейся в одном выборочном наблюдении?

14. Какую точечную оценку называют асимптотически эффективной?

15. Покажите, что выборочная дисперсия является состоятельной оценкой дисперсии генеральной совокупности.

16. Покажите, что если $\tilde{\theta}$ – несмещённая оценка параметра θ, то $\tilde{\theta}^2$ будет смещённой оценкой параметра θ^2.

17. Покажите, что статистика $S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ является смещённой оценкой среднеквадратичного отклонения σ генеральной совокупности.

18. Покажите, что статистика X_1, равная первому элементу случайной выборки $X_1, \ldots, X_n$, является несмещённой. Состоятельна ли ли эта оценка? Является ли эта оценка эффективной для нормально распределённой генеральной совокупности?

§ 7. Методы получения точечных оценок

Точечной оценкой неизвестного параметра θ, вообще говоря, может являться любая статистика. Однако на практике интерес представляют лишь наиболее качественные оценки, для которых вероятность того, что при реализации случайной выборки они примут значение, максимально близкое к неизвестному значению θ, наибольшая. Такие оценки должны быть несмещёнными, состоятельными и эффективными. Возникает вопрос: как получить качественную оценку для произвольного параметра θ наблюдаемой случайной величины X?

Ниже рассмотрены три основных метода получения точечных оценок.

1. *Метод подстановки*. Метод состоит в том, что в качестве оценки $\tilde{\theta}$ неизвестного параметра θ выбирается соответствующая выборочная характеристика:

$$\tilde{\theta} = \theta^*. \qquad (2.15)$$

Например, согласно методу подстановки оценкой математического ожидания будет выборочное среднее, а оценкой дисперсии – выборочная дисперсия.

Все оценки, полученные по методу подстановки, являются состоятельными, однако их несмещённость и эффективность не гарантированы. Примером смещённой оценки, рассмотренной ранее, является выборочная дисперсия.

2. *Метод моментов*. Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей распределение $F_X(x, \theta)$ с вектором неизвестных параметров $\theta = (\theta_1, ..., \theta_k)$. Предположим, что для этого распределения могут быть рассчитаны начальные $\alpha_r = \alpha_r(\theta_1, ..., \theta_k)$ и центральные $\mu_r = \mu_r(\theta_1, ..., \theta_k)$ моменты некоторых порядков r. Эти моменты являются функциями неизвестных параметров $\theta_1, \ldots, \theta_k$. В то же время, для выборки могут быть рассчитаны выборочные начальные α_r^* и центральные μ_r^* моменты тех же порядков r.

Метод моментов состоит в нахождении такого вектора параметров θ, при котором теоретические моменты равны выборочным моментам, т.е. в решении системы уравнений вида

$$\begin{cases} \alpha_{r_i}(\theta_1, ..., \theta_k) = \alpha_{r_i}^*, \quad i = 1, 2, ...; \\ \mu_{r_j}(\theta_1, ..., \theta_k) = \mu_{r_j}^*, \quad j = 1, 2, ... \end{cases} \qquad (2.16)$$

Число уравнений в системе (2.16) выбирается равным числу неизвестных параметров k. Для получения оценок по методу моментов, вообще говоря, могут быть выбраны любые моменты произвольных порядков, однако, как правило, на практике используют лишь моменты низших порядков.

Все оценки, рассчитанные по методу моментов, являются состоятельными, однако их несмещённость и эффективность так же, как и в случае метода подстановки, не гарантированы.

Точечные оценки, полученные по методу моментов, называются *ММ-оценками*.

Пример 2.5. Найти ММ-оценку параметров a и b по выборке $x_1, \ldots, x_n$ из равномерно распределённой генеральной совокупности $X \sim R(a, b)$.

Для построения системы уравнений (2.16) выберем два момента, например математическое ожидание и дисперсию. Известно, что для случайной величины $X \sim R(a, b)$:

$$\mathbf{M}[X] = \frac{a+b}{2}, \qquad \mathbf{D}[X] = \frac{(b-a)^2}{12}.$$

Составим систему уравнений, приравняв теоретические моменты выборочным:

$$\begin{cases} \dfrac{a+b}{2} = \overline{x}; \\ \dfrac{(b-a)^2}{12} = d_X^*, \end{cases}$$

решая которую, получим точечные оценки:

$$\tilde{a} = \overline{x} - \sqrt{3d_X^*}, \quad \tilde{b} = \overline{x} + \sqrt{3d_X^*}.$$

3. *Метод максимального правдоподобия* (*maximum likelihood estimation, MLE*). Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей распределение $F_X(x, \theta)$ с вектором неизвестных параметров $\theta = (\theta_1, \ldots, \theta_k)$. *Функцией правдоподобия* выборки $x_1, \ldots, x_n$ из генеральной совокупности X называется совместная функция плотности распределения случайного вектора $X = (X_1, \ldots, X_n)$ при условии, что его реализация равна $x = (x_1, \ldots, x_n)$:

$$L(x_1, \ldots, x_n; \theta) = f_{X_1 \ldots X_n}(x_1, \ldots, x_n; \theta). \qquad (2.17)$$

Для независимой случайной выборки $X_1, \ldots, X_n$ многомерная функция плотности есть произведение одномерных функций плотностей:

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta). \qquad (2.18)$$

В (2.18) учтено, что все компоненты X_1, …, X_n имеют одинаковое распределение, совпадающее с распределением генеральной совокупности X.

Функция правдоподобия выборки x_1, …, x_n является функцией только вектора неизвестных параметров θ.

Аналогично определяется функция правдоподобия для случая дискретной генеральной совокупности с распределением вероятностей $P(x, \theta)$, $\sum_x P(x, \theta) = 1$:

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} P(X_i = x_i; \theta) = \prod_{i=1}^{n} P(x_i; \theta). \qquad (2.19)$$

Метод максимального правдоподобия состоит в том, что в качестве оценки вектора неизвестных параметров $\theta = (\theta_1, ..., \theta_k)$ принимается вектор $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_k)$, доставляющий максимум функции правдоподобия, т.е.

$$\tilde{\theta} = \arg\max_{\theta} L(x_1, ..., x_n; \theta). \qquad (2.20)$$

Иными словами, метод максимального правдоподобия состоит в нахождении такого вектора параметров $\tilde{\theta}$, при котором данная реализация x_1, …, x_n случайной выборки X_1, …, X_n была бы наиболее вероятной.

Запишем необходимое условие экстремума функции правдоподобия:

$$\frac{\partial L(x_1, ..., x_n; \theta)}{\partial \theta_i} = 0, \quad i = \overline{1,k}. \qquad (2.21)$$

Это система k уравнений с k неизвестными θ_1, …, θ_k, решая которую, получаем оценки $\tilde{\theta}_1, ..., \tilde{\theta}_k$ неизвестных параметров распределения.

На практике бывает удобно вместо системы уравнений (2.21) составить систему уравнений

$$\frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial \theta_i} = 0, \quad i = \overline{1,k}, \qquad (2.22)$$

которая имеет те же решения. Функция $\ln L(x_1, ..., x_n; \theta)$ называется *логарифмической функцией правдоподобия*.

Все оценки, рассчитанные по методу максимального правдоподобия, являются состоятельными и, по крайней мере, асимптотически несмещёнными и асимптотически эффективными, в связи с чем метод максимального правдоподобия получил широкое распространение на практике. Если для неизвестного параметра существует эффективная оценка, то метод максимального правдоподобия даёт именно эту оценку.

Точечные оценки, полученные по методу максимального правдоподобия, называются *МП-оценками*.

Пример 2.6. Найти МП-оценки параметров m и σ^2 по выборке $x_1, \ldots, x_n$ из нормально распределённой генеральной совокупности $X \sim N(m, \sigma)$.

Найдём функцию правдоподобия выборки $x_1, \ldots, x_n$ из генеральной совокупности $X \sim N(m, \sigma)$:

$$L(x_1, ..., x_n; m, \sigma^2) = \prod_{i=1}^{n} f_X(x_i; m, \sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right) =$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right);$$

$$\ln L(x_1, ..., x_n; m, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2.$$

Необходимое условие максимума функции правдоподобия:

$$\begin{cases} \dfrac{\partial \ln L(x_1, ..., x_n; m, \sigma^2)}{\partial m} = 0; \\ \dfrac{\partial \ln L(x_1, ..., x_n; m, \sigma^2)}{\partial \sigma^2} = 0; \end{cases}$$

$$\begin{cases} \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i - m) = 0; \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(x_i - m)^2 = 0. \end{cases}$$

Решая эту систему уравнений, получаем МП-оценки:

$$\tilde{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}\,, \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = D_X^*\,.$$

Контрольные вопросы и задачи

1. В чём состоит метод подстановки для получения точечных оценок параметров генеральной совокупности? Какими свойствами обладают оценки, полученные по методу подстановки?
2. В чём состоит метод моментов для получения точечных оценок параметров генеральной совокупности? Какими свойствами обладают ММ-оценки?
3. Что называется функцией правдоподобия выборки? В чём её статистический смысл?
4. В чём состоит метод максимального правдоподобия для получения точечных оценок параметров генеральной совокупности? Какими свойствами обладают МП-оценки?
5. Найдите ММ-оценки и МП-оценки параметров равномерного распределения, биномиального распределения, распределения Пуассона. Сравните полученные выражения.

§ 8. Точечные оценки математического ожидания и дисперсии

I. *Оценки математического ожидания.*

1. Оптимальная оценка математического ожидания – выборочное среднее:

$$\tilde{m} = \overline{X}\,.$$

Оценка является несмещённой, состоятельной, эффективной.

2. На практике нередко возникает необходимость быстрой оценки математического ожидания. Такой оценкой может быть, например, статистика

$$\tilde{m} = \frac{X_{\min} + X_{\max}}{2},$$

где $X_{\min}$ и $X_{\max}$ – экстремальные порядковые статистики.

Оценка является состоятельной, асимптотически несмещённой, но неэффективной.

3. В качестве оценки математического ожидания симметричного распределения может быть использована выборочная медиана:

$$\tilde{m} = x^*_{0,5}.$$

Можно показать, что при больших объёмах выборки распределение статистики $X^*_{0,5}$ аппроксимируется нормальным распределением $N\left(m, \sigma\sqrt{\frac{\pi}{2n}}\right)$. Таким образом, эффективность выборочной медианы как оценки математического ожидания нормально распределенной генеральной совокупности равна

$$e(X^*_{0,5}) = \frac{1/I_n(m)}{\pi\sigma^2/2n} = \frac{\sigma^2/n}{\pi\sigma^2/2n} = \frac{2}{\pi} \approx 64\%.$$

Оценка является состоятельной, несмещённой, но неэффективной.

4. Пусть даны две выборки объёмов n_1 и n_2 из одной генеральной совокупности, $\bar{X}_1$ и $\bar{X}_2$ – выборочные средние. Тогда *агрегированная оценка* математического ожидания генеральной совокупности

$$\tilde{m} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

является несмещённой, состоятельной, эффективной.

II. *Оценки дисперсии.*

1. Оптимальная оценка дисперсии при неизвестном математическом ожидании генеральной совокупности – исправленная выборочная дисперсия:

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Оценка является несмещённой, состоятельной, эффективной.

2. Выборочная дисперсия

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

Оценка является асимптотически несмещённой, состоятельной, асимптотически эффективной.

3. На практике нередко возникает необходимость быстрой оценки дисперсии. Такой оценкой может быть, например, статистика

$$\tilde{\sigma}^2 = \left(\frac{X_{\max} - X_{\min}}{5}\right)^2,$$

где $X_{\min}$ и $X_{\max}$ – экстремальные порядковые статистики.

Оценка является грубой, для большинства распределений смещённой и неэффективной.

4. В случае если известно математическое ожидание m генеральной совокупности, оптимальная оценка дисперсии – статистика

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 .$$

Оценка является несмещённой, состоятельной, эффективной.

5. Пусть даны две выборки объёмов n_1 и n_2 из одной генеральной совокупности, S_1^2 и S_2^2 – исправленные выборочные дисперсии. Тогда *агрегированная оценка* дисперсии генеральной совокупности

$$\tilde{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

является несмещённой, состоятельной, эффективной.

Контрольные вопросы и задачи

1. Сравните различные оценки математического ожидания. Какими свойствами они обладают?
2. Сравните различные оценки дисперсии. Какими свойствами они обладают?
3. Приведите примеры оценок среднеквадратичного отклонения. Какими свойствами они обладают?

Глава 3. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

§ 9. Понятие доверительного интервала

Точечная оценка $\tilde{\theta}$ неизвестного параметра θ является случайной величиной, определяемой как некоторая функция случайной выборки $X_1, \ldots, X_n$. Это означает, что для каждой новой реализации $x_1, \ldots, x_n$ этой выборки точечная оценка $\tilde{\theta}$ каждый раз будет принимать новое значение. Использование точечных оценок не даёт ответа на вопрос, насколько для данной выборки $x_1, \ldots, x_n$ рассчитанная реализация точечной оценки $\tilde{\theta}$ близка к значению оцениваемого параметра θ. Ответ на этот вопрос могут дать интервальные оценки, позволяющие получить вероятностную характеристику точности оценивания неизвестного параметра θ.

Пусть $X_1, \ldots, X_n$ – случайная выборка объёма n из генеральной совокупности X с функцией распределения $F_X(x, \theta)$, зависящей от параметра θ, значение которого неизвестно. *Доверительным интервалом* для параметра θ называется интервал $(\theta_1; \theta_2)$, содержащий (накрывающий) истинное значение θ с заданной вероятностью γ, т.е.

$$P(\theta_1 < \theta < \theta_2) = \gamma, \qquad (3.1)$$

где $\theta_1 = \theta_1(X_1, ..., X_n)$ и $\theta_2 = \theta_2(X_1, ..., X_n)$ – некоторые статистики. Вероятность γ называется *доверительной вероятностью*, а вероятность $\alpha = 1 - \gamma$ – *уровнем значимости*. Доверительный интервал с доверительной вероятностью γ называют также γ-*доверительным интервалом*, или γ-*доверительной интервальной оценкой* параметра θ. Статистики θ_1 и θ_2 называются *нижней* и *верхней границами* доверительного интервала соответственно.

Для каждой новой реализации $x_1, \ldots, x_n$ случайной выборки $X_1, \ldots, X_n$ границы доверительного интервала – случайные величины θ_1 и θ_2 – будут принимать новые значения. Однако, согласно определению, для данной реализации $x_1, \ldots, x_n$ соответствующая

реализация доверительного интервала $(\theta_1; \theta_2)$ накроет истинное значение неизвестного параметра θ с заданной вероятностью γ. Это означает, что доля реализаций случайной выборки $X_1, \ldots, X_n$, для которых доверительный интервал накроет θ, в среднем равна доверительной вероятности γ.

Пример 3.1. Исследуется качество партии выпускаемых предприятием изделий. Пусть θ – доля бракованных изделий в партии, которую оценивают независимо друг от друга в N различных лабораториях по результатам обследования нескольких случайно выбранных деталей из партии. Иначе говоря, долю бракованных изделий в партии в каждой лаборатории оценивают по «своей» выборке деталей, и в каждой лаборатории получают свои значения верхней и нижней границ γ-доверительного интервала.

Возможны случаи, когда γ-доверительный интервал не накрывает истинного значения θ. Если M – число таких случаев, то их доля будет стремиться к уровню значимости α при увеличении N, т.е. $\frac{M}{N} \to \alpha$ при $N \to \infty$.

Ширина доверительного интервала, характеризующая точность интервального оценивания, зависит от объёма выборки n и доверительной вероятности γ: при увеличении объёма выборки ширина доверительного интервала уменьшается. Причина этого состоит в том, что в выборке большего объёма содержится больше информации об оцениваемом параметре, что позволяет более точно определить область, в которой он находится. При увеличении доверительной вероятности предъявляется более «жёсткое» требование к вероятности нахождения неизвестного параметра внутри доверительного интервала, вследствие чего его ширина увеличивается.

Границы доверительного интервала θ_1 и θ_2 могут быть выбраны множеством способов. Единственное требование, предъявляемое к этим статистикам – это выполнение условия (3.1). Однако на практике, как правило, эти статистики выбирают, исходя из некоторых соображений симметрии, которые будут рассмотрены далее.

Иногда требуется оценить параметр θ только снизу или только сверху. Если выполняется условие

$$P(\theta_1 < \theta) = \gamma, \qquad (3.2)$$

то доверительный интервал $(\theta_1; \infty)$ называется *правосторонним*, а статистика $\theta_1 = \theta_1(X_1, ..., X_n)$ – *односторонней нижней границей* доверительного интервала.

Если же

$$P(\theta < \theta_2) = \gamma, \qquad (3.3)$$

то доверительный интервал $(-\infty; \theta_2)$ называется *левосторонним*, а статистика $\theta_2 = \theta_2(X_1, ..., X_n)$ – *односторонней верхней границей* доверительного интервала.

Контрольные вопросы и задачи

1. Дайте определение доверительного интервала для неизвестного параметра распределения генеральной совокупности.
2. Что называется доверительной вероятностью?
3. Что называется уровнем значимости?
4. В чём основное преимущество интервального оценивания по сравнению с точечным оцениванием?
5. Как объём выборки влияет на ширину доверительного интервала?
6. Как уровень значимости влияет на ширину доверительного интервала?
7. Какой доверительный интервал называется односторонним?

§ 10. Метод построения доверительных интервалов

Пусть $X_1, \ldots, X_n$ – случайная выборка объёма n из генеральной совокупности X с функцией распределения $F_X(x, \theta)$, зависящей от параметра θ, значение которого неизвестно. Наиболее простым и популярным методом построения доверительного интервала $(\theta_1; \theta_2)$ для неизвестного параметра θ является метод, основанный на использовании так называемой центральной статистики.

Центральной статистикой случайной выборки $X_1, \ldots, X_n$ называется любая статистика $Z = Z(X_1, ..., X_n; \theta)$, зависящая от неизвестного параметра θ, удовлетворяющая следующим свойствам:

1) закон распределения $F_Z(z)$ статистики Z известен и не зависит от θ;

2) статистика Z непрерывна и строго монотонна по θ.

Из определения квантили следует, что для любой случайной величины, в том числе и для статистики $Z = Z(X_1, ..., X_n; \theta)$, справедливо равенство

$$P\left(z_{\alpha/2} < Z(X_1, ..., X_n; \theta) < z_{1-\alpha/2}\right) = 1 - \alpha, \qquad (3.4)$$

где $z_{\alpha/2}$ и $z_{1-\alpha/2}$ – квантили случайной величины Z на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

При построении односторонних доверительных интервалов рассматриваются другие равенства:

$$P\left(z_{\alpha} < Z(X_1, ..., X_n; \theta)\right) = 1 - \alpha, \qquad (3.5)$$

$$P\left(Z(X_1, ..., X_n; \theta) < z_{1-\alpha}\right) = 1 - \alpha. \qquad (3.6)$$

Задача нахождения доверительного интервала состоит в разрешении неравенства, стоящего под знаком вероятности в выражении (3.4) (или выражениях (3.5), (3.6)), относительно неизвестного параметра θ. В результате получим эквивалентное выражение

$$P\left(\theta_1(X_1, ..., X_n) < \theta < \theta_2(X_1, ..., X_n)\right) = 1 - \alpha, \qquad (3.7)$$

из которого следует, что интервал $\left(\theta_1(X_1, ..., X_n); \theta_2(X_1, ..., X_n)\right)$ является доверительным на уровне значимости α.

Таким образом, алгоритм построения доверительного интервала для неизвестного параметра θ на основе случайной выборки $X_1, \ldots, X_n$ состоит в следующем.

1. Выбор центральной статистики $Z = Z(X_1, ..., X_n; \theta)$ и определение её закона распределения $F_Z(z)$. Знание закона распределения необходимо для расчёта квантилей $z_{\alpha/2}$ и $z_{1-\alpha/2}$ (или z_α и $z_{1-\alpha}$).

2. Разрешение неравенства под знаком вероятности в выражении (3.4) (или выражениях (3.5) и (3.6)) относительно θ.

Очевидно, что для случайной выборки $X_1, \ldots, X_n$ в общем случае может быть построено бесконечно много центральных статистик Z.

Возникает вопрос: какую центральную статистику выбрать, чтобы полученный с её помощью доверительный интервал был бы наиболее узким, а следовательно, наиболее точным, при фиксированной доверительной вероятности $\gamma = 1 - \alpha$?

Как правило, центральные статистики связывают с некоторой точечной оценкой $\tilde{\theta}$ неизвестного параметра θ. Чем меньше дисперсия точечной оценки $\tilde{\theta}$, тем меньшей дисперсией будет обладать и центральная статистика Z, построенная на основе $\tilde{\theta}$. А для случайной величины с меньшей дисперсией интервал $(z_{\alpha/2}; z_{1-\alpha/2})$ будет у́же при прочих равных условиях (рис. 3.1). Учитывая монотонность зависимости центральной статистики Z от параметра θ, заключаем, что чем у́же интервал $(z_{\alpha/2}; z_{1-\alpha/2})$, тем у́же соответствующий доверительный интервал $(\theta_1; \theta_2)$. Поскольку наименьшей дисперсией обладают эффективные оценки, то *центральную статистику* $Z(X_1, ..., X_n; \theta)$ *целесообразно выбирать связанной с эффективной оценкой* $\tilde{\theta}(X_1, ..., X_n)$ *неизвестного параметра* θ.

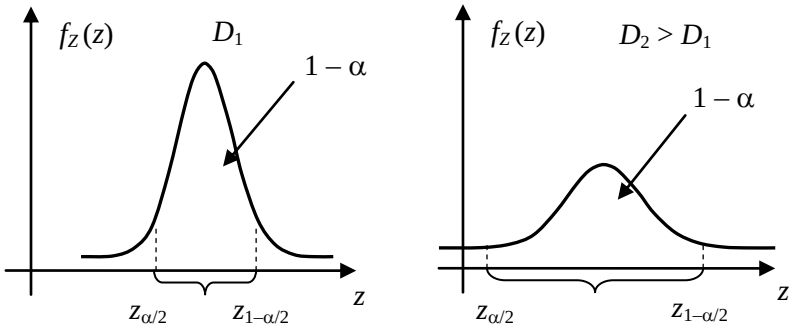

Рис. 3.1. Распределения центральных статистик с различными дисперсиями

Пример 3.2. Пусть $x_1, \ldots, x_n$ – выборка наблюдений нормально распределённой случайной величины $X \sim N(m, \sigma)$. Найти доверительный интервал для математического ожидания m при довери-

тельной вероятности $1 - \alpha$, если дисперсия генеральной совокупности σ^2 известна.

Будем строить центральную статистику на основе эффективной оценки математического ожидания $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, где $X_1, \ldots, X_n$ – случайная выборка, все элементы которой имеют распределение генеральной совокупности, т.е. $X_i \sim N(m, \sigma)$, $i = \overline{1,n}$. Учитывая композиционную устойчивость нормального распределения, случайная величина $\bar{X}$ также будет иметь нормальное распределение. Применяя свойства математического ожидания и дисперсии, получим $\mathbf{M}[\bar{X}] = m$, $\mathbf{D}[\bar{X}] = \frac{\sigma^2}{n}$ (см. пример 2.1), т.е. $\bar{X} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Введём статистику $Z(X_1, ..., X_n; m) = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$, имеющую распределение $N(0, 1)$. Для неё выполнены все требования, предъявляемые к центральным статистикам. Запишем тождество (3.4) для статистики Z:

$$P\left(u_{\alpha/2} < \frac{\bar{X} - m}{\sigma/\sqrt{n}} < u_{1-\alpha/2}\right) = 1 - \alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили стандартизованного нормального распределения на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно m и учитывая симметричность нормального распределения ($u_{\alpha/2} = -u_{1-\alpha/2}$), получим

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < m < \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right) = 1 - \alpha,$$

откуда следует, что интервал $\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}; \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right)$ является доверительным для m с доверительной вероятностью $1 - \alpha$.

Контрольные вопросы и задачи

1. Дайте определение центральной статистики.
2. Сформулируйте алгоритм построения доверительного интервала на основе центральной статистики.
3. Как выбирается центральная статистика при построении доверительного интервала?
4. Используя метод центральной статистики, рассчитайте лево- и правосторонний доверительные интервалы для математического ожидания m при доверительной вероятности $1 - \alpha$, если дисперсия генеральной совокупности σ^2 известна.

§ 11. Законы распределения некоторых статистик нормальной выборки

Пусть $X_1, \ldots, X_n$ – случайная выборка объёма n из нормально распределённой генеральной совокупности $N(m, \sigma)$. Для вывода выражений для расчёта границ доверительных интервалов найдём законы распределения некоторых статистик, которые могут быть использованы как центральные.

1. Статистика $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (*среднее арифметическое*).

В силу композиционной устойчивости нормального закона статистика $\bar{X}$ имеет распределение $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ (см. пример 3.2).

2. Статистика $U = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$ (*стандартизованное среднее арифметическое при известной дисперсии*).

Статистика U получается путём стандартизации статистики $\bar{X}$ и имеет распределение $U \sim N(0,1)$.

3. Статистика $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ (*оценка дисперсии при известном математическом ожидании*).

Для вывода закона распределения домножим и разделим каждое слагаемое на σ^2:

$$S_0^2 = \frac{1}{n}\sigma^2 \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \frac{1}{n}\sigma^2 \sum_{i=1}^{n} U_i^{\,2},$$

где случайные величины U_i, $i = \overline{1,n}$, независимы и имеют стандартизованное нормальное распределение $N(0, 1)$. По определению закона распределения хи-квадрат, случайная величина $\frac{n}{\sigma^2} S_0^2 \sim \chi^2(n)$. Далее будем записывать, что статистика $S_0^2 \sim \frac{\sigma^2}{n}\chi^2(n)$.

4. Статистика $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (*оценка дисперсии при неизвестном математическом ожидании*).

Теорема Фишера. Пусть $X_1, \ldots, X_n$ – независимые случайные величины, имеющие нормальное распределение $N(m, \sigma)$. Тогда случайные величины $\bar{X}$ и S^2 независимы, и случайная величина $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.

5. Статистика $T = \frac{\bar{X} - m}{S/\sqrt{n}}$ (*стандартизованное среднее арифметическое при неизвестной дисперсии*).

Учитывая выражения для статистик U (см. п. 2) и S (см. п. 4), запишем

$$T = \frac{\sigma}{S}\frac{\bar{X} - m}{\sigma/\sqrt{n}} = \frac{\sigma U}{S} = \frac{\sigma U}{\sqrt{\frac{\sigma^2}{n-1}\chi^2(n-1)}} = \frac{U}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}.$$

По определению закона распределения Стьюдента статистика $T \sim T(n-1)$.

Рассмотрим теперь законы распределения некоторых статистик, связанных с двумя случайными выборками. Пусть $X_{11}, \ldots, X_{1,n_1}$ и $X_{21}, \ldots, X_{2,n_2}$ – случайные выборки объёмов n_1 и n_2 из нормально

распределённых генеральных совокупностей $N(m_1, \sigma_1)$ и $N(m_2, \sigma_2)$ соответственно.

6. Статистика $\bar{X} = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$ (*агрегированное среднее*).

Поскольку средние арифметические выборок имеют нормальные распределения $\bar{X}_1 \sim N\left(m_1, \dfrac{\sigma_1}{\sqrt{n_1}}\right)$ и $\bar{X}_2 \sim N\left(m_2, \dfrac{\sigma_2}{\sqrt{n_2}}\right)$ (см. п. 1), а нормальный закон композиционно устойчив, то статистика $\bar{X}$ также будет иметь нормальное распределение. Применяя свойства операторов математического ожидания и дисперсии, находим его параметры:

$$\mathbf{M}[\bar{X}] = \frac{1}{n_1 + n_2}\left(n_1\mathbf{M}[\bar{X}_1] + n_2\mathbf{M}[\bar{X}_2]\right) = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2},$$

$$\mathbf{D}[\bar{X}] = \frac{1}{(n_1 + n_2)^2}\left(n_1^2\mathbf{D}[\bar{X}_1] + n_2^2\mathbf{D}[\bar{X}_2]\right) = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{(n_1 + n_2)^2}.$$

Таким образом, $\bar{X} \sim N\left(\dfrac{n_1 m_1 + n_2 m_2}{n_1 + n_2}, \dfrac{\sqrt{n_1\sigma_1^2 + n_2\sigma_2^2}}{n_1 + n_2}\right)$.

7. Статистика $U = \dfrac{n_1(\bar{X}_1 - m_1) + n_2(\bar{X}_2 - m_2)}{\sqrt{n_1\sigma_1^2 + n_2\sigma_2^2}}$ (*стандартизованное агрегированное среднее при известных дисперсиях*).

Статистика U получается путём стандартизации статистики $\bar{X}$ и имеет распределение $U \sim N(0,1)$.

8. Статистика $S_0^2 = \dfrac{n_1 S_{01}^2 + n_2 S_{02}^2}{n_1 + n_2}$ (*агрегированная оценка дисперсии при известном математическом ожидании*).

Если $\sigma_1 = \sigma_2 = \sigma$, то в силу композиционной устойчивости распределения хи-квадрат статистика имеет распределение $S_0^2 \sim \dfrac{\sigma^2}{n_1 + n_2}\chi^2(n_1 + n_2)$.

9. Статистика $S^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ (*агрегированная оценка дисперсии при неизвестных математическых ожиданиях*).

Если $\sigma_1 = \sigma_2 = \sigma$, то в силу композиционной устойчивости распределения хи-квадрат статистика имеет распределение $S^2 \sim \dfrac{\sigma^2}{n_1+n_2-2}\chi^2(n_1+n_2-2)$.

10. Статистика $\Delta = \bar{X}_1 - \bar{X}_2$ (*разность средних при известных дисперсиях*).

В связи с композиционной устойчивостью нормального распределения статистика Δ также будет иметь нормальное распределение. Применяя свойства операторов математического ожидания и дисперсии, находим его параметры:

$$\mathbf{M}[\Delta] = \mathbf{M}[\bar{X}_1] - \mathbf{M}[\bar{X}_2] = m_1 - m_2,$$

$$\mathbf{D}[\Delta] = \mathbf{D}[\bar{X}_1] + \mathbf{D}[\bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Таким образом, $\Delta \sim N\left(m_1 - m_2, \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$.

11. Статистика $U = \dfrac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ (*стандартизованная разность средних при известных дисперсиях*).

Статистика U получается путём стандартизации статистики Δ и имеет распределение $U \sim N(0,1)$.

В частном случае, если $\sigma_1 = \sigma_2 = \sigma$, то $U = \dfrac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

12. Статистика $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ (*стандартизованная разность средних при неизвестных дисперсиях*).

Если $\sigma_1 = \sigma_2 = \sigma$, то

$$T = \frac{\sigma}{S}\frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{\sigma U}{S} =$$

$$= \frac{\sigma U}{\sqrt{\dfrac{\sigma^2}{n_1 + n_2 - 2}\chi^2(n_1 + n_2 - 2)}} = \frac{U}{\sqrt{\dfrac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}}.$$

По определению закона распределения Стьюдента статистика $T \sim T(n_1 + n_2 - 2)$.

13. Статистика $F_0 = \dfrac{S_{01}^2 / \sigma_1^2}{S_{02}^2 / \sigma_2^2}$ (*стандартизованное отношение дисперсий при известных математических ожиданиях*).

Применяя выражение для статистики S_0^2 (см. п. 3), запишем:

$$F_0 = \frac{\dfrac{\sigma_1^2}{n_1}\chi^2(n_1)/\sigma_1^2}{\dfrac{\sigma_2^2}{n_2}\chi^2(n_2)/\sigma_2^2} = \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}.$$

По определению закона распределения Фишера статистика $F_0 \sim F(n_1, n_2)$.

14. Статистика $F = \dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$ (*стандартизованное отношение дисперсий при неизвестных математических ожиданиях*).

Применяя выражение для статистики S^2 (см. п. 4), запишем:

$$F = \frac{\frac{\sigma_1^2}{n_1 - 1}\chi^2(n_1 - 1)/\sigma_1^2}{\frac{\sigma_2^2}{n_2 - 1}\chi^2(n_2 - 1)/\sigma_2^2} = \frac{\chi^2(n_1 - 1)/(n_1 - 1)}{\chi^2(n_2 - 1)/(n_2 - 1)} .$$

По определению закона распределения Фишера статистика $F \sim F(n_1 - 1, n_2 - 1)$.

Контрольные вопросы и задачи

1. Сформулируйте теорему Фишера для нормальной случайной выборки.

2. Покажите, что агрегированная оценка математического ожидания $\bar{X}$, рассчитанная по двум выборкам из одной генеральной совокупности, является состоятельной и несмещённой.

3. Покажите, что агрегированная оценка дисперсии S_0^2, рассчитанная по двум выборкам из одной генеральной совокупности, является состоятельной и несмещённой.

4. Покажите, что агрегированная оценка дисперсии S^2, рассчитанная по двум выборкам из одной генеральной совокупности, является состоятельной и несмещённой.

5. Покажите, что агрегированная оценка математического ожидания $\bar{X}$, рассчитанная по двум выборкам из генеральных совокупностей с равными математическими ожиданиями $m_1 = m_2 = m$, более эффективна, чем каждая из оценок $\bar{X}_1$ и $\bar{X}_2$, вычисляемых по одной из выборок.

6. Покажите, что агрегированная оценка дисперсии S_0^2, рассчитанная по двум выборкам из генеральных совокупностей с равными дисперсиями $\sigma_1 = \sigma_2 = \sigma$, более эффективна, чем каждая из оценок S_{01}^2 и S_{02}^2, вычисляемых по одной из выборок.

7. Покажите, что агрегированная оценка дисперсии S^2, рассчитанная по двум выборкам из генеральных совокупностей с равными дисперсиями $\sigma_1 = \sigma_2 = \sigma$, более эффективна, чем каждая из оценок S_1^2 и S_2^2, вычисляемых по одной из выборок.

§ 12. Примеры построения интервальных оценок параметров нормального распределения

Пусть $X_1, \ldots, X_n$ – случайная выборка объёма n из нормально распределённой генеральной совокупности $N(m, \sigma)$. Рассмотрим варианты построения доверительных интервалов для математического ожидания m и дисперсии σ^2.

1. Доверительный интервал для математического ожидания m при известной дисперсии σ^2.

В качестве центральной статистики выберем стандартизованное среднее $U = \dfrac{\bar{X} - m}{\sigma/\sqrt{n}} \sim N(0,1)$ (см. п. 2 § 11). При таком выборе центральной статистики доверительный интервал для математического ожидания m на уровне значимости α имеет вид (см. пример 3.2)

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}; \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right).$$

2. Доверительный интервал для математического ожидания m при неизвестной дисперсии σ^2.

В качестве центральной статистики выберем стандартизованное среднее $T = \dfrac{\bar{X} - m}{S/\sqrt{n}} \sim T(n-1)$ (см. п. 5 § 11). Запишем тождество (3.4) для статистики T:

$$P\left(t_{\alpha/2}(n-1) < \frac{\bar{X} - m}{S/\sqrt{n}} < t_{1-\alpha/2}(n-1) \right) = 1 - \alpha,$$

где $t_{\alpha/2}(n-1)$ и $t_{1-\alpha/2}(n-1)$ – квантили распределения Стьюдента с $n-1$ степенями свободы на уровнях $\alpha/2$ и $1-\alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно m и учитывая симметричность распределения Стьюдента ($t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$), получим:

$$P\left(\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) < m < \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right) = 1 - \alpha,$$

откуда следует, что интервал

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1);\ \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right)$$

является доверительным для m на уровне значимости α.

3. Доверительный интервал для дисперсии σ^2 при известном математическом ожидании m.

В качестве центральной статистики выберем статистику $\frac{n}{\sigma^2} S_0^2 \sim \chi^2(n)$ (см. п. 3 § 11). Запишем для неё тождество (3.4):

$$P\left(\chi^2_{\alpha/2}(n) < \frac{n}{\sigma^2} S_0^2 < \chi^2_{1-\alpha/2}(n) \right) = 1 - \alpha,$$

где $\chi^2_{\alpha/2}(n)$ и $\chi^2_{1-\alpha/2}(n)$ – квантили распределения хи-квадрат с n степенями свободы на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно σ^2, получим:

$$P\left(\frac{nS_0^2}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{nS_0^2}{\chi^2_{\alpha/2}(n)} \right) = 1 - \alpha,$$

откуда следует, что интервал $\left(\frac{nS_0^2}{\chi^2_{1-\alpha/2}(n)}; \frac{nS_0^2}{\chi^2_{\alpha/2}(n)} \right)$ является доверительным для σ^2 на уровне значимости α.

4. Доверительный интервал для дисперсии σ^2 при неизвестном математическом ожидании m.

В качестве центральной статистики выберем статистику $\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$ (см. п. 4 § 11). Запишем для неё тождество (3.4):

$$P\left(\chi^2_{\alpha/2}(n-1) < \frac{n-1}{\sigma^2} S^2 < \chi^2_{1-\alpha/2}(n-1) \right) = 1 - \alpha,$$

где $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$ – квантили распределения хи-квадрат с $n - 1$ степенями свободы на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно σ^2, получим:

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha,$$

откуда следует, что интервал $\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}; \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right)$ является доверительным для σ^2 на уровне значимости α.

Рассмотрим теперь варианты построения доверительных интервалов, связанных с двумя выборками. Пусть $X_{11}, ..., X_{1,n_1}$ и $X_{21}, ..., X_{2,n_2}$ – случайные выборки объёмов n_1 и n_2 из нормально распределённых генеральных совокупностей $N(m_1, \sigma_1)$ и $N(m_2, \sigma_2)$ соответственно.

5. Доверительный интервал для разности математических ожиданий $m_1 - m_2$ при известных дисперсиях σ_1^2 и σ_2^2.

В качестве центральной статистики выберем стандартизованную разность средних при известных дисперсиях (см. п. 11 § 11)

$$U = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Запишем для статистики U тождество (3.4):

$$P\left(u_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < u_{1-\alpha/2}\right) = 1-\alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили стандартизованного нормального распределения на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно $m_1 - m_2$ и учитывая симметричность нормального распределения ($u_{\alpha/2} = -u_{1-\alpha/2}$), получим:

$$P\left((\bar{X}_1 - \bar{X}_2) - u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < m_1 - m_2 < \right.$$

$$\left. < (\bar{X}_1 - \bar{X}_2) + u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha,$$

откуда следует, что интервал

$$\left((\bar{X}_1 - \bar{X}_2) - u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; (\bar{X}_1 - \bar{X}_2) + u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

является доверительным для $m_1 - m_2$ на уровне значимости α.

6. Доверительный интервал для разности математических ожиданий $m_1 - m_2$ при неизвестных равных дисперсиях $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

В качестве центральной статистики выберем стандартизованную разность средних при неизвестных равных дисперсиях (см. п. 12 § 11)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2).$$

Запишем для статистики T тождество (3.4):

$$P\left(t_{\alpha/2}(n_1 + n_2 - 2) < \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\alpha/2}(n_1 + n_2 - 2)\right) = 1 - \alpha,$$

где $t_{\alpha/2}(n_1 + n_2 - 2)$ и $t_{1-\alpha/2}(n_1 + n_2 - 2)$ – квантили распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно $m_1 - m_2$ и учитывая симметричность распределения Стьюдента, получим:

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < m_1 - m_2 < \right.$$

$$\left. < (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 1 - \alpha,$$

откуда следует, что интервал

$$\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; \right.$$

$$\left. (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

является доверительным для $m_1 - m_2$ на уровне значимости α.

7. Доверительный интервал для отношения дисперсий $\frac{\sigma_1^2}{\sigma_2^2}$ *при известных математических ожиданиях* m_1 *и* m_2.

В качестве центральной статистики выберем статистику $F_0 = \frac{S_{01}^2 / \sigma_1^2}{S_{02}^2 / \sigma_2^2} \sim F(n_1, n_2)$ (см. п. 13 § 11). Запишем для неё тождество (3.4):

$$P\left(f_{\alpha/2}(n_1, n_2) < \frac{S_{01}^2 / \sigma_1^2}{S_{02}^2 / \sigma_2^2} < f_{1-\alpha/2}(n_1, n_2) \right) = 1 - \alpha,$$

где $f_{\alpha/2}(n_1, n_2)$ и $f_{1-\alpha/2}(n_1, n_2)$ – квантили распределения Фишера с n_1 и n_2 степенями свободы в числителе и знаменателе на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно $\frac{\sigma_1^2}{\sigma_2^2}$ и учитывая, что $f_{\alpha/2}(n_1, n_2) = \frac{1}{f_{1-\alpha/2}(n_2, n_1)}$, получим:

$$P\left(\frac{S_{01}^2}{S_{02}^2} f_{\alpha/2}(n_2, n_1) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_{01}^2}{S_{02}^2} f_{1-\alpha/2}(n_2, n_1) \right) = 1 - \alpha,$$

откуда следует, что интервал $\left(\frac{S_{01}^2}{S_{02}^2} f_{\alpha/2}(n_2, n_1); \frac{S_{01}^2}{S_{02}^2} f_{1-\alpha/2}(n_2, n_1) \right)$ является доверительным для $\frac{\sigma_1^2}{\sigma_2^2}$ на уровне значимости α.

8. Доверительный интервал для отношения дисперсий $\frac{\sigma_1^2}{\sigma_2^2}$ при неизвестных математических ожиданиях m_1 и m_2.

В качестве центральной статистики выберем статистику $F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$ (см. п. 14 § 11). Запишем для неё тождество (3.4):

$$P\left(f_{\alpha/2}(n_1 - 1, n_2 - 1) < \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} < f_{1-\alpha/2}(n_1 - 1, n_2 - 1) \right) = 1 - \alpha,$$

где $f_{\alpha/2}(n_1 - 1, n_2 - 1)$ и $f_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ – квантили распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы в числителе и знаменателе на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно.

Решая неравенство под знаком вероятности относительно $\frac{\sigma_1^2}{\sigma_2^2}$ и учитывая, что $f_{\alpha/2}(n_1, n_2) = \frac{1}{f_{1-\alpha/2}(n_2, n_1)}$, получим:

$$P\left(\frac{S_1^2}{S_2^2} f_{\alpha/2}(n_2 - 1, n_1 - 1) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{1-\alpha/2}(n_2 - 1, n_1 - 1) \right) = 1 - \alpha,$$

откуда следует, что интервал

$$\left(\frac{S_1^2}{S_2^2} f_{\alpha/2}(n_2 - 1, n_1 - 1); \frac{S_1^2}{S_2^2} f_{1-\alpha/2}(n_2 - 1, n_1 - 1) \right)$$

является доверительным для $\frac{\sigma_1^2}{\sigma_2^2}$ на уровне значимости α.

Пример 3.3. В условиях примера 1.4 найти доверительный интервал для средней по генеральной совокупности продолжительности горения лампочек с доверительной вероятностью 95%.

Поскольку дисперсия времени горения лампочек неизвестна, то доверительный интервал для математического ожидания времени горения лампочек (среднего времени по генеральной совокупности) имеет вид (см. п. 2):

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1);\ \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right).$$

Для расчёта реализации доверительного интервала для данной выборки необходимо рассчитать реализации $\bar{x}$ и s случайных величин $\bar{X}$ и S соответственно.

Средняя по выборке продолжительность горения лампочек равна среднему арифметическому середин интервалов группированной выборки, взвешенному объёмами групп (см. пример 1.7):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} c_i n_i = \frac{910 \cdot 8 + 930 \cdot 15 + ... + 1010 \cdot 7}{100} = 960.$$

Аналогично для исправленной дисперсии:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} (c_i - \bar{x})^2 n_i = \frac{50^2 \cdot 8 + 30^2 \cdot 15 + ... + 50^2 \cdot 7}{99} \approx 683;$$

$$s = \sqrt{683} \approx 26.$$

По таблице (табл. П4 приложения) находим квантиль распределения Стьюдента с $n-1=99$ степенями свободы на уровне $1 - \alpha/2$: $t_{0,975}(99) = 1,99$. Таким образом, с вероятностью 95% можем утверждать, что средняя продолжительность горения лампочек во всей партии лежит в интервале:

$$m \in \left(960 - \frac{26}{10} 1,99;\ 960 + \frac{26}{10} 1,99 \right) \approx (954,8;\ 965,2).$$

Пример 3.4. Исследуется качество производства элементов интегральной микросхемы на двух технологических линиях. Мерой качества производства считается дисперсия размера элементов. В

таблице представлены группированные выборочные данные размеров элементов:

Размер, мкм	0,23–0,24	0,24–0,25	0,25–0,26	0,26–0,27	0,27–0,28
Линия 1	5	30	22	11	2
Линия 2	10	34	16	8	2

Предполагая, что размеры элементов микросхемы являются нормально распределёнными случайными величинами, построить доверительный интервал для отношения дисперсий размеров элементов микросхемы, произведённых на двух технологических линиях, на уровне значимости 5%.

Поскольку математические ожидания размеров производимых элементов не известны, доверительный интервал рассчитываем по следующей формуле (см. п. 8):

$$\left(\frac{S_1^2}{S_2^2} f_{\alpha/2}(n_2 - 1, n_1 - 1); \frac{S_1^2}{S_2^2} f_{1-\alpha/2}(n_2 - 1, n_1 - 1) \right).$$

Рассчитаем исправленные оценки дисперсий s_1^2 и s_2^2:

$$\overline{x}_1 = \frac{1}{70}\left(0,235 \cdot 5 + ... + 0,275 \cdot 2\right) \approx 0,251;$$

$$\overline{x}_2 = \frac{1}{70}\left(0,235 \cdot 10 + ... + 0,275 \cdot 2\right) \approx 0,249;$$

$$s_1^2 = \frac{1}{69}\left((0,235 - 0,251)^2 \cdot 5 + ... + (0,275 - 0,251)^2 \cdot 2\right) \approx 8,71 \cdot 10^{-5};$$

$$s_2^2 = \frac{1}{69}\left((0,235 - 0,249)^2 \cdot 10 + ... + (0,275 - 0,249)^2 \cdot 2\right) \approx 9,39 \cdot 10^{-5}.$$

По таблице находим квантили распределения Фишера: $f_{0,975}(69,69) = 1,61$, $f_{0,025}(69,69) = 1/1,61 = 0,62$. Таким образом, с вероятностью 95% можем утверждать, что отношение дисперсий размеров элементов микросхемы лежит в интервале:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{8,71 \cdot 10^{-5}}{9,39 \cdot 10^{-5}} \cdot 0,62; \frac{8,71 \cdot 10^{-5}}{9,39 \cdot 10^{-5}} \cdot 1,61 \right) \approx (0,58; 1,49).$$

Контрольные вопросы и задачи

1. Какую центральную статистику используют для построения интервальной оценки математического ожидания в случае нормальной модели при известной дисперсии?
2. Возможно ли использование центральной статистики T для построения интервальной оценки математического ожидания в случае нормальной модели при известной дисперсии?
3. Какую статистику используют для построения интервальной оценки дисперсии нормально распределённой генеральной совокупности при известном математическом ожидании? По какому закону она распределена?
4. Выведите формулу расчёта одностороннего доверительного интервала для математического ожидания нормально распределённой генеральной совокупности при неизвестной дисперсии.
5. Выведите формулу расчёта одностороннего доверительного интервала для дисперсии нормально распределённой генеральной совокупности при известном математическом ожидании.
6. Пусть $X_1, \ldots, X_n$ – случайная выборка объёма n из равномерно распределённой генеральной совокупности $R(0, b)$, где b – неизвестный параметр. Предложите какую-либо центральную статистику для построения доверительного интервала для b. Какой закон распределения она имеет?

§ 13. Интервальная оценка вероятности «успеха» в схеме Бернулли

Пусть проводится серия из n испытаний по схеме Бернулли, и случайная величина X_i – исход i-го испытания ($X_i = 1$, если «успех», и $X_i = 0$, если «отказ»), $i = \overline{1,n}$. Для интервального оценивания вероятности p «успеха» в каждом отдельном испытании рассмотрим число «успехов» в серии из n испытаний, т.е. случайную величину

$$K = X_1 + \ldots + X_n, \qquad (3.8)$$

которая имеет биномиальное распределение $K \sim B(n, p)$. Как известно, математическое ожидание $m_K = np$ и дисперсия $d_K = np(1-p)$.

В соответствии с предельной теоремой Муавра–Лапласа при больших объёмах n случайной выборки статистика K имеет закон распределения, близкий к нормальному: $K \sim N\left(np, \sqrt{np(1-p)}\right)$.

Для построения доверительного интервала введём центральную статистику

$$U = \frac{K - np}{\sqrt{np(1-p)}},$$

которая представляет собой стандартизованное число «успехов» K в серии из n испытаний и при больших n имеет распределение, близкое к $N(0, 1)$.

Запишем тождество (3.4) для статистики U:

$$P\left(u_{\alpha/2} < \frac{K - np}{\sqrt{np(1-p)}} < u_{1-\alpha/2} \right) = 1 - \alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили стандартизованного нормального распределения на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно. Преобразуя неравенство под знаком вероятности, запишем:

$$P\left(\frac{K}{n} - u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \frac{K}{n} + u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha.$$

Это выражение ещё не даёт интервальной оценки параметра p, так как левая и правая части неравенства под знаком вероятности содержат этот параметр. На практике в указанные части неравенства подставляют вместо неизвестного точного значения p его эффективную оценку $H = \frac{K}{n}$. В результате получают следующий интервал для вероятности p:

$$\left(H - u_{1-\alpha/2}\sqrt{\frac{H(1-H)}{n}}; H + u_{1-\alpha/2}\sqrt{\frac{H(1-H)}{n}} \right),$$

являющийся доверительным на уровне значимости α.

Указанные границы доверительного интервала получены в результате аппроксимации и могут использоваться лишь при достаточно больших объёмах наблюдений n.

Пример 3.5. При исследовании качества выпускаемой предприятием продукции проведено обследование 100 случайно отобранных изделий. Оказалось, что шесть из них имеют брак. Определить с вероятностью 95% максимальное число бракованных изделий в партии из 1000 изделий, выпущенных тем же предприятием.

В результате статистического наблюдения случайная величина K – число бракованных изделий в выборке объёма $n = 100$ – приняла значение $k = 6$. Таким образом, реализация случайной величины H (оценки вероятности p брака изделия) приняла значение $h = \frac{k}{n} = 0,06$. По таблице (табл. П1 приложения) находим квантиль стандартизованного нормального распределения $u_{0,975} = 1,96$ и рассчитываем границы доверительного интервала для p:

$$\left(0,06 - 1,96\sqrt{\frac{0,06 \cdot 0,94}{100}}; 0,06 + 1,96\sqrt{\frac{0,06 \cdot 0,94}{100}}\right) \approx (0,014; 0,107).$$

Максимальная доля бракованных изделий в генеральной совокупности равна 0,107, следовательно, делаем вывод, что число бракованных изделий в партии из 1000 изделий не превосходит $0,107 \cdot 1000 = 107$ с вероятностью 95%.

Пусть теперь проводятся две серии испытаний по схеме Бернулли, и требуется построить доверительный интервал для разности вероятностей «успехов» p_1 и p_2 в этих сериях. Случайные величины $K_1 = X_1 + ... + X_{n_1}$ и $K_2 = Y_1 + ... + Y_{n_2}$, означающие число «успехов» в первой и второй сериях соответственно, имеют биномиальные распределения $K_1 \sim B(n_1, p_1)$, $K_2 \sim B(n_2, p_2)$, где n_1 и n_2 – число испытаний в сериях.

В соответствии с предельной теоремой Муавра–Лапласа при больших объёмах n_1 и n_2 случайных выборок статистики K_1 и K_2 имеют законы распределения, близкие к нормальному:

$K_1 \sim N\left(n_1 p_1, \sqrt{n_1 p_1 (1-p_1)}\right)$, $K_2 \sim N\left(n_2 p_2, \sqrt{n_2 p_2 (1-p_2)}\right)$. Перейдём от числа «успехов» K_1 и K_2 к относительным частотам «успехов» H_1 и H_2:

$$H_1 = \frac{K_1}{n_1} \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right),$$

$$H_2 = \frac{K_2}{n_2} \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right).$$

В силу композиционной устойчивости нормального распределения, разность относительных частот $H = H_1 - H_2$ также будет иметь нормальное распределение, параметры которого находим, используя свойства операторов математического ожидания и дисперсии:

$$H \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Для построения доверительного интервала введём центральную статистику

$$U = \frac{H - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

которая представляет собой стандартизованную разность числа «успехов» в двух сериях испытаний и при больших n_1 и n_2 имеет распределение, близкое к $N(0, 1)$.

Запишем тождество (3.4) для статистики U:

$$P\left(u_{\alpha/2} < \frac{H - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} < u_{1-\alpha/2}\right) = 1 - \alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили стандартизованного нормального распределения на уровнях $\alpha/2$ и $1 - \alpha/2$ соответственно. Преобразуя неравенство под знаком вероятности, запишем:

$$P\left(H - u_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < p_1 - p_2 < \right.$$

$$\left. < H + u_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right) = 1 - \alpha.$$

Это выражение ещё не даёт интервальной оценки разности вероятностей $p_1 - p_2$, так как левая и правая части неравенства под знаком вероятности содержат эти параметры. Подставляя в указанные части неравенства вместо неизвестных точных значений p_1 и p_2 их эффективные оценки $H_1 = \frac{K_1}{n_1}$ и $H_2 = \frac{K_2}{n_2}$, получаем следующий интервал для разности вероятностей $p_1 - p_2$:

$$\left(H_1 - H_2 - u_{1-\alpha/2}\sqrt{\frac{H_1(1-H_1)}{n_1} + \frac{H_2(1-H_2)}{n_2}}; \right.$$

$$\left. H_1 - H_2 + u_{1-\alpha/2}\sqrt{\frac{H_1(1-H_1)}{n_1} + \frac{H_2(1-H_2)}{n_2}} \right),$$

являющийся доверительным на уровне значимости α.

Указанные границы доверительного интервала получены в результате аппроксимации и могут использоваться лишь при достаточно больших объёмах наблюдений n_1 и n_2.

Контрольные вопросы и задачи

1. Сформулируйте теорему Муавра–Лапласа.
2. Какую центральную статистику используют при построении доверительного интервала неизвестного параметра p биномиального распределения? По какому закону она распределена при больших объёмах выборки?
3. Выведите формулу расчёта одностороннего доверительного интервала для параметра p биномиального распределения.
4. Какую центральную статистику используют при построении доверительного интервала для разности $p_1 - p_2$ параметров биномиального распределения? По какому закону она распределена при больших объёмах выборки?

Глава 4. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

§ 14. Основные понятия и определения

В практических задачах часто требуется проверить то или иное предположение относительно каких-нибудь свойств закона распределения наблюдаемой случайной величины X. Для проверки этого предположения исследователь проводит эксперимент, в результате которого получает реализацию $x_1, \ldots, x_n$ случайной выборки $X_1, \ldots, X_n$ из генеральной совокупности X. По этим данным ему нужно дать ответ на вопрос: согласуется ли его гипотеза с результатами эксперимента или нет? Другими словами, исследователю нужно решить, можно ли принять выдвинутую гипотезу или её нужно отклонить как противоречащую результатам эксперимента.

Любое предположение относительно параметров или закона распределения наблюдаемой случайной величины (или нескольких величин) называется *статистической гипотезой*. Проверяемую статистическую гипотезу также называют *основной*, или *нулевой*, статистической гипотезой и, как правило, обозначают H_0.

Наряду с проверяемой статистической гипотезой H_0 выдвигают также конкурирующую гипотезу, противоречащую H_0. Конкурирующая гипотеза называется *альтернативной* и, как правило, обозначается H_1 или H'. Если в результате статистического анализа делается вывод, что основная гипотеза H_0 должна быть отвергнута, то решение принимается в пользу альтернативной гипотезы H'. В простейшем случае альтернативная гипотеза – это отрицание основной гипотезы.

Статистическая гипотеза H_0 называется *простой*, если она однозначно определяет параметр или распределение наблюдаемой случайной величины X. В противном случае гипотеза H_0 называется *сложной*.

Если статистическая гипотеза H_0 представляет утверждение о некотором параметре θ известного распределения случайной вели-

чины X, то гипотеза называется *параметрической*. В противном случае – *непараметрической*.

Пример 4.1. При исследовании качества выпускаемой предприятием продукции проведено обследование 100 случайно отобранных изделий. Оказалось, что шесть из них имеют брак. Пусть случайная величина X – число бракованных изделий в партии из 1000 изделий, выпущенных тем же предприятием. Относительно случайной величины X могут быть сформулированы, например, следующие предположения:

1) случайная величина X имеет биномиальное распределение $B(1000; 0{,}06)$;

2) случайная величина X имеет биномиальное распределение $B(1000; p)$, где $0{,}04 \le p \le 0{,}08$;

3) математическое ожидание случайной величины X равно 70;

4) дисперсия случайной величины X не более 50;

5) вероятность того, что во всей партии будет более 80 бракованных изделий, не превосходит 90%;

6) вероятность того, что во всей партии будет ровно 60 бракованных изделий, не менее 5%.

Запишем эти гипотезы формально:

1) $H_0 : X \sim B(1000; 0{,}06)$;

2) $H_0 : X \sim B(1000; p), \quad 0{,}04 \le p \le 0{,}08$;

3) $H_0 : m_X = 70$;

4) $H_0 : d_X \le 50$;

5) $H_0 : P(X > 80) \le 0{,}9$;

6) $H_0 : P(X = 60) \ge 0{,}05$.

Все приведённые гипотезы являются параметрическими, поскольку распределение случайной величины X известно априорно из условий эксперимента, а все гипотезы связаны так или иначе с неизвестным параметром p биномиального распределения. Гипотезы 1 и 3 – простые, поскольку содержат утверждения, однозначно определяющие значение оцениваемого параметра.

Пример 4.2. Исследуется качество производства элементов интегральной микросхемы на двух технологических линиях. Мерой качества производства считается дисперсия размера элементов. Результаты выборочного наблюдения размеров выпущенных элементов на двух технологических линиях приведены в примере 3.4. Пусть случайные величины X_1 и X_2 – размеры элементов микросхем на первой и второй линиях соответственно. Относительно этих случайных величин могут быть сформулированы, например, следующие предположения:

1) размер элементов микросхемы, произведённых на первой линии, имеет нормальное распределение;

2) размер элементов микросхемы, произведённых на второй линии, распределён по закону $N(0{,}25;\ 0{,}01)$;

3) математические ожидания размеров элементов микросхем, произведённых на обеих линиях, равны;

4) качество производства элементов микросхем на второй линии выше, чем на первой.

Запишем эти гипотезы формально:

1) $H_0 : X_1 \sim N(m_1, \sigma_1)$;

2) $H_0 : X_2 \sim N(0{,}25; 0{,}01)$;

3) $H_0 : m_1 = m_2$;

4) $H_0 : \sigma_1^2 > \sigma_2^2$.

Здесь гипотезы 3 и 4 являются параметрическими, 1 и 2 – непараметрическими. Гипотезы 2 и 3 – простые, 1 и 4 – сложные.

Статистическое решение, т.е. решение о принятии или отклонении основной гипотезы H_0, проводится в соответствии с некоторым критерием.

Статистическим критерием, или *решающим правилом*, при проверке статистической гипотезы H_0 называется правило, в соответствии с которым гипотеза H_0 принимается или отклоняется.

Статистическая гипотеза – всегда утверждение о свойствах наблюдаемой генеральной совокупности, а задача проверки статистической гипотезы состоит в проверке соответствия результатов эксперимента $x_1, \ldots, x_n$ выдвинутой гипотезе. Иными словами, задача проверки статистической гипотезы состоит в ответе на во-

прос: *могло ли случиться так, что выборка $x_1, \ldots, x_n$ была получена из генеральной совокупности с указанными в гипотезе свойствами*?

Как правило, статистический критерий связывают с некоторой статистикой Z, являющейся функцией случайной выборки $X_1, \ldots, X_n$. Эта статистика служит мерой того, насколько наблюдаемые выборочные значения могли быть получены из генеральной совокупности с указанными в основной гипотезе свойствами. Вопрос о том, какую статистику Z следует взять для проверки той или иной статистической гипотезы, не имеет однозначного ответа. Это может быть любая статистика, удовлетворяющая двум требованиям:

1) закон распределения $F_Z(z \mid H_0)$ при условии истинности основной гипотезы H_0 должен быть известен;

2) распределение статистики Z должно быть чувствительно к факту справедливости основной или альтернативной гипотезы, т.е. законы распределения $F_Z(z \mid H_0)$ и $F_Z(z \mid H')$ должны существенно различаться.

Для реализации $x_1, \ldots, x_n$ случайной выборки $X_1, \ldots, X_n$, статистика Z примет реализацию z. Предположим, что гипотеза H_0 верна. В связи с тем, что закон распределения статистики Z при условии истинности основной гипотезы H_0 известен, то возможно рассчитать вероятность её попадания в некоторую окрестность точки z. Если эта вероятность высока, это означает, что ничто не противоречит предположению об истинности гипотезы H_0. Если же эта вероятность мала или близка к нулю, то это может означать один из двух вариантов:

1) в условиях основной гипотезы H_0 произошло маловероятное или практически невозможное событие;

2) статистика Z на самом деле имеет некоторый другой закон распределения, отличный от $F_Z(z \mid H_0)$, при котором вероятность её попадания в окрестность точки z много больше нуля – это означает, что предположение об истинности гипотезы H_0 сделано неверно.

Статистика $Z = Z(X_1, ..., X_n)$, на основе реализации которой $z = Z(x_1, ..., x_n)$ выдвигается статистическое решение, называется

статистикой критерия (*test statistic*). Реализация статистики критерия $z = Z(x_1, ..., x_n)$, рассчитанная для выборки x_1, …, x_n, называется *выборочным значением статистики критерия*.

Проверка статистических гипотез основывается на принципе, в соответствии с которым *маловероятные события относительно статистики критерия Z считаются невозможными*. В соответствии с этим принципом, если вероятность попадания статистики критерия Z в окрестность рассчитанного выборочного значения z мала, то должен выбираться второй вариант, т.е. основная гипотеза H_0 должна отклоняться.

Область Ω_0 наиболее вероятных значений статистики критерия Z, при попадании выборочных значений z в которую основная гипотеза H_0 принимается, называется *областью допустимых значений* (*region of acceptance*) статистики критерия Z.

Область Ω' маловероятных значений статистики критерия Z, при попадании в которую выборочных значений z основная гипотеза H_0 отклоняется, называется *критической областью* (*critical region*) значений статистики критерия Z. Множество $\Omega_0 \cup \Omega'$ должно являться множеством всех возможных значений статистики критерия Z.

Из определений области допустимых значений и критической области следует статистический критерий проверки гипотезы H_0: *если выборочное значение статистики критерия* $z \in \Omega_0$, *то основная гипотеза* H_0 *принимается; если выборочное значение статистики критерия* $z \in \Omega'$, *то основная гипотеза* H_0 *отвергается*.

Пусть для выборки x_1, …, x_n статистика критерия Z приняла выборочное значение z, лежащее в критической области Ω'. В соответствии со статистическим критерием основная гипотеза H_0 должна быть отвергнута. Однако событие $z \in \Omega'$, хоть и с малой вероятностью, но всё же могло произойти в условиях основной гипотезы H_0. Если это так, то статистическое решение об отклонении гипотезы H_0 будет ошибочным.

С другой стороны, если для выборки x_1, …, x_n статистика критерия Z приняла выборочное значение z, лежащее в области допустимых значений Ω_0, это могло случиться как в условиях основной гипотезы H_0 (с высокой вероятностью), так и в условиях альтерна-

тивной гипотезы H' (с малой вероятностью). В соответствии со статистическим критерием основная гипотеза H_0 принимается. Если же событие $z \in \Omega_0$ на самом деле произошло в условиях альтернативной гипотезы H', то статистическое решение о принятии гипотезы H_0 также будет ошибочным. В обоих случаях говорят об ошибках принятия статистического решения.

Ошибкой 1-го рода (*type* I *error*) при принятии статистического решения называется событие, состоящее в том, что основная гипотеза H_0 отклоняется, в то время как она верна.

Ошибкой 2-го рода (*type* II *error*) при принятии статистического решения называется событие, состоящее в том, что основная гипотеза H_0 принимается, в то время как верна альтернативная гипотеза H'.

Пример 4.3. Наблюдаемый объект может быть либо своим, либо объектом противника. Система обнаружения относит объект к одному из классов по результатам нескольких замеров определённых характеристик. Основная гипотеза H_0: объект свой; альтернативная гипотеза H': объект чужой. В чём состоят ошибки первого и второго рода при принятии статистического решения на основе замеров характеристик объекта?

Результат замера определённой характеристики объекта является случайной величиной вследствие погрешности измерительного прибора, влияния на результат измерения внешних случайных факторов или вследствие иных причин. Однако вывод о том, является ли объект своим или чужим, должен проводиться на основе истинных значений этих характеристик. Для этой цели выдвигается статистическая гипотеза.

Ошибка первого рода возникнет, если в результате проверки статистического критерия будет принято решение о том, что характеристики объекта соответствуют объекту противника, в то время как на самом деле объект является своим («уничтожен свой»).

Ошибка второго рода возникнет, если в результате проверки статистического критерия будет принято решение о том, что характеристики объекта соответствуют своему объекту, в то время как

на самом деле объект является объектом противника («пропущен чужой»).

Пример 4.4. Технология производства элемента интегральной микросхемы удовлетворяет производственным нормам, если вероятность брака в элементе не более 0,01. Соответствие производственным нормам проводится на основе выборочного наблюдения 1000 элементов. Если не более чем 15 элементов имеют брак, то считается, что производственные нормы соблюдены. В противном случае делается вывод о несоответствии технологии производства нормам.

Пусть p – вероятность брака в элементе интегральной микросхемы. Сформулируем основную и альтернативную гипотезы:

$$H_0 : p \le 0,01,$$

$$H' : p > 0,01.$$

Ответить на следующие вопросы:

1) какая статистика критерия используется в данной задаче, каковы её распределение и область значений;

2) какое решающее правило для проверки основной гипотезы используется в данной задаче, какова область допустимых значений и критическая область;

3) в чём состоят ошибки первого и второго рода?

По условию задачи статистическое решение принимается на основе значения случайной величины Z – числа бракованных элементов в серии из 1000. Таким образом, случайная величина Z является статистикой критерия. Очевидно, что $Z \sim B(1000, p)$. Возможные значения статистики Z: 0, 1, …, 1000.

Решающее правило: если выборочное значение статистики $z \le 15$, то H_0 принимается; если $z > 15$, то H_0 отклоняется. Таким образом, область допустимых значений $\Omega_0 = \{0, ..., 15\}$, критическая область $\Omega' = \{16, ..., 1000\}$.

Ошибка первого рода возникнет, если число бракованных элементов в выборке из 1000 будет более 15 (гипотеза H_0 будет отвергнута), при этом вероятность брака в отдельном элементе

$p \le 0,01$, т.е. будет принято решение о несоответствии производственным нормам, в то время как на самом деле соответствие есть.

Ошибка второго рода возникнет, если число бракованных элементов в выборке из 1000 окажется не более 15 (гипотеза H_0 будет принята), при этом вероятность брака в отдельном элементе $p > 0,01$, т.е. будет принято решение о соответствии производственным нормам, в то время как на самом деле соответствия нет.

Уровнем значимости (*significance level*) α при проверке статистической гипотезы называется вероятность ошибки первого рода:

$$\alpha = P\left(Z \in \Omega' \mid H_0\right). \qquad (4.1)$$

Вероятность β ошибки второго рода равна

$$\beta = P\left(Z \in \Omega_0 \mid H'\right). \qquad (4.2)$$

С уменьшением вероятности ошибки первого рода возрастает вероятность ошибки второго рода и наоборот. Это означает, что при выборе критической области и области допустимых значений статистики критерия должен соблюдаться определённый компромисс.

Пусть основная H_0 и альтернативная H' гипотезы являются простыми. Пусть статистика критерия Z при условии истинности основной гипотезы H_0 имеет распределение $F_Z(z \mid H_0) \sim N(m_1, \sigma_1)$, а при условии истинности H' – распределение $F_Z(z \mid H') \sim N(m_2, \sigma_2)$. В качестве критической области Ω' выберем хвосты распределения $f_Z(z \mid H_0)$ (как наименее вероятные области значений статистики Z при условии истинности гипотезы H_0), площадь каждого из которых равна $\alpha/2$ (рис. 4.1). Вероятность попадания статистики критерия Z, имеющей распределение $f_Z(z \mid H_0)$, в критическую область, таким образом, равна вероятности ошибки первого рода α. Вероятность ошибки второго рода β равна площади под графиком функции плотности распределения $f_Z(z \mid H')$ внутри области допустимых значений Ω_0. Из графиков видно, что, уменьшая ширину области допустимых значений, площадь α будет увеличиваться, в то время как площадь β уменьшаться, и наоборот.

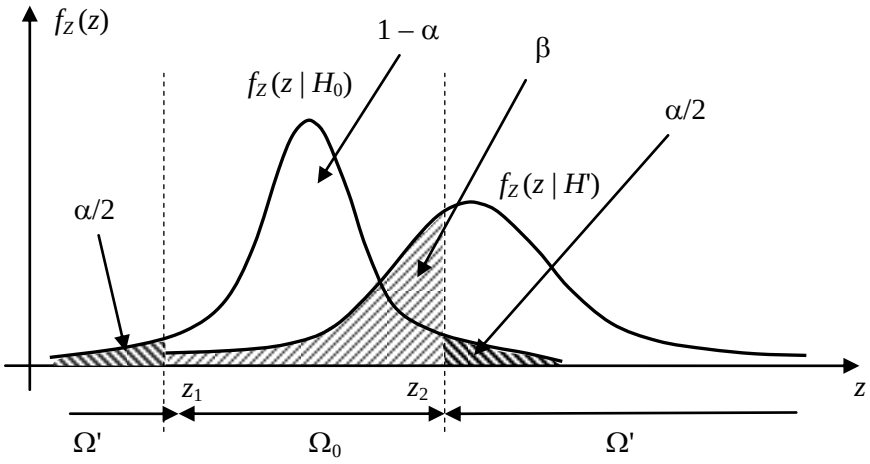

Рис. 4.1. Плотности распределения вероятностей статистики критерия Z при условии истинности основной и альтернативной гипотез

Точки на оси значений статистики критерия z, разделяющие область допустимых значений Ω_0 и критическую область Ω', называются *критическими*. На рис. 4.1 это точки z_1 и z_2, являющиеся квантилями распределения $f_Z(z \mid H_0)$ на уровнях $\alpha/2$ и $1-\alpha/2$ соответственно.

В случае если основная и альтернативная гипотезы H_0 и H' – простые, величина $\mu = 1-\beta$ называется *мощностью критерия* (*power of test*).

При заданном значении вероятности α ошибки первого рода выбор критической области Ω' может быть сделан неоднозначно. Единственное требование, предъявляемое к критической области, состоит в том, что площадь под графиком функции плотности известного распределения статистики критерия $f_Z(z \mid H_0)$ в критической области должна быть равна заданной вероятности α. Однако критерии, использующие различные критические области, будут иметь, вообще говоря, различные вероятности β ошибок второго рода (рис. 4.2).

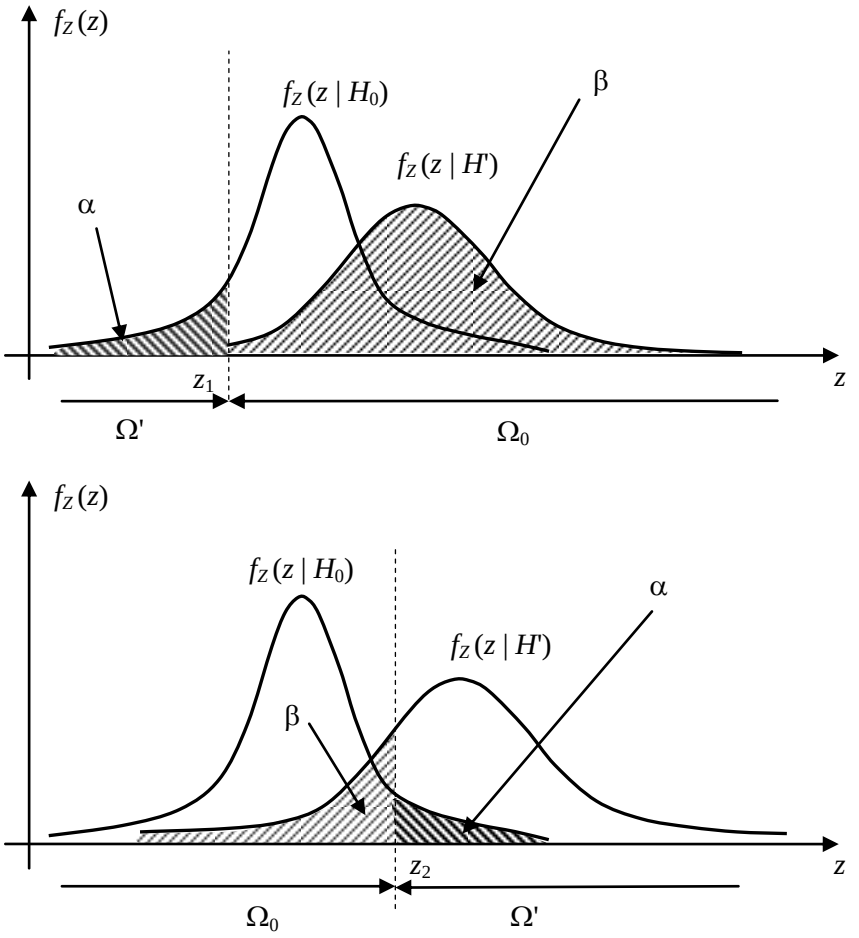

Рис. 4.2. Различные варианты выбора критической области

Наилучшей критической областью (*НКО*) называют критическую область, которая при заданном уровне значимости α обеспечивает минимальную вероятность β ошибки второго рода. Критерий, использующий наилучшую критическую область, имеет максимальную мощность.

Если альтернативная гипотеза является сложной, т.е. не определяет однозначно функцию распределения $F_X(x)$ генеральной совокупности X, а следовательно, и функцию распределения статистики критерия $F_Z(z \mid H')$, а определяет её с точностью до значения некоторого параметра θ, то вводят *функцию мощности критерия* μ(θ) как функцию параметра θ. Значение функции мощности критерия μ(θ) в точке θ определяется как

$$\mu(\theta) = 1 - \beta(\theta), \qquad (4.3)$$

где β(θ) – вероятность ошибки второго рода при условии, что неизвестный параметр принял значение θ, $\theta \in \Theta$, где Θ – область возможных значений параметра θ, определяемая альтернативной гипотезой H'.

Пример 4.5. В условиях примера 4.4 выдвигаются следующие основная и альтернативная гипотезы относительно вероятности p брака в элементе интегральной микросхемы:

$$H_0 : p = 0{,}01,$$

$$H' : p > 0{,}01.$$

Построить функцию мощности статистического критерия: если выборочное значение z статистики критерия Z – числа бракованных изделий из $n = 1000$ – не более 12, то H_0 принимается; если z более 12, то H_0 отвергается.

Запишем выражение для вероятности β ошибки второго рода при условии, что вероятность брака $p = p_1$, где $p_1 \in (0{,}01; \infty)$:

$$\beta(p_1) = P\left(Z \in \Omega_0 \mid p = p_1\right).$$

Статистика критерия Z при условии, что $p = p_1$, имеет биномиальное распределение $B(1000, p_1)$. Согласно теореме Муавра–Лапласа, при больших n биномиальное распределение может быть аппроксимировано нормальным:

$$Z|_{p=p_1} \sim N\left(m_Z(p_1), \sigma_Z(p_1)\right),$$

где $m_Z(p_1) = np_1$ и $\sigma_Z(p_1) = \sqrt{np_1(1-p_1)}$.

Учитывая, что область допустимых значений статистики критерия $\Omega_0 = \{0, ..., 12\}$, запишем:

$$\beta(p_1) = P(0 \le Z \le 12 \mid p = p_1) = P(0 \le \sigma_Z(p_1)U + m_Z(p_1) \le 12) =$$

$$= P\left(-\frac{m_Z(p_1)}{\sigma_Z(p_1)} \le U \le \frac{12 - m_Z(p_1)}{\sigma_Z(p_1)} \right) =$$

$$= P\left(\frac{-np_1}{\sqrt{np_1(1-p_1)}} \le U \le \frac{12 - np_1}{\sqrt{np_1(1-p_1)}} \right) =$$

$$= \Phi\left(\frac{12 - np_1}{\sqrt{np_1(1-p_1)}} \right) - \Phi\left(\frac{-np_1}{\sqrt{np_1(1-p_1)}} \right),$$

где $U \sim N(0,1)$ – стандартизованная нормально распределённая случайная величина, а Ф – функция Лапласа. Вычисляя с помощью таблиц математической статистики вероятность $\beta(p_1)$ для нескольких значений $p_1 \in (0,01; \infty)$, строим функцию мощности критерия $\mu(p_1) = 1 - \beta(p_1)$ поточечно (рис. 4.3).

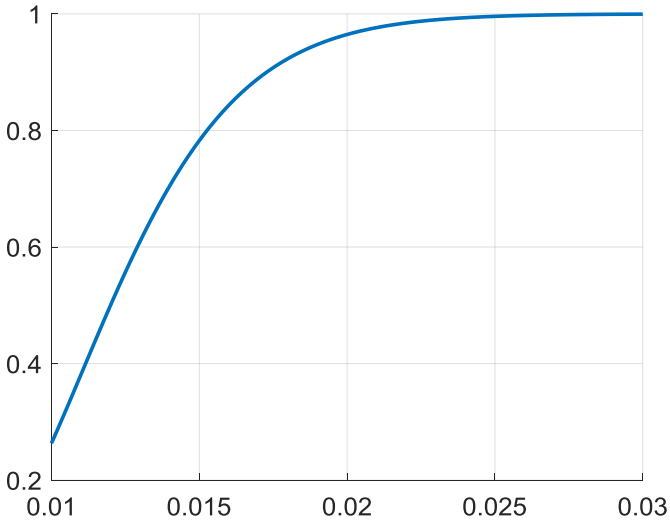

Рис. 4.3. График функции мощности критерия $\mu(p_1)$

Вероятность ошибки первого рода:

$$\alpha = P(Z \in \Omega' \mid H_0) = P(Z > 12 \mid p = 0{,}01) =$$
$$= 1 - \beta(0{,}01) = \mu(0{,}01) \approx 0{,}26.$$

Из графика функции мощности видно, что предложенный в задаче критерий имеет следующее свойство: с увеличением истинного значения вероятности p брака в элементе микросхемы от 0,01 до 1 (область значений p, определяемая гипотезой H') вероятность β ошибки второго рода при фиксированной вероятности ошибки первого рода $\alpha \approx 0{,}26$ уменьшается практически до нуля. В самом деле, с увеличением вероятности p брака вероятность ошибочно принять гипотезу H_0, используя предложенный критерий, стремится к нулю.

Функция мощности имеет важное значение в задачах, связанных с оценкой необходимого объёма выборки для обеспечения требуемой вероятности ошибки второго рода принятия статистического решения при заданной вероятности ошибки первого рода.

Контрольные вопросы и задачи

1. Что такое статистическая гипотеза?
2. Какую статистическую гипотезу называют параметрической?
3. Какую гипотезу называют основной, альтернативной, простой, сложной?
4. Что такое статистическое решение?
5. Что такое статистический критерий?
6. Что называется статистикой критерия при проверке статистической гипотезы? Какими свойствами должна обладать статистика критерия?
7. Какой принцип лежит в основе проверки статистических гипотез с использованием статистики критерия?
8. Что называется областью допустимых значений статистики критерия?
9. Что называется критической областью значений статистики критерия?
10. Что такое критические точки?

11. Сформулируйте критерий проверки статистических гипотез с использованием статистики критерия.

12. В чём состоит ошибка первого рода, второго рода?

13. Что такое уровень значимости при проверке статистической гипотезы?

14. Что называют мощностью критерия?

15. Какую критическую область называют наилучшей?

16. Какую функцию называют функцией мощности критерия? Как строится эта функция?

17. При каком значении параметра p_1 функция мощности в примере 4.5 принимает значение 0,5? Что показывает это значение?

18. Как изменится функция мощности в примере 4.5, если объём выборки будет $n = 100$?

19. Постройте функцию мощности критерия в условиях примера 4.5, если альтернативная гипотеза имеет вид $H': p < 0,01$.

§ 15. Алгоритм проверки статистических гипотез

Алгоритм проверки простой статистической гипотезы включает следующие этапы.

1. Сформулировать проверяемую гипотезу H_0 и альтернативную гипотезу H'. Гипотезы формулируются, исходя из условия задачи и особенностей рассматриваемой проблемной области.

2. Выбрать уровень значимости α, на котором будет сделано статистическое решение. Уровень значимости выбирается исследователем как допустимая вероятность ошибки первого рода при принятии статистического решения. Обычно уровень значимости выбирается небольшим, например $\alpha = 0,1$ или $\alpha = 0,01$, однако следует помнить, что выбор слишком малого уровня значимости приведёт к увеличению вероятности ошибки второго рода при принятии статистического решения.

3. Выбрать статистику критерия Z для проверки гипотезы H_0. Для большинства встречающихся на практике статистических гипотез H_0 выражение для статистики критерия Z, обеспечивающей минимальное или близкое к минимальному значение вероятности ошибки второго рода при фиксированном уровне значимости, из-

вестно. От исследователя, как правило, не требуется придумывать оригинальное выражение для расчёта статистики критерия.

4. Найти закон распределения $f_Z(z \mid H_0)$ выбранной статистики критерия Z при условии истинности основной гипотезы H_0. Законы распределения большинства используемых на практике статистик критерия также известны.

5. Построить область допустимых значений Ω_0 и критическую область Ω'. Критическая область Ω' зависит от вида статистики критерия Z, альтернативной гипотезы H' и уровня значимости α.

Простая основная параметрическая гипотеза имеет вид $H_0 : \theta = \theta_0$, где θ – неизвестный параметр генеральной совокупности; θ_0 – некоторая константа из области возможных значений параметра θ. Для такой основной гипотезы возможны следующие варианты формулировок альтернативных гипотез:

а) $H' : \theta < \theta_0$;

б) $H' : \theta > \theta_0$;

в) $H' : \theta \neq \theta_0$.

Как правило, наилучшая критическая область представляет собой область маловероятных значений статистики критерия в хвостах распределения $f_Z(z \mid H_0)$. Если критическая область расположена в левом хвосте распределения $f_Z(z \mid H_0)$, то такая критическая область называется *левосторонней*, если в правом хвосте – то *правосторонней*, если в обоих хвостах – то *двусторонней* (рис. 4.4). В случае двусторонней критической области площади каждого из хвостов, как правило, выбираются равными. Уровень значимости α определяет ширину критической области.

6. Вычислить выборочное значение z статистики критерия Z на основе имеющихся выборочных наблюдений из генеральной совокупности.

7. Принять статистическое решение, используя решающее правило: если выборочное значение статистики критерия $z \in \Omega_0$, то основная гипотеза H_0 принимается; если выборочное значение статистики критерия $z \in \Omega'$, то основная гипотеза H_0 отклоняется в пользу альтернативной гипотезы H'.

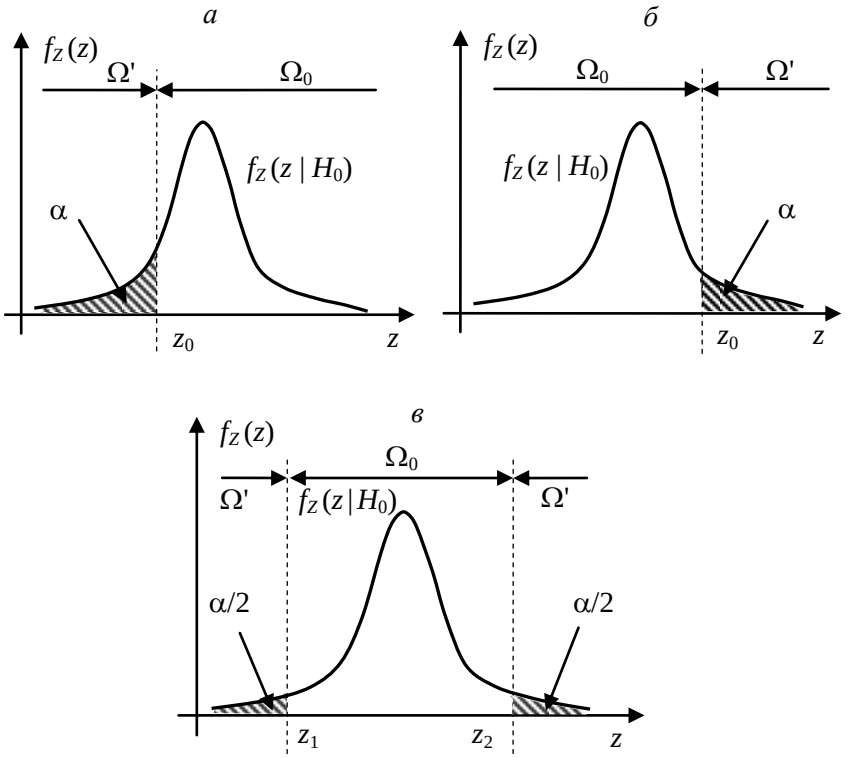

Рис. 4.4. Типы критических областей:
а – левосторонняя; *б* – правосторонняя; *в* – двусторонняя

Иногда при использовании статистических пакетов для проверки гипотез процедура статистического анализа не возвращает в явном виде выборочное значение *z* статистики критерия *Z*. В этом случае статистическое решение принимается на основе так называемого значения *p-value*.

Если альтернативная гипотеза имеет вид $H': \theta < \theta_0$ или $H': \theta > \theta_0$, то значение *p-value* – это площадь под графиком функции плотности распределения статистики критерия $f_Z(z \mid H_0)$, расположенная левее / правее выборочного значения статистики критерия *z*:

$$H': \theta < \theta_0 \Rightarrow p\text{-}value = F_Z(z \mid H_0),$$
$$H': \theta > \theta_0 \Rightarrow p\text{-}value = 1 - F_Z(z \mid H_0).$$

Иными словами, *p-value* – вероятность того, что статистика критерия Z примет более «экстремальные» значения в левом / правом хвосте критической области, чем рассчитанное для данной выборки значение z.

Если альтернативная гипотеза имеет вид $H': \theta \neq \theta_0$, то *p-value* рассчитывается по следующей формуле:

$$H': \theta \neq \theta_0 \Rightarrow p\text{-}value = 2\min\left(F_Z(z \mid H_0), 1 - F_Z(z \mid H_0)\right).$$

В этом случае *p-value* – вероятность того, что статистика критерия Z примет более «экстремальные» значения, чем z, в любом из хвостов двусторонней критической области.

Если значение *p-value* мало, это свидетельствует о том, что выборочное значение статистики критерия z уже приняло довольно «экстремальное» значение, что может говорить о противоречии выборочных данных основной гипотезе. Если значение *p-value* велико, то оснований отклонять основную гипотезу нет.

При использовании значения *p-value* критерий проверки статистической гипотезы формулируется следующим образом: *если значение p-value больше уровня значимости* α*, то основная гипотеза* H_0 *принимается; если значение p-value меньше уровня значимости* α*, то основная гипотеза* H_0 *отклоняется*.

Если основная гипотеза H_0 отклоняется, то делается вывод, что выборочные наблюдения противоречат основной гипотезе; если же H_0 принимается, то выборочные данные могли быть получены из генеральной совокупности со свойствами, указанными в H_0, что, впрочем, не означает, что генеральная совокупность в самом деле имеет эти свойства.

Пример 4.6. Для лечения больного врачи применяют инъекции биоактивного препарата. При этом проводится ежедневный анализ крови на содержание в ней нитроплазмидов, который дает случайное значение x в силу погрешности используемой методики измерения. Среднеквадратичное отклонение этого показателя известно и равно 0,35 ед. Средний уровень нитроплазмидов в крови, кото-

рый допускает продолжение лечения, составляет 9 ед. При превышении этого уровня лечение должно быть прекращено. Приводятся результаты анализов крови, выполненные в течение недели:

День	пн	вт	ср	чт	пт	сб	вс
x, ед.	9,22	9,06	8,85	8,70	9,35	9,26	9,33

На основании этих данных сделать вывод о необходимости прекращения лечения, предполагая, что измеряемый уровень нитроплазмидов в крови имеет нормальное распределение. Принять уровень значимости $\alpha = 0{,}01$.

Пусть случайная величина X – уровень нитроплазмидов в крови. По условию задачи $X \sim N(m; 0{,}35)$. Сформулируем основную и альтернативную гипотезы:

$$H_0 : m = 9,$$

$$H' : m > 9.$$

В альтернативной гипотезе выбран знак «больше», поскольку «критичной» является ситуация, когда математическое ожидание уровня нитроплазмидов в крови превосходит номинальный уровень 9 ед. – именно в этом случае будет принята альтернативная гипотеза и сделан вывод, что лечение должно быть прекращено. В случае же если в результате статистического анализа будет принята основная гипотеза, то оснований считать, что математическое ожидание уровня нитроплазмидов в крови превосходит номинальный уровень 9 ед., нет.

Известно, что для статистической модели $X \sim N(m; 0{,}35)$ с известной дисперсией в качестве статистики критерия используется статистика

$$Z = \frac{\bar{X} - m_0}{\sigma / \sqrt{n}},$$

распределённая по стандартизованному нормальному закону $N(0, 1)$ при условии истинности основной гипотезы H_0. Здесь $m_0 = 9$, $\sigma = 0{,}35$, $n = 7$.

Определим тип критической области из следующих соображений. Маловероятными значениями статистики критерия Z при условии истинности основной гипотезы H_0 являются значения в хвостах нормального распределения $N(0, 1)$. Однако сформулированной альтернативной гипотезе H' соответствуют лишь те значения статистики критерия Z, которые находятся в правом хвосте её распределения: если m значимо больше 9 ед., то с высокой вероятностью $\bar{X} > 9$, т.е. статистика Z будет принимать значения из правого хвоста распределения. Таким образом, область допустимых значений $\Omega_0 = (-\infty; z_0)$, а критическая область $\Omega' = [z_0; +\infty)$ является правосторонней (см. рис. 4.4, *б*), z_0 – критическая точка.

Рассчитаем выборочное значение статистики критерия:

$$z = \frac{\bar{x} - 9}{0{,}35/\sqrt{7}} \approx \frac{9{,}11 - 9}{0{,}13} \approx 0{,}83.$$

Критическая точка правосторонней критической области является квантилью стандартизованного нормального распределения на уровне $1 - \alpha$. По таблице квантилей (табл. П1 приложения) находим

$$z_0 = u_{0{,}99} \approx 2{,}33.$$

Так как $z \in \Omega_0$, то основная гипотеза H_0 должна приниматься. Таким образом, оснований считать, что средний уровень нитроплазмидов в крови превышает 9 ед., нет.

Контрольные вопросы и задачи

1. Перечислите шаги алгоритма проверки простой статистической гипотезы.
2. Почему нельзя выбирать слишком маленький уровень значимости при проверке статистических гипотез?
3. Какие типы критических областей вам известны?
4. Как определяется тип критической области?
5. Что такое значение *p-value*?
6. Как рассчитывается значение *p-value* для случая левосторонней / правосторонней / двусторонней критической области?
7. Сформулируйте критерий проверки статистической гипотезы на основе значения *p-value*.

§ 16. Проверка гипотез о параметрах нормально распределённой генеральной совокупности

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей нормальное распределение $N(m, \sigma)$. Ниже приводятся наилучшие по мощности статистики критерия для различных вариантов гипотез относительно параметров m и σ. Как правило, эти статистики связаны с эффективными оценками параметров, относительно которых выдвигаются гипотезы.

1. Гипотеза о значении математического ожидания при известной дисперсии (*one-sample z-test*):

$$H_0 : m = m_0 .$$

При условии истинности гипотезы H_0 случайная величина $\bar{X}$ имеет распределение $N\left(m_0, \dfrac{\sigma}{\sqrt{n}}\right)$ (см. пример 3.2). В качестве статистики критерия выберем стандартизованное среднее

$$Z = \frac{\bar{X} - m_0}{\sigma / \sqrt{n}}, \qquad (4.4)$$

которое при условии истинности H_0 распределено по стандартизованному нормальному закону $N(0, 1)$.

2. Гипотеза о значении математического ожидания при неизвестной дисперсии (*one-sample t-test*):

$$H_0 : m = m_0 .$$

В связи с тем, что σ не известно, статистику (4.4) здесь использовать нельзя. Вместо σ в (4.4) подставляется оценка S среднеквадратичного отклонения:

$$Z = \frac{\bar{X} - m_0}{S / \sqrt{n}}, \qquad (4.5)$$

при этом в условиях истинности гипотезы H_0 статистика Z будет иметь распределение Стьюдента с $n - 1$ степенью свободы (см. § 11).

3. Гипотеза о значении дисперсии (*или с.к.о.*) *при известном математическом ожидании* (*chi-squared test*):

$$H_0 : \sigma = \sigma_0 .$$

Эффективной оценкой дисперсии при известном математическом ожидании является статистика $S_0^2 \sim \frac{\sigma^2}{n}\chi^2(n)$ (см. § 11). В качестве статистики критерия выберем статистику

$$Z = \frac{nS_0^2}{\sigma_0^2}, \tag{4.6}$$

имеющую в условиях истинности гипотезы H_0 распределение хи-квадрат с n степенями свободы:

$$Z|_{H_0} \sim \chi^2(n).$$

4. Гипотеза о значении дисперсии (*или с.к.о.*) *при неизвестном математическом ожидании* (*chi-squared test*):

$$H_0 : \sigma = \sigma_0.$$

Эффективной оценкой дисперсии при неизвестном математическом ожидании является статистика $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$ (см. § 11). В качестве статистики критерия выберем статистику

$$Z = \frac{(n-1)S^2}{\sigma_0^2}, \tag{4.7}$$

имеющую в условиях истинности гипотезы H_0 распределение хи-квадрат с $n - 1$ степенями свободы:

$$Z|_{H_0} \sim \chi^2(n-1).$$

Рассмотрим теперь статистики критерия для гипотез, связанных с параметрами двух генеральных совокупностей. Пусть $x_{11}, ..., x_{1,n_1}$ и $x_{21}, ..., x_{2,n_2}$ – выборки объёмов n_1 и n_2 из нормально распределённых генеральных совокупностей $N(m_1, \sigma_1)$ и $N(m_2, \sigma_2)$ соответственно.

5. Гипотеза о равенстве математических ожиданий при известных дисперсиях (*two-sample z-test*):

$$H_0 : m_1 = m_2.$$

Поскольку средние значения выборок имеют нормальные распределения $\bar{X}_1 \sim N\left(m_1, \sigma_1/\sqrt{n_1}\right)$, $\bar{X}_2 \sim N\left(m_2, \sigma_2/\sqrt{n_2}\right)$, то неслож-

но показать (см. § 11), что при условии истинности гипотезы H_0 статистика

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{4.8}$$

имеет стандартизованное нормальное распределение $N(0, 1)$.

6. Гипотеза о равенстве дисперсий (*или с.к.о.*) *при известных математических ожиданиях* (*two-sample F-test*):

$$H_0 : \sigma_1 = \sigma_2 .$$

В качестве статистики критерия используется отношение оценок дисперсий при известных математических ожиданиях

$$Z = F_0 \,|_{H_0} = \frac{S_{01}^2}{S_{02}^2}, \tag{4.9}$$

которое при условии истинности H_0 распределено по закону Фишера $F(n_1, n_2)$ (см. § 11).

7. Гипотеза о равенстве дисперсий (*или с.к.о.*) *при неизвестных математических ожиданиях* (*two-sample F-test*):

$$H_0 : \sigma_1 = \sigma_2 .$$

В качестве статистики критерия используется отношение оценок дисперсий при неизвестных математических ожиданиях

$$Z = F \,|_{H_0} = \frac{S_1^2}{S_2^2}, \tag{4.10}$$

которое при условии истинности H_0 распределено по закону Фишера $F(n_1 - 1, n_2 - 1)$ (см. § 11).

8. Гипотеза о равенстве математических ожиданий при неизвестных дисперсиях (*two-sample unpooled t-test*):

$$H_0 : m_1 = m_2 .$$

8а. Дисперсии генеральных совокупностей равны $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (это может быть известно априорно, исходя из особенностей предметной области, или в случае, если гипотеза $H_0 : \sigma_1 = \sigma_2$ при неизвестных математических ожиданиях принимается).

Объединённая оценка дисперсии σ^2 по двум выборкам имеет вид

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

При условии истинности H_0 статистика S^2 имеет распределение

$$S^2 \sim \frac{\sigma^2}{n_1 + n_2 - 2}\chi^2(n_1 + n_2 - 2).$$

Несложно показать (см. § 11), что статистика

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (4.11)$$

при условии истинности H_0 имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

8б. Оснований считать, что дисперсии генеральных совокупностей равны, нет (*критерий Уэлча*, *Welch's t-test*).

Для каждой из дисперсий вычисляются свои оценки S_1^2 и S_2^2. Статистика критерия имеет вид

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \qquad (4.12)$$

Показано, что при условии истинности H_0 статистика Z имеет распределение Стьюдента с числом степеней свободы, равным целой части от величины $\frac{1}{\nu}$, где

$$\nu = \frac{\left(\frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2/n_2}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_2 - 1}.$$

Основные статистики критерия, используемые при проверке статистических гипотез о параметрах нормально распределённой генеральной совокупности, и их законы распределения при условии истинности гипотезы H_0 приведены в табл. 4.1.

Таблица 4.1

Статистики критерия для проверки статистических гипотез о параметрах нормально распределённой генеральной совокупности

Основная гипотеза H_0	Математическое ожидание	Дисперсия	Статистика критерия Z	Закон распределения $f_Z(z \mid H_0)$
$H_0 : m = m_0$	Не известно	Известна	$\dfrac{\bar{X} - m_0}{\sigma / \sqrt{n}}$	$N(0,1)$
$H_0 : m = m_0$	Не известно	Не известна	$\dfrac{\bar{X} - m_0}{S / \sqrt{n}}$	$T(n-1)$
$H_0 : \sigma = \sigma_0$	Известно	Не известна	$\dfrac{nS_0^2}{\sigma_0^2}$	$\chi^2(n)$
$H_0 : \sigma = \sigma_0$	Не известно	Не известна	$\dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$
$H_0 : m_1 = m_2$	Не известны	Известны	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$H_0 : m_1 = m_2$	Не известны	Не известны, равны	$\dfrac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$T(n_1 + n_2 - 2)$
$H_0 : \sigma_1 = \sigma_2$	Известны	Не известны	$\dfrac{S_{01}^2}{S_{02}^2}$	$F(n_1, n_2)$
$H_0 : \sigma_1 = \sigma_2$	Не известны	Не известны	$\dfrac{S_1^2}{S_2^2}$	$F(n_1 - 1, n_2 - 1)$

Все приведённые выше выражения для статистик критерия и их законов распределения справедливы, если случайная выборка $X_1, ..., X_n$ или выборки $X_{11}, ..., X_{1,n_1}$ и $X_{21}, ..., X_{2,n_2}$ получены из нормально распределённых генеральных совокупностей. Однако поскольку все статистики основаны на оценках $\overline{X}$ и S^2, представляющих собой суммы случайных величин, то согласно центральной предельной теореме теории вероятностей распределение этих статистик при больших объёмах выборок будет близко к нормальному, даже если распределение каждого входящего в них слагаемого отлично от нормального. В то же время, если генеральная совокупность распределена нормально, то статистика S^2 имеет распределение хи-квадрат, которое при больших объёмах выборки также может быть аппроксимировано нормальным распределением. Это означает, что приведённые в табл. 4.1 законы распределения статистик критерия остаются справедливыми при больших объёмах выборок в случае распределения генеральной совокупности, отличного от нормального.

Пример 4.7. Точность работы станка-автомата проверяется по среднеквадратичному отклонению контролируемого размера деталей, которое не должно превышать $\sigma_0 = 0,25$ мм. Взята проба из $n = 25$ случайно отобранных деталей, получены следующие результаты измерений:

Размер деталей, мм	3,0–3,3	3,3–3,6	3,6–3,9	3,9–4,2	4,2–4,5
Частоты	3	5	10	6	1

Предполагая, что размер деталей имеет нормальное распределение, проверить на уровне значимости 10%, обеспечивает ли станок требуемую точность.

Рассчитаем вначале оценку среднеквадратичного отклонения размера деталей:

$$\overline{x} = \frac{1}{25}\left(3,15 \cdot 3 + ... + 4,35 \cdot 1\right) \approx 3,71;$$

$$s^2 = \frac{1}{24}\left((3{,}15-3{,}71)^2 \cdot 3 + \ldots + (4{,}35-3{,}71)^2 \cdot 1\right) \approx 0{,}1;$$

$$s = \sqrt{0{,}1} \approx 0{,}32.$$

Полученное значение превышает допустимое $\sigma_0 = 0{,}25$ мм, однако делать вывод о том, что станок не обеспечивает требуемую точность, ещё нельзя. Превышение допустимого значения среднеквадратичного отклонения могло быть связано с особенностями конкретной выборки, а не генеральной совокупности, из которой она получена (т.е. особенностями станка).

Сформулируем основную и альтернативную статистические гипотезы:

$$H_0 : \sigma = 0{,}25,$$

$$H' : \sigma > 0{,}25.$$

Если основная гипотеза H_0 будет отвергнута в пользу альтернативной H', это будет означать, что выборочные наблюдения противоречат утверждению о точности работы станка-автомата. Если же основная гипотеза H_0 будет принята, то оснований считать, что станок не обеспечивает требуемую точность, не будет.

Из табл. 4.1 находим, что статистика критерия для данной гипотезы имеет вид $Z = \dfrac{(n-1)S^2}{\sigma_0^2}$. Рассчитаем её выборочное значение:

$$z = \frac{24 \cdot 0{,}1}{0{,}0625} \approx 38{,}4.$$

Гипотеза H_0 должна отвергаться в пользу альтернативной гипотезы H', если среднеквадратичное отклонение $\sigma > \sigma_0 = 0{,}25$ мм. В этом случае с высокой вероятностью оценка S среднеквадратичного отклонения будет больше 0,25 мм, т.е. статистика критерия Z будет принимать значения, большие единицы. Это означает, что критическая область должна выбираться в области больших маловероятных при условии истинности H_0 значений, т.е. являться правосторонней.

Критическую точку при правосторонней критической области – квантиль распределения $\chi^2(24)$ на уровне $(1 - \alpha)$ – находим по таблице (табл. П3 приложения):

$$z_{0,9} \approx 33,2 .$$

Таким образом, область допустимых значений $\Omega_0 = (-\infty; 33,2)$, критическая область $\Omega' = [33,2; +\infty)$. Поскольку $z = 38,4 \in \Omega'$, то гипотеза H_0 должна быть отвергнута. Делаем вывод, что экспериментальные данные противоречат гипотезе о том, что станок-автомат обеспечивает требуемую точность.

Пример 4.8. В условиях примера 3.4 выдвигается предположение о том, что элементы микросхем, выпускаемые на первой технологической линии, имеют положительную систематическую погрешность в размере по сравнению с элементами, выпускаемыми на второй линии. Проверить эту гипотезу на уровне значимости $\alpha = 0,05$.

Сформулируем основную и альтернативную статистические гипотезы:

$$H_0 : m_1 = m_2,$$
$$H' : m_1 > m_2.$$

Если основная гипотеза H_0 будет отвергнута в пользу альтернативной H', это будет означать, что выборочные наблюдения размеров элементов противоречат утверждению об отсутствии систематического превышения размеров элементов, выпущенных на первой линии. Если же основная гипотеза H_0 будет принята, то оснований считать, что размер элементов, выпущенных на первой технологической линии, имеет систематическую положительную погрешность, не будет.

В связи с тем, что дисперсии размеров элементов не известны, проверим предварительно гипотезу о равенстве дисперсий на том же уровне значимости α:

$$G_0 : \sigma_1 = \sigma_2,$$
$$G' : \sigma_1 \neq \sigma_2.$$

Используем статистику критерия $Z = \dfrac{S_1^2}{S_2^2}$, выборочное значение которой равно

$$z \approx \frac{8,71 \cdot 10^{-5}}{9,39 \cdot 10^{-5}} \approx 0,93.$$

Объёмы выборок $n_1 = 70$, $n_2 = 70$. При условии истинности гипотезы G_0 статистика Z имеет распределение Фишера: $Z|_{G_0} \sim F(69,69)$.

Поскольку альтернативная гипотеза G' содержит знак неравенства, критическая область для статистики Z должна выбираться двусторонней. Критические точки – квантили распределения Фишера $F(69;69)$ на уровнях $\alpha/2$ и $(1-\alpha/2)$ – находим по таблице (табл. П5 приложения):

$$z_{0,025} \approx 1/1,61 \approx 0,62,$$

$$z_{0,975} \approx 1,61.$$

Таким образом, область допустимых значений $\Omega_0 = (0,62;1,61)$. Поскольку $z = 0,93 \in \Omega_0$, то оснований считать, что гипотеза G_0 не согласуется с экспериментальными данными, нет.

Считая, что дисперсия является мерой точности технологической линии, полученное статистическое решение означает, что нет оснований утверждать, что одна из технологических линий более точна по сравнению с другой.

Рассчитаем объединённую оценку дисперсии $\sigma_1^2 = \sigma_2^2 = \sigma^2$ по двум выборкам:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \approx \frac{69 \cdot 8,71 \cdot 10^{-5} + 69 \cdot 9,39 \cdot 10^{-5}}{70+70-2} =$$

$$= \frac{8,71+9,39}{2} \cdot 10^{-5} = 9,05 \cdot 10^{-5}$$

и выборочное значение статистики критерия

$$z = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx \frac{0,2514 - 0,2490}{0,0095 \cdot 0,169} \approx 1,51.$$

При условии истинности H_0 статистика Z имеет распределение Стьюдента $T(138)$, которое с высокой точностью может быть ап-

проксимировано стандартизованным нормальным распределением $N(0, 1)$.

Альтернативная гипотеза H' должна приниматься, если $m_1 > m_2$. В этом случае с высокой вероятностью выполнится неравенство $\overline{x}_1 > \overline{x}_2$, т.е. статистика Z примет положительные значения, причём тем большие, чем больше разница между $\overline{x}_1$ и $\overline{x}_2$. Таким образом, критическая область должна выбираться в области положительных маловероятных при условии истинности H_0 значений, т.е. являться правосторонней.

Критическая точка правосторонней критической области – квантиль стандартизованного нормального распределения на уровне $(1 - \alpha)$ – находим по таблице (табл. П1 приложения):

$$z_{0,95} \approx 1,65.$$

Таким образом, область допустимых значений $\Omega_0 = (-\infty; 1,65)$. Поскольку $z = 1,51 \in \Omega_0$, то гипотеза H_0 должна приниматься. Делаем вывод, что экспериментальные данные не противоречат гипотезе о том, что положительного смещения в размере элементов микросхем, выпускаемых на первой технологической линии, нет. Иными словами, по данной выборке наблюдений нет оснований считать, что это положительное смещение имеется.

В некоторых случаях для проверки параметрических статистических гипотез может быть использован *метод доверительных интервалов*. Пусть основная гипотеза $H_0 : \theta = \theta_0$, альтернативная гипотеза $H' : \theta \neq \theta_0$. Если для неизвестного параметра θ может быть построен доверительный интервал $(\theta_1; \theta_2)$, то проверка статистической гипотезы H_0 сводится к проверке попадания значения θ_0 в этот интервал. Критерий проверки гипотез при использовании метода доверительных интервалов состоит в следующем: *если* $\theta_0 \in (\theta_1; \theta_2)$, *то основная гипотеза* H_0 *должна приниматься, в противном случае – отклоняться*. Если альтернативная гипотеза H' имеет вид $H' : \theta < \theta_0$ или $H' : \theta > \theta_0$, то строится соответствующий односторонний доверительный интервал $(-\infty; \theta_2)$ или $(\theta_1; +\infty)$.

При проверке статистической гипотезы о равенстве математических ожиданий $H_0 : m_1 = m_2$ строится доверительный интервал для разности $m_1 - m_2$ (см. § 12). Если интервал накрывает 0, то основная гипотеза принимается, в противном случае – отклоняется.

При проверке статистической гипотезы о равенстве среднеквадратичных отклонений $H_0 : \sigma_1 = \sigma_2$ строится доверительный интервал для отношения σ_1^2 / σ_2^2 (см. § 12). Если интервал накрывает 1, то основная гипотеза принимается, в противном случае – отклоняется.

Пример 4.9. В условиях примера 4.8 проверить гипотезу о равенстве размеров элементов микросхем, выпускаемых на двух технологических линиях, методом доверительных интервалов на уровне значимости $\alpha = 0{,}05$.

Правосторонний доверительный интервал для разности математических ожиданий $m_1 - m_2$ при равных дисперсиях на уровне значимости α имеет вид (см. § 12)

$$\left((\bar{X}_1 - \bar{X}_2) - t_{1-\alpha}(n_1 + n_2 - 2) S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; +\infty \right).$$

Рассчитаем его границу:

$$\bar{x}_1 - \bar{x}_2 = 0{,}2514 - 0{,}2490 = 0{,}0024;$$

$$t_{0{,}95}(138) \cdot S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \approx 1{,}65 \cdot 0{,}0095 \cdot 0{,}169 \approx 0{,}0026;$$

$$(0{,}0024 - 0{,}0026; +\infty) = (-0{,}0002; +\infty).$$

Так как доверительный интервал накрывает 0, то основная гипотеза H_0 должна приниматься. Оснований считать, что есть положительное смещение размеров элементов, выпускаемых на первой технологической линии, нет, что совпадает с результатом, полученным в примере 4.8.

Контрольные вопросы и задачи

1. Объясните принцип выбора статистик критерия для оценки параметров нормально распределённой генеральной совокупности.

2. Возможно ли использование приведённых в табл. 4.1 статистик критерия для проверки статистических гипотез о параметрах генеральной совокупности, имеющей распределение, отличное от нормального?

3. В чём состоит метод доверительных интервалов при проверке статистических гипотез?

§ 17. Проверка гипотез о вероятности «успеха» в схеме Бернулли

При статистическом анализе данных, связанных с повторными независимыми испытаниями (схемой Бернулли), обычно рассматривают два вида задач: сравнение вероятности «успеха» p в одном испытании с заданным значением p_0 и сравнение вероятностей «успеха» в двух сериях испытаний.

Пусть проводится серия из n испытаний по схеме Бернулли, и случайная величина K – число «успехов». Тогда K имеет биномиальное распределение $K \sim B(n, p)$. Как известно, математическое ожидание $m_K = np$ и дисперсия $d_K = np(1-p)$. В соответствии с предельной теоремой Муавра–Лапласа при большом числе испытаний n статистика K имеет закон распределения, близкий к нормальному:

$$K \sim N\left(np, \sqrt{np(1-p)}\right).$$

Частота «успеха» $H = K/n$ также имеет нормальное распределение

$$H \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Для проверки статистической гипотезы (*one-proportion z-test*)

$$H_0 : p = p_0$$

в качестве статистики критерия используем стандартизованную частоту

$$Z = \frac{H - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \qquad (4.13)$$

которая при условии истинности H_0 имеет распределение $f_Z(z \mid H_0) \sim N(0,1)$.

Если альтернативная гипотеза $H': p \neq p_0$, то критическая область для статистики критерия выбирается двусторонней; если $H': p < p_0$ или $H': p > p_0$, то – лево- или правосторонней соответственно.

Пусть теперь проводится две серии испытаний и требуется проверить гипотезу о равенстве вероятностей «успехов» p_1 и p_2 в этих сериях (*two-proportion z-test*):

$$H_0 : p_1 = p_2 .$$

Частота «успеха» в первой серии $H_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$, во второй серии $H_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$, где n_1 и n_2 – число испытаний в первой и второй сериях соответственно. В силу композиционной устойчивости нормального распределения разность частот $H = H_1 - H_2$ также будет иметь нормальное распределение $H \sim N(m_H, \sigma_H)$, где

$$m_H = p_1 - p_2 ,$$

$$\sigma_H^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} .$$

При условии истинности H_0 (т.е. при $p_1 = p_2 = p$) стандартизованная разность частот (см. § 11)

$$Z = \frac{H_1 - H_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

имеет стандартизованное нормальное распределение $N(0, 1)$. Заменяя в знаменателе неизвестную истинную вероятность p на её эффективную оценку – агрегированную частоту

$$H = \frac{n_1 H_1 + n_2 H_2}{n_1 + n_2} ,$$

получим приближённое выражение для статистики критерия:

$$Z = \frac{H_1 - H_2}{\sqrt{H(1-H)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \qquad (4.14)$$

Подчеркнём, что указанная статистика может использоваться лишь при достаточно больших объёмах наблюдений n_1 и n_2.

Если альтернативная гипотеза $H': p_1 \neq p_2$, то критическая область для статистики критерия выбирается двусторонней; если $H': p_1 < p_2$ или $H': p_1 > p_2$, то – лево- или правосторонней соответственно.

Пример 4.10. При исследовании качества выпускаемой двумя предприятиями продукции проведено обследование 100 случайно отобранных изделий, произведённых каждым из предприятий. Среди изделий первого предприятия обнаружено 6 бракованных, среди изделий второго – 11. Можно ли утверждать на основе данного выборочного наблюдения, что процент брака в изделиях, выпускаемых обоими предприятиями, одинаков? Принять уровень значимости 10%.

Рассчитаем вначале оценки долей бракованных изделий:

$$h_1 = \frac{6}{100} = 0,06; \qquad h_2 = \frac{11}{100} = 0,11.$$

Полученные значения не равны, однако делать вывод о том, что процент брака среди изделий двух предприятий различен, ещё нельзя. Превышение процента бракованных изделий у второго предприятия могло быть связано с особенностями конкретных выборок, а не генеральных совокупностей, из которых они получены (т.е. особенностями технологических линий предприятий).

Сформулируем основную и альтернативную статистические гипотезы:

$$H_0: p_1 = p_2,$$
$$H': p_1 \neq p_2.$$

Если основная гипотеза H_0 будет отвергнута, это будет означать, что результаты наблюдений противоречат утверждению об одина-

ковом качестве производства изделий обоими предприятиями. Если же основная гипотеза H_0 будет принята, то оснований считать, что процент брака на предприятиях различен, не будет.

Рассчитаем выборочное значение статистики критерия:

$$h = \frac{6+11}{200} = 0,085;$$

$$z = \frac{0,06 - 0,11}{\sqrt{0,085 \cdot 0,915}\sqrt{0,02}} \approx \frac{-0,05}{0,04} = -1,25.$$

Так как альтернативная гипотеза содержит знак неравенства, то критическая область выбирается двусторонней. Используя таблицу квантилей стандартизованного нормального распределения (табл. П1 приложения), находим критические точки:

$$z_{0,05} \approx -1,65; \quad z_{0,95} \approx 1,65.$$

Таким образом, область допустимых значений $\Omega_0 = (-1,65; 1,65)$. Поскольку $z = -1,25 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Иными словами, утверждать, что процент брака в изделиях, выпускаемых обоими предприятиями, различен, нельзя. Превышение числа бракованных изделий среди выпущенных вторым предприятием обусловлено свойствами данных выборок, а не генеральных совокупностей, из которых они получены.

Контрольные вопросы и задачи

1. Какая статистика критерия используется при проверке статистической гипотезы о равенстве вероятности «успеха» в серии испытаний по схеме Бернулли фиксированному значению? Какой закон распределения имеет эта статистика при больших объёмах выборки?
2. Какая статистика критерия используется при проверке статистической гипотезы о равенстве вероятностей «успеха» в двух сериях испытаний по схеме Бернулли? Какой закон распределения имеет эта статистика при больших объёмах выборок?
3. Объясните принцип проверки статистических гипотез о вероятностях «успеха» в серии испытаний по схеме Бернулли методом доверительных интервалов.

Глава 5. КРИТЕРИИ СОГЛАСИЯ

§ 18. Проверка гипотез о виде распределения. Критерий Колмогорова

Статистические методы, изложенные в предыдущих главах, опираются на различные априорные допущения о виде исследуемой статистической модели. Например, основные формулы расчёта доверительных интервалов и статистик критерия для проверки параметрических статистических гипотез выведены в предположениях о нормальности распределения генеральной совокупности и независимости элементов наблюдаемой случайной выборки.

В практических приложениях может возникнуть вопрос о соответствии выборочных наблюдений предполагаемой статистической модели. Эти предположения могут быть сформулированы как статистические гипотезы и проверены с помощью статистических критериев.

Критериями согласия (*goodness-of-fit tests*) называют статистические критерии, предназначенные для проверки гипотез о виде распределения наблюдаемой генеральной совокупности. Критерии согласия отвечают на вопрос, насколько хорошо экспериментальные данные согласуются с предполагаемой статистической моделью генеральной совокупности.

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей неизвестное распределение $F_X(x)$. Рассмотрим задачу проверки статистической гипотезы о том, что функция распределения $F_X(x)$ совпадает с некоторой известной функцией $G(x)$. Сформулируем основную и альтернативную гипотезы:

$$H_0 : F_X(x) = G(x),$$

$$H' : F_X(x) \neq G(x).$$

Оценкой неизвестной функции распределения $F_X(x)$, рассчитанной по выборке $x_1, \ldots, x_n$, является эмпирическая функция распределения (ЭФР) $F_n^*(x)$ (см. § 3), которую можно рассматривать как

реализацию случайной ЭФР $\mathcal{F}_n^*(x)$ соответствующей случайной выборки $X_1, \ldots, X_n$. Поскольку $\mathcal{F}_n^*(x)$ – состоятельная оценка функции распределения $F_X(x)$, то при каждом фиксированном x случайная величина $\mathcal{F}_n^*(x)$ стремится по вероятности к значению функции распределения $F_X(x)$ в точке x при $n \to \infty$. Следовательно, при условии истинности основной гипотезы вероятность того, что рассогласование $\Delta\left(F_n^*(x), G(x)\right)$ между $F_n^*(x)$ и $G(x)$ примет достаточно большие значения, стремится к нулю с ростом объёма выборки n. Меру рассогласования между двумя распределениями можно выбрать многими способами, и в зависимости от этого выбора получаем различные статистики критерия для проверки гипотезы H_0.

Критерий Колмогорова (*one-sample KS-test*), называемый также *критерием Колмогорова–Смирнова* (*A.N. Kolmogorov, N.V. Smirnov*, 1933), основан на результатах сравнения ЭФР $F_n^*(x)$ с предполагаемой функцией распределения $G(x)$ с помощью метрики

$$\Delta\left(F_n^*(x), G(x)\right) = D_n = \sup_x \left|F_n^*(x) - G(x)\right|. \qquad (5.1)$$

Если функции $F_n^*(x)$ и $G(x)$ близки с точки зрения указанной метрики, то оснований отклонять основную гипотезу H_0 нет. Если расхождение между этими функциями велико, то распределение случайной величины X значимо отлично от предполагаемого распределения $G(x)$, следовательно, основная гипотеза H_0 должна быть отвергнута в пользу альтернативной.

В критерии Колмогорова используется статистика критерия

$$Z_n = \sqrt{n} D_n, \qquad (5.2)$$

для которой показано, что при условии истинности основной гипотезы H_0 её распределение не зависит от вида функции $G(x)$ и стремится при $n \to \infty$ к *распределению Колмогорова* с функцией распределения

$$K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}. \qquad (5.3)$$

Приближённо полагая при больших n ($n > 40$), что статистика критерия Z_n имеет распределение Колмогорова, для неё может быть

рассчитана любая квантиль, используя формулу (5.3) или таблицу квантилей распределения Колмогорова.

В случае истинности альтернативной гипотезы H' рассогласование D_n между ЭФР $F_n^*(x)$ и $G(x)$ при $n \to \infty$ будет отлично от нуля, причём с увеличением D_n статистика критерия Z_n более вероятно будет принимать большие значения. Следовательно, основная гипотеза H_0 должна отвергаться в области больших значений Z_n, т.е. критическая область должна выбираться правосторонней.

На практике для вычисления рассогласования D_n между ЭФР $F_n^*(x)$ и $G(x)$ по выборке $x_1, \ldots, x_n$ удобно использовать формулу

$$D_n = \max_{i=1,n} \left\{ \frac{i}{n} - G(x_{(i)}), G(x_{(i)}) - \frac{i-1}{n} \right\}, \qquad (5.4)$$

которую также можно записать в виде

$$D_n = \max_{i=1,n} \left\{ \left| G(x_{(i)}) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right\}, \qquad (5.5)$$

где $x_{(1)}, ..., x_{(n)}$ – вариационный ряд выборки.

Пример 5.1. Используя критерий Колмогорова, проверить на уровне значимости 10% гипотезу о том, что выборка

0,90; 0,56; 0,05; 0,21; 0,97; 0,80; 0,04; 0,12; 0,73; 0,49

получена из равномерно распределённой генеральной совокупности $R(0, 1)$.

Сформулируем основную и альтернативную статистические гипотезы:

$$H_0 : F_X(x) = R(x),$$
$$H' : F_X(x) \neq R(x),$$

где X – наблюдаемая случайная величина, а предполагаемая функция распределения $R(x)$ имеет вид

$$R(x) = \begin{cases} 0, & x < 0; \\ x, & 0 \le x \le 1; \\ 1, & otherwise. \end{cases}$$

Составим вариационный ряд выборки:

0,04; 0,05; 0,12; 0,21; 0,49; 0,56; 0,73; 0,80; 0,90; 0,97.

Значения ЭФР $F_n^*(x)$ в этих точках:

0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9.

Графики эмпирической функции распределения $F_n^*(x)$ и функции распределения $R(x)$ приведены на рис. 5.1.

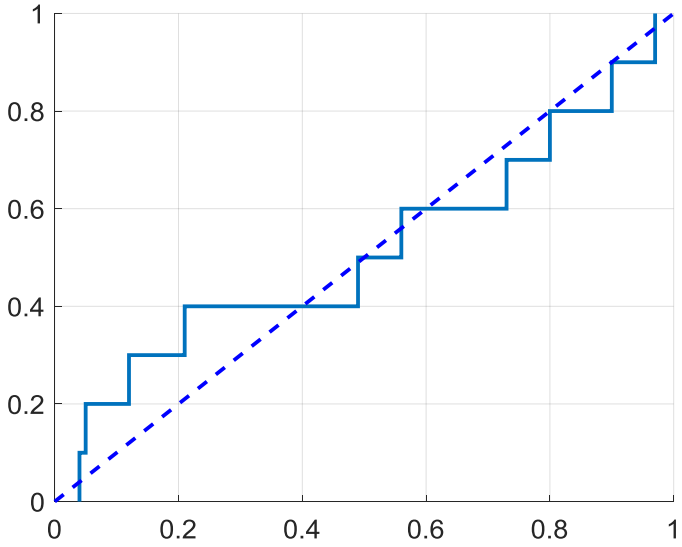

Рис. 5.1. Графики ЭФР $F_n^*(x)$ (сплошная линия) и функции распределения $R(x)$ (пунктирная линия)

По формуле (5.5) находим рассогласования между $F_n^*(x)$ и $R(x)$:

0,06; 0,15; 0,18; 0,19; 0,09; 0,06; 0,13; 0,10; 0,10; 0,07

и максимальное из них $D_n = 0{,}19$. По формуле (5.2) рассчитываем выборочное значение статистики критерия:

$$z = \sqrt{10} \cdot 0{,}19 \approx 0{,}6 .$$

По таблице квантилей распределения Колмогорова (табл. П2 приложения) находим квантиль на уровне $1 - \alpha$:

$$z_{0,9} = 1,22.$$

Таким образом, область допустимых значений $\Omega_0 = [0; 1,22)$, критическая область $\Omega' = [1,22; +\infty)$. Поскольку $z = 0,6 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Отвергать, что данная выборка могла быть получена из равномерного распределения $R(0, 1)$, нельзя.

Критерий Колмогорова имеет существенное ограничение применимости: предполагаемая функция распределения $G(x)$ должна быть полностью определена. Так, если распределение $G(x)$ имеет неизвестные параметры, то использование вместо них выборочных оценок может существенно изменить распределение статистики критерия и расположение критической точки.

Контрольные вопросы и задачи

1. Что называется критерием согласия?
2. Покажите, что эмпирическая функция распределения $\mathcal{F}_n^*(x)$ является состоятельной оценкой функции распределения $F_X(x)$.
3. Какая метрика на пространстве распределений вероятностей используется в критерии согласия Колмогорова?
4. Какая статистика критерия используется в критерии согласия Колмогорова? По какому закону она распределена при $n \to \infty$?
5. Какой тип критической области выбирается для критерия согласия Колмогорова?

§ 19. Критерий «омега-квадрат»

Из вида метрики Колмогорова (5.1) следует, что она хорошо различает функции распределения $F_n^*(x)$ и $G(x)$, отличающиеся друг от друга достаточно сильно пусть даже в небольшом числе точек. Если же $F_n^*(x)$ отличается от $G(x)$ на довольно широком интервале (или на всей числовой оси), но везде не очень сильно, то величина D_n будет невелика, и критерий Колмогорова может ложно принять основную гипотезу H_0, в то время как на самом деле рас-

пределения $F_X(x)$ и $G(x)$ различны. Этот факт свидетельствует о высокой вероятности ошибки второго рода при проверке статистической гипотезы о равенстве распределений, что делает критерий Колмогорова маломощным.

Указанный недостаток может быть устранён при использовании другой метрики для расчёта рассогласования между двумя распределениями, называемой *метрикой «омега-квадрат»*, в непрерывном случае:

$$\Delta\left(F_n^*(x), G(x)\right) = \omega_n^2 = \int_{-\infty}^{\infty} \left|F_n^*(x) - G(x)\right|^2 dx, \qquad (5.6)$$

в дискретном случае:

$$\Delta\left(F_n^*(x), G(x)\right) = \omega_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left|F_n^*(x_i) - G(x_i)\right|^2. \qquad (5.7)$$

Статистика критерия, основанная на данной метрике, называется *статистикой Крамера–Мизеса* (*Harald Cramer, Richard Edler von Mises, 1930*), или *статистикой «омега-квадрат»*:

$$Z_n = n\omega_n^2, \qquad (5.8)$$

для которой показано, что при условии истинности основной гипотезы H_0 при $n \to \infty$ её закон распределения не зависит от вида функции $G(x)$ и стремится к распределению «омега-квадрат».

Приближённо полагая при больших n ($n > 40$), что $Z_n \sim \omega^2$, для статистики критерия может быть рассчитана любая квантиль, используя таблицу. Некоторые критические точки распределения «омега-квадрат» и соответствующие им уровни значимости приведены в табл. 5.1.

Таблица 5.1

Таблица квантилей распределения «омега-квадрат»

α	0,005	0,01	0,025	0,05	0,10	0,15	0,20	0,25
$z_{1-\alpha}$	0,87	0,75	0,58	0,46	0,35	0,28	0,24	0,21

Аналогично критерию Колмогорова, в критерии «омега-квадрат» критическая область выбирается правосторонней.

На практике для вычисления рассогласования ω_n^2 между ЭФР $F_n^*(x)$ и предполагаемой функцией распределения $G(x)$ по выборке $x_1, \ldots, x_n$ удобно использовать формулу

$$\omega_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(G(x_{(i)}) - \frac{2i-1}{2n} \right)^2 , \qquad (5.9)$$

где $x_{(1)}, ..., x_{(n)}$ – вариационный ряд выборки.

Пример 5.2. В условиях примера 5.1 проверить на уровне значимости 10% гипотезу о том, что выборка получена из равномерно распределённой генеральной совокупности $X \sim R(0, 1)$, используя критерий «омега-квадрат».

По формуле (5.9) находим рассогласование между $F_n^*(x)$ и $G(x)$:

$$\omega_n^2 \approx 0,006$$

и выборочное значение статистики критерия:

$$z = 10 \cdot 0,006 \approx 0,06 .$$

По табл. 5.1 находим квантиль распределения «омега-квадрат» на уровне $1 - \alpha$:

$$z_{0,9} = 0,35 .$$

Таким образом, область допустимых значений $\Omega_0 = [0; 0,35)$, критическая область $\Omega' = [0,35; +\infty)$. Поскольку $z = 0,06 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Отвергать, что данная выборка могла быть получена из равномерного распределения $R(0, 1)$, нельзя.

Если требуется проверить принадлежность функции распределения $F_X(x, \theta)$ заданному параметрическому множеству распределений $G(x, \theta)$, где θ – вектор параметров, то проверяется согласие эмпирической функции распределения $F_n^*(x)$ лишь с максимально правдоподобным для данной выборки распределением $G(x, \tilde{\theta})$, где $\tilde{\theta}$ – вектор МП-оценок параметров.

Контрольные вопросы и задачи

1. Каким недостатком обладает критерий Колмогорова?
2. Какая метрика на пространстве распределений вероятностей используется в критерии согласия «омега-квадрат»?
3. Какая статистика критерия используется в критерии согласия «омега-квадрат»? По какому закону она распределена при $n \to \infty$?
4. Какой тип критической области выбирается в критерии согласия «омега-квадрат»?
5. Объясните, почему при проверке принадлежности функции распределения $F_X(x, \theta)$ заданному параметрическому множеству распределений $G(x, \theta)$ проверяется лишь согласие с распределением $G(x, \tilde{\theta})$, где $\tilde{\theta}$ – МП-оценка параметра θ. Может ли быть использована ММ-оценка параметра θ?

§ 20. Критерий Пирсона

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины X, имеющей неизвестное распределение $F_X(x)$. Наряду с критерием Колмогорова и критерием «омега-квадрат» для проверки гипотезы о совпадении функции распределения $F_X(x)$ с некоторой известной функцией $G(x)$:

$$H_0 : F_X(x) = G(x),$$
$$H' : F_X(x) \neq G(x);$$

может быть также использован критерий Пирсона.

Критерий Пирсона (*Karl Pearson*, 1900), или *критерий «хи-квадрат»* (*Pearson's chi-squared test*), основан на оценке степени близости гистограммы относительных частот выборки и известной плотности распределения $g(x) = \dfrac{dG(x)}{dx}$. Для построения гистограммы проводят группировку выборочных значений на k интервалов $J_1, \ldots, J_k$, как правило, одинаковой ширины h (см. § 2):

$$J_1 = [\alpha_0 = x_{(1)}; \alpha_1),\ J_2 = [\alpha_1; \alpha_2),\ \ldots,\ J_k = [\alpha_{k-1}; \alpha_k = x_{(n)}].$$

Пусть n_i – число элементов выборки, принадлежащих интервалу J_i, $i=\overline{1,k}$, $\sum_{i=1}^{k} n_i = n$.

На основе известной функции плотности распределения $g(x)$ рассчитываются ожидаемые вероятности p_i попадания в каждый интервал J_i:

$$p_i = P(X \in J_i) = \int_{\alpha_{i-1}}^{\alpha_i} g(x)dx, \quad i=\overline{1,k}.$$

Полученные результаты представим в виде таблицы:

Событие	Частоты			Всего
	J_1	...	J_k	
Наблюдаемое	n_1	...	n_k	n
Ожидаемое	np_1	...	np_k	n

При условии истинности основной гипотезы H_0 относительная частота $\tilde{p}_i = \frac{n_i}{n}$ попадания в интервал J_i будет состоятельной оценкой вероятности p_i, $i=\overline{1,k}$. Это означает, что для каждого интервала J_i, $i=\overline{1,k}$, вероятность того, что рассогласование между $\tilde{p}_i$ и p_i примет достаточно большие значения, стремится к нулю при $n \to \infty$.

В качестве меры рассогласования между $\tilde{p}_i$ и p_i используется статистика

$$Z = n\sum_{i=1}^{k} \frac{(\tilde{p}_i - p_i)^2}{p_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (5.10)$$

для которой показано, что при условии истинности основной гипотезы H_0 при $n \to \infty$ её закон распределения не зависит от вида функции $G(x)$ и стремится к распределению «хи-квадрат» с $k-r-1$ степенями свободы, где r – число оцененных по выборке параметров распределения $G(x)$ (*теорема Пирсона*).

Использование статистики Z возможно также для проверки согласия выборочных данных с дискретным распределением гене-

ральной совокупности. В этом случае в качестве вероятностей $p_1, \ldots, p_k$ следует брать предполагаемые вероятности дискретных значений генеральной совокупности ($\sum_{i=1}^{k} p_i = 1$), а в качестве оценок $\tilde{p}_1, ..., \tilde{p}_k$ – относительные частоты этих значений в выборке. При необходимости близкие дискретные значения могут быть сгруппированы.

Если требуется проверить принадлежность функции распределения $F_X(x, \theta)$ заданному параметрическому множеству распределений $G(x, \theta)$, где θ – вектор параметров, то проверяется согласие лишь с максимально правдоподобным для данной выборки распределением $G(x, \tilde{\theta})$, где $\tilde{\theta}$ – вектор МП-оценок параметров.

Аппроксимация закона распределения статистики Z при условии истинности основной гипотезы H_0 законом $\chi^2(k-r-1)$ с высокой точностью возможна лишь при больших значениях ожидаемых абсолютных частот np_i, $i = \overline{1,k}$. В случае если для некоторых интервалов $np_i < 5$, то такие интервалы рекомендуется объединить с соседними.

Аналогично критериям Колмогорова и «омега-квадрат», в критерии Пирсона критическая область выбирается правосторонней.

Пример 5.3. Проведено $n = 1502$ наблюдений работы системы обмена данными в некоторый промежуток времени в часы «пик». В течение контролируемого промежутка времени фиксировалось число передач ошибочных данных из-за перегруженности канала связи. Данные наблюдений приведены в таблице:

Число ошибочных передач x_i	0	1	2	3	4	5	6
Число наблюдений n_i	322	511	370	200	75	20	4

Используя критерий согласия Пирсона, проверить гипотезу на уровне значимости $\alpha = 0{,}05$ о том, что случайная величина X – чис-

ло ошибочных передач данных в единицу времени – распределена по закону Пуассона.

Закон распределения Пуассона имеет единственный неизвестный параметр λ, следовательно, $r = 1$. МП-оценкой параметра λ является среднее арифметическое числа ошибочных передач:

$$\tilde{\lambda} = \bar{x} = \frac{1}{1502}(0 \cdot 322 + ... + 6 \cdot 4) \approx 1{,}51.$$

Таким образом, задача сводится к проверке статистической гипотезы:

$$H_0 : X \sim Poisson(1{,}51),$$

$$H' : X \not\sim Poisson(1{,}51).$$

По формуле вероятностей распределения Пуассона $p_i = P(X = i) = \frac{\tilde{\lambda}^i}{i!} e^{-\tilde{\lambda}}$ находим ожидаемые при условии истинности основной гипотезы H_0 вероятности p_0, …, p_6:

$$p_0 = \frac{1}{1!} e^{-1{,}51} \approx 0{,}220;$$

$$p_1 = \frac{1{,}51}{1!} e^{-1{,}51} \approx 0{,}333;$$

…

$$p_6 = \frac{1{,}51^6}{6!} e^{-1{,}51} \approx 0{,}004.$$

Рассчитанные значения сведены в таблицу:

x_i	0	1	2	3	4	5	6
n_i	322	511	370	200	75	20	4
p_i	0,220	0,333	0,252	0,127	0,048	0,015	0,004
np_i	330,1	500,2	378,8	191,4	72,4	21,9	5,5

Найдём выборочное значение статистики Пирсона:

$$z = \sum_{i=0}^{6} \frac{(n_i - np_i)^2}{np_i} = \frac{(322 - 330{,}1)^2}{330{,}1} + ... + \frac{(4 - 5{,}5)^2}{5{,}5} \approx 1{,}7.$$

При условии истинности основной гипотезы статистика $Z \sim \chi^2(7-1-1) = \chi^2(5)$.

По таблице квантилей распределения «хи-квадрат» с пятью степенями свободы (табл. П3 приложения) находим квантиль на уровне $1 - \alpha$:

$$z_{0,95} = 11{,}1.$$

Таким образом, область допустимых значений $\Omega_0 = [0; 11{,}1)$, критическая область $\Omega' = [11{,}1; +\infty)$. Поскольку $z = 1{,}7 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Отвергать, что число передач ошибочных данных из-за перегруженности канала связи имеет распределение Пуассона, нельзя.

Контрольные вопросы и задачи

1. Какая метрика на пространстве распределений вероятностей используется в критерии согласия Пирсона?
2. Как выбирается число интервалов группировки в критерии согласия Пирсона?
3. Можно ли при расчёте статистики критерия Пирсона использовать интервалы различной ширины?
4. Покажите, что относительная частота попадания выборочных значений в интервал группировки является состоятельной и несмещённой оценкой вероятности попадания наблюдаемой случайной величины в этот интервал.
5. Какая статистика критерия используется в критерии согласия Пирсона?
6. Сформулируйте теорему Пирсона.
7. Объясните принцип выбора типа критической области в критерии Пирсона.
8. Докажите, что статистика критерия Пирсона имеет распределение «хи-квадрат» при $n \to \infty$.
9. Какое условие накладывается на интервалы группировки при использовании критерия Пирсона?
10. Проведите сравнительный анализ критериев согласия Колмогорова, «омега-квадрат» и Пирсона.

§ 21. Проверка гипотезы о нормальности распределения

Частным случаем статистической гипотезы о виде распределения является гипотеза о нормальности распределения наблюдаемой случайной величины X:

$$H_0 : X \sim N(m, \sigma), \qquad H' : X \not\sim N(m, \sigma). \tag{5.11}$$

Для проверки этой гипотезы может быть использован любой из рассмотренных выше критериев согласия (критерий Колмогорова, критерий «омега-квадрат», критерий Пирсона) или один из специальных критериев проверки на нормальность, в частности *критерий Харке–Бера* (*Carlos Jarque, Anil K. Bera*, 1980).

Для нормально распределённой случайной величины X известно, что её коэффициент асимметрии γ и эксцесс ε равны нулю. Выборочные значения этих характеристик γ^* и ε^*, рассчитываемые по формулам (1.21) и (1.22) соответственно, независимы и являются состоятельными и асимптотически несмещёнными оценками, а в случае нормально распределённой генеральной совокупности при $n \to \infty$ имеют распределения $N\left(0, \sqrt{\frac{6}{n}}\right)$ и $N\left(0, \sqrt{\frac{24}{n}}\right)$. Следовательно, *стандартизованный коэффициент асимметрии*

$$\gamma^*_{\text{ст}} = \frac{\gamma^*}{\sqrt{6/n}} \tag{5.12}$$

и *стандартизованный эксцесс*

$$\varepsilon^*_{\text{ст}} = \frac{\varepsilon^*}{\sqrt{24/n}} \tag{5.13}$$

имеют стандартизованные нормальные распределения $N(0, 1)$.

По определению закона «хи-квадрат» статистика

$$Z = \left(\gamma^*_{\text{ст}}\right)^2 + \left(\varepsilon^*_{\text{ст}}\right)^2 = \frac{n}{6}\left((\gamma^*)^2 + \frac{(\varepsilon^*)^2}{4}\right), \tag{5.14}$$

используемая в качестве статистики критерия, имеет распределение $\chi^2(2)$.

При отклонении распределения генеральной совокупности от нормального, сопровождающегося отклонением коэффициента асимметрии и эксцесса от нулевых значений, происходит увеличение выборочных значений статистики критерия. В связи с этим критическая область в критерии Харке–Бера должна выбираться правосторонней.

Контрольные вопросы и задачи

1. Какой закон распределения имеют выборочные коэффициент асимметрии и эксцесс для выборки из нормально распределённой генеральной совокупности при $n \to \infty$?
2. Что называется стандартизованным коэффициентом асимметрии и стандартизованным эксцессом?
3. Какая статистика критерия используется в критерии Харке–Бера? По какому закону она распределена при $n \to \infty$?
4. Может ли гипотеза о нормальности распределения генеральной совокупности быть сведена к проверке двух параметрических гипотез: о равенстве нулю коэффициента асимметрии и эксцесса?
5. Какие статистические критерии могут быть использованы для проверки нормальности распределения генеральной совокупности?

§ 22. Проверка гипотез об однородности выборок. Критерий знаков

В практических приложениях наряду с задачей о соответствии выборочных наблюдений предполагаемому закону распределения встречается задача о проверке соответствия распределений двух генеральных совокупностей по результатам выборочных наблюдений.

Пусть $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$ – выборки объёмов n_X и n_Y наблюдений случайных величин X и Y, имеющих неизвестные распределения $F_X(\xi)$ и $F_Y(\xi)$ соответственно. Выборки $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$ называются *однородными*, если $F_X(\xi) = F_Y(\xi) \quad \forall \xi \in \mathbb{R}$. Иными словами, выборки однородные, если они получены из одной и той же

генеральной совокупности, или являются наблюдениями одной и той же случайной величины.

Сформулируем основную и альтернативную гипотезы однородности:

$$H_0 : F_X(\xi) = F_Y(\xi), \qquad H' : F_X(\xi) \neq F_Y(\xi). \tag{5.15}$$

Одним из наиболее простых и грубых критериев проверки этих гипотез является критерий знаков.

Критерий знаков (*sign test*) используется для проверки однородности двух *связанных выборок* (*paired samples*). Такие выборки получаются в результате наблюдений двумерного случайного вектора (*X, Y*). Объёмы связанных выборок всегда равны.

Критерий знаков – пример *непараметрического критерия* математической статистики, т.е. критерия, использующего не сами численные значения элементов выборки, а структурные свойства выборки (например, отношения порядка между её элементами, знаки и пр.). Мощность непараметрических критериев, как правило, меньше, чем мощность их параметрических аналогов. Причина этого связана с неизбежной потерей части информации, содержащейся в выборке. Однако непараметрические методы могут применяться при менее строгих предположениях о свойствах наблюдаемых случайных величин и, как правило, более просты с вычислительной точки зрения.

Если выборки получены из одной и той же генеральной совокупности, то значения x_i и y_i, $i = \overline{1,n}$, взаимозаменяемы, и, следовательно, вероятности появления положительных и отрицательных разностей между x_i и y_i равны, т.е.

$$P(X_i - Y_i > 0) = P(X_i - Y_i < 0) = 1/2. \tag{5.16}$$

Пусть K – число знаков «+» в последовательности знаков разностей $x_1 - y_1, \ldots, x_n - y_n$. Если в этой последовательности разностей содержатся нулевые элементы, то они исключаются из рассмотрения. Далее для простоты будем считать, что нулевых элементов нет. При условии, что основная гипотеза H_0 верна, а пары наблюдений (X_i, Y_i), $i = \overline{1,n}$, и, следовательно, знаки разностей $X_i - Y_i$ независимы, число K знаков «+» имеет биномиальное распределение

$B(n, 1/2)$. Таким образом, проверка гипотезы однородности (5.15) сводится к проверке гипотезы о параметре p биномиального распределения:

$$H_0 : p = 1/2,$$
$$H' : p \neq 1/2.$$

Несложно показать, что эта гипотеза эквивалентна гипотезе о равенстве медиан распределений $F_X(\xi)$ и $F_Y(\xi)$.

Как известно, для биномиального распределения математическое ожидание $m_K = np$ и дисперсия $d_K = np(1-p)$. В соответствии с предельной теоремой Муавра–Лапласа при большом числе испытаний n (на практике уже при $n > 30$) статистика K имеет закон распределения, близкий к нормальному:

$$K \sim N\left(np, \sqrt{np(1-p)}\right).$$

Частота «успеха» $H = K / n$ также имеет нормальное распределение:

$$H \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

В качестве статистики критерия используется стандартизованная частота H при $p = 1/2$ (см. § 17):

$$Z = \frac{H - 1/2}{\sqrt{\frac{1}{4n}}} = 2\sqrt{n}(H - 1/2), \qquad (5.17)$$

которая при условии истинности H_0 имеет распределение $f_Z(z \mid H_0) \sim N(0,1)$.

Основная гипотеза H_0 должна отклоняться при больших отличиях частоты знаков «+» от значения $1/2$ как в меньшую, так и в большую сторону, т.е. в области больших абсолютных значений статистики критерия Z. Таким образом, критическая область для статистики Z должна выбираться двусторонней.

Условие (5.16) является необходимым, но не достаточным условием однородности выборок $x_1, ..., x_n$ и $y_1, ..., y_n$. Это означает, что из принятия основной гипотезы критерия знаков не следует одно-

родность выборок, а следует лишь возможность однородности. Если же основная гипотеза критерия знаков отклоняется, то отклоняется и гипотеза однородности выборок.

Пример 5.4. Исследуется действие психологического тренинга на уровень тревожности испытуемых, измеряемый с помощью специального теста. В результате выборочного тестирования $n = 14$ испытуемых получены следующие результаты:

Номер испытуемого	1	2	3	4	5	6	7
Уровень до тренинга	30	39	35	34	40	35	22
Уровень после тренинга	34	39	26	33	34	40	25
Номер испытуемого	8	9	10	11	12	13	14
Уровень до тренинга	22	32	23	16	34	33	34
Уровень после тренинга	21	30	24	15	27	35	30

Определить, является ли действие тренинга на уровень тревожности статистически значимым при уровне значимости $\alpha = 10\%$, используя критерий знаков.

Действие тренинга на уровень тревожности будет статистически значимым, если случайные величины X и Y – уровни тревожности испытуемых до и после тренинга соответственно – имеют различные законы распределения.

Сформулируем основную и альтернативную статистические гипотезы:

$$H_0 : F_X(\xi) = F_Y(\xi),$$
$$H' : F_X(\xi) \neq F_Y(\xi),$$

где $F_X(\xi)$ и $F_Y(\xi)$ – функции распределения случайных величин X и Y соответственно.

Составим выборку из знаков разностей между выборочными значениями уровней тревожности после и до тренинга:

+, 0, –, –, –, +, +, –, –, +, –, –, +, –

Нуль из рассмотрения исключаем, далее полагаем $n = 13$. Относительная частота знаков «+»:

$$h = \frac{5}{13} \approx 0,38 .$$

Найдём выборочное значение статистики критерия (5.17):

$$z = 2\sqrt{13} \cdot (0,38 - 0,5) \approx -0,87 .$$

Аппроксимируем закон распределения статистики критерия при условии истинности основной гипотезы H_0 стандартизованным нормальным распределением $N(0, 1)$. По таблице математической статистики (табл. П1 приложения) находим квантиль на уровне $1 - \alpha/2$:

$$z_{0,95} = 1,65 .$$

Таким образом, область допустимых значений $\Omega_0 = (-1,65; 1,65)$. Поскольку $z = -0,87 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Утверждать, что психологический тренинг оказал значимое действие на уровень тревожности испытуемых, нельзя.

Контрольные вопросы и задачи

1. Какие выборки называются однородными?
2. Как формулируются основная и альтернативная гипотезы в статистических гипотезах об однородности выборок?
3. Какие статистические критерии называются непараметрическими?
4. В чём состоит критерий знаков для проверки гипотезы об однородности выборок?
5. Каким свойством должны обладать выборки при использовании критерия знаков?
6. Покажите, что основная гипотеза критерия знаков эквивалентна статистической гипотезе о равенстве медиан двух распределений.
7. Какой закон распределения имеет статистика критерия, используемая в критерии знаков? Каким законом он аппроксимируется при $n \to \infty$?
8. Какой тип критической области выбирается при проверке гипотезы однородности с помощью критерия знаков?

§ 23. Критерий Манна–Уитни

Одним из наиболее популярных непараметрических критериев проверки статистической гипотезы (5.15) об однородности выборок является критерий Манна–Уитни. Критерий был предложен *Уилкоксоном* (*Frank Wilcoxon*, 1945) и существенно переработан и расширен *Манном* и *Уитни* (*Henry Mann*, *Donald Whitney*, 1947). Другие названия критерия: критерий Манна–Уитни–Уилкоксона, критерий суммы рангов Уилкоксона (*Mann–Whitney–Wilcoxon test, MWW-test, U-test*).

Критерий Манна–Уитни использует тот факт, что если выборки $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$ получены из одной и той же генеральной совокупности, то элементы как первой, так и второй выборок в вариационном ряду $z_{(1)}, ..., z_{(N)}$ объединённой выборки $x_1, ..., x_{n_X}$, $y_1, ..., y_{n_Y}$, где $N = n_X + n_Y$ – суммарный объём выборок, перемешаны равномерно.

Для оценки степени перемешивания данных двух выборок проводится *ранжирование* объединённой выборки $z_1, ..., z_N$. *Рангом* элемента z_i в выборке $z_1, ..., z_N$ называется его порядковый номер в вариационном ряду $z_{(1)}, ..., z_{(N)}$. Минимальный элемент $z_{(1)}$ выборки имеет ранг 1, максимальный элемент $z_{(N)}$ – ранг *N*. Если несколько выборочных значений в вариационном ряду равны, то им приписываются одинаковые ранги, равные среднему арифметическому из их порядковых номеров.

В результате ранжирования для выборки $z_1, ..., z_N$ получаем выборку соответствующих рангов $r_1, ..., r_N$. Обозначим через R_X сумму рангов в ряду $r_1, ..., r_N$, соответствующих элементам из первой выборки, R_Y – элементам из второй выборки.

Несложно показать, что

$$\frac{n_X(n_X+1)}{2} \le R_X \le \frac{n_X(n_X+1)}{2} + n_X n_Y,$$

$$\frac{n_Y(n_Y+1)}{2} \le R_Y \le \frac{n_Y(n_Y+1)}{2} + n_X n_Y.$$

Минимальное значение $\frac{n_X(n_X+1)}{2}$ суммы рангов R_X элементов первой выборки достигается, когда все они расположены на левом конце объединённого вариационного ряда, максимальное значение $n_X n_Y + \frac{n_X(n_X+1)}{2}$ – когда на правом. Аналогичные значения достигаются для суммы рангов R_Y.

Введём статистики U_X и U_Y, линейно связанные с суммами рангов R_X и R_Y:

$$U_X = n_X n_Y + \frac{n_X(n_X+1)}{2} - R_X, \quad 0 \le U_X \le n_X n_Y,$$

$$U_Y = n_X n_Y + \frac{n_Y(n_Y+1)}{2} - R_Y, \quad 0 \le U_Y \le n_X n_Y,$$

$$U_X + U_Y = n_X n_Y.$$

В качестве меры степени перемешивания элементов двух выборок в критерии Манна–Уитни используется любая из статистик U_X или U_Y. При равномерном перемешивании элементов выборочные значения статистик U_X и U_Y будут близки к их средним по диапазону изменения значениям $\frac{n_X n_Y}{2}$.

В условиях истинности основной гипотезы H_0 при $n_X \to \infty$, $n_Y \to \infty$ закон распределения статистик U_X и U_Y не зависит от вида функций $F_X(\xi)$ и $F_Y(\xi)$ и стремится к нормальному

$$N\left(\frac{n_X n_Y}{2}, \sqrt{\frac{n_X n_Y (n_X + n_Y + 1)}{12}}\right).$$

В качестве статистики критерия выберем любую из стандартизованных статистик U_X или U_Y (например, U_X):

$$Z = \frac{U_X - \frac{n_X n_Y}{2}}{\sqrt{\frac{n_X n_Y (n_X + n_Y + 1)}{12}}}, \qquad (5.18)$$

которая при условии истинности H_0 имеет распределение $f_Z(z \mid H_0) \sim N(0,1)$. На практике распределение статистики Z мож-

но аппроксимировать нормальным уже при $n_X > 10$, $n_Y > 10$, что делает критерий Манна–Уитни применимым для проверки гипотезы об однородности для малых выборок.

Основная гипотеза H_0 должна отклоняться при больших отличиях статистики U_X от среднего значения $\frac{n_X n_Y}{2}$, т.е. в области больших по модулю значений статистики критерия Z. Таким образом, критическая область для статистики Z должна выбираться двусторонней.

Пример 5.5. В условиях примера 5.4 определить, является ли действие тренинга на уровень тревожности статистически значимым при уровне значимости $\alpha = 10\%$, используя критерий Манна–Уитни.

Составим вариационный ряд объединённой выборки уровней тревожности до (случайная величина X) и после (случайная величина Y) тренинга:

№ п/п	1	2	3	4	5	6	7
Уровни тревожности	**15**	16	**21**	22	22	23	**24**
№ п/п	8	9	10	11	12	13	14
Уровни тревожности	**25**	**26**	**27**	30	**30**	**30**	32
№ п/п	15	16	17	18	19	20	21
Уровни тревожности	33	**33**	34	34	34	**34**	**34**
№ п/п	22	23	24	25	26	27	28
Уровни тревожности	35	35	**35**	39	**39**	40	**40**

Жирным шрифтом в таблице выделены выборочные значения уровней тревожности после тренинга.

Построим ряд рангов для объединённой выборки и представим их в виде таблицы:

Номер испытуемого	Уровень до тренинга	Ранг уровня до тренинга	Уровень после тренинга	Ранг уровня после тренинга
1	30	12	**34**	**19**
2	39	25,5	**39**	**25,5**
3	35	23	**26**	**9**
4	34	19	**33**	**15,5**
5	40	27,5	**34**	**19**
6	35	23	**40**	**27,5**
7	22	4,5	**25**	**8**
8	22	4,5	**21**	**3**
9	32	14	**30**	**12**
10	23	6	**24**	**7**
11	16	2	**15**	**1**
12	34	19	**27**	**10**
13	33	15,5	**35**	**23**
14	34	19	**30**	**12**
Сумма	–	214,5	**–**	**191,5**

Рассчитаем выборочное значение статистики U_X:

$$u_X = 14 \cdot 14 + \frac{14 \cdot 15}{2} - 214{,}5 = 86{,}5$$

и выборочное значение статистики критерия Z:

$$z = \frac{86{,}5 - 14 \cdot 14/2}{\sqrt{14 \cdot 14 \cdot 29/12}} \approx -0{,}53.$$

Аппроксимируем закон распределения статистики критерия при условии истинности основной гипотезы H_0 стандартизованным нормальным распределением $N(0, 1)$. По таблице математической статистики (табл. П1 приложения) находим квантиль на уровне $1 - \alpha/2$:

$$z_{0{,}95} = 1{,}65.$$

Таким образом, область допустимых значений $\Omega_0 = (-1{,}65; 1{,}65)$. Поскольку $z = -0{,}53 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Утверждать,

что психологический тренинг оказал значимое действие на уровень тревожности испытуемых, нельзя.

Контрольные вопросы и задачи

1. Что называется рангом элемента в выборке?
2. В чём состоит критерий Манна–Уитни для проверки гипотезы об однородности выборок?
3. Покажите, что математическое ожидание случайной величины U_X при условии истинности основной гипотезы равно $\frac{n_X n_Y}{2}$.
4. Каким законом аппроксимируется распределение статистики критерия Манна–Уитни при $n_X \to \infty$, $n_Y \to \infty$?
5. Какой тип критической области выбирается при проверке гипотезы однородности с помощью критерия Манна–Уитни?

§ 24. Модификации критериев Колмогорова, «омега-квадрат» и Пирсона для проверки гипотез об однородности выборок

Пусть $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$ – выборки объёмов n_X и n_Y наблюдений случайных величин X и Y, имеющих неизвестные распределения $F(\xi)$ и $G(\xi)$ соответственно.

Параметрические критерии проверки статистической гипотезы (5.15) об однородности основаны на оценке рассогласования между эмпирическими распределениями наблюдаемых случайных величин. Здесь могут быть использованы те же самые метрики, что и в критериях Колмогорова, «омега-квадрат» и Пирсона. Такие критерии, модифицированные для случая двух выборок, называются *двухвыборочными* (*two-sample tests*).

1. *Двухвыборочный критерий Колмогорова* (*two-sample KS-test*).

В критерии используется статистика

$$Z_{n_X, n_Y} = \sqrt{\frac{n_X n_Y}{n_X + n_Y}} D_{n_X, n_Y}, \qquad (5.19)$$

где D_{n_X,n_Y} – расстояние по Колмогорову между эмпирическими функциями распределения $F^*_{n_X}(\xi)$ и $G^*_{n_Y}(\xi)$ случайных величин X и Y соответственно:

$$D_{n_X,n_Y} = \max_{\xi} \left| F^*_{n_X}(\xi) - G^*_{n_Y}(\xi) \right|.$$

Для статистики Z_{n_X,n_Y} показано, что при условии истинности основной гипотезы H_0 при $n_X \to \infty$, $n_Y \to \infty$ её закон распределения не зависит от вида функций $F(\xi)$ и $G(\xi)$, причём её распределение стремится к распределению Колмогорова (см. § 18). Аппроксимация распределения статистики Z_{n_X,n_Y} распределением Колмогорова даёт хорошие результаты уже при $n_X > 40$, $n_Y > 40$.

Так же, как и в критерии согласия Колмогорова, здесь основная гипотеза H_0 должна отклоняться в области больших значений Z_{n_X,n_Y}, т.е. критическая область должна выбираться правосторонней.

На практике для вычисления рассогласования D_{n_X,n_Y} между ЭФР $F^*_{n_X}(\xi)$ и $G^*_{n_Y}(\xi)$ удобно использовать формулу:

$$D_{n_X,n_Y} = \max_{i=1,N} \left| F^*_{n_X}(z_{(i)}) - G^*_{n_Y}(z_{(i)}) \right|, \qquad (5.20)$$

где $z_{(1)}, ..., z_{(N)}$ – вариационный ряд объединённой выборки $x_1, ..., x_{n_X}, y_1, ..., y_{n_Y}$; $N = n_X + n_Y$ – суммарный объём выборок.

Пример 5.6. В условиях примера 5.4 определить, является ли действие тренинга на уровень тревожности статистически значимым при уровне значимости $\alpha = 10\%$, используя критерий Колмогорова.

Графики эмпирических функций распределения $F^*_{n_X}(\xi)$ и $G^*_{n_Y}(\xi)$ приведены на рис. 5.2.

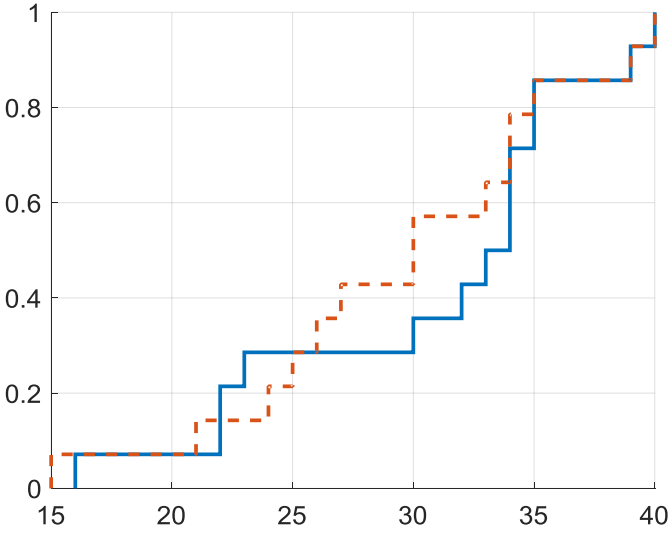

Рис. 5.2. Эмпирические функции распределения $F^*_{n_X}(\xi)$ (сплошная линия) и $G^*_{n_Y}(\xi)$ (пунктирная линия)

Максимальное рассогласование между эмпирическими функциями распределения во всех точках объединённого вариационного ряда $D_{n_X,n_Y} = 0,21$. Выборочное значение статистики критерия:

$$z_{n_X,n_Y} = \sqrt{\frac{14 \cdot 14}{14+14}} \cdot 0,21 \approx 0,56.$$

По таблице математической статистики (табл. П2 приложения) находим квантиль распределения Колмогорова на уровне $1 - \alpha$:

$$z_{0,9} = 1,22.$$

Таким образом, область допустимых значений $\Omega_0 = [0;1,22)$, критическая область $\Omega' = [1,22;+\infty)$. Поскольку $z = 0,56 \in \Omega_0$, то оснований считать, что гипотеза H_0 не согласуется с экспериментальными данными, нет. Утверждать, что психологический тренинг оказал значимое действие на уровень тревожности испытуемых, нельзя.

2. *Двухвыборочный критерий «омега-квадрат»* (*two-sample omega-squared test*).

Метрика «омега-квадрат» для расчёта рассогласования между распределениями случайных величин X и Y на основе результатов наблюдений $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$ имеет вид

$$\omega^2_{n_X, n_Y} = \frac{1}{N} \sum_{i=1}^{N} \left| F^*_{n_X}(z_i) - G^*_{n_X}(z_i) \right|^2, \qquad (5.21)$$

где $z_1, ..., z_N$ – объединённая выборка $x_1, ..., x_{n_X}$, $y_1, ..., y_{n_Y}$; $N = n_X + n_Y$ – суммарный объём выборок; $F^*_{n_X}(\xi)$ и $G^*_{n_Y}(\xi)$ – эмпирические функции распределения случайных величин X и Y соответственно.

В двухвыборочном критерии «омега-квадрат» (критерии Крамера–Мизеса) используется статистика

$$Z_{n_X, n_Y} = \frac{n_X n_Y}{n_X + n_Y} \omega^2_{n_X, n_Y}, \qquad (5.22)$$

для которой показано, что при условии истинности основной гипотезы H_0 при $n_X \to \infty$, $n_Y \to \infty$ её закон распределения не зависит от вида функций $F(\xi)$ и $G(\xi)$ и стремится к распределению «омега-квадрат» (см. § 19).

Аналогично критерию согласия «омега-квадрат», основная гипотеза H_0 должна отклоняться в области больших значений Z_{n_X, n_Y}, т.е. критическая область должна выбираться правосторонней.

На практике для вычисления выборочного значения статистики Крамера–Мизеса удобно использовать формулу:

$$z = \frac{w}{n_X n_Y (n_X + n_Y)} - \frac{4 n_X n_Y - 1}{6(n_X + n_Y)}, \qquad (5.23)$$

где

$$w = n_X \sum_{i=1}^{n_X} (r_i - i)^2 + n_Y \sum_{j=1}^{n_Y} (s_j - j)^2,$$

а r_i и s_j – ранги элементов x_i и y_j соответственно в объединённой выборке $z_1, ..., z_N$, $i = \overline{1, n_X}$, $j = \overline{1, n_Y}$.

3. *Двухвыборочный критерий Пирсона* (*two-sample chi-squared test*).

Двухвыборочный критерий Пирсона, или критерий «хи-квадрат», основан на оценке степени близости гистограмм относительных частот выборок $x_1, ..., x_{n_X}$ и $y_1, ..., y_{n_Y}$. Для построения гистограмм проводят совместную группировку выборочных значений обеих выборок на k интервалов J_1, …, J_k, как правило, одинаковой ширины h (см. § 2):

$$J_1 = [\alpha_0 = z_{(1)}; \alpha_1),\ J_2 = [\alpha_1; \alpha_2),\ \ldots,\ J_k = [\alpha_{k-1}; \alpha_k = z_{(N)}],$$

где $z_{(1)}$ и $z_{(N)}$ – крайние члены вариационного ряда объединённой выборки $x_1, ..., x_{n_X}$, $y_1, ..., y_{n_Y}$; $N = n_X + n_Y$ – суммарный объём выборок.

Результаты группировки представим в виде таблицы:

Наблюдаемая случайная величина	Число наблюдений			Всего
	J_1	...	J_k	
X	$n_1^{(X)}$	...	$n_k^{(X)}$	n_X
Y	$n_1^{(Y)}$	...	$n_k^{(Y)}$	n_Y

В качестве меры рассогласования между относительными частотами $\frac{n_i^{(X)}}{n_X}$ и $\frac{n_i^{(Y)}}{n_Y}$ используется статистика

$$Z_{n_X, n_Y} = n_X n_Y \sum_{i=1}^{k} \frac{1}{n_i^{(X)} + n_i^{(Y)}} \left(\frac{n_i^{(X)}}{n_X} - \frac{n_i^{(Y)}}{n_Y} \right)^2, \qquad (5.24)$$

для которой показано, что в условиях истинности основной гипотезы H_0 при $n_X \to \infty$, $n_Y \to \infty$ её распределение не зависит от вида функций $F(\xi)$ и $G(\xi)$ и стремится к распределению «хи-квадрат» с $k - 1$ степенью свободы.

Аналогично критерию согласия Пирсона основная гипотеза H_0 должна отклоняться в области больших значений Z_{n_X, n_Y}, т.е. критическая область должна выбираться правосторонней.

Аппроксимация закона распределения статистики Z_{n_X,n_Y} при условии истинности основной гипотезы H_0 законом $\chi^2(k-1)$ с высокой точностью возможна лишь при больших значениях частот $n_i^{(X)}$ и $n_i^{(Y)}$, $i=\overline{1,k}$. В случае если для некоторых интервалов $n_i^{(X)}<3$ или $n_i^{(Y)}<3$, то такие интервалы рекомендуется объединить с соседними.

Пример 5.7. Исследуются произведения американских писателей XIX в.: «Всадник без головы» Майн Рида и «Зверобой» Фенимора Купера. Пусть случайные величины *X* и *Y* – количества слов в предложении в этих произведениях. Проверить на уровне значимости $\alpha=5\%$ гипотезу о том, что случайные величины *X* и *Y* имеют одинаковые законы распределения, если по результатам выборочного подсчёта количества слов в 100 предложениях из каждого произведения получены следующие результаты:

№ п/п	Количество слов в предложении	Число предложений	
		«Всадник без головы»	«Зверобой»
1	От 1 до 3	8	7
2	От 4 до 6	15	8
3	От 7 до 9	25	13
4	От 10 до 12	21	13
5	От 13 до 15	9	16
6	От 16 до 18	4	6
7	От 19 до 21	3	19
8	От 22 до 24	7	7
9	От 25 до 27	2	4
10	От 28 до 30	5	1
11	Свыше 30	1	6

Для расчёта статистики критерия Пирсона рассчитаем относительные частоты встречаемости предложений с различным количеством слов и представим их в виде таблицы:

№ п/п	$n_i^{(X)} / n_X$	$n_i^{(Y)} / n_Y$	$n_i^{(X)} + n_i^{(Y)}$
1	0,08	0,07	15
2	0,15	0,08	23
3	0,25	0,13	38
4	0,21	0,13	34
5	0,09	0,16	25
6	0,04	0,6	10
7	0,03	0,19	22
8	0,07	0,07	14
9	0,08	0,11	19

Последние три строки таблицы были объединены в одну из-за малого числа наблюдений в них.

Используя формулу (5.24), рассчитываем выборочное значение статистики Пирсона: $z_{n_X,n_Y} \approx 22,3$. При условии истинности основной гипотезы статистика $Z_{n_X,n_Y} \sim \chi^2(9-1) = \chi^2(8)$.

По таблице квантилей распределения «хи-квадрат» с восемью степенями свободы (табл. П3 приложения) находим квантиль на уровне $1 - \alpha$:

$$z_{0,95} = 15,5.$$

Таким образом, критическая область $\Omega' = [15,5; +\infty)$. Поскольку $z = 22,3 \in \Omega'$, то основная гипотеза H_0 не согласуется с экспериментальными данными и должна быть отвергнута. Количество слов в предложениях, встречающихся в рассматриваемых произведениях, имеет значимо различные законы распределения.

Контрольные вопросы и задачи

1. Какие статистические критерии называются двухвыборочными?
2. В чём состоит двухвыборочный критерий Колмогорова?
3. В чём состоит двухвыборочный критерий «омега-квадрат»?
4. В чём состоит двухвыборочный критерий Пирсона?
5. Проведите сравнительный анализ двухвыборочных критериев Колмогорова, «омега-квадрат» и Пирсона.

Глава 6. АНАЛИЗ СТАТИСТИЧЕСКИХ ВЗАИМОСВЯЗЕЙ

§ 25. Виды связей между величинами

При изучении объектов и явлений исследователю, как правило, приходится иметь дело с несколькими некоторым образом связанными статистическими признаками. Например, объём продукции предприятия связан с численностью работников, мощностью оборудования, стоимостью производственных фондов и еще многими другими величинами. Признаки «пол» и «число лейкоцитов в крови» могли бы рассматриваться как зависимые, если бы большинство мужчин имело высокий уровень лейкоцитов, а большинство женщин – низкий, или наоборот. Рост связан с весом, потому что обычно высокие индивиды тяжелее низких; IQ (коэффициент интеллекта) связан с количеством ошибок в тесте, так как люди с высоким уровнем IQ, как правило, делают меньше ошибок и т.д.

Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других особенностей связей. Поэтому методы исследования, направленные на измерение связей, составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистического.

При исследовании причинно-следственных связей статистические признаки разделяют на факторные и результативные. *Факторные признаки*, или *факторы*, – это признаки, обусловливающие изменение других, связанных с ними, признаков. *Результативными* называются признаки, изменяющиеся под воздействием факторных признаков.

Различают два типа связей между факторными и результативными признаками: функциональную и статистическую. *Функциональной* называют такую связь, при которой каждому определённому значению x факторного признака соответствует одно и только одно значение y результативного признака:

$$y = f(x).$$

Функциональная связь двух величин возможна лишь при условии, что вторая из них зависит только от первой и ни от чего более. Такие связи являются абстракциями, в реальной жизни они встречаются редко, но находят широкое применение в точных науках, и в первую очередь в математике. Например, зависимость площади круга от радиуса $S(R) = \pi R^2$.

Функциональная зависимость результативного признака y от многих факторов $x_1, \ldots, x_k$ возможна только в том случае, если признак y всегда зависит от перечисленного набора факторов и ни от чего более. Такие связи также являются абстракциями, поскольку большинство явлений и процессов безграничного реального мира связано между собой, и нет такого конечного числа переменных, которые абсолютно полно определяли бы собою зависимую величину. Тем не менее, на практике нередко используют представление реальных связей как функциональных. Например, продолжительность года (период обращения Земли вокруг Солнца) почти функционально зависит только от массы Солнца и расстояния Земли от него. На самом деле она зависит в очень слабой степени и от масс, и от расстояний других планет от Земли, но вносимые ими (и тем более в миллионы раз более далекими звездами) искажения функциональной связи для всех практических целей, кроме космонавтики, пренебрежимо малы.

Статистической связью между результативным и факторным признаками называется связь, при которой каждому определённому значению x факторного признака соответствует некоторое распределение $F_Y(y \mid x)$ вероятностей значений результативного признака y.

Такие связи имеют место, например, если на результативный признак действуют несколько факторных признаков, а для описания связи используется один или несколько определяющих (учтённых) факторов.

Частным случаем статистической связи между результативным и факторным признаками y и x является *корреляционная связь*. При корреляционной связи от значения x факторного признака зависит не всё распределение вероятностей $F_Y(y)$, а лишь математическое ожидание величины Y. Математическое ожидание случайной вели-

чины Y при фиксированном значении случайной величины $X = x$ называется *условным математическим ожиданием* и обозначается $\mathbf{M}[Y \mid x]$, а уравнение

$$\mathbf{M}[Y \mid x] = f(x)$$

называется *уравнением регрессии Y на X*.

В зависимости от типа рассматриваемых статистических признаков для анализа статистических связей между ними используются различные статистические методы (табл. 6.1).

Таблица 6.1

Методы исследования статистических связей

Факторный признак	Результативный признак	
	Номинальный	Количественный
Номинальный	Таблицы сопряжённости	Дисперсионный анализ
Количественный	Таблицы сопряжённости	Корреляционный, регрессионный анализы

Для анализа степени тесноты связи между *количественными* факторным и результативным признаками, т.е. признаками, варианты которых имеют числовое выражение, используются методы *корреляционного анализа*, для анализа уравнения регрессии – методы *регрессионного анализа*. Корреляционный и регрессионный анализы также могут быть применены для случая качественных *порядковых*, или *ординальных*, признаков, т.е. признаков, варианты которых могут быть некоторым образом упорядочены. Для таких признаков можно сказать, какие значения больше или меньше, но нельзя сказать на сколько.

В случае если факторный признак является *номинальным* (*категориальным*, или *атрибутивным*), т.е. признаком, варианты которого могут быть измерены только в терминах принадлежности к некоторым категориям, а результативный – количественным, то

для анализа статистической связи между ними используются методы *дисперсионного анализа*.

Если же оба признака – и факторный, и результативный – номинальные, то используется метод *таблиц сопряжённости*.

Если факторный признак количественный, а результативный номинальный, то задачу, как правило, сводят к случаю двух номинальных признаков путём группировки значений факторного признака.

Контрольные вопросы и задачи

1. Какие признаки называются факторными, а какие – результативными?
2. Назовите возможные виды связей между признаками.
3. В чём отличие статистической связи от функциональной?
4. Какая связь называется корреляционной?
5. Что называется уравнением регрессии?
6. Какие признаки называются количественными, порядковыми, номинальными?
7. Какие статистические методы используются для анализа статистических связей между признаками различных видов?

§ 26. Анализ статистической связи между номинальными величинами. Таблицы сопряжённости

Пусть (X, Y) – вектор номинальных случайных величин X и Y, т.е. величин, значения которых нельзя выразить количественно (например, это может быть имя, город, национальность и т.п.). Номинальные случайные величины обязательно являются случайными величинами дискретного типа. Пусть $x_1, \ldots, x_k$ и $y_1, \ldots, y_l$ – варианты случайных величин X и Y соответственно. Распределение случайного вектора (X, Y) представим в виде таблицы (табл. 6.2).

Вероятность пары вариантов (x_i, y_j) обозначена через $p_{ij} = P(X = x_i, Y = y_j)$, $i = \overline{1,k}$, $j = \overline{1,l}$, а в последнем столбце и последней строке таблицы приведены маргинальные распределения случайных величин X и Y соответственно.

Таблица 6.2

Таблица распределения двумерного случайного вектора

Варианты	y_1	…	y_j	…	y_l	Σ
x_1	p_{11}	…	p_{1j}	…	p_{1l}	$\sum_{j=1}^{l} p_{1j}$
…	…	…	…	…	…	…
x_i	p_{i1}	…	p_{ij}	…	p_{il}	$\sum_{j=1}^{l} p_{ij}$
…	…	…	…	…	…	…
x_k	p_{k1}	…	p_{kj}	…	p_{kl}	$\sum_{j=1}^{l} p_{kj}$
Σ	$\sum_{i=1}^{k} p_{i1}$	…	$\sum_{i=1}^{k} p_{ij}$	…	$\sum_{i=1}^{k} p_{il}$	1

Будем считать, что признак *x* является факторным, а признак *y* – результативным. При каждом фиксированном варианте x_i случайной величины *X*, $i=\overline{1,k}$, случайная величина *Y* имеет распределение вероятностей, представленное в *i*-й строке таблицы. При отсутствии статистической связи между *X* и *Y* распределение вероятностей случайной величины *Y* не зависит от значений случайной величины *X* и совпадает с её маргинальным распределением, т.е. $\forall j=\overline{1,l}$ должно выполняться равенство:

$$P(Y=y_j \mid X=x_1)=...=P(Y=y_j \mid X=x_k)=P(Y=y_j). \qquad (6.1)$$

Используя определение условной вероятности

$$P(Y=y_j \mid X=x_i)=\frac{P(X=x_i, Y=y_j)}{P(X=x_i)}$$

и учитывая, что $P(X=x_i)=\sum_{j=1}^{l} p_{ij}$ и $P(Y=y_j)=\sum_{i=1}^{k} p_{ij}$, запишем эквивалентное равенству (6.1) условие:

$$\frac{p_{ij}}{\sum_{j=1}^{l} p_{ij}} = \sum_{i=1}^{k} p_{ij}, \quad \forall i = \overline{1,k}, \quad \forall j = \overline{1,l}. \tag{6.2}$$

Пусть $\left(x^{(1)}, y^{(1)}\right), ..., \left(x^{(n)}, y^{(n)}\right)$ – выборка наблюдений случайного вектора (X, Y) объёма n. Обозначим через n_{ij} частоту пары вариантов (x_i, y_j) в выборке, $i = \overline{1,k}$, $j = \overline{1,l}$. Таблица, составленная из этих частот, называется (*эмпирической*) *таблицей сопряжённости* (*contingency table, cross tabulation*) (табл. 6.3).

Таблица 6.3

Таблица сопряжённости

Варианты	y_1	…	y_j	…	y_l	Σ
x_1	n_{11}	…	n_{1j}	…	n_{1l}	$\sum_{j=1}^{l} n_{1j}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{il}	$\sum_{j=1}^{l} n_{ij}$
…	…	…	…	…	…	…
x_k	n_{k1}	…	n_{kj}	…	n_{kl}	$\sum_{j=1}^{l} n_{kj}$
Σ	$\sum_{i=1}^{k} n_{i1}$	…	$\sum_{i=1}^{k} n_{ij}$	…	$\sum_{i=1}^{k} n_{il}$	n

Сформулируем статистическую гипотезу об отсутствии статистической связи между случайными величинами X и Y:

$$H_0 : F_Y(y \mid X = x_1) = ... = F_Y(y \mid X = x_k) = F_Y(y), \quad H' : \neg H_0. \tag{6.3}$$

В случае если основная гипотеза H_0 верна, т.е. справедливы равенства (6.2), в таблице сопряжённости вместо наблюдаемых час-

тот n_{ij}, $i=\overline{1,k}$, $j=\overline{1,l}$, будут стоять теоретические частоты $m_{ij}=np_{ij}$:

$$m_{ij}=np_{ij}=n\sum_{j=1}^{l}p_{ij}\sum_{i=1}^{k}p_{ij}=\frac{1}{n}\sum_{j=1}^{l}n_{ij}\sum_{i=1}^{k}n_{ij}\,, \qquad (6.4)$$

из которых можно составить теоретическую таблицу сопряжённости.

Для проверки статистической гипотезы (6.3) используется критерий, основанный на оценке степени близости между частотами в эмпирической и теоретической таблицах сопряжённости. В качестве меры рассогласования используется статистика

$$Z=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\,, \qquad (6.5)$$

для которой показано, что при условии истинности основной гипотезы H_0 при $n\to\infty$ её закон распределения стремится к распределению «хи-квадрат» с $(k-1)(l-1)$ степенями свободы. На практике закон распределения статистики критерия Z может быть аппроксимирован с высокой точностью законом $\chi^2\left((k-1)(l-1)\right)$, если выполняется условие $m_{ij}>5$ для всех $i=\overline{1,k}$, $j=\overline{1,l}$.

В связи с тем, что основная гипотеза H_0 должна отвергаться при больших рассогласованиях между частотами в эмпирической и теоретической таблицах сопряжённости, то критическая область для статистики критерия Z должна выбираться правосторонней.

Статистика критерия (6.5) может быть применена также для анализа значимости статистической связи между двумя количественными признаками. В этом случае признаки должны быть предварительно сгруппированы, а результаты группировки представлены в виде корреляционной таблицы (см. § 5).

Пример 6.1. Исследуется эффект от действия прививки против гриппа на факт заболеваемости в возрастной категории 20–40 лет. По результатам выборочного наблюдения 850-ти участников исследования получены следующие данные:

Прививка	Заболеваемость	
	Не заболели	Заболели
Не прививались	240	150
Прививались	375	85

Определить, является ли статистически значимым эффект от прививки на уровне значимости $\alpha = 0{,}01$.

Фактически требуется проверить, имеется ли статистическая связь между факторным признаком X – «прививка» – и результативным признаком Y – «факт заболевания». Сформулируем основную гипотезу об отсутствии статистической связи между этими признаками:

$$H_0 : F_Y(y \mid X = \text{"не прив."}) = F_Y(y \mid X = \text{"прив."}) = F_Y(y).$$

Построим теоретическую таблицу сопряжённости:

Прививка	Заболеваемость		Σ
	Не заболели	Заболели	
Не прививались	282,2	107,8	390
Прививались	332,8	127,2	460
Σ	615	235	850

При расчёте частот в теоретической таблице сопряжённости достаточно рассчитать лишь одно значение теоретической частоты, например $m_{11} = \dfrac{390 \cdot 615}{850} \approx 282,2$, остальные частоты могут быть восстановлены из условия сохранения суммарных частот в последней строке и в последнем столбце таблицы.

Рассчитаем выборочное значение статистики критерия Z:

$$z = \frac{(240 - 282,2)^2}{282,2} + \ldots + \frac{(85 - 127,2)^2}{127,2} \approx 42,1.$$

При условии истинности основной гипотезы статистика $Z \sim \chi^2\left((2-1)\cdot(2-1)\right) = \chi^2(1)$.

По таблице распределения «хи-квадрат» с одной степенью свободы (табл. П3 приложения) находим квантиль на уровне $1 - \alpha$:

$$z_{0,99} = 6,63.$$

Таким образом, критическая область $\Omega' = [6,63; +\infty)$. Поскольку $z = 42,1 \in \Omega'$, то гипотеза H_0 не согласуется с экспериментальными данными и должна быть отклонена. Следовательно, фактор «прививка» оказывает влияние на распределение результативного признака «факт заболевания», т.е. между этими признаками имеется значимая статистическая связь на выбранном уровне значимости.

Контрольные вопросы и задачи

1. Что называется таблицей сопряжённости?
2. Для проверки какой статистической гипотезы используется метод таблиц сопряжённости?
3. Объясните принцип построения теоретической таблицы сопряжённости.
4. Какая статистика критерия используется в методе таблиц сопряжённости? Какой закон распределения она имеет при условии истинности основной гипотезы?
5. Объясните принцип выбора типа критической области в методе таблиц сопряжённости.
6. Возможно ли использование таблиц сопряжённости для анализа статистической связи между двумя количественными признаками?

§ 27. Виды дисперсий в совокупности, разделённой на части

Пусть исследуемая генеральная совокупность разделена по некоторому номинальному признаку на группы. Например, при исследовании доходов предприятий в различных регионах страны множество предприятий разделено на группы по признаку «территориальное расположение», при исследовании качества продукции генеральная совокупность разделена на группы по признаку «производитель» и т.п. Пусть в каждой группе проведено выборочное наблюдение, в результате которого получена выборка значений интересующего количественного признака.

Ставится задача определить, есть ли значимая статистическая связь между группировочным признаком (фактором) и результативным признаком.

Введём следующие обозначения: G – номинальный группировочный признак, имеющий k вариантов; X – количественный результативный признак; $x_1^{(i)}, ..., x_{n_i}^{(i)}$ – выборка наблюдений случайной величины X объёма n_i, соответствующая i-му варианту группировочного признака, $i = \overline{1,k}$. Для выборки из каждой группы могут быть рассчитаны выборочные характеристики:

$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^{(i)}$ – частное (групповое) среднее, $i = \overline{1,k}$;

$\tilde{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} (x_j^{(i)} - \overline{x}_i)^2$ – частная (групповая) дисперсия, $i = \overline{1,k}$.

Выборочные характеристики объединённой выборки $x_1^{(1)}, ..., x_{n_1}^{(1)}$, …, $x_1^{(k)}, ..., x_{n_k}^{(k)}$:

$\overline{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_j^{(i)}$ – общее среднее;

$D_X^* = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_j^{(i)} - \overline{x})^2$ – общая дисперсия,

где $n = \sum_{i=1}^{k} n_i$ – общий объём выборки.

Несложно показать, что общее среднее представляет собой среднее арифметическое групповых средних, взвешенное объёмами выборок:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{x}_i .$$

Аналогично введём среднее арифметическое групповых дисперсий, взвешенное объёмами выборок:

$$D_{\text{внутр}}^* = \frac{1}{n}\sum_{i=1}^{k} n_i \tilde{\sigma}_i^2 . \qquad (6.6)$$

Величина, рассчитываемая по формуле (6.6), называется *внутригрупповой дисперсией* выборок $x_1^{(i)},...,x_{n_i}^{(i)}$, $i=\overline{1,k}$.

Выборочная общая дисперсия D_X^* является мерой разброса (вариации) выборочных данных объединённой выборки относительно общего среднего, внутригрупповая дисперсия $D_{внутр}^*$ – мерой разброса данных внутри каждой группы относительно соответствующего группового среднего. Мерой разброса групповых средних относительно общего среднего является *межгрупповая дисперсия*, определяемая выражением

$$D_{\text{меж}}^* = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 . \qquad (6.7)$$

Можно показать, что для внутригрупповой, межгрупповой и общей дисперсий справедливо *правило сложения дисперсий*:

$$D_X^* = D_{\text{внутр}}^* + D_{\text{меж}}^* . \qquad (6.8)$$

Правило сложения дисперсий имеет следующую интерпретацию: общая вариация результативного признака *X* складывается из его вариации внутри каждой группы (при каждом фиксированном значении группировочного признака *G*) и вариации групповых средних. Вариация значений признака *X* внутри каждой группы не может быть обусловлена признаком *G* (поскольку внутри каждой группы он имеет фиксированное значение) и связана с действием других факторов, называемых *остаточными*. В то же время вариация групповых средних связана именно с действием фактора *G*. Таким образом, может быть предложна ещё одна интерпретация правила сложения дисперсий: *вариация результативного признака складывается из вариации, обусловленной действием остаточных факторов, и вариации, связанной с группировочным признаком*.

Отношение межгрупповой дисперсии к общей дисперсии называется *эмпирическим коэффициентом детерминации* (*ЭКД*):

$$\eta_{\text{эмп}}^2 = \frac{D_{\text{меж}}^*}{D_X^*} . \qquad (6.9)$$

Возможные значения ЭКД: $0 \le \eta_{\text{эмп}}^2 \le 1$. ЭКД показывает, какая доля в общей вариации результативного признака *X* связана с дей-

ствием группировочного признака *G*. ЭКД называют также *показателем* «*эта-квадрат*» (*eta-squared*).

Отношение межгруппового среднеквадратичного отклонения к общему среднеквадратичному отклонению называется *эмпирическим корреляционным отношением* (*ЭКО*):

$$\eta_{эмп} = \sqrt{\frac{D^*_{меж}}{D^*_X}}\,. \qquad (6.10)$$

Возможные значения ЭКО: $0 \le \eta_{эмп} \le 1$. На основе ЭКО судят о степени тесноты статистической связи между факторным признаком *G* и результативным признаком *X*. Для характеристики степени тесноты связи может быть использована *шкала Чеддока* (*R.E. Chaddock*, 1925) (табл. 6.4).

Таблица 6.4

Шкала Чеддока

$\eta_{эмп}$	Степень тесноты связи
0,1–0,3	Слабая
0,3–0,5	Умеренная
0,5–0,7	Заметная
0,7–0,9	Высокая
0,9–0,99	Сильная
0,99–1	Функциональная

Пример 6.2. Для оценки производительности микропроцессора применены три методики. Каждая методика включает прогон специального теста, по результатам которого рассчитывается производительность. Для каждой методики прогон тестов выполнен многократно, рассчитанные средние значения производительности и среднеквадратичного отклонения приведены в таблице:

Методика	Средняя производительность	С.к.о	Количество тестов
1	1544	12	25
2	1606	24	15
3	1572	18	10

Можно ли утверждать, что методика измерения производительности микропроцессоров влияет на результаты измерений?

Фактически требуется определить, в какой степени фактор «методика измерения» оказывает влияние на признак «производительность». Для оценки степени тесноты статистической связи рассчитаем ЭКД и ЭКО.

Число групп $k=3$, общее количество проведённых тестов $n=25+15+10=50$. Общая средняя производительность:

$$\overline{x}=\frac{1}{50}\left(25\cdot 1544+15\cdot 1606+10\cdot 1572\right)=1568,2.$$

Внутригрупповая дисперсия:

$$D^{*}_{\text{внутр}}=\frac{1}{50}\left(25\cdot 12^{2}+15\cdot 24^{2}+10\cdot 18^{2}\right)=309,6.$$

Межгрупповая дисперсия:

$$D^{*}_{\text{меж}}=\frac{1}{50}\Big[25\cdot(1544-1568,2)^{2}+15\cdot(1606-1568,2)^{2}+$$
$$+10\cdot(1572-1568,2)^{2}\Big]=724,36.$$

Общая дисперсия:

$$D^{*}_{X}=309,6+724,36=1033,96.$$

Эмпирический коэффициент детерминации:

$$\eta^{2}_{\text{эмп}}=\frac{724,36}{1033,96}\approx 0,7.$$

Эмпирическое корреляционное отношение:

$$\eta_{\text{эмп}}=\sqrt{0,7}\approx 0,84.$$

Полученное значение ЭКД означает, что 70% общей вариации признака «производительность» связана с признаком «методика измерения». Полученное значение ЭКО позволяет утверждать, что

методика измерения оказывает высокое влияние на результаты оценки производительности.

При расчётах внутригрупповой, межгрупповой и общей дисперсий, а также ЭКД и ЭКО по результатам выборочного наблюдения необходимо иметь в виду, что все получаемые значения являются смещёнными оценками соответствующих теоретических значений, характеризующих генеральную совокупность. Показатели вариации, а также их несмещённые оценки сведены в таблицу, называемую *таблицей дисперсионного анализа* (табл. 6.5).

Таблица 6.5

Таблица дисперсионного анализа

Источник вариации	Показатель вариации	Число степеней свободы	Несмещённая оценка
Группировочный признак	$D^*_{\text{меж}}$	$k-1$	$\frac{n}{k-1}D^*_{\text{меж}}$
Остаточные признаки	$D^*_{\text{внутр}}$	$n-k$	$\frac{n}{n-k}D^*_{\text{внутр}}$
Все признаки	D^*_X	$n-1$	$\frac{n}{n-1}D^*_X$

Оценка ЭКД, рассчитываемая по формуле (6.9), имеет положительное смещение, т.е. такая оценка в среднем даёт завышенную долю объяснённой дисперсии. Однако с ростом объёма выборки величина смещения уменьшается. При малом объёме выборки вместо оценки ЭКД (6.9) рекомендуется использовать другую оценку, обладающую меньшим смещением:

$$\omega^2_{\text{эмп}} = \frac{D^*_{\text{меж}} - \frac{k-1}{n-k}D^*_{\text{внутр}}}{D^*_X + \frac{k-1}{n-k}D^*_{\text{внутр}}}. \qquad (6.11)$$

Оценка ЭКД, рассчитываемая по формуле (6.11), всегда меньше оценки, рассчитываемой по формуле (6.9).

Контрольные вопросы и задачи

1. Что показывает внутригрупповая дисперсия результативного признака?
2. Что показывает межгрупповая дисперсия результативного признака?
3. Докажите правило сложения дисперсий.
4. Предложите интерпретацию правила сложения дисперсий, используя понятия вариации факторного и результативного признаков.
5. Что показывают эмпирический коэффициент детерминации и эмпирическое корреляционное отношение? Какие значения могут принимать эти показатели?
6. Что можно сказать о влиянии группировочного признака на результативный, если значение ЭКД, рассчитанное по выборке, оказалось равным 0? Равным 1?
7. Для чего используется шкала Чеддока?
8. В чём недостаток оценки $\eta^2_{\text{эмп}}$ эмпирического коэффициента детерминации? Какую оценку предпочтительнее использовать для случая выборки малого объёма?

§ 28. Однофакторный дисперсионный анализ

При исследовании влияния номинального группировочного признака G на количественный результативный признак X задача проверки значимости статистической связи между этими признаками в некоторых случаях может быть сведена к задаче проверки статистической гипотезы о равенстве математических ожиданий случайных величин $X_1, \ldots, X_k$, соответствующих каждому варианту группировочного признака G. Для проверки такой гипотезы используется *дисперсионный анализ* (*Analysis of Variance, ANOVA*).

Поскольку рассматривается единственный группировочный признак G (фактор), то дисперсионный анализ называется *однофакторным* (*one-way ANOVA*).

Пусть $x_1^{(i)},...,x_{n_i}^{(i)}$ – выборка объёма n_i из i-й группы, т.е. результаты наблюдений случайной величины X_i, $i=\overline{1,k}$. В дисперсионном анализе выдвигаются следующие предположения:

1) все случайные величины X_1, …, X_k имеют нормальное распределение;

2) выборки из каждой группы являются независимыми;

3) дисперсии случайных величин X_1, …, X_k равны (такие случайные величины называются *гомоскедастичными*).

Учитывая эти предположения, гипотеза об отсутствии статистической связи между группировочным и результативным признаками

$$H_0 : F_{X_1}(x) = ... = F_{X_k}(x) = F_X(x)$$

эквивалентна гипотезе о равенстве математических ожиданий:

$$H_0 : m_1 = ... = m_k,$$

$$H' : \neg H_0.$$

Для проверки этой гипотезы используется статистика

$$F = \frac{D^*_{\text{меж}} / (k-1)}{D^*_{\text{внутр}} / (n-k)}, \qquad (6.12)$$

которая при условии истинности основной гипотезы H_0 имеет распределение Фишера $F(k-1, n-k)$. Фактически статистика F представляет собой отношение несмещённых оценок межгрупповой и внутригрупповой дисперсий. При наличии статистической связи между группировочным и исследуемым признаками (случай отклонения гипотезы H_0) межгрупповая дисперсия много больше внутригрупповой, из чего следует, что критическая область должна выбираться правосторонней.

Дисперсионный анализ слабо чувствителен к требованию о нормальности распределения наблюдаемых случайных величин при больших и сбалансированных объёмах выборок, а нарушение требования гомоскедастичности наблюдаемых случайных величин может приводить к росту вероятности ошибки второго рода при принятии статистического решения.

Пример 6.3. Исследуется производительность трёх процессоров при выполнении специальных тестов. Результаты измерений представляются в баллах. В таблице приведены средние баллы и среднеквадратичные отклонения, рассчитанные по результатам многократных тестирований:

Тип процессора	Средняя производительность	С.к.о	Количество тестов
Intel Core 2 Duo E6600	1582	24	10
Intel Core 2 Duo E6700	1663	32	8
Intel Core 2 Extreme X6800	1716	27	12

Предполагая, что результаты измерений производительности каждого процессора имеют нормальное распределение с равными дисперсиями, проверить, имеется ли значимое различие в средних производительностях исследуемых процессоров на уровне значимости $\alpha = 0{,}01$.

Сформулируем основную и альтернативную гипотезы:

$$H_0 : m_1 = m_2 = m_3,$$

$$H' : \neg H_0,$$

где m_1, m_2, m_3 – математические ожидания производительностей процессоров.

Число групп $k = 3$, общее количество тестов $n = 10 + 8 + 12 = 30$. Общая средняя производительность:

$$\overline{x} = \frac{1}{30}\left(10 \cdot 1582 + 8 \cdot 1663 + 12 \cdot 1716\right) = 1657{,}2.$$

Внутригрупповая дисперсия:

$$D^*_{\text{внутр}} = \frac{1}{30}\left(10 \cdot 12^2 + 8 \cdot 32^2 + 12 \cdot 27^2\right) \approx 756{,}7.$$

Межгрупповая дисперсия:

$$D^*_{\text{меж}} = \frac{1}{30}\Big[10\cdot(1582-1657,2)^2 + 8\cdot(1663-1657,2)^2 +$$

$$+12\cdot(1716-1657,2)^2\Big] \approx 3277.$$

Выборочное значение статистики Фишера:

$$f = \frac{3277/(3-1)}{756,7/(30-3)} \approx 58,5\,.$$

По таблице математической статистики (табл. П5 приложения) находим квантиль распределения Фишера с 2 и 27 степенями свободы в числителе и знаменателе соответственно на уровне $1 - \alpha$:

$$f_{0,99}(2;27) = 5,49\,.$$

Таким образом, критическая область $\Omega' = [5,49;+\infty)$. Поскольку $f \in \Omega'$, то основная гипотеза H_0 должна быть отвергнута, т.е. различие в средних производительностях процессоров значимо.

В частном случае число вариантов k группировочного признака может быть равно 2. Тогда основная гипотеза дисперсионного анализа представляет собой двухвыборочную параметрическую гипотезу (см. § 16):

$$H_0 : m_1 = m_2,$$

для проверки которой может быть использован критерий Стьюдента (см. табл. 4.1).

Основная гипотеза H_0 дисперсионного анализа состоит в том, что математические ожидания в каждой из k групп равны против альтернативной гипотезы, состоящей в том, что математические ожидания хотя бы в двух группах окажутся различными. Такая альтернатива включает множество вариантов. Основная гипотеза дисперсионного анализа будет отклонена как в случае значимого различия математических ожиданий лишь в двух группах, так и в случае значимого различия математических ожиданий всех групп.

Если основная гипотеза H_0 в результате дисперсионного анализа отклоняется, то бывает необходимо узнать, какие именно математические ожидания значимо отличаются, а какие равны. Возможным способом такой проверки является проведение попарных

сравнений математических ожиданий для каждой пары групп, т.е. проверка множества статистических гипотез вида

$$H_0 : m_i = m_j,$$

$$H' : m_i \neq m_j,$$

где $i = \overline{1,k}$, $j = \overline{1,k}$, $i \neq j$.

Однако такой способ проверки имеет существенный недостаток. При проверке одной параметрической гипотезы задаётся некоторый уровень значимости α, определяющий вероятность ошибки первого рода, т.е. отклонения основной гипотезы при условии её истинности. При проверке множества параметрических гипотез, каждую на уровне значимости α, вероятность ошибочно обнаружить различие в математических ожиданиях будет расти с числом проверяемых гипотез.

Вероятность α_k ошибки первого рода при проверке k независимых статистических гипотез равна

$$\alpha_k = 1 - (1 - \alpha)^k.$$

В случае зависимых гипотез может быть рассчитана оценка эффективной вероятности α_k ошибки первого рода, используя различные корректирующие поправки (например, поправку Бонферрони).

Для того чтобы обеспечить заданную вероятность α ошибки первого рода при проверке множества параметрических гипотез вида $H_0 : m_i = m_j$, $i = \overline{1,k}$, $j = \overline{1,k}$, $i \neq j$, на практике используются *методы множественного сравнения* (*multiple comparison tests*).

Одним из методов множественного сравнения является *метод Шеффе* (*Henry Scheffe*, 1953), называемый также *методом линейных контрастов*. С помощью метода Шеффе проверяется основная гипотеза вида

$$H_0 : \sum_{i=1}^{k} \beta_i m_i = 0,$$

$$H' : \sum_{i=1}^{k} \beta_i m_i \neq 0,$$

где $\beta_1, \ldots, \beta_k$ – весовые коэффициенты, причём $\sum_{i=1}^{k}\beta_i = 0$. Величина $c = \sum_{i=1}^{k}\beta_i m_i$ называется *линейным контрастом*.

Для проверки гипотезы H_0 используем метод доверительных интервалов. Точечной оценкой линейного контраста является линейная комбинация групповых средних

$$\tilde{C} = \sum_{i=1}^{k}\beta_i \bar{X}_i ,$$

которая для конкретной выборки примет выборочное значение

$$\tilde{c} = \sum_{i=1}^{k}\beta_i \bar{x}_i .$$

Можно показать, что оценка дисперсии линейного контраста равна

$$\tilde{\sigma}_C^2 = \frac{n}{n-k} D^*_{\text{внутр}} \sum_{i=1}^{k}\frac{\beta_i^2}{n_i} ,$$

а границы доверительного интервала имеют вид

$$\tilde{C} \pm \tilde{\sigma}_C \sqrt{(k-1) f_{1-\alpha}(k-1, n-k)} ,$$

где $f_{1-\alpha}(k-1, n-k)$ – квантиль распределения Фишера с $k-1$ и $n-k$ степенями свободы в числителе и знаменателе соответственно на уровне значимости $1 - \alpha$.

Если доверительный интервал накрывает нулевое значение, то нет оснований отвергать основную гипотезу H_0 о равенстве нулю линейного контраста.

В частном случае при $\beta_i = -\beta_j$, $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, k\}$, $i \neq j$, и остальных нулевых коэффициентах линейный контраст $C = m_i - m_j$, а проверяемая гипотеза имеет вид

$$H_0 : m_i = m_j ,$$

$$H' : m_i \neq m_j .$$

Пример 6.4. В условиях примера 6.3 определить, в каких именно средних производительностях процессоров имеются значимые различия.

Проверим три гипотезы о попарном равенстве математических ожиданий производительности процессоров на уровне значимости α:

$$H_0^{12}: m_1 = m_2, \qquad H_0^{13}: m_1 = m_3, \qquad H_0^{23}: m_2 = m_3,$$
$$H': m_1 \neq m_2; \qquad H': m_1 \neq m_3; \qquad H': m_2 \neq m_3.$$

Для линейных контрастов $c_{12} = m_1 - m_2$, $c_{13} = m_1 - m_3$, $c_{23} = m_2 - m_3$ рассчитаем выборочные значения и оценки дисперсий:

$$\tilde{c}_{12} = \overline{x}_1 - \overline{x}_2 = 1582 - 1663 = -81;$$
$$\tilde{c}_{13} = \overline{x}_1 - \overline{x}_3 = 1582 - 1716 = -134;$$
$$\tilde{c}_{23} = \overline{x}_2 - \overline{x}_3 = 1663 - 1716 = -53;$$
$$\tilde{\sigma}_{12}^2 = \frac{30 \cdot 756,7}{27}\left(\frac{1}{10} + \frac{1}{8}\right) \approx 189,2;$$
$$\tilde{\sigma}_{13}^2 = \frac{30 \cdot 756,7}{27}\left(\frac{1}{10} + \frac{1}{12}\right) \approx 154,1;$$
$$\tilde{\sigma}_{23}^2 = \frac{30 \cdot 756,7}{27}\left(\frac{1}{8} + \frac{1}{12}\right) \approx 175,1.$$

Квантиль распределения Фишера $f_{0,99}(2; 27) = 5,49$.

Границы доверительных интервалов для линейных контрастов:

$$c_{12}: \quad -81 \pm \sqrt{189,2 \cdot 2 \cdot 5,49} \approx -81 \pm 45,5 = (-126,6; -35,5);$$
$$c_{13}: \quad -134 \pm \sqrt{154,2 \cdot 2 \cdot 5,49} \approx -134 \pm 41,1 = (-175,1; -92,9);$$
$$c_{23}: \quad -53 \pm \sqrt{175,2 \cdot 2 \cdot 5,49} \approx -53 \pm 43,8 = (-96,9; -9,2).$$

Ни один из доверительных интервалов не накрыл нулевое значение, следовательно, все выдвинутые гипотезы должны быть *одновременно* отклонены на уровне значимости $\alpha = 0,01$. Таким образом, значимое различие в средних производительностях имеется для каждой пары исследуемых процессоров.

Контрольные вопросы и задачи

1. Какая задача ставится в дисперсионном анализе?
2. Какие предположения о свойствах выборки выдвигаются в дисперсионном анализе? Как гипотеза дисперсионного анализа связана со статистической независимостью исследуемых признаков?
3. Какая статистика критерия используется в дисперсионном анализе? Какой закон распределения она имеет при условии истинности основной гипотезы? Какой тип критической области выбирается при проверке гипотезы дисперсионного анализа?
4. Покажите, что распределение Фишера с одной степенью свободы в числителе равно квадрату распределения Стьюдента: $F(1, n) = T^2(n)$.
5. Покажите, что при числе вариантов группировочного признака $k = 2$ статистика Фишера равна квадрату статистики Стьюдента, используемой при проверке статистической гипотезы о равенстве математических ожиданий двух выборок при неизвестных, но равных дисперсиях.
6. Объясните, почему проверка гипотезы дисперсионного анализа не может быть сведена к проверке по критерию Стюдента множества гипотез о попарном равенстве математических ожиданий в группах.
7. Какую задачу решают методы множественного сравнения?
8. В чем состоит метод Шеффе? Что называется линейным контрастом? Какое ограничение накладывается на его весовые коэффициенты? Как проводится проверка гипотезы о нулевом значении линейного контраста?

§ 29. Статистическая связь между компонентами нормально распределённого случайного вектора

Частным случаем статистической связи между количественными признаками *x* и *y* является корреляционная связь. При корреляционной связи от значения *x* факторного признака зависит лишь условное математическое ожидание $\mathbf{M}[Y \mid x]$ случайной величины *Y*, при этом все остальные характеристики её распределения оста-

ются неизменными. Функция $f(x)$, описывающая эту зависимость, называется *функцией регрессии Y на X*.

Частным случаем корреляционной связи между признаками *x* и *y* является *линейная корреляционная связь*, когда функция регрессии *Y* на *X* линейно зависит от *x*:

$$f(x) = \beta_0 + \beta_1 x. \tag{6.13}$$

Пусть случайный вектор $Z = (X, Y)^T$ имеет двумерное нормальное распределение, $N_2(m, C)$, где $m = (m_X, m_Y)^T$ – вектор математических ожиданий; $C = \begin{pmatrix} \sigma_X^2 & k_{XY} \\ k_{YX} & \sigma_Y^2 \end{pmatrix}$ – ковариационная матрица. Покажем, что если между случайными величинами *X* и *Y* есть статистическая связь, то она является линейной корреляционной.

Запишем двумерную функцию плотности распределения случайного вектора *Z* в матричном виде:

$$f_Z(z) = \frac{1}{2\pi\sqrt{\det C}} \exp\left\{-\frac{1}{2}(z-m)^T C^{-1}(z-m)\right\}, \tag{6.14}$$

где $z = (x, y)^T$.

В скалярном виде:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left\{-\frac{1}{(1-\rho_{XY}^2)} \times \right.$$
$$\left. \times \left[\frac{(x-m_X)^2}{2\sigma_X^2} - \rho_{XY}\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{2\sigma_Y^2}\right]\right\}, \tag{6.15}$$

где $\rho_{XY} = \dfrac{k_{XY}}{\sigma_X\sigma_Y}$ – коэффициент корреляции.

В результате интегрирования двумерной функции плотности (6.15) по соответствующей переменной получим маргинальные распределения случайных величин *X* и *Y*:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left\{-\frac{(x-m_X)^2}{2\sigma_X^2}\right\}, \tag{6.16}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left\{-\frac{(y-m_Y)^2}{2\sigma_Y^2}\right\}, \qquad (6.17)$$

Из равенств (6.15), (6.16) и (6.17) видно, что в случае если $\rho_{XY} = 0$, выполняется тождество

$$f_{XY}(x,y) = f_X(x)\, f_Y(y), \qquad (6.18)$$

из которого следует, что *если нормально распределённые случайные величины X и Y некоррелированы, то они независимы*. Для произвольного закона распределения это утверждение в общем случае неверно.

По определению условная плотность распределения случайной величины *Y* равна

$$f_Y(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}. \qquad (6.19)$$

Подставляя выражения (6.15) и (6.16), получим

$$f_Y(y \mid x) = \frac{1}{\sigma_{Y|X}\sqrt{2\pi}} \exp\left\{-\frac{(y-m_{Y|X}(x))^2}{2\sigma_{Y|X}^2}\right\}, \qquad (6.20)$$

где использованы обозначения

$$m_{Y|X}(x) = m_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - m_X), \qquad (6.21)$$

$$\sigma_{Y|X} = \sigma_Y\sqrt{1-\rho_{XY}^2}. \qquad (6.22)$$

Из (6.20)–(6.22) следует, что при любом фиксированном значении *x* случайная величина *Y* распределена нормально с математическим ожиданием $m_{Y|X}(x)$ и постоянной дисперсией $\sigma_{Y|X}^2$. Это означает, что статистическая связь между величинами *X* и *Y* является корреляционной, а поскольку $m_{Y|X}(x) = \mathbf{M}\left[Y \mid x\right]$ линейно зависит от *x*, то функция регрессии *Y* на *X* линейна. Это означает, что *статистическая связь между нормально распределёнными случайными величинами может быть только линейной корреляционной*. Из (6.21) получаем коэффициенты линейной регрессии:

$$\beta_0 = m_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}m_X,$$
$$\beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}. \qquad (6.23)$$

Из этих равенств видно, что функция регрессии не зависит от x ($\beta_1 = 0$), если коэффициент корреляции $\rho_{XY} = 0$. Таким образом, *некоррелированность нормально распределённых случайных величин означает отсутствие статистической связи между ними*.

Можно показать, что если $|\rho_{XY}| = 1$, то между случайными величинами X и Y имеется линейная функциональная связь:

$$Y = \pm\frac{\sigma_Y}{\sigma_X}X + \left(m_Y \mp \frac{\sigma_Y}{\sigma_X}m_X\right). \qquad (6.24)$$

Следовательно, коэффициент корреляции можно рассматривать как показатель тесноты статистической связи между нормально распределёнными случайными величинами X и Y.

На рис. 6.1 показаны примеры диаграмм рассеяния (в корреляционном анализе называемые *корреляционными полями*) двумерных нормальных распределений с нулевыми математическими ожиданиями и с.к.о. компонентов $\sigma_X = 1$, $\sigma_Y = 1$ при различных значениях коэффициента корреляции ρ_{XY}.

С изменением коэффициента корреляции от –1 до 1 прямая регрессии изменяет угол наклона от минимального значения $-\arctan\left(\frac{\sigma_Y}{\sigma_X}\right) = -\frac{\pi}{4}$ до максимального $\arctan\left(\frac{\sigma_Y}{\sigma_X}\right) = \frac{\pi}{4}$, кроме того, изменяется степень рассеяния выборочных значений относительно прямой регрессии: при $\rho_{XY} = 0$ рассеяние максимально, а при $|\rho_{XY}| = 1$ – отсутствует.

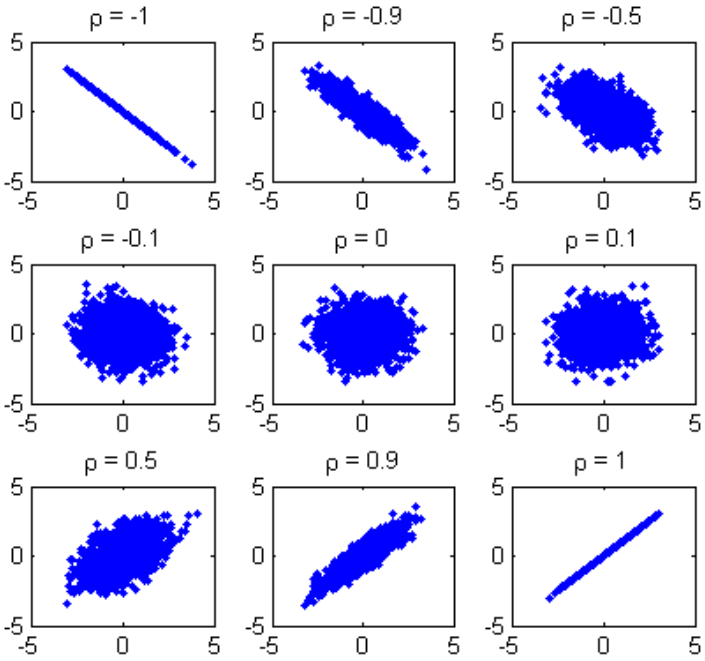

Рис. 6.1. Корреляционные поля двумерных нормальных распределений с нулевыми математическими ожиданиями и с.к.о. компонентов $\sigma_X = 1$, $\sigma_Y = 1$ при различных значениях коэффициента корреляции ρ_{XY}

Согласно (6.23), угол наклона линии регрессии зависит не только от коэффициента корреляции ρ_{XY}, но и от отношения среднеквадратичных отклонений $\frac{\sigma_Y}{\sigma_X}$. На рис. 6.2 показаны корреляционные поля двумерных нормальных распределений при фиксированном значении ρ_{XY} и различных значениях отношения $\frac{\sigma_Y}{\sigma_X}$.

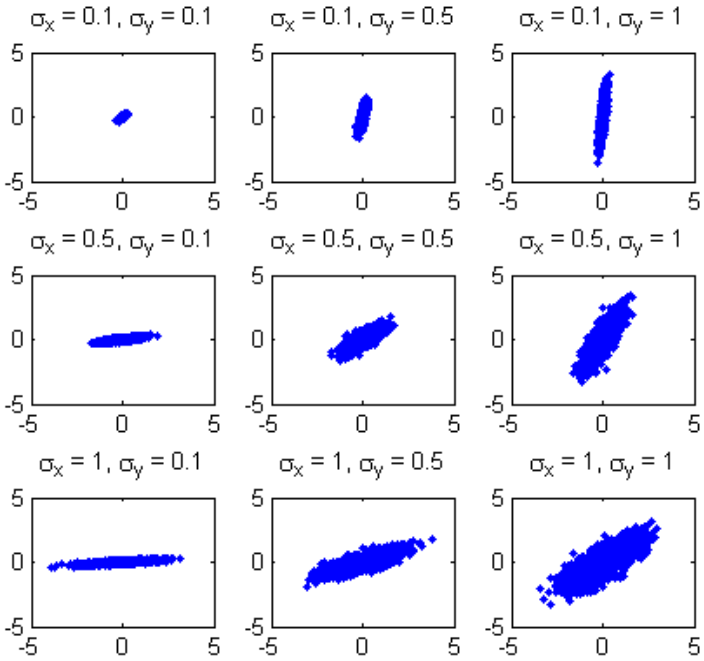

Рис. 6.2. Корреляционные поля двумерных нормальных распределений с нулевыми математическими ожиданиями и $\rho_{XY} = 0{,}8$ при различных значениях с.к.о. компонентов

Из диаграмм рассеяния видно, что угол наклона прямой регрессии может изменяться практически от 0 до $\pi/2$ в зависимости от соотношения $\dfrac{\sigma_Y}{\sigma_X}$ при фиксированном положительном значении коэффициента корреляции ρ_{XY}. Кроме того, изменяется степень рассеяния выборочных значений относительно прямой регрессии, определяемая условными среднеквадратичными отклонениями $\sigma_{Y|X}$ и $\sigma_{X|Y}$.

Контрольные вопросы и задачи

1. Как связаны понятия некоррелированности и независимости компонентов двумерного нормально распределённого случайного вектора?
2. Назовите критерий отсутствия статистической связи между компонентами двумерного нормально распределённого случайного вектора.
3. Покажите, что если $|\rho_{XY}| = 1$, то между случайными величинами X и Y имеется линейная функциональная связь вида (6.24).
4. Чем определяется степень разброса выборочных данных относительно линии регрессии Y на X?
5. В каких пределах может изменяться угол наклона линии регрессии Y на X при изменении σ_X и σ_Y, если коэффициент корреляции $\rho_{XY} < 0$?
6. В каких пределах может изменяться угол наклона линии регрессии Y на X при изменении коэффициента корреляции ρ_{XY} от –1 до 1, если $\sigma_X = 1$, $\sigma_Y = 2$?
7. Что называется корреляционным полем двумерного распределения?

§ 30. Корреляционное отношение

Если распределение случайного вектора $Z = (X, Y)^T$ отлично от нормального, то характер изменения условного математического ожидания $\mathbf{M}[Y \mid x] = f(x)$ в общем случае нелинейный, а условная дисперсия $\mathbf{D}[Y \mid x]$ зависит от x. При каждом фиксированном x условная дисперсия является мерой рассеяния условного распределения $F_Y(y \mid x)$ относительно условного математического ожидания $\mathbf{M}[Y \mid x]$, т.е. значения функции регрессии в точке x.

Рассеяние случайной величины Y относительно её математического ожидания m_Y складывается из двух слагаемых, а именно: рассеяния случайной величины Y относительно функции регрессии и рассеяния значений функции регрессии относительно математического ожидания случайной величины Y, т.е.

$$\mathbf{M}\left[(Y-m_Y)^2\right]=\mathbf{M}\left[(Y-f(X))^2\right]+\mathbf{M}\left[(f(X)-m_Y)^2\right]. \quad (6.25)$$

В этой формуле аргументом функции регрессии является случайная величина X, таким образом, случайная величина $f(X)$ образует генеральную совокупность значений функции регрессии.

Для доказательства преобразуем левую часть равенства:

$$\begin{aligned}\mathbf{M}\left[(Y-m_Y)^2\right]&=\mathbf{M}\left[\left((Y-f(X))+(f(X)-m_Y)\right)^2\right]=\\&=\mathbf{M}\left[(Y-f(X))^2\right]+\\&+2\mathbf{M}\left[(Y-f(X))(f(X)-m_Y)\right]+\\&+\mathbf{M}\left[(f(X)-m_Y)^2\right].\end{aligned}$$

Учитывая, что

$$\mathbf{M}\left[(Y-f(X))(f(X)-m_Y)\right]=0, \quad (6.26)$$

получаем верное равенство (6.25).

Величина $D_{\text{ост}\,Y}=\mathbf{M}\left[(Y-f(X))^2\right]$ характеризует степень разброса значений случайной величины Y относительно линии регрессии и называется *остаточной дисперсией случайной величины Y* (*residual variance*).

Величина $D_{\text{регр}\,Y|X}=\mathbf{M}\left[(f(X)-m_Y)^2\right]$ характеризует степень разброса значений, принадлежащих линии регрессии, относительно математического ожидания случайной величины Y и называется *дисперсией, обусловленной регрессией Y на X* (*variance explained by regression*).

Таким образом, для общей дисперсии Y, дисперсии, обусловленной регрессией Y на X, и остаточной дисперсии Y справедливо *правило сложения дисперсий*:

$$D_Y=D_{\text{ост}\,Y}+D_{\text{регр}\,Y|X}. \quad (6.27)$$

Отношение дисперсии, обусловленной регрессией Y на X, к общей дисперсии случайной величины Y называется *коэффициентом детерминации* (*КД*) *Y на X*:

$$R^2_{Y|X}=\frac{D_{\text{регр}\,Y|X}}{D_Y}. \quad (6.28)$$

Возможные значения КД: $0 \le R^2_{Y|X} \le 1$. КД показывает, какая доля в общей вариации результативного признака *Y* связана с вариацией линии регрессии. Иными словами, КД – доля вариации, *объяснённой регрессией*, в общей вариации значений признака *Y*.

Отношение среднеквадратичного отклонения, обусловленного регрессией *Y* на *X*, к среднеквадратичному отклонению случайной величины *Y* называется *корреляционным отношением* (*КО*) *Y на X*:

$$R_{Y|X} = \sqrt{\frac{D_{\text{регр } Y|X}}{D_Y}}\,. \qquad (6.29)$$

Возможные значения КО: $0 \le R_{Y|X} \le 1$. На основе КО судят о степени тесноты корреляционной связи между факторным признаком *X* и результативным признаком *Y*. Для качественной характеристики степени тесноты связи может быть использована *шкала Чеддока* (см. табл. 6.3).

Равенство $R_{Y|X} = 0$ означает, что вариация значений функции регрессии $f(x)$ при различных значениях *x* полностью отсутствует, линия регрессии является горизонтальной прямой $f(x) = \text{const}$. Другими словами, между случайными величинами *X* и *Y* отсутствует корреляционная связь.

Равенство $R_{Y|X} = 1$ будет иметь место, если остаточная дисперсия $D_{\text{ост } Y} = 0$, т.е. если вариация признака *Y* относительно линии регрессии при каждом фиксированном значении *x* полностью отсутствует. Это означает, что при каждом значении *x* значение признака *Y* однозначно определено и равно $f(x)$. Иными словами, между случайными величинами *X* и *Y* имеется функциональная связь.

Аналогично выражению (6.28) определяется КД $R^2_{X|Y}$ *X* на *Y*.

Между $R^2_{Y|X}$ и $R^2_{X|Y}$ в общем случае нет какой-либо простой зависимости. Возможны ситуации, когда один из этих показателей принимает нулевое значение, в то время как другой равен единице. Так, на рис. 6.3 приведён пример диаграммы рассеяния, когда $R^2_{Y|X} \approx 1$, в то время как $R^2_{X|Y} \approx 0$. В первом случае (на рисунке слева) линия регрессии *Y* на *X* – парабола (на рисунке сплошная ли-

ния), разброс данных относительно неё небольшой, т.е. $D_{\text{ост}\,Y} \approx 0$, в то время как разброс значений, принадлежащих линии регрессии, относительно среднего значения случайной величины Y (на рисунке горизонтальная пунктирная линия) много больше, т.е. $D_{\text{регр}\,Y|X} \approx D_Y$. Во втором случае (на рисунке справа) линия регрессии X на Y – прямая, параллельная оси Y – проходит почти на уровне среднего значения случайной величины X, т.е. $D_{\text{регр}\,X|Y} \approx 0$, в то время как разброс значений случайной величины X относительно линии регрессии практически равен разбросу относительно среднего значения случайной величины X, т.е. $D_{\text{ост}\,X} \approx D_X$.

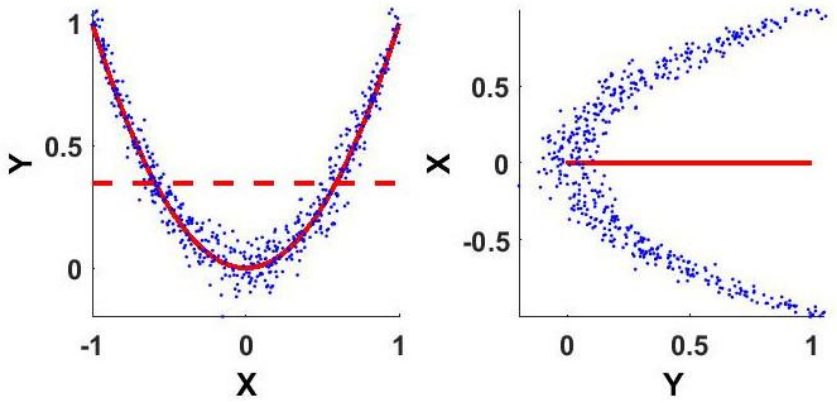

Рис. 6.3. Иллюстрация к расчёту КД $R^2_{Y|X}$ (слева) и $R^2_{X|Y}$ (справа)

Рассмотрим, как связаны между собой коэффициент корреляции ρ_{XY} и корреляционное отношение $R_{Y|X}$. Из теории вероятностей известно, что из независимости случайных величин следует их некоррелированность. В терминах статистической связи это утверждение формулируется так: если между признаками отсутствует статистическая связь, то между ними отсутствует линейная корреляционная связь. Справедливость этого утверждения очевидна, поскольку линейная корреляционная связь – частный случай стати-

стической связи. Обратное утверждение в общем случае неверно: отсутствие линейной корреляционной связи ещё не означает отсутствие статистической связи какого-либо другого вида.

В § 29 показано, что между компонентами двумерного нормально распределённого случайного вектора статистическая связь может быть лишь линейной корреляционной связью. Из этого следует, что термины «статистическая связь» и «линейная корреляционная связь» для нормального распределения эквивалентны. В теории вероятностей это утверждение известно как «из некоррелированности нормально распределённых случайных величин следует их независимость».

Рассчитаем КД для нормально распределённых случайных величин X и Y. По определению коэффициент детерминации равен

$$R^2_{Y|X} = \frac{D_{\text{регр } Y|X}}{D_Y} = 1 - \frac{D_{\text{ост } Y}}{D_Y} = 1 - \frac{\mathbf{M}\left[(Y - f(X))^2\right]}{D_Y}.$$

где $f(x) = \mathbf{M}[Y \mid x]$ – функция регрессии.

Учитывая, что условная дисперсия $\sigma^2_{Y|X}$ не зависит от x (см. § 29):

$$\mathbf{D}[Y \mid x] = \mathbf{M}\left[(Y - f(x))^2\right] = \sigma_Y \sqrt{1 - \rho^2_{XY}} = \sigma^2_{Y|X},$$

запишем

$$R^2_{Y|X} = 1 - \frac{\mathbf{D}[Y \mid X]}{D_Y} = 1 - \frac{\sigma^2_{Y|X}}{D_Y} = 1 - \frac{\sigma^2_Y\left(1 - \rho^2_{XY}\right)}{\sigma^2_Y} = \rho^2_{XY}.$$

Таким образом, для *нормально распределённых случайных величин коэффициент корреляции и корреляционное отношение совпадают с точностью до знака*. Это означает, что использование корреляционного отношения в качестве показателя статистической связи имеет смысл лишь для признаков, распределения которых отличны от нормального.

При любом законе распределения случайного вектора $Z = (X, Y)^T$ для КД и коэффициента корреляции справедливо неравенство:

$$0 \le \rho^2_{XY} \le R^2_{Y|X} \le 1, \qquad (6.30)$$

при этом возможны, в частности, следующие варианты:

а) $\rho_{XY}^2 = 0$ тогда и только тогда, когда линейная корреляционная связь между X и Y отсутствует;

б) $\rho_{XY}^2 = R_{Y|X}^2 = 1$ тогда и только тогда, когда имеется линейная функциональная связь между X и Y;

в) $\rho_{XY}^2 < R_{Y|X}^2 = 1$ тогда и только тогда, когда имеется нелинейная функциональная связь между X и Y;

г) $\rho_{XY}^2 = R_{Y|X}^2 < 1$ тогда и только тогда, когда регрессия Y на X линейна, но функциональная связь отсутствует.

Контрольные вопросы и задачи

1. Что показывает дисперсия признака Y, обусловленная регрессией Y на X?
2. Что показывает остаточная дисперсия признака Y?
3. Докажите равенство (6.26) для случайных величин X и Y: а) дискретного типа; б) непрерывного типа.
4. Докажите правило сложения дисперсий.
5. Предложите интерпретацию правила сложения дисперсий, используя понятия вариации результативного признака и вариации линии регрессии.
6. Что можно сказать о характере связи между признаками X и Y, если значение КД $R_{Y|X}^2$ оказалось равным 0? Равным 1?
7. Что можно сказать о характере связи между признаками X и Y, если значения КД:

а) $R_{Y|X}^2 \approx 0$ и $R_{X|Y}^2 \approx 1$;

б) $R_{Y|X}^2 \approx 0$ и $R_{X|Y}^2 \approx 0$;

в) $R_{Y|X}^2 \approx 1$ и $R_{X|Y}^2 \approx 1$;

г) $R_{Y|X}^2 = 0$ и $R_{X|Y}^2 = 1$;

д) $R_{Y|X}^2 = 0$ и $R_{X|Y}^2 = 0$;

е) $R_{Y|X}^2 = 1$ и $R_{X|Y}^2 = 1$;

ж) $R_{Y|X}^2 \approx R_{X|Y}^2$?

Приведите примеры диаграмм рассеяния значений признаков X и Y для каждого из этих случаев.

8. Коэффициент корреляции между нормально распределёнными случайными величинами X и Y равен $\rho_{XY} = -0,4$. Чему равны КО $R_{Y|X}$ и КО $R_{X|Y}$?

9. Предложите интерпретацию значений КО и коэффициента корреляции для произвольно распределённых признаков X и Y.

10. Что можно сказать о характере связи между признаками X и Y, если:

а) $0 = \rho_{XY}^2 = R_{Y|X}^2$;

б) $0 = \rho_{XY}^2 < R_{Y|X}^2 = 1$;

в) $0 = \rho_{XY}^2 < R_{Y|X}^2 < 1$;

г) $0 < \rho_{XY}^2 < R_{Y|X}^2 < 1$?

Приведите примеры диаграмм рассеяния значений признаков X и Y для каждого из этих случаев.

§ 31. Оценивание коэффициента корреляции по выборочным данным

Пусть $(x_1, y_1), \ldots, (x_n, y_n)$ – выборка наблюдений двумерного случайного вектора (X, Y), имеющего неизвестное распределение $F_{XY}(x, y)$.

1. Точечная оценка коэффициента корреляции. На практике в качестве точечной оценки коэффициента корреляции ρ_{XY} используется выборочный коэффициент корреляции (см. § 5):

$$\rho_{XY}^* = \frac{k_{XY}^*}{\sigma_X^* \sigma_Y^*}, \tag{6.31}$$

где k_{XY}^* – выборочный ковариационный момент:

$$k_{XY}^* = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}). \tag{6.32}$$

Выборочный коэффициент корреляции ρ_{XY}^* является состоятельной смещённой оценкой коэффициента корреляции ρ_{XY} со

смещением, равным $-\frac{\rho_{XY}(1-\rho_{XY}^2)}{2n}$. Величина смещения убывает с увеличением объёма выборки и при $n > 30$ уже становится практически пренебрежимой.

2. Интервальная оценка коэффициента корреляции. Пусть случайный вектор (*X*, *Y*) распределён по двумерному нормальному закону. В этом случае точечная оценка коэффициента корреляции ρ_{XY}^*, рассчитываемая по формуле (6.31), имеет асимптотически нормальный закон распределения с математическим ожиданием

$$\mathbf{M}\left[\rho_{XY}^*\right] \approx \rho_{XY} - \frac{\rho_{XY}\left(1-\rho_{XY}^2\right)}{2n}$$

и дисперсией

$$\mathbf{D}\left[\rho_{XY}^*\right] \approx \frac{\left(1-\rho_{XY}^2\right)^2}{n}.$$

В качестве центральной статистики при построении доверительного интервала (см. § 12) выберем стандартизованную оценку коэффициента корреляции:

$$U = \frac{\rho_{XY}^* - \mathbf{M}\left[\rho_{XY}^*\right]}{\sqrt{\mathbf{D}\left[\rho_{XY}^*\right]}} = \frac{\rho_{XY}^* - \left(\rho_{XY} - \frac{\rho_{XY}\left(1-\rho_{XY}^2\right)}{2n}\right)}{\frac{1-\rho_{XY}^2}{\sqrt{n}}} \sim N(0,1).$$

Запишем верное тождество для статистики *U*:

$$P\left(u_{\alpha/2} < \frac{\left(\rho_{XY}^* + \frac{\rho_{XY}\left(1-\rho_{XY}^2\right)}{2n}\right) - \rho_{XY}}{\frac{1-\rho_{XY}^2}{\sqrt{n}}} < u_{1-\alpha/2}\right) = 1-\alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили стандартизованного нормального распределения на уровнях $\alpha/2$ и $1-\alpha/2$ соответственно. Преобразуя неравенство под знаком вероятности, получим:

$$P\left(\rho^*_{XY}+\frac{\rho_{XY}\left(1-\rho^2_{XY}\right)}{2n}-u_{1-\alpha/2}\frac{1-\rho^2_{XY}}{\sqrt{n}}<\rho_{XY}<\right.$$

$$\left.<\rho^*_{XY}+\frac{\rho_{XY}\left(1-\rho^2_{XY}\right)}{2n}+u_{1-\alpha/2}\frac{1-\rho^2_{XY}}{\sqrt{n}}\right)=1-\alpha.$$

Это выражение ещё не даёт интервальной оценки коэффициента корреляции ρ_{XY}, так как левая и правая части неравенства под знаком вероятности содержат этот параметр. На практике в указанные части неравенств подставляют вместо неизвестного точного значения ρ_{XY} его оценку ρ^*_{XY}. В результате получается интервал

$$\left(\rho^*_{XY}+\frac{\rho^*_{XY}\left(1-(\rho^*_{XY})^2\right)}{2n}-u_{1-\alpha/2}\frac{1-(\rho^*_{XY})^2}{\sqrt{n}};\right.$$

$$\left.\rho^*_{XY}+\frac{\rho^*_{XY}\left(1-(\rho^*_{XY})^2\right)}{2n}+u_{1-\alpha/2}\frac{1-(\rho^*_{XY})^2}{\sqrt{n}}\right). \qquad (6.33)$$

Подчеркнём, что границы доверительного интервала (6.33) получены в результате аппроксимации и могут использоваться лишь при достаточно больших объёмах выборки ($n>500$).

При малых объёмах выборки границы доверительного интервала для ρ_{XY} на уровне значимости α рассчитываются по следующим приближённым формулам:

$$\left(\tanh\left(\frac{1}{2}\ln\frac{1+\rho^*_{XY}}{1-\rho^*_{XY}}+\frac{\rho^*_{XY}}{2(n-1)}-\frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right);\right.$$

$$\left.\tanh\left(\frac{1}{2}\ln\frac{1+\rho^*_{XY}}{1-\rho^*_{XY}}+\frac{\rho^*_{XY}}{2(n-1)}+\frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)\right), \qquad (6.34)$$

где $\tanh x=\dfrac{e^x-e^{-x}}{e^x+e^{-x}}$ – функция гиперболического тангенса.

Формулы (6.33) и (6.34) выведены в условиях нормальности распределения генеральной совокупности. Однако в случае отклонения от нормальности уже при объёмах выборки $n>30$ возникающая неточность расчёта практически пренебрежима.

3. Проверка значимости коэффициента корреляции. Для проверки статистической гипотезы

$$H_0 : \rho_{XY} = 0$$

в качестве статистики критерия используется статистика

$$Z = \frac{\rho^*_{XY}}{\sqrt{1-(\rho^*_{XY})^2}}\sqrt{n-2}\,, \qquad (6.35)$$

которая при условии истинности H_0 имеет распределение Стьюдента с $n-2$ степенями свободы $f_Z(z \mid H_0) \sim T(n-2)$.

Если альтернативная гипотеза $H' : \rho_{XY} \neq 0$, то критическая область для статистики критерия выбирается двусторонней; если $H' : \rho_{XY} < 0$ или $H' : \rho_{XY} > 0$, то лево- или правосторонней соответственно.

Пример 6.5. Проводятся наблюдения числа посетителей развивающегося сайта (признак *X*) и его средневзвешенной позиции по основным запросам в поисковой системе (признак *Y*). В результате наблюдения получены следующие данные:

№ п/п	1	2	3	4	5	6	7	8
X	500	750	820	1550	2420	2230	1890	1630
Y	12,2	10,5	8,4	6,2	4,7	4,9	5,1	5,5

Рассчитать коэффициент корреляции между указанными признаками и ответить на вопрос: можно ли утверждать, что между числом посетителей сайта и его средневзвешенной позицией в поисковой системе есть значимая ($\alpha = 0{,}1$) линейная корреляционная связь?

Для предварительного анализа данных построим корреляционное поле имеющихся наблюдений (рис. 6.4). Уже по виду корреляционного поля иногда можно сделать вывод о наличии и характере связи между признаками *X* и *Y*. Так, из рисунка видно, что некоторая нелинейная зависимость (например, квадратичная) лучше описывает имеющиеся данные, чем линейная зависимость, измеряемая коэффициентом корреляции.

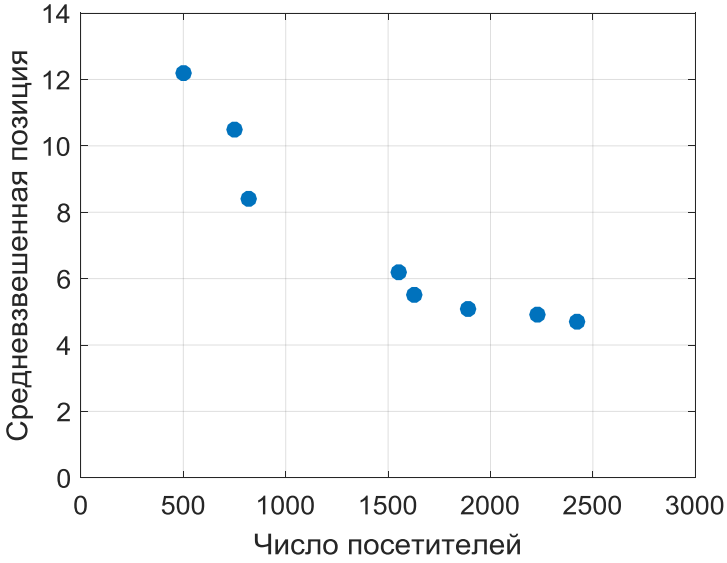

Рис. 6.4. Корреляционное поле

Рассчитаем точечную оценку коэффициента корреляции по формуле (6.31):

$$\overline{x} = \frac{1}{8}(500 + ... + 1630) \approx 1474; \qquad \overline{y} = \frac{1}{8}(12,2 + ... + 5,5) \approx 7;$$

$$\sigma_X^* = \sqrt{\frac{1}{8}\left((500 - 1474)^2 + ... + (1630 - 1474)^2\right)} \approx 668;$$

$$\sigma_Y^* = \sqrt{\frac{1}{8}\left((12,2 - 7)^2 + ... + (5,5 - 7)^2\right)} \approx 2,7;$$

$$k_{XY}^* = \frac{1}{8}\left((500 - 1474)(12,2 - 7) + ... + (1630 - 1474)(5,5 - 7)\right) \approx -1670;$$

$$\rho_{XY}^* = \frac{-1670}{668 \cdot 2,7} \approx -0,93.$$

Полученное значение говорит о сильной линейной корреляционной связи. Однако вследствие небольшого объёма выборки и

смещённости этой оценки, значение коэффициента корреляции ρ_{XY} генеральной совокупности может существенно отличаться от значения ρ^*_{XY}.

Найдём границы доверительного интервала для ρ_{XY}. Учитывая небольшой объём выборки, используем формулу (6.34):

$$\underline{\rho}_{XY} = \tanh\left(\frac{1}{2}\ln\frac{0,07}{1,93} + \frac{-0,93}{14} - \frac{1,64}{2,34}\right) \approx \tanh(-2,5) \approx -0,985;$$

$$\overline{\rho}_{XY} = \tanh\left(\frac{1}{2}\ln\frac{0,07}{1,93} + \frac{-0,93}{14} + \frac{1,64}{2,34}\right) \approx \tanh(-1,1) \approx -0,772,$$

где $u_{0,95} = 1,64$ – квантиль стандартизованного нормального распределения на уровне $1 - \alpha/2$.

Для проверки гипотезы о значимости коэффициента корреляции $H_0 : \rho_{XY} = 0$ против альтернативной гипотезы $H' : \rho_{XY} < 0$ рассчитаем выборочное значение статистики критерия по формуле (6.35):

$$z = \frac{-0,93 \cdot 2,45}{\sqrt{1 - 0,93^2}} \approx -6,2 \ ,$$

которому соответствует значение *p-value*

$$p = F_{T(6)}(-6,2) \approx 0,0004 .$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза должна быть отклонена, т.е. линейная корреляционная связь между рассматриваемыми признаками значима.

Контрольные вопросы и задачи

1. Какими свойствами точечных оценок обладает выборочное значение коэффициента корреляции? Является ли оценка несмещённой? Чему равно её смещение?
2. При каком значении коэффициента корреляции генеральной совокупности его выборочное значение имеет минимальное смещение?
3. Докажите состоятельность выборочного значения коэффициента корреляции.

4. Что называется стандартизованным выборочным значением коэффициента корреляции? Какое распределение имеет эта статистика при объёме выборки $n \to \infty$?

5. Выведите формулу расчёта одностороннего доверительного интервала для коэффициента корреляции, используя в качестве центральной статистики стандартизованный выборочный коэффициент корреляции.

6. Какая статистика критерия используется при проверке гипотезы о равенстве нулю коэффициента корреляции? Какой закон распределения имеет эта статистика при условии истинности основной гипотезы?

7. Сформулируйте метод доверительных интервалов для проверки гипотезы о равенстве нулю коэффициента корреляции.

§ 32. Оценивание коэффициента детерминации и корреляционного отношения по выборочным данным

Пусть (x_1, y_1), …, (x_n, y_n) – выборка наблюдений двумерного случайного вектора (X, Y), имеющего неизвестное распределение $F_{XY}(x, y)$.

1. Точечные оценки КД и КО. Для расчёта КД и КО необходимо знать функцию регрессии результативного признака Y на фактор X. Пусть эта функция имеет вид $f(x, \beta_0, ..., \beta_{k-1})$, где $\beta_0, ..., \beta_{k-1}$ – известные параметры. Тогда, учитывая определения дисперсии, обусловленной регрессией, и остаточной дисперсии (см. § 30), запишем выражения для их выборочных значений:

$$D^*_{\text{регр } Y|X} = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i, \beta_0, ..., \beta_{k-1}) - \bar{y}\right)^2, \qquad (6.36)$$

$$D^*_{\text{ост } Y} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i, \beta_0, ..., \beta_{k-1})\right)^2. \qquad (6.37)$$

Если же для функции регрессии задан только её вид, а параметры $\beta_0, ..., \beta_{k-1}$ оцениваются на основе результатов наблюдений (x_1, y_1), …, (x_n, y_n), то в формулах (6.36) и (6.37) вместо параметров $\beta_0, ..., \beta_{k-1}$ подставляются их точечные оценки $\tilde{\beta}_0, ..., \tilde{\beta}_{k-1}$.

Выборочная дисперсия признака Y не зависит от вида функции регрессии и рассчитывается по известной формуле

$$D_Y^* = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 .$$

Можно показать, что для выборочных оценок общей дисперсии Y, дисперсии, обусловленной регрессией Y на X, и остаточной дисперсии Y справедливо *правило сложения дисперсий*:

$$D_Y^* = D_{\text{ост}\,Y}^* + D_{\text{регр}\,Y|X}^* . \qquad (6.38)$$

При точечном оценивании коэффициента детерминации (6.28) вместо теоретических дисперсий используются соответствующие выборочные значения. Учитывая правило сложения дисперсий, запишем выражение для точечной оценки КД:

$$R_{Y|X}^{2\,*} = \frac{D_{\text{регр}\,Y|X}^*}{D_Y^*} = 1 - \frac{D_{\text{ост}\,Y}^*}{D_Y^*} . \qquad (6.39)$$

Эту оценку называют также *показателем* «*эр-квадрат*» (*R-squared*).

Точечной оценкой корреляционного отношения служит корень из показателя «эр-квадрат»:

$$R_{Y|X}^* = \sqrt{\frac{D_{\text{регр}\,Y|X}^*}{D_Y^*}} = \sqrt{1 - \frac{D_{\text{ост}\,Y}^*}{D_Y^*}} . \qquad (6.40)$$

При обработке реальных данных встречаются случаи, когда ни вид, ни параметры функции регрессии априорно не известны. Тогда функция регрессии может быть оценена непосредственно по результатам наблюдений. Для этого проводится группировка выборочных значений $x_1, \ldots, x_n$ (см. § 2). Обозначим: $J_1, \ldots, J_k$ – интервалы группировки; n_i – число выборочных точек, попавших в интервал J_i, $i = \overline{1,k}$; k – число интервалов.

Пусть $(x_{i1}, y_{i1}), ..., (x_{i,n_i}, y_{i,n_i})$ – выборочные наблюдения, попавшие в интервал J_i, $i = \overline{1,k}$. Для этих наблюдений рассчитываются групповые средние $(\bar{x}_i, \bar{y}_i)$:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}.$$

Линия, соединяющая все групповые средние $(\overline{x}_1, \overline{y}_1), ..., (\overline{x}_k, \overline{y}_k)$, и будет являться оценкой линии регрессии.

На практике для упрощения вычислений при расчёте оценки дисперсии, обусловленной регрессией Y на X, предполагается, что функция регрессии кусочно-постоянна (рис. 6.5):

$$\forall x \in J_i \rightarrow f(x) = \overline{y}_i, \qquad i = \overline{1,k}. \tag{6.41}$$

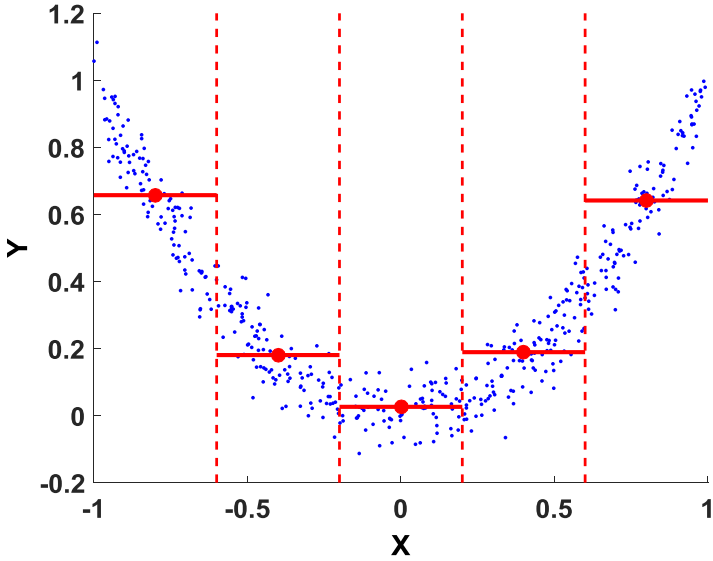

Рис. 6.5. Оценивание линии регрессии по выборочным данным. Крупными точками отмечены групповые средние $(\overline{x}_1, \overline{y}_1), ..., (\overline{x}_k, \overline{y}_k)$

Число интервалов группировки k не должно быть слишком мало – в этом случае кусочно-постоянная аппроксимация функции регрессии будет неточной. С другой стороны, при слишком большом числе интервалов группировки становятся неточными оценки групповых средних.

Учитывая (6.36), (6.37) и (6.41), запишем выражения для выборочных дисперсии, обусловленной регрессией Y на X, и остаточной дисперсии Y для случая кусочно-постоянной функции регрессии:

$$D^*_{\text{регр}\,Y|X} = \frac{1}{n}\sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2 . \qquad (6.42)$$

$$D^*_{\text{ост}\,Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2 . \qquad (6.43)$$

Эти выражения совпадают с выражениями (6.7) и (6.6) для расчёта межгрупповой и внутригрупповой дисперсий в дисперсионном анализе (см. § 27). Таким образом, дисперсия, обусловленная регрессией Y на X, является аналогом межгрупповой дисперсии и показывает меру вариации результативного признака, объяснённую факторным признаком, а остаточная дисперсия – аналог внутригрупповой дисперсии – меру вариации, связанной с действием остаточных факторов.

При расчётах дисперсии, обусловленной регрессией, остаточной дисперсии и общей дисперсии, а также КД и КО по результатам выборочного наблюдения необходимо иметь в виду, что все получаемые значения являются смещёнными оценками соответствующих теоретических значений, характеризующих генеральную совокупность. Показатели вариации, а также их несмещённые оценки сведены в таблицу, называемую *таблицей регрессионного анализа* (табл. 6.6).

Таблица 6.6

Таблица регрессионного анализа

Источник вариации	Показатель вариации	Число степеней свободы	Несмещённая оценка
Регрессия	$D^*_{\text{регр}\,Y\|X}$	$k-1$	$\frac{n}{k-1}D^*_{\text{регр}\,Y\|X}$
Остаточные признаки	$D^*_{\text{ост}\,Y}$	$n-k$	$\frac{n}{n-k}D^*_{\text{ост}\,Y}$
Все признаки	D^*_Y	$n-1$	$\frac{n}{n-1}D^*_Y$

Смещение точечной оценки КД, рассчитываемой по формуле (6.39), равно

$$\mathbf{M}\left[R^{2\,*}_{Y|X}\right] - R^2_{Y|X} = \frac{1-R^2_{Y|X}}{n}\left(k-(1-R^2_{Y|X})(1+2R^2_{Y|X})\right),$$

где k – число оцениваемых параметров функции регрессии ($k>1$). Если используется кусочно-постоянная аппроксимация функции регрессии, то это число равно числу интервалов группировки.

Это смещение всегда положительно, т.е. оценка КД (6.39) в среднем даёт завышенную долю дисперсии, объясненной регрессией. При больших k и малых n это смещение может достигать существенных значений и приводить к серьёзным ошибкам в интерпретации получаемых результатов. В частности, при $R^2_{Y|X}=0$ смещение оценки КД равно

$$\mathbf{M}\left[R^{2\,*}_{Y|X} \mid R^2_{Y|X}=0\right] = \frac{k-1}{n}.$$

Пренебрегая единицей в числителе, это смещение имеет смысл величины, обратной числу наблюдений, приходящихся на один оцениваемый параметр уравнения регрессии. Например, для выборки объёма $n=18$ из генеральной совокупности с КД, равным нулю, при числе оцениваемых параметров уравнения регрессии $k=6$ (таким образом, три наблюдения на параметр), оценка КД в среднем будет равна $5/18 \approx 0{,}278$. При $n/k>100$ смещение выборочного значения КД становится менее 0,01.

Оценкой КД, имеющей меньшее смещение, является отношение несмещённых оценок остаточной дисперсии и общей дисперсии признака Y за вычетом из единицы:

$$\bar{R}^2_{Y|X} = 1 - \frac{\tilde{D}_{\text{ост}\,Y}}{\tilde{D}_Y}, \qquad (6.44)$$

где

$$\tilde{D}_{\text{ост}\,Y} = \frac{n}{n-k} D^*_{\text{ост}\,Y},$$

$$\tilde{D}_Y = \frac{n}{n-1} D^*_Y.$$

Учитывая выражение (6.39) для расчёта показателя $R^{2\,*}_{Y|X}$, запишем:

$$\bar{R}^2_{Y|X} = 1 - \frac{D^*_{\text{ост}\,Y}\dfrac{1}{n-k}}{D^*_Y\dfrac{1}{n-1}} = 1 - \left(1 - R^{2\,*}_{Y|X}\right)\frac{n-1}{n-k}. \qquad (6.45)$$

Такая оценка называется *скорректированной оценкой коэффициента детерминации*, или *показателем* «*эр-бар-квадрат*» (*adjusted R-squared*). Эта оценка по-прежнему смещённая, поскольку отношение двух несмещённых оценок в общем случае не является несмещённой оценкой отношения.

Показатели «эр-квадрат» и «эр-бар-квадрат» имеют принципиально различную интерпретацию. Показатель $R^{2\,*}_{Y|X}$ является мерой вариации признака *Y*, объяснённой регрессией $f(x)$. Если вариация выборочных данных относительно линии регрессии отсутствует, т.е. все выборочные наблюдения лежат на линии регрессии, то $R^{2\,*}_{Y|X} = 1$. Если вариация самой линии регрессии отсутствует, т.е. $f(x) = \text{const}$, то $R^{2\,*}_{Y|X} = 0$.

Показатель $\bar{R}^2_{Y|X}$ всегда меньше показателя $R^{2\,*}_{Y|X}$ и может даже принимать отрицательные значения. Этот показатель можно рассматривать как сравнительную меру «объяснительных» способностей различных уравнений регрессии с поправкой на число параметров *k*.

При высоком отношении n/k объёма выборки к числу параметров уравнения регрессии разница между $R^{2\,*}_{Y|X}$ и $\bar{R}^2_{Y|X}$ становится практически пренебрежимой.

2. Интервальные оценки КД и КО. При расчёте границ доверительных интервалов для КД и КО используются различные аппроксимации. Если случайный вектор (*X*, *Y*) распределён по двумерному нормальному закону, то приближённый доверительный интервал для КД $R^2_{Y|X}$ на уровне значимости α может быть рассчитан по формуле

$$\left(R^{2\,*}_{Y|X} - t_{1-\alpha/2}(n-k)s\left[R^{2\,*}_{Y|X}\right]; R^{2\,*}_{Y|X} + t_{1-\alpha/2}(n-k)s\left[R^{2\,*}_{Y|X}\right]\right),$$

где $t_{1-\alpha/2}(n-k)$ – квантиль распределения Стьюдента с $n-k$ степенями свободы на уровне $1-\alpha/2$, а $s\left[R^{2\,*}_{Y|X}\right]$ – оценка среднеквадратичного отклонения показателя «эр-квадрат», рассчитываемая по формуле

$$s\left[R^{2\,*}_{Y|X}\right] = \sqrt{\frac{4R^{2\,*}_{Y|X}\left(1-R^{2\,*}_{Y|X}\right)^2\left(n-k\right)^2}{\left(n^2-1\right)\left(n+3\right)}},$$

которая при $n >> k$ аппроксимируется выражением

$$s\left[R^{2\,*}_{Y|X}\right] \approx \sqrt{\frac{4R^{2\,*}_{Y|X}\left(1-R^{2\,*}_{Y|X}\right)^2}{n}}.$$

Приближённый доверительный интервал для КО $R_{Y|X}$ на уровне значимости α имеет вид

$$\left(\sqrt{\frac{(n-k)R^{2\,*}_{Y|X}}{n\left(1-R^{2\,*}_{Y|X}\right)f_{1-\alpha/2}(r_1,r_2)} - \frac{k-1}{n}}; \sqrt{\frac{(n-k)R^{2\,*}_{Y|X}}{n\left(1-R^{2\,*}_{Y|X}\right)f_{\alpha/2}(r_1,r_2)} - \frac{k-1}{n}}\right),$$

где $f_{\alpha/2}(r_1,r_2)$ и $f_{1-\alpha/2}(r_1,r_2)$ – квантили распределения Фишера с r_1 и r_2 степенями свободы в числителе и знаменателе на уровнях $\alpha/2$ и $1-\alpha/2$ соответственно. Степени свободы вычисляются по формулам:

$$r_1 = \left[\frac{\left(k-1+nR^{2\,*}_{Y|X}\right)^2}{k-1+2nR^{2\,*}_{Y|X}}\right],$$

$$r_2 = n-k,$$

где $[\cdot]$ – целая часть числа.

На практике указанные аппроксимации применяются и для случая, когда распределение наблюдаемого случайного вектора (X, Y) отличается от нормального, причём, чем больше отношение n/k, тем выше точность аппроксимации.

3. Проверка значимости КД и КО. Для проверки статистической гипотезы

$$H_0 : R^2_{Y|X} = 0 \text{ (или } H_0 : R_{Y|X} = 0)$$

в качестве статистики критерия используется статистика

$$Z = \frac{R^{2\,*}_{Y|X} / (k-1)}{\left(1 - R^{2\,*}_{Y|X}\right) / (n-k)}, \qquad (6.46)$$

которая при условии истинности H_0 имеет распределение Фишера с $k-1$ и $n-k$ степенями свободы в числителе и знаменателе соответственно: $f_Z(z \mid H_0) \sim F(k-1, n-k)$.

Критическая область для статистики критерия выбирается правосторонней.

Пример 6.6. Сравниваются баллы, полученные в результате тестирования школьников по математике (по 10-балльной системе), и годовые оценки по алгебре за прошлый год. В результате статистического наблюдения получены следующие данные:

№ п/п	1	2	3	4	5	6	7	8
Оценки	4	5	4	3	3	5	4	2
Баллы	7	9	8	7	5	9	10	6

№ п/п	9	10	11	12	13	14	15	16
Оценки	2	4	5	3	3	5	4	3
Баллы	5	9	8	7	5	8	9	6

Определить степень влияния годовых оценок школьников по алгебре на баллы, полученные ими при тестировании. Можно ли утверждать, что между этими признаками имеется значимая ($\alpha = 0{,}1$) линейная корреляционная связь?

Обозначим: X – годовые оценки по алгебре; Y – баллы, полученные в результате тестирования. Объём выборки $n = 16$. На рис. 6.6 представлена диаграмма рассеяния полученных наблюдений.

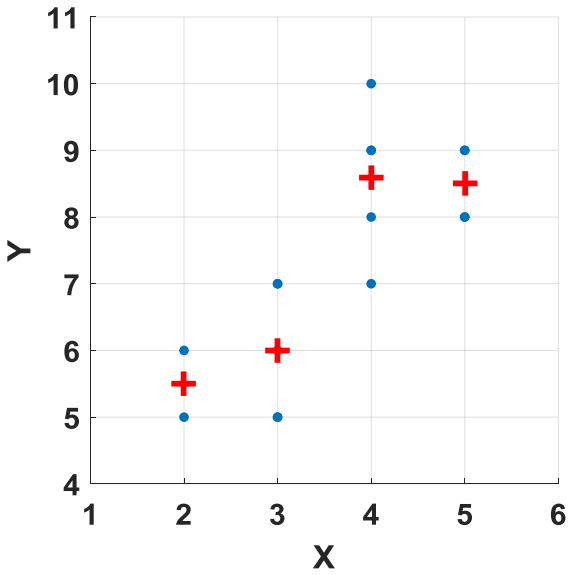

Рис. 6.6. Диаграмма рассеяния. Знаками «+» отмечены средние значения $\overline{y}_i$ признака Y при каждом фиксированном значении x_i признака X, $i = \overline{1,k}$

Рассчитаем оценки корреляционного отношения Y на X и коэффициента корреляции между X и Y. В связи с тем, что уравнение регрессии изначально неизвестно, оценим его на основе выборочных значений. Для этого сгруппируем результаты наблюдений по оценкам. Число групп k выберем равным числу вариантов оценок, т.е. $k = 4$:

Оценки	2	3	4	5
Баллы	6; 5	7; 5; 7; 5; 6	7; 8; 10; 9; 9	9; 9; 8; 8
n_i	2	5	5	4
$\overline{y}_i$	5,5	6	8,6	8,5

Из рисунка видно, что зависимость между групповыми средними скорее нелинейная, однако она может быть аппроксимирована прямой с высокой степенью точности. Это говорит о том, что корреляционное отношение Y на X и коэффициент корреляции между X и Y должны быть достаточно близки.

Рассчитаем общий средний балл:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i = \frac{1}{16}\left(2 \cdot 5{,}5 + \ldots + 4 \cdot 8{,}5\right) \approx 7{,}38$$

и оценки дисперсии, обусловленной регрессией Y на X, и общей дисперсии признака Y:

$$D^*_{\text{регр}\,Y|X} = \frac{1}{16}\left(2 \cdot (5{,}5 - 7{,}38)^2 + \ldots + 4 \cdot (8{,}5 - 7{,}38)^2\right) \approx 1{,}82;$$

$$D^*_Y = \frac{1}{16}\left((7 - 7{,}38)^2 + \ldots + (6 - 7{,}38)^2\right) \approx 2{,}65.$$

Показатель «эр-квадрат» и корреляционное отношение:

$$R^{2\,*}_{Y|X} = \frac{1{,}82}{2{,}65} \approx 0{,}69;$$

$$R^*_{Y|X} = \sqrt{0{,}69} \approx 0{,}83.$$

Показатель «эр-бар-квадрат»:

$$\bar{R}^2_{Y|X} = 1 - (1 - 0{,}69)\frac{15}{12} \approx 0{,}61.$$

По шкале Чеддока (см. § 27) для корреляционного отношения определяем, что имеется высокое влияние годовых оценок школьников по алгебре на баллы, полученные ими при тестировании.

Рассчитаем оценку коэффициента корреляции:

$$\rho^*_{XY} \approx 0{,}76.$$

Полученное значение $\rho^*_{XY} \approx R^*_{Y|X}$, что говорит о возможности аппроксимации оценённой по выборке линии регрессии прямой с высокой степенью точности.

Рассчитаем доверительный интервал для $R^2_{Y|X}$:

$$s^2\left[R^{2\,*}_{Y|X}\right] = \frac{4 \cdot 0{,}69 \cdot 0{,}31^2 \cdot 12^2}{255 \cdot 19} \approx 0{,}0079;$$

$$\left(0,69-1,8\sqrt{0,0079};0,69+1,8\sqrt{0,0079}\right)\approx(0,53;0,85),$$

где $t_{0,95}(12)=1,8$ – квантиль распределения Стьюдента с $n-k$ степенями свободы на уровне 1 – $\alpha/2$.

Для проверки гипотезы $H_0: R_{Y|X}=0$ о незначимости КО против альтернативной гипотезы $H_0: R_{Y|X}>0$ рассчитаем выборочное значение статистики критерия по формуле (6.46):

$$z=\frac{0,69/3}{0,31/12}\approx 8,9,$$

которому соответствует значение *p-value*

$$p=1-F_{F(3,12)}(8,9)\approx 0,002.$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза должна быть отклонена, т.е. корреляционная связь между рассматриваемыми признаками значима.

Контрольные вопросы и задачи

1. Сколько степеней свободы имеет остаточная дисперсия $D^*_{\text{ост}\,Y}$, если функция регрессии *Y* на *X* априорно задана? Если задан вид функции регрессии, а её параметры оцениваются по результатам выборочного наблюдения?
2. Объясните принцип оценивания уравнения регрессии по выборочным данным, если априорная информация о её виде неизвестна.
3. Как выбирается число интервалов группировки при оценивании уравнения регрессии по выборочным данным?
4. Какими свойствами точечных оценок обладает выборочное значение коэффициента детерминации? Является ли оценка несмещённой? Чему равно её смещение?
5. При каком значении КД генеральной совокупности смещение выборочного значения КД максимально?
6. В каком случае смещением выборочного значения КД можно пренебречь?

7. Объясните различие между показателями «эр-квадрат» и «эр-бар-квадрат». Является ли скорректированная оценка КД несмещённой?

8. Покажите, что отношение несмещённых оценок параметров θ_1 и θ_2 в общем случае не является несмещённой оценкой отношения θ_1 / θ_2.

9. По каким формулам рассчитываются несмещённые оценки общей дисперсии *Y*, дисперсии, обусловленной регрессией *Y* на *X*, и остаточной дисперсии *Y*? Поясните все использованные обозначения.

10. Докажите правило сложения дисперсий для выборочных дисперсий D_Y^*, $D_{\text{регр}\,Y|X}^*$ и $D_{\text{ост}\,Y}^*$.

11. Справедливо ли правило сложения дисперсий для несмещённых оценок $\tilde{D}_Y$, $\tilde{D}_{\text{регр}\,Y|X}$ и $\tilde{D}_{\text{ост}\,Y}$?

12. Как связаны между собой значения показателей «эр-квадрат» и «эр-бар-квадрат»? В каком случае эти значения совпадают?

13. Предложите интерпретацию значений показателей «эр-квадрат» и «эр-бар-квадрат».

14. Что можно сказать о выборочных данных, если $R_{Y|X}^{2\,*} = 1$? $R_{Y|X}^{2\,*} = 0$?

15. Какая статистика критерия используется при проверке гипотезы о равенстве нулю КД (КО)? Какой закон распределения имеет эта статистика при условии истинности основной гипотезы?

16. Объясните принцип выбора типа критической области при проверке гипотезы о равенстве нулю КД (КО).

§ 33. Ранговый коэффициент корреляции по Спирмену

Пусть $x_1, \ldots, x_n$ – выборка наблюдений случайной величины *X*, имеющей распределение $F_X(x)$, $x_{(1)}, ..., x_{(n)}$ – её вариационный ряд.

Рангом r_i элемента x_i выборки $x_1, \ldots, x_n$ называется его порядковый номер в вариационном ряду выборки, т.е.

$$x_{(r_i)} = x_i, \quad i = \overline{1,n}.$$

Ранг r_i элемента x_i можно рассматривать как реализацию случайной величины $R_i = R_i(X_1, ..., X_n)$, $i = \overline{1,n}$, определяемой как ранг случайной величины X_i в случайной выборке X_1, …, X_n.

Ранговой статистикой Z называется произвольная функция от рангов R_1, …, R_n:

$$Z = \varphi(R_1, ..., R_n).$$

В связи с тем, что статистика Z – функция случайных аргументов, то она сама является случайной величиной. Для каждой реализации x_1, …, x_n случайной выборки X_1, …, X_n получим соответствующие ей реализацию рангов r_1, …, r_n и реализацию z ранговой статистики Z:

$$z = \varphi(r_1, ..., r_n).$$

Если выборка x_1, …, x_n содержит одинаковые элементы, то им, как правило, приписывают одинаковый ранг, равный среднему из порядковых номеров этих элементов в вариационном ряду.

Пример 6.7. Рассчитать ранги элементов выборки:

x_i	1,2	1,6	2,2	0,8	1,2	1,4	1,6	1,6

Построим вариационный ряд и соответствующую выборку рангов:

i	1	2	3	4	5	6	7	8
$x_{(i)}$	0,8	1,2	1,2	1,4	1,6	1,6	1,6	2,2
r_i	1	2,5	2,5	4	6	6	6	8

В связи с тем, что значение 1,2 встречается в вариационном ряду 2 раза (под номерами 2 и 3), ранги соответствующих элементов выборки полагаются равными $\frac{2+3}{2} = 2,5$. Аналогично рассчитываются ранги элементов, равных 1,6.

Пусть (x_1, y_1), …, (x_n, y_n) – выборка наблюдений двумерного случайного вектора (X, Y); r_1, …, r_n – ранги элементов выборки x_1,…, x_n; s_1, …, s_n – ранги элементов выборки y_1, …, y_n.

Ранговым коэффициентом корреляции по Спирмену (*Charles Spearman*, 1904) называется ранговая статистика, определяемая следующим выражением:

$$\rho_{sXY} = \frac{\sum_{i=1}^{n}(R_i - \bar{r})(S_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{r})^2 \sum_{i=1}^{n}(S_i - \bar{s})^2}}, \qquad (6.47)$$

где $\bar{r}$ и $\bar{s}$ – средние значения рангов:

$$\bar{r} = \bar{s} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}. \qquad (6.48)$$

Выборочное значение этой статистики для выборки (x_1, y_1), …, (x_n, y_n) равно:

$$\rho^*_{sXY} = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2 \sum_{i=1}^{n}(s_i - \bar{s})^2}} = \frac{\mu^*_{RS}}{\sigma^*_S \sigma^*_R}, \qquad (6.49)$$

где σ^*_S и σ^*_R – среднеквадратичные отклонения выборок рангов r_1, …, r_n и s_1, …, s_n соответственно; μ^*_{RS} – их ковариационный момент.

Фактически значение рангового коэффициента корреляции по Спирмену для выборки (x_1, y_1), …, (x_n, y_n) равно линейному коэффициенту корреляции для соответствующей выборки рангов (r_1, s_1), …, (r_n, s_n).

Учитывая (6.48), выражение (6.49) можно упростить:

$$\rho^*_{sXY} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_i - s_i)^2 . \qquad (6.50)$$

Известно, что линейный коэффициент корреляции ρ_{XY} используется для обнаружения линейной корреляционной связи между величинами X и Y (см. § 29). Так, если $|\rho_{XY}| = 1$, то между X и Y име-

ется линейная функциональная связь. Если $\rho_{XY} = 0$, то между X и Y отсутствует линейная корреляционная связь.

Ситуация $|\rho_{sXY}| = 1$ будет означать, что между рангами случайных величин X и Y имеется линейная функциональная связь. Если же $\rho_{sXY} = 0$, то между рангами отсутствует линейная корреляционная связь.

Рассмотрим, что означают эти случаи в пространстве признаков X и Y. Если X и Y связаны линейной функциональной зависимостью $Y = aX + b$, то между рангами также будет линейная зависимость. В самом деле, при $a > 0$ бóльшим значениям X будут соответствовать бóльшие значения Y, таким образом, для отсортированной в порядке возрастания по X выборки (x_1, y_1), …, (x_n, y_n) соответствующая выборка рангов будет иметь вид:

r_i	1	…	i	…	n
s_i	1	…	i	…	n

При $a < 0$:

r_i	1	…	i	…	n
s_i	n	…	$n-i+1$	…	1

Рассчитывая ранговый коэффициент по формуле (6.50), получим, что при $a > 0$: $\rho^*_{sXY} = 1$, при $a < 0$: $\rho^*_{sXY} = -1$.

Если $Y = \varphi(X)$, где $\varphi(X)$ – монотонно возрастающая функция, то для отсортированной по X выборки (x_1, y_1), …, (x_n, y_n) соответствующая выборка рангов будет такой же, что и для случая линейной функциональной зависимости между X и Y при $a > 0$. Если $\varphi(X)$ – монотонно убывающая функция, то – такой же, что для случая $a < 0$ (рис. 6.7).

Из рисунка видно, что переход к рангам «выпрямляет» монотонную нелинейную зависимость исходных признаков.

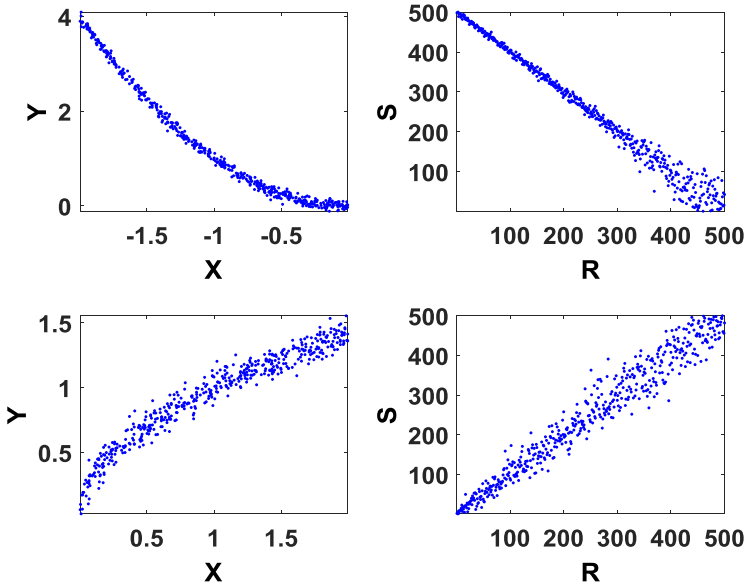

Рис. 6.7. Диаграммы рассеяния выборочных наблюдений (слева) и соответствующих выборочных рангов (справа)

Если признаки X и Y независимы, случайный вектор рангов $(S_1, ..., S_n)$, составленный для случайной выборки $Y_1, ..., Y_n$, соответствующей отсортированным по возрастанию значениям выборки $x_1, ..., x_n$, с равной вероятностью является любой из возможных $n!$ перестановок, составленных из чисел 1, …, n. Следовательно, математическое ожидание рангового коэффициента корреляции по Спирмену (6.47) будет равно нулю, т.е. $\mathbf{M}\left[\rho_{sXY}\right] = 0$. Можно показать, что дисперсия $\mathbf{D}\left[\rho_{sXY}\right] = \dfrac{1}{n-1}$. Это означает, что значения выборочного рангового коэффициента корреляции по Спирмену ρ^{*}_{sXY} при условии независимости случайных величин X и Y и большом объёме выборки будут группироваться вблизи нуля.

Из рис. 6.8 видно, что для независимых случайных величин X и Y выборочные ранги рассеяны практически равномерно внутри квадрата $n \times n$.

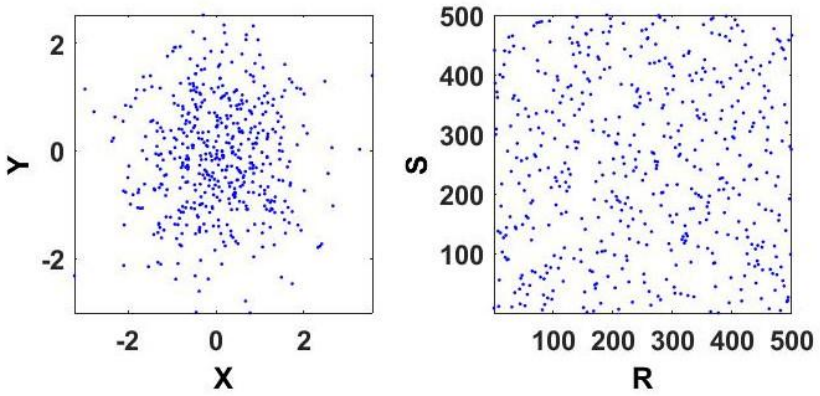

Рис. 6.8. Диаграммы рассеяния выборочных наблюдений (слева) и соответствующих выборочных рангов (справа) для случая независимых признаков

Для проверки значимости рангового коэффициента корреляции по Спирмену сформулируем основную гипотезу:

$$H_0 : \rho_{sXY} = 0 .$$

В качестве статистики критерия используется статистика

$$Z = \frac{\rho^*_{sXY}}{\sqrt{1 - \left(\rho^*_{sXY}\right)^2}} \sqrt{n-2}, \qquad (6.51)$$

которая при условии истинности H_0 имеет распределение Стьюдента с $n-2$ степенями свободы: $f_Z(z \mid H_0) \sim T(n-2)$.

Критическая область для статистики критерия выбирается, исходя из вида альтернативной гипотезы.

Пример 6.8. Десяти школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Результаты измерений представлены в таблице:

Номер школьника	Среднее время решения тестов на наглядно-образное мышление X, с	Среднее время решения тестов на вербальное мышление Y, с
1	19	15
2	12	7
3	32	17
4	17	14
5	14	8
6	25	15
7	15	8
8	35	17
9	29	16
10	27	16

Психолога интересует вопрос: существует ли взаимосвязь между временем решения этих задач?

Обозначим: X – среднее время решения тестов на наглядно-образное мышление; Y – среднее время решения тестов на вербальное мышление. Объём выборки $n = 10$.

Рассчитаем выборочное значение коэффициента корреляции между признаками X и Y:

$$\rho^*_{XY} \approx 0,87 .$$

Полученное значение говорит о довольно сильной линейной корреляционной связи между X и Y.

Проверим связь между X и Y на монотонность. Для этого составим выборку рангов и рассчитаем ранговый коэффициент корреляции по Спирмену по формуле (6.50):

i	1	2	3	4	5	6	7	8	9	10
r_i	5	1	9	4	2	6	3	10	8	7
s_i	5,5	1	9,5	4	2,5	5,5	2,5	9,5	7,5	7,5

$$\rho^*_{sXY} = 1 - \frac{6}{10 \cdot 99}\left((5-5,5)^2 + ... + (7-7,5)^2\right) \approx 0,99 .$$

Полученное значение практически равно 1, следовательно, между признаками X и Y имеется практически функциональная монотонно возрастающая зависимость. Близость полученного значения к значению коэффициента линейной корреляции означает, что эта монотонная зависимость может быть аппроксимирована линейной зависимостью с высокой степенью точности.

Таким образом, между средним временем решения заданий теста на наглядно-образное и вербальное мышление имеется практически функциональная монотонно возрастающая связь.

Контрольные вопросы и задачи

1. Что называется рангом элемента выборки? Как рассчитываются ранги одинаковых элементов?
2. Что называется ранговой статистикой?
3. Что называется ранговым коэффициентом корреляции по Спирмену? Какие значения он может принимать?
4. Докажите формулу (6.50) расчёта рангового коэффициента корреляции по Спирмену.
5. Чему равен ранговый коэффициент корреляции по Спирмену для линейно связанных случайных величин? Для независимых случайных величин?
6. Предложите интерпретацию значения рангового коэффициента корреляции по Спирмену. В чём отличие от интерпретации линейного коэффициента корреляции?
7. Следует ли из равенства нулю рангового коэффициента корреляции по Спирмену равенство нулю линейного коэффициента корреляции? Верно ли обратное утверждение?
8. Следует ли из равенства единице рангового коэффициента корреляции по Спирмену равенство единице линейного коэффициента корреляции? Верно ли обратное утверждение?
9. Что можно сказать о выборочных данных, если:

а) $\rho^*_{XY} < \rho^*_{sXY}$;

б) $\rho^*_{XY} > \rho^*_{sXY}$;

в) $\rho^*_{XY} = \rho^*_{sXY}$?

§ 34. Ранговый коэффициент корреляции по Кендаллу

Пусть (x_1, y_1), …, (x_n, y_n) – выборка наблюдений двумерного случайного вектора (X, Y).

Ранговым коэффициентом корреляции по Кендаллу (*Maurice Kendall*, 1938) называется ранговая статистика, определяемая следующим выражением:

$$\tau_{XY} = \frac{N^+ - N^-}{\frac{1}{2}n(n-1)}, \qquad (6.52)$$

где N^+ – число пар наблюдений (x_i, y_i), (x_j, y_j), $i > j$, для которых выполнено условие $(x_i - x_j)(y_i - y_j) > 0$; N^- – число пар наблюдений (x_i, y_i), (x_j, y_j), $i > j$, для которых выполнено условие $(x_i - x_j)(y_i - y_j) < 0$. Иными словами, N^+ – число наблюдаемых пар, у которых имеется одинаковая тенденция к изменению по обоим признакам: либо при увеличении значения одного увеличивается значение другого, либо при уменьшении значения одного уменьшается значение другого; N^- – число наблюдаемых пар с противоположными тенденциями к изменению. Ранговый коэффициент корреляции по Кендаллу также называется «*тау Кендалла*» (*Kendall's tau coefficient*).

Отсортируем результаты наблюдений в порядке возрастания значений признака X. Тогда выборкой рангов признака X будет последовательность натуральных чисел 1, 2, …, n (если все наблюдения x_1, …, x_n различны). Соответствующую выборку рангов признака Y обозначим через s_1, …, s_n.

На практике для расчёта выборочного значения рангового коэффициента корреляции по Кендаллу используют формулу:

$$\tau^*_{XY} = \frac{4Q}{n(n-1)} - 1, \qquad (6.53)$$

где $Q = \sum_{i=1}^{n-1} Q_i$; $Q_i = \sum_{j=i+1}^{n} \left[s_j > s_i \right]$ – количество рангов в выборке $s_{i+1}, ..., s_n$ больших, чем s_i, $i = \overline{1, n-1}$.

Использование формулы (6.53) даёт верный результат лишь для случая, когда в выборках $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ отсутствуют повторяющиеся элементы. Однако при небольшом их количестве погрешностью расчёта на практике можно пренебречь.

Свойства и интерпретация рангового коэффициента корреляции по Кендаллу аналогичны свойствам и интерпретации рангового коэффициента корреляции по Спирмену. Так, при функциональной монотонно возрастающей зависимости между случайными величинами X и Y значение $\tau_{XY} = 1$, при монотонно убывающей $\tau_{XY} = -1$. Для независимых случайных величин X и Y математическое ожидание $\mathbf{M}\left[\tau_{XY}\right] = 0$.

Выборочные значения коэффициента корреляции по Спирмену, как правило, получаются выше (по абсолютной величине) выборочных значений коэффициента корреляции по Кендаллу. Этот эффект связан с большей чувствительностью первого коэффициента к несоответствию в тенденциях изменений значений признаков.

Для проверки значимости рангового коэффициента корреляции по Кендаллу сформулируем основную гипотезу:

$$H_0 : \tau_{XY} = 0 .$$

В качестве статистики критерия используется статистика

$$Z = \tau^*_{XY} \sqrt{\frac{9n(n-1)}{2(2n+5)}}, \qquad (6.54)$$

которая при условии истинности H_0 и большом объёме выборки ($n > 30$) аппроксимируется стандартизованным нормальным распределением: $f_Z(z \mid H_0) \sim N(0,1)$.

Критическая область для статистики критерия выбирается, исходя из вида альтернативной гипотезы.

Пример 6.9. В условиях примера 6.8 рассчитать ранговый коэффициент корреляции по Кендаллу и сравнить полученное значение с ранговым коэффициентом корреляции по Спирмену.

Для расчёта рангового коэффициента корреляции по Кендаллу переупорядочим выборку рангов в порядке возрастания рангов r_i, $i = \overline{1,10}$:

r_i	1	2	3	4	5	6	7	8	9	10
s_i	1	2,5	2,5	4	5,5	5,5	7,5	7,5	9,5	9,5

Рассчитаем показатель Q:

$$Q = 9 + 7 + 7 + 6 + 4 + 4 + 2 + 2 = 41$$

и выборочное значение «тау Кендалла»:

$$\tau^*_{XY} = \frac{4 \cdot 41}{10 \cdot 9} - 1 \approx 0,82 .$$

Полученное значение оказалось намного меньше значения рангового коэффициента корреляции по Спирмену ($\rho^*_{sXY} \approx 0,99$), что связано с наличием в выборке рангов $s_1, ..., s_n$ повторяющихся значений при её небольшом объёме.

Контрольные вопросы и задачи

1. Предложите интерпретацию значения рангового коэффициента корреляции по Кендаллу. В чём отличие от интерпретации рангового коэффициента корреляции по Спирмену?
2. Чему равен ранговый коэффициент корреляции по Кендаллу для линейно связанных случайных величин? Для независимых случайных величин?
3. Следует ли из равенства нулю рангового коэффициента корреляции по Спирмену равенство нулю рангового коэффициента корреляции по Кендаллу? Верно ли обратное утверждение?
4. Следует ли из равенства единице рангового коэффициента корреляции по Спирмену равенство единице рангового коэффициента корреляции по Кендаллу? Верно ли обратное утверждение?
5. Какая статистика критерия используется при проверке гипотезы о равенстве нулю рангового коэффициента корреляции по Кендаллу? Какой закон распределения имеет эта статистика при условии истинности основной гипотезы и большом объёме выборки?

Глава 7. РЕГРЕССИОННЫЙ АНАЛИЗ

§ 35. Статистические модели

Для применения математических методов описания явлений необходимо, прежде всего, установить соотношения между величинами, характеризующими рассматриваемые явления. Каждое такое соотношение представляет собой математическую модель явления. Так, законы Ньютона представляют совокупность моделей механических явлений, уравнения Максвелла – математическую модель электродинамических явлений, волновое уравнение и уравнение теплопроводности – модели колебательных процессов в сплошных средах и распространения тепла в заданной области пространства соответственно.

Пусть поведение моделируемой системы описывается некоторой совокупностью величин, причём одни величины носят характер внешних воздействий на систему и называются её *входными воздействиями* (*input variables*), а другие представляют собой результат работы системы и называются *откликами системы* (*responses*) на входные воздействия (рис. 7.1).

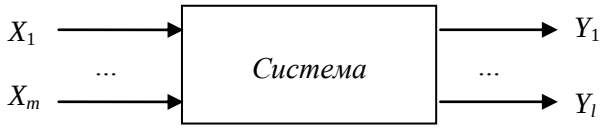

Рис. 7.1. Входные воздействия и отклики системы

В ряде случаев математическую модель системы можно построить чисто теоретическим путём на основе известных законов механики, физики и других дисциплин, использующих количественные соотношения (такую модель будем называть *теоретической*). Например, различные модели управляемого летательного

аппарата можно построить математически, пользуясь законами аэродинамики и связанных с ней разделов механики.

Однако существуют и такие системы, для которых принципиально невозможно построить адекватные модели чисто теоретическим путём. Причиной этому может быть как отсутствие точных сведений о структуре или параметрах системы, так и её высокая сложность. Примерами таких систем могут служить общество, завод, отрасль промышленности, экономика и т.п. В этом случае приходится прибегать к экспериментальному исследованию самих систем или входящих в них подсистем и строить соответствующие модели, используя собранные статистические данные.

Модель, построенная на основе статистической обработки результатов экспериментального исследования функционирования системы, называется *статистической моделью* системы.

К числу задач, при решении которых используются статистические модели, относятся прогнозирование погоды по измеренным значениями параметров состояния атмосферы в различных точках пространства и в различные моменты времени, в медицинской практике задача диагностики болезни пациента по результатам обследования и назначение соответствующих методов лечения, распознавание рукописных символов и цифр на изображении и многие другие задачи, для решения которых применение классических математических методов оказывается практически невозможным или неэффективным.

Задача построения статистической модели явления, процесса или системы состоит в нахождении соотношений между величинами, описывающими течение данного явления, процесса или функционирование системы. Если эти соотношения позволяют по данным значениям входных величин однозначно определить значения выходных, то описываемая ими модель называется *детерминированной*. Если же выходы модели являются случайными величинами, то модель называется *стохастической*.

Как теоретические модели, выводимые математически из законов физики, химии, экономики или других областей науки, так и статистические модели, получаемые на основе статистической обработки результатов наблюдений, могут быть детерминированными или стохастическими.

Одному и тому же явлению могут соответствовать различные модели. Проблема построения статистической модели включает выбор подходящего вида модели, обладающей разумной степенью сложности, и определение её параметров.

Контрольные вопросы и задачи

1. Что называется статистической моделью системы?
2. В чем отличие статистической модели от теоретической?
3. Приведите примеры статистических и теоретических моделей.
4. Какие модели называются стохастическими? В чем отличие стохастической модели от детерминированной?
5. Приведите примеры детерминированных и стохастических моделей.
6. С чем связана необходимость построения статистических моделей? В чем недостатки и преимущества статистических моделей перед теоретическими?

§ 36. Задачи регрессионного анализа

Поставим задачу определения значения случайной величины по данным значениям другой величины. Эту задачу можно рассматривать как задачу построения детерминированной статистической модели стохастической системы. Причиной стохастичности моделируемой системы может являться как присущая ей внутренняя случайность, так и отсутствие информации о влияющих на функционирование системы внешних факторах. Например, при моделировании дохода предприятия как функции объёма произведённой продукции и цены, моделируемая величина будет обладать стохастическим поведением, поскольку в модели не учтён фактор внереализационных доходов. Рыночную цену на продукцию (например, металл или нефть), в свою очередь, также можно рассматривать как случайную величину, поскольку при её моделировании практически невозможно учесть все факторы, оказывающие влияние на её значение.

Пусть Y – случайная величина, значение которой требуется определить; x – известная величина, представляющая собой значение

некоторой случайной величины X или заданное значение некоторой переменной. Предположим, что между величинами Y и X имеется статистическая связь, т.е. распределение случайной величины Y зависит от значения x (см. § 25).

С точки зрения математической статистики рассматриваемая задача представляет собой задачу нахождения оценки $\hat{y}(x)$ значения случайной величины Y при данном значении x. В связи с тем, что x – фиксированное значение, то оценка $\hat{y}(x)$ не является случайной величиной. Случайная величина – ошибка этой оценки:

$$\varepsilon(x) = Y \mid x - \hat{y}(x). \tag{7.1}$$

В качестве меры точности оценки $\hat{y}(x)$ будем использовать математическое ожидание квадрата ошибки $\varepsilon(x)$:

$$\mathbf{M}\left[\varepsilon^2(x)\right] = \mathbf{M}\left[\left(Y - \hat{y}(x)\right)^2 \mid x\right]. \tag{7.2}$$

Наилучшей оценкой значения случайной величины Y при данном значении x будет оценка, минимизирующая ошибку (7.2):

$$\mathbf{M}\left[\left(Y - \hat{y}(x)\right)^2 \mid x\right] \to \min_{\hat{y}(x)}. \tag{7.3}$$

Из известного в теории вероятностей равенства

$$\mathbf{M}\left[\left(Y - a\right)^2\right] = \mathbf{D}[Y] + \left(\mathbf{M}[Y] - a\right)^2 \tag{7.4}$$

следует, что математическое ожидание квадрата ошибки (7.1) будет минимальным, если $\hat{y}(x)$ будет математическим ожиданием случайной величины Y при данном значении x:

$$\hat{y}(x) = \mathbf{M}\left[Y \mid x\right]. \tag{7.5}$$

Следовательно, зависимость оценки $\hat{y}(x)$ значения случайной величины Y при данном значении x представляет собой регрессию Y на X. Таким образом, *оптимальной с точки зрения среднего квадрата ошибки* (7.1) *оценкой зависимости Y от x служит регрессия Y на X*. В частности, *оптимальным прогнозом величины Y по данному значению x будет прогноз по регрессии*.

Статистическая модель, определяемая регрессией Y на X, называется *регрессионной*. Построение и исследование регрессионных моделей составляет предмет *регрессионного анализа*.

Регрессионную модель имеет смысл строить, если априорно или по результатам предварительного анализа выявлено, что между входными и выходными величинами имеется статистическая связь. В терминах регрессионного анализа входные переменные называются *регрессорами*, или *предикторами*, а моделируемые величины – *откликами* модели.

Ниже перечислены основные задачи регрессионного анализа.

1. Выбор класса функций для описания зависимости откликов модели $Y_1, ..., Y_l$ от регрессоров $X_1, ..., X_m$.

2. Нахождение оценок неизвестных параметров функции из выбранного класса.

3. Статистический анализ найденной зависимости откликов от регрессоров модели.

4. Предсказание значений откликов модели по результатам наблюдения регрессоров на основе найденной зависимости.

Рассмотрим случай одного регрессора X и скалярного отклика Y. Как показано выше, оптимальной функцией, описывающей зависимость отклика модели Y от регрессора X, является функция регрессии Y на X. При этом возможны следующие ситуации (рис. 7.2).

I. Вид функции регрессии известен, исходя из априорной информации о наблюдаемых величинах. Например, если известно, что случайные величины X и Y имеют нормальный закон распределения, то уравнение регрессии Y на X (как и X на Y) может быть только линейным (см. § 29).

II. Вид функции регрессии не известен или эта функция слишком сложна. В этой ситуации возможны следующие подходы к определению вида функции регрессии.

1. Исследователь задаёт некоторый ограниченный класс функций Ψ (например, линейные или полиномиальные функции), в котором предлагается искать функцию регрессии. Если этот класс функций не содержит «истинную» функцию регрессии, то минимум среднего квадрата ошибки $\varepsilon(x)$ при каждом значении x не может быть обеспечен. Для выбора класса функций, в котором целесообразно искать функцию регрессии, нередко требуется проведение дополнительного анализа исходных данных.

2. Функция регрессии оценивается по результатам наблюдений. Такое оценивание основано на расчёте множества условных средних значений наблюдений отклика Y и аппроксимации линии регрессии по рассчитанным точкам (см. § 32).

Рис. 7.2. Схема оценивания функции регрессии

Контрольные вопросы и задачи

1. В чём состоит задача оценивания значения случайной величины Y при данном значении x?
2. Какой критерий точности используется для оценки значения случайной величины Y при данном значении x?
3. При каком значении оценки случайной величины Y при данном значении x критерий точности модели принимает наименьшее значение?

4. Докажите равенство (7.4).

5. Что называется регрессионной моделью? Приведите примеры регрессионных моделей.

6. Что называется регрессорами модели? Что называется откликами модели?

7. Перечислите основные задачи регрессионного анализа.

8. Перечислите подходы к построению регрессионных моделей, если вид функции регрессии априорно не известен.

§ 37. Оценивание параметров уравнения регрессии. Метод наименьших квадратов

Как только определён класс функций Ψ, в котором предполагается искать функцию регрессии, возникает задача оценивания её параметров.

Рассмотрим сначала случай одного регрессора X и скалярного отклика Y. Пусть $f(x, \beta_0, ..., \beta_{k-1}) \in \Psi$ – предполагаемая скалярная функция регрессии, $\beta = (\beta_0, ..., \beta_{k-1})^T$ – вектор неизвестных параметров. Из (7.1) и (7.5) следует, что случайная величина Y при фиксированном значении x складывается из двух слагаемых: неслучайной величины $f(x, \beta_0, ..., \beta_{k-1})$ и случайной ошибки $\varepsilon(x)$:

$$Y \mid x = f(x, \beta_0, ..., \beta_{k-1}) + \varepsilon(x). \qquad (7.6)$$

Модель (7.6) представляет собой регрессионную модель отклика Y, а случайная величина $\varepsilon(x)$ называется *ошибкой регрессионной модели*.

При оценивании и проведении статистического анализа регрессионной модели (7.6), как правило, выдвигаются следующие ключевые требования.

1°. Математическое ожидание ошибки модели $\varepsilon(x)$ для всех x из рассматриваемой области изменения равно нулю:

$$\mathbf{M}\left[\varepsilon(x)\right] = 0. \qquad (7.7)$$

Если математическое ожидание $\mathbf{M}\left[\varepsilon(x)\right] = \text{const} \neq 0$, то требование (7.7) может быть обеспечено для любых предполагаемых функций регрессии со свободным членом, поскольку он может

взять на себя ненулевое математическое ожидание ошибок. В связи с этим выбор моделей со свободным членом, как правило, предпочтительнее.

Нарушение требования (7.7) в общем случае приводит к смещённости оценок регрессионной модели.

2°. Вход модели X и ошибки модели $\varepsilon(x)$ для всех x из рассматриваемой области изменения – независимые случайные величины. Это требование называется требованием *экзогенности* входа модели.

Экзогенность входа X означает независимость случайной величины X от функционирования моделируемой системы. Значения экзогенных переменных определяются вне модели и не связаны с результатами работы системы.

Неэкзогенной является, например, модель

$$Y_t = f(Y_{t-1}, \beta_0, ..., \beta_{k-1}) + \varepsilon_t,$$

у которой вход Y_{t-1}, очевидно, зависит от ошибки модели ε_{t-1}.

Нарушение требования экзогенности приводит к существенному ухудшению статистических свойств оценок регрессионной модели.

Пусть (x_1, y_1), …, (x_n, y_n) – выборка наблюдений двумерного случайного вектора (X, Y), имеющего распределение $F_{XY}(x, y)$.

Оптимальным вектором параметров $\beta_0, ..., \beta_{k-1}$ будет вектор, при котором достигается минимум критерия (7.3) при каждом x. На практике вместо математического ожидания в критерии используется его оценка – среднее арифметическое. Таким образом, мерой *точности* регрессионной модели служит средний квадрат ошибок модели на выборочных значениях $x_1, ..., x_n$ (рис. 7.3):

$$\overline{\tilde{\varepsilon}^2} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}^2(x_i) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i, \beta_0, ..., \beta_{k-1})\right)^2 = D^*_{\text{ост}\,Y}, \qquad (7.8)$$

где $\tilde{\varepsilon}(x_i) = y_i - f(x_i, \beta_0, ..., \beta_{k-1})$ – реализация случайной ошибки $\varepsilon(x_i)$ модели в точке x_i, $i = \overline{1,n}$.

Несложно заметить, что формула (7.8) в точности совпадает с выражением (6.36) для остаточной дисперсии признака Y (см. § 32).

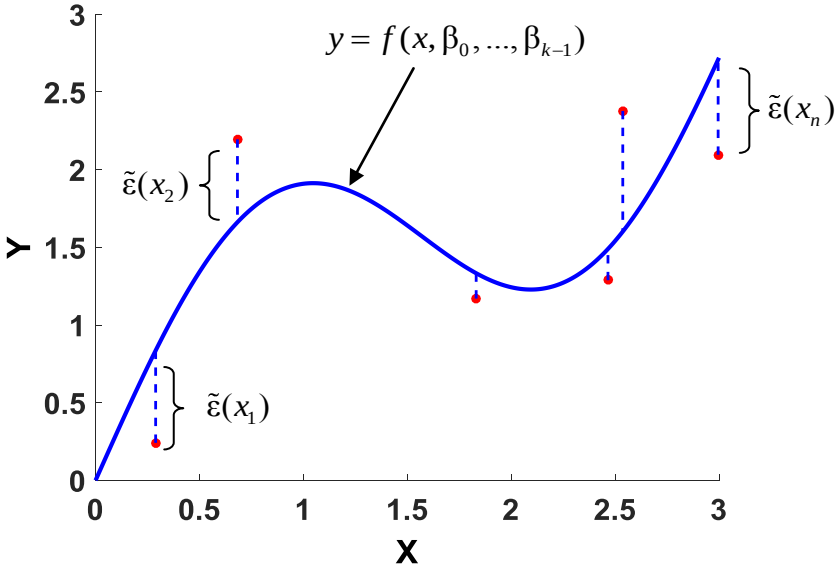

Рис. 7.3. Иллюстрация к методу наименьших квадратов

Таким образом, задача нахождения вектора параметров $\beta_0, ..., \beta_{k-1}$ представляет собой задачу минимизации остаточной дисперсии:

$$D^*_{\text{ост}\,Y}(\beta_0, ..., \beta_{k-1}) \to \min_{\beta_0, ..., \beta_{k-1}} . \qquad (7.9)$$

Вектор $\tilde{\beta} = (\tilde{\beta}_0, ..., \tilde{\beta}_{k-1})^T$, минимизирующий критерий (7.9), является рассчитанной по выборке $(x_1, y_1), \ldots, (x_n, y_n)$ точечной оценкой вектора параметров $\beta = (\beta_0, ..., \beta_{k-1})^T$ функции регрессии $f(x, \beta_0, ..., \beta_{k-1}) \in \Psi$. Метод расчёта вектора $\tilde{\beta}$, основанный на минимизации критерия (7.9), называется *методом наименьших квадратов* (*МНК*), а рассчитанные с его помощью оценки $\tilde{\beta}_0, ..., \tilde{\beta}_{k-1}$ – *МНК-оценками* (*Least Squares Estimates*, *LSE*).

Учитывая, что выборочная дисперсия D^*_Y случайной величины Y не зависит от функции регрессии и параметров $\beta_0, ..., \beta_{k-1}$, крите-

рий минимизации остаточной дисперсии (7.9) эквивалентен критерию максимизации показателя «эр-квадрат» (см. (6.39)):

$$R^{2\,*}_{Y|X}(\beta_0,...,\beta_{k-1}) \to \max_{\beta_0,...,\beta_{k-1}}. \qquad (7.10)$$

Поскольку показатель «эр-квадрат» характеризует долю вариации случайной величины *Y*, объяснённую функцией регрессии, суть метода наименьших квадратов состоит в подборе таких параметров функции регрессии из заданного класса функций Ψ, при которых она объясняет максимально возможную долю вариации признака *Y*.

Необходимым условием минимума функции $D^{*}_{\text{ост}\,Y}(\beta_0,...,\beta_{k-1})$ является равенство нулю частных производных:

$$\frac{\partial D^{*}_{\text{ост}\,Y}(\beta_0,...,\beta_{k-1})}{\partial \beta_j} = 0, \qquad j = \overline{0,k-1}. \qquad (7.11)$$

Подставляя выражение для остаточной дисперсии (7.8) в (7.11), получим систему *k* уравнений с *k* неизвестными:

$$\sum_{i=1}^{n}\left(y_i - f(x_i,\beta_0,...,\beta_{k-1})\right)\frac{\partial f(x_i,\beta_0,...,\beta_{k-1})}{\partial \beta_j} = 0, \quad j = \overline{0,k-1}, \quad (7.12)$$

решая которую относительно $\beta_0,...,\beta_{k-1}$, находим МНК-оценки $\tilde{\beta}_0,...,\tilde{\beta}_{k-1}$ параметров функции регрессии $f(x,\beta_0,...,\beta_{k-1})$. Значение $\tilde{f}(x) = f(x,\tilde{\beta}_0,...,\tilde{\beta}_{k-1})$ представляет собой МНК-оценку значения функции регрессии в точке *x*.

Подставляя в регрессионную модель (7.6) вместо функции регрессии $f(x)$ её оценку $\tilde{f}(x)$, получим построенную по выборке регрессионную модель:

$$\tilde{Y} \mid x = \tilde{f}(x) + \varepsilon(x), \qquad (7.13)$$

где $\tilde{Y} \mid x$ – оценка предсказанного значения случайной величины *Y* при фиксированном значении *x*; $\varepsilon(x)$ – случайная ошибка модели в точке *x*.

Подставляя выборочные значения $x_1,...,x_n$ в модель (7.13), получим множество случайных величин $\tilde{Y}_1,...,\tilde{Y}_n$, предсказанных моделью:

$$\tilde{Y}_i = \tilde{f}(x_i) + \varepsilon(x_i), \qquad (7.14)$$

реализациями которых являются выборочные значения $y_1, ..., y_n$. Разности между значениями $y_1, ..., y_n$ и расчётными значениями функции регрессии $\tilde{f}(x_1), ..., \tilde{f}(x_n)$ называются *регрессионными остатками* (*residuals*):

$$\tilde{\varepsilon}(x_i) = y_i - \tilde{f}(x_i), \quad i = \overline{1,n}. \qquad (7.15)$$

Регрессионные остатки $\tilde{\varepsilon}(x_1), ..., \tilde{\varepsilon}(x_n)$ являются реализациями случайных ошибок $\varepsilon(x_1), ..., \varepsilon(x_n)$ регрессионной модели (7.13) при значениях её входа, равных $x_1, ..., x_n$ соответственно.

Можно показать, что при соблюдении требований 1° и 2° к регрессионной модели оценки параметров $\tilde{\beta}_0, ..., \tilde{\beta}_{k-1}$ функции регрессии и её значения $\tilde{f}(x)$ в произвольной точке x являются состоятельными и несмещёнными.

Контрольные вопросы и задачи

1. Какие требования выдвигаются к регрессионным моделям? К чему может привести нарушение этих требований?
2. Почему в регрессионные модели рекомендуется включать свободный член?
3. Что называется экзогенностью входа модели?
4. В чём суть метода наименьших квадратов? Какой критерий оптимальности оценок регрессионной модели используется в методе наименьших квадратов?
5. Что является мерой точности регрессионной модели?
6. Какими свойствами обладают оценки параметров регрессионной модели, рассчитанные по методу наименьших квадратов?

§ 38. Простейшая линейная регрессионная модель

Пусть функция регрессии Y на X линейна:

$$f(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x. \qquad (7.16)$$

Тогда регрессионная модель (7.6) отклика Y будет иметь вид

$$Y \mid x = \beta_0 + \beta_1 x + \varepsilon(x). \qquad (7.17)$$

Такая модель называется *простейшей линейной регрессионной моделью* (*simple linear regression*).

Используя метод наименьших квадратов, найдём оценки параметров модели β_0, β_1. Запишем систему уравнений (7.12):

$$\begin{cases} \beta_0 n + \beta_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i; \\ \beta_0 \sum\limits_{i=1}^{n} x_i + \beta_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i, \end{cases} \tag{7.18}$$

решением которой являются оценки $\tilde{\beta}_0, \tilde{\beta}_1$:

$$\begin{aligned} \tilde{\beta}_0 &= \overline{y} - \rho^*_{XY} \frac{\sigma^*_Y}{\sigma^*_X} \overline{x}; \\ \tilde{\beta}_1 &= \rho^*_{XY} \frac{\sigma^*_Y}{\sigma^*_X}. \end{aligned} \tag{7.19}$$

Подставляя оценки (7.19) в (7.16), получим выражение для оценки значения простейшей линейной функции регрессии *Y* на *X* в точке *x*:

$$\tilde{f}(x) = f(x, \tilde{\beta}_0, \tilde{\beta}_1) = \overline{y} + \rho^*_{XY} \frac{\sigma^*_Y}{\sigma^*_X} (x - \overline{x}). \tag{7.20}$$

Заметим, что найденные МНК-оценки параметров простейшей линейной регрессии являются выборочными оценками теоретических значений (6.23), рассчитанных для функции регрессии нормально распределённых случайных величин (см. § 29).

МНК-оценки $\tilde{\beta}_0, \tilde{\beta}_1$ имеют следующие свойства:

1) линейная зависимость от результатов наблюдений $y_1, ..., y_n$;
2) состоятельность, т.е $\tilde{\beta}_i \xrightarrow{P} \beta_i$ при $n \to \infty$, $i = 0,1$;
3) несмещённость, т.е. $\mathbf{M}\left[\tilde{\beta}_i\right] = \beta_i$, $i = 0,1$.

Пусть к регрессионной модели (7.17) наряду с требованиями 1° и 2° (см. § 37) налагаются следующие дополнительные требования (*условия Гаусса–Маркова*).

3°. Дисперсии ошибок $\varepsilon(x)$ неизменны для всех *x* из рассматриваемой области определения:

$$\mathbf{D}\left[\varepsilon(x)\right] = \sigma^2 . \qquad (7.21)$$

Требование постоянства дисперсии произвольных случайных величин $\xi_1, ..., \xi_n$ называется требованием их *гомоскедастичности*. Если дисперсии случайных величин $\xi_1, ..., \xi_n$ различны, то такие величины называются *гетероскедастичными*.

Регрессионная модель называется *гомоскедастичной*, если гомоскедастичны её ошибки, т.е. если выполнено условие (7.21) для всех x из рассматриваемой области определения.

Гомоскедастичность ошибок простейшей линейной регрессионной модели связана с гомоскедастичностью наблюдаемой случайной величины Y при различных значениях x. Так, если при различных x дисперсии случайных величин $Y \mid x$ различны, то регрессионная модель будет гетероскедастичной.

Иногда гетероскедастичность данных можно обнаружить визуально (например, рис. 7.4, справа). Если диаграммы рассеяния не дают явной информации, тогда применяются статистические тесты на гомоскедастичность.

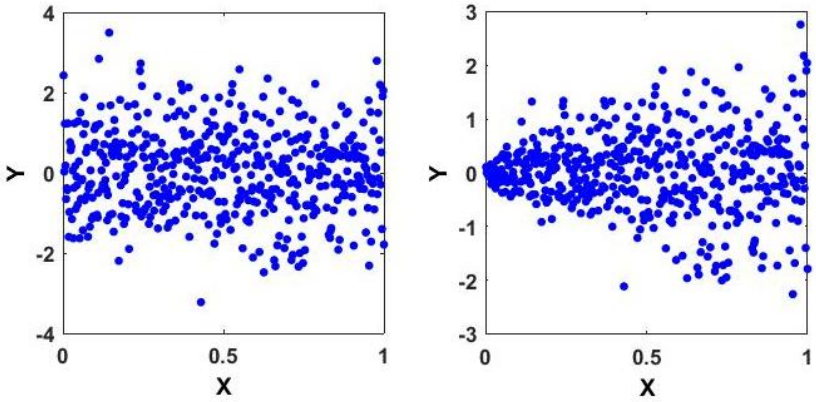

Рис. 7.4. Пример гомоскедастичных (слева)
и гетероскедастичных (справа) данных

Наличие гетероскедастичности в наблюдениях случайной величины Y приводит к неэффективности МНК-оценок $\tilde{\beta}_0, \tilde{\beta}_1$ и $\tilde{f}(x)$.

4°. Ошибки $\varepsilon(x)$ и $\varepsilon(x')$ модели независимы для всех x и x' из рассматриваемой области определения.

Если известно, что ошибки модели имеют нормальное распределение $\varepsilon(x) \sim N(0, \sigma)$ при всех x, то требование независимости эквивалентно требованию некоррелированности:

$$\operatorname{cov}\left[\varepsilon(x), \varepsilon(x')\right] = 0. \qquad (7.22)$$

Независимость остатков простейшей линейной регрессионной модели связана с независимостью наблюдаемых значений случайной величины Y при различных значениях x. Иными словами, выборка наблюдений $y_1, ..., y_n$ должна быть реализацией независимой случайной выборки $Y_1, ..., Y_n$. Если это требование не выполняется, то МНК-оценки $\tilde{\beta}_0, \tilde{\beta}_1$ и $\tilde{f}(x)$ являются неэффективными (*теорема Гаусса–Маркова*).

Требования 3° и 4° эквивалентны выполнению условия

$$V_\varepsilon = \sigma^2 I, \qquad (7.23)$$

где V_ε – ковариационная матрица ошибок регрессионной модели; I – единичная матрица.

При соблюдении требований 1°–4° оценки $\tilde{\beta}_0, \tilde{\beta}_1$ и $\tilde{f}(x)$ являются эффективными, т.е. оценками с наименьшей дисперсией в классе всех линейных несмещённых оценок. Эти дисперсии равны:

$$\mathbf{D}\left[\tilde{\beta}_0\right] = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n^2 D_X^*}, \qquad \mathbf{D}\left[\tilde{\beta}_1\right] = \frac{\sigma^2}{n D_X^*}. \qquad (7.24)$$

Если ошибки $\varepsilon(x)$ модели распределены нормально при всех x (что эквивалентно нормальности распределения случайной величины Y при всех x), то при соблюдении требований 1°–4° оценки $\tilde{\beta}_0, \tilde{\beta}_1$ также имеют нормальные законы распределения:

$$\tilde{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 D_X^*}}\right), \qquad \tilde{\beta}_1 \sim N\left(\beta_1, \sigma\sqrt{\frac{1}{nD_X^*}}\right).$$

Используя в качестве центральных статистик (см. § 10) стьюдентизованные случайные величины $\tilde{\beta}_0, \tilde{\beta}_1$ и учитывая, что оценкой дисперсии регрессионных остатков σ^2 является остаточная дисперсия признака Y (см. (7.8)), запишем формулы расчёта границ доверительных интервалов для параметров β_0, β_1 на уровне значимости α:

$$\left(\tilde{\beta}_0 - t_{1-\alpha/2}\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 D_X^*}}; \tilde{\beta}_0 + t_{1-\alpha/2}\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 D_X^*}}\right), \qquad (7.25)$$

$$\left(\tilde{\beta}_1 - t_{1-\alpha/2}\sqrt{\tilde{D}_{ост\,Y}}\sqrt{\frac{1}{nD_X^*}}; \tilde{\beta}_1 + t_{1-\alpha/2}\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\frac{1}{nD_X^*}}\right), \qquad (7.26)$$

где $t_{1-\alpha/2}$ – квантиль распределения Стьюдента с $n-2$ степенями свободы на уровне $1 - \alpha/2$; D_X^* – выборочная дисперсия случайной величины X; $\tilde{D}_{\text{ост}\,Y}$ – несмещённая оценка остаточной дисперсии случайной величины Y:

$$\tilde{D}_{\text{ост}\,Y} = \frac{1}{n-2}\sum_{i=1}^{n}\left(\tilde{f}(x_i) - y_i\right)^2. \qquad (7.27)$$

Доверительный интервал на уровне значимости α для функции регрессии $f(x) = \mathbf{M}[Y \mid x]$ вида (7.16) в точке x имеет вид

$$\left(\tilde{f}(x) - t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{nD_X^*}}; \right.$$
$$\left. \tilde{f}(x) + t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{nD_X^*}}\right). \qquad (7.28)$$

Отметим, что границы доверительного интервала для функции регрессии $f(x)$ нелинейно зависят от x, расширяясь при удалении x от $\bar{x}$ (рис. 7.5).

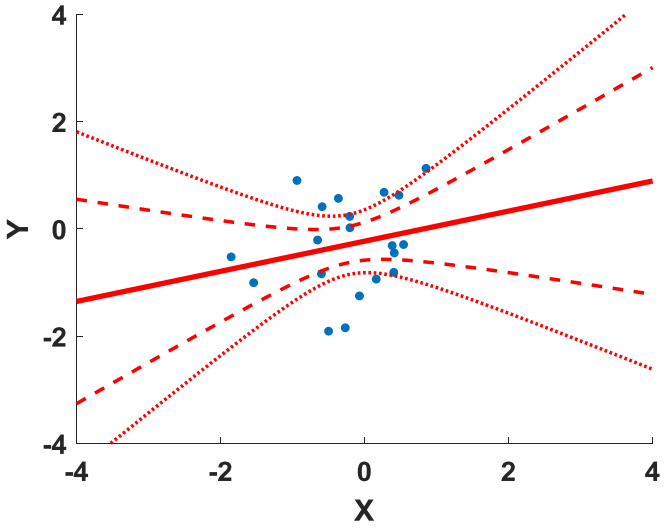

Рис. 7.5. Простейшая линейная регрессия

На рисунке сплошной линией изображена оцененная функция регрессии $\tilde{f}(x)$, пунктирными и точечными линиями – границы доверительных интервалов для $f(x)$ на уровнях значимости $\alpha = 0{,}1$ и $\alpha = 0{,}01$ соответственно.

Простейшая регрессионная модель (7.17) называется *значимой*, если $\beta_1 \neq 0$. Для проверки значимости модели (7.17) сформулируем основную и альтернативную гипотезы:

$$H_0 : \beta_1 = 0,$$

$$H' : \beta_1 \neq 0.$$

В качестве статистики критерия используется статистика

$$Z = \frac{R^{2\,*}_{Y|X}}{\left(1 - R^{2\,*}_{Y|X}\right) / (n-2)}, \qquad (7.29)$$

которая при условии истинности H_0 имеет распределение Фишера с 1 и $n-2$ степенями свободы в числителе и знаменателе соответственно: $f_Z(z \mid H_0) \sim F(1, n-2)$.

Критическая область для статистики критерия выбирается правосторонней.

Статистика критерия (7.29), используемая при проверке значимости простейшей линейной регрессионной модели, представляет собой статистику (6.46), используемую при проверке гипотезы о равенстве нулю коэффициента детерминации Y на X при числе неизвестных параметров функции регрессии $k = 2$ (см. § 32). Таким образом, гипотеза о значимости регрессионной модели эквивалентна гипотезе о равенстве нулю коэффициента детерминации для функции регрессии вида (7.16).

Пример 7.1. По выборочным данным примера 6.8 построить простейшую линейную регрессионную модель среднего времени решения вербальных заданий тестов и проверить её значимость на уровне $\alpha = 0{,}1$.

По условию задачи зависимой переменной является случайная величина Y – среднее время решения вербальных заданий тестов, регрессор модели – среднее время решения наглядно-образных заданий (случайная величина X).

Точечные оценки $\tilde{\beta}_0, \tilde{\beta}_1$ параметров β_0, β_1 простейшей линейной регрессионной модели (7.17) находим по методу наименьших квадратов (7.19):

$$\overline{x} = 22{,}5; \qquad \overline{y} = 13{,}3;$$

$$D_X^* \approx 59{,}6; \qquad D_Y^* \approx 14{,}4;$$

$$\rho_{XY}^* \approx 0{,}87;$$

$$\tilde{\beta}_0 = 13{,}3 - 0{,}87 \cdot \sqrt{\frac{14{,}4}{59{,}6}} \cdot 22{,}5 \approx 3{,}62;$$

$$\tilde{\beta}_1 = 0{,}87 \cdot \sqrt{\frac{14{,}4}{59{,}6}} \approx 0{,}43.$$

Таким образом, оценка функции регрессии имеет вид

$$\tilde{f}(x) = 3,62 + 0,43x.$$

Рассчитаем значения функции регрессии $\tilde{f}(x_i)$ в выборочных точках $x_1, ..., x_n$:

i	x_i	y_i	$\tilde{f}(x_i)$
1	19	15	11,80
2	12	7	8,79
3	32	17	17,39
4	17	14	10,94
5	14	8	9,65
6	25	15	14,38
7	15	8	10,08
8	35	17	18,68
9	29	16	16,10
10	27	16	15,24

По формулам (7.25), (7.26), (7.28) находим несмещённую оценку остаточной дисперсии и интервальные оценки параметров модели и значений функции регрессии:

$$D^*_{\text{ост } Y} = \frac{1}{10}\left((15-11,8)^2 + ... + (16-15,24)^2\right) \approx 3,38;$$

$$\tilde{D}_{\text{ост } Y} = \frac{10}{8} \cdot 3,38 \approx 4,23;$$

$$\underline{\beta}_0 = 3,62 - 1,86\sqrt{4,23}\sqrt{5659/(100 \cdot 59,6)} \approx -0,11;$$

$$\overline{\beta}_0 = 3,62 + 1,86\sqrt{4,23}\sqrt{5659/(100 \cdot 59,6)} \approx 7,34;$$

$$\beta_0 \in (-0,11; 7,34);$$

$$\underline{\beta}_1 = 0,43 - 1,86\sqrt{4,23}\sqrt{1/(10 \cdot 59,6)} \approx 0,27;$$

$$\overline{\beta}_1 = 0,43 + 1,86\sqrt{4,23}\sqrt{1/(10 \cdot 59,6)} \approx 0,59;$$

$$\beta_1 \in (0,27; 0,59);$$

$$1{,}86\sqrt{4{,}23}\sqrt{\frac{1}{10}+\frac{(x-22{,}5)^2}{10\cdot 59{,}6}} \approx 0{,}16\sqrt{59{,}6+(x-22{,}5)^2};$$

$$f(x)\in\Big(3{,}62+0{,}43x-0{,}16\sqrt{59{,}6+(x-22{,}5)^2};$$

$$3{,}62+0{,}43x+0{,}16\sqrt{59{,}6+(x-22{,}5)^2}\Big).$$

Визуальное представление выборочной функции регрессии и границ доверительных интервалов приведено на рис. 7.6.

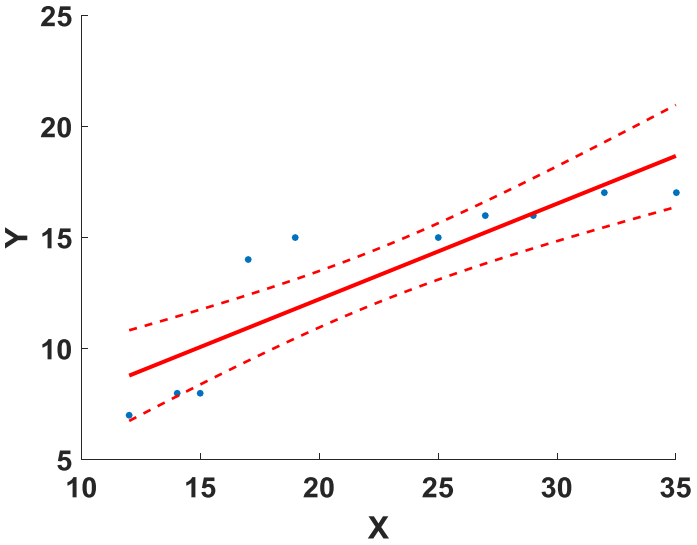

Рис. 7.6. Функция регрессии $\tilde{f}(x)=3{,}62+0{,}43x$
и границы доверительных интервалов при $\alpha=0{,}1$

Рассчитаем показатель «эр-квадрат» по формуле (6.39):

$$R^{2\,*}_{Y|X}=1-\frac{3{,}38}{14{,}4}\approx 0{,}77\,.$$

В силу линейности функции регрессии убеждаемся, что

$$R^{2\,*}_{Y|X}=\rho^{*2}_{XY}=0{,}87^2\approx 0{,}77\,.$$

Скорректированный показатель «эр-квадрат» находим по формуле (6.45):

$$\bar{R}^2_{Y|X} = 1-(1-0,77)\frac{9}{8} \approx 0,74 .$$

Для проверки гипотезы о незначимости регрессионной модели $H_0 : \beta_1 = 0$ рассчитаем выборочное значение статистики Фишера (7.29):

$$z = \frac{0,77}{0,23/8} \approx 26,7 ,$$

которому соответствует значение *p-value*

$$p = 1 - F_{F(1,8)}(26,7) \approx 0,0009 .$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза должна быть отклонена, т.е. рассматриваемая линейная регрессионная модель среднего времени решения вербальных заданий тестов значима.

Контрольные вопросы и задачи

1. Какая регрессионная модель называется простейшей линейной?
2. Какими свойствами обладают МНК-оценки параметров простейшей линейной регрессионной модели?
3. Покажите, что МНК-оценки параметров простейшей линейной регрессионной модели линейно зависят от выборочных значений $y_1, ..., y_n$.
4. Покажите, что МНК-оценки параметров простейшей линейной регрессионной модели являются несмещёнными.
5. Сформулируйте условия Гаусса–Маркова. Какими свойствами обладают МНК-оценки параметров простейшей линейной регрессионной модели при их выполнении?
6. Какие случайные величины называются гомоскедастичными?
7. Какая регрессионная модель называется гомоскедастичной?
8. Какая регрессионная модель называется значимой?
9. Какая статистика критерия используется при проверке гипотезы о значимости простейшей линейной регрессионной модели?

Какой закон распределения имеет эта статистика при условии истинности основной гипотезы?

10. Объясните принцип выбора типа критической области при проверке гипотезы о значимости простейшей линейной регрессионной модели.

11. Сформулируйте метод доверительных интервалов для проверки гипотезы о значимости простейшей линейной регрессионной модели.

§ 39. Линейная регрессионная модель общего вида

Пусть функция регрессии имеет вид

$$f(x, \beta_0, ..., \beta_{k-1}) = \beta_0 \varphi_0(x) + ... + \beta_{k-1} \varphi_{k-1}(x), \qquad (7.30)$$

где $\varphi_0(x), ..., \varphi_{k-1}(x)$ – некоторая система функций (не обязательно линейных).

Тогда регрессионная модель (7.6) отклика Y на входное воздействие x выглядит следующим образом:

$$Y \mid x = \beta_0 \varphi_0(x) + ... + \beta_{k-1} \varphi_{k-1}(x) + \varepsilon(x), \qquad (7.31)$$

где $\varepsilon(x)$ – случайная ошибка модели.

Такая модель называется *линейной регрессионной моделью общего вида* (*Generalized Linear Model, GLM*), или просто *линейной регрессионной моделью*. Под линейностью регрессионной модели понимается линейность по её параметрам $\beta_0, ..., \beta_{k-1}$.

Как правило, в качестве функции $\varphi_0(x)$ выбирается тождественная единица:

$$\varphi_0(x) \equiv 1. \qquad (7.32)$$

При условии постоянства математического ожидания ошибок модели $\mathbf{M}[\varepsilon(x)] = \text{const} \neq 0$ такой выбор обеспечивает выполнение требования 1° (см. § 37), предъявляемого к регрессионным моделям.

Используя метод наименьших квадратов, найдём оценки параметров модели $\beta_0, ..., \beta_{k-1}$. Запишем систему уравнений (7.12) в матричном виде:

$$F^T F \beta = F^T y, \qquad (7.33)$$

где $\beta = (\beta_0, ..., \beta_{k-1})^T$ – вектор параметров модели; $y = (y_1, ..., y_n)^T$ – вектор откликов модели; F – матрица размерности $n \times k$, составленная из значений функций $\varphi_0(x), ..., \varphi_{k-1}(x)$ в выборочных точках $x_1, ..., x_n$:

$$F = \begin{pmatrix} \varphi_0(x_1) & ... & \varphi_{k-1}(x_1) \\ \varphi_0(x_2) & ... & \varphi_{k-1}(x_2) \\ ... & ... & ... \\ \varphi_0(x_n) & ... & \varphi_{k-1}(x_n) \end{pmatrix}, \qquad (7.34)$$

называемая *регрессионной матрицей*, или *матрицей плана* (*design matrix*).

Решая систему (7.33), получаем вектор $\tilde{\beta} = (\tilde{\beta}_0, ..., \tilde{\beta}_{k-1})^T$ МНК-оценок параметров модели (7.31):

$$\tilde{\beta} = (F^T F)^{-1} F^T y. \qquad (7.35)$$

МНК-оценка $\tilde{f}(x)$ линейной функции регрессии (7.30) Y на X в точке x имеет вид

$$\tilde{f}(x) = f(x, \tilde{\beta}_0, ..., \tilde{\beta}_{k-1}) = \sum_{j=0}^{k-1} \tilde{\beta}_j \varphi_j(x). \qquad (7.36)$$

При соблюдении требований 1°–4° к регрессионным моделям (см. § 37, 38) МНК-оценки (7.35) и (7.36) имеют те же свойства, что и МНК-оценки простейшей линейной регрессионной модели, а ковариационная матрица вектора МНК-оценок $\tilde{\beta}$ равна

$$\operatorname{cov}\left[\tilde{\beta}\right] = \sigma^2 (F^T F)^{-1}, \qquad (7.37)$$

где σ^2 – дисперсия ошибок модели (7.31). Матрица $(F^T F)^{-1}$ называется *дисперсионной матрицей Фишера*.

Если ошибки $\varepsilon(x)$ модели распределены нормально при всех x (что эквивалентно нормальности распределения случайной величины Y при всех x), то при соблюдении требований 1°–4° вектор МНК-оценок $\tilde{\beta}$ имеет k-мерное нормальное распределение с вектором математических ожиданий β и ковариационной матрицей (7.37).

Границы доверительного интервала на уровне значимости α для параметра β_j регрессионной модели (7.31), $j = \overline{0, k-1}$, рассчитываются по формуле

$$\left(\tilde{\beta}_j - t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{c_{jj}}; \tilde{\beta}_j + t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{c_{jj}}\right), \quad (7.38)$$

где c_{jj} – j-й (считая от нуля) диагональный элемент матрицы $(F^T F)^{-1}$, $j = \overline{0, k-1}$; $t_{1-\alpha/2}(n-k)$ – квантиль распределения Стьюдента с $n-k$ степенями свободы на уровне $1 - \alpha/2$; $\tilde{D}_{\text{ост}\,Y}$ – несмещённая оценка остаточной дисперсии случайной величины Y:

$$\tilde{D}_{\text{ост}\,Y} = \frac{1}{n-k}\sum_{i=1}^{n}\left(\tilde{f}(x_i) - y_i\right)^2, \quad (7.39)$$

Доверительный интервал на уровне значимости α для значения функции регрессии $f(x) = \mathbf{M}[Y \mid x]$ в точке x имеет вид:

$$\begin{aligned}\Big(&\tilde{f}(x) - t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\varphi^T(x)(F^T F)^{-1}\varphi(x)};\\ &\tilde{f}(x) + t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{\text{ост}\,Y}}\sqrt{\varphi^T(x)(F^T F)^{-1}\varphi(x)}\Big),\end{aligned} \quad (7.40)$$

где $\varphi(x) = \left(\varphi_0(x), ..., \varphi_{k-1}(x)\right)^T$ – вектор значений системы функций в точке x.

Пример функции регрессии, линейной по параметрам, приведён на рис. 7.7.

На рисунке сплошной линией изображена оцененная по выборке функция регрессии $\tilde{f}(x) = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 x^2 + \tilde{\beta}_3 x^3$, пунктирными и точечными линиями – границы доверительных интервалов для $f(x)$ на уровнях значимости $\alpha = 0{,}1$ и $\alpha = 0{,}01$ соответственно.

Линейная регрессионная модель (7.31) называется *значимой*, если соответствующая ей функция регрессии зависит от x. В частности, если выполнено условие (7.32), то модель значима, если хотя бы один из коэффициентов $\beta_1, ..., \beta_{k-1}$ отличен от нуля. Если все $\beta_1 = ... = \beta_{k-1} = 0$, то модель называется *незначимой*.

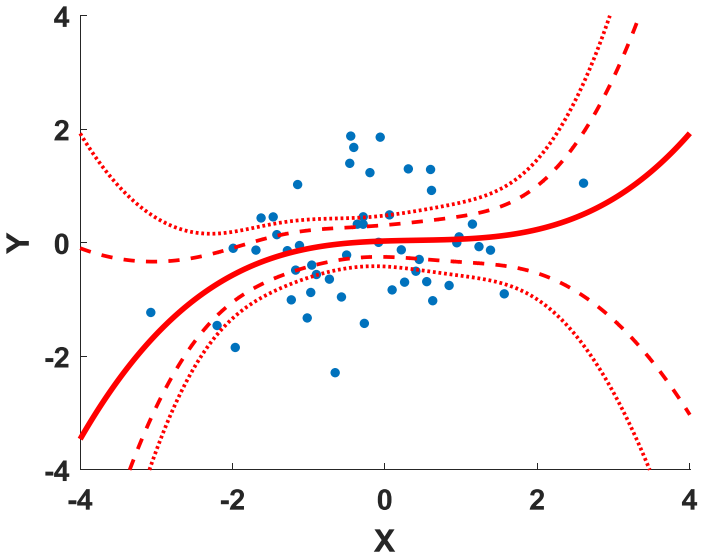

Рис. 7.7. Линейная регрессия общего вида

Проверка значимости линейной регрессионной модели означает проверку статистической гипотезы

$$H_0 : \beta_1 = \ldots = \beta_{k-1} = 0$$

против альтернативной гипотезы

$$H' : \sum_{j=1}^{k-1} \beta_j^2 > 0 .$$

В качестве статистики критерия используется статистика

$$Z = \frac{R^{2\,*}_{Y|X} \big/ (k-1)}{\left(1 - R^{2\,*}_{Y|X}\right) \big/ (n-k)}, \qquad (7.41)$$

которая при условии истинности H_0 имеет распределение Фишера с $k-1$ и $n-k$ степенями свободы в числителе и знаменателе соответственно: $f_Z(z \mid H_0) \sim F(k-1, n-k)$.

Критическая область для статистики критерия выбирается правосторонней.

Статистика критерия (7.41), используемая при проверке значимости линейной регрессионной модели, представляет собой статистику (6.46), используемую при проверке гипотезы о равенстве нулю коэффициента детерминации Y на X при числе неизвестных параметров функции регрессии, равном k (см. § 32). Таким образом, гипотеза о значимости регрессионной модели (7.31) эквивалентна гипотезе о равенстве нулю коэффициента детерминации для функции регрессии вида (7.30).

Для проверки гипотезы о равенстве нулю параметра β_j линейной регрессионной модели

$$H_0 : \beta_j = 0, \qquad j = \overline{0, k-1},$$

используется статистика критерия

$$Z = \frac{\tilde{\beta}_j}{\sqrt{\tilde{D}_{\text{ост } Y}}\sqrt{c_{jj}}}, \tag{7.42}$$

которая при условии истинности H_0 имеет распределение Стьюдента с $n-k$ степенями свободы: $f_Z(z \mid H_0) \sim T(n-k)$.

Критическая область для статистики критерия выбирается, исходя из вида альтернативной гипотезы.

Система функций $\varphi_0(x), ..., \varphi_{k-1}(x)$, вообще говоря, может быть выбрана произвольным образом, однако на практике удобно использовать некоторую систему ортогональных функций.

Система функций $\varphi_0(x), ..., \varphi_{k-1}(x)$ называется *ортогональной* на выборке $x_1, ..., x_n$, если

$$\sum_{i=1}^{n} \varphi_m(x_i)\varphi_l(x_i) = 0, \qquad \forall m, l = \overline{0, k-1}, \qquad m \neq l. \tag{7.43}$$

Если $\varphi_0(x) \equiv 1$, то, полагая $m = 0$, из (7.43) получим:

$$\sum_{i=1}^{n} \varphi_l(x_i) = 0, \qquad \forall l = \overline{1, k-1}, \tag{7.44}$$

т.е. функции, образующие ортогональную систему с константой, центрированы. Следовательно, условие ортогональности (7.43) в терминах математической статистики означает условие некоррелированности функций $\varphi_m(x)$ и $\varphi_l(x)$ при $m \neq l$:

$$\operatorname{cov}\left[\varphi_m(x), \varphi_l(x)\right] = 0, \quad \forall m, l = \overline{1, k-1}, \quad m \neq l. \qquad (7.45)$$

Можно показать, что если система функций $\varphi_0(x), ..., \varphi_{k-1}(x)$ ортогональна на выборке $x_1, ..., x_n$, то МНК-оценки параметров $\beta_0, ..., \beta_{k-1}$ регрессионной модели (7.31) вычисляются по формуле

$$\tilde{\beta}_j = \frac{\sum_{i=1}^{n} y_i \varphi_j(x_i)}{\sum_{i=1}^{n} \varphi_j^2(x_i)}, \quad j = \overline{0, k-1}. \qquad (7.46)$$

В качестве системы ортогональных функций $\varphi_0(x), ..., \varphi_{k-1}(x)$ могут быть выбраны, например, ортогональные полиномы Чебышева или Эрмита.

Пример 7.2. По выборочным данным примера 6.8 построить линейную регрессионную модель вида

$$Y \mid x = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon(x),$$

где случайная величина Y – среднее время решения вербальных заданий тестов, а регрессор – среднее время решения наглядно-образных заданий (случайная величина X), и проверить её значимость на уровне $\alpha = 0{,}1$.

Из заданного вида регрессионной модели получаем, что вектор значений системы функций $\varphi_0(x), ..., \varphi_{k-1}(x)$ в точке x равен $\varphi(x) = \left(1, x, x^2, x^3\right)^T$. Число неизвестных параметров модели $k = 4$.

Запишем регрессионную матрицу F:

1	x	x^2	x^3
1	19	361	6859
1	12	144	1728
1	32	1024	32768
1	17	289	4913
1	14	196	2744
1	25	625	15625
1	15	225	3375

1	x	x^2	x^3
1	35	1225	42875
1	29	841	24389
1	27	729	19683

Дисперсионная матрица Фишера $(F^T F)^{-1}$ имеет размерность 4×4 и равна

$$(F^T F)^{-1} \approx \begin{pmatrix} 150,5 & -21,0 & 0,91 & -0,01 \\ -21,0 & 2,96 & -0,13 & 0,002 \\ 0,91 & -0,13 & 0,006 & 0 \\ -0,01 & 0,002 & 0 & 0 \end{pmatrix}.$$

По формуле (7.35) получаем точечную оценку вектора параметров модели:

$$\tilde{\beta} = \left(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3\right)^T \approx \left(-34,1; 5,2; -0,18; 0,0021\right)^T.$$

Таким образом, оценка функции регрессии имеет вид

$$\tilde{f}(x) = -34,1 + 5,2x - 0,18x^2 + 0,0021x^3.$$

Рассчитаем значения функции регрессии $\tilde{f}(x_i)$ в выборочных точках $x_1, ..., x_n$:

i	x_i	y_i	$\tilde{f}(x_i)$
1	19	15	13,83
2	12	7	5,85
3	32	17	16,45
4	17	14	12,34
5	14	8	8,99
6	25	15	15,85
7	15	8	10,27
8	35	17	17,16
9	29	16	16,17
10	27	16	16,04

Показатели «эр-квадрат» и «эр-бар-квадрат» рассчитываем по формулам (6.39) и (6.45):

$$R^{2\,*}_{Y|X} \approx 0{,}91;$$

$$\overline{R}^{2}_{Y|X} \approx 0{,}87.$$

Сравнивая эти показатели с полученными в примере 7.1, делаем вывод, что рассматриваемая линейная модель обладает лучшими объяснительными возможностями, чем простейшая линейная регрессионная модель.

Визуальное представление выборочной функции регрессии и доверительных интервалов для неё приведено на рис. 7.8.

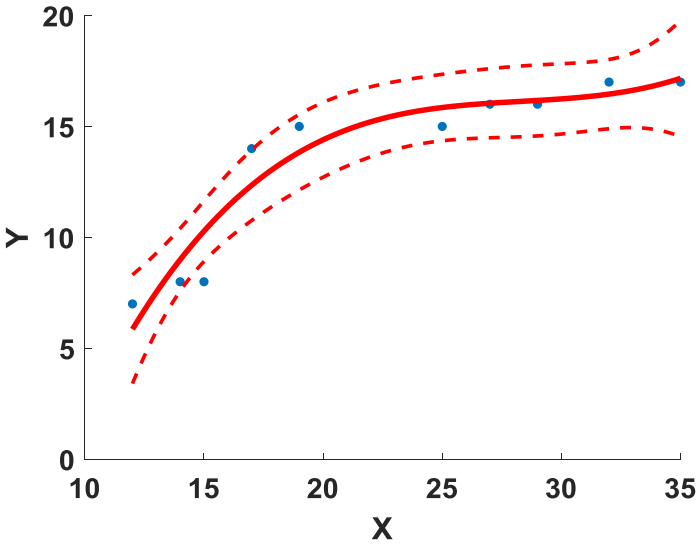

Рис. 7.8. Функция регрессии $\tilde{f}(x) = -34{,}1 + 5{,}2x - 0{,}18x^2 + 0{,}0021x^3$

Для проверки гипотезы $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ о незначимости регрессионной модели рассчитаем выборочное значение статистики Фишера (7.41):

$$z = \frac{0,91/(4-1)}{0,09/(10-4)} \approx 20,2,$$

которому соответствует значение *p-value*

$$p = 1 - F_{F(3,6)}(20,2) \approx 0,0014.$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза должна быть отклонена, т.е. рассматриваемая линейная регрессионная модель значима.

Контрольные вопросы и задачи

1. Какая регрессионная модель называется линейной регрессионной моделью общего вида? Что означает линейность модели?
2. Для чего функция $\varphi_0(x)$ в линейной регрессионной модели, как правило, выбирается равной константе?
3. Что называется регрессионной матрицей модели?
4. Что называется дисперсионной матрицей Фишера?
5. Какими свойствами обладают МНК-оценки параметров линейной регрессионной модели общего вида?
6. Какая регрессионная модель называется значимой?
7. Какая статистика критерия используется при проверке гипотезы о значимости линейной регрессионной модели? Какой закон распределения имеет эта статистика при условии истинности основной гипотезы?
8. Какая статистика критерия используется при проверке гипотезы о равенстве нулю коэффициента линейной регрессионной модели? Какой закон распределения имеет эта статистика при условии истинности основной гипотезы?
9. Может ли гипотеза об одновременном равенстве нулю коэффициентов регрессионной модели быть сведена к последовательной проверке гипотез о равенстве нулю отдельных коэффициентов модели при заданном уровне значимости?
10. Какая система функций называется ортогональной на выборке $x_1, ..., x_n$? Какой статистический смысл имеет ортогональность?
11. По значению какого показателя сравниваются объяснительные способности двух регрессионных моделей?

§ 40. Множественная линейная регрессия

Если на систему действует множество факторов $X_1, ..., X_m$, то её регрессионная модель имеет вид

$$Y \mid x = f(x, \beta_0, ..., \beta_{k-1}) + \varepsilon(x), \qquad (7.47)$$

где $f(x, \beta_0, ..., \beta_{k-1})$ – функция регрессии; $\beta_0, ..., \beta_{k-1}$ – параметры модели; $\varepsilon(x)$ – случайная ошибка модели; $x = (x_1, ..., x_m)$ – вектор входных воздействий.

Пусть функция регрессии является линейной (по параметрам):

$$f(x, \beta_0, ..., \beta_{k-1}) = \beta_0 \varphi_0(x) + ... + \beta_{k-1} \varphi_{k-1}(x), \qquad (7.48)$$

где $\varphi_0(x), ..., \varphi_{k-1}(x)$ – некоторая система скалярных функций (не обязательно линейных) *m* переменных. В этом случае матрица плана регрессионной модели (7.47) строится аналогично матрице плана (7.34) линейной регрессионной модели общего вида. МНК-оценки параметров $\beta_0, ..., \beta_{k-1}$ функции регрессии (7.48) рассчитываются по формуле (7.35).

Для расчёта доверительных интервалов параметров модели и проверки значимости модели используются те же формулы, что и для линейной регрессионной модели (7.31).

Рассмотрим частный случай функции регрессии (7.48). Пусть $k-1=m$, а функции $\varphi_0(x), ..., \varphi_{k-1}(x)$ заданы следующим образом:

$$\begin{gathered} \varphi_0(x_1, ..., x_m) \equiv 1, \\ \varphi_j(x_1, ..., x_m) = x_j, \quad j = \overline{1, m}. \end{gathered} \qquad (7.49)$$

Тогда функция регрессии (7.48) определяет гиперплоскость в пространстве признаков $(x_1, ..., x_m, y)$:

$$f(x, \beta_0, ..., \beta_m) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m. \qquad (7.50)$$

Пусть $(x_{11}, ..., x_{m1}, y_1)$, …, $(x_{1n}, ..., x_{mn}, y_n)$ – выборка наблюдений случайного вектора $(X_1, ..., X_m, Y)$. В соответствии с формулой (6.39) по этим данным может быть рассчитан показатель «эр-квадрат»:

$$R^{2\,*}_{Y|X_1,...,X_m} = \frac{D^*_{\text{регр}\,Y|X_1,...,X_m}}{D^*_Y} = 1 - \frac{D^*_{\text{ост}\,Y}}{D^*_Y}. \qquad (7.51)$$

Этот показатель следует интерпретировать как долю вариации выборочных данных, объяснённую линейной функцией регрессии (7.48). Величина остаточной дисперсии $D^*_{\text{ост}\,Y}$ характеризует разброс выборочных значений относительно гиперплоскости регрессии.

При анализе линейного уравнения регрессии (7.50) выборочное корреляционное отношение $R^*_{Y|X_1,...,X_m}$ называют также *множественным коэффициентом корреляции* (*multiple correlation*). В отличие от линейного коэффициента корреляции Пирсона, изменяющегося от –1 до 1, множественный коэффициент корреляции принимает значения в диапазоне от 0 до 1.

Можно показать, что множественный коэффициент корреляции $R^*_{Y|X_1,...,X_m}$ выражается через парные коэффициенты корреляции следующим образом:

$$R^*_{Y|X_1,...,X_m} = \sqrt{c^T R^{-1}_{XX} c}\,, \tag{7.52}$$

где R_{XX} – корреляционная матрица регрессоров $X_1, ..., X_m$ размерности $m \times m$; c – вектор-столбец корреляций отклика Y с регрессорами $X_1, ..., X_m$.

В частном случае при $m = 2$ формула (7.52) имеет вид

$$R^*_{Y|X_1,X_2} = \sqrt{\frac{(\rho^*_{YX_1})^2 + (\rho^*_{YX_2})^2 - 2\rho^*_{YX_1}\rho^*_{YX_2}\rho^*_{X_1X_2}}{1-(\rho^*_{X_1X_2})^2}}\,. \tag{7.53}$$

Добавление в регрессионную модель новых регрессоров всегда увеличивает или оставляет неизменным значение показателя «эр-квадрат». Связано это с тем, что с увеличением размерности пространства признаков ошибка линейной аппроксимации n точек может только уменьшиться или остаться неизменной (при нулевом коэффициенте перед добавляемым признаком). Эта особенность является недостатком показателя «эр-квадрат», поскольку подобное увеличение его значения может быть не связано с наличием статистической связи между рассматриваемым откликом Y модели и факторами $X_1, ..., X_m$.

Показателем, компенсирующим этот эффект, является скорректированное корреляционное отношение (см. § 32):

$$\bar{R}_{Y|X_1,\ldots,X_m} = \sqrt{1 - \frac{D^*_{\text{ост}\,Y}/(n-k)}{D^*_Y/(n-1)}} = \sqrt{1 - \left(1 - R^{2\,*}_{Y|X_1,\ldots,X_m}\right)\frac{n-1}{n-k}}, \quad (7.54)$$

при анализе линейного уравнения регрессии (7.50) называемое также *скорректированным множественным коэффициентом корреляции* (*adjusted multiple correlation*).

Пример 7.3. Используя выборочные данные примера 6.8, построить линейную регрессионную модель вида

$$Y \mid x, z = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon(x, z),$$

где Y – среднее время решения вербальных заданий тестов, а регрессоры: X – среднее время решения наглядно-образных заданий, Z – рост школьника, и проверить её значимость на уровне $\alpha = 0{,}1$. Выборочные значения признака Z представлены в таблице:

Номер школьника	Рост Z, см
1	156
2	160
3	159
4	158
5	154
6	157
7	151
8	156
9	156
10	154

Определить, даёт ли рост школьника значимый вклад в построенную регрессионную модель.

Запишем регрессионную матрицу F:

1	x	z
1	19	156
1	12	160
1	32	159
1	17	158
1	14	154
1	25	157
1	15	151
1	35	156
1	29	156
1	27	154

Дисперсионная матрица Фишера $(F^T F)^{-1}$ имеет размерность 3×3 и равна

$$(F^T F)^{-1} \approx \begin{pmatrix} 389,7 & 0,06 & -2,5 \\ 0,06 & 0,002 & -0,001 \\ -2,5 & -0,001 & 0,016 \end{pmatrix}.$$

По формуле (7.35) получаем точечную оценку вектора параметров модели:

$$\tilde{\beta} = \left(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2\right)^T \approx (-22,9; 0,42; 0,17)^T.$$

Таким образом, оценка функции регрессии имеет вид

$$\tilde{f}(x,z) = -22,9 + 0,42x + 0,17z.$$

Показатели «эр-квадрат» и «эр-бар-квадрат» рассчитываем по формулам (6.39) и (6.45):

$$R^{2\,*}_{Y|X_1X_2} \approx 0,78;$$

$$\overline{R}^2_{Y|X_1X_2} \approx 0,72.$$

Сравнивая эти показатели с полученными в примере 7.1 ($R^{2\,*}_{Y|X} \approx 0,77$, $\overline{R}^2_{Y|X} \approx 0,74$), делаем вывод, что рассматриваемая плоскость регрессии в пространстве признаков (x, z, y) немного лучше аппроксимирует выборочные данные, чем прямая регрессии

на плоскости признаков (*x*, *y*), однако её объяснительные возможности хуже. Этот результат говорит о том, что наблюдаемое увеличение показателя «эр-квадрат» является ложным и не связано с наличием статистической зависимости между признаками *y* и *z*.

Для проверки гипотезы $H_0 : \beta_1 = \beta_2 = 0$ о незначимости регрессионной модели рассчитаем выборочное значение статистики Фишера (7.41):

$$z = \frac{0,78 / (3-1)}{(1-0,78) / (10-3)} \approx 12,4,$$

которому соответствует значение *p-value*

$$p = 1 - F_{F(2,7)}(12,4) \approx 0,005.$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза должна быть отклонена, т.е. рассматриваемая линейная регрессионная модель значима.

Проверим теперь гипотезу об отсутствии вклада фактора *Z* в построенную регрессионную модель:

$$H_0 : \beta_2 = 0$$

против альтернативной гипотезы $H' : \beta_2 \neq 0$.

Рассчитаем выборочное значение остаточной дисперсии и её несмещённую оценку:

$$D^*_{\text{ост}\,Y} = \frac{1}{10}\left(\left(15 - \tilde{f}(19,156)\right)^2 + \ldots + \left(16 - \tilde{f}(27,154)\right)^2\right) \approx 3,2;$$

$$\tilde{D}_{\text{ост}\,Y} = \frac{10}{7} \cdot 3,2 \approx 4,57.$$

По формуле (7.42) находим выборочное значение статистики критерия:

$$z = \frac{0,17}{\sqrt{4,57 \cdot 0,016}} \approx 0,63,$$

которому соответствует значение *p-value*

$$p = 2(1 - F_{T(7)}(0,63)) \approx 0,55.$$

Согласно критерию проверки статистических гипотез, делаем вывод, что основная гипотеза H_0 должна быть принята, т.е. вклад фактора «рост школьника» в регрессионную модель незначим.

Контрольные вопросы и задачи

1. Какая регрессионная модель называется множественной линейной? Что означает линейность модели?
2. Что называется множественным коэффициентом корреляции? Какие возможные значения он может принимать?
3. Как множественный коэффициент корреляции связан с парными коэффициентами корреляции отклика и регрессоров модели?
4. Какой недостаток имеет множественный коэффициент корреляции при увеличении числа регрессоров? В каком случае этим недостатком можно пренебречь?
5. Что называется скорректированным множественным коэффициентом корреляции? Как следует интерпретировать его значение?
6. Следует ли из равенства нулю множественного коэффициента корреляции $R^*_{Y|X_1,X_2}$ равенство нулю коэффициентов $R^*_{X_1|Y,X_2}$ и $R^*_{X_2|Y,X_1}$?
7. Следует ли из равенства нулю множественного коэффициента корреляции $R^*_{Y|X_1,X_2}$ равенство нулю парных коэффициентов корреляции $\rho^*_{YX_1}$, $\rho^*_{YX_2}$ и $\rho^*_{X_1X_2}$? Верно ли обратное утверждение?

§ 41. Некоторые регрессионные модели, сводящиеся к линейным

В ряде практических приложений нередко возникает необходимость построения регрессионных моделей, нелинейных по параметрам. Такие модели позволяют описывать более широкий класс статистических зависимостей, при этом, как правило, обладая меньшим числом параметров. Кроме того, линейные модели могут противоречить самой природе моделируемого объекта.

Для нахождения оптимальных МНК-параметров нелинейных регрессионных моделей могут быть использованы различные численные методы оптимизации, в частности методы градиентного спуска, метод сопряжённых градиентов, метод Левенберга-Маркардта и пр.

В частных случаях нелинейные модели могут быть сведены к линейным путём замены переменных. При переходе к новым переменным следует иметь в виду, что при вычислении оценок параметров модели по методу наименьших квадратов минимизируется сумма квадратов преобразованных, а не исходных данных. Свойства полученных оценок зависят от того, выполняются ли требования 1°–4° (см. § 37, 38) именно для преобразованных переменных.

Рассмотрим некоторые наиболее часто встречающиеся нелинейные модели с одним фактором, сводящиеся к линейным. Все модели могут быть легко обобщены на случай многих факторов.

1. Экспоненциальная модель. Функция регрессии имеет вид

$$f(x, \beta_0, \beta_1) = \beta_0 e^{\beta_1 x}. \qquad (7.55)$$

Эта функция может отражать, например, зависимость концентрации вещества, участвующего в химической реакции (величина Y), от времени (величина X), зависимость скорости распада радиоактивных ядер от времени, а также ряд других физических, химических и экономических зависимостей.

Для сведения модели к линейной прологарифмируем обе части равенства (7.55):

$$\ln f(x, \beta_0, \beta_1) = \ln \beta_0 + \beta_1 x.$$

Делая замену переменных:

$$g(x, \beta_0, \beta_1) = \ln f(x, \beta_0, \beta_1),$$

$$\gamma_0 = \ln \beta_0, \qquad \gamma_1 = \beta_1,$$

получаем линейную функцию регрессии:

$$g(x, \beta_0, \beta_1) = \gamma_0 + \gamma_1 x,$$

для которой находим МНК-оценки $\tilde{\gamma}_0, \tilde{\gamma}_1$ параметров γ_0, γ_1, используя преобразованную выборку наблюдений

$$(x_1, \ln y_1), \ldots, (x_n, \ln y_n).$$

Оценки параметров β_0, β_1 функции регрессии (7.55) связаны с оценками линейной регрессионной модели соотношениями:

$$\tilde{\beta}_0 = e^{\tilde{\gamma}_0}, \qquad \tilde{\beta}_1 = \tilde{\gamma}_1.$$

2. Степенная модель. Функция регрессии имеет вид

$$f(x, \beta_0, \beta_1) = \beta_0 x^{\beta_1}. \qquad (7.56)$$

Эта функция может отражать, например, зависимость спроса на товар (величина *Y*) от его цены (величина *X*), зависимость объёма производства от объёма использованных ресурсов.

Прологарифмируем обе части равенства (7.56):

$$\ln f(x, \beta_0, \beta_1) = \ln \beta_0 + \beta_1 \ln x .$$

Делая замену переменных:

$$g(x, \beta_0, \beta_1) = \ln f(x, \beta_0, \beta_1),$$

$$z = \ln x , \qquad \gamma_0 = \ln \beta_0 , \qquad \gamma_1 = \beta_1 ,$$

получаем линейную функцию регрессии:

$$g(z, \beta_0, \beta_1) = \gamma_0 + \gamma_1 z ,$$

для которой находим МНК-оценки $\tilde{\gamma}_0, \tilde{\gamma}_1$ параметров γ_0, γ_1, используя преобразованную выборку наблюдений

$$(\ln x_1, \ln y_1), \ldots, (\ln x_n, \ln y_n).$$

Оценки параметров β_0, β_1 функции регрессии (7.56) связаны с оценками линейной регрессионной модели соотношениями:

$$\tilde{\beta}_0 = e^{\tilde{\gamma}_0} , \qquad \tilde{\beta}_1 = \tilde{\gamma}_1 .$$

3. Обратная модель. Функция регрессии имеет вид

$$f(x, \beta_0, \beta_1) = \frac{1}{\beta_0 + \beta_1 x} . \qquad (7.57)$$

Эта функция может отражать, например, зависимость спроса на малоценный товар (величина *Y*) от уровня доходов (величина *X*), зависимость заработной платы от уровня безработицы.

Делая замену переменных:

$$g(x, \beta_0, \beta_1) = \frac{1}{f(x, \beta_0, \beta_1)} ,$$

получаем линейную функцию регрессии:

$$g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x ,$$

для которой находим МНК-оценки $\tilde{\beta}_0, \tilde{\beta}_1$ параметров β_0, β_1, используя преобразованную выборку наблюдений

$$\left(x_1, \frac{1}{y_1} \right), \ldots, \left(x_n, \frac{1}{y_n} \right).$$

4. Логистическая модель. Функция регрессии имеет вид

$$f(x, \beta_0, \beta_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \qquad (7.58)$$

Логистическая кривая используется для описания поведения показателей, имеющих определённые «уровни насыщения», например, зависимость результата, приносимого от внедрения новой технологии (величина *Y*), от затрат на её развитие (величина *X*), зависимость численности популяции от времени при ограниченных ресурсах среды. Логистические модели находят также широкое применение в задачах оценивания вероятностей возникновения событий в зависимости от значения факторного признака *X*.

Преобразуя выражение (7.58), запишем:

$$\ln\left(\frac{f(x, \beta_0, \beta_1)}{1 - f(x, \beta_0, \beta_1)}\right) = \beta_0 + \beta_1 x.$$

Делая замену переменных:

$$g(x, \beta_0, \beta_1) = \ln\left(\frac{f(x, \beta_0, \beta_1)}{1 - f(x, \beta_0, \beta_1)}\right),$$

получаем линейную функцию регрессии:

$$g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x,$$

для которой находим МНК-оценки $\tilde{\beta}_0, \tilde{\beta}_1$ параметров β_0, β_1, используя преобразованную выборку наблюдений

$$\left(x_1, \ln\frac{y_1}{1 - y_1}\right), \ldots, \left(x_n, \ln\frac{y_n}{1 - y_n}\right).$$

Контрольные вопросы и задачи

1. Какая регрессионная модель называется нелинейной?
2. Какие методы используются для поиска оптимальных параметров нелинейной регрессионной модели?
3. Следует ли оптимальность оценок параметров нелинейной модели из оптимальности МНК-оценок параметров соответствующей линейной модели, полученной путём замены переменных?
4. Приведите примеры процессов, моделируемых с помощью экспоненциальной, степенной, обратной, логистической регрессионных моделей.

ПРИЛОЖЕНИЕ. ТАБЛИЦЫ КВАНТИЛЕЙ РАСПРЕДЕЛЕНИЙ

Таблица П1

Квантили нормального распределения $N(0, 1)$

p	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
x_p	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Таблица П2

Квантили распределения Колмогорова

p	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
x_p	1,224	1,358	1,480	1,628	1,731	1,949	2,036

Таблица ПЗ

Квантили распределения хи-квадрат $\chi^2(k)$

k	**0,0005**	**0,001**	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**
1	4e-07	2e-06	4e-05	2e-04	0,001	0,004	0,02
2	0,001	0,002	0,01	0,02	0,05	0,10	0,21
3	0,02	0,02	0,07	0,11	0,22	0,35	0,58
4	0,06	0,09	0,21	0,30	0,48	0,71	1,06
5	0,16	0,21	0,41	0,55	0,83	1,15	1,61
6	0,30	0,38	0,68	0,87	1,24	1,64	2,20
7	0,48	0,60	0,99	1,24	1,69	2,17	2,83
8	0,71	0,86	1,34	1,65	2,18	2,73	3,49
9	0,97	1,15	1,73	2,09	2,70	3,33	4,17
10	1,26	1,48	2,16	2,56	3,25	3,94	4,87
11	1,59	1,83	2,60	3,05	3,82	4,57	5,58
12	1,93	2,21	3,07	3,57	4,40	5,23	6,30
13	2,31	2,62	3,57	4,11	5,01	5,89	7,04
14	2,70	3,04	4,07	4,66	5,63	6,57	7,79
15	3,11	3,48	4,60	5,23	6,26	7,26	8,55
16	3,54	3,94	5,14	5,81	6,91	7,96	9,31
17	3,98	4,42	5,70	6,41	7,56	8,67	10,09
18	4,44	4,90	6,26	7,01	8,23	9,39	10,86
19	4,91	5,41	6,84	7,63	8,91	10,12	11,65
20	5,40	5,92	7,43	8,26	9,59	10,85	12,44
21	5,90	6,45	8,03	8,90	10,28	11,59	13,24
22	6,40	6,98	8,64	9,54	10,98	12,34	14,04
23	6,92	7,53	9,26	10,20	11,69	13,09	14,85
24	7,45	8,08	9,89	10,86	12,40	13,85	15,66
25	7,99	8,65	10,52	11,52	13,12	14,61	16,47
26	8,54	9,22	11,16	12,20	13,84	15,38	17,29
27	9,09	9,80	11,81	12,88	14,57	16,15	18,11
28	9,66	10,39	12,46	13,56	15,31	16,93	18,94
29	10,23	10,99	13,12	14,26	16,05	17,71	19,77
30	10,80	11,59	13,79	14,95	16,79	18,49	20,60
35	13,79	14,69	17,19	18,51	20,57	22,47	24,80
40	16,91	17,92	20,71	22,16	24,43	26,51	29,05
45	20,14	21,25	24,31	25,90	28,37	30,61	33,35
50	23,46	24,67	27,99	29,71	32,36	34,76	37,69
100	59,90	61,92	67,33	70,06	74,22	77,93	82,36

Окончание табл. П3

Квантили распределения хи-квадрат $\chi^2(k)$

k	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	2,71	3,84	5,02	6,63	7,88	10,83	12,12
2	4,61	5,99	7,38	9,21	10,60	13,82	15,20
3	6,25	7,81	9,35	11,34	12,84	16,27	17,73
4	7,78	9,49	11,14	13,28	14,86	18,47	20,00
5	9,24	11,07	12,83	15,09	16,75	20,52	22,11
6	10,64	12,59	14,45	16,81	18,55	22,46	24,10
7	12,02	14,07	16,01	18,48	20,28	24,32	26,02
8	13,36	15,51	17,53	20,09	21,95	26,12	27,87
9	14,68	16,92	19,02	21,67	23,59	27,88	29,67
10	15,99	18,31	20,48	23,21	25,19	29,59	31,42
11	17,28	19,68	21,92	24,72	26,76	31,26	33,14
12	18,55	21,03	23,34	26,22	28,30	32,91	34,82
13	19,81	22,36	24,74	27,69	29,82	34,53	36,48
14	21,06	23,68	26,12	29,14	31,32	36,12	38,11
15	22,31	25,00	27,49	30,58	32,80	37,70	39,72
16	23,54	26,30	28,85	32,00	34,27	39,25	41,31
17	24,77	27,59	30,19	33,41	35,72	40,79	42,88
18	25,99	28,87	31,53	34,81	37,16	42,31	44,43
19	27,20	30,14	32,85	36,19	38,58	43,82	45,97
20	28,41	31,41	34,17	37,57	40,00	45,31	47,50
21	29,62	32,67	35,48	38,93	41,40	46,80	49,01
22	30,81	33,92	36,78	40,29	42,80	48,27	50,51
23	32,01	35,17	38,08	41,64	44,18	49,73	52,00
24	33,20	36,42	39,36	42,98	45,56	51,18	53,48
25	34,38	37,65	40,65	44,31	46,93	52,62	54,95
26	35,56	38,89	41,92	45,64	48,29	54,05	56,41
27	36,74	40,11	43,19	46,96	49,64	55,48	57,86
28	37,92	41,34	44,46	48,28	50,99	56,89	59,30
29	39,09	42,56	45,72	49,59	52,34	58,30	60,73
30	40,26	43,77	46,98	50,89	53,67	59,70	62,16
35	46,06	49,80	53,20	57,34	60,27	66,62	69,20
40	51,81	55,76	59,34	63,69	66,77	73,40	76,09
45	57,51	61,66	65,41	69,96	73,17	80,08	82,88
50	63,17	67,50	71,42	76,15	79,49	86,66	89,56
100	118,50	124,34	129,56	135,81	140,17	149,45	153,17

Таблица П4

Квантили распределения Стьюдента $T(k)$

k	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	3,078	6,314	12,706	31,821	63,657	318,309	636,619
2	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373
200	1,286	1,653	1,972	2,345	2,601	3,131	3,340
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Таблица П5

Квантили распределения Фишера $F(k_1, k_2)$,
значения параметра k_2 указаны в первой строке таблицы

а) $p = 0{,}9$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	39,9	8,53	5,54	4,54	4,06	3,78	3,59	3,46	3,36	3,29	3,07	2,97	2,88	2,84	2,81	2,76	2,73
2	49,5	9,00	5,46	4,32	3,78	3,46	3,26	3,11	3,01	2,92	2,70	2,59	2,49	2,44	2,41	2,36	2,33
3	53,6	9,16	5,39	4,19	3,62	3,29	3,07	2,92	2,81	2,73	2,49	2,38	2,28	2,23	2,20	2,14	2,11
4	55,9	9,24	5,34	4,11	3,52	3,18	2,96	2,81	2,69	2,61	2,36	2,25	2,14	2,09	2,06	2,00	1,97
5	57,2	9,29	5,31	4,05	3,45	3,11	2,88	2,73	2,61	2,52	2,27	2,16	2,05	2,00	1,97	1,91	1,88
6	58,2	9,33	5,28	4,01	3,40	3,05	2,83	2,67	2,55	2,46	2,21	2,09	1,98	1,93	1,90	1,83	1,80
7	58,9	9,35	5,27	3,98	3,37	3,01	2,78	2,62	2,51	2,41	2,16	2,04	1,93	1,87	1,84	1,78	1,75
8	59,4	9,37	5,25	3,95	3,34	2,98	2,75	2,59	2,47	2,38	2,12	2,00	1,88	1,83	1,80	1,73	1,70
9	59,9	9,38	5,24	3,94	3,32	2,96	2,72	2,56	2,44	2,35	2,09	1,96	1,85	1,79	1,76	1,69	1,66
10	60,2	9,39	5,23	3,92	3,30	2,94	2,70	2,54	2,42	2,32	2,06	1,94	1,82	1,76	1,73	1,66	1,63
11	60,5	9,40	5,22	3,91	3,28	2,92	2,68	2,52	2,40	2,30	2,04	1,91	1,79	1,74	1,70	1,64	1,60
12	60,7	9,41	5,22	3,90	3,27	2,90	2,67	2,50	2,38	2,28	2,02	1,89	1,77	1,71	1,68	1,61	1,58
13	60,9	9,41	5,21	3,89	3,26	2,89	2,65	2,49	2,36	2,27	2,00	1,87	1,75	1,70	1,66	1,59	1,56
14	61,1	9,42	5,20	3,88	3,25	2,88	2,64	2,48	2,35	2,26	1,99	1,86	1,74	1,68	1,64	1,57	1,54
15	61,2	9,42	5,20	3,87	3,24	2,87	2,63	2,46	2,34	2,24	1,97	1,84	1,72	1,66	1,63	1,56	1,52
20	61,7	9,44	5,18	3,84	3,21	2,84	2,59	2,42	2,30	2,20	1,92	1,79	1,67	1,61	1,57	1,49	1,46
25	62,1	9,45	5,17	3,83	3,19	2,81	2,57	2,40	2,27	2,17	1,89	1,76	1,63	1,57	1,53	1,45	1,41
30	62,3	9,46	5,17	3,82	3,17	2,80	2,56	2,38	2,25	2,16	1,87	1,74	1,61	1,54	1,50	1,42	1,38
40	62,5	9,47	5,16	3,80	3,16	2,78	2,54	2,36	2,23	2,13	1,85	1,71	1,57	1,51	1,46	1,38	1,34
50	62,7	9,47	5,15	3,80	3,15	2,77	2,52	2,35	2,22	2,12	1,83	1,69	1,55	1,48	1,44	1,35	1,31

Продолжение 1 табл. П5

б) $p = 0{,}95$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	161	18,5	10,1	7,71	6,61	5,99	5,59	5,32	5,12	4,96	4,54	4,35	4,17	4,08	4,03	3,94	3,89
2	200	19,0	9,55	6,94	5,79	5,14	4,74	4,46	4,26	4,10	3,68	3,49	3,32	3,23	3,18	3,09	3,04
3	216	19,2	9,28	6,59	5,41	4,76	4,35	4,07	3,86	3,71	3,29	3,10	2,92	2,84	2,79	2,70	2,65
4	225	19,3	9,12	6,39	5,19	4,53	4,12	3,84	3,63	3,48	3,06	2,87	2,69	2,61	2,56	2,46	2,42
5	230	19,3	9,01	6,26	5,05	4,39	3,97	3,69	3,48	3,33	2,90	2,71	2,53	2,45	2,40	2,31	2,26
6	234	19,3	8,94	6,16	4,95	4,28	3,87	3,58	3,37	3,22	2,79	2,60	2,42	2,34	2,29	2,19	2,14
7	237	19,4	8,89	6,09	4,88	4,21	3,79	3,50	3,29	3,14	2,71	2,51	2,33	2,25	2,20	2,10	2,06
8	239	19,4	8,85	6,04	4,82	4,15	3,73	3,44	3,23	3,07	2,64	2,45	2,27	2,18	2,13	2,03	1,98
9	241	19,4	8,81	6,00	4,77	4,10	3,68	3,39	3,18	3,02	2,59	2,39	2,21	2,12	2,07	1,97	1,93
10	242	19,4	8,79	5,96	4,74	4,06	3,64	3,35	3,14	2,98	2,54	2,35	2,16	2,08	2,03	1,93	1,88
11	243	19,4	8,76	5,94	4,70	4,03	3,60	3,31	3,10	2,94	2,51	2,31	2,13	2,04	1,99	1,89	1,84
12	244	19,4	8,74	5,91	4,68	4,00	3,57	3,28	3,07	2,91	2,48	2,28	2,09	2,00	1,95	1,85	1,80
13	245	19,4	8,73	5,89	4,66	3,98	3,55	3,26	3,05	2,89	2,45	2,25	2,06	1,97	1,92	1,82	1,77
14	246	19,4	8,71	5,87	4,64	3,96	3,53	3,24	3,03	2,86	2,42	2,22	2,04	1,95	1,89	1,79	1,74
15	246	19,4	8,70	5,86	4,62	3,94	3,51	3,22	3,01	2,85	2,40	2,20	2,01	1,92	1,87	1,77	1,72
20	248	19,5	8,66	5,80	4,56	3,87	3,44	3,15	2,94	2,77	2,33	2,12	1,93	1,84	1,78	1,68	1,62
25	249	19,5	8,63	5,77	4,52	3,83	3,40	3,11	2,89	2,73	2,28	2,07	1,88	1,78	1,73	1,62	1,56
30	250	19,5	8,62	5,75	4,50	3,81	3,38	3,08	2,86	2,70	2,25	2,04	1,84	1,74	1,69	1,57	1,52
40	251	19,5	8,59	5,72	4,46	3,77	3,34	3,04	2,83	2,66	2,20	1,99	1,79	1,69	1,63	1,52	1,46
50	252	19,5	8,58	5,70	4,44	3,75	3,32	3,02	2,80	2,64	2,18	1,97	1,76	1,66	1,60	1,48	1,41

Продолжение 2 табл. П5

в) $p = 0{,}975$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	647	39	17,4	12,2	10,0	8,81	8,07	7,57	7,21	6,94	6,20	5,87	5,57	5,42	5,34	5,18	5,10
2	799	39	16,0	10,7	8,43	7,26	6,54	6,06	5,71	5,46	4,77	4,46	4,18	4,05	3,97	3,83	3,76
3	864	39	15,4	9,98	7,76	6,60	5,89	5,42	5,08	4,83	4,15	3,86	3,59	3,46	3,39	3,25	3,18
4	899	39	15,1	9,60	7,39	6,23	5,52	5,05	4,72	4,47	3,80	3,51	3,25	3,13	3,05	2,92	2,85
5	921	39	14,9	9,36	7,15	5,99	5,29	4,82	4,48	4,24	3,58	3,29	3,03	2,90	2,83	2,70	2,63
6	937	39	14,7	9,20	6,98	5,82	5,12	4,65	4,32	4,07	3,41	3,13	2,87	2,74	2,67	2,54	2,47
7	948	39	14,6	9,07	6,85	5,70	4,99	4,53	4,20	3,95	3,29	3,01	2,75	2,62	2,55	2,42	2,35
8	956	39	14,5	8,98	6,76	5,60	4,90	4,43	4,10	3,85	3,20	2,91	2,65	2,53	2,46	2,32	2,26
9	963	39	14,5	8,90	6,68	5,52	4,82	4,36	4,03	3,78	3,12	2,84	2,57	2,45	2,38	2,24	2,18
10	968	39	14,4	8,84	6,62	5,46	4,76	4,30	3,96	3,72	3,06	2,77	2,51	2,39	2,32	2,18	2,11
11	973	39	14,4	8,79	6,57	5,41	4,71	4,24	3,91	3,66	3,01	2,72	2,46	2,33	2,26	2,12	2,06
12	976	39	14,3	8,75	6,52	5,37	4,67	4,20	3,87	3,62	2,96	2,68	2,41	2,29	2,22	2,08	2,01
13	979	39	14,3	8,71	6,49	5,33	4,63	4,16	3,83	3,58	2,92	2,64	2,37	2,25	2,18	2,04	1,97
14	982	39	14,3	8,68	6,46	5,30	4,60	4,13	3,80	3,55	2,89	2,60	2,34	2,21	2,14	2,00	1,93
15	984	39	14,3	8,66	6,43	5,27	4,57	4,10	3,77	3,52	2,86	2,57	2,31	2,18	2,11	1,97	1,90
20	993	40	14,2	8,56	6,33	5,17	4,47	4,00	3,67	3,42	2,76	2,46	2,20	2,07	1,99	1,85	1,78
25	0,1*	40	14,1	8,50	6,27	5,11	4,40	3,94	3,60	3,35	2,69	2,40	2,12	1,99	1,92	1,77	1,70
30	0,1*	40	14,1	8,46	6,23	5,07	4,36	3,89	3,56	3,31	2,64	2,35	2,07	1,94	1,87	1,71	1,64
40	0,1*	40	14,0	8,41	6,18	5,01	4,31	3,84	3,51	3,26	2,59	2,29	2,01	1,88	1,80	1,64	1,56
50	0,1*	40	14,0	8,38	6,14	4,98	4,28	3,81	3,47	3,22	2,55	2,25	1,97	1,83	1,75	1,59	1,51

г) $p = 0,99$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	0,4*	99	34,1	21,2	16,3	13,8	12,3	11,3	10,6	10,0	8,68	8,10	7,56	7,31	7,17	6,90	6,76
2	0,5*	99	30,0	18,0	13,3	10,9	9,55	8,65	8,02	7,56	6,36	5,85	5,39	5,18	5,06	4,82	4,71
3	0,5*	99	29,5	16,7	12,1	9,78	8,45	7,59	6,99	6,55	5,42	4,94	4,51	4,31	4,20	3,98	3,88
4	0,6*	99	28,7	16,0	11,4	9,15	7,85	7,01	6,42	5,99	4,89	4,43	4,02	3,83	3,72	3,51	3,41
5	0,6*	99	28,2	15,6	11,0	8,75	7,46	6,63	6,06	5,64	4,56	4,10	3,70	3,51	3,41	3,21	3,11
6	0,6*	99	27,9	15,2	10,7	8,47	7,19	6,37	5,80	5,39	4,32	3,87	3,47	3,29	3,19	2,99	2,89
7	0,6*	99	27,7	15,0	10,5	8,26	6,99	6,18	5,61	5,20	4,14	3,70	3,30	3,12	3,02	2,82	2,73
8	0,6*	99	27,5	14,8	10,3	8,10	6,84	6,03	5,47	5,06	4,00	3,56	3,17	2,99	2,89	2,69	2,60
9	0,6*	99	27,4	14,7	10,2	7,98	6,72	5,91	5,35	4,94	3,89	3,46	3,07	2,89	2,78	2,59	2,50
10	0,6*	99	27,2	14,6	10,1	7,87	6,62	5,81	5,26	4,85	3,80	3,37	2,98	2,80	2,70	2,50	2,41
11	0,6*	99	27,1	14,5	10,0	7,79	6,54	5,73	5,18	4,77	3,73	3,29	2,91	2,73	2,63	2,43	2,34
12	0,6*	99	27,1	14,4	9,89	7,72	6,47	5,67	5,11	4,71	3,67	3,23	2,84	2,66	2,56	2,37	2,27
13	0,6*	99	27,0	14,3	9,82	7,66	6,41	5,61	5,05	4,65	3,61	3,18	2,79	2,61	2,51	2,31	2,22
14	0,6*	99	26,9	14,3	9,77	7,60	6,36	5,56	5,01	4,60	3,56	3,13	2,74	2,56	2,46	2,27	2,17
15	0,6*	99	26,9	14,2	9,72	7,56	6,31	5,52	4,96	4,56	3,52	3,09	2,70	2,52	2,42	2,22	2,13
20	0,6*	99	26,7	14,0	9,55	7,40	6,16	5,36	4,81	4,41	3,37	2,94	2,55	2,37	2,27	2,07	1,97
25	0,6*	99	26,6	13,9	9,45	7,30	6,06	5,26	4,71	4,31	3,28	2,84	2,45	2,27	2,17	1,97	1,87
30	0,6*	99	26,5	13,9	9,38	7,23	5,99	5,20	4,65	4,25	3,21	2,78	2,39	2,20	2,10	1,89	1,79
40	0,6*	99	26,4	13,8	9,29	7,14	5,91	5,12	4,57	4,17	3,13	2,69	2,30	2,11	2,01	1,80	1,69
50	0,6*	99	26,4	13,7	9,24	7,09	5,86	5,07	4,52	4,12	3,08	2,64	2,25	2,06	1,95	1,74	1,63

Продолжение 4 табл. П5

д) $p = 0{,}995$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	1,6*	199	56	31,3	22,8	18,6	16,2	14,7	13,6	12,8	10,8	9,94	9,18	8,83	8,63	8,24	8,06
2	2,0*	199	50	26,3	18,3	14,5	12,4	11,0	10,1	9,43	7,70	6,99	6,35	6,07	5,90	5,59	5,44
3	2,2*	199	48	24,3	16,5	12,9	10,9	9,60	8,72	8,08	6,48	5,82	5,24	4,98	4,83	4,54	4,41
4	2,2*	199	46	23,2	15,6	12,0	10,1	8,81	7,96	7,34	5,80	5,17	4,62	4,37	4,23	3,96	3,84
5	2,3*	199	45	22,5	14,9	11,5	9,52	8,30	7,47	6,87	5,37	4,76	4,23	3,99	3,85	3,59	3,47
6	2,3*	199	45	22,0	14,5	11,1	9,16	7,95	7,13	6,54	5,07	4,47	3,95	3,71	3,58	3,33	3,21
7	2,4*	199	44	21,6	14,2	10,8	8,89	7,69	6,88	6,30	4,85	4,26	3,74	3,51	3,38	3,13	3,01
8	2,4*	199	44	21,4	14,90	10,6	8,68	7,50	6,69	6,12	4,67	4,09	3,58	3,35	3,22	2,97	2,86
9	2,4*	199	44	21,1	13,8	10,4	8,51	7,34	6,54	5,97	4,54	3,96	3,45	3,22	3,09	2,85	2,73
10	2,4*	199	44	21,0	13,6	10,3	8,38	7,21	6,42	5,85	4,42	3,85	3,34	3,12	2,99	2,74	2,63
11	2,4*	199	44	20,8	13,5	10,1	8,27	7,10	6,31	5,75	4,33	3,76	3,25	3,03	2,90	2,66	2,54
12	2,4*	199	43	20,7	13,4	10,0	8,18	7,01	6,23	5,66	4,25	3,68	3,18	2,95	2,82	2,58	2,47
13	2,5*	199	43	20,6	13,3	9,95	8,10	6,94	6,15	5,59	4,18	3,61	3,11	2,89	2,76	2,52	2,40
14	2,5*	199	43	20,5	13,2	9,88	8,03	6,87	6,09	5,53	4,12	3,55	3,06	2,83	2,70	2,46	2,35
15	2,5*	199	43	20,4	13,2	9,81	7,97	6,81	6,03	5,47	4,07	3,50	3,01	2,78	2,65	2,41	2,30
20	2,5*	199	43	20,2	12,9	9,59	7,75	6,61	5,83	5,27	3,88	3,32	2,82	2,60	2,47	2,23	2,11
25	2,5*	199	43	20,0	12,8	9,45	7,62	6,48	5,71	5,15	3,77	3,20	2,71	2,48	2,35	2,11	1,99
30	2,5*	199	43	19,9	12,7	9,36	7,53	6,40	5,62	5,07	3,69	3,12	2,63	2,40	2,27	2,02	1,91
40	2,5*	199	42	19,8	12,6	9,24	7,42	6,29	5,52	4,97	3,58	3,02	2,52	2,30	2,16	1,91	1,79
50	2,5*	199	42	19,7	12,5	9,17	7,35	6,22	5,45	4,90	3,52	2,96	2,46	2,23	2,10	1,84	1,71

е) $p = 0{,}999$

k_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	200
1	40*	999	167	74,1	47,2	35,5	29,2	25,4	22,9	21,0	16,6	14,8	13,3	12,6	12,2	11,5	11,1
2	50*	999	148	61,2	37,1	27,0	21,7	18,5	16,4	14,9	11,3	9,95	8,77	8,25	7,96	7,41	7,15
3	54*	999	141	56,2	33,2	23,7	18,8	15,8	13,9	12,6	9,34	8,10	7,05	6,59	6,34	5,86	5,63
4	56*	999	137	53,4	31,1	21,9	17,2	14,4	12,6	11,3	8,25	7,10	6,12	5,70	5,46	5,02	4,81
5	58*	999	135	51,7	29,8	20,8	16,2	13,5	11,7	10,5	7,57	6,46	5,53	5,13	4,90	4,48	4,29
6	59*	999	133	50,5	28,8	20,0	15,5	12,9	11,1	9,93	7,09	6,02	5,12	4,73	4,51	4,11	3,92
7	59*	999	132	49,7	28,2	19,5	15,0	12,4	10,7	9,52	6,74	5,69	4,82	4,44	4,22	3,83	3,65
8	60*	999	131	49,0	27,6	19,0	14,6	12,0	10,4	9,20	6,47	5,44	4,58	4,21	4,00	3,61	3,43
9	60*	999	130	48,5	27,2	18,7	14,3	11,8	10,1	8,96	6,26	5,24	4,39	4,02	3,82	3,44	3,26
10	61*	999	129	48,1	26,9	18,4	14,1	11,5	9,9	8,75	6,08	5,08	4,24	3,87	3,67	3,30	3,12
11	61*	999	129	47,7	26,6	18,2	13,9	11,4	9,7	8,59	5,94	4,94	4,11	3,75	3,55	3,18	3,00
12	61*	999	128	47,4	26,4	18,0	13,7	11,2	9,6	8,45	5,81	4,82	4,00	3,64	3,44	3,07	2,90
13	61*	999	128	47,2	26,2	17,8	13,6	11,1	9,4	8,32	5,71	4,72	3,91	3,55	3,35	2,99	2,82
14	61*	999	128	46,9	26,1	17,7	13,4	10,9	9,3	8,22	5,62	4,64	3,82	3,47	3,27	2,91	2,74
15	62*	999	127	46,8	25,9	17,6	13,3	10,8	9,2	8,13	5,54	4,56	3,75	3,40	3,20	2,84	2,67
20	62*	999	126	46,1	25,4	17,1	12,9	10,5	8,9	7,80	5,25	4,29	3,49	3,14	2,95	2,59	2,42
25	62*	999	126	45,7	25,1	16,9	12,7	10,3	8,7	7,60	5,07	4,12	3,33	2,98	2,79	2,43	2,26
30	63*	999	125	45,4	24,9	16,7	12,5	10,1	8,5	7,47	4,95	4,00	3,22	2,87	2,68	2,32	2,15
40	63*	999	125	45,1	24,6	16,4	12,3	9,9	8,4	7,30	4,80	3,86	3,07	2,73	2,53	2,17	2,00
50	63*	999	125	44,9	24,4	16,3	12,2	9,8	8,3	7,19	4,70	3,77	2,98	2,64	2,44	2,08	1,90

Примечание: * – умножение на 10^4.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М. : ЮНИТИ-ДАНА, 2012.
2. Математическая статистика: Учеб. для вузов / Под ред. В.С. Зарубина, А.П. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002.
3. Сборник задач по математике для втузов / Под ред. А.В. Ефимова. Ч.4. Теория вероятностей и математическая статистика. – М.: Физматлит, 2003.
4. Мишулина О.А. Основы теории вероятностей: Учебн. пособие. – М.: НИЯУ МИФИ, 2011.
5. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2003.
6. Куликов Е.И. Прикладной статистический анализ. – М.: Радио и связь, 2003.
7. Пугачев В.С. Теория вероятностей и математическая статистика: Учебник. – М.: Физматлит, 2011.
8. Кибзун А.И., Горяинова Е.Р., Наумов А.В. и др. Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами: Учебн. пособие. – М.: Физматлит, 2005.
9. Ватутин В.А., Ивченко Г.И., Медведев Ю.И. и др. Теория вероятностей и математическая статистика в задачах. – М.: Дрофа, 2005.

Александр Геннадьевич Трофимов

Основы
математической статистики

Учебное пособие

Редактор М.В. Макарова

Подписано в печать 20.11.2015. Формат 60×84 1/16
Уч.-изд. л. 16,0. Печ. л. 16,0. Изд. № 1/49.

Национальный исследовательский ядерный университет «МИФИ».
115409, Москва, Каширское ш., 31.