

Федеральное агентство
научных организаций

Российская
академия наук

Московский государственный университет имени М.В. Ломоносова
Московская школа экономики
Федеральное государственное бюджетное учреждение науки
Институт социально-экономического развития территорий
Российской академии наук



Е.А. Ивин, А.Н. Курбацкий, Д.В. Артамонов

МЕТОДИЧЕСКОЕ ПОСОБИЕ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

для социально-экономических специальностей

Вологда
2016

УДК 330.43
ББК 65в631
И25

Ивин, Е. А.

И25 Методическое пособие по математической статистике: для социально-экономических специальностей [Текст] / Е. А. Ивин, А. Н. Курбацкий, Д. В. Артамонов. – Вологда: ИСЭРТ РАН, 2016. – 141 с.

ISBN 978-5-93299-331-6

Настоящее пособие предназначено для самостоятельного знакомства и освоения основных методик эконометрических исследований в области анализа экономических и социально-экономических данных. Это пособие написано на основе курса, прочтенного авторами для аспирантов и сотрудников ИСЭРТ РАН, а также обсуждения изложенного материала слушателями. Каждый раздел снабжен кратким сводом основных понятий и теорем, необходимых для решения задач, а также решениями типовых задач по ключевым темам.

Книга адресована студентам и аспирантам социально-экономических специальностей, а также преподавателям.

УДК 330.43
ББК 65в631

Рецензент:

В.А. Ильин

доктор экономических наук, профессор, научный руководитель
Института социально-экономического развития территорий
Российской академии наук, заслуженный деятель науки РФ

ISBN 978-5-93299-331-6

© Ивин Е.А., Курбацкий А.Н., Артамонов Д.В., 2016
© ИСЭРТ РАН, 2016

Оглавление

Введение	6
1 Типы данных	7
1.1 Данные	7
2 Описательная статистика	9
2.1 Графическое представление данных	9
2.1.1 Гистограмма	10
2.1.2 Эмпирическая функция распределения	15
2.2 Характеристики среднего	17
2.2.1 Мода	17
2.2.2 Медиана	18
2.2.3 Среднее	18
2.3 Разброс и симметрия данных	19
2.3.1 Размах	19
2.3.2 Дисперсия и стандартное отклонение	21
2.3.3 Асимметрия и эксцесс	22
3 Оценки параметров	25
3.1 Оценивание параметров	25
3.1.1 Выборка	25
3.1.2 Примеры оценок	26
3.2 Свойства оценок	28
3.2.1 Три ключевых свойства	28
3.3 Методы оценивания	30
3.3.1 Метод моментов	30
3.3.2 Метод максимального правдоподобия	32
4 Доверительные интервалы	37
4.1 Точечные и интервальные оценки	37
4.2 Доверительный интервал для среднего	38
4.2.1 Случай известной дисперсии	38
4.2.2 Случай неизвестной дисперсии и объём выборки $n > 30$	40
4.2.3 Случай малой выборки и неизвестной дисперсии	40
4.2.4 Минимальный объём выборки	41
4.3 Доверительный интервал для доли и дисперсии	42

4.3.1	Доверительный интервал для доли	42
4.3.2	Доверительный интервал для дисперсии	43
5	Проверка гипотез	45
5.1	Понятие статистической гипотезы	45
5.1.1	Ошибки первого и второго родов	46
5.1.2	Статистика критерия	46
5.1.3	Критическая область	47
5.1.4	Минимальный уровень значимости	48
5.2	Проверка гипотезы о среднем	48
5.2.1	Случай известной дисперсии	48
5.2.2	Дисперсия неизвестна и объём выборки $n > 30$	50
5.2.3	Случай неизвестной дисперсии и маленькой выборки	51
5.3	Проверка гипотезы о доле и дисперсии	52
5.3.1	Гипотеза о доле	52
5.3.2	Гипотеза о дисперсии	54
6	Сравнение выборок	56
6.1	Равенство средних для независимых выборок	56
6.1.1	Дисперсии известны	56
6.1.2	Дисперсии неизвестны, но равны	57
6.1.3	Дисперсии неизвестны и не предполагаются равными	59
6.1.4	Доверительный интервал для разности средних	60
6.2	Равенство средних в случае зависимых выборок	61
6.2.1	Доверительный интервал для разности средних	63
6.3	Равенство долей и дисперсий	64
6.3.1	Проверка гипотезы о равенстве дисперсий	64
6.3.2	Проверка гипотезы о равенстве долей	66
6.3.3	Доверительный интервал для разности двух долей	67
7	Корреляция	69
7.1	Парный коэффициент корреляции	69
7.1.1	Коэффициент корреляции	69
7.1.2	Выборочный коэффициент корреляции	71
8	Упражнения для самопроверки	75
8.1	Описательная статистика и эмпирическая функция распределения	75
8.2	Свойства оценок	76
8.3	Доверительные интервалы	80
8.4	Проверка гипотез (одна выборка)	82
8.5	Сравнение выборок	85
8.6	Коэффициент корреляции	88

9	Дополнение. Непараметрические методы	91
9.1	Критерии случайности	92
9.1.1	Критерий серий	92
9.1.2	Критерий поворотных точек	93
9.1.3	Критерий Кендалла	94
9.2	Критерии согласия	95
9.2.1	Критерий Пирсона χ^2	96
9.2.2	Критерий Колмогорова–Смирнова	98
9.3	Критерии нормальности	100
9.3.1	Критерий Лиллиефорса	100
9.3.2	Критерий Андерсона–Дарлинга	102
9.3.3	Критерий Харке–Бэра	103
9.4	Критерии однородности	104
9.4.1	Критерий знаков	104
9.4.2	Критерий знаков для проверки гипотез о медиане	106
9.4.3	Критерий знаков для проверки гипотез о вероятности успеха	107
9.4.4	Критерий рангов	108
9.4.5	Критерий Манна–Уитни	110
9.4.6	Критерий Вилкоксона	112
9.5	Исследование взаимосвязей между выборками	114
9.5.1	Коэффициент ранговой корреляции Спирмена	114
9.5.2	Коэффициент Кендалла	117
9.6	Факторный анализ	119
9.6.1	Однофакторный анализ. Критерий Краскелла–Уоллиса	119
9.6.2	Двухфакторный анализ. Критерий Фридмана	121
A	Таблицы	124
B	Как работать с таблицами	133
B.1	Таблица нормального распределения	133
B.2	Распределение хи-квадрат	136
B.3	Таблица распределения Стьюдента	137
B.4	Распределение Фишера	138
B.5	Использование компьютера	139
	Литература	140

Введение

Настоящее пособие основано на материалах восстановительных лекций и практических занятий, которые авторы читали и вели для аспирантов и сотрудников ИСЭРТ РАН, подготавливая их, в первую очередь, к восприятию курса эконометрики.

Авторы стремились донести до читателей и слушателей тот непреложный факт, что методы математической статистики не ограничиваются инструментарием для эконометрики, а призваны и могут решать целый ряд самостоятельных задач. При этом в дополнении большое внимание уделено некоторым методам непараметрической статистики.

Целью настоящего издания является создание пособия для первоначальной работы с собираемыми данными. А именно предоставления возможности получить справку и руководство к действию в решении наиболее часто встречающихся проблем (задач) обработки данных.

Для дополнительного чтения мы советуем книги [1, 3, 5, 6, 9]. Для читателей без должной математической подготовки можно предложить книги [8, 4, 10] в качестве альтернативы для первоначального ознакомления с методом.

Благодарности

Авторы выражают благодарность руководству МШЭ МГУ и ИСЭРТ РАН за постановку задач по обучению будущих ученых и прекрасное обеспечение его реализации, а также сотрудникам и аспирантам ИСЭРТ РАН за возможность ознакомления и обсуждения круга решаемых задач и возникающих при их решении проблем.

Мы также благодарны аспирантам ИСЭРТ РАН (г.Вологда) за задаваемые ими вопросы и студентам Московской школы экономики.

Глава 1

Типы данных

1.1 Данные

Статистика имеет дело с данными – результатами наблюдений. Поэтому, прежде чем переходить к методам математической статистики, мы определимся, какие типы данных будем рассматривать. Приведём следующую классификацию [8].

Результаты опросов, измерений бывают двух типов: дискретные и непрерывные.

Определение. *Дискретные данные* представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счетным, т.е. может быть занумеровано натуральными числами от одного до бесконечности.

Определение. *Непрерывные данные* могут принимать любое значение в некотором интервале числовой прямой.

Этим типам данных в свою очередь соответствуют несколько шкал, которые зависят уже от природы исходных данных. Перечислим основные их виды:

- номинальная шкала;
- порядковая шкала;
- интервальная шкала;
- относительная шкала;
- дихотомическая шкала.

Определение. *Номинальная шкала* состоит из названий или категорий для сортировки или классификации объектов по некоторому признаку. Результаты измерений, полученные при помощи номинальной шкалы, не могут быть упорядочены, и с ними не могут производиться арифметические операции.

Примерами номинальной шкалы служат пол, семейное положение, профессия.

Определение. *Порядковая шкала* означает, что числа присваиваются объектам, чтобы обозначить относительные позиции объектов, но не величину различий между ними.

Результаты измерений, полученные при помощи порядковой шкалы, могут быть определенным образом упорядочены, однако не могут указать на величину разницы между ними. Например, итоговые места спортсменов в соревновании.

Определение. *Дихотомическая шкала* - номинальная шкала, которая состоит из двух категорий.

К дихотомической шкале применимы некоторые арифметические операции. Например, если среди 100 опрошенных людей 20 поддерживают некоего кандидата в президенты, то, разделив число сторонников на общее количество опрошенных 100, получим 0.2. Значение 0.2 есть доля сторонников данного кандидата в выборке.

Номинальные, порядковые, дихотомические шкалы являются дискретными. Рассмотрим два вида непрерывных шкал.

Определение. *Интервальная шкала* позволяет находить разницу между двумя величинами. Обладает всеми свойствами номинальной и порядковой, но она позволяет указать количественное значение измеряемого признака. Недостатком служит отсутствие абсолютного нуля в качестве точки отсчета.

Интервальная шкала состоит из интервалов одинаковой длины, называемых единицей измерения. Каждый единичный интервал может быть поделен на некоторое количество интервалов. Интервальные шкалы делимы. Шкала времени, например, может быть разделена на годы, каждый год разделен на дни, дни на часы и далее.

Непрерывные шкалы позволяют производить точные измерения значений признака, с этими значениями можно проводить арифметические операции: складывать, вычитать, умножать и делить.

Определение. *Относительная шкала* обладает абсолютным нулем в качестве точки отсчета, что позволяет ей иметь все свойства интервальной шкалы.

Для данных этой шкалы осмысленными являются все операции, включая вычитание и деление. Именно работе с этими данными будет посвящена основная часть данного пособия.

Глава 2

Описательная статистика

Описательная статистика занимается начальным анализом данных. В этой главе описываются методы предварительного анализа данных. Они включают построение гистограмм, которые являются приближениями к графикам плотностей распределения заданных выборок. Кроме того, мы обсуждаем методы построения приближённого среднего значения, анализ разброса выборки и определение наличия выбросов.

2.1 Графическое представление данных

Для начала мы познакомимся с базовыми способами представления данных. Графические изображения дают возможность сразу получить представление о поведении и распределении данных.

Прежде чем переходить к графическому представлению данных, их надо упорядочить по возрастанию и разбить на группы. Полученный упорядоченный набор называется **вариационным рядом**, как правило, именно с ним мы будем работать. Группы, на которые разбивается множество значений, будем называть **интервалами группировки**.

Пусть мы упорядочили наши n наблюдений x_1, \dots, x_n . Они лежат в некотором интервале, который мы разбиваем еще на m интервалов. Последние и называются интервалами группировки. Их длины обозначим через $\Delta_1, \dots, \Delta_m$, а середины интервалов группировки - через c_1, \dots, c_m .

Сразу отметим, что очень часто выборочные значения не могут быть записаны в числовой форме. Например, трудно численно измерить вкусы напитков, но вы можете сравнивать их между собой и говорить, какой вам нравится больше. В случае, когда у нас нет единицы измерения, мы просто упорядочиваем наблюдения. Такая процедура называется ранжированием, а номер, который получили наблюдения, называются рангами.

Рангом наблюдения называется порядковый номер наблюдения в вариационном ряду. Если значения наблюдаемых величин повторяются, то каждому из этих значений (наблюдений), присваивается одинаковый ранг, равный среднему арифметическому номеров занимаемых мест.

Переход от самих наблюдений к последовательности их рангов называется *ранжированием*.

В основном мы будем иметь дело с численными выборками, которые упорядочиваются по возрастанию.

Базовыми графическими инструментами представления данных являются гистограммы, полигоны и кумуляты (накопительные гистограммы). Рассмотрим их по порядку.

2.1.1 Гистограмма

Графическое изображение числа наблюдений n_i выборки, соответствующих каждому интервалу, называется **гистограммой** выборки. По горизонтальной оси откладываются значения наблюдаемой величины, по вертикальной — частота их появления.

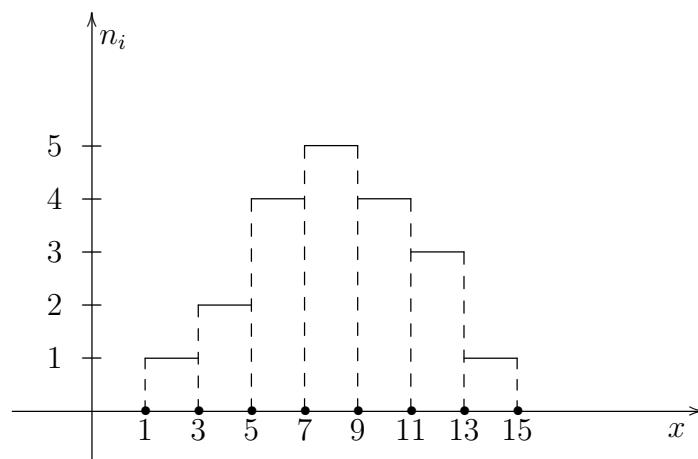
Пример. Дана выборка 3.3, 3.8, 5.1, 2.4, 6.5, 6.9, 5.7, 9.2, 9.7, 10.1, 12.8, 8.5, 8.6, 10.4, 11.2, 14.8, 7.2, 7.8, 8.1, 12.1. Построим гистограмму.

Решение. Чтобы это сделать, сперва упорядочим выборку, выписав вариационный ряд 2.4, 3.3, 3.8, 5.1, 5.7, 6.5, 6.9, 7.2, 7.8, 8.1, 8.5, 8.6, 9.2, 9.7, 10.1, 10.4, 11.2, 12.1, 12.8, 14.8. Объем выборки $n = 20$.

Все наблюдения попадают в интервал от 1 до 15. Построим гистограмму с шагом 2.

1. В интервал (1;3) попадает одно наблюдение 2.4, поэтому $n_1 = 1$.
2. В интервал (3;5) попадает два наблюдения 3.3, 3.8, поэтому $n_2 = 2$.
3. В интервал (5;7) попадает четыре наблюдения 5.1, 5.7, 6.5, 6.9, поэтому $n_3 = 4$.
4. В интервал (7;9) попадает пять наблюдений 7.2, 7.8, 8.1, 8.5, 8.6, поэтому $n_4 = 5$.
5. В интервал (9;11) попадает четыре наблюдения 9.2, 9.7, 10.1, 10.4, поэтому $n_5 = 4$.
6. В интервал (11;13) попадает три наблюдения 11.2, 12.1, 12.8, поэтому $n_6 = 3$.
7. В интервал (13;15) попадает одно наблюдение 14.8, поэтому $n_7 = 1$.

Изобразим это на графике:



Графическое изображение зависимости частоты $h_i = \frac{n_i}{n}$ попадания элементов выборки от соответствующего интервала называется гистограммой частот выборки.

Гистограмма частот есть графическое представление, в котором по горизонтальной оси откладываются значения переменной, а по вертикальной оси - соответствующие им частоты. Гистограмма строится в виде прямоугольников, высота которых соответствует частоте наблюдений в интервале группировки.

Пример. Дана выборка 3.3, 3.8, 5.1, 2.4, 6.5, 6.9, 5.7, 9.2, 9.7, 10.1, 12.8, 8.5, 8.6, 10.4, 11.2, 14.8, 7.2, 7.8, 8.1, 12.1. Построим гистограмму частот.

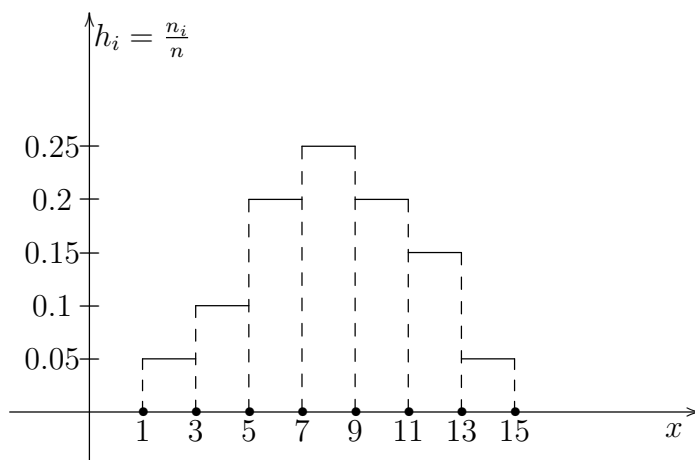
Решение. Чтобы это сделать, сперва упорядочим выборку, выписав вариационный ряд 2.4, 3.3, 3.8, 5.1, 5.7, 6.5, 6.9, 7.2, 7.8, 8.1, 8.5, 8.6, 9.2, 9.7, 10.1, 10.4, 11.2, 12.1, 12.8, 14.8. Объем выборки $n = 20$.

Все наблюдения попадают в интервал от 1 до 15. Построим гистограмму с шагом 2.

1. В интервал (1;3) попадает одно наблюдение 2.4, поэтому $h_1 = \frac{1}{20} = 0.05$.
2. В интервал (3;5) попадает два наблюдения 3.3, 3.8, поэтому $h_2 = \frac{2}{20} = 0.1$.
3. В интервал (5;7) попадает четыре наблюдения 5.1, 5.7, 6.5, 6.9, поэтому $h_3 = \frac{4}{20} = 0.2$.
4. В интервал (7;9) попадает пять наблюдений 7.2, 7.8, 8.1, 8.5, 8.6, поэтому $h_4 = \frac{5}{20} = 0.25$.
5. В интервал (9;11) попадает четыре наблюдения 9.2, 9.7, 10.1, 10.4, поэтому $h_5 = \frac{4}{20} = 0.2$.
6. В интервал (11;13) попадает три наблюдения 11.2, 12.1, 12.8, поэтому $h_6 = \frac{3}{20} = 0.15$.

7. В интервал (13;15) попадает одно наблюдение 14.8, поэтому $h_7 = \frac{1}{20} = 0.05$.

Изобразим это на графике:



Такую гистограмму ещё называют **гистограммой относительных частот**. Отличие гистограммы относительных частот от обычной гистограммы частот состоит в том, что на оси y вместо количества наблюдений на данном интервале отмечены их доли (или процент) от общего числа.

Чтобы каждый раз не думать о том, какой длины выбирать интервал группировки, можно пользоваться формулой Стерджеса $m \approx 1 + \log_2 n$. Длина каждого интервала будет равна $\Delta = \frac{x_{\max} - x_{\min}}{m}$.

В предыдущем примере число интервалов по формуле Стерджеса равно $m \approx 1 + \log_2 20 \approx 5.32$. Округлим до 5, тогда длина каждого интервала будет равна $\frac{15-1}{5} = 2.8$.

Можно избавиться от влияния размера интервала группировки, поделив частоты h_j на соответствующие длины Δ_j . В таком случае площадь фигуры под гистограммой становится равной единице и поэтому её можно назвать эмпирической функцией плотности.

Пример. Для той же выборки 2.4, 3.3, 3.8, 5.1, 5.7, 6.5, 6.9, 7.2, 7.8, 8.1, 8.5, 8.6, 9.2, 9.7, 10.1, 10.4, 11.2, 12.1, 12.8, 14.8, построим гистограмму относительных частот, делённую на длину Δ интервала группировки, то есть на 2.

Решение. 1. В интервал (1;3) попадает одно наблюдение 2.4, поэтому $\frac{h_1}{\Delta} = \frac{0.05}{2} = 0.025$.

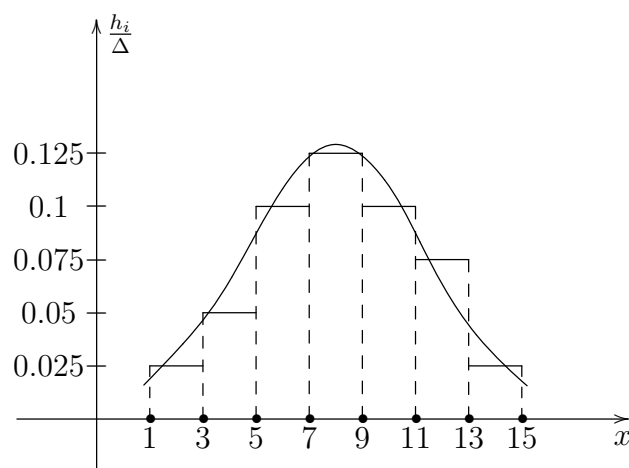
2. В интервал (3;5) попадает два наблюдения 3.3, 3.8, поэтому $\frac{h_2}{\Delta} = \frac{0.1}{2} = 0.05$.

3. В интервал (5;7) попадает четыре наблюдения 5.1, 5.7, 6.5, 6.9, поэтому $\frac{h_3}{\Delta} = \frac{0.2}{2} = 0.1$.

4. В интервал (7;9) попадает пять наблюдений 7.2, 7.8, 8.1, 8.5, 8.6, поэтому $\frac{h_4}{\Delta} = \frac{0.25}{2} = 0.125$.

5. В интервал (9;11) попадает четыре наблюдения 9.2, 9.7, 10.1, 10.4, поэтому $\frac{h_5}{\Delta} = \frac{0.2}{2} = 0.1$.
6. В интервал (11;13) попадает три наблюдения 11.2, 12.1, 12.8, поэтому $\frac{h_6}{\Delta} = \frac{0.15}{2} = 0.075$.
7. В интервал (13;15) попадает одно наблюдение 14.8, поэтому $\frac{h_7}{\Delta} = \frac{0.05}{2} = 0.025$.

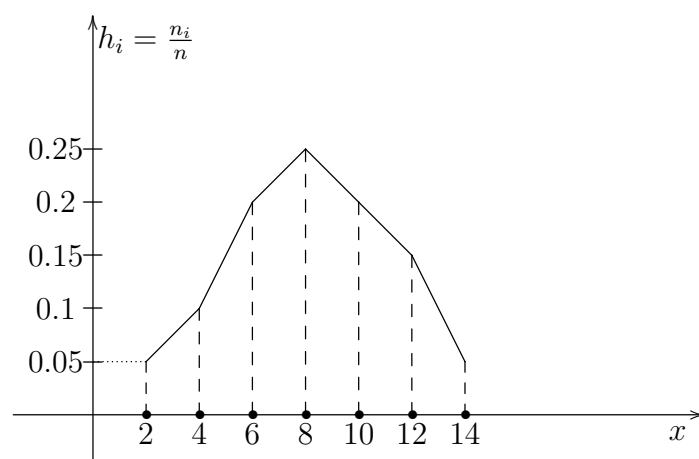
Изобразим это на графике вместе со сглаженной гистограммой, которую также часто рисуют, чтобы лучше представлять, какому непрерывному распределению приблизительно соответствует распределение относительных частот:



Несколько иное графическое представление данных дает *полигон*.

Полигон строится в виде области, ограниченной линией, которая проходит через точки $(c_i; h_i)$, где c_i - середина интервала, а h_i - частота.

Построим полигон для той же выборки, для которой мы построили гистограмму. Для этого надо отметить середины интервалов группировки c_i : $c_1 = 2$, $c_2 = 4$, $c_3 = 6$, $c_4 = 8$, $c_5 = 10$, $c_6 = 12$, $c_7 = 14$. И соединить точки $(c_i; h_i)$.



Оба представления данных – и гистограмма, и полигон – позволяют одним взглядом охватить все данные. Какое из этих представлений лучше - зависит от ваших предпочтений.

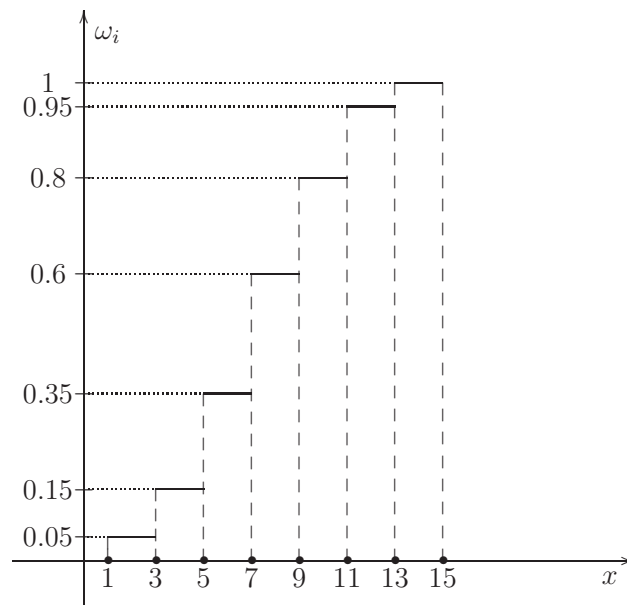
И ещё одно ключевое графическое изображение данных - это *кумулята* или *накопительная гистограмма*. Графическое изображение зависимости накопленных частот $\omega_i = \sum_{j=1}^i h_j$ называется **кумулятой** выборки.

Пример. Возьмём всё ту же выборку: 2.4, 3.3, 3.8, 5.1, 5.7, 6.5, 6.9, 7.2, 7.8, 8.1, 8.5, 8.6, 9.2, 9.7, 10.1, 10.4, 11.2, 12.1, 12.8, 14.8. Построим накопительную гистограмму.

Решение. Как и выше, строим гистограмму с шагом 2.

1. В интервал (1;3) попадает одно наблюдение 2.4, поэтому $h_1 = \frac{1}{20} = 0.05$, откуда $\omega_1 = h_1 = 0.05$.
2. В интервал (3;5) попадает два наблюдения 3.3, 3.8, поэтому $h_2 = \frac{2}{20} = 0.1$, откуда $\omega_2 = h_1 + h_2 = 0.15$.
3. В интервал (5;7) попадает четыре наблюдения 5.1, 5.7, 6.5, 6.9, поэтому $h_3 = \frac{4}{20} = 0.2$, откуда $\omega_3 = h_1 + h_2 + h_3 = 0.35$.
4. В интервал (7;9) попадает пять наблюдений 7.2, 7.8, 8.1, 8.5, 8.6, поэтому $h_4 = \frac{5}{20} = 0.25$, откуда $\omega_4 = h_1 + h_2 + h_3 + h_4 = 0.6$.
5. В интервал (9;11) попадает четыре наблюдения 9.2, 9.7, 10.1, 10.4, поэтому $h_5 = \frac{4}{20} = 0.2$, откуда $\omega_5 = \omega_4 + h_5 = 0.8$.
6. В интервал (11;13) попадает три наблюдения 11.2, 12.1, 12.8, поэтому $h_6 = \frac{3}{20} = 0.15$, откуда $\omega_6 = \omega_5 + h_6 = 0.95$.
7. В интервал (13;15) попадает одно наблюдение 14.8, поэтому $h_7 = \frac{1}{20} = 0.05$, откуда $\omega_7 = \omega_6 + h_7 = 1$.

Построим график:



2.1.2 Эмпирическая функция распределения

А теперь поговорим, пожалуй, о ключевом понятии для всего дальнейшего статистического анализа, об *эмпирической функции распределения*.

Эмпирической функцией распределения случайной величины, построенной по выборке x_1, \dots, x_n , называется функция $F_n(x)$, которая равна доле таких значений x_i , для которых $x_i \leq x$. То есть $F_n(x) = n_x/n$, где n_x - число наблюдений меньших или равных x , а n - объем выборки.

Эмпирическую функцию распределения $F_n(x)$ еще называют функцией распределения выборки, а функцию распределения $F(x)$ генеральной совокупности называют теоретической функцией распределения.

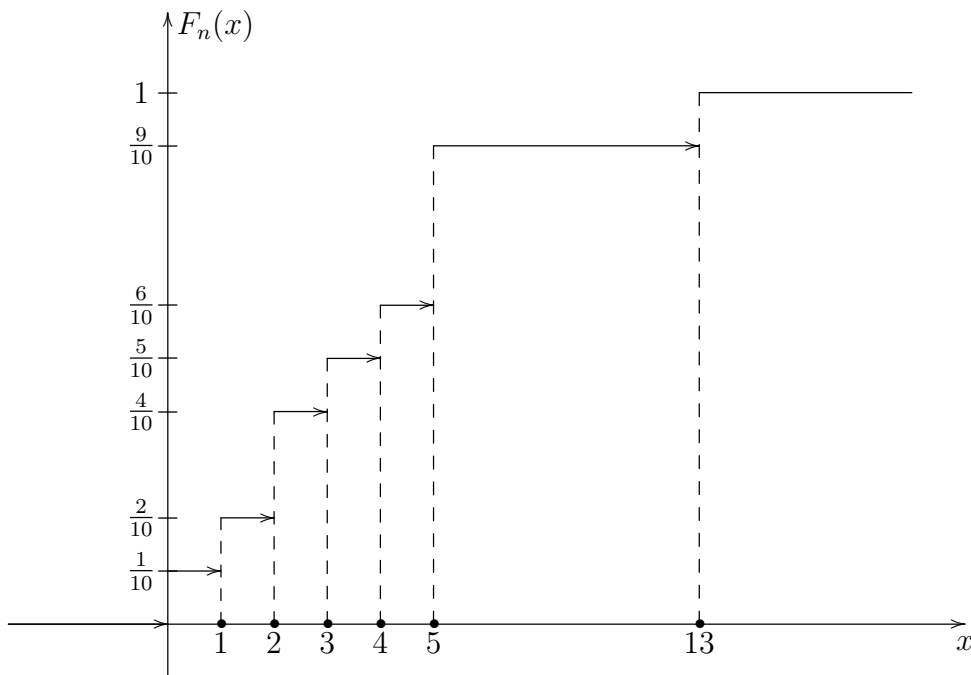
С ростом объема выборки эмпирическая функция распределения приближается к теоретической функции распределения, более точно $P(\lim_{n \rightarrow +\infty} \sup |F_n(x) - F(x)| = 0) = 1$. Этот замечательный факт доставляет нам теорема Гливенко-Кантелли.

Чем больше выборка, тем больше мы знаем о распределении исследуемого множества. Соответственно, выборочная функция приближается к теоретической.

Пример. Дана выборка: 0, 2, 13, 5, 5, 3, 1, 4, 5, 2. Найти эмпирическую функцию распределения.

Решение. Эмпирическая функция распределения имеет вид $F_n(x) = n_x/n$.

1. При $x < 0$ $F_n(x) = \frac{0}{10} = 0$, так как нет наблюдений меньших нуля.
2. При $x \in [0; 1)$ $F_n(x) = \frac{1}{10}$, так как одно наблюдение лежит левее 1.
3. При $x \in [1; 2)$ $F_n(x) = \frac{2}{10}$, так как два наблюдения лежат левее 2.
4. При $x \in [2; 3)$ $F_n(x) = \frac{4}{10}$, так как четыре наблюдения лежат левее 3.
5. При $x \in [3; 4)$ $F_n(x) = \frac{5}{10}$, так как пять наблюдений лежат левее 4.
6. При $x \in [4; 5)$ $F_n(x) = \frac{6}{10}$, так как шесть наблюдений лежат левее 5.
7. При $x \in [5; 13)$ $F_n(x) = \frac{9}{10}$, так как девять наблюдений лежат левее 13.
8. При $x \geq 13$ $F_n(x) = 1$, так как все наблюдения не превосходят 13.



Обратите внимание, что эмпирическая функция распределения похожа на кумуляту. Но при построении гистограмм мы вправе выбирать удобные для нас интервалы группировки наблюдений и, соответственно, менять вид диаграммы. А график эмпирической функции распределения однозначно определен.

Пример. Дана выборка: -1, 2, 2, 5, 4. Найти эмпирическую функцию распределения.

Решение. Эмпирическая функция распределения имеет вид:

1. при $x < -1$ $F_n(x) = \frac{0}{5} = 0$, так как нет наблюдений меньших -1.
2. При $x \in [-1; 2)$ $F_n(x) = \frac{1}{5}$, так как только одно значение лежит левее 2.
3. При $x \in [2; 4)$ $F_n(x) = \frac{3}{5}$, так как три наблюдения -1, 2, 2 лежат левее 4.
4. При $x \in [4; 5)$ $F_n(x) = \frac{4}{5}$, так как четыре наблюдения лежат левее 5.
5. При $x \geq 5$ $F_n(x) = 1$.

Перечислим свойства эмпирической функции распределения:

1. $0 \leq F_n(x) \leq 1$.
2. $F_n(x)$ - неубывающая функция.¹
3. $F_n(x)$ непрерывна слева.
4. $F_n(x) = 0$ при $x < x_{min}$ и $F_n(x) = 1$ при $x \geq x_{max}$.

Пример. Дана выборка из нормального распределения со средним значением 4 и дисперсией 1: 2, 4, 1.5, 5, 3, 6.5, 5.5, 2. Найти разность $F_n(x) - F(x)$ при $x = 5.1$, где $F_n(x)$ и $F(x)$ эмпирическая и теоретическая функции распределения.

Решение. Левее точки 5.1 лежит 6 наблюдений, поэтому $F_n = \frac{6}{8} = 0.75$. Найдём значение теоретической функции распределения по таблице. Для этого надо провести процедуру стандартизации $z = \frac{x-\mu}{\sigma} = \frac{5.1-4}{1} = 1.1$, $F(5.1) = \Phi(1.1) \approx 0.8643$.

Окончательно находим $F_n(5.1) - F(5.1) \approx 0.75 - 0.8643 = -0.1143$.

Пример. Дана выборка из нормального распределения (об этом распределении можно посмотреть [4]) со средним значением 5 и дисперсией 4: 0, 4, 1, 5, 10, 4, 6, 3, -1, 7. Найти разность $F_n(x) - F(x)$ при $x = 5.5$, где $F_n(x)$ и $F(x)$ эмпирическая и теоретическая функции распределения (с точностью до четвертого знака).

Решение. Левее точки 5.5 лежит 6 наблюдений, поэтому $F_n = \frac{7}{10} = 0.7$. Найдём значение теоретической функции распределения по таблице. Для этого надо провести процедуру стандартизации $z = \frac{x-\mu}{\sigma} = \frac{5.5-5}{2} = 0.25$, $F(5.5) = \Phi(0.25) \approx 0.5987$. Откуда $F_n(5.5) - F(5.5) = 0.7 - 0.5987 = 0.1013$.

Остановимся на графическом представлении данных и перейдем к следующей теме, в которой мы рассмотрим выборку не одним взглядом, а более детально.

2.2 Характеристики среднего

В этой главе мы хотим узнать, каким образом можно **оценить** срединное значение набора наблюдений. Для этого существует несколько числовых характеристик, каждая из которых дает свою особенную информацию о распределении данных.

2.2.1 Мода

Наша текущая задача состоит в выборе одного числа, которое можно было бы назвать центральным значением для набора данных. Первое из них имеет название *мода*.

Мода Mo – наиболее часто встречающееся значение в выборке.

В выборке 7, 2, 5, -12, -6, 13, 8, 2 двойка встречается чаще остальных наблюдений, поэтому $Mo = 2$.

Мода может быть не одна.

В выборке 1, 3, 4, -1, 2, 3, 5, 4 есть две моды 3 и 4. В таком случае распределение будет называться *бимодальным*.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения: 0, 2, 6, 5, 5, 3, 1, 4, 6, 2. Укажите количество мод данной выборки.

Решение. Значения 2, 5, 6 встречаются по два раза, поэтому в данной выборки 3 моды.

2.2.2 Медиана

Еще одна характеристика среднего - это **медиана** (оценка медианы)[4], которая определяется как значение, которое делит упорядоченную выборку пополам по количеству наблюдений. Медиана определяется по-разному для выборок с четным и нечетным числом наблюдений. Для нечетного числа наблюдений медиана есть просто центральное наблюдение $x_{(n+1)/2}$. Для четного числа наблюдений медиана - это среднее арифметическое двух соседних центральных наблюдений $x_{\frac{n}{2}}$ и $x_{\frac{n}{2}+1}$.

Рассмотрим выборку 1, 0, 3, 6, -1, 2, 7, 5, 4. Выпишем её вариационный ряд -1, 0, 1, 2, 3, 4, 5, 6, 7. Объем выборки равен 9, поэтому медиана - это просто центральный (пятый) элемент в выборке $Me = x_{(9+1)/2} = x_5 = 3$.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения: 0, 2, 6, 5, 5, 3, 1, 4, 6, 2. Вычислите медиану.

Решение. Объем выборки равен 10, поэтому надо взять среднее арифметическое двух соседних центральных элементов $Me = \frac{x_5+x_6}{2} = \frac{3+4}{2} = 3.5$.

2.2.3 Среднее

Наиболее распространённой характеристикой среднего безусловного математического ожидания [4] при работе с числовыми данными является **среднее арифметическое**.

Среднее значение выборки объема n вычисляется по формуле:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i.$$

Рассмотрим выборку 1, 0, 3, 6, -1, 2, 7, 5, 4. Медиану мы уже находили $Me = 3$, теперь найдем среднее \bar{x} :

$$\bar{x} = \frac{1}{9}(1 + 0 + 3 + 6 - 1 + 2 + 5 + 7 + 4) = 3.$$

Медиана и среднее значение бывают равны, но редко.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения: 0, 2, 6, 5, 5, 3, 1, 4, 6, 2. Вычислите среднее.

Решение. Среднее равно $\bar{x} = \frac{1}{10}(0 + 2 + 6 + 5 + 5 + 3 + 1 + 4 + 6 + 2) = 3.4$. А медиана, кстати, равна 3.5.

Пример. Рассмотрим выборку $0, 3, 9, -1, 2, 5, 2, 4$. Найдите моду, медиану и среднее значение.

Решение. Упорядочим выборку $-1, 0, 2, 2, 3, 4, 5, 9$. Наиболее часто встречающееся значение здесь 2, то есть мода $Mo = 2$.

Так как объем выборки равен 8, то медиана равна среднему арифметическому двух центральных элементов $Me = (2 + 3)/2 = 2.5$.

Среднее значение

$$\bar{x} = \frac{1}{8}(-1 + 0 + 2 + 2 + 3 + 4 + 5 + 9) = 3.$$

Среднее значение, как и медиана, сами по себе малоценны в качестве информации о выборке. Примером может служить средняя температура по больнице. Необходимы и характеристики разброса данных.

2.3 Разброс и симметрия данных

2.3.1 Размах

Простейшей мерой разброса является **размах** (range).

Размах - это разность между минимальным и максимальным значениями выборки, то есть $x_{\max} - x_{\min}$.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения: $-1, 2, 4, 6, 5, 7, 1, 4, 0, 2$. Чему равен размах?

Решение. Минимальный элемент равен -1 , а максимальный равен 7 . Значит, размах равен $7 - (-1) = 8$.

Чтобы ввести ещё одну меру разброса, нам потребуется определить понятие выборочной квантили.

Определение. Выборочной квантилью x_p называется решение уравнения $F_n(x) = p$, где $F_n(x)$ - это эмпирическая функция распределения. Смысл квантили состоит в том, что левее точки x_p лежит приблизительно $100p\%$ наблюдений.

Наиболее используемыми в описательной статистике являются:

- квантиль $x_{0.5}$, называемая медианой;
- квантиль $x_{0.25}$, называемая нижней квартилью;
- квантиль $x_{0.75}$, называемая верхней квартилью;
- квантили $x_{0.1}, x_{0.2}, x_{0.3}, x_{0.4}, x_{0.5}, x_{0.6}, x_{0.7}, x_{0.8}, x_{0.9}$, называемые децилями.

Кроме того, есть перцентили - это квантили $x_{0.01}, x_{0.02}, \dots, x_{0.99}$. А децили даже в жизни встречаются!

Пример. Отношение минимального дохода 10% самых богатых граждан и максимального дохода 10% самых бедных часто называют децильным коэффициентом, то есть $\frac{x_{0.9}}{x_{0.1}}$ - это показатель степени расслоения доходов.

Стоит отметить, что с определением квантилей возникают проблемы из-за того, что уравнение $F_n(x) = p$ не всегда разрешимо. Поэтому нам придётся договориться, как действовать в таких ситуациях.

С медианой мы уже разобрались в прошлом параграфе.

С квартилями при ручном счёте будем поступать следующим образом:

- сначала находится медиана, которая разбивает выборку на две равные подвыборки;
- для каждой из подвыборок ищем еще раз медиану. Медиану верхней подвыборки называем верхней квартилью, а медиану нижней подвыборки - нижней квартилью.

Замечание. Если выборка **нечётная**, то медиана включается в нижнюю и верхнюю подвыборки. Данными не разбрасываемся!

Пример. Пусть имеется выборка $-1, 3, 5, 6, 7, 8.5, 9$.

Определим медиану и квартили.

Решение. Медиана $Me = 6$, так как этот элемент разбивает выборку на две равные по объёму подвыборки. В каждой подвыборке находим медиану - это и будут квартили.

Для нижней подвыборки $-1, 3, 5, 6$ медиана равна 3, то есть $Q_{0.25} = 4$. Для верхней подвыборки $6, 7, 8.5, 9$ $Q_{0.75} = 7.75$

Вернёмся к показателям разброса данных и введём ещё одну меру вариации данных, называемую **межквартильным размахом**.

Межквартильный размах d - это разность между верхней и нижней квартилями, то есть $d = Q_{0.75} - Q_{0.25}$. Иногда используется обозначение IR (interquartile range).

В отличие от размаха, который полностью игнорирует распределение данных между минимальным и максимальным элементами, межквартильный размах показывает, где расположены 50% центральных данных. Крайние же значения выпадают из обзора.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения:

0, 2, 6, 5, 5, 3, 1, 4, 6, 2.

Вычислите межквартильный размах.

Решение. Выпишем вариационный ряд 0, 1, 2, 2, 3, 4, 5, 5, 6, 6. Медиана равна 3.5, она разбивает выборку на две подвыборки 0, 1, 2, 2, 3 и 4, 5, 5, 6, 6.

Для верхней подвыборки 4, 5, 5, 6, 6 медиана равна 5, поэтому $Q_{0.75} = 5$ Для нижней подвыборки 0, 1, 2, 2, 3 срединное значение равно 2, поэтому $Q_{0.25} = 2$

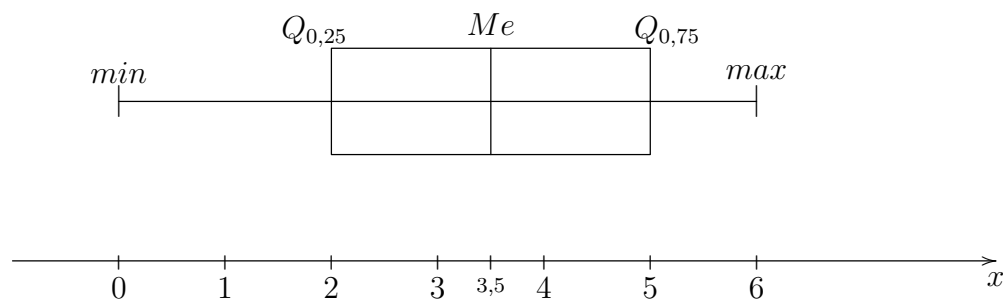
Откуда находим межквартильный размах $d = 5 - 2 = 3$.

С квантилями связаны два понятия: **коробчатая диаграмма** и **выбросы**.

Коробчатая диаграмма представляет собой необычный рисунок, так называемый "ящик с усами" (boxplot):

- отрезок прямой от минимального до максимального значения;
- коробочку, в которой заключены 50% наблюдений между нижней и верхней квартилью;
- в этой коробочке отмечена медиана;
- иногда особо выделяют выбросы.

Коробчатая диаграмма



Выбросы - это те наблюдения, которые меньше $Q_{0,25} - 1.5d$ или больше $Q_{0,75} + 1.5d$, то есть значения $x \notin [Q_{0,25} - 1.5d; Q_{0,75} + 1.5d]$.

Пример. В результате независимых наблюдений случайной величины были получены следующие ее значения: 0, 2, 6, 5, 5, 3, 1, 4, 6, 2. Определите выбросы.

Решение. Квантили и межквартильный размах уже были вычислены в предыдущей задаче. Нам остаётся посчитать $Q_{0,25} - 1.5d = 2 - 1.5 \cdot 3 = -2.5$ и $Q_{0,75} + 1.5d = 5 + 1.5 \cdot 3 = 9.5$, поэтому выбросов нет.

2.3.2 Дисперсия и стандартное отклонение

Когда речь идет о так называемых параметрических методах статистики, то на первый план среди различных мер разброса данных выходят выборочные *дисперсия* и *стандартное отклонение*.

Выборочная дисперсия вычисляется по формуле

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

а выборочное стандартное отклонение - это корень из дисперсии, то есть

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Пример. Для выборки: -1, 6, 0, 3, 6, 7, -2, 0, 8 найти среднее, дисперсию и стандартное отклонение.

Решение. Среднее равно $\bar{x} = \frac{1}{9}(-1 + 6 + 0 + 3 + 6 + 7 - 2 + 0 + 8) = 3$. Дисперсия

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9-1} ((-2-3)^2 + (-1-3)^2 + (0-3)^2 + (0-3)^2 + (3-3)^2 + (6-3)^2 + (6-3)^2 + (7-3)^2 + (8-3)^2) = \frac{118}{8} = \frac{59}{4}. \quad (2.1)$$

Стандартное отклонение соответственно равно $s = \sqrt{\frac{59}{4}} \approx 3,84$.

Стандартное отклонение хорошо тем, что измеряется в тех же единицах измерения, что и случайная величина. А дисперсия используется хотя бы потому, что сначала удобнее её посчитать, а потом уж извлекать корень.

2.3.3 Асимметрия и эксцесс

В заключении введем ещё два коэффициента, которыми часто руководствуются, чтобы делать вывод о принадлежности данных некоторому семейству распределения. Как правило, всех интересует, является ли распределение нормальным или нет.

Коэффициент асимметрии характеризует симметричность в распределении наблюдений и равен $As = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3}$.

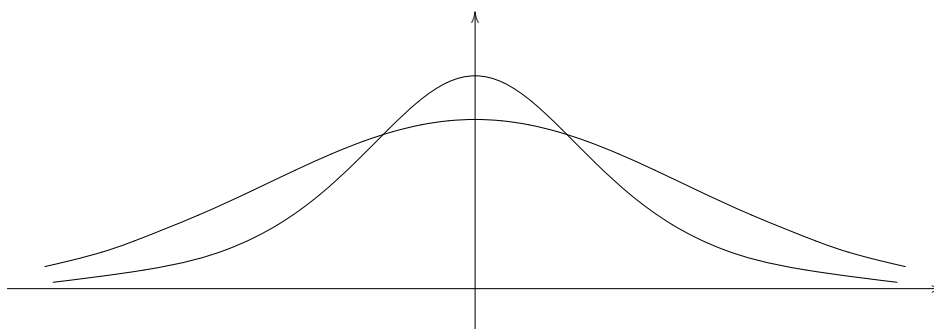
Коэффициент эксцесса характеризует вероятность появления больших (по модулю) значений и равен $Kurt = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4}$.

То есть это оценки для третьего и четвёртого центральных нормированных моментов.

Наличие симметрии характеризуется близостью коэффициента асимметрии к нулю.

Из курса теории вероятностей вам знакомы нормальное распределение и распределение Стюдента, которые являются симметричными. Примерами несимметричных распределений служат распределения Фишера и χ^2 .

Эксцесс характеризует частоту появления значений, которые удалены от среднего, то есть насколько много наблюдений находится в "хвостах" распределения.



А знаете, как проверить нормальность? Для нормального распределения коэффициент асимметрии равен нулю, а эксцесс - трем.

Если эксцесс сильно отличается от трёх, то говорят о наличии "тяжёлых хвостов".

Для закрепления всего пройденного в этой теме решите задачу.

Пример. Дана выборка: 0, 2, 13, 5, 5, 3, 1, 4, 5, 2. Построить вариационный ряд, определить ранги наблюдений, вычислить моду, медиану, квартили, межквартильный размах, размах, выбросы, среднее, дисперсию и среднеквадратичное отклонение, коэффициенты асимметрии и эксцесса.

Решение. 1. Вариационный ряд - это упорядоченная выборка, он имеет вид:

$$0, 1, 2, 2, 3, 4, 5, 5, 5, 13$$

2. Ранги наблюдений - это их порядковые номера в вариационном ряду

$$1, 2.5, 10, 8, 8, 5, 2, 6, 8, 2.5$$

3. Мода - это наиболее часто встречающееся значение $Mo = 5$.

4. Медиана - это среднее арифметическое двух центральных элементов для чётного числа наблюдений: $M_e = \frac{3+4}{2} = 3.5$

5. Нижняя квартиль - это точка, ниже которой лежит примерно 25% наблюдений. Для её вычисления надо определить медиану для меньшей подвыборки 0,1,2,2,3, на которые разбивает выборку медиана $Me: Q_{0.25} = 2$.

Верхняя квартиль - это точка, выше которой лежит примерно 75% наблюдений. Для того чтобы её найти, надо определить медиану для большей подвыборки 4,5,5,5,13: $Q_{0.75} = 5$.

6. Межквартильный размах $d = Q_{0.75} - Q_{0.25} = 5 - 2 = 3$.

7. Размах данных, то есть разность между максимальным и минимальным элементами, составляет $r = 13 - 0 = 13$

8. Среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{0+1+2+2+3+4+5+5+5+13}{10} = 4$.

9. Дисперсия

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{10-1} ((0-4)^2 + (1-4)^2 + (2-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + \\ &\quad + (5-4)^2 + (5-4)^2 + (13-4)^2) = \frac{118}{9} \approx 13.1. \quad (2.2) \end{aligned}$$

Среднеквадратическое отклонение

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{118}{9}} \approx 3.62.$$

10. Выбросы - это те наблюдения, которые меньше $Q_{0.25} - 1.5d$ или больше $Q_{0.75} + 1.5d$. В данном случае, $Q_{0.25} - 1.5d = 2 - 1.5 \cdot 3 = -2.5$ и $Q_{0.75} + 1.5d = 5 + 1.5 \cdot 3 = 9.5$, поэтому выбросы следующие: $x = 13$.

11. Коэффициент асимметрии равен $As = \frac{68.8}{40.53} \approx 1.7$.

12. Коэффициент эксцесса равен $Kurt = \frac{667.8}{139.24} \approx 4.8$.

Глава 3

Оценки параметров

Теперь перейдём к более точным способам анализа. А именно, к исследованию выборок с помощью статистических тестов.

Возможны две принципиально различные ситуации. В первом случае мы предполагаем, что нам известно, что выборка имеет распределение известной функциональной формы (например, известно, что выборка взята из какого-то нормального распределения), но в этом распределении есть неизвестные параметры. Тогда исследование сводится к оценке этих параметров и дальнейшему их анализу (построению доверительных интервалов, проверке гипотез). Этот случай называется параметрическим.

Во втором случае у нас нет информации о том, что выборка имеет распределение из какого-то класса. Этот случай называется непараметрическим.

3.1 Оценивание параметров

Не имея возможности исследовать всё множество объектов, нам приходится делать выборку, анализировать её и делать выводы о всей совокупности.

Как правило, выводы касаются некоторых характеристик генеральной совокупности (среднего, дисперсии, доли признака), поэтому по выборке мы ищем оценки для этих характеристик.

3.1.1 Выборка

Мы уже знаем, что в статистике есть два базовых понятия: **генеральная совокупность** и **выборка**. Интересующее нас множество объектов называется генеральной совокупностью (ГС), но так как оно, как правило, большое, то работать мы будем только с некоторой выборкой объектов из этой ГС.

Если изучается распределение роста взрослых мужчин в Вологде, то генеральная совокупность – это все возможные значения роста жителей Вологды. Если у нас нет информации о росте жителей, то можно взять и самим начать измерять рост людей в городе, пока не надоест. Данные о росте этих жителей называются выборкой из генеральной совокупности.

Выборка объёма n из заданной генеральной совокупности получается при случайном выборе n субъектов из генеральной совокупности и фиксации значений исследуемого признака. Короче, выборка - это набор x_1, \dots, x_n наблюдений значений изучаемой характеристики.

Случайная выборка x_1, \dots, x_n рассматривается как совокупность независимых одинаково распределённых случайных величин, имеющих то же распределение, что и генеральная совокупность. В частности, далее будет часто использоваться такой факт, что их математические ожидания и дисперсии равны $E(x_i) = \mu$, $Var(x_i) = \sigma^2$ при каждом i .

Статистика разрабатывает и изучает методы получения выводов о генеральной совокупности на основании выборок x_1, \dots, x_n .

Если мы хотим оценить теоретическое среднее значение $\mu = E(X)$ рассматриваемой генеральной совокупности на основании выборки x_1, \dots, x_n , то можно посчитать выборочное среднее

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Потом мы узнаем, почему эта оценка часто лучше остальных.

Оценка параметров генеральной совокупности на основании выборки - это ключевая задача математической статистики. По традиции параметр обозначается греческой буквой θ , а её оценка - $\hat{\theta}$.

Далее мы обсудим, каким образом выбирать оценки, чтобы они были информативными.

Мы сможем понять, в частности, почему выборочная дисперсия в первой теме вычислялась по формуле $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, а не по формуле $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. На первый взгляд, деление на n более логично.

3.1.2 Примеры оценок

Не всякое число имеет смысл рассматривать в качестве оценки некоторого параметра.

В качестве оценки среднего роста жителей Минска глупо брать величину меньше 100 сантиметров или более 250 сантиметров. Методы математической статистики позволяют не только отсекать заведомо нерациональные оценки, но и делать более качественные выводы. Например, какая из оценок среднего роста - 150 или 180 - является релевантной данной выборке.

Так как оценка параметра, который мы оцениваем, это, как правило, некоторая усреднённая величина, то указать идеальный способ её нахождения невозможно. Многое зависит от конкретной задачи.

В примере с ростом жителей в качестве оценки среднего роста можно взять несколько величин:

- самый популярный рост;

- величину роста, при которой у половины людей рост меньше этой величины, а у половины больше.
- среднее арифметическое, то есть сумму величин роста всех измеренных людей, делённую на их количество.

Мы уже знаем, что приведенные выше оценки – это, соответственно, мода, медиана и среднее.

Ещё одним из самых популярных параметров, которые пытаются оценить, является показатель разброса.

Помимо среднего роста жителей нас может заинтересовать, насколько сильно отличается рост остальных людей относительно этого среднего. В этом случае мы тоже можем оценивать разброс разными способами:

- посчитать среднее арифметическое квадратов отклонений от среднего $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$;
- посчитать среднее арифметическое модулей отклонений от среднего $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$.

Как правило, разброс относительно среднего считают так, как мы делали в первой теме, а именно, вычисляют дисперсию

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Пример. Пусть исследователь захотел оценить среднюю стоимость телефонов у студентов хорошего вуза и величину разброса относительно этой средней величины. Он опросил достаточно много студентов с разных курсов и факультетов и думает над несколькими способами оценки:

- посчитать среднее арифметическое \bar{x} цен телефонов, а в качестве меры разброса взять среднее арифметическое суммы отклонений от этой средней величины, то есть $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})$;
- определить самую часто встречаемую цену и сравнить с максимальной ценой (разбросом будет разность между ними);
- посчитать среднее арифметическое цен телефонов \bar{x} , а в качестве меры разброса взять среднее арифметическое суммы квадратов отклонений от этой средней величины, то есть $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Что вы ему посоветуете сделать?

Решение. Пункт 1) плох тем, что мера разброса может оказаться маленькой даже при больших отклонениях. Например, если взять двух студентов с телефонами за 3000 и 7000 рублей, то средняя цена телефона 5000 рублей, а сумма отклонений относительного среднего равна $(5000-3000)+(5000-7000)=0$.

Пункт 2) даёт ненадёжную оценку разброса, потому что она сильно зависит от максимальной цены. На неё может повлиять всего лишь один человек. Если у него цена оказалось намного больше, чем у остальных, то и разброс будет велик, даже при условии, что у остальных одинаковые цены телефонов.

А пункт 3) не имеет таких явных недостатков, посоветуем его.

3.2 Свойства оценок

3.2.1 Три ключевых свойства

Поскольку возможны различные оценки для одного и того же параметра, то надо уметь выбирать те, которые лучше. «Хорошие» оценки должны удовлетворять следующим свойствам. Отметим, что оценка является случайной величиной, значение которой рассчитывается по выборке, поэтому у неё можно находить математическое ожидание, дисперсию и другие моменты.

Определение. Величина $\hat{\theta}$ называется несмещённой оценкой параметра θ , если $E(\hat{\theta}) = \theta$. Для любой оценки $\hat{\theta}$ параметра θ разность $bias = E(\hat{\theta}) - \theta$ будем называть смещением.

Если говорить неформально, то отсутствие смещения означает, что метод не имеет систематической ошибки.

Примерами несмещённых оценок служат выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и выборочная дисперсия $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Докажите, что другая оценка дисперсии $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ является смещённой.

Поэтому мы и не считали по ней дисперсию.

Свойство несмещённости говорит нам, что в среднем наша оценка совпадает с параметром. Но таких оценок тоже много и далеко не всегда они будут хорошими. Следующие же свойство позволяет выбирать среди несмещённых наиболее подходящую.

Определение. Несмещённая оценка называется **эффективной** среди рассматриваемых оценок, если она имеет минимальную дисперсию.

Если требуется сравнить оценки, которые не обязательно являются несмещёнными, то вычисляют величину $MSE = E(\hat{\theta} - \theta)^2$. Эффективной в этом случае называют ту оценку, у которой эта величина минимальна.

Величина $E(\hat{\theta} - \theta)^2$ называется среднеквадратической ошибкой. Для несмещённых оценок она совпадает с дисперсией.

Пример. По выборке x_1, x_2, x_3, x_4, x_5 из нормального распределения $N(\theta, \sigma^2)$ построена следующая оценка параметра θ : $\hat{\theta} = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.3x_4 + 0.1x_5$.

- а) Выяснить является ли оценка $\hat{\theta}$ несмещенной;
- б) Найти дисперсию оценки $\hat{\theta}$;
- в) Является ли оценка $\hat{\theta}$ эффективной среди всех линейных оценок?

Решение. а) Оценка $\hat{\theta}$ является несмещенной, если $E(\hat{\theta}) = \theta$:

$$\begin{aligned} E(\hat{\theta}) &= E(0.1x_1 + 0.2x_2 + 0.3x_3 + 0.3x_4 + 0.1x_5) = \\ &= 0.1E(x_1) + 0.2E(x_2) + 0.3E(x_3) + 0.3E(x_4) + 0.1E(x_5) \\ &= 0.1\theta + 0.2\theta + 0.3\theta + 0.3\theta + 0.1\theta = \theta. \end{aligned} \quad (3.1)$$

Следовательно, оценка несмещённая.

б) Дисперсия оценки:

$$\begin{aligned} D(\hat{\theta}) &= E(0.1x_1 + 0.2x_2 + 0.3x_3 + 0.3x_4 + 0.1x_5) = \\ &= 0.1^2D(x_1) + 0.2^2D(x_2) + 0.3^2D(x_3) + 0.3^2D(x_4) + 0.1^2D(x_5) = \\ &= 0.01\sigma^2 + 0.04\sigma^2 + 0.09\sigma^2 + 0.09\sigma^2 + 0.01\sigma^2 = \\ &= 0.24\sigma^2. \end{aligned} \quad (3.2)$$

в) Оценка не является эффективной, так как среди линейных несмещённых оценок эффективной оценкой является среднее \bar{x} . Можно это проверить явно для данного случая, найдя её дисперсию: $D(\bar{x}) = D(\frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)) = \frac{1}{25}(\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{5}\sigma^2 = 0.2\sigma^2$.

Откуда получаем, что $0.24\sigma^2 = D(\hat{\theta}) > D(\bar{x}) = 0.2\sigma^2$.

Пример. По выборке x_1, x_2, x_3, x_4 из пуассоновского распределения с параметром θ построена следующая оценка параметра θ : $\hat{\theta} = 0.2x_1 + 0.3x_2 + 0.3x_3 + 0.2x_4$.

Вычислите смещение оценки $\hat{\theta}$.

Решение. Смещение оценки $\hat{\theta}$ вычисляется так $bias = E(\hat{\theta}) - \theta$:

$$\begin{aligned} E(\hat{\theta}) &= E(0.2x_1 + 0.3x_2 + 0.2x_3 + 0.3x_4) = \\ &= 0.2E(x_1) + 0.3E(x_2) + 0.2E(x_3) + 0.3E(x_4) = 0.2\theta + 0.3\theta + 0.2\theta + 0.3\theta = \theta, \end{aligned} \quad (3.3)$$

поэтому оценка несмещённая и смещение равно нулю.

Последнее свойство, которое мы рассмотрим, является, пожалуй, наиболее важным.

Определение. Оценка $\hat{\theta}$ называется **состоятельной** оценкой параметра θ , если $\hat{\theta} \xrightarrow{P} \theta$ при $n \rightarrow \infty$ (то есть оценка стремится к параметру по вероятности). Еще это записывается так: $\forall \varepsilon > 0 \lim_{n \rightarrow +\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$.

Состоятельность оценки означает, что при увеличении объема выборки значение оценки приближается к значению оцениваемого параметра генеральной совокупности.

Чем больше n , тем больше у нас информации о генеральной совокупности. Значит, и оценка должна быть ближе к неизвестному параметру. Если это не так, то оценка точно "несостоятельна".

Эмпирическая функция распределения приближается к теоретической функции распределения с ростом объема выборки, то есть является состоятельной оценкой.

Замечание. Помните теорему Гливенко-Кантелли?

Из неравенства Чебышёва следует:

Теорема. Если $\hat{\theta}_n$ – несмещённая оценка параметра θ и её дисперсия стремится к нулю при $n \rightarrow \infty$, то оценка состоятельна.

3.3 Методы оценивания

3.3.1 Метод моментов

Начнём с *метода моментов*. Метод моментов заключается в приравнивании некоторого числа выборочных моментов к соответствующим теоретическим, которые являются функциями неизвестных параметров $\theta_1, \theta_2, \dots, \theta_m$.

Из курса теории вероятности мы знаем, что k -ым моментом случайной переменной X является математическое ожидание случайной переменной, возведенное в k -ую степень. Таким образом, $E(X^k) = \mu_k$ – k -ый момент случайной переменной X .

k -ый момент можно оценить состоятельно, используя выборочный аналог (с объемом выборки, равным n):

$$\hat{\mu}_k = \frac{1}{n} \sum x_i^k$$

это и есть k -ый выборочный момент случайной переменной X .

Процедура оценивания методом моментов состоит в приравнивании m популяционных моментов к m выборочным моментам для оценивания m неизвестных параметров модели.

Систему, правда, не всегда удаётся легко решить. И моменты можно брать разные, но мы договоримся брать моменты по порядку.

В качестве простого примера пусть X – случайная переменная со средним значением μ^1 , и выражением дисперсии из уравнения:

$$\text{Var}(X) = \sigma^2 = E(X^2) - \mu^2.$$

¹Отметим, что в первом популяционном моменте μ_1 принято опускать подстрочный индекс и для обозначения среднего значения X просто применять μ .

Для оценивания методом моментов двух популяционных параметров μ и σ^2 мы должны приравнять два популяционных момента к двум выборочным моментам.

С помощью этих двух моментов мы можем получить соответствующие решения для параметров неизвестного среднего значения и неизвестной дисперсии. Чтобы получить оценку популяционного среднего значения, приравняем первый выборочный момент к первому популяционному моменту:

$$\hat{\mu} = \bar{x}$$

Затем применим выражение, заменяя второй популяционный момент его выборочным значением и заменяя первый момент:

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

В качестве оценки популяционного среднего значения метод моментов привел нас к выборочному среднему значению. Оценивание методом моментов дисперсии имеет в знаменателе выражения объем выборки n , а не обычное $n - 1$, и, таким образом, это не в точности выборочная дисперсия, которую мы используем. Но в больших выборках это не существенно.

В общем, оценки методом моментов состоятельны и в больших выборках сходятся к истинным значениям параметров, но нет никакой гарантии, что они являются “наилучшими” в каком-либо смысле.

Замечание. Метод был предложен английским статистиком Карлом Пирсоном в 1894 году.

Пример. Дана выборка из пуассоновского распределения с параметром λ : 1, 4, 3, 2, 3, 0, 1, 1, 0, 5. Методом моментов найдите оценку параметра λ .

Решение. Так как у нас только один параметр, то рассмотрим момент первого порядка, то есть математическое ожидание EX . Из курса теории вероятностей известно, что $EX = \lambda$, откуда мы и получаем уравнение на неизвестный параметр. Вместо теоретического момента EX подставим эмпирический момент - выборочное среднее \bar{x} - и получаем, что $\hat{\lambda} = \bar{x} = \frac{1}{10}(1 + 4 + 3 + 2 + 3 + 0 + 1 + 1 + 0 + 5) = 2$.

Пример. Дана выборка из равномерного распределения на отрезке $[a; 4]$: 0, 3, 2, 1.5, 1. Методом моментов найдите оценку параметра a .

Решение. Так как у нас один параметр, то рассмотрим момент первого порядка, то есть математическое ожидание EX . Так как для равномерного распределения $EX = \frac{a+4}{2}$, откуда мы и получаем уравнение на неизвестный параметр. Вместо теоретического момента EX подставим выборочное среднее \bar{x} и получим, что $\frac{a+4}{2} = \bar{x} = \frac{1}{5}(0 + 3 + 2 + 1.5 + 1) = 1.5$.

Откуда находим $a = -1$.

Пример. Дана выборка из нормального распределения $N(\mu, \sigma^2 = 1)$: 1, 3, -2, 2, 1. Методом моментов найдите оценку параметра μ .

Решение. Неизвестным является только один параметр, поэтому вычислим момент первого порядка, то есть математическое ожидание EX , который и равен параметру μ . Вместо теоретического момента EX подставляем эмпирический момент и получаем, что $\mu = \bar{x} = \frac{1}{5}(1 + 3 - 2 + 2 + 1) = 1$.

3.3.2 Метод максимального правдоподобия

Теперь перейдем, пожалуй, к самому популярному методу для нахождения оценок. При определенных условиях *метод максимального правдоподобия* является наилучшим.

Рассмотрим выборку x_1, x_2, \dots, x_n . Пусть плотность распределения генеральной совокупности $p(x, \theta)$ в точке x зависит от параметра θ . Рассмотрим совместную плотность выборки, которая равна произведению плотностей в силу независимости наблюдений: $L(\theta) = p(x_1, \theta) \cdot \dots \cdot p(x_n, \theta)$.

Метод максимального правдоподобия состоит в том, чтобы при конкретных значениях выборки x_1, \dots, x_n найти такое значение θ , при котором функция $L(\theta)$ принимает максимальное значение. Идея метода заключается в том, что мы максимизируем вероятности получения тех наблюдений, что нам даны. Ведь раз они есть, значит они наиболее вероятны.

Сама функция L называется *функцией правдоподобия*.

Замечание. Надо иметь в виду, что оценка параметра зависит от выборки, хотя при записи функции правдоподобия мы этого явно не указали.

Так как при разных значениях x_1, \dots, x_n могут получаться разные значения оценки $\hat{\theta}$, то она является случайной величиной (а не просто числом).

Как правило, при нахождении максимума функции правдоподобия L рассматривают не её саму, а её логарифм $\ln L$. Связано это с тем, что, логарифмируя функцию правдоподобия, произведение превращается в сумму и становится проще находить производную. Максимумы L и $\ln L$ достигаются при одном и том же значении параметра θ .

Пример. Дана выборка из пуассоновского распределения с параметром λ : 1, 4, 1, 1, 0, 5.

- а) Выпишите функцию правдоподобия и ее логарифм.
- б) Вычислите оценку максимального правдоподобия.

Решение. Такую задачу мы уже умеем решать методом моментов, теперь попробуем найти оценку параметра θ методом максимального правдоподобия.

- а) Распределение Пуассона задаётся формулой $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Выпишем функцию правдоподобия для нашей выборки 1, 4, 1, 1, 0, 5:

$$P(X = 1) = \frac{\lambda e^{-\lambda}}{1!}, P(X = 4) = \frac{\lambda^4 e^{-\lambda}}{4!}, P(X = 0) = \frac{\lambda^0 e^{-\lambda}}{0!}, P(X = 5) = \frac{\lambda^5 e^{-\lambda}}{5!}.$$

Функция правдоподобия имеет вид:

$$L(\lambda) = \left(\frac{\lambda e^{-\lambda}}{1!} \right)^3 \frac{\lambda^4 e^{-\lambda}}{4!} \frac{\lambda^0 e^{-\lambda}}{0!} \frac{\lambda^5 e^{-\lambda}}{5!} = \frac{\lambda^{12} e^{-6\lambda}}{4!5!}.$$

Теперь, чтобы удобнее было найти максимум функции, рассмотрим её натуральный логарифм:

$$\ln L = \ln \left(\frac{\lambda^{12} e^{-6\lambda}}{4!5!} \right)$$

И воспользуемся свойствами логарифма:

$$\ln L = 12 \ln \lambda - 6\lambda - \ln(4!5!).$$

- б) Чтобы найти максимум функции, вычислим производную по параметру λ и приравняем ее к нулю:

$$(\ln L)' = \frac{12}{\lambda} - 6 = 0 \Rightarrow \lambda = 2.$$

Теперь убедимся, что это действительно максимум, проверим достаточное условие:

$$(\ln L)'' = -\frac{12}{\lambda^2} < 0.$$

Ответ: 2.

Обратите внимание, что если сделать по методу моментов, то мы получим такой же ответ и намного быстрее.

Пример. Для случайной величины с распределением

X	0	1	2	4
P	$0,5 + \theta$	$0,1 - \theta$	$0,2$	$0,2$

получена выборка: 1, 4, 2, 2, 0, 1.

- Выпишите функцию правдоподобия и ее логарифм.
- Вычислите оценку максимального правдоподобия.

Решение. • а) Функция правдоподобия имеет вид:

$$\begin{aligned} L(\lambda) &= P(X = 1) \cdot P(X = 4) \cdot P(X = 2) \cdot P(X = 2) \cdot P(X = 0) \cdot P(X = 1) = \\ &= (0.1 - \theta)^2 (0.2)^3 (0.5 + \theta). \end{aligned} \quad (3.4)$$

Выпишем натуральный логарифм функции правдоподобия:

$$\ln L = \ln ((0.1 - \theta)^2 (0.2)^3 (0.5 + \theta))$$

И воспользуемся свойствами логарифма:

$$\ln L = 2 \ln(0.1 - \theta) + \ln(0.2)^3 + \ln(0.5 + \theta).$$

- **б)** Чтобы найти максимум функции, вычислим производную по параметру θ и приравняем ее к нулю:

$$(\ln L)' = -\frac{2}{0.1 - \theta} + \frac{1}{0.5 + \theta} = 0 \Rightarrow 1 + 2\theta = 0.1 - \theta \Rightarrow \theta = -0.3.$$

Остаётся убедиться, что это действительно максимум, для этого проверим достаточное условие:

$$(\ln L)'' = -\frac{2}{(0.1 - \theta)^2} - \frac{1}{(0.5 + \theta)^2} < 0.$$

Ответ: -0.3 .

Пример. Дана выборка из нормального распределения со средним θ и дисперсией 1: 0, 1, 2, 1. Выпишите функцию правдоподобия и вычислите оценку максимального правдоподобия для среднего.

Решение. Плотность нормального распределения в данном случае имеет вид:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}.$$

Поэтому функцию правдоподобия для нашей выборки равна

$$\begin{aligned} L &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(0-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(1-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(2-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(1-\theta)^2}{2}} = \\ &= \frac{1}{(2\pi)^2} e^{-\frac{(0-\theta)^2 + (1-\theta)^2 + (2-\theta)^2 + (1-\theta)^2}{2}} = \frac{1}{(2\pi)^2} e^{-2\theta^2 + 4\theta - 3}. \end{aligned} \quad (3.5)$$

Перейдем к логарифму:

$$\ln L = \ln \left(\frac{1}{(2\pi)^2} e^{-2\theta^2 + 4\theta - 3} \right).$$

По свойствам логарифма:

$$\ln L = \ln \left(\frac{1}{2\pi} \right)^2 - 2\theta^2 + 4\theta - 3.$$

Чтобы найти максимум функции, вычислим производную по параметру θ и приравняем ее к нулю:

$$(\ln L)' = -4\theta + 4 = 0 \Rightarrow \theta = 1.$$

Чтобы убедиться, что это действительно максимум, проверим достаточное условие: $(\ln L)'' = -4 < 0$.

Ответ: 1.

В заключение решим задачу с двумя неизвестными параметрами.

Задача. Дана выборка из нормального распределения со средним θ и дисперсией σ^2 : -2, 2, 3, 0, 2.

Вычислите оценку максимального правдоподобия для среднего и дисперсии.

Решение. Функцию правдоподобия для нашей выборки равна

$$L = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^5 e^{-\frac{(-2-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(2-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(3-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(0-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(2-\theta)^2}{2\sigma^2}} \quad (3.6)$$

То есть $L = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^5 e^{\frac{-5\theta^2+10\theta-21}{2\sigma^2}}$.

Перейдем к логарифму:

$$\ln L = \ln \left(\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^5 e^{\frac{-5\theta^2+10\theta-21}{2\sigma^2}} \right)$$

По свойствам логарифма:

$$\ln L = \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^5 + \frac{-5\theta^2+10\theta-21}{2\sigma^2} = -2.5 \ln(2\pi\sigma^2) + \frac{-5\theta^2+10\theta-21}{2\sigma^2}.$$

Чтобы найти максимум функции, вычислим частные производные по параметрам θ и σ^2 и приравняем их к нулю:

$$(\ln L)'_{\theta} = \frac{-10\theta+10}{2\sigma^2} = 0 \Rightarrow \theta = 1.$$

$$\begin{aligned} (\ln L)'_{\sigma^2} &= -\frac{5\pi}{2\pi\sigma^2} - \frac{-5\theta^2+10\theta-21}{2\sigma^4} = -\frac{5}{2\sigma^2} - \frac{-5\theta^2+10\theta-21}{2\sigma^4} = 0 \Rightarrow \\ &\Rightarrow -5\sigma^2+5\theta^2-10\theta+21=0 \Rightarrow \\ &\Rightarrow 5\sigma^2=5\theta^2-10\theta+21=16 \Rightarrow \sigma^2=3.2. \end{aligned} \quad (3.7)$$

Проверьте, что это действительно максимум функции!

Приведём в конце условия, при которых ММП-оценки точно обладают очень хорошими свойствами.

Условия регулярности

1. Область $D_n = \{x : p(x; \theta) > 0\}$ не зависит от θ ;
2. $\int_{-\infty}^{+\infty} p(x; \theta) dx = 1$ можно дважды дифференцировать под знаком интеграла, а $E(\hat{\theta}_n) = \int_{D_n} \hat{\theta}_n p(x; \theta) dx$ можно один раз дифференцировать под знаком интеграла.
3. Математическое ожидание $I(\theta) = E \left(\frac{\partial \ln p(\xi, \theta)}{\partial \theta} \right)^2$ конечно и положительно.

Теорема. Если для выборки объёма n выполнены условия регулярности, то

- решение $\hat{\theta}_n$ единственно;
- $\hat{\theta}_n$ – состоятельная оценка параметра θ ;
- $\hat{\theta}_n$ – асимптотически нормальна с математическим ожиданием θ и дисперсией $\frac{1}{nI(\theta)}$;
- ММП-оценка асимптотически эффективна.

Глава 4

Доверительные интервалы

4.1 Точечные и интервальные оценки

Основной задачей математической статистики является возможность получения выводов о исследуемой группе объектов, которую мы называем генеральной совокупностью. Выводы же, как правило, касаются некоторых параметров, например, среднего или дисперсии.

Сбережения населения представляют интерес многих исследователей и не только исследователей. С целью получения выводов, как правило, проводится выборочное обследование, потому что получить данные о всей генеральной совокупности не представляется возможным. По выборочным данным мы уже умеем считать среднее и стандартное отклонение, с помощью них мы теперь будем пытаться оценить теоретические параметры.

В этой теме мы научимся оценивать параметры не только одним числом, а сможем находить интервал, в который исследуемая характеристика попадает с заданной вероятностью.

В примере со сбережениями можно опросить знакомых и делать на этом основании некоторые выводы. Можно оценить средний доход населения по полученным данным. А откинув по 5% самых богатых и самых бедных, можно указать интервал, в который попадает доход 90% человек.

При статистическом исследовании требуется, чтобы выборка была *репрезентативна*, то есть была уменьшенной копией всей генеральной совокупности. Иными словами, выборка должна быть представительной, чтобы в ней были отражены все категории генеральной совокупности в соответствующих пропорциях. Везде далее мы предполагаем, что это условие выполнено.

Как мы уже обсуждали в первой теме, одним из способов делать такие выводы является оценивание параметров генеральной совокупности. Основные параметры, с которыми мы уже встретились раньше, - это *среднее*, *дисперсия*, *стандартное отклонение*, *доля*.

Например, допустим, что доля неженатых мужчин в стране составляет 40%, а их средняя заработная плата равна 5000 со стандартным отклонением 1000. Такую точную информацию получить обычно невозможно, поэтому все эти параметры мы будем оценивать по выборочным данным.

Точечной оценкой называется число, которое используют для оценки параметра ГС.

Пусть мы провели опрос с целью оценить долю избирателей некоего кандидата в президенты страны. Опросили 1000 человек, и 275 его поддерживают, значит, выборочная доля сторонников кандидата равна 27.5%. Это и есть точечная оценка для доли его сторонников во всей стране, которую мы заранее не знаем.

Но может так оказаться, что точечная оценка малоинформативна.

В примере с кандидатом в президенты намного лучше знать не просто точечную оценку 27.5%, а интервал, в котором с большой вероятностью находится доля сторонников кандидата. Скажем с 95% вероятностью доля сторонников лежит в пределах от 25% до 30%.

Рассмотрим ещё примеры, демонстрирующие малую эффективность точечных оценок.

Пример. Средняя температура по больнице ни о чем не говорит. Или средний доход на душу населения при большем расслоении не показателен, так как, к примеру, 5 человек могут получать по 500 тыс.рублей, а остальные 95 - по 10 тыс. рублей. В среднем будет по 34500, но никто такой зарплаты получать не будет.

Для того чтобы делать содержательные выводы, стараются находить не точечные, а интервальные оценки.

Определение. Доверительный интервал - это интервал, который с заданной вероятностью содержит оцениваемый параметр ГС.

Средняя температура по больнице равна 37 (точечная оценка), но с 95% вероятностью она лежит в пределах от 36.5 до 38.2 (интервальная оценка). Ясно, что интервальная оценка намного информативней.

Имейте в виду, что для разных выборок одной и той же ГС могут получаться разные доверительные интервалы!

При работе с доверительными интервалами часто используют два термина.

- Уровень значимости α - это вероятность, с которой значение параметра не попадает в доверительный интервал.
- Уровень доверия $\beta = 1 - \alpha$ - это вероятность того, что доверительный интервал содержит значение параметра.

Обычно уровень значимости равен 0.01, 0.05, 0.1, что соответствует уровню доверия 0.99, 0.95, 0.9. Очень часто уровни значимости и доверия измеряются в процентах, то есть уровень доверия 0.99 и 99% - это одно и то же.

4.2 Доверительный интервал для среднего

4.2.1 Случай известной дисперсии

Важнейшей характеристикой генеральной совокупности является среднее значение. Что же необходимо сделать, чтобы построить для него доверительный

интервал?

Здесь и далее предполагается, что генеральная совокупность имеет нормальный закон распределения.

В курсе теории вероятностей доказывается очень важная теорема, она даже называется центральной предельной теоремой (ЦПТ). Так вот по этой теореме среднее значение одинаково распределенных случайных величин стремится к нормальному распределению. Более того, верна следующая теорема.

Теорема. Если распределение генеральной совокупности имеет конечные математическое ожидание и дисперсию, то при $n \rightarrow \infty$ основные выборочные характеристики (среднее, дисперсия, эмпирическая функция распределения) являются нормальными.

Итак, рассмотрим случайную выборку объема n из генеральной совокупности, вычислим среднее значение \bar{x} по выборке и зададим уровень доверия β .

Доверительный интервал для среднего имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$, где Δ - это точность интервальной оценки.

Правило для вычисления точности зависит от того, что мы знаем о генеральной совокупности и с какой выборкой мы имеем дело.

Пусть нам известно стандартное отклонение σ генеральной совокупности.

Тогда $\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha$, где z_α - это квантиль нормального распределения уровня $1 - \frac{\alpha}{2}$ (то есть мы ищем это число в таблице нормального распределения).

Доверительный интервал для среднего с известной дисперсией имеет вид $(\bar{x} - \frac{\sigma}{\sqrt{n}} z_\alpha; \bar{x} + \frac{\sigma}{\sqrt{n}} z_\alpha)$.

Пример. Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$. Постройте 99% доверительный интервал.

Решение. • Среднее значение равно $\bar{x} = \frac{9+5+7+7+4+10}{6} = 7$.

- Итак, доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.995$ и определяем квантиль $z_\alpha = 2.58$. Теперь можем найти точность $\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha = \frac{1}{\sqrt{6}} 2.58 \approx 1.05$ (здесь мы воспользовались тем, что известна дисперсия генеральной совокупности).
- Искомый 99%-доверительный интервал имеет вид $(7 - 1.05; 7 + 1.05) = (5.95; 8.05)$.

Пример. Пусть для выборки объема $n = 25$ вычислено среднее $\bar{x} = 130$. Из предыдущих исследований известно стандартное отклонение $\sigma = 12$. Постройте 98% доверительный интервал для среднего значения. В ответе укажите точность интервальной оценки.

Решение. • Доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$. Уровень доверия равен $\beta = 0.98$, поэтому $\alpha = 0.02$. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.99$ и определяем квантиль $z_\alpha = 2.33$. Теперь можем найти точность $\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha = \frac{12}{\sqrt{25}} 2.33 \approx 5.59$.

- Искомый 98%-доверительный интервал имеет вид $(130 - 5.59; 130 + 5.59) = (124.41; 135.59)$.

4.2.2 Случай неизвестной дисперсии и объём выборки $n > 30$

Если выборка больше 30, но стандартное отклонение нам неизвестно, то вместо σ мы будем использовать выборочное стандартное отклонение

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Доверительный интервал для среднего при неизвестной дисперсии, но большой выборке ($n > 30$), имеет вид $\left(\bar{x} - \frac{s}{\sqrt{n}}z_{\alpha}; \bar{x} + \frac{s}{\sqrt{n}}z_{\alpha}\right)$.

Пример. Пусть объём выборки $n = 49$, выборочное среднее $\bar{x} = 4$, выборочная дисперсия $s^2 = 9$. Постройте 95% доверительный интервал.

Решение. • Среднее значение равно $\bar{x} = 4$, а выборочная дисперсия $s^2 = 9$.

- Так как в нашем случае неизвестна дисперсия генеральной совокупности, но $n \geq 30$, то доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$, где точность интервальной оценки $\Delta = \frac{s}{\sqrt{n}}z_{\alpha}$, z_{α} — это квантиль нормального распределения уровня $1 - \frac{\alpha}{2}$.
- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.975$ и определяем квантиль $z_{\alpha} = 1.96$. Теперь можем найти точность $\Delta = \frac{s}{\sqrt{n}}z_{\alpha} = \frac{3}{\sqrt{49}}1.96 \approx 0.84$.
- Искомый 95% доверительный интервал имеет вид $(4 - 0.84; 4 + 0.84) = (3.16; 4.84)$.

4.2.3 Случай малой выборки и неизвестной дисперсии

Теперь рассмотрим самый проблемный случай для любого исследователя, когда выборка маленькая и про её параметры ничего неизвестно. Если дисперсия неизвестна и объём выборки небольшой ($n \leq 30$), тогда вместо нормального распределения теперь используется t -распределение.

Доверительный интервал в этом случае имеет вид

$$\left(\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha}(n-1); \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha}(n-1)\right).$$

Здесь $t_{\alpha}(n-1)$ — это квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2}$ с $n-1$ степенью свободы (то есть мы ищем это число в таблице t -распределения).

Замечание. Распределение Стьюдента стремится к нормальному распределению при $n \rightarrow \infty$, поэтому при больших выборках доверительные интервалы для среднего, посчитанные по любой из наших формул, будут почти совпадать.

Число степеней свободы зависит от того, сколько имеется связей между наблюдениями. Так как мы знаем среднее, то наблюдения связаны одним равенством и степеней свободы становится на одну меньше. То, что других связей нет, надо доказывать, но их действительно нет.

Замечание. Распределение Стьюдента было введено в 1908 году В.С.Госсетом, ирландским служащим пивоваренного завода, который участвовал в разработке новых технологий производства пива и никаким студентом не был. Передать известности результаты исследований означало открыть корпоративную тайну, поэтому Госсет напечатал свои материалы под псевдонимом Стьюдент, откуда и t -распределение часто называют распределением Стьюдента.

Пример. Пусть объем выборки $n = 16$, выборочное среднее $\bar{x} = 5$, выборочная дисперсия $s^2 = 4$. Постройте 99% доверительный интервал.

Решение. • Среднее значение равно $\bar{x} = 5$, а выборочная дисперсия $s^2 = 4$.

- Так как неизвестна дисперсия генеральной совокупности и $n < 30$, поэтому точность интервальной оценки $\Delta = \frac{s}{\sqrt{n}} t_\alpha$.
- По таблице распределения Стьюдента находим $1 - \frac{\alpha}{2} = 0.995$ и, так как у нас $n - 1 = 16 - 1 = 15$ степеней свободы, определяем квантиль $t_\alpha = 3.29$. Теперь можем найти точность $\Delta = \frac{s}{\sqrt{n}} t_\alpha = \frac{2}{\sqrt{16}} 3.29 \approx 1.645$.
- Искомый 99% доверительный интервал имеет вид $(5 - 1.645; 5 + 1.645) = (3.355; 6.645)$.

4.2.4 Минимальный объем выборки

Благодаря тому, что мы знаем формулу для доверительного интервала, можно решить интересную задачу: найти минимальный необходимый объем выборки для того, чтобы с заданной точностью и уровнем доверия найти среднее значение.

Для того чтобы найти минимальный необходимый объем выборки для построения доверительного интервала для среднего значения с заданной точностью Δ и уровнем значимости α , достаточно применить формулу $n = \left(\frac{z_\alpha \sigma}{\Delta} \right)^2$.

Пример. Найдём минимально необходимый объем выборки для построения интервальной оценки среднего с точностью $\Delta = 3$, дисперсией $\sigma^2 = 225$ и уровнем доверия $\beta = 0.95$.

Решение. • Для доверительной вероятности $\beta = 0.95$ вычислим $\alpha = 1 - \beta = 0.05$. В таблице нормального распределения находим квантиль уровня $1 - \alpha/2 = 0.975$ $z_\alpha = 1.96$.

- Точность интервальной оценки $\Delta = 2$, поэтому теперь можем найти минимально необходимый объем выборки по формуле:

$$n = \left(\frac{z_\alpha \sigma}{\Delta} \right)^2 = \left(\frac{1.96 \cdot 15}{3} \right)^2 = 96.04 \approx 97$$

Теперь понятно, как найти минимально необходимый объем выборки при проведении собственных исследований!

4.3 Доверительный интервал для доли и дисперсии

4.3.1 Доверительный интервал для доли

Следующим популярным параметром, который часто требует оценивания, является доля признака p в ГС.

По выборке мы можем определить долю \hat{p} того или иного признака, просто посчитав число объектов m с этим признаком и поделив на объем выборки n , то есть $\hat{p} = \frac{m}{n}$. Долю объектов, не обладающих этим признаком, обозначают $\hat{q} = 1 - \hat{p}$.

Доверительный интервал для доли имеет вид $\left(\hat{p} - \sqrt{\frac{\hat{p}\hat{q}}{n}}z_\alpha; \hat{p} + \sqrt{\frac{\hat{p}\hat{q}}{n}}z_\alpha\right)$. Обратите внимание, что для использования этой формулы требуют выполнения условий $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$.

Пример. Объем выборки $n = 100$, выборочная доля $\hat{p} = 0.2$, $\alpha = 0.05$.

Решение. • Выборочная доля $\hat{p} = 0.2$, поэтому $\hat{q} = 1 - \hat{p} = 0.8$. Убеждаемся, что выполнены условия надёжности использования этих формул $n\hat{p} = 20 \geq 5$ и $n\hat{q} = 80 \geq 5$.

- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.975$ и определяем квантиль $z_\alpha = 1.96$. Теперь можем найти точность $\Delta = \sqrt{\frac{\hat{p}\hat{q}}{n}}z_\alpha = \sqrt{\frac{0.2 \cdot 0.8}{100}}1.96 \approx 0.078$.
- Искомый 95% доверительный интервал имеет вид $(0.2 - 0.078; 0.2 + 0.078) = (0.122; 0.278)$.

Если мы хотим узнать минимально необходимый объем выборки для того, чтобы с заданными точностью и уровнем доверия оценить долю признака в ГС, то сделать это можно по формуле $n = \hat{p} \cdot (1 - \hat{p}) \left(\frac{z_\alpha}{\Delta}\right)^2$.

Пример. Определить минимально необходимый объем выборки для оценивания доли с точностью $\Delta = 0.1$, если выборочная доля $\hat{p} = 0.4$, а уровень доверия $\beta = 0.99$.

Решение. • Для доверительной вероятности $\beta = 0.99$ вычислим $\alpha = 1 - \beta = 0.01$. В таблице нормального распределения находим квантиль уровня $1 - \alpha/2 = 0.995$ $z_\alpha = 2.58$.

- Точность интервальной оценки $\Delta = 0.1$, поэтому теперь можем найти минимально необходимый объем выборки по формуле:

$$n = \hat{p} \cdot (1 - \hat{p}) \left(\frac{z_\alpha}{\Delta}\right)^2 = 0.4 \cdot (1 - 0.4) \left(\frac{2.58}{0.1}\right)^2 \approx 159.75 \approx 160.$$

Пример. Какой минимальный объем выборки необходим для построения доверительного интервала для доли с точностью $\Delta = 0.02$, если выборочная доля $\hat{p} = 0.13$, а доверительная вероятность равна $\beta = 0.9$?

Решение. • Для доверительной вероятности $\beta = 0.9$ вычислим $\alpha = 1 - \beta = 0.1$. В таблице нормального распределения находим квантиль уровня $1 - \alpha/2 = 0.95$ $z_\alpha = 1.65$.

- Точность интервальной оценки $\Delta = 0.02$, поэтому теперь можем найти минимально необходимый объем выборки по формуле:

$$n = \hat{p} \cdot (1 - \hat{p}) \left(\frac{z_\alpha}{\Delta} \right)^2 = 0.13 \cdot (1 - 0.13) \left(\frac{1.65}{0.02} \right)^2 \approx 770.$$

Имейте в виду, что выборочная доля \hat{p} может быть неизвестна. В таких случаях её кладут равной 0.5, потому что при этом выражение $\hat{p} \cdot (1 - \hat{p}) \left(\frac{z_\alpha}{\Delta} \right)^2$ принимает наибольшее значение. При остальных значениях \hat{p} объем выборки был бы меньше, но если мы её не знаем, то берём крайний вариант.

4.3.2 Доверительный интервал для дисперсии

Теперь мы переходим к ещё одному важнейшему параметру генеральной совокупности - к дисперсии.

Которая, напомним, характеризует разброс случайной величины относительно ее среднего значения. Поэтому сейчас мы научимся строить для неё доверительный интервал.

Доверительный интервал для дисперсии имеет вид $\left(\frac{(n-1)s^2}{\chi_r^2(\beta)}; \frac{(n-1)s^2}{\chi_l^2(\beta)} \right)$.

Здесь значения $\chi_r^2(\alpha)$ и $\chi_l^2(\alpha)$ находятся по таблицам χ^2 -распределения (читается хи-квадрат) с $n - 1$ степенью свободы, причем в таблице мы ищем $\alpha/2$ и $1 - \frac{\alpha}{2}$.

Пример. Построить 90% доверительный интервал для дисперсии по выборке объема $n = 20$, если выборочная дисперсия $s^2 = 196$.

Решение. • Выборочная дисперсия $s^2 = 196$.

- По таблице χ^2 -распределения находим $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, число степеней свободы $n - 1 = 20 - 1 = 19$ и определяем критические точки $\chi_l^2 = 10.12$, $\chi_r^2 = 30.14$.

- Искомый 90% доверительный интервал имеет вид $\left(\frac{(20-1)196}{30.14}; \frac{(20-1)196}{10.16} \right) = (123.56; 366.54)$.

Часто для поиска дисперсии вручную удобнее использовать формулу $s^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)$.

Для доказательства формулы достаточно раскрыть скобки в исходном выражении для дисперсии.

Пример. По данным выборки объема $n = 12$ было найдено, что $\sum x_i = 216$, $\sum x_i^2 = 4046$. Постройте 90% доверительный интервал для теоретической дисперсии. В ответ укажите длину доверительного интервала, округлив до целого числа.

Решение. • Выборочное среднее равно $\bar{x} = \frac{216}{12} = 18$, а выборочная дисперсия $s^2 = \frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2) = \frac{1}{11}(4046 - 12 \cdot 18^2) \approx 14.36$.

- По таблице χ^2 -распределения находим $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, число степеней свободы $n - 1 = 12 - 1 = 11$ и определяем критические точки $\chi_l^2 = 4.57$, $\chi_r^2 = 19.675$.
- Искомый 90% доверительный интервал имеет вид $\left(\frac{(12-1)14.36}{19.675}; \frac{(12-1)14.36}{4.57} \right) = (8; 34.56)$. Длина интервала приблизительно равна 27.

Глава 5

Проверка гипотез

5.1 Понятие статистической гипотезы

Определение. *Статистическая гипотеза* - это некоторое предположение о свойствах и характеристиках исследуемых генеральных совокупностей.

Это предположение проверяется на основе анализа выборок. В этой теме мы будем иметь дело с параметрическими гипотезами, то есть с гипотезами о параметрах исследуемой ГС (о среднем, доле, дисперсии и т.п.). Непараметрические гипотезы будут рассмотрены позже.

Если однокурсница сказала вам, что средний вес студентки в университете равен 60 кг, то это статистическая гипотеза, которую вы можете проверить, опросив знакомых девушек. Если у опрошенных девушек вес окажется значительно выше 60, то, видимо, эта гипотеза неверна.

Но может оказаться, что вы просто общаетесь с крупными девушками! То есть ваша выборка не является показательной. Выборка должна представлять весь университет в уменьшенном масштабе или, как говорят социологи, быть **репрезентативной**. Везде далее мы будем предполагать выполнение этого условия.

Возникает несколько вопросов, с которыми нам предстоит разобраться. Во-первых, с какой величиной будем сравнивать гипотетический результат, то есть что вычислять по выборке? Во-вторых, когда начинается это "значительно выше", про которое было сказано в примере?

Определение. Основная или нулевая гипотеза H_0 - это гипотеза, которой мы придерживаемся, пока наблюдения не заставят признать обратное. Ей всегда сопутствует альтернативная гипотеза H_1 .

Вообще говоря, статистические методы не позволяют доказать гипотезу. По наблюдениям, которыми мы располагаем, мы можем гипотезу опровергнуть. И проблема состоит в том, что проверяем мы некоторое следствие, которое верно при выдвинутой гипотезе. Если следствие не соответствует имеющимся данным, то и гипотеза неверна. Но если данные согласуются со следствием, то это не означает справедливости гипотезы.

Пример. Допустим, вы хотите проверить гипотезу о том, что вы умный. Для проверки гипотезы вы взяли результаты своего Единого Государственного Экзамена по математике. Если у вас оказалось 25 баллов, то гипотезу, к сожалению, придётся отвергнуть. Если же баллы высокие, то гипотеза данными не опровергается, но при этом нельзя утверждать, что вы умный.

5.1.1 Ошибки первого и второго родов

Чтобы определиться, когда гипотезу отвергать, а когда не отвергать, введем еще два понятия.

Определение. • Ошибка первого рода - это ситуация, когда H_0 отвергается, хотя она, на самом деле, верна.

- Ошибка второго рода - это ситуация, когда H_0 принимается, хотя она неверна.

Пример. Мы совершаем ошибку первого рода, когда не берем съедобный гриб, думая, что он несъедобный.

Ошибка второго рода выглядит так. Например, суд выдвигает гипотезу H_0 : подсудимый невиновен. А он на самом деле виновен, но суд признает его невиновным за отсутствием улик (презумпция невиновности). То есть суд принимает гипотезу, хотя она неверна.

Определение. Буквой α обозначается уровень значимости или вероятность ошибки первого рода. Буквой β - вероятность ошибки второго рода.

Естественно, хочется сделать как можно меньше сразу обе ошибки, но это, к сожалению, невозможно. При уменьшении ошибки первого рода, увеличивается ошибка второго рода и наоборот. Обычно α берут 0.1, 0.05 или 0.01.

5.1.2 Статистика критерия

Для проверки гипотез используется функция, называемая **статистикой критерия**, которая зависит от выборки.

Определение. *Статистикой критерия* называется случайная величина, значение которой вычисляется по выборке.

Теперь под статистикой будем подразумевать некоторую функцию от выборки.

Для каждой задачи мы будем выбирать уровень значимости и статистику критерия, по значению которой будем делать вывод о справедливости гипотезы. При справедливости основной гипотезы будет известно, с какой вероятностью какое значение принимает статистика критерия. Если эта вероятность очень маленькая, то гипотезу придётся отвергнуть.

Определение. Мощностью критерия называется вероятность не совершить ошибку второго рода, то есть $1 - \beta$. А наиболее мощным критерием из всех критериев с уровнем значимости α называется тот, который обладает наибольшей мощностью.

Гипотеза либо отвергается, либо не отвергается. Старайтесь не употреблять слов "принимаем гипотезу", потому что невозможность отвергнуть гипотезу не означает, что она верна и ее стоит придерживаться. Может быть, просто недостаточно оснований или наблюдений, чтобы её отвергнуть. Но если очень хочется или удобно работать именно при таком предположении, то, конечно, вместо "не отвергаем гипотезу" можно сказать, что мы "принимаем гипотезу".

5.1.3 Критическая область

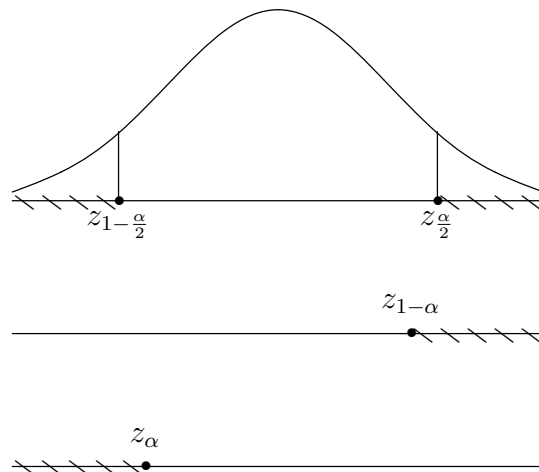
Теперь надо определиться, когда наступает момент, с которого мы будем отвергать гипотезу.

Определение. Критической областью называется область значений статистики критерия, при которых отвергается H_0 . *Акритические значения* - это граница критической области.

Существует три вида критических областей: левосторонняя, правосторонняя и двусторонняя.

Вид критической области определяется видом альтернативной гипотезы.

Виды критических областей





Если $H_1 : \theta \neq \theta_0$, то критическая область является двусторонней. При $H_1 : \theta > \theta_0$ критическая область является правосторонней, а при $H_1 : \theta < \theta_0$ - левосторонней.

Теперь мы готовы узнать, как проверить статистическую гипотезу. Процедура проверки гипотезы состоит из нескольких этапов:

- Сформулировать основную и альтернативную гипотезы и задать уровень значимости α .
- Найти критические значения и построить критическую область.
- Вычислить по выборке значение статистики и посмотреть, попало ли оно в критическую область.
- Сделать вывод. Если значение попало в критическую область, то основная гипотеза отвергается, в противном случае – не отвергается.

5.1.4 Минимальный уровень значимости

Теперь обсудим вопрос, который обычно возникает в случае, когда в задаче не дан уровень значимости. Какой уровень значимости всё-таки лучше 1%, 2%, 5% или 10%? А может  угой? Проверять каждый раз все трудоёмко. И вообще получается, что ответ зависит от того, какой уровень значимости взяли.

Допустим, мы не отвергли гипотезу при 5% уровне значимости. Но нам хочется знать, с какой вероятностью ошибки первого рода мы её можем отвергнуть. Ошибка в 6% может быть вполне допустимой, а ошибка в 25%  уж слишком много.

Нам нужна величина, которая позволит указать пороговое значение уровня значимости, с которого гипотеза отвергается. То есть по одному числу определить, с какой вероятностью гипотезу можно отвергать, а с какой нет.

Определение. Минимальный уровень значимости (p – value) – это уровень значимости, начиная с которого гипотеза отвергается.

5.2 Проверка гипотезы о среднем

5.2.1 Случай известной дисперсии

В этом параграфе будем учиться проверять гипотезу $H_0 : \mu = \mu_0$ (μ означает среднее ГС, а μ_0 – некоторое предполагаемое нами фиксированное значение). Мы уже знаем, что для проверки гипотезы нужно знать статистику и ее распределение. Оказывается, в зависимости от условий, статистика имеет разный вид.

Если стандартное отклонение ГС известно, то в этом случае статистика критерия $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ имеет нормальное распределение в предположении справедливости гипотезы H_0 . Убедитесь в этом, вспомнив центральную предельную теорему из курса теории вероятностей!

Для левосторонней области в таблице нормального распределения ищется квантиль уровня α , для правосторонней – квантиль уровня $1 - \alpha$, а для двухсторонней – квантили уровня $\frac{\alpha}{2}$ и $1 - \frac{\alpha}{2}$. При пользовании таблицами надо не забывать, что $z_\alpha = -z_{1-\alpha}$, в силу чётности функции плотности нормального распределения.

Пример. Пусть студенты университета в начале учебного года сдают предварительное тестирование, оцениваемое по десятибалльной шкале. Вы предполагаете, что средняя оценка равна 6, и решаете это проверить, опросив несколько человек. Получилась следующая выборка: 9, 5, 7, 7, 4, 10. Из наблюдений прошлых лет известно, что дисперсия $\sigma^2 = 1$. Проверим гипотезу, что среднее равно 6, на уровне значимости $\alpha = 0.01$ против односторонних альтернатив $\mu > 6$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu = 6; \quad H_1 : \mu > 6.$$

- Так как известна дисперсия генеральной совокупности, то для проверки гипотезы используется статистика $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
- Критическая область является правосторонней. По таблице нормального распределения находим $1 - \alpha = 0.99$ и определяем критическое значение $z_{cr} = 2.33$. Критическая область имеет вид $(2.33; +\infty)$.
- Вычислим значение статистики критерия. Среднее значение

$$\bar{x} = \frac{9 + 5 + 7 + 7 + 4 + 10}{6} = 7,$$

стандартное отклонение $\sigma = 1$.

Значение статистики критерия равно $z = \frac{7-6}{1/\sqrt{6}} \approx 2.45$.

- Вывод. Так как $z \in (2.33; +\infty)$, то основная гипотеза H_0 отвергается.
- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $1 - z^{-1}(2.45) \approx 0.007$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Пример. Оценим среднее время, которое тратят студенты в день на дорогу до университета и обратно. Пусть дисперсия генеральной совокупности известна и равна 144. Для выборки объема $n = 25$ посчитано среднее $\bar{x} = 130$ минут. На уровне значимости 0.01% проверьте гипотезу, что среднее равно 140. В ответ укажите значение статистики критерия с точностью до сотых.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu = 140; \quad H_1 : \mu < 140.$$

- Так как известна дисперсия генеральной совокупности, то для проверки гипотезы используется статистика $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
- Критическая область является левосторонней. По таблице нормального распределения находим $1 - \alpha = 1 - 0.0001 = 0.9999$ и определяем критическое значение $z_{cr} = -3.719$. Критическая область имеет вид $(-3.719; +\infty)$.

- Вычислим значение статистики критерия. Среднее значение $\bar{x} = 130$, стандартное отклонение $\sigma = 12$.

Значение статистики критерия равно $z = \frac{130-140}{12/\sqrt{25}} \approx -4.17$.

- Вывод. Так как $z \in (-\infty; -3.719)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

5.2.2 Дисперсия неизвестна и объём выборки $n > 30$

Если стандартное отклонение ГС неизвестно, но объём выборки большой ($n > 30$), то гипотеза проверяется с помощью той же статистики, только в ней теоретическое стандартное отклонение σ меняется на выборочное s , которое можно посчитать по формуле $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)}$.

Если стандартное отклонение ГС неизвестно, но $n > 30$, то в этом случае для проверки гипотезы о среднем используют статистику критерия $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, которая имеет нормальное распределение.

Стандартное отклонение σ надо заменить на выборочное стандартное отклонение s и решать так же, как и раньше.

Пример. Пусть $n = 49$, выборочное среднее равно 5, выборочная дисперсия – 4. Проверьте гипотезу, что среднее равно 6, на уровне значимости $\alpha = 0.01$ против двусторонних альтернатив $\mu \neq 6$.

Решение. • Сформулируем основную и альтернативную гипотезы: $H_0 : \mu = 6$; $H_1 : \mu \neq 6$.

- Определимся с критерием. Так как неизвестна дисперсия, но объём выборки $n > 30$, то для проверки гипотезы используется статистика $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.
- Критическая область является двусторонней. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.995$ и определяем критическое значение $z_{cr} = 2.58$. Критическая область имеет вид

$$(-\infty; -2.58) \cup (2.58; +\infty)$$

- Вычислим значение статистики критерия. Среднее значение $\bar{x} = 5$, выборочное стандартное отклонение $s = 2$. Значение статистики критерия равно $z = \frac{5-6}{2/\sqrt{49}} = -3.5$.
- Вывод. Так как $z \in (-\infty; -2.58) \cup (2.58; +\infty)$, то основная гипотеза H_0 отвергается.
- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $2 \cdot z^{-1}(-3.5) \approx 0.0004$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

5.2.3 Случай неизвестной дисперсии и маленькой выборки

Последний случай, который осталось рассмотреть, когда стандартное отклонение ГС неизвестно и объем выборки небольшой ($n \leq 30$). В этом случае гипотеза проверяется с помощью, так называемой, t -статистики.

Если стандартное отклонение ГС неизвестно и $n \leq 30$, то статистика критерия имеет вид $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

Эта статистика имеет распределение Стьюдента с $n - 1$ степенью свободы.

Пример. Выборка 1, 0, 3, 5, 4, основная гипотеза $\mu = 3$, $\alpha = 0.01$, односторонние альтернативы $\mu < 3$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu = 3; \quad H_1 : \mu < 3.$$

- Так как неизвестна дисперсия, а объем выборки $n < 30$, поэтому для проверки гипотезы используется статистика $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.
- Критическая область является левосторонней. По таблице распределения Стьюдента находим $1 - \alpha = 0.99$, число степеней свободы $n - 1 = 4$ и определяем критическое значение $t_{cr} = -4.6$. Критическая область имеет вид $(-\infty; -4.6)$.
- Вычислим значение статистики. Среднее значение $\bar{x} = \frac{1+0+3+5+4}{5} = 2.6$, выборочная дисперсия $s^2 = \frac{1}{5-1}((1-2.6)^2 + (0-2.6)^2 + (3-2.6)^2 + (5-2.6)^2 + (4-2.6)^2) \approx 4.28$, откуда выборочное стандартное отклонение $s \approx 2.07$. Значение статистики критерия равно $t = \frac{2.6-3}{2.07/\sqrt{5}} \approx 0.64$.
- Вывод. Так как $t \notin (-\infty; -4.6)$, то основная гипотеза H_0 не отвергается.
- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $1 - t^{-1}(0.64) \approx 0.51$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Пример. Преподаватель по информатике заскучал на контрольной. Чтобы немного отвлечься, он стал наблюдать рабочие столы студентов у себя на мониторе и решил оценить, сколько запросов в поисковой системе будет ими введено, если их не останавливать. Не имея возможности следить за всеми, он следил за 12 студентами. Он насчитал общее число обращений $\sum x_i = 216$ и вычислил $\sum x_i^2 = 4046$. На уровне значимости 5% проверьте гипотезу, что среднее число обращений равно 20 против двусторонних альтернатив.

Решение. • Сформулируем основную и альтернативную гипотезы: $H_0 : \mu = 20; \quad H_1 : \mu \neq 20$.

- Определимся с критерием. Так как неизвестна дисперсия и объем выборки $n \leq 30$, то для проверки гипотезы используется статистика $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

- Критическая область является двусторонней. По таблице распределения Стьюдента находим $1 - \frac{\alpha}{2} = 0.975$ и определяем критическое значение $t_{cr} = 2.2$, так как число степеней свободы равно $n - 1 = 11$. Критическая область имеет вид

$$(-\infty; -2.2) \cup (2.2; +\infty).$$

- Вычислим значение статистики критерия. Среднее значение $\bar{x} = \frac{216}{12} = 18$, выборочная дисперсия равна $s^2 = \frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2) = \frac{1}{11}(4046 - 12 \cdot 324) \approx 14.36$, а выборочное стандартное отклонение $s \approx 3.79$. Значение статистики критерия равно $t = \frac{18-20}{3.79/\sqrt{12}} \approx -1.83$.

- Вывод. Так как $t \notin (-\infty; -2.2) \cup (2.2; +\infty)$, то основная гипотеза H_0 не отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

5.3 Проверка гипотезы о доле и дисперсии

5.3.1 Гипотеза о доле

Теперь мы научимся отвечать на вопрос, какая доля объектов в генеральной совокупности обладает определенным признаком.

Например, вы сомневаетесь в том, что доля избирателей некоторого кандидата на предстоящих выборах равна 0.6 или 60%, и хотите проверить эту информацию. Для этого недостаточно просто опросить большое количество людей, надо ещё уметь определять, насколько сильно полученные результаты должны отличаться от заявленных 60%, чтобы иметь основания опровергать эту информацию.

Проверка гипотезы $H_0 : p = p_0$ о доле p признака в ГС проводится с помощью z -статистики.

Статистика критерия для проверки гипотезы о доле равна $z = \frac{m - np_0}{\sqrt{np_0q_0}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0q_0}{n}}}$, которая сходится по распределению к стандартной нормальной величине при $n \rightarrow \infty$.

Здесь n - это объем выборки, m - число объектов в выборке с данным признаком или число "успехов", p_0 - это предполагаемая доля признака в генеральной совокупности, $\hat{p} = \frac{m}{n}$ - это доля признака в выборке и $q_0 = 1 - p_0$.

Несмотря на то, что значения статистики мы будем искать в таблице нормального распределения, надо не забывать, что статистика имеет лишь асимптотически нормальное распределение. То есть использовать её стоит только при больших выборках и дополнительно проверять условия $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$.

Пример. Проверим гипотезу, что доля признака в ГС равна 0.1 на уровне значимости $\alpha = 0.05$, против односторонних альтернатив $p > 0.1$. Объем выборки $n = 100$ и пусть выборочная доля составила $\hat{p} = 0.2$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : p = 0.1; \quad H_1 : p > 0.1$$

- Условия надёжности использования этих формул $n\hat{p} = 100 \cdot 0.2 = 20 \geq 5$ и $n\hat{q} = 100 \cdot 0.8 = 80 \geq 5$ выполнены.
- Критическая область является правосторонней. По таблице нормального распределения находим $1 - \alpha = 0.95$ и определяем критическое значение $z_{cr} = 1.65$. Критическая область имеет вид $(1.65; +\infty)$.
- Значение статистики критерия равно $z = \frac{\hat{p}-p_0}{\sqrt{p_0q_0/n}} = \frac{0.2-0.1}{\sqrt{0.1 \cdot 0.9/100}} \approx 3.3$.
- Вывод. Так как $z \in (1.65; +\infty)$, то основная гипотеза H_0 отвергается.
- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $1 - z^{-1}(3.3) \approx 0.0005$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Пример. Допустим, вы думаете, что 25% студентов вашего университета ни разу не пропустили ни одной лекции. В выборочном опросе из 75 случайных студентов таких оказалось 15 человек. Проверьте свою гипотезу на уровне значимости $\alpha = 0.02$ против двусторонних альтернатив. В ответ запишите минимальный уровень значимости с точностью до четвертого знака.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : p = 0.25; \quad H_1 : p \neq 0.25$$

- Выборочная доля равна $\hat{p} = \frac{15}{75} = 0.2$. Условия надёжности использования z -статистики выполнены, так как $n\hat{p} = 75 \cdot 0.2 = 15 \geq 5$ и $n\hat{q} = 75 \cdot 0.8 = 60 \geq 5$.
- Критическая область является двусторонней. По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.99$ и определяем критические значения $z_{cr} = \pm 2.33$. Критическая область имеет вид $(-\infty; -2.33) \cup (2.33; +\infty)$.
- Значение статистики критерия равно $z = \frac{\hat{p}-p_0}{\sqrt{p_0q_0/n}} = \frac{0.2-0.25}{\sqrt{0.25 \cdot 0.75/75}} = -1$.
- Вывод. Так как $z \notin (-\infty; -2.33) \cup (2.33; +\infty)$, то основная гипотеза H_0 отвергается.
- Минимальный уровень значимости равен $2(1 - z^{-1}(1)) \approx 0.3174$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

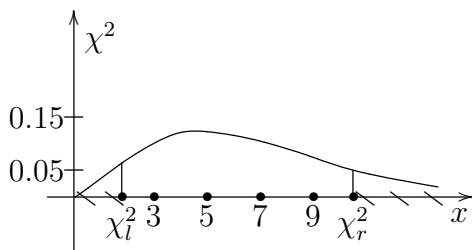
5.3.2 Гипотеза о дисперсии

Наконец, перейдем к проверке гипотезы о равенстве дисперсии σ^2 некоторому значению σ_0^2 . Это необходимо делать, когда приходится пользоваться предположениями о дисперсии.

Когда проверяли гипотезу о среднем, откуда-то предполагали, что известна дисперсия. Теперь хоть можем это проверить!

Чтобы проверить гипотезу $H_0 : \sigma^2 = \sigma_0^2$ надо рассмотреть статистику $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, имеющую χ^2 -распределения с $n - 1$ степенью свободы.

Обратите внимание, что в таблице в случае левосторонней критической области по значению $1 - \alpha$ находят квантиль, обозначаемую χ_r^2 , в случае правосторонней по значению α - χ_l^2 , в случае двусторонней критической области в таблице по значениям $1 - \frac{\alpha}{2}$ и $\frac{\alpha}{2}$ - χ_r^2 и χ_l^2 соответственно.



Пример. Допустим, мы предполагаем, что стандартное отклонение в стобалльном рейтинге студентов равно 15. И решаем проверить это, оценив рейтинг знакомых. Получилась выборка объема $n = 20$, у которой выборочная дисперсия равна 196. Наша задача эквивалентна проверке гипотезы о равенстве дисперсии 225. Уровень значимости возьмём $\alpha = 0.1$, а альтернативы рассмотрим двусторонние.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \sigma^2 = 225; \quad H_1 : \sigma^2 \neq 225.$$

- Критическая область является двусторонней. По таблице χ^2 -распределения находим $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, число степеней свободы $n - 1 = 20 - 1 = 19$ и определяем критические точки $\chi_l^2 = 10.12$, $\chi_r^2 = 30.14$. Критическая область имеет вид $(0; 10.12) \cup (30.14; +\infty)$.
- Значение статистики критерия равно $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(20-1) \cdot 196}{225} \approx 16.55$.
- Вывод. Так как $\chi^2 \notin (0; 10.12) \cup (30.14; +\infty)$, то основная гипотеза H_0 не отвергается.

- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $2 \cdot (1 - (\chi^2)^{-1}(16.55)) \approx 0.76$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Пример. Пусть объем выборки $n = 25$, её выборочная дисперсия равна 0.015. Проверьте гипотезу о равенстве дисперсии 0.01 на уровне значимости $\alpha = 0.05$ против односторонних альтернатив. В ответ введите разность между значением статистики и критическим значением, округлив до целого числа.

Решение. • В условии не сказано, какой именно должна быть альтернативная гипотеза, но так как выборочное значение превышает гипотетическое, то логично проверять гипотезу против правосторонних альтернатив:

$$H_0 : \sigma^2 = 0.01; H_1 : \sigma^2 > 0.01.$$

- Критическая область является правосторонней. По таблице χ^2 -распределения находим $1 - \alpha = 0.95$, число степеней свободы $n - 1 = 25 - 1 = 24$ и определяем критическую точку $\chi_r^2 = 36.415$. Критическая область имеет вид $(36.415; +\infty)$.
- Значение статистики критерия равно $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1) \cdot 0.015}{0.01} = 36$.
- Вывод. Так как $\chi^2 \notin (36.415; +\infty)$, то основная гипотеза H_0 не отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Глава 6

Сравнение выборок

6.1 Равенство средних для независимых выборок

В предыдущей теме мы занимались одной генеральной совокупностью и делали выводы о её параметрах. Но часто исследователю приходится сравнивать две выборки.

Пример. Как сравнить, у кого средний доход на душу населения больше, у жителей Одессы или Ростова-на-Дону?

В этом параграфе мы будем иметь дело с независимыми выборками.

6.1.1 Дисперсии известны

В случае, когда дисперсии известны, для проверки гипотезы о равенстве разности средних некоторому значению применяется статистика:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

где

\bar{x}_1 и \bar{x}_2 - выборочные средние,

μ_1 и μ_2 - гипотетические генеральные средние,

n_1 и n_2 - объемы выборок,

σ_1^2 и σ_2^2 - известные генеральные дисперсии.

Статистика z имеет стандартное нормальное распределение.

Так как

$$D(\bar{x}) = D\left(\frac{1}{n} \sum x_i\right) = \frac{1}{n^2} \sum D(x_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}.$$

Поэтому в знаменателе стоит стандартное отклонение разности средних двух выборок

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = D(\bar{x}_1 - \bar{x}_2) = D(\bar{x}_1) + D(\bar{x}_2) = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Пример. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

С помощью критерия Стьюдента проверить гипотезу о равенстве средних значений этих выборок на 95% уровне доверия против односторонней альтернативы.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 > \mu_2.$$

- Для проверки гипотезы о равенстве средних при известных дисперсиях используется z -статистика

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

- Критическая область является правосторонней. По таблице z -распределения находим $1 - \alpha = 0.95$ и определяем критическую точку $z_{cr} = 1.64$. Критическая область имеет вид $(1.64; +\infty)$.

- Значение статистики критерия равно $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{25 - 21}{\sqrt{\frac{2}{9} + \frac{1}{6}}} \approx 6.45$.

- Вывод. Так как $z \in (1.64; +\infty)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Пример. В условиях предыдущей задачи найдите минимальный уровень значимости.

Решение. Значение статистики уже посчитано и равно 6.45. Поэтому минимальный уровень значимости составляет $(1 - z^{-1}(6.45)) \approx 0.000$.

6.1.2 Дисперсии неизвестны, но равны

Теперь мы рассмотрим ситуацию, когда дисперсии неизвестны, но предполагаются равными.

В случае, когда дисперсии неизвестны, но равны, для проверки гипотезы применяется статистика:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

где

\bar{x}_1 и \bar{x}_2 – выборочные средние,

μ_1 и μ_2 – гипотетические генеральные средние,

n_1 и n_2 – объемы выборок,

s_p^2 – объединённая оценка дисперсии.

Вычисляется объединённая оценка дисперсии по формуле:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2},$$

где s_1^2 и s_2^2 – выборочные дисперсии.

Статистика имеет t -распределение с числом степеней свободы $n_1 + n_2 - 2$.

Пример. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	выборочная дисперсия
X	10	15	2
Y	7	12	1

С помощью критерия Стьюдента проверить гипотезу о равенстве средних значений этих выборок (считая их дисперсии равными) при 95% уровне доверия против двусторонних альтернатив.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2.$$

- Для проверки гипотезы о равенстве средних при известных дисперсиях используется t -статистика $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.
- Критическая область является двусторонней. По таблице t -распределения находим $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$, число степеней свободы $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ и определяем критические точки $t_{cr} = \pm 2.13$. Критическая область имеет вид $(-\infty; -2.13) \cup (2.13; +\infty)$.
- Значение статистики критерия равно

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{15 - 12}{\sqrt{\frac{1.6}{10} + \frac{1.6}{7}}} \approx 4.81,$$

$$\text{так как } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{2 \cdot 9 + 1 \cdot 6}{15} = 1.6.$$

- Вывод. Так как $t \in (-\infty; -2.13) \cup (2.13; +\infty)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Естественно, возникает вопрос, почему можно предполагать, что дисперсии равны¹. Поэтому необходимо рассмотреть общий случай, когда ничего неизвестно.

¹Ведь лучше сразу предположить, что средние равны, и ничего не делать.

6.1.3 Дисперсии неизвестны и не предполагаются равными

В самом общем случае, когда дисперсии неизвестны и не равны, точный критерий для проверки гипотезы о равенстве средних указать трудно. В этом случае пользуются приближительными формулами.

Как и следовало ожидать, для проверки гипотезы применяется t -статистика, в которой вместо теоретических значений дисперсий стоят выборочные оценки.

В случае, когда дисперсии неизвестны и не предполагаются равными, для проверки гипотезы о равенстве средних некоторому значению применяется статистика

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ где}$$

\bar{x}_1 и \bar{x}_2 - выборочные средние, μ_1 и μ_2 - гипотетические генеральные средние, n_1 и n_2 - объемы выборок, s_1^2 и s_2^2 - выборочные дисперсии.

Статистика близка к t -распределению с числом степеней свободы $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$.

Для упрощения вычислений число степеней свободы часто вычисляют по формуле $\min(n_1 - 1, n_2 - 1)$.

Пример. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	выборочная дисперсия
X	16	6	3
Y	6	8	4

С помощью критерия Стьюдента проверить гипотезу о равенстве средних значений этих выборок при 90% уровне доверия против односторонней альтернативы.

Решение. • 1. Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 < \mu_2$$

- Для проверки гипотезы о равенстве средних при известных дисперсиях используется t -статистика

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Критическая область является левосторонней. По таблице t -распределения находим $\alpha = 0.1$, число степеней свободы равно 5 и определяем критические точки $t_{cr} = -1.48$. Критическая область имеет вид $(-\infty; -1.48)$.
- Значение статистики критерия равно $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{6-8}{\sqrt{\frac{3}{16} + \frac{4}{6}}} = -\frac{4}{\sqrt{3}} \approx -2.3$.

- Вывод. Так как $t \in (-\infty; -1.48)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

6.1.4 Доверительный интервал для разности средних

В заключение построим доверительный интервал для разности средних двух генеральных совокупностей. Рассмотрим сперва случай построения доверительного интервала для разности средних, когда дисперсии генеральных совокупностей известны.

Доверительный интервал для разности средних, когда дисперсии генеральных совокупностей известны, имеет вид $((\bar{x}_1 - \bar{x}_2) - \Delta; (\bar{x}_1 - \bar{x}_2) + \Delta)$, где $\Delta = z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, z_α -это квантиль нормального распределения уровня $1 - \frac{\alpha}{2}$.

Пример. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

Построим 90% доверительный интервал для разности средних.

Решение. • Выборочные средние $\bar{x}_1 = 25$ и $\bar{x}_2 = 21$, а дисперсии равны $\sigma_1^2 = 2$ и $\sigma_2^2 = 1$.

- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.95$ и определяем квантиль $z_\alpha = 1.64$. Теперь можем найти точность $\Delta = z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.64 \sqrt{\frac{2}{9} + \frac{1}{6}} \approx 1.02$.
- Искомый 90% доверительный интервал имеет вид $((\bar{x}_1 - \bar{x}_2) - \Delta; (\bar{x}_1 - \bar{x}_2) + \Delta) = (4 - 1.02; 4 + 1.02) = (2.98; 5.02)$.

Ответ: (2.98; 5.02).

В случае, когда дисперсии неизвестны, но предполагаются равными, точность доверительного интервала находится по формуле $\Delta = t_\alpha \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$, где $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, t_α -это квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2}$ с $n_1 + n_2 - 2$ степенью свободы.

Пример. Даны две нормальные выборки числа подтягиваний студентов-менеджеров и студентов-экономистов со следующими характеристиками

	объем выборки	выборочное среднее	выборочная дисперсия
X	10	15	2
Y	7	12	1

Построим 95% доверительный интервал для разности средних, если дисперсии предполагаются равными.

Решение. • Выборочные средние $\bar{x}_1 = 10$ и $\bar{x}_2 = 7$, а дисперсии равны $s_1^2 = 2$ и $s_2^2 = 1$. Смешанная дисперсия равна $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1) + (n_2-1)} = \frac{2 \cdot 9 + 1 \cdot 6}{15} = 1.6$.

• По таблице t -распределения находим $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$, число степеней свободы $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ и определяем квантиль $t_\alpha \approx 2.13$. Теперь можем найти точность $\Delta = t_\alpha \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \approx 2.13 \sqrt{\frac{1.6}{10} + \frac{1.6}{7}} \approx 1.33$.

• Искомый 95% доверительный интервал имеет вид $(3 - 1.33; 3 + 1.33) = (1.67; 4.33)$.

Ответ: (1.67; 4.33).

6.2 Равенство средних в случае зависимых выборок

Теперь мы будем рассматривать случай зависимых (парных) наблюдений.

Как правило, парные данные возникают, когда работают с одним и тем же набором объектов и наблюдения над ними производят дважды (до и после некоторого воздействия/эксперимента). Требуется выяснить, есть ли эффект от этого воздействия.

Формализуем задачу следующим образом. Пусть имеется совокупность n пар наблюдений $(x_1, y_1), \dots, (x_n, y_n)$. Составим разности $d_i = y_i - x_i$ и проверим гипотезу о равенстве нулю среднего разностей μ_d : $H_0 : \mu_d = 0$; $H_1 : \mu_d \neq 0$.

В предыдущих параграфах выборки были независимыми, а здесь выборка часто одна и та же, просто в разные моменты времени. При выводе формул факт независимости использовался, например, дисперсия разности средних в сумму дисперсий распадалась.

Для проверки гипотезы применяется следующая статистика

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}},$$

где

d - разность между двумя значениями в одной паре,

\bar{d} - выборочное среднее для парных разностей,

μ_d - среднее для парных разностей генеральной совокупности,

s_d - стандартное отклонение разностей для выборки,

n - количество пар.

Стандартное отклонение разностей для выборки вычисляется по одной из формул:

$$s_d = \sqrt{\frac{1}{n-1} \left(\sum (d_i - \bar{d})^2 \right)} = \sqrt{\frac{1}{n-1} \left(\sum d^2 - n\bar{d}^2 \right)} = \sqrt{\frac{1}{n-1} \left(\sum d^2 - \frac{(\sum d)^2}{n} \right)}.$$

Статистика имеет t -распределение с числом степеней свободы $n-1$.

Пример. 10 абитуриентов пришли на подготовительные курсы по ЕГЭ и написали тестирование в начале обучения и после. Результаты теста приведены в таблице

	1	2	3	4	5	6	7	8	9	10
До	7	6	5	4	6	2	10	3	8	5
После	9	6	4	5	7	4	10	6	9	6
Разность d	2	0	-1	1	1	2	0	3	1	1

Проверим гипотезу об отсутствии влияния подготовительных курсов на подготовку абитуриентов на уровне значимости 0.01.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu_d = 0; \quad H_1 : \mu_d \neq 0$$

- Для проверки гипотезы используется t -статистика

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}},$$

где $\bar{d} = \frac{\sum d}{n} = \frac{10}{10} = 1$, $s_d = \sqrt{\frac{1}{n-1} \left(\sum d^2 - \frac{(\sum d)^2}{n} \right)} = \sqrt{\frac{22 - \frac{1}{10}(10)^2}{10-1}} = 1.15$.

- Найдем критическое значение и построим критическую область. Критическая область является двусторонней. По таблице t -распределения находим $\alpha/2 = 0.005$, $1 - \alpha/2 = 0.995$, число степеней свободы $n-1 = 9$ и определяем критические точки $t_{cr} = \pm 3.25$. Критическая область имеет вид $(-\infty; -3.25) \cup (3.25; +\infty)$.
- Вычислим значение статистики критерия. Значение статистики критерия равно $t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{1-0}{\frac{1.15}{\sqrt{10}}} = 2.75$.
- Вывод. Так как $t \notin (-\infty; -3.25) \cup (3.25; +\infty)$, то основная гипотеза H_0 не отвергается.

Значение статистики уже известно $t = 2.75$. Так как альтернатива была двусторонней, то минимальный уровень значимости равен $2 \cdot (1 - t^{-1}(2.75)) \approx 0.022$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Пример. При исследовании эффекта воздействия была взята выборка объема $n = 9$ и получены данные до и после воздействия. Значения разностей d_i следующие: 3, 11, 7, 5, 9, 3, -5, 2, -8. На уровне значимости 5% проверить гипотезу о существенности влияния воздействия.

Решение. • Так как заметно, что разности в основном положительные, сформулируем основную и альтернативную гипотезы следующим образом:

$$H_0 : \mu_d = 0; \quad H_1 : \mu_d > 0$$

- Для проверки гипотезы используем t -статистику

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}},$$

где $\bar{d} = \frac{\Sigma d}{n} = \frac{1}{9}(3+11+7+5+9+3-5+2-8) = 3$, $s_d = \sqrt{\frac{1}{8}(387 - 9 \cdot 3^2)} = 6.18$, так как $\sum x_i^2 = 387$.

- Критическая область является правосторонней. По таблице t -распределения находим $1 - \alpha = 0.95$, число степеней свободы $n - 1 = 8$ и определяем критическую точку $t_{cr} = 1.86$. Критическая область имеет вид $(1.86; +\infty)$.

- Значение статистики критерия равно $t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{3-0}{\frac{6.18}{\sqrt{9}}} = 3.45$.

- Вывод. Так как $t \in (1.86; +\infty)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

6.2.1 Доверительный интервал для разности средних

Теперь для парных данных мы построим доверительный интервал для разности.

Доверительный интервал для среднего разностей имеет вид: $\bar{d} - \Delta < \mu_d < \bar{d} + \Delta$. Точность оценки находится по формуле $\Delta = t_\alpha \frac{s_d}{\sqrt{n}}$, где d - разность между двумя значениями в одной паре, \bar{d} - среднее для парных разностей для выборки, μ_d - среднее для парных разностей генеральной совокупности, s_d - стандартное отклонение разностей для выборки, n - количество пар.

Пример. 10 абитуриентов пришли на подготовительные курсы по ЕГЭ и написали тестирование в начале обучения и после. Результаты теста занесены в таблице

	1	2	3	4	5	6	7	8	9	10
До	7	6	5	4	6	2	10	3	8	5
После	9	6	4	5	7	4	10	6	9	6
Разность d	2	0	-1	1	1	2	0	3	1	1

Построим 99% доверительный интервал для разности средних.

Решение. • Среднее значение разностей равно $\bar{d} = \frac{\Sigma d}{n} = \frac{10}{10} = 1$, а выборочное стандартное отклонение $s_d = \sqrt{\frac{1}{n-1} \left(\sum d^2 - \frac{(\Sigma d)^2}{n} \right)} = \sqrt{\frac{22 - \frac{1}{10}(10)^2}{10-1}} \approx 1.15$.

- По таблице распределения Стьюдента находим $1 - \frac{\alpha}{2} = 0.995$, у нас 9 степеней свободы, поэтому квантиль $t_\alpha = 3.25$. Теперь можем найти точность $\Delta = t_\alpha \frac{s_d}{\sqrt{n}} = 3.25 \frac{1.15}{\sqrt{10}} \approx 1.18$.
 - Искомый 99% доверительный интервал имеет вид $(1 - 1.18; 1 + 1.18) = (-0.18; 2.18)$.
- Ответ:** $(-0.18; 2.18)$.

Пример. При исследовании эффекта воздействия была взята выборка объема $n = 9$ и получены данные до и после воздействия. Значения разностей d_i следующие: 3, 11, 7, 5, 9, 3, -5, 2, -8. Постройте 90% доверительный интервал для среднего разностей. В ответ укажите точность интервальной оценки с точностью до сотых.

Решение. • Среднее значение разностей равно $\bar{d} = \frac{\sum d}{n} = \frac{1}{9}(3 + 11 + 7 + 5 + 9 + 3 - 5 + 2 - 8) = 3$, а выборочное стандартное отклонение $s_d = \sqrt{\frac{1}{n-1} (\sum d^2 - n\bar{d}^2)} = \sqrt{\frac{1}{8}(387 - 9 \cdot 3^2)} \approx 6.18$, так как $\sum x_i^2 = 387$.

- По таблице t -распределения находим $1 - \alpha = 0.95$, число степеней свободы $n - 1 = 8$, определяем квантиль $t_\alpha = 1.86$. Точность $\Delta = t_\alpha \frac{s_d}{\sqrt{n}} = 1.86 \frac{6.18}{\sqrt{9}} \approx 3.83$.
- Искомый 99%-доверительный интервал имеет вид $(3 - 3.83; 3 + 3.83) = (-0.83; 6.83)$.

6.3 Равенство долей и дисперсий

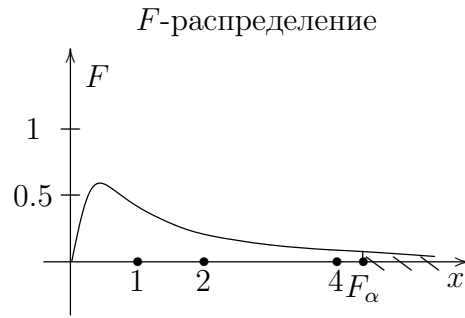
6.3.1 Проверка гипотезы о равенстве дисперсий

Сейчас мы узнаем, каким образом проверить гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей. Это интересно не только само по себе, но и необходимо нам, когда мы проверяем гипотезу о равенстве средних и предполагаем, что дисперсии равны. Такие предположения надо проверять.

Если генеральные совокупности имеют нормальное распределение, то гипотезу о равенстве их дисперсий можно проверить с помощью F -критерия, называемого также критерием Фишера.

Итак, мы хотим проверить гипотезу о равенстве дисперсий: $H_0 : \sigma_1^2 = \sigma_2^2$, где, как обычно, σ_1^2 и σ_2^2 - дисперсии генеральных совокупностей, s_1^2 и s_2^2 - выборочные дисперсии, n_1 и n_2 - объемы выборок. Будем считать, что $s_1^2 > s_2^2$.

Для проверки гипотезы о равенстве дисперсий используется статистика $F = \frac{s_1^2}{s_2^2}$, которая имеет распределение Фишера с числом степеней свободы числителя $n_1 - 1$ и знаменателя $n_2 - 1$.



Пример. Для нормальных выборок объемами 9 и 17 известны выборочные дисперсии 5 и 4 соответственно. Проверим гипотезу о равенстве дисперсий, на уровне значимости $\alpha = 0.05$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \sigma_1^2 = \sigma_2^2; \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

- Для проверки гипотезы о равенстве дисперсией используется F -статистика $F = \frac{s_1^2}{s_2^2}$.
- Критическая область является правосторонней. По таблице F -распределения находим $\alpha = 0.05$, число степеней свободы $n_1 - 1 = 8$ и знаменателя $n_2 - 1 = 16$ и определяем критические точки $F_{cr} = 2.59$. Критическая область имеет вид $(2.59; +\infty)$.
- Значение статистики критерия равно $F = \frac{s_1^2}{s_2^2} = 1.25$.
- Так как $F \notin (2.59; +\infty)$, то основная гипотеза H_0 не отвергается.
- Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $F(1.25) \approx 0.33$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Пример. Для нормальных выборок объемами 16 и 11 известны выборочные стандартные отклонения 28 и 35 соответственно. Проверьте гипотезу о равенстве дисперсий, на уровне значимости $\alpha = 0.1$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : \sigma_1^2 = \sigma_2^2; \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

- Для проверки гипотезы о равенстве дисперсией используется F -статистика $F = \frac{s_1^2}{s_2^2}$, где $s_1^2 > s_2^2$. Поэтому $s_1^2 = 35$, а $s_2^2 = 28$.

- Критическая область является правосторонней. По таблице F -распределения находим $\alpha/2 = 0.05$, число степеней свободы $n_1 \setminus 1 = 10$ и знаменателя $n_2 \setminus 1 = 15$ и находим критическую точку $F_{cr} = 2.55$. Критическая область имеет вид $(2.55; +\infty)$.
- Значение статистики критерия равно $F = \frac{s_1^2}{s_2^2} = \frac{35^2}{28^2} \approx 1.96$.
- Так как $F \notin (2.55; +\infty)$, то основная гипотеза H_0 не отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

6.3.2 Проверка гипотезы о равенстве долей

Наконец, перейдем к сравнению долей признака в двух генеральных совокупностях. То есть мы хотим сравнить долю p_1 некоторого признака в первой генеральной совокупности с долей этого признака p_2 во второй генеральной совокупности. И для этого мы научимся проверять гипотезу $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$.

Здесь мы предполагаем, что выборки независимы и для них выполняются условия $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$. Иначе выводы будут ненадежными.

Проверка гипотезы о равенстве долей осуществляется с помощью z -статистики.

Для проверки гипотезы о равенстве двух долей некоторому значению используется статистика $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$, имеющая нормальный закон распределения.

Здесь p_1 и p_2 - это генеральные доли признака, n_1 и n_2 - объемы выборок, m_1 и m_2 - число «успехов» в каждой выборке, $\hat{p}_1 = \frac{m_1}{n_1}$ и $\hat{p}_2 = \frac{m_2}{n_2}$ - доля «успехов» в каждой выборке, $\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}$ - общая доля «успехов» в двух выборках.

Пример. Пусть из 200 случайно отобранных студентов экономического факультета 86 ездят в университет на машине, а из 300 студентов химического факультета 135 ездят в университет на велосипеде. Проверим гипотезу о равенстве соответствующих долей на уровне значимости $\alpha = 0.05$.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : p_1 = p_2; \quad H_1 : p_1 \neq p_2.$$

- Для проверки гипотезы о равенстве долей используется z -статистика

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}},$$

где

$$\hat{p}_1 = \frac{m_1}{n_1} = \frac{86}{200} = 0.43, \quad \hat{p}_2 = \frac{m_2}{n_2} = \frac{135}{300} = 0.45, \quad \hat{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{86 + 135}{200 + 300} = 0.442.$$

- Значение статистики критерия равно

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{0.42 - 0.45}{\sqrt{\frac{0.442 \cdot 0.558}{200} + \frac{0.442 \cdot 0.558}{300}}} \approx -0.441.$$

- Критическая область является двухсторонней. По таблице z -распределения находим $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$ и определяем критические точки $z_{cr} = \pm 1.96$. Критическая область имеет вид $(-\infty; -1.96) \cup (1.96; +\infty)$.
- Вывод. Так как $z \notin (-\infty; -1.96) \cup (1.96; +\infty)$, то основная гипотеза H_0 не отвергается.

Пример. На уровне значимости 5% проверьте гипотезу о том, что доля нелюбителей математического анализа на мехмате МГУ меньше доли нелюбителей математического анализа на матмехе СПбГУ. Опрошено было по 400 студентов и выборочные доли оказались равными $\hat{p}_1 = 0.12$ и $\hat{p}_2 = 0.16$. В ответ укажите значение статистики критерия.

Решение. • Сформулируем основную и альтернативную гипотезы:

$$H_0 : p_1 = p_2; \quad H_1 : p_1 < p_2.$$

- Для проверки гипотезы о равенстве долей используется z -статистика

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}},$$

где

$$\hat{p}_1 = 0.12, \hat{p}_2 = 0.16, \hat{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{0.12 \cdot 400 + 0.16 \cdot 400}{400 + 400} = \frac{48 + 64}{800} = 0.1425.$$

- Значение статистики критерия равно $z = \frac{0.12 - 0.16}{\sqrt{\frac{0.1425 \cdot 0.8575}{400} + \frac{0.1425 \cdot 0.8575}{400}}} \approx -1.62$.
- Критическая область является левосторонней. По таблице z -распределения находим $1 - \alpha = 0.95$ и определяем критическое значение $z_{cr} = -1.64$. Критическая область имеет вид $(-\infty; -1.64)$.
- Вывод. Так как $z \notin (-\infty; -1.64)$, то основная гипотеза H_0 не отвергается.

Ответ: на уровне значимости 5% нет оснований утверждать, что среди мехматян меньше нелюбителей матана.

6.3.3 Доверительный интервал для разности двух долей

Последней задачей этого параграфа будет нахождение доверительного интервала для разности долей.

Доверительный интервал для разности между долями некоторого признака в двух независимых нормальных генеральных совокупностях имеет вид $((\hat{p}_1 - \hat{p}_2) - \Delta) < p_1 - p_2 < ((\hat{p}_1 - \hat{p}_2) + \Delta)$.

Точность оценки Δ вычисляется по формуле: $\Delta = z_\alpha \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$, где $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{q}_2 = 1 - \hat{p}_2$, а z_α - это квантиль нормального распределения уровня $1 - \frac{\alpha}{2}$.

Пример. Пусть объемы выборок равны $n_1 = 100$ и $n_2 = 200$, выборочные доли $\hat{p}_1 = 0.2$, $\hat{p}_2 = 0.25$. Построим 95% доверительный интервал для разности долей генеральных совокупностей.

Решение. • Выборочные доли $\hat{p}_1 = 0.2$ и $\hat{p}_2 = 0.25$, поэтому $\hat{q}_1 = 1 - \hat{p}_1 = 0.8$ и $\hat{q}_2 = 1 - \hat{p}_2 = 0.75$. Проверьте, что условия надёжности использования этих формул $n\hat{p}_i \geq 5$ и $n\hat{q}_i \geq 5$ действительно выполнены.

- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.975$ и определяем квантиль $z_\alpha = 1.96$. Теперь можем найти точность

$$\Delta = z_\alpha \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = 1.96 \sqrt{\frac{0.2 \cdot 0.8}{100} + \frac{0.25 \cdot 0.75}{200}} \approx 0.099.$$

- Искомый 95% доверительный интервал имеет вид $((\hat{p}_1 - \hat{p}_2) - \Delta; (\hat{p}_1 - \hat{p}_2) + \Delta) = (0.05 - 0.099; 0.05 + 0.099) = (-0.049; 0.0149)$.

Ответ: $(-0.049; 0.0149)$.

Пример. Пусть из 100 жителей Калуги, которые ходят в библиотеку, 75 имеют высшее образование, а 25 не имеют. Постройте 95% доверительный интервал для разности долей.

Решение. Внимание! Пользоваться формулой доверительного интервала для разности долей, приведенной выше, нельзя, так как выборки зависимы (человек либо имеет высшее образование, либо нет). Поэтому решать эту задачу следует следующим образом.

Построим доверительный интервал для доли людей с высшим образованием, которые посещают библиотеку. Выборочная доля равна $\hat{p} = \frac{75}{100} = 0.75$, по таблице нормального распределения найдём квантиль уровня $1 - \alpha/2 = 0.975$ $z_\alpha = 1.96$. Откуда находим точность оценки $\Delta = z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \sqrt{\frac{0.75 \cdot 0.25}{100}} = 0.085$, а сам 95% доверительный интервал имеет вид $(0.75 - 0.085; 0.75 + 0.085) = (0.665; 0.835)$.

Если доля людей с высшим образованием равна p , то доля людей без высшего образования равна $1 - p$, поэтому разность долей равна $p - (1 - p) = 2p - 1$. Доверительный интервал для разности получается равным $(2 \cdot 0.665 - 1; 2 \cdot 0.835 - 1) = (0.33; 0.67)$.

Ответ: $(0.33; 0.67)$.

Глава 7

Корреляция

7.1 Парный коэффициент корреляции

7.1.1 Коэффициент корреляции Пирсона

Корреляционный анализ – метод математической статистики, используемый для изучения, исследования взаимосвязи между (генеральными) экономическими показателями на основе их наблюдаемых статистических (выборочных) аналогов. При этом сами показатели считаются случайными величинами. Парный корреляционный анализ – изучение взаимосвязи между двумя экономическими показателями, описывающими свойства однотипных объектов из некоторой совокупности.

Приведём несколько базовых фактов из книги [2] о коэффициенте корреляции, которые необходимы для понимания его значимости и границ его применимости.

Пусть (X, Y) – двумерная **нормально распределенная** случайная величина. Тогда «степень зависимости» случайных величин X и Y характеризуется парным коэффициентом корреляции

$$\rho = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{E(XY) - EX \cdot EY}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}.$$

Из определения коэффициента корреляции следует, что

1. всегда $-1 \leq \rho \leq 1$;
2. не меняется при линейных преобразованиях величин, т.е.

$$\text{corr}(X, Y) = \text{corr}(a_0 + a_1X, b_0 + b_1Y), \quad a_1, b_1 \neq 0.$$

Коэффициент корреляции принимает крайние значения ± 1 в том и только том случае, когда между случайными величинами X и Y существует **линейная функциональная** зависимость, т.е.

$$\rho = \pm 1 \Leftrightarrow Y = \beta_0 + \beta_1 X, \quad \beta_1 \neq 0,$$

причем

$$\beta_1 = \rho \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}},$$

т.е. знак коэффициента β_1 совпадает со знаком коэффициента корреляции.

В общем случае коэффициент корреляции возникает при решении следующей экстремальной задачи: подобрать линейную функцию $l(x) = \beta_0 + \beta_1 x$ так, чтобы случайная величина $l(X)$ меньше всего отклонялась от Y в среднеквадратичном, т.е.

$$\mathbb{E}(Y - \beta_0 - \beta_1 X)^2 \xrightarrow{\beta_0, \beta_1} \min.$$

Решение этой задачи задается равенствами

$$\beta_1^* = \frac{\text{cov}(X, Y)}{\text{Var}(X)} = \rho \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}}, \quad \beta_0^* = \mathbb{E}Y - \beta_1^* \cdot \mathbb{E}X$$

и наименьшее среднеквадратичное отклонение равно

$$\mathbb{E}(Y - \beta_0^* - \beta_1^* X)^2 = (1 - \rho^2) \text{Var}(Y).$$

Кроме того, для всех $x \in \mathbb{R}$ верно

$$\mathbb{E}(Y|X = x) = \beta_0^* + \beta_1^* x,$$

т.е. наилучший прогноз случайной величины Y , при условии, что известно значение случайной величины $X = x$, равен $\hat{Y} = \beta_0^* + \beta_1^* x$. Рассмотрим три случая:

1. $\rho > 0$. Тогда $\beta_1^* > 0$ и при увеличении x ожидаемое (среднее) значение $\mathbb{E}(Y|X = x)$ случайной величины Y также увеличивается. В этом случае говорят о **прямой линейной зависимости** между величинами.
2. $\rho < 0$. Тогда $\beta_1^* < 0$ и при увеличении x ожидаемое (среднее) значение $\mathbb{E}(Y|X = x)$ случайной величины y уменьшается. В этом случае говорят об **обратной линейной зависимости** между величинами.
3. $\rho = 0$. Тогда $\beta_1^* = 0$, $\mathbb{E}(Y|X = x) = \beta_0^*$ и знание значения случайной величины X не улучшает прогноз Y .

Важное значение коэффициента корреляции обусловлено следующей теоремой.

Теорема. Пусть (X, Y) – двумерная нормально распределенная случайная величина. Тогда случайные величины X и Y независимы, тогда и только тогда, когда $\text{corr}(X, Y) = 0$.

Таким образом, парный коэффициент корреляции можно рассматривать как *меру зависимости* двух случайных величин (факторов), имеющих совместное нормальное распределение, причем:

- $\rho = 0 \Leftrightarrow$ величины независимы;
- $\rho = \pm 1 \Leftrightarrow$ между величинами линейная функциональная зависимость: $y = \beta_0^* + \beta_1^* x$.

7.1.2 Выборочный коэффициент корреляции

Пусть $(x_i, y_i)_{i=1}^n$ – выборка из двумерной нормально распределенной случайной величины, n – объем выборки.

Напомним, что выборочные (неисправленные) оценки дисперсий случайных величин X и Y определяются как

$$\widehat{\text{Var}}(X) = \hat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \overline{(x^2)} - (\bar{x})^2$$

$$\widehat{\text{Var}}(Y) = \hat{\sigma}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \overline{(y^2)} - (\bar{y})^2,$$

где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \overline{(x^2)} = \frac{1}{n} \sum_{i=1}^n x_i^2.$$

Напомним также, что $\widehat{\text{Var}}(X)$ и $\widehat{\text{Var}}(Y)$ – состоятельные, но смещенные оценки дисперсий $\text{Var}(X)$ и $\text{Var}(Y)$ соответственно.

Выборочный коэффициент ковариации определяется как¹

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y},$$

а *выборочный коэффициент корреляции* определяется равенством²

$$r = \widehat{\text{corr}}(X, Y) = \frac{\widehat{\text{cov}}(X, Y)}{\sqrt{\widehat{\text{Var}}(X) \cdot \widehat{\text{Var}}(Y)}}, \quad -1 \leq r \leq 1$$

Выборочные коэффициенты ковариации и корреляции являются состоятельными оценками коэффициентов ковариации и корреляции в генеральной совокупности. Выборочный коэффициент корреляции может рассматриваться как выборочная «мера линейной зависимости» между случайными величинами.

Пример. Вычислите выборочный коэффициент корреляции для следующих пар данных:

x	1	2	3	4
y	4	1	3	2

Решение. Воспользуемся формулой

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

Для этого вычислим суммы:

$$\sum x_i y_i = 1 \cdot 4 + 2 \cdot 1 + 3 \cdot 3 + 4 \cdot 2 = 23;$$

¹В MS Excel функция КОВАР(\cdot, \cdot)

²В MS Excel функция КОРРЕЛ(\cdot, \cdot)

$$\begin{aligned}\sum x_i &= 1 + 2 + 3 + 4 = 10; \\ \sum y_i &= 4 + 1 + 3 + 2 = 10; \\ \sum x_i^2 &= 1^2 + 2^2 + 3^2 + 4^2 = 30; \\ \sum y_i^2 &= 4^2 + 1^2 + 3^2 + 2^2 = 30.\end{aligned}$$

Подставляем их в формулу

$$r = \frac{4 \cdot 23 - 10 \cdot 10}{\sqrt{4 \cdot 30 - 10^2} \sqrt{4 \cdot 30 - 10^2}} = \frac{-8}{20} = -0.4.$$

Для того чтобы не забывать о том, что коэффициент корреляции представляет собой меру линейной связи между переменными, рассмотрим выборку

x	-2	-1	0	1	2
y	4	1	0	1	4

Очевидно, что переменные могут быть связаны соотношением $y = x^2$. А вот коэффициент корреляции при этом равен нулю. Проверьте!

В то же время для следующих пар данных

x	0	1	2	4
y	-1	0	1	3

можно заметить, что x и y связаны линейной положительной зависимостью $y = x - 1$, поэтому коэффициент корреляции равен 1.

Проверка значимости коэффициента корреляции

Проверка значимости подразумевает проверку статистической гипотезы

$$H_0 : \rho = 0$$

против двусторонней альтернативы

$$H_0 : \rho \neq 0.$$

Другими словами, проверяется статистическая гипотеза, что в генеральной совокупности случайные величины (факторы) X и Y **некоррелируют**. Так как двумерная случайная величина (X, Y) по предположению имеет совместное нормальное распределение, то некоррелируемость означает независимость факторов. Проверка гипотезы о независимости факторов основана на следующем результате: при справедливости нулевой гипотезы t -статистика

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{H_0}{\sim} t_{n-2}$$

имеет распределение Стьюдента с $(n-2)$ степенями свободы. Следовательно, получаем следующий статистический критерий проверки нулевой гипотезы:

при заданном уровне значимости α гипотеза H_0 отвергается в пользу альтернативы H_1 при $|t| > t$,

где $t = t(\alpha; n - 2)$ есть **двустороннее** критическое значение распределения Стьюдента t_{n-2} . Напомним, что двустороннее критическое значение определяется как решение уравнения

$$P(|t_{n-2}| > t) = \alpha.$$

При $|t| < t$ говорят, что данные *согласуются* или *не противоречат* нулевой гипотезой, H_0 *не отвергается*.

Пример. Был рассчитан выборочный коэффициент корреляции $r = 0.68$ между дневными логарифмическими доходностями³ биржевых индексов NASDAQ и FTSE на основе $n = 62$ выборочных данных. Проверим значимость коэффициента корреляции, т.е. проверим статистическую гипотезу H_0 о **независимости** доходностей обоих биржевых индексов (в предположении их **нормальной распределенности!**). Вычислим значение t -статистики:

$$t = \frac{0.68 \cdot \sqrt{62 - 2}}{\sqrt{1 - 0.68^2}} \approx 7.1838.$$

Критическое значение распределения Стьюдента при уровне значимости $\alpha = 5\%$ равно: $t_{cr} = t(5\%; 62 - 2) \approx 2.003$. Так как $|t| > t_{cr}$, то гипотеза H_0 о независимости доходностей **отвергается**, коэффициент корреляции значим.

Пример. Пусть для выборки объема $n = 66$ выборочный коэффициент корреляции равен $r = 0.8$. На уровне значимости $\alpha = 0.05$ проверим гипотезу о значимости коэффициента корреляции.

Решение. Сформулируем основную и альтернативную гипотезы:

$$H_0 : \rho = 0 \quad H_1 : \rho \neq 0$$

Для проверки гипотезы используется t -статистика $t = r \sqrt{\frac{n-2}{1-r^2}}$ с $n - 2$ степенями свободы.

Найдем критическое значение и построим критическую область. Критическая область является двухсторонней. В таблице t -распределения находим $\alpha/2 = 0.025$, число степеней свободы $n - 2 = 64$ и определяем критические точки $t_{cr} = \pm 1.99$. Критическая область имеет вид $(-\infty; -1.99) \cup (1.99; +\infty)$.

Вычислим значение статистики критерия. Значение статистики критерия равно $t = r \sqrt{\frac{n-2}{1-r^2}} = 0.8 \sqrt{\frac{64}{1-0.64}} \approx 10.67$.

Вывод. Так как $t \in (-\infty; -1.99) \cup (1.99; +\infty)$, то основная гипотеза H_0 отвергается.

Замечание. Минимальный уровень значимости, начиная с которого гипотеза отвергается, составляет $2 \cdot (1 - t^{-1}(10.67)) \approx 0$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

³Логарифмическая доходность рассчитывается как $h_t = \ln(S_t/S_{t-1})$.

Пример. Пусть в результате наблюдений над некоторой двумерной случайной величиной (X, Y) получены методом случайной выборки некоторые совокупности пар значений

i	1	2	3	4	5	6	7	8	9	10
x_i	1,5	3,5	0,5	1,5	2	1,5	4,5	3	3,5	3
y_i	6,5	6	6,5	4,5	8	5,5	6,5	6	8	4,5

Проверить значимость выборочного коэффициента корреляции.

Вычислим парный коэффициент корреляции. Имеем

$$\bar{x} = 2.45; \quad (\bar{x})^2 = 6.0025; \quad \overline{x^2} = 7,375; \quad \bar{y} = 6.2; \quad (\bar{y})^2 = 38.44; \quad \overline{y^2} = 39.75; \quad \overline{xy} = 15.375;$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \sqrt{\overline{y^2} - (\bar{y})^2}} \approx 0.161.$$

Проверим значимость выборочного коэффициента корреляции. Вычислим t -статистику

$$t = \frac{0.161 \cdot \sqrt{10-2}}{\sqrt{1-0.161^2}} \approx 0.46.$$

При уровне значимости $\alpha = 0.05$ критическое значение распределения Стьюдента равно $t_{cr} = t(0.05; 10-2) \approx 2.31$. Так как $|t| < t_{cr}$, то коэффициент незначим, данные согласуются с нулевой гипотезой о независимости факторов при уровне значимости 5%.

Итак, мы научились с помощью коэффициента корреляции отвечать на вопрос, существует ли линейная связь между двумя переменными x и y . Если такая связь наблюдается, то нам хотелось бы указать приблизительный вид этой связи, а это уже позволяет делать прогнозы. Но эта тема будет предметом курса эконометрики.

Замечание. Стоит отметить, что с помощью статистики мы пытаемся выявить наличие связей, а не причины, по которым они возникают. Никогда не надо забывать о возможности ложной корреляции.

Глава 8

Упражнения для самопроверки

8.1 Описательная статистика и эмпирическая функция распределения

1. Дискретная случайная величина X задана следующим рядом распределения:

X	-8	-4	1	8
p	0.25	0.25		0.5

Дополнить таблицу и вычислить математическое ожидание, дисперсию, моду, медиану, нижнюю квартиль. Построить функцию распределения.

2. Для выборки -8, -4, 8, -8, -4, 8, 8, 8 вычислить моду, медиану, квартили, межквартильный размах, среднее (двумя способами) и среднеквадратичное отклонение этой величины (двумя способами). Нарисовать гистограмму частот, эмпирическую функцию распределения.

3. Дискретная случайная величина X задана следующим рядом распределения:

X	-1	0	2	3
p	0.1		0.1	0.3

Дополнить таблицу и вычислить математическое ожидание, дисперсию, моду, медиану, квартили. Построить функцию распределения.

4. Для выборки 0, -1, 3, 2, 0, 0, 0, 3, 3, 0 вычислить моду, медиану, квартили, межквартильный размах, среднее (двумя способами) и среднеквадратичное отклонение этой величины (двумя способами). Нарисовать гистограмму частот, эмпирическую функцию распределения.

5. В результате независимых наблюдений случайной величины были получены следующие ее значения: -1, 2, 3, 0, 1, 2, 7. Вычислить моду, медиану, квартили, межквартильный размах, выбросы, среднее и среднее квадратичное отклонение этой величины. Нарисовать коробчатую диаграмму, эмпирическую функцию распределения и гистограмму частот.
6. В результате независимых наблюдений случайной величины были получены следующие ее значения: 0, 2, 6, 5, 5, 3, 1, 4, 6, 2. Вычислить моду, медиану, квартили, межквартильный размах, выбросы, ранги, среднее и среднее квадратичное отклонение этой величины. Нарисовать эмпирическую функцию распределения, полигон и гистограмму частот с шагом 2.
7. Дана выборка из нормального распределения со средним значением 1 и дисперсией 1: 2, -1, 0, 2, 3, 4. Найти разность $|F_n(x) - F(x)|$ при $x = 2.5$, где $F_n(x)$ и $F(x)$ эмпирическая и теоретическая функции распределений.
8. Дана выборка из равномерного распределения на отрезке $[-4; 10]$: 2, -1, 0, 3, 2, 1, 7. Найти разность $F_n(x) - F(x)$ при $x = 4$, где $F_n(x)$ и $F(x)$ эмпирическая и теоретическая функции распределений.
9. Покажите, что $s_n^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \bar{x}^2 - (\bar{x})^2$.
10. * Докажите, что при любом $\varepsilon > 0 \lim_{n \rightarrow +\infty} P(|F_n(x) - F(x)| < \varepsilon) = 1$ для всех x . Как это соотносится с теоремой Гливенко-Кантелли?

8.2 Свойства оценок

11. По выборке x_1, x_2, x_3, x_4, x_5 из нормального распределения $N(\theta, \sigma^2)$ построена следующая оценка параметра θ : $\hat{\theta} = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.3x_4 + 0.1x_5$.
 - а) Является ли оценка $\hat{\theta}$ несмещенной?
 - б) Найти дисперсию оценки $\hat{\theta}$.
 - в) Является ли оценка $\hat{\theta}$ эффективной?
12. По выборке x_1, x_2, x_3, x_4, x_5 из нормального распределения $N(\theta, \sigma^2)$ построена следующая оценка параметра θ : $\hat{\theta} = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_5$. Выяснить, является ли оценка $\hat{\theta}$ несмещенной и эффективной? Предложите смещенную оценку с меньшей дисперсией.

13. Дискретная случайная величина X задана следующим рядом распределения:

X	0	1	2
p	0.5	θ	$0.5 - \theta$

Имеется два наблюдения этой случайной величины x_1 и x_2 .

- а) Рассмотрим оценку $\hat{\theta} = 2 - 0.4x_1 - 0.6x_2$ для параметра θ . Найдите смещение этой оценки.
- б) При каких условиях на коэффициенты $\alpha_0, \alpha_1, \alpha_2$ оценка $\hat{\theta} = \alpha_0 + \alpha_1x_1 + \alpha_2x_2$ будет несмещённой для параметра θ .
- в) Найдите эффективную оценку параметра θ среди оценок такого вида.

14. Дискретная случайная величина X задана следующим рядом распределения:

X	-1	0	1
p	0.5	$0.1 - \theta$	$0.4 + \theta$

Имеется два наблюдения этой случайной величины x_1 и x_2 .

- а) Рассмотрим оценку $\hat{\theta} = 0.2x_1 + 0.8x_2$ для параметра θ . Найдите смещение этой оценки.
- б) При каких условиях на коэффициенты $\alpha_0, \alpha_1, \alpha_2$ оценка $\hat{\theta} = \alpha_0 + \alpha_1x_1 + \alpha_2x_2$ будет несмещённой для параметра θ .
- в) Найдите эффективную оценку параметра θ среди оценок такого вида.

15. По выборке x_1, x_2 из нормальной генеральной совокупности с параметрами θ и σ^2 построена следующая оценка параметра θ : $\hat{\theta} = \frac{1}{2}(x_1 + x_2)$. Докажите, что она является эффективной оценкой среди всех линейных несмещённых оценок, решив соответствующую оптимизационную задачу. Будет ли её дисперсия минимальной среди всех линейных оценок?
16. По выборке x_1, x_2, x_3 из нормальной генеральной совокупности с параметрами θ и σ^2 построена следующая оценка параметра θ : $\hat{\theta} = \frac{1}{3}(x_1 + x_2 + x_3)$. Докажите, что она является эффективной оценкой среди всех линейных несмещённых оценок, решив соответствующую оптимизационную задачу. Будет ли её дисперсия минимальной среди всех линейных оценок?

17. Проверить, являются ли несмещёнными оценки $\hat{\theta}_n$ для выборки из генеральной совокупности с математическим ожиданием θ :
- $\hat{\theta}_n = \frac{1}{n} \sum (x_i - \bar{x})$;
 - $\hat{\theta}_n = \sum \omega_i x_i$, если $\sum \omega_i = 1$.
18. Доказать, что $MSE = E(\hat{\theta} - \theta)^2 = D(\hat{\theta}) + bias^2(\hat{\theta})$.
19. Пусть $\hat{\theta}$ - несмещённая оценка параметра θ с конечной положительной дисперсией $D(\hat{\theta})$. Найти смещение оценки $\hat{\theta}^2$ для θ^2 ?
20. Дана выборка x_1, x_2, \dots из распределения Бернулли с параметром p . Проверить, что x_1 является несмещённой оценкой для p . Является ли эта оценка состоятельной?
21. Дана выборка x_1, x_2, \dots из распределения Пуассона с параметром λ . Проверить, что x_1 является несмещённой оценкой для λ . Является ли эта оценка состоятельной?
22. Доказать, что выборочное среднее является эффективной оценкой математического ожидания нормального распределения, когда дисперсия известна.
23. Для выборки x_1, x_2, \dots, x_n из генеральной совокупности с равномерным распределением на интервале $(0; \theta)$ проверить несмещённость и эффективность оценки $\hat{\theta} = \frac{n+1}{n} x_{\max}$.
24. Доказать, что в схеме Бернулли \hat{p} является эффективной оценкой неизвестной вероятности p .
25. Пусть случайная величина имеет математическое ожидание a . Проверить, что $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$ есть несмещённая оценка дисперсии.
26. Пусть оценка $\hat{\theta}_n$ есть *асимптотически* несмещённая оценка, дисперсия ее стремится к нулю, тогда оценка состоятельна.
27. Доказать, что $s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$ является несмещённой оценкой дисперсии генеральной совокупности.

28. Доказать, что эмпирическая функция распределения является несмещённой и состоятельной оценкой функции распределения генеральной совокупности.

29. Для случайной величины с распределением

X	-1	0	2	4
P	$0,6 + \theta$	$0,1 - \theta$	$0,2$	$0,1$

получена выборка: 4, 0, 2, -1, 4. Вычислите оценку максимального правдоподобия и методом моментов.

30. Для случайной величины с распределением

X	0	1	2
P	θ	3θ	$1 - 4\theta$

получена выборка: 0, 1, 2, 2. Вычислите оценку методом максимального правдоподобия и методом моментов.

31. Для случайной величины с распределением

значения	0	1	2	4
вероятности	$0,5 + \theta$	$0,1 - \theta$	$0,2$	$0,2$

получена выборка: 1, 4, 2, 2, 0, 1.

а) Выпишите функцию правдоподобия и ее логарифм.

б) Вычислите оценку максимального правдоподобия.

32. Для случайной величины с равномерным распределением на отрезке $[a; b]$ получена выборка: 1, 1, 3, 3. Вычислите оценки параметров методом моментов.

33. Дана выборка из пуассоновского распределения с параметром θ : 1, 1, 3, 2, 3, 2. Вычислите оценку максимального правдоподобия.

34. Дана выборка из пуассоновского распределения с параметром θ : 1, 4, 3, 2, 3, 0, 1, 1, 0, 5.

а) Выпишите функцию правдоподобия и ее логарифм.

б) Вычислите оценку максимального правдоподобия и по методу моментов.

35. Дана выборка из нормального распределения со средним θ и дисперсией 1: 1, 3, -2, -1, 0, 4, 1, 2, 1.
- Выпишите функцию правдоподобия и ее логарифм.
 - Вычислите оценку максимального правдоподобия для среднего.
36. Дана выборка из нормального распределения со средним θ и дисперсией σ^2 : 1, 0, 1, 2, 1.
- Выпишите функцию правдоподобия и ее логарифм.
 - Вычислите оценки максимального правдоподобия для среднего и дисперсии.
 - Вычислите оценки для среднего и дисперсии методом моментов.
37. Методом максимального правдоподобия найдите вероятность успеха θ в схеме испытаний Бернулли. Докажите, что эта оценка является несмещенной и состоятельной.
38. По выборке x_1, x_2, \dots, x_n в случае равномерного распределения на отрезке $[0; \theta]$ методом максимального правдоподобия найти оценку параметра θ . Найдите её смещение. Является ли оценка асимптотически несмещенной? А состоятельной?

8.3 Доверительные интервалы

39. Для стандартной нормальной величины X найдите интервал $(a; b)$ минимальной длины такой, что $P(a < X < b) = 0.9545$.
40. Для случайной величины $X \sim N(\mu; \sigma^2)$ найдите интервал $(a; b)$ минимальной длины такой, что $P(a < \frac{X-\mu}{\sigma} < b) = 0.9973$.
41. Пусть среднее значение выборки из нормальной генеральной совокупности равно 130. Постройте 98% доверительный интервал для среднего значения, если объем выборки равен 130, а дисперсия генеральной совокупности равна 12.
42. Оценка среднего значения нормальной выборки объема $n = 9$ равна 24, а оценка дисперсии равна 36. Построить 95% доверительный интервал для среднего значения.
43. Для выборки 13, 17, 18, 15, 18, 15 из нормальной генеральной совокупности построить 95% доверительный интервал для среднего.

44. Дана выборка из нормального распределения с неизвестным средним и дисперсией: 4, 1, 3, 6, 2, 2.
- а) Найти оценку медианы этой выборки;
 - б) найти среднее значение этой выборки;
 - в) найти оценку дисперсии;
 - г) Построить 90% доверительный интервал для среднего значения этой выборки.
45. Из предыдущих исследований известно, что месячный доход преподавателей города N имеет нормальное распределение со стандартным отклонением 200 у.е. Опрошено случайным образом 100 преподавателей. Их средний доход составил 1500 у.е. Найти 95% доверительный интервал для среднего месячного дохода преподавателей города N.
46. Из предыдущих исследований известно, что месячный доход преподавателей города N имеет нормальное распределение со стандартным отклонением 200 у.е. Опрошено случайным образом 100 преподавателей. Их средний доход составил 1500 у.е. Определите минимально необходимый объем выборки для построения 95% доверительного интервала для среднего длины 20.
47. У студентов на контрольной сердце бьётся со скоростью 96 ударов в минуту. Известно, что стандартное отклонение равно 15 ударам в минуту. Определите минимально необходимый объем выборки для определения среднего с надёжностью 98% и точностью 1.
48. У 20 студентов на контрольной сердце билось со скоростью 96 ударов в минуту. Известно, что стандартное отклонение равно 5 ударам в минуту. Найти 98% доверительный интервал для среднего.
49. Выборочный опрос 80 студентов показал, что 24 из них не поддерживают введение двухбалльной системы экзаменационной оценки: "ХОРОШО" и "ОТЛИЧНО". Найти 96%-ый доверительный интервал для фактической доли студентов, поддерживающих это нововведение.
50. Опрос 100 иногородних студентов показал, что 73 из них живут в общежитии, а остальные снимают квартиры. Найдите 94% доверительный интервал для фактической доли иногородних студентов, которые снимают квартиры.
51. Постройте 90% доверительный интервал для стандартного отклонения, если выборочная дисперсия равна 625, объем выборки равен 19.
52. Найдите 96% доверительный интервал для дисперсии по выборке объема $n = 18$, если выборочное стандартное отклонение равно 10.

53. По результатам обследования была оценена доля людей, страдающих сердечно-сосудистыми заболеваниями и получен симметричный 90% доверительный интервал $(0.11; 0.15)$. Какой объем выборки при этом использовался?
54. Найти минимальный объем выборки, при котором с надёжностью 0.975 точность оценки математического ожидания по выборочному среднему равна 0.3, если известно среднее квадратическое отклонение $\sigma = 1.2$ нормальной генеральной совокупности.
55. Оцените объём выборки, который необходим для построение 90% доверительного интервала для доли длины 0.02.
56. По выборке из нормального распределения с известной дисперсией σ^2 постройте односторонний 95% доверительный интервал для среднего вида $(\theta; +\infty)$.
57. По выборке из нормального распределения с известной дисперсией σ^2 постройте односторонний 99% доверительный интервал для среднего вида $(-\infty; \theta)$.

8.4 Проверка гипотез (одна выборка)

58. Проводится медицинский тест(анализ) для проверки наличия болезни. Изначально она предполагается. Что будет означать для больного ошибка первого рода? А второго?
59. Вы собираете грибы и априори, считает, что они съедобные (презумпция съедобности грибов). Что такое в этом случае ошибка первого рода? А второго?
60. Дана случайная выборка из нормального распределения $N(\mu, \sigma^2)$. Вычислите математическое ожидание и дисперсию величины $z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma}$. Докажите, что z стремится к стандартному нормальному распределению при $n \rightarrow \infty$.
61. Дана случайная выборка из биномиального распределения с вероятностью успеха p . Вычислите математическое ожидание и дисперсию величины $z = \frac{(\hat{p} - p)\sqrt{n}}{\sqrt{pq}}$. Докажите, что z стремится к стандартному нормальному распределению при $n \rightarrow \infty$.
62. По случайной выборке из нормального распределения $N(\theta, \sigma^2 = 1)$: 9, 5, 7, 7, 4, 10 проверить нулевую гипотезу о том, что $\theta = 6$, при 1% уровне значимости против односторонних альтернатив $\theta > 6$.
63. По случайной выборке из нормального распределения $N(\theta, \sigma^2 = 1)$: 3, 2, 4, 5, 1, 3 проверить нулевую гипотезу о том, что $\theta = 4$, при 5% уровне значимости против односторонних альтернатив $\theta < 4$.

64. Дана выборка из нормального распределения с неизвестным средним и дисперсией: 9, 6, 8, 11, 7, 7. На 5% уровне значимости проверьте гипотезу о равенстве среднего значения пяти против двусторонних альтернатив (двумя способами).
65. Дана выборка из нормального распределения с неизвестным средним и дисперсией: 4, 1, 3, 6, 2, 2. На 10% уровне значимости проверьте гипотезу о равенстве среднего значения двум против двусторонних альтернатив (двумя способами).
66. Студенты ВУЗа предполагают, что более 10% всех студенток в этом ВУЗе страдают от избыточного веса. Проверьте предположение на уровне значимости 5%, если в выборке из 90 студенток избыточный вес оказался у 18.
67. Студентки ВУЗа предполагают, что более половины всех студентов в этом ВУЗе курят. Проверьте предположение на уровне значимости 1%, если в выборке из 80 человек в курении признались только 20.
68. Среди студентов проводился опрос с целью выяснить среднюю стоимость мобильного телефона. Результаты получились такими: 4900, 6500, 5000, 7300, 6400, 6300, 5800, 6200, 6100, 5500, 6300, 6200, 6000. На уровне значимости 14% проверить гипотезу о том, что у 60% студентов мобильный телефон стоит дороже 6000 рублей. Можно ли доверять полученным выводам?
69. Среди покупателей проводился опрос с целью выяснить среднюю стоимость наручных часов. Результаты таковы: 2900, 1500, 5000, 10000, 7000, 6300, 1600, 5200, 3800, 1200, 1700, 1100. На уровне значимости 12% проверить гипотезу о том, что у 55% покупателей часы стоят дороже 3000 рублей. Можно ли доверять полученным выводам?
70. Проверьте гипотезу о равенстве дисперсии 75 на уровне значимости 0.05 против односторонних альтернатив, если выборочная дисперсия получилась равной 81. Объём выборки равен 26.
71. Проверьте гипотезу о равенстве дисперсии 361 на уровне значимости 0.01 против односторонних альтернатив, если выборочная дисперсия получилась равной 225. Объём выборки равен 16.
72. Если при проверке гипотезы о среднем $H_0 : \mu = 0.02$ против альтернативы $H_1 : \mu \neq 0.02$ было получено $p - value$ 0.03, то нулевая гипотеза отвергается на уровне значимости
- (a) 0.01
 - (b) 0.05
 - (c) 0.1

73. Если при проверке гипотезы о среднем $H_0 : \mu = 4$ против альтернативы $H_1 : \mu > 4$ было получено p -value 0.072, то нулевая гипотеза отвергается на уровне значимости
- 0.01
 - 0.05
 - 0.1
74. По выборке x_1, \dots, x_{100} из нормальной генеральной совокупности $N(\mu; \sigma^2 = 9)$ для проверки гипотезы о среднем $H_0 : \mu = 3$ против альтернативы $H_a : \mu < 3$ был получен минимальный уровень значимости 0.007. Стоит ли отвергнуть нулевую гипотезу? Вычислите сумму $x_1 + \dots + x_{100}$.
75. По выборке x_1, \dots, x_{121} из нормальной генеральной совокупности $N(\mu; \sigma^2 = 16)$ для проверки гипотезы о среднем $H_0 : \mu = 5$ против альтернативы $H_a : \mu < 5$ был получен минимальный уровень значимости 0.15. Стоит ли отвергнуть нулевую гипотезу? Вычислите сумму $x_1 + \dots + x_{121}$.
76. Исследователь заподозрил, что монета не является симметричной (ему кажется, что герб выпадает реже, чем решка) и выдвинул гипотезу $H_0 : p = 1/2$ против альтернативы $H_a : p = 1/3$. Для проверки гипотезы он подбрасывает монету 10 раз.
- Если число гербов больше 7, то нулевая гипотеза отвергается и принимается H_a .
- Вычислите ошибку первого рода;
 - Вычислите ошибку второго рода и мощность теста.
 - Как изменяться ошибки первого и второго родов, если исследователь увеличит пороговое значение с 7 до 8?
77. Маша долго подбрасывала монету и заметила, что герб выпадает приблизительно в 25% случаев. Она выдвинула гипотезу $H_0 : p = 1/2$ против альтернативы $H_a : p = 1/4$. Чтобы на уровне значимости $\alpha = 0.1$ проверить гипотезу о том, что монета не является симметричной, она устраивает итоговый тест: подбрасывает монету 10 раз и если число гербов превышает C , то она заключает, что монета не симметричная.
- Найдите критическое множество.
 - Вычислите ошибку второго рода и мощность теста.
 - Как изменится мощность теста, если Маша уменьшит уровень значимости до $\alpha = 0.01$?
78. Костя долго подбрасывал монету и заметил, что герб выпадает приблизительно в 75% случаев. Чтобы проверить гипотезу о том, что монета не является симметричной, он устраивает итоговый тест: подбрасывает монету 15 раз и если число гербов превышает 9, то он заключает, что монета не симметричная.

- Сформулируйте основную и альтернативную гипотезу.
- Вычислите ошибки первого и второго родов.
- Как изменятся ошибки первого и второго родов, если Костя увеличит пороговое значение с 9 до 10?

79. Пусть p - это доля сторонников некоторого кандидата в президенты. Исследователь хочет проверить, что его поддерживает половина избирателей и выдвигает основную гипотезу $H_0 : p = 1/2$ против односторонних альтернатив $H_a : p > 1/2$. Для этого он опрашивает 20 человек и если число сторонников в выборке оказывается больше 12, то основная гипотеза отвергается в пользу альтернативы.

- Найдите ошибку первого рода.
- Вычислите мощность теста для разных значений p (например, 0.6, 0.7, 0.8, 0.9) и нарисуйте приблизительный график функции мощности.

80. Рассматривается выборка объёма 64 из нормальной генеральной совокупности $N(\mu, 4)$. Для проверки гипотезы $H_0 : \mu = 0$ против двусторонних альтернатив $H_a : \mu \neq 0$ исследователь делает следующее:

вычисляет среднее \bar{x} и если $\bar{x} \in [-1; 1]$, то нулевая гипотеза не отвергается, если $\bar{x} \notin [-1; 1]$, то отвергает H_0 в пользу H_a .

- Вычислите уровень значимости;
- Вычислите мощность теста, если $\mu = 1$.

81. Рассматривается выборка объёма 100 из нормальной генеральной совокупности $N(\mu, 9)$. Для проверки гипотезы $H_0 : \mu = 2$ против односторонних альтернатив $H_a : \mu < 2$ на 5% уровне значимости исследователь делает следующее:

вычисляет среднее \bar{x} и если $\bar{x} \geq C$, то нулевая гипотеза не отвергается, если $\bar{x} < C$, то он отвергает H_0 и принимает H_a .

- Вычислите C ;
- Вычислите мощность теста, если $\mu = 0$.

8.5 Сравнение выборок

82. Даны две независимые выборки x_1, \dots, x_{n_1} и x_1, \dots, x_{n_2} из нормальных генеральных совокупностей с одинаковым распределением $N(0; \sigma^2)$. Доказать, что

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

имеет стандартное нормальное распределение.

83. Даны две независимые выборки x_1, \dots, x_{n_1} и x_1, \dots, x_{n_2} из нормальных генеральных совокупностей $N(\mu_1; \sigma_1^2)$ и $N(\mu_2; \sigma_2^2)$ соответственно. Доказать, что

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

имеет стандартное нормальное распределение.

84. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

Проверьте гипотезу о равенстве средних значений этих выборок при 95% уровне доверия против односторонней альтернативы. В ответе укажите значение статистики критерия.

85. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	выборочная дисперсия
X	10	15	2
Y	7	12	1

С помощью критерия Стьюдента проверьте гипотезу о равенстве средних значений этих выборок (считая их дисперсии равными) при 95% уровне доверия против двусторонних альтернатив. В ответ укажите минимальный уровень значимости.

86. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	выборочные дисперсии
X	10	15	2
Y	7	12	1

Построить 95%-доверительный интервал для разности средних, если дисперсии предполагаются равными.

87. Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

Построить 90%-доверительный интервал для разности средних.

88. Средний уровень ежемесячных продаж компании в прошедшем году составил 15 у.е. В новом году компания сменила рекламную стратегию. При этом были зафиксированы следующие ежемесячные показатели продаж: 13, 17, 18, 15, 15, 18. При 95% уровне доверия проверьте гипотезу о том, что средний уровень продаж после рекламы не изменился.
89. 10 абитуриентов пришли на подготовительные курсы по ЕГЭ и написали тестирование в начале обучения и после. Результаты теста занесены в таблице

	1	2	3	4	5	6	7	8	9	10
До	7	6	5	4	6	2	10	3	8	5
После	9	6	4	5	7	4	10	6	9	6

Проверьте гипотезу об отсутствии влияния подготовительных курсов на подготовку студентов на уровне значимости 0.01.

90. Для исследования эффективности препарата для похудения было проведено исследование. Был записан вес пациента до приёма препарата и после. В результате получились следующие разности для веса каждого пациента: 2, -3, 4, 1, 1, -3, -1, 5, 3, 1, 2, 6, 5, -2, 0. На уровне значимости 5% проверьте гипотезу об отсутствии влияния препарата против двусторонних альтернатив.
91. Для нормальных выборок объемами 9 и 17 известны выборочные дисперсии 5 и 4 соответственно. Проверьте гипотезу о равенстве дисперсий, на уровне значимости $\alpha = 0.05$ против односторонних альтернатив.
92. Для нормальных выборок объемами 26 и 18 известны выборочные дисперсии 36 и 10 соответственно. Проверьте гипотезу о равенстве дисперсий, на уровне значимости $\alpha = 0.05$ против односторонних альтернатив.
93. Даны две нормальные выборки. Пусть из 200 случайно отобранных студентов экономического факультета 86 ездят в университет на машине, а из 300 студентов химического факультета - 135 человек. Проверьте гипотезу о равенстве долей студентов каждого факультета, которые ездят в университет на машине, на уровне значимости $\alpha = 0,05$. Укажите минимальный уровень значимости.
94. Объем выборок $n_1 = 100$ и $n_2 = 200$, выборочные доли $\hat{p}_1 = 0.2$, $\hat{p}_2 = 0.25$. Постройте 95% доверительный интервал для разности долей.
95. Пусть объем выборок $n_1 = 100$ и $n_2 = 200$, выборочные доли $\hat{p}_1 = 0.43$, $\hat{p}_2 = 0.45$. Постройте 95% доверительный интервал для разности долей.
96. Из 200 случайных прохожих, 80 ходят в наушниках, а остальные нет. Постройте 96% доверительный интервал для разности долей.

97. Из опрошенных 100 случайных фанатов, идущих на футбольный матч 80 знают, кто будет играть, 20 нет. Постройте 99% доверительный интервал для разности долей.

8.6 Коэффициент корреляции

98. Рассчитанный по выборочным данным коэффициент корреляции оказался равным -1. Это означает что:
- (a) между изучаемыми переменными есть слабая отрицательная линейная связь;
 - (b) между изучаемыми переменными есть связь, но она не является линейной;
 - (c) между изучаемыми переменными есть функциональная линейная отрицательная связь;
 - (d) между изучаемыми переменными отсутствует связь;
 - (e) полученное число никак не интерпретируется, допущена ошибка в вычислениях;
 - (f) между переменными есть функциональная положительная зависимость;
 - (g) между изучаемыми переменными достаточно сильная прямая линейная корреляционная зависимость;
 - (h) ни один из предложенных ответов не является правильным.
99. Финансовая ситуация вынудила фирму резко сократить расходы на рекламу. В скором времени упали объемы продаж, но в меньшей степени, чем ожидалось. Какому выборочному значению коэффициента корреляции между затратами на рекламу и объемом продаж может соответствовать данная ситуация:
- (a) $\hat{\rho} = -0.6$;
 - (b) $\hat{\rho} = 0.9$;
 - (c) $\hat{\rho} = 0.5$;
 - (d) $\hat{\rho} = -0.3$?

Ответ обосновать.

100. Два сотрудника нефтяной компании изучали зависимость объема добычи нефти и мировой цены на нефть. Каждый из них вычислил показатель ковариации и коэффициент корреляции. Первый сотрудник объем добычи считал в баррелях и цену в долларах, а второй – в тоннах и рублях соответственно. Потом они сравнили результаты. Одинаковыми или различными были получены у них результаты? Ответ поясните.

101. На основе опроса 27 семей был вычислен коэффициент корреляции между доходами и расходами на питание: $r = 0.26$. Значимо ли рост доходов влияет на рост расходов на питание семьи (при уровне значимости 2%)?
102. По данным 24 магазинов был вычислен коэффициент корреляции между ценой и объемом продаж некоторого товара: $r = -0.37$. Значимо ли увеличение цены влияет на уменьшение объема продаж (при уровне значимости 0.1%)?
103. На основе опроса 27 семей был вычислен коэффициент корреляции между доходами и накоплениями: $r = 0.62$. Значимо ли рост доходов влияет на рост накоплений (при уровне значимости 10%)?
104. По 25 предприятиям был вычислен коэффициент корреляции между объемом продаж и затратами на рекламу: $r = 0.42$. Значимо ли рост затрат на рекламу влияет на рост продаж (при уровне значимости 1%)?
105. Даны две выборки:

X	4	1	6	2	2
Y	5	1	7	1	1

Вычислить коэффициент корреляции.

106. Даны две выборки:

X	4	1	3	6	2	2
Y	7	2	3	8	2	2

Вычислить коэффициент корреляции Пирсона.

107. Исследователь хочет определить, существует ли связь между возрастом человека и тем, сколько часов в день он или она смотрит телевизор. Уровень значимости $\alpha = 0,1$.

возраст	18	24	36	40	58
количество часов	3,9	2,6	2	2,3	1,2

108. Менеджер магазина хотел бы узнать, существует ли какая-либо связь между возрастом работников и количеством больничных, которые они берут каждый год. Уровень значимости $\alpha = 0,01$.

возраст	18	26	39	48	53	58
дни болезни	16	12	9	5	6	2

109. Преподавателю необходимо узнать, какова связь между IQ студента и его успеваемостью. Уровень значимости $\alpha = 0,05$.

IQ	98	105	100	100	106	95	116	112
средний балл	2.1	2.4	3.2	2.7	2.2	2.3	3.8	3.4

110. Офис-менеджер хочет определить, есть ли связь между тем, сколько лет уже прослужила копировальная машина, и тем, во сколько обходится ее ремонтное обслуживание в течение месяца. Уровень значимости $\alpha = 0,2$.

возраст ксерокса	3	5	2	1	2	4	3
стоимость обслуживания	80	100	75	60	80	93	84

111. Даны девять пар наблюдений

x	-1	0	1	2	3	4	5	6	7
y	1	-2	-1	-4	-4	-6	-5	-8	-7

Вычислите коэффициент корреляции. Проверьте его значимость на уровне $\alpha = 0.05$.

112. Наблюдения 16 пар (x, y) дали следующие результаты:

$$\sum y^2 = 526, \sum x^2 = 657, \sum xy = 492, \sum y = 64, \sum x = 96.$$

Найдите коэффициент корреляции.

113. По 10 наблюдениям показателей x и y были получены следующие данные:

$$\begin{aligned} \sum x_i &= 1700, \quad \sum y_i = 1100, \quad \sum x_i y_i = 204400 \\ \sum x_i^2 &= 316000, \quad \sum y_i^2 = 135000 \end{aligned}$$

Вычислите коэффициент корреляции.

114. Какое наименьшее (по абсолютной величине) значение выборочного коэффициента корреляции следует считать значимым на 5% уровне значимости, если объем выборки $n = 38$?

Глава 9

Дополнение. Непараметрические методы

В предыдущих разделах рассматривались тесты для проверок гипотез о математических ожиданиях, дисперсиях и пр. Однако при построении этих тестов предполагалось, что тестируемые генеральные совокупности имеют нормальное распределение. Желательно иметь возможность работать и с выборками, которые не имеют нормального распределения и которые применимы для номинальных и порядковых данных.

Кроме того, параметрические методы не позволяют ответить на вопрос о том, как протестировать тот факт, что выборка имеет заданное распределение.

Разумеется, при отказе от предположения о том, что выборка имеет нормальное распределение, должна снизиться мощность критерия. Это означает, что для того, чтобы нулевая гипотеза была отвергнута, требуются значительные отклонения статистики критерия.

Для обеспечения той же мощности приходится брать больший объём выборки. Так, например, для критерия знаков, являющегося непараметрическим аналогом t -теста, приходится брать выборку примерно в полтора раза большую. Для других приводимых ниже критериев (кроме критериев случайности и нормальности, не имеющих параметрических аналогов) - примерно на 10 процентов большие выборки.

Ниже приведены непараметрические критерии следующих трёх основных типов. Во-первых, это критерии случайности, проверяющие гипотезы о том, что выборка взята из одного распределения, и критерии нормальности, проверяющие гипотезы о том, что выборка взята из нормального распределения. Во-вторых – критерии однородности, проверяющие гипотезы о том, что две выборки взяты из одного и того же распределения. В третьих, мы обсудим непараметрические аналоги коэффициента корреляции.

В конце будет рассмотрен непараметрический аналог факторного анализа.

9.1 Критерии случайности

Критериями случайности называют критерии проверки того, что заданная выборка x_1, \dots, x_n есть последовательность независимых наблюдений одной и той же случайной величины.

9.1.1 Критерий серий

Предположим, что выборка x_1, \dots, x_n случайна, то есть представляет собой независимую выборку из распределения, задаваемого функцией распределения F . Пусть M – медиана этого распределения, то есть такое число, что $F(M) = \frac{1}{2}$. Интуитивно кажется ясным, что при последовательном переборе элементов выборки мы не должны слишком долго задерживаться по одну сторону от M . Критерий серий основан на этом наблюдении.

Чтобы его сформулировать точно, введём такие обозначения. Пусть \bar{M} – выборочная медиана. Определяется она следующим образом. Расположим элементы выборки в порядке возрастания (т.е. образуем вариационный ряд): $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, тогда

$$\bar{M} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{если } n \text{ нечетно,} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{если } n \text{ четно.} \end{cases}$$

Теперь пройдемся по всем элементам выборки x_1, \dots, x_n и сопоставим x_i знак $+$, если $x_i > \bar{M}$ и знак $-$, если $x_i < \bar{M}$. Последовательность идущих подряд плюсов назовём положительной серией, а последовательность идущих подряд минусов – отрицательной серией. Пусть n_+ – количество плюсов, n_- – количество минусов, а z – общее количество серий.

Если верна гипотеза о случайности, то

$$E(z) = \frac{n_+ n_-}{n} + 1,$$
$$V(z) = \frac{2n_+ n_- (2n_+ n_- - n)}{n^2 (n - 1)}.$$

Величина же

$$z^* = \frac{z - E(z)}{\sqrt{V(z)}}$$

имеет приблизительно стандартное нормальное распределение.

Таким образом, гипотеза о случайности должна быть отвергнута на уровне значимости α , если величина $|z^*| > z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Пусть имеется выборка

0,07 0,15 0,74 0,75 0,62 0,06 0,66 0,97 0,69 0,49

1. Упорядочив элементы выборки по возрастанию, построим вариационный ряд

0,06 0,07 0,15 0,49 0,62 0,66 0,69 0,74 0,75 0,97

2. Найдем выборочную медиану по формуле $\bar{M} = \frac{1}{2}(0,62 + 0,66) = 0,64$.

3. Построим по выборке последовательность знаков

— — + + — — + + + —

4. Число плюсов n_+ равно 5, число минусов n_- равно 5, число серий $z = 5$. Используя это, вычисляем

$$\frac{n_+n_-}{n} + 1 = \frac{5 \cdot 5}{10} + 1 = 3,5$$

$$\frac{2n_+n_-(2n_+n_- - n)}{n^2(n-1)} = \frac{2 \cdot 5 \cdot 5 \cdot (2 \cdot 5 \cdot 5 - 10)}{10^2 \cdot 9} \approx 2,22$$

5. Таким образом,

$$z^* = \frac{5 - 3.5}{\sqrt{2.22}} \approx 1.006.$$

6. Если взять $\alpha = 0.05$, то $z_{0.975} = 1.96$. Так как $1.006 < 1.96$, то гипотеза о случайности не отвергается.

Задачи

Проверьте гипотезу случайности для выборок

0,07 0,12 0,72 0,75 0,31 0,73 0,46 0,71 0,46

0,88 0,67 0,28 0,30 0,57 1,20 0,51 0,92 1,05 1,59

9.1.2 Критерий поворотных точек

Построим по выборке x_1, \dots, x_n график. Назовём поворотной точкой локальные минимумы и максимумы этого графика. Идея теста заключается в том, что в случае, когда имеет место гипотеза случайности, поворотных точек не должно быть слишком мало или слишком много. Дадим точные формулировки.

Итак, пусть z – количество элементов выборки x_s , для которых либо $x_{s-1} < x_s > x_{s+1}$ или $x_{s-1} > x_s < x_{s+1}$. Если верна гипотеза о случайности, то

$$E(z) = (n-2)\frac{2}{3}, \quad V(z) = \frac{16n-29}{90},$$

и распределение z близко к нормальному. Положим

$$z^* = \frac{z - (n-2)\frac{2}{3}}{\sqrt{\frac{16n-29}{90}}}.$$

Гипотеза о случайности должна быть отвергнута на уровне значимости α , если величина $|z^*| > z_{1-\frac{\alpha}{2}}$, где $z_{1-\alpha/2}$ есть $1 - \alpha/2$ -квантиль стандартного нормального распределения.

Пример

Рассмотрим выборку

76 6 34 42 36 50 51 39 18 99 93 48 74 39 64 87 17 23 35 15

1. Выделим поворотные точки

76 **6** 34 **42** **36** 50 **51** 39 **18** **99** 93 **48** **74** **39** 64 **87** **17** 23 **35** 15

2. Таким образом, $z = 12$, $z^* = \frac{12 - \frac{36}{3}}{\sqrt{\frac{32-29}{90}}} = 0$.

3. Статистика z^* меньше по модулю, чем $z_{0.975} = 1,96$, поэтому гипотеза о случайности не отвергается.

Задачи

Проверьте гипотезу о случайности методом поворотных точек для выборок

76 6 34 42 36 50 50 39 18 99 93 48 74 39 64 87 17 23 35 15

5 11 12 6 8 7 10 14 15 16 17 16 14 83 14 11 10 9 10 15

9.1.3 Критерий Кендалла

Рассмотрим всевозможные пары элементов x_i, x_j из выборки x_1, \dots, x_n . Идея теста заключается в том, что если гипотеза об однородности имеет место, то примерно в половине случаев будем иметь $x_i < x_j$, а в другой половине будем иметь $x_i > x_j$.

Точнее, пусть z – количество пар i, j , где $i < j$, таких, что $x_i < x_j$. Если верна гипотеза об однородности, то эта величина имеет приблизительно нормальное распределение со средним

$$E(z) = \frac{n(n-1)}{4},$$

построим величину τ , называемую τ -Кендалла

$$\tau = 1 - \frac{4z}{n(n-1)},$$

тогда

$$V(\tau) = \frac{2(2n+5)}{n(n-1)},$$

так что величина

$$\tau^* = \frac{\tau}{\sqrt{\frac{2(2n+5)}{n(n-1)}}}$$

имеет приблизительно нормальное распределение.

Таким образом, гипотеза о случайности должна быть отвергнута на уровне значимости α , если $|\tau^*| > z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Дана выборка

4 9 10 2 3 1 5 8 11 7 14 13

1. Количество пар i, j , $i < j$, для которых $x_i < x_j$, равно 47.
2. Значит, $\tau = 1 - \frac{188}{132} \simeq 0,42$, тогда $\tau^* = \frac{0,42}{0,43} \simeq 0,63$.
3. Вычисленная статистика τ^* меньше по модулю, чем $z_{0,975} = 1,96$, поэтому гипотеза об однородности не отвергается.

Задачи

Проверить гипотезу об однородности для выборок

5 10 11 2 3 1 4 8 9 19 18 13

14 15 16 17 18 13 83 14 11 10 9 8 12

9.2 Критерии согласия

Критерии согласия проверяют гипотезу о том, что выборка имеет данное распределение или распределение из данного параметрического класса.

9.2.1 Критерий Пирсона χ^2

Рассмотрим сначала ситуацию, когда проверяется гипотеза о том, что выборка имеет заданное распределение с функцией $F(x)$.

Зафиксируем целое положительное число k . Разобьём числовую прямую на k промежутков

$$\mathbb{R} = (-\infty, a_1) \cup [a_1, a_2) \cup \dots \cup [a_{k-1}, \infty).$$

Пусть в первый промежуток попало m_1 элементов выборки, во второй – m_2 и т.д.

Рассчитаем теоретические вероятности попадания в эти промежутки

$$p_i = F(a_i) - F(a_{i-1}).$$

После этого составим статистику

$$X = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}.$$

Если выполнена нулевая гипотеза, то эта статистика имеет распределение, близкое к χ_{n-k}^2 , таким образом, нулевая гипотеза должна быть отвергнута, если $X > \chi_{n-k}^{2, 1-\alpha}$, где $\chi_{n-k}^{2, 1-\alpha}$ есть $(1 - \alpha)$ -квантиль распределения χ_{n-k}^2 .

Теперь рассмотрим ситуацию, когда проверяется гипотеза о том, что выборка имеет распределение с функцией $F(x, \theta_1, \dots, \theta_k)$, зависящей от неизвестных параметров $\theta_1, \dots, \theta_p$.

Тогда прежде всего получают асимптотически нормальные оценки $\hat{\theta}_1, \dots, \hat{\theta}_p$, а затем вычисляются оценочные теоретические вероятности попадания в промежутки

$$\hat{p}_i = F(a_i, \hat{\theta}_1, \dots, \hat{\theta}_p) - F(a_{i-1}, \hat{\theta}_1, \dots, \hat{\theta}_p).$$

Составляется статистика

$$X = \sum_{i=1}^k \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i}.$$

Отличие от предыдущей ситуации состоит в том, что эта статистика имеет при нулевой гипотезе распределение χ_{n-k-p}^2 . Соответственно, нулевая гипотеза должна быть отвергнута, если $X > \chi_{n-k-p}^{2, 1-\alpha}$.

Пример 1

Рассмотрим выборку

$$0.79, 0.21, 0.58, 0.02, 0.63, 0.24, 0.13, 0.46, 0.28, 0.48, 0.84, 0.15, 0.45, 0.79, 0.40.$$

Проверим гипотезу о том, что эта выборка имеет равномерное распределение на отрезке $[0, 1]$.

1. Зафиксируем число $k = 5$, разобьём числовую ось на 5 промежутков

$$(-\infty, 0, 2) \sqcup [0, 2, 0, 4) \sqcup [0, 4, 0, 6) \sqcup [0, 6, 0, 8) \sqcup [0, 8, \infty).$$

2. Вычислим количество элементов выборки, попадающих в каждый из этих промежутков

$$m_1 = 3, \quad m_2 = 3, \quad m_3 = 5, \quad m_4 = 3, \quad m_5 = 1,$$

рассчитаем также теоретические вероятности попадания в эти промежутки, в нашем случае они все равны $\frac{1}{5}$.

3. Вычислим статистику

$$X = \frac{(3 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(3 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(5 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(3 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(1 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} = \frac{8}{3}$$

4. Вычисленное значение статистики меньше критического значения $\chi_{15-5}^{2,0,95} = 18,3$. Так что гипотеза о том, что выборка взята из равномерного на отрезке $[0, 1]$ распределения, не отвергается.

Пример 2

Рассмотрим теперь выборку

$$0.31, 0.49, 0.36, 0.80, 0.61, 0.03, 0.71, 0.01, 0.54, 0.79, 0.59, 0.84, 0.65, 0.62, 0.14.$$

Проверим гипотезу о том, что эта выборка имеет равномерное распределение.

1. Прежде всего найдем оценки параметров этого распределения. Для этого воспользуемся методом моментов, он приводит к формулам

$$\hat{a} = \bar{x} - \sqrt{3}s = 0,03, \quad \hat{b} = \bar{x} + \sqrt{3}s = 0,97.$$

2. Зафиксируем число $k = 5$, разобьём числовую ось на 5 промежутков, для удобства сделаем это так, чтобы промежуток $[0,03, 0,97]$ разбился на 5 равных промежутков

$$(-\infty, 0, 22) \sqcup [0, 22, 0, 40) \sqcup [0, 40, 0, 59) \sqcup [0, 59, 0, 78) \sqcup [0, 78, \infty).$$

3. Вычислим количество элементов выборки, попадающих в каждый из этих промежутков

$$m_1 = 3, \quad m_2 = 2, \quad m_3 = 2, \quad m_4 = 4, \quad m_5 = 4,$$

рассчитаем также теоретические вероятности попадания в эти промежутки, в нашем случае они все равны $\frac{1}{5}$.

4. Вычислим статистику

$$X = \frac{(3 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(2 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(2 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(4 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} + \frac{(4 - 15 \cdot \frac{1}{5})^2}{15 \cdot \frac{1}{5}} = \frac{4}{3}$$

5. Вычисленное значение статистики меньше критического значения $\chi_{15-5-2}^{2,0,95} = 15,5$. Так что гипотеза о том, что выборка взята из равномерного на отрезке $[0, 1]$ распределения не отвергается.

Задача 1

Проверьте гипотезу о том, что выборка

0.99, 0.01, 0.41, 0.93, 0.00, 0.16, 0.14, 0.72, 0.91, 0.29, 0.41, 0.12, 0.10, 0.79, 0.80.

имеет показательное распределение с параметром $\lambda = 1$.

Задача 2

Проверьте гипотезу о том, что выборка

0.14, 0.01, -0.64, 0.47, 0.16, -0.25, -0.26, 0.35, 0.59, 0.22, -0.58, 0.72, -0.26, -0.10, 0.80.

имеет нормальное распределение.

9.2.2 Критерий Колмогорова–Смирнова

Данный критерий также является критерием согласия, применяется он только в случае, когда распределение, с которым сравнивается выборка, задаётся функцией $F(x)$, которая не содержит неизвестных параметров.

Пусть дана выборка x_1, \dots, x_n , определим эмпирическую функцию распределения

$$F_n(x) = \frac{1}{n} \# \{x_i : x_i < x\},$$

где $\#$ - обозначает количество элементов в множестве. Эта функция кусочно постоянна, в точках x_i у нее разрывы, в которых значение подскакивает на $\frac{1}{n}$. Значение в точке $x_{(i)}$ (i -ый член вариационного ряда) справа равно $\frac{i}{n}$, а слева $-\frac{i-1}{n}$.

Статистика Колмогорова–Смирнова имеет вид

$$K_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

так как функция $F(x)$ возрастает, а $F_n(x)$ кусочно постоянна, то точная верхняя грань разности достигается в точках разрыва x_i , поэтому

$$K_n = \max_{i=1, \dots, n} \{|\frac{i}{n} - F(x_{(i)})|, |\frac{i-1}{n} - F(x_{(i)})|\}.$$

Если выполнена нулевая гипотеза, то распределение K_n не зависит от функции $F(x)$. Нулевая гипотеза должна быть отвергнута, если $K_n > K_{n,\text{критич}}$, 5-процентные критические точки для малых n заданы в таблице

n	$\alpha = 0.05$
1	0.98
2	0.84
3	0.71
4	0.62
5	0.56
6	0.52
7	0.48
8	0.45
9	0.43
10	0.41
11	0.39
12	0.38
13	0.36
14	0.35
15	0.34
16	0.33
17	0.32
18	0.31
19	0.30
20	0.29

Если же n велико, то используется тот факт, что случайная величина $\sqrt{n}K_n$ имеет предел при $n \rightarrow \infty$. Поэтому гипотеза должна быть отвергнута, если $\sqrt{n}K_n > Q_{\text{критич}}$. Пятипроцентная критическая точка распределения Q равна 1,36.

Пример

Рассмотрим выборку объема $n = 10$.

0,05 0,72 0,73 0,40 0,51 0,12 0,87 0,47 0,10 0,68

Проверим гипотезу о том, что это есть выборка из равномерного на отрезке $[0, 1]$ распределения.

1. Располагаем наблюдения в порядке возрастания

0,05 0,10 0,12 0,40 0,47 0,51 0,68 0,72 0,73 0,87

2. Для каждого элемента построенного вариационного ряда вычисляем выражения $\{|\frac{i}{n} - F(x_{(i)})|, |\frac{i-1}{n} - F(x_{(i)})|\}$:

$\{0,05, 0,05\}$ $\{0,10, 0,00\}$ $\{0,18, 0,08\}$ $\{0,00, 0,10\}$ $\{0,03, 0,07\}$
 $\{0,09, 0,01\}$ $\{0,02, 0,08\}$ $\{0,08, 0,02\}$ $\{0,17, 0,07\}$ $\{0,13, 0,03\}$

3. Вычисляем максимальное из полученных чисел $K_{10} = 0,18$.

4. Так как $K_{10} < 0,41$, то гипотеза о равномерном на отрезке $[0, 1]$ распределении не отвергается.

Задача 1

Дана выборка:

0.83, −0.15, 0.84, 0.09, −0.49, −0.06, 0.83, 0.58, 0.37, −0.66, −0.25, −0.93, 0.70, 0.10, 0.91.

Проверьте, что данная выборка имеет стандартное распределение.

9.3 Критерии нормальности

Гипотеза о том, что данная выборка имеет нормальное распределение, может быть проверена с помощью критерия χ^2 в варианте, где проверяется принадлежность распределения параметрическому семейству.

Обсудим теперь специфические критерии, тестирующие гипотезу о том, что данная выборка имеет нормальное распределение.

9.3.1 Критерий Лиллиефорса

Данный критерий представляет собой модификацию критерия Колмогорова–Смирнова.

Для проверки гипотезы о нормальности прежде всего оценим параметры распределения – математическое ожидание \hat{a} и дисперсию $\hat{\sigma}^2$. После этого вычислим статистику L по формуле, аналогичной (9.2.2), однако при этом в качестве $F(x)$ берём функцию распределения для нормального распределения, в которую подставлены оценённые параметры \hat{a} и $\hat{\sigma}^2$.

Из-за того, что используется не настоящая функция распределения, а функция, в которую подставлены неизвестные параметры, распределение статистики K_n другое, нежели распределение Колмогорова–Смирнова. Соответственно, меняются критические точки.

Таким образом, если $n < 30$, то нулевая гипотеза должна быть отвергнута, если $K_n > L_{n,\text{критич}}$. Пятипроцентные критические точки для малых n заданы в таблице

n	$\alpha = 0.05$
4	0.381
5	0.337
6	0.319
7	0.300
8	0.285
9	0.271
10	0.258
11	0.230
12	0.242
13	0.234
14	0.227
15	0.220
16	0.213
17	0.206
18	0.200
19	0.195
20	0.190
25	0.180
30	0.161

Если же n велико, то используется тот факт, что случайная величина $\sqrt{n}K_n$ имеет предел при $n \rightarrow \infty$. Поэтому гипотеза должна быть отвергнута, если $\sqrt{n}K_n > L_{\text{критич}}$. Пятипроцентная критическая точка распределения L равна 0,886.

Пример

Рассмотрим выборку

1,47 -0,80 -3,34 -0,72 2,73 2,81 1,81 0,75 -5,00 4,49

1. Упорядочим выборку по возрастанию

-5,00 -3,34 -0,80 -0,72 0,75 1,47 1,81 2,73 2,81 4,49

2. Оценим параметры нормального распределения, получим

$$\hat{a} = \bar{x} = 0,42, \quad \hat{\sigma} = 8,5.$$

3. По формуле

$$\hat{F}(x_{(i)}) = \Phi\left(\frac{x_{(i)} - \hat{a}}{\hat{\sigma}}\right)$$

вычислим значение оценённой теоретической функции распределения, получим:

0,26 0,33 0,44 0,44 0,51 0,54 0,56 0,60 0,60 0,68

4. Теперь найдём выражения $\{|\frac{i}{n} - \hat{F}(x_{(i)})|, |\frac{i-1}{n} - \hat{F}(x_{(i)})|\}$:

$$\{0.26; 0.16\}, \{0.23; 0.13\}, \{0.24; 0.14\}, \{0.14; 0.04\}, \{0.11; 0.01\}, \\ \{0.04; 0.05\}, \{0.03; 0.13\}, \{0.09; 0.19\}, \{0.19; 0.29\}, \{0.21; 0.31\}.$$

5. Максимальное из найденных чисел есть 0,31. Пятипроцентная критическая точка распределения Лиллиефорса для $n = 10$ есть 0,258. Так как $0,31 > 0,258$, то гипотеза о нормальности отвергается.

Задача

Проверьте, что выборка

$$0,09 \quad 5,10 \quad 5,61 \quad -0,53 \quad 1,07 \quad 2,02 \quad 5,54 \quad 2,73 \quad 3,81 \quad 1,69$$

распределена нормально.

9.3.2 Критерий Андерсона-Дарлинга

Данный критерий¹ считается весьма мощным и может быть использован на выборках малых объёмов. Опять в качестве $F(x)$ берём функцию распределения для нормального распределения, в которую подставлены оценённые параметры \hat{a} и $\hat{\sigma}^2$.

Определим величину

$$A^2 = g \int_{-\infty}^{+\infty} \frac{1}{F(x)(1-F(x))} |F_n(x) - F(x)|^2 dF(x).$$

Она может быть вычислена по следующей формуле. Определим

$$y_i = \frac{x_{(i)} - \bar{x}}{s},$$

тогда оказывается, что

$$A^2 = -n - \sum_{k=1}^n \frac{2k-1}{n} (\ln \Phi(y_k) + \ln(1 - \Phi(y_{n+1-k}))).$$

В критерии используется не сама величина A^2 , а величина A_*^2 , скорректированная на объём выборки

$$A_*^2 = A^2 \left(1 + \frac{0,75}{n} + \frac{2,25}{n^2} \right).$$

Нулевая гипотеза должна быть отвергнута на 5-процентном уровне значимости, если $A_*^2 > 0,752$.

¹Существует вариант этого теста, проверяющий гипотезу о том, что выборка имеет заданное распределение.

Пример

Возьмём выборку

0,30 0,75 0,24 0,21 0,75 0,67 0,10 0,56 0,14 0,46

1. Оценим параметры нормального распределения, получим

$$\hat{a} = \bar{x} = 0,42, \quad \hat{s} = 0,25.$$

2. Вычислим величины $y_i \frac{x_{(i)} - \bar{x}}{s}$

-1,27 -1,08 -0,81 -0,71 -0,46 0,17 0,55 0,99 1,30 1,32

3. Теперь вычислим величину $A^2 = -n - \sum_{k=1}^n \frac{2k-1}{n} (\ln \Phi(y_k) + \ln(1 - \Phi(y_{n+1-k}))) = 0,38$.
4. Вычислим $A_*^2 = A^2 \left(1 + \frac{0,75}{n} + \frac{2,25}{n^2}\right) = 0,42$.
5. Как видно, $0,42 < 0,752$, поэтому гипотеза о нормальности не отвергается.

Задача

Проверьте на нормальность выборку

0,20 0,76 0,84 0,61 0,36 0,98 0,63 0,25 0,62 0,63

9.3.3 Критерий Харке–Бэра

Критерий основан на том, что у нормального распределения выборочный куртуозис равен 0, а выборочный эксцесс равен 3.

Для проверки гипотезы о нормальности прежде всего вычисляются выборочные центральные моменты m_3 и m_4 , где

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k,$$

а затем вычисляется выборочный куртуозис

$$\hat{\kappa} = \frac{m_4}{m_2^2}$$

и выборочный эксцесс

$$\hat{\gamma} = \frac{m_3}{m_2^{\frac{3}{2}}}.$$

Статистика Харке–Бэра имеет вид

$$J = n \left(\frac{\hat{\gamma}^2}{6} + \frac{(\hat{\kappa} - 3)^2}{24} \right),$$

если выполнена гипотеза о нормальности, то она имеет распределение, близкое к χ_2^2 , поэтому гипотеза о нормальности должна быть отвергнута, если $J > \chi_2^{2,1-\alpha}$. Здесь $\chi_2^{2,1-\alpha}$ есть $1 - \alpha$ квантиль распределения χ_2^2 .

Пример

Дана выборка

0,20 0,94 0,39 0,12 0,15 0,57 0,33 0,91 0,31 0,84

1. Её выборочный эксцесс равен $-1,48$, а куртуозис $0,52$.
2. Статистика Харке-Бэра равна таким образом

$$JB = 0,62$$

Критическая точка равна $7,81$, поэтому гипотеза о нормальности не отвергается.

Задача

Проверить на нормальность выборку

0,45 0,81 0,22 0,23 0,72 0,30 0,13 0,73 0,85 0,93

9.4 Критерии однородности

Критерии однородности проверяют гипотезу о том, что данные выборки представляют собой выборки из одного и того же распределения.

9.4.1 Критерий знаков

Пусть даны выборки x_1, \dots, x_n и y_1, \dots, y_n одного объёма. Проверим гипотезу о том, что они представляют собой две выборки из одного и того же распределения.

Для проверки образуем разности $s_1 = x_1 - y_1, \dots, s_n = x_n - y_n$ и каждой разности сопоставляем знак $+$, если она положительна, и $-$, если она отрицательна. Разности, равные нулю, просто отбрасываем. Идея теста состоит в том, что если выполнено предположение об однородности, то число минусов и плюсов должно быть примерно одинаково.

Рассмотрим величину

$$z = \min\{\text{количество } +, \text{ количество } -\},$$

В случае, когда n мало (скажем, что n мало, если $n < 25$) эта величина непосредственно используется как статистика критерия. Гипотеза об однородности отвергается, если $z < z_{\text{критич.}}$. Таблица 5% и 10% критических точек в данном случае следующая:

n	$\alpha = 0.05$	$\alpha = 0.1$
9	1	1
10	1	1
11	1	2
12	2	2
13	2	3
14	2	3
15	3	3
16	3	4
17	4	4
18	4	5
19	4	5
20	5	5
25	7	7
30	9	10

В случае, когда n велико (скажем, что n велико, если $n \geq 25$), статистика z имеет приблизительно нормальное распределение с математическим ожиданием и дисперсией

$$E(z) = \frac{n-1}{2}, \quad V(z) = \frac{n}{4},$$

таким образом, разумно в качестве статистики критерия взять величину

$$z^* = \frac{z - \frac{n-1}{2}}{\sqrt{\frac{n}{4}}}.$$

Так как при построении величины z брался минимум из числа плюсов и минусов и нулевая гипотеза отвергалась при малом значении z , то разумно сформулировать правило принятия решения так. Гипотеза об однородности отвергается на уровне значимости α , если величина $z^* < -z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Имеются выборки

0,01 0,78 0,17 0,50 0,78 0,38 0,22 0,24 0,12 0,14

0,91 0,42 0,93 0,72 0,62 0,11 0,99 0,67 0,42 0,96

Проверим, что они взяты из одного и того же распределения.

1. Образует выборку из разностей

$-0,89 \quad 0,35 \quad -0,75 \quad -0,22 \quad 0,15 \quad 0,27 \quad -0,77 \quad -0,43 \quad -0,30 \quad -0,81$

Таким образом, последовательность знаков будет такой

$- \quad + \quad - \quad - \quad + \quad + \quad - \quad - \quad - \quad -$

2. Вычисляем статистику $z = \min\{3, 7\} = 3$.
3. Объём выборки $n = 10$ мал, так что используем первый вариант теста. Так как $z > z_{\text{критич}} = 1$, то нулевая гипотеза не отвергается, так что делаем вывод, что выборки взяты из одного и того же распределения.

Задача 1

Проверьте гипотезу о том, что две следующие выборки однородны

$0,20 \quad 1,72 \quad 0,27 \quad 0,88 \quad 1,51 \quad 0,68 \quad 1,86 \quad 0,59 \quad 1,12 \quad 0,30$

$0,89 \quad 0,04 \quad 0,75 \quad 0,88 \quad 0,24 \quad 0,03 \quad 0,47 \quad 0,60 \quad 0,37 \quad 0,77$

Задача 2

До реорганизации доход предприятия в течение 10 месяцев был равен следующим величинам

$0,32 \quad 0,12 \quad 0,58 \quad 0,99 \quad 0,77 \quad 0,35 \quad 0,65 \quad 0,28 \quad 0,17 \quad 0,67$

А в следующие 10 месяцев - величинам

$0,05 \quad 0,52 \quad 0,94 \quad 0,54 \quad 0,34 \quad 0,41 \quad 0,82 \quad 0,36 \quad 0,50 \quad 0,27$

Проверить гипотезу о том, что реорганизация изменила доходность.

9.4.2 Критерий знаков для проверки гипотез о медиане

Критерий знаков может быть использован для проверки гипотезы о значении медианы. Именно нулевая гипотеза состоит в том, что медиана распределения равна M .

Сопоставим каждому элементу выборки $+$ или $-$ в зависимости от того, что имеет место $x_i > M$ или $x_i < M$ (если случилось так, что $x_i = M$, то наблюдение отбрасывается). Далее, так же как и раньше, составляется статистика z при $n < 25$ или статистика z^* при $n \geq 25$, после чего проверка гипотез ведётся так же, как раньше.

Пример

Дана выборка

0,90 0,94 0,63 0,37 0,32 0,56 0,99 0,44 0,06 0,07

Проверим гипотезу о том, что медиана соответствующего распределения равна 0,5.

1. Составим последовательность знаков

+ + + - - + + - - -

2. Вычисляем статистику $z = \min\{5, 5\} = 5$.

3. Так как $z > z_{\text{критич}} = 1$, то нулевая гипотеза не отвергается, так что делаем вывод, что медиана равна 0,5.

Задача

Из 10 человек 6 собираются идти на выборы. Проверьте гипотезу о том, что большинство пойдёт на выборы.

9.4.3 Критерий знаков для проверки гипотез о вероятности успеха

Также критерий знаков может быть использован для проверки гипотезы о вероятности успеха в испытании Бернулли. Именно, проверяется гипотеза о том, что вероятность успеха равна $p = 0.5$. Сопоставим + тем элементам выборки, для которых соответствующее испытание Бернулли приводит к удачному исходу и – тем, где к неудачному. Так же как и раньше строится статистика z .

Тогда можно составить статистику

$$z^* = \frac{z - \frac{n}{2} + \frac{1}{2}}{\sqrt{\frac{n}{4}}},$$

при $n \geq 25$ можно считать, что она приблизительно имеет нормальное распределение.

Как и раньше, гипотеза отвергается на уровне значимости α , если $z^* < -z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1-\frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Рассмотрим последовательность из 25 испытаний, пусть результаты этих испытаний следующие (пишем 1 в случае успеха и 0 в случае неудачи)

1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

Проверим гипотезу о том, что вероятность успеха равна 0.5.

1. Вычислим статистику

$$z = \min\{\text{количество } 0, \text{ количество } 1\} = \min\{6, 19\} = 6.$$

2. Вычислим статистику

$$z^* = \frac{12 - 25 + 1}{\sqrt{25}} = -2.4.$$

3. Так как $z^* < -z_{0.975} = -1.96$, то гипотеза о том, что вероятность успеха равна 0.5, отвергается.

Задача 1

Проверьте, что вероятность успеха равна 0.7, если результат 20 испытаний следующий

1 0 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 0 1 1 1

Задача 2

Проверьте, что вероятность успеха равна 0.4, если результат 20 испытаний следующий

1 0 1 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 1 0 1

9.4.4 Критерий рангов

Как и критерий знаков, этот критерий проверяет гипотезу однородности.

Пусть даны выборки одинакового объема x_1, \dots, x_n и y_1, \dots, y_n , составим разности $s_1 = x_1 - y_1, \dots, s_n = x_n - y_n$. Разности, равные нулю, просто отбрасываем.

Ранжируем разности

$$s_{i_1} \leq s_{i_2} \leq \dots \leq s_{i_n},$$

те индексы i , для которых $s_i < 0$, назовём отрицательными рангами, а те индексы i , для которых $s_i > 0$, назовём положительными рангами.

Положим

$$z = \min\{\text{сумма отрицательных рангов, сумма положительных рангов}\}.$$

Если n мало (считаем, что n мало, если $n \leq 30$), то гипотеза об однородности отвергается, если $z < z_{\text{критич.}}$. Таблица 5% и 10%-критических точек в данном случае следующая:

n	$\alpha = 0.05$	$\alpha = 0.1$
9	8	3
10	11	5
11	14	7
12	17	10
13	21	13
14	26	16
15	30	20
16	36	24
17	41	28
18	47	33
19	54	38
20	60	43
25	101	77
30	125	120

Если же n достаточно большое (считаем, что n мало, если $n > 30$), то статистика z имеем приблизительно нормальное распределение с математическим ожиданием и дисперсией

$$E(z) = \frac{n(n+1)}{4}, \quad V(z) = \frac{n(n+1)(2n+1)}{24},$$

таким образом, разумно в качестве статистики критерия взять величину

$$z^* = \frac{z - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

Гипотеза об однородности отвергается на уровне значимости α , если величина $z^* < -z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Проверим гипотезу об однородности для выборок

0,45 0,28 0,14 0,44 0,74 0,51 0,09 0,25 0,10 0,03

0,64 0,63 0,91 0,09 0,27 0,52 0,52 0,34 0,59 0,34

1. Составляем ряд разностей

$$s_1 = -0,18 \quad s_2 = -0,35 \quad s_3 = -0,77 \quad s_4 = 0,34 \quad s_5 = 0,47$$

$$s_6 = -0,01 \quad s_7 = -0,42 \quad s_8 = -0,09 \quad s_9 = -0,48 \quad s_{10} = -0,31$$

2. Ранжируем разности

$$\begin{aligned}s_3 &= -0,77 & s_9 &= -0,48 & s_7 &= -0,42 & s_2 &= -0,35 & s_{10} &= -0,31 \\s_1 &= -0,18 & s_8 &= -0,09 & s_6 &= -0,01 & s_4 &= 0,34 & s_5 &= 0,47\end{aligned}$$

3. Сумма положительных рангов равна $3 + 9 + 7 + 2 + 10 + 1 + 8 + 6 = 46$, сумма отрицательных равна $4 + 5 = 9$, таким образом $z = 9$.
4. Так как $z < z_{\text{критич.}} = 11$, то гипотеза об однородности не отвергается.

Задача 1

Проверить гипотезу однородности для выборок

$$0,11 \quad 0,92 \quad 0,13 \quad 0,73 \quad 0,43 \quad 0,79 \quad 0,25 \quad 0,25 \quad 0,26 \quad 0,06$$

$$0,24 \quad 0,77 \quad 0,62 \quad 0,17 \quad 0,16 \quad 0,74 \quad 0,65 \quad 0,97 \quad 0,45 \quad 0,82$$

Задача 2

Цена одного квадратного метра в 10 квартирах в первом городе равна

$$0,69 \quad 0,02 \quad 0,89 \quad 0,48 \quad 0,33 \quad 0,12 \quad 0,26 \quad 0,01 \quad 0,36 \quad 0,03$$

а во втором равна

$$0,03 \quad 0,64 \quad 0,89 \quad 0,57 \quad 0,41 \quad 0,41 \quad 0,79 \quad 0,57 \quad 0,73 \quad 0,69$$

Проверьте гипотезу о том, что цены на жильё в двух городах одинаковы.

9.4.5 Критерий Манна–Уитни

Пусть даны две выборки возможно разных объемов x_1, \dots, x_n и y_1, \dots, y_m . Мы перебираем всевозможные пары x_i, y_j . Пусть z_+ – число пар, для которых $x_i > y_j$, а z_- – число пар, для которых $x_i < y_j$. При этом если имеется пара, для которой $x_i = y_j$, то мы увеличиваем и z_+ , и z_- на $\frac{1}{2}$.

Положим

$$z = \min\{z_-, z_+\}.$$

Тогда гипотеза об однородности не отвергается, если $z > z_{\text{критич.}}$.

Таблица 5-процентных критических точек Манна–Уитни следующая.

$m \setminus n$	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Пример

Рассмотрим выборки

0,32 0,96 0,58

0,28 0,24 0,67 0,39 0,00 0,41 0,08

Проверим гипотезу об однородности.

1. Перечислим все пары значений, отметив знаком $+$, те, где $x_i > y_j$, и минусом $-$ те пары, где $x_i < y_j$

$y \setminus x$	0,32	0,96	0,58
0,28	+	+	+
0,24	+	+	+
0,67	-	+	-
0,39	-	+	+
0,00	+	+	+
0,41	-	+	+
0,08	+	+	+

2. Таким образом получаем, что $z_+ = 17$, $z_- = 4$. Поэтому $z = \min\{17, 4\} = 4$.
3. Так как $z > z_{\text{критич}} = 1$, гипотеза об однородности не отвергается.

Задача 1

Проверить гипотезу об однородности для выборок

0,43 0,55 0,45 0,43 0,68 0,88 0,56

0,69 0,96 0,87

Задача 2

Цена за квадратный метр в объявлениях о продаже квартир в городе равна

0,85 0,96 0,95 0,03 0,33 0,63 0,12

а в поселке-спутнике - равна

0,51 0,63 0,76

Проверьте гипотезу о том, что цены на жильё в городе и поселке-спутнике одинаковы.

9.4.6 Критерий Вилкоксона

Проверяется гипотеза об однородности для двух выборок возможно разного объема x_1, \dots, x_n и y_1, \dots, y_m . Для проверки гипотезы выборки объединяются и в одну выборку s_1, \dots, s_{n+m} , после чего объединённая выборка ранжируется

$$s_{i_1} \leq \dots \leq s_{i_{n+m}}.$$

Подсчитываем сумму рангов (то есть индексов i_k), относящихся к первой выборке и ко второй выборке. Если $n + m$ мало (меньше 10), то положим

W = минимальная из двух сумм рангов.

Гипотеза об однородности не отвергается, если значение статистики W больше критического, таблица 5-процентных критических точек следующая

$n + m$	
5	0
6	2
7	3
8	5
9	8
10	10

Если же $n + m$ большое (больше 10), то статистика z имеет приблизительно нормальное распределение с параметрами

$$E(W) = \frac{n(n+m+1)}{2}, \quad V(W) = \frac{nm(n+m+1)}{12}.$$

Тогда

$$z^* = \frac{W - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

имеет приблизительно стандартное нормальное распределение.

Гипотеза отвергается на уровне значимости α , если величина $z^* < -z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Возьмём две выборки

0.12, 0.52, 0.45 и 0.68, 0.85, 0.83, 0.07, 0.73, 0.03, 0.65.

Проверим гипотезу об однородности для $\alpha = 0.05$.

1. Строим объединённую выборку, выделим первую выборку

$\mathbf{x}_1 = \mathbf{0.12}, \mathbf{x}_2 = \mathbf{0.52}, \mathbf{x}_3 = \mathbf{0.45}, x_4 = 0.68, x_5 = 0.85,$
 $x_6 = 0.83, x_7 = 0.07, x_8 = 0.73, x_9 = 0.03, x_{10} = 0.65.$

2. Ранжируем объединённую выборку

$x_{(1)} = 0.03, x_{(2)} = 0.07, \mathbf{x_{(3)} = 0.12}, \mathbf{x_{(4)} = 0.45}, \mathbf{x_{(5)} = 0.52},$
 $x_{(6)} = 0.65, x_{(7)} = 0.68, x_{(8)} = 0.73, x_{(9)} = 0.85, x_{(10)} = 0.83.$

3. Найдём сумму рангов, относящуюся к первой подвыборке, получаем 12, а также ко второй подвыборке - 43. Таким образом, $W = 12$.

4. Так как $W > W_{\text{критич}} = 10$, то гипотеза об однородности не отвергается.

Задача 1

Проверьте гипотезу об однородности для выборок

0,37 0,01 0,54

и

0,73 0,36 0,35 0,07 0,20 0,31 0,10

Задача 2.

Доля замужних женщин на трех предприятиях

0.52, 0.42, 0.31,

а в 10 офисах

0.28 0.94 0.71 0.00 0.21 0.07 0.66

Проверьте гипотезу о том, что число замужних женщин имеет распределение, не зависящее от места работы.

9.5 Исследование взаимосвязей между выборками

Для двух выборок одного объёма может быть вычислен коэффициент корреляции, измеряющий силу линейной связи между двумя случайными величинами. Однако имеющаяся методика исследования его на значимость работает лишь в случае двух нормальных выборок.

Приводимые ниже коэффициенты Спирмена и Кендалла могут быть использованы и для исследования связи двух выборок, не имеющих нормальных распределений.

9.5.1 Коэффициент ранговой корреляции Спирмена

Пусть даны две выборки одного и того же объёма x_1, \dots, x_n и y_1, \dots, y_n , ранжируем их и положим

$$d_i := R_{x_i} - R_{y_i},$$

где R_{x_i} и R_{y_i} - ранги соответствующих наблюдений.

Определим **коэффициент ранговой корреляции Спирмена** формулой

$$r_S := 1 - \frac{6}{n^3 - n} \sum_{i=1}^n d_i^2.$$

Замечание. Коэффициент корреляции Спирмена принимает значения в отрезке $[-1, 1]$, значение 0 получается, если связь между выборками отсутствует. Единица получится в случае совпадения рангов выборок.

Проверка данного коэффициента на значимость осуществляется одним из следующих двух способов.

1. Если n мало (считаем, что n мало, если $n < 30$), то гипотеза о незначимости не отвергается, если $|r_S| < r_{\text{критич}}$, где 5-процентная критическая точка $r_{\text{критич}}$ находится с помощью таблицы

n	$\alpha = 0.05$	$\alpha = 0.1$
6	0,886	0,829
7	0,786	0,714
8	0,738	0,643
9	0,683	0,600
10	0,648	0,564
11	0,623	0,523
12	0,591	0,497
13	0,566	0,475
14	0,545	0,457
15	0,525	0,441
16	0,507	0,425
17	0,490	0,412
18	0,476	0,399
19	0,462	0,388
20	0,475	0,377
21	0,438	0,368
22	0,428	0,359
23	0,418	0,351
24	0,409	0,343
25	0,400	0,336
26	0,392	0,329
27	0,385	0,323
28	0,377	0,317
29	0,370	0,317
30	0,364	0,305

Если же n велико, то гипотеза о незначимости не отвергается, если $|r_S| < r_{\text{критич}}$. При этом 5-процентная критическая точка $r_{\text{критич}}$, которая находится теперь по формуле

$$r_{\text{критич}} = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}},$$

где $z_{1-\frac{\alpha}{2}}$ – квантиль стандартного нормального распределения.

2. Второй способ заключается в следующем. Составляется дробь

$$t = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}},$$

которая, если верна гипотеза о незначимости коэффициента Спирмена, имеет распределению Стьюдента t_{n-2} .

Соответственно, гипотеза о незначимости отвергается, если

$$|t| < t_{\text{критич}},$$

где $t_{\text{критич}} = t_{n-2}^{1-\frac{\alpha}{2}}$ есть $1-\frac{\alpha}{2}$ - квантиль распределения Стьюдента с $(n-2)$ степенями свободы.

Пример

Рассмотрим выборки 0.09, 0.42, 0.85, 0.59, 0.10, 0.23, 0.32 и 0.8, 0.18, 0.25, 0.29, 0.89, 0.93, 0.39.

Проверим гипотезу о том, что коэффициент корреляции Спирмена незначим ($\alpha = 0.05$).

1. Ранжируем две выборки и выписываем ранги

$$R_{x_1} = 1, R_{x_2} = 5, R_{x_3} = 7, R_{x_4} = 6, R_{x_5} = 2, R_{x_6} = 3, R_{x_7} = 4 \text{ и } R_{y_1} = 5, R_{y_2} = 1, R_{y_3} = 2, R_{y_4} = 3, R_{y_5} = 6, R_{y_6} = 7, R_{y_7} = 4$$

2. Образует ряд разностей рангов

$$d_1 = -4, d_2 = 4, d_3 = 5, d_4 = 3, d_5 = -4, d_6 = -4, d_7 = 0$$

3. Вычисляем коэффициент Спирмена

$$r_S = 1 - \frac{6}{7^3 - 7}(16 + 16 + 25 + 9 + 16 + 16 + 0) = -0.75$$

4. Вычисляем статистику критерия $t = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}} = -2.53$.

5. Поскольку $|t| < t_{\text{критич}} = 2.57$, гипотеза о незначимости не отвергается.

Задача 1

Проверьте, значим ли коэффициент корреляции Спирмена между выборками

$$0,35 \quad 0,39 \quad 0,37 \quad 0,22 \quad 0,85 \quad 0,05 \quad 0,87$$

$$0,12 \quad 0,24 \quad 0,32 \quad 0,56 \quad 0,87 \quad 0,30 \quad 0,89$$

Задача 2

Проверьте, что между расходами на рекламу, заданными в выборке

$$0,37 \quad 0,13 \quad 0,25 \quad 0,36 \quad 0,59 \quad 0,28 \quad 0,78,$$

и объёмом продаж, заданным в выборке

$$0,19 \quad 0,22 \quad 0,19 \quad 0,33 \quad 0,54 \quad 0,47 \quad 0,71,$$

есть связь.

9.5.2 Коэффициент Кендалла

Пусть даны выборки x_1, \dots, x_n и y_1, \dots, y_n одинаковых объёмов.

Пусть P – число пар (i, j) , таких, что $x_i < x_j$ и $y_i < y_j$, или $x_i > x_j$ и $y_i > y_j$.

Пусть I – число пар (i, j) , таких, что $x_i > x_j$ и $y_i < y_j$, или $x_i < x_j$ и $y_i > y_j$.

Образуем величину

$$\tau = \frac{P - I}{P + I}.$$

Можно привести и другие формулы для τ . Именно, так как $P + I = \frac{n(n-1)}{2}$ (общее число пар индексов), то

$$\tau = 1 - \frac{4I}{n(n-1)} = \frac{4P}{n(n-1)} - 1.$$

Как и обычный коэффициент корреляции, коэффициент Кендалла принимает значения в отрезке $[-1, 1]$, значение 0 получается, если связь между выборками отсутствует.

Для проверки гипотезы о равенстве коэффициента нулю

$$z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}.$$

Если верна гипотеза о равенстве коэффициента Кендалла нулю, то эта статистика имеет приблизительно стандартное нормальное распределение.

Таким образом, гипотеза о равенстве нулю отвергается, если $|z^*| > z_{1-\frac{\alpha}{2}}$, где $z_{1-\frac{\alpha}{2}}$ есть $1 - \frac{\alpha}{2}$ -квантиль стандартного нормального распределения.

Пример

Имеются две выборки

0.83 0.72 0.67 0.52 0.01

0.52 0.95 0.86 0.93 0.80

Проверим, есть ли связь между ними.

1. Составим таблицу. Поставим в ячейке, соответствующей паре (i, j) , где $i < j$, плюс, если $x_i < x_j$, и поставим – иначе.

$y \setminus x$	0,83	0,72	0,67	0,52	0,01
0,83		+	+	+	+
0,72			+	+	+
0,67				+	+
0,52					+
0,01					

2. Составим таблицу. Поставим в ячейке, соответствующей паре (i, j) , где $i < j$, плюс, если $y_i < y_j$, и поставим $-$ иначе.

$y \setminus x$	0,52	0,95	0,86	0,93	0,80
0,52		-	-	-	-
0,95			+	+	+
0,86				-	+
0,93					+
0,80					

3. Составим таблицу. Поставим в ячейке, соответствующей паре (i, j) , где $i < j$ плюс, если $x_i < x_j$ и $y_i < y_j$, или $x_i > x_j$ и $y_i > y_j$. И поставим $-$ иначе.

$y \setminus x$	1	2	3	4	5
1		-	-	-	-
2			+	+	+
3				-	+
4					+
5					

4. Число пар первого типа P равно 5, число пар второго типа Q равно 5.

5. Найдём коэффициент Кендалла $\tau = \frac{P-I}{P+I} = 0$

6. Вычислим статистику $z^* = \frac{0}{\sqrt{\frac{30}{180}}} = 0$.

7. Так как $|z^*| < z^{0,975} = 1,96$, то гипотеза о равенстве нулю коэффициента Кендалла не отвергается.

Задача 1

В двух выборках ниже приведены результаты экспертных оценок некоторой величины. Проверьте с помощью коэффициента Кендалла гипотезу о том, что экспертные мнения адекватны (между оценками есть связь).

0,73 0,82 0,69 0,59 0,91

0,55 0,98 0,76 0,73 0,08

Задача 2

В двух выборках приведены доли правильно выполненных заданий по математике и русскому языку. Проверьте с использованием коэффициента Кендалла, что между результатами есть связь.

0,73 0,82 0,69 0,59 0,91

0,74 0,72 0,60 0,50 0,61

9.6 Факторный анализ

В данном разделе речь пойдёт о двух критериях, которые являются непараметрическими аналогами дисперсионного анализа. Первый критерий является непараметрическим аналогом однофакторного анализа, а второй критерий - аналог двухфакторного анализа.

9.6.1 Однофакторный анализ. Критерий Краскелла-Уоллиса

Данный критерий представляет собой непараметрический аналог одномерному дисперсионному анализу. Он проверяет гипотезу о том, что все выборки имеют одинаковое распределение.

Пусть даны k независимых выборок объёмов n_1, \dots, n_k :

$$\begin{array}{c} x_{1,1}, \dots, x_{n_1,1} \\ \dots \\ x_{1,k}, \dots, x_{n_k,k} \end{array}$$

Проверяется гипотеза, что на распределение $x_{i,j}$ номер j влияния не оказывает. Таким образом, мы проверяем отсутствие влияния на распределение x_i^j одного фактора. Этим объясняется термин "однофакторный анализ".

Приводимый ниже критерий является аналогом параметрического F -критерия.

Для построения статистики критерия все выборки объединяются в выборку объёма $N = n_1 + \dots + n_k$ и ранжируются.

Пусть R_i - сумма рангов, относящихся к i -ой выборке. Также пусть $\bar{R}_i = \frac{1}{n_i} R_i$ - средний ранг в i -ой выборке, $\bar{\bar{R}} = \frac{N-1}{2}$ - средний ранг в объединённой выборке.

Определим

$$\begin{aligned} X &= \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{\bar{R}})^2 = \\ &= \frac{12}{N(N+1)} \left(\frac{\bar{R}_1^2}{n_1} + \dots + \frac{\bar{R}_k^2}{n_k} \right) - 3(N+1) \end{aligned}$$

Если верна нулевая гипотеза, то данная статистика имеет приблизительно распределение χ_{k-1}^2 . Поэтому нулевая гипотеза должна быть отвергнута, если $X > \chi_{k-1}^{2,1-\alpha}$, где $\chi_{k-1}^{2,1-\alpha}$ есть $1 - \alpha$ квантиль распределения χ_{k-1}^2 .

Пример

Рассмотрим три выборки объёмов 5, 6, 3

$$x_1 = 0,82 \quad x_2 = 0,85 \quad x_3 = 0,27 \quad x_4 = 0,00 \quad x_5 = 0,95$$

$$y_1 = 0,68 \quad y_2 = 0,88 \quad y_3 = 0,36 \quad y_4 = 0,39 \quad y_5 = 0,23 \quad y_6 = 0,51$$

$$z_1 = 0,75 \quad z_2 = 0,77 \quad z_3 = 0,29$$

Проверим гипотезу о том, что распределение не меняется при переходе от одной выборки к другой.

1. Объединяем выборки в одну выборку объема $N = 14$ и ранжируем

$$x_4 = 0,00 \quad y_5 = 0,23 \quad x_3 = 0,27 \quad z_3 = 0,29 \quad y_3 = 0,36 \quad y_4 = 0,39 \quad y_6 = 0,51 \quad y_1 = 0,68 \\ z_1 = 0,75 \quad z_2 = 0,77 \quad x_1 = 0,82 \quad x_2 = 0,85 \quad y_2 = 0,88 \quad x_5 = 0,95$$

2. К первой выборке относятся ранги $\{1, 3, 11, 12, 14\}$, ко второй выборке относятся ранги $\{2, 5, 6, 7, 8, 13\}$, к третьей выборке относятся ранги $\{4, 9, 10\}$
3. Средний ранг по первой выборке $\bar{R}_1 = 8,2$, средний ранг по второй выборке $\bar{R}_2 = 6,8$, средний ранг по первой третьей выборке равен $\bar{R}_3 = 7,6$
4. Вычисляем статистику $X = \frac{12}{14(14+1)}(\frac{8,2^2}{5} + \frac{6,8^2}{6} + \frac{7,6^2}{3}) - 3(14 + 1) = 0,29$
5. Так как $X > \chi_{k-1}^{2,1-\alpha} = 5,9$, то гипотеза о том, что распределение не меняется при переходе от одной выборки к другой не отвергается.

Задача 1

Фермер продаёт свою продукцию в трёх городах, в каждом в нескольких торговых точках.

Объём продаж в первом городе представлен в выборке

$$0,57 \quad 0,17 \quad 0,64 \quad 0,93 \quad 0,09$$

Объём продаж во втором городе представлен в выборке

$$0,25 \quad 0,51 \quad 0,12 \quad 0,03 \quad 0,95$$

Объём продаж в третьем городе представлен в выборке

$$0,51 \quad 0,33 \quad 0,53$$

Проверьте гипотезу о том, что продукция во всех городах продаётся одинаково.

Задача 2

После сезонного снижения цены на продукцию объёмы продаж изменились. В трех городах в различных торговых точках они теперь соответствуют следующим величинам

$$2,57 \quad 0,57 \quad 1,64 \quad 1,93 \quad 0,09$$

$$1,51 \quad 0,93 \quad 1,53$$

$$0,25 \quad 0,51 \quad 0,12 \quad 0,03$$

Проверьте гипотезу о том, что снижение цены подействовало одинаково на объём продаж во всех трех городах.

9.6.2 Двухфакторный анализ. Критерий Фридмана

Данный критерий является непараметрическим аналогом двухфакторного дисперсионного анализа.

Предполагается, что на x могут оказывать влияние два фактора. Объясним, что это значит.

Пусть имеется выборка, занумерованная двумя индексами $x_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, k$. Предположим, что

$$x_{i,j} = \mu + \beta_i + \alpha_j + \epsilon_{i,j},$$

здесь μ – константа, β – влияние на x фактора B , α – влияние на x фактора A , ϵ – случайная составляющая x .

Проверим гипотезу о том, что влияние фактора A отсутствует, то есть $\alpha_1 = \dots = \alpha_k = 0$.

Ранжируем выборку $x_{i,j}$, пусть $r_{i,j}$ – ранг наблюдения (i, j) . Введём обозначения:

$$\bar{r}_i = \frac{1}{n} \sum_{j=1}^k r_{i,j}, \quad \bar{\bar{r}} = \frac{1}{nm} \sum_{i,j} r_{i,j} = \frac{k+1}{2}.$$

Для проверки гипотезы составляется величина

$$X = \frac{12n}{k(k+1)} \sum_{i=1}^k (\bar{r}_i - \bar{\bar{r}})^2.$$

При малых k гипотеза отвергается, если $X > X_{\text{критич}}$. Таблица 5-процентных критических точек при $k = 3, 4$ и малых n приведена ниже.

n	$k = 3$	$k = 4$
2		6,000
3	6,000	7,000
4	6,500	7,500
5	6,400	7,320
6	6,333	7,400
7	6,000	7,629
8	6,250	7,650
9	6,222	7,667
10	6,200	7,688
11	6,545	
12	6,167	
13	6,000	
14	6,143	
15	6,400	
16	6,125	
17	6,118	
18	6,333	
19	6,000	
20	6,100	
21	6,000	
22	5,818	
23	5,826	
24	6,083	
25	6,080	

Если $k > 4$, то при нулевой гипотезе статистика X имеет приблизительно распределение χ^2_{k-1} . Поэтому нулевая гипотеза должна быть отвергнута, если $X > \chi^2_{k-1}^{2,1-\alpha}$, где $\chi^2_{k-1}^{2,1-\alpha}$ есть $1 - \alpha$ квантиль распределения χ^2_{k-1} .

Пример

Даны выборки

0,92 0,01 0,08 0,02

0,49 0,03 0,43 0,23

0,21 0,35 0,32 0,20

Требуется провести двухфакторный анализ.

1. Составляем таблицу, по строкам записав ранжировки первой, второй и третьей выборок соответственно

4	1	2	3
4	1	3	2
2	4	3	1

2. Находим средние значения по столбцам : $\bar{r}_1 = 3,3$, $\bar{r}_2 = 2$, $\bar{r}_3 = 2,6$, $\bar{r}_4 = 2$.
3. Находим $X = \frac{12 \cdot 4}{3(3+1)} \sum_{i=1}^k (\bar{r}_i - \bar{\bar{r}})^2 = 1,62$.
4. Так как $X < X_{\text{критич}} = 6,500$, то гипотеза об отсутствии влияния фактора не отвергается.

Задача 1

Картофель, капуста и помидоры продаются в четырёх торговых точках. В трёх выборках приведены объёмы продаж в этих торговых точках.

0,57 0,17 0,64 0,93

0,25 0,51 0,09 0,53

0,51 0,12 0,33 0,34

Проверьте гипотезу о том, что объём продаж не зависит от типа товара.

Задача 2

Картофель, капуста и помидоры продаются в четырёх торговых точках. В трёх выборках ниже приведены объёмы продаж в этих торговых точках.

1,47 0,67 1,68 0,73

0,54 0,63 1,57 0,79

0,25 0,61 0,42 0,03

Проверьте гипотезу о том, что объём продаж не зависит от торговой точки.

Приложение А

Таблицы

Таблица А.1: Критические значения стандартного нормального распределения

	Уровень значимости α								
two-side	0.400	0.200	0.100	0.050	0.020	0.010	0.005	0.002	0.001
one-side	0.200	0.100	0.050	0.025	0.010	0.005	0.0025	0.001	0.0005
z	0.842	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Таблица А.2: Критические значения распределения χ_k^2

	Уровень значимости α									
k	0.100	0.050	0.025	0.020	0.010	0.005	0.001	0.95	0.975	0.995
1	2.706	3.841	5.024	5.412	6.635	7.879	10.828	0.004	0.001	0.000
2	4.605	5.991	7.378	7.824	9.210	10.597	13.816	0.103	0.051	0.01
3	6.251	7.815	9.348	9.837	11.345	12.838	16.266	0.352	0.216	0.072
4	7.779	9.488	11.143	11.668	13.277	14.860	18.467	0.711	0.484	0.207
5	9.236	11.070	12.833	13.388	15.086	16.750	20.515	1.145	0.831	0.412
6	10.645	12.592	14.449	15.033	16.812	18.548	22.458	1.635	1.237	0.676
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322	2.167	1.69	0.989
8	13.362	15.507	17.535	18.168	20.090	21.955	26.124	2.733	2.18	1.344
9	14.684	16.919	19.023	19.679	21.666	23.589	27.877	3.325	2.7	1.735
10	15.987	18.307	20.483	21.161	23.209	25.188	29.588	3.94	3.24	2.156
11	17.275	19.675	21.920	22.618	24.725	26.757	31.264	4.57	3.816	2.603
12	18.549	21.026	23.337	24.054	26.217	28.300	32.909	5.226	4.404	3.074
13	19.812	22.362	24.736	25.472	27.688	29.819	34.528	5.892	5.009	3.565
14	21.064	23.685	26.119	26.873	29.141	31.319	36.123	6.571	5.629	4.075
15	22.307	24.996	27.488	28.259	30.578	32.801	37.697	7.261	6.262	4.601
16	23.542	26.296	28.845	29.633	32.000	34.267	39.252	7.962	6.908	5.142
17	24.769	27.587	30.191	30.995	33.409	35.718	40.790	8.672	7.564	5.697
18	25.989	28.869	31.526	32.346	34.805	37.156	42.312	9.39	8.231	6.265
19	27.204	30.144	32.852	33.687	36.191	38.582	43.820	10.117	8.907	6.844
20	28.412	31.41	34.17	35.02	36.191	39.997	45.315	10.851	9.591	7.434
21	29.615	32.671	35.479	36.343	38.932	41.401	46.797	11.591	10.283	8.034
22	30.813	33.924	36.781	37.659	40.289	42.796	48.268	12.338	10.982	8.643
23	32.007	35.172	38.076	38.968	41.638	44.181	49.728	13.091	11.689	9.26
24	33.196	36.415	39.364	40.270	42.980	45.559	51.179	13.848	12.401	9.886
25	34.382	37.652	40.646	41.566	44.314	46.928	52.620	14.611	13.12	10.52
30	40.256	43.773	46.979	47.962	50.892	53.672	59.703	18.49	16.79	13.787
40	51.805	55.758	59.342	60.436	63.691	66.766	73.402	26.51	24.433	20.707
50	63.167	67.505	71.420	72.613	76.154	79.490	86.661	34.764	32.357	27.991
60	74.397	79.082	83.298	84.580	88.379	91.952	99.607	43.188	40.482	35.534
70	85.527	90.531	95.023	96.388	100.425	104.215	112.317	51.739	48.758	43.275
80	96.578	101.879	106.629	108.069	112.329	116.321	124.839	60.391	57.153	51.172
90	107.565	113.145	118.136	119.648	124.116	128.299	137.208	69.126	65.647	59.196
100	118.498	124.342	129.561	131.142	135.807	140.169	149.449	77.93	74.222	67.328

Таблица А.3: Критические значения распределения t_k (распределения Стьюдента)

k	Уровень значимости α					
two-side	0.100	0.050	0.025	0.010	0.005	0.001
one-side	0.050	0.0250	0.0125	0.0050	0.0025	0.0005
1	6.314	12.706	25.452	63.657	127.321	636.619
2	2.920	4.303	6.205	9.925	14.089	31.599
3	2.353	3.182	4.177	5.841	7.453	12.924
4	2.132	2.776	3.495	4.604	5.598	8.610
5	2.015	2.571	3.163	4.032	4.773	6.869
6	1.943	2.447	2.969	3.707	4.317	5.959
7	1.895	2.365	2.841	3.499	4.029	5.408
8	1.860	2.306	2.752	3.355	3.833	5.041
9	1.833	2.262	2.685	3.250	3.690	4.781
10	1.812	2.228	2.634	3.169	3.581	4.587
11	1.796	2.201	2.593	3.106	3.497	4.437
12	1.782	2.179	2.560	3.055	3.428	4.318
13	1.771	2.160	2.533	3.012	3.372	4.221
14	1.761	2.145	2.510	2.977	3.326	4.140
15	1.753	2.131	2.490	2.947	3.286	4.073
16	1.746	2.120	2.473	2.921	3.252	4.015
17	1.740	2.110	2.458	2.898	3.222	3.965
18	1.734	2.101	2.445	2.878	3.197	3.922
19	1.729	2.093	2.433	2.861	3.174	3.883
20	1.725	2.086	2.423	2.845	3.153	3.850
21	1.721	2.080	2.414	2.831	3.135	3.819
22	1.717	2.074	2.405	2.819	3.119	3.792
23	1.714	2.069	2.398	2.807	3.104	3.768
24	1.711	2.064	2.391	2.797	3.091	3.745
25	1.708	2.060	2.385	2.787	3.078	3.725
26	1.706	2.056	2.379	2.779	3.067	3.707
27	1.703	2.052	2.373	2.771	3.057	3.690
28	1.701	2.048	2.368	2.763	3.047	3.674
29	1.699	2.045	2.364	2.756	3.038	3.659
30	1.697	2.042	2.360	2.750	3.030	3.646
40	1.684	2.021	2.329	2.704	2.971	3.551
50	1.676	2.009	2.311	2.678	2.937	3.496
60	1.671	2.000	2.299	2.660	2.915	3.460
70	1.667	1.994	2.291	2.648	2.899	3.435
80	1.664	1.990	2.284	2.639	2.887	3.416
90	1.662	1.987	2.280	2.632	2.878	3.402
100	1.660	1.984	2.276	2.626	2.871	3.390
120	1.658	1.980	2.270	2.617	2.860	3.373
200	1.653	1.972	2.258	2.601	2.839	3.340
300	1.650	1.968	2.253	2.592	2.828	3.323
500	1.648	1.965	2.248	2.586	2.820	3.310
∞	1.645	1.960	2.241	2.576	2.807	3.291

Таблица А.4: 5% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1									
k_2	1	2	3	4	5	6	7	8	9	10
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Таблица А.5: 5% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1								
k_2	15	20	25	30	40	50	60	100	120
2	19.429	19.446	19.456	19.462	19.471	19.476	19.479	19.486	19.487
3	8.703	8.660	8.634	8.617	8.594	8.581	8.572	8.554	8.549
4	5.858	5.803	5.769	5.746	5.717	5.699	5.688	5.664	5.658
5	4.619	4.558	4.521	4.496	4.464	4.444	4.431	4.405	4.398
6	3.938	3.874	3.835	3.808	3.774	3.754	3.740	3.712	3.705
7	3.511	3.445	3.404	3.376	3.340	3.319	3.304	3.275	3.267
8	3.218	3.150	3.108	3.079	3.043	3.020	3.005	2.975	2.967
9	3.006	2.936	2.893	2.864	2.826	2.803	2.787	2.756	2.748
10	2.845	2.774	2.730	2.700	2.661	2.637	2.621	2.588	2.580
11	2.719	2.646	2.601	2.570	2.531	2.507	2.490	2.457	2.448
12	2.617	2.544	2.498	2.466	2.426	2.401	2.384	2.350	2.341
13	2.533	2.459	2.412	2.380	2.339	2.314	2.297	2.261	2.252
14	2.463	2.388	2.341	2.308	2.266	2.241	2.223	2.187	2.178
15	2.403	2.328	2.280	2.247	2.204	2.178	2.160	2.123	2.114
16	2.352	2.276	2.227	2.194	2.151	2.124	2.106	2.068	2.059
17	2.308	2.230	2.181	2.148	2.104	2.077	2.058	2.020	2.011
18	2.269	2.191	2.141	2.107	2.063	2.035	2.017	1.978	1.968
19	2.234	2.155	2.106	2.071	2.026	1.999	1.980	1.940	1.930
20	2.203	2.124	2.074	2.039	1.994	1.966	1.946	1.907	1.896
21	2.176	2.096	2.045	2.010	1.965	1.936	1.916	1.876	1.866
22	2.151	2.071	2.020	1.984	1.938	1.909	1.889	1.849	1.838
23	2.128	2.048	1.996	1.961	1.914	1.885	1.865	1.823	1.813
24	2.108	2.027	1.975	1.939	1.892	1.863	1.842	1.800	1.790
25	2.089	2.007	1.955	1.919	1.872	1.842	1.822	1.779	1.768
26	2.072	1.990	1.938	1.901	1.853	1.823	1.803	1.760	1.749
27	2.056	1.974	1.921	1.884	1.836	1.806	1.785	1.742	1.731
28	2.041	1.959	1.906	1.869	1.820	1.790	1.769	1.725	1.714
29	2.027	1.945	1.891	1.854	1.806	1.775	1.754	1.710	1.698
30	2.015	1.932	1.878	1.841	1.792	1.761	1.740	1.695	1.683
40	1.924	1.839	1.783	1.744	1.693	1.660	1.637	1.589	1.577
50	1.871	1.784	1.727	1.687	1.634	1.599	1.576	1.525	1.511
60	1.836	1.748	1.690	1.649	1.594	1.559	1.534	1.481	1.467
100	1.768	1.676	1.616	1.573	1.515	1.477	1.450	1.392	1.376
120	1.750	1.659	1.598	1.554	1.495	1.457	1.429	1.369	1.352
500	1.686	1.592	1.528	1.482	1.419	1.376	1.345	1.275	1.255
∞	1.666	1.571	1.506	1.459	1.394	1.350	1.318	1.243	1.221

Таблица А.6: 10% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1									
k_2	1	2	3	4	5	6	7	8	9	10
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663
120	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644	1.612
∞	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599

Таблица А.7: 10% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1								
k_2	15	20	25	30	40	50	60	100	120
2	9.425	9.441	9.451	9.458	9.466	9.471	9.475	9.481	9.483
3	5.200	5.184	5.175	5.168	5.160	5.155	5.151	5.144	5.143
4	3.870	3.844	3.828	3.817	3.804	3.795	3.790	3.778	3.775
5	3.238	3.207	3.187	3.174	3.157	3.147	3.140	3.126	3.123
6	2.871	2.836	2.815	2.800	2.781	2.770	2.762	2.746	2.742
7	2.632	2.595	2.571	2.555	2.535	2.523	2.514	2.497	2.493
8	2.464	2.425	2.400	2.383	2.361	2.348	2.339	2.321	2.316
9	2.340	2.298	2.272	2.255	2.232	2.218	2.208	2.189	2.184
10	2.244	2.201	2.174	2.155	2.132	2.117	2.107	2.087	2.082
11	2.167	2.123	2.095	2.076	2.052	2.036	2.026	2.005	2.000
12	2.105	2.060	2.031	2.011	1.986	1.970	1.960	1.938	1.932
13	2.053	2.007	1.978	1.958	1.931	1.915	1.904	1.882	1.876
14	2.010	1.962	1.933	1.912	1.885	1.869	1.857	1.834	1.828
15	1.972	1.924	1.894	1.873	1.845	1.828	1.817	1.793	1.787
16	1.940	1.891	1.860	1.839	1.811	1.793	1.782	1.757	1.751
17	1.912	1.862	1.831	1.809	1.781	1.763	1.751	1.726	1.719
18	1.887	1.837	1.805	1.783	1.754	1.736	1.723	1.698	1.691
19	1.865	1.814	1.782	1.759	1.730	1.711	1.699	1.673	1.666
20	1.845	1.794	1.761	1.738	1.708	1.690	1.677	1.650	1.643
21	1.827	1.776	1.742	1.719	1.689	1.670	1.657	1.630	1.623
22	1.811	1.759	1.726	1.702	1.671	1.652	1.639	1.611	1.604
23	1.796	1.744	1.710	1.686	1.655	1.636	1.622	1.594	1.587
24	1.783	1.730	1.696	1.672	1.641	1.621	1.607	1.579	1.571
25	1.771	1.718	1.683	1.659	1.627	1.607	1.593	1.565	1.557
26	1.760	1.706	1.671	1.647	1.615	1.594	1.581	1.551	1.544
27	1.749	1.695	1.660	1.636	1.603	1.583	1.569	1.539	1.531
28	1.740	1.685	1.650	1.625	1.592	1.572	1.558	1.528	1.520
29	1.731	1.676	1.640	1.616	1.583	1.562	1.547	1.517	1.509
30	1.722	1.667	1.632	1.606	1.573	1.552	1.538	1.507	1.499
40	1.662	1.605	1.568	1.541	1.506	1.483	1.467	1.434	1.425
50	1.627	1.568	1.529	1.502	1.465	1.441	1.424	1.388	1.379
60	1.603	1.543	1.504	1.476	1.437	1.413	1.395	1.358	1.348
100	1.557	1.494	1.453	1.423	1.382	1.355	1.336	1.293	1.282
120	1.545	1.482	1.440	1.409	1.368	1.340	1.320	1.277	1.265
500	1.501	1.435	1.391	1.358	1.313	1.282	1.260	1.209	1.194
∞	1.487	1.421	1.375	1.342	1.295	1.263	1.240	1.185	1.169

Таблица А.8: 1% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1									
k_2	1	2	3	4	5	6	7	8	9	10
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Таблица А.9: 1% критические значения распределения F_{k_1, k_2} (распределения Фишера)

	k_1								
k_2	15	20	25	30	40	50	60	100	120
2	99.433	99.449	99.459	99.466	99.474	99.479	99.482	99.489	99.491
3	26.872	26.690	26.579	26.505	26.411	26.354	26.316	26.240	26.221
4	14.198	14.020	13.911	13.838	13.745	13.690	13.652	13.577	13.558
5	9.722	9.553	9.449	9.379	9.291	9.238	9.202	9.130	9.112
6	7.559	7.396	7.296	7.229	7.143	7.091	7.057	6.987	6.969
7	6.314	6.155	6.058	5.992	5.908	5.858	5.824	5.755	5.737
8	5.515	5.359	5.263	5.198	5.116	5.065	5.032	4.963	4.946
9	4.962	4.808	4.713	4.649	4.567	4.517	4.483	4.415	4.398
10	4.558	4.405	4.311	4.247	4.165	4.115	4.082	4.014	3.996
11	4.251	4.099	4.005	3.941	3.860	3.810	3.776	3.708	3.690
12	4.010	3.858	3.765	3.701	3.619	3.569	3.535	3.467	3.449
13	3.815	3.665	3.571	3.507	3.425	3.375	3.341	3.272	3.255
14	3.656	3.505	3.412	3.348	3.266	3.215	3.181	3.112	3.094
15	3.522	3.372	3.278	3.214	3.132	3.081	3.047	2.977	2.959
16	3.409	3.259	3.165	3.101	3.018	2.967	2.933	2.863	2.845
17	3.312	3.162	3.068	3.003	2.920	2.869	2.835	2.764	2.746
18	3.227	3.077	2.983	2.919	2.835	2.784	2.749	2.678	2.660
19	3.153	3.003	2.909	2.844	2.761	2.709	2.674	2.602	2.584
20	3.088	2.938	2.843	2.778	2.695	2.643	2.608	2.535	2.517
21	3.030	2.880	2.785	2.720	2.636	2.584	2.548	2.475	2.457
22	2.978	2.827	2.733	2.667	2.583	2.531	2.495	2.422	2.403
23	2.931	2.781	2.686	2.620	2.535	2.483	2.447	2.373	2.354
24	2.889	2.738	2.643	2.577	2.492	2.440	2.403	2.329	2.310
25	2.850	2.699	2.604	2.538	2.453	2.400	2.364	2.289	2.270
26	2.815	2.664	2.569	2.503	2.417	2.364	2.327	2.252	2.233
27	2.783	2.632	2.536	2.470	2.384	2.330	2.294	2.218	2.198
28	2.753	2.602	2.506	2.440	2.354	2.300	2.263	2.187	2.167
29	2.726	2.574	2.478	2.412	2.325	2.271	2.234	2.158	2.138
30	2.700	2.549	2.453	2.386	2.299	2.245	2.208	2.131	2.111
40	2.522	2.369	2.271	2.203	2.114	2.058	2.019	1.938	1.917
50	2.419	2.265	2.167	2.098	2.007	1.949	1.909	1.825	1.803
60	2.352	2.198	2.098	2.028	1.936	1.877	1.836	1.749	1.726
100	2.223	2.067	1.965	1.893	1.797	1.735	1.692	1.598	1.572
120	2.192	2.035	1.932	1.860	1.763	1.700	1.656	1.559	1.533
500	2.075	1.915	1.810	1.735	1.633	1.566	1.517	1.408	1.377
∞	2.039	1.878	1.773	1.696	1.592	1.523	1.473	1.358	1.325

Приложение В

Как работать с таблицами

В.1 Таблица нормального распределения

В этом уроке мы научимся находить значения вероятностей нормальной величины, а также её квантили. Функцию распределения стандартной нормальной случайной величины $Z \sim N(0, 1)$ обозначают $\Phi(x)$. По определению она есть:

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

Первообразную не найти! Значение интеграла вычисляют численно. Для функции $\Phi(x)$ составлены подробные таблицы, позволяющие находить вероятность того, что случайная величина Z попадает в некоторый отрезок $[a; b]$:

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

По таблице проще всего вычислить вероятность того, что случайная величина Z попадает левее точки a :

$$P(Z \leq a) = \Phi(a).$$

А чтобы вычислить вероятность того, что случайная величина Z попадает правее точки a , пользуются тем, что сумма вероятностей равна 1:

$$P(Z \geq a) = 1 - \Phi(a).$$

Для отрицательных значений пользуются чётностью функции плотности, то есть

$$P(Z \leq -a) = 1 - P(Z \leq a) = 1 - \Phi(a).$$

Можно пользоваться таблицей покороче!

Как правило, в задачах используются следующие значения:

Критические значения стандартного нормального распределения

	Уровень значимости α								
two-side	0.400	0.200	0.100	0.050	0.020	0.010	0.005	0.002	0.001
one-side	0.200	0.100	0.050	0.025	0.010	0.005	0.0025	0.001	0.0005
z	0.842	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Покажем на примере, как вычисляются вероятности для стандартной нормальной величины.

Пусть нам требуется вычислить следующую вероятность $P(z \leq 0.52)$.

Решение.

- $P(z \leq 0.52) = 0.6985$.

z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0, 6554	0, 6591	0, 6628	0, 6664	0, 6700
0.5	0, 6915	0, 6950	0, 6985	0, 7019	0, 7054
0.6	0, 7257	0, 7291	0, 7324	0, 7357	0, 7389
0.7	0, 7580	0, 7611	0, 7642	0, 7673	0, 7704

Пример. Пусть нам требуется вычислить следующие вероятности $P(z \geq 0.52)$, $P(z \leq -1.23)$, $P(z \geq -2.33)$, $P(0.12 \leq z \leq 2.52)$, $P(-1.12 \leq z \leq 1.65)$.

Решение. • $P(z \geq 0.52) = 1 - P(z \leq 0.52) = 1 - 0.6985 = 0.3015$;

- $P(z \leq -1.23) = 1 - P(z \leq 1.23) = 1 - 0.8907 = 0.1093$;
- $P(z \geq -2.33) = P(z \leq 2.33) = 0.9901$;
- $P(0.12 \leq z \leq 2.52) = P(z \leq 2.52) - P(z \leq 0.12) = 0.9941 - 0.5478$;
- $P(-1.12 \leq z \leq 1.65) = P(z \leq 1.65) - P(z \leq -1.12) = P(z \leq 1.65) - (1 - P(z \leq 1.12)) = 0.9505 - (1 - 0.8686) = 0.8191$.

Для произвольной случайной нормальной величины X вероятность $P(X \leq x)$ находится следующим образом (эта процедура называется стандартизацией):

- вычисляем $z = \frac{x-a}{\sigma}$;
- находим в таблице $\Phi(z)$.

Квантилью уровня p для с.в. X называется такое значение x_p , что $P(X \leq x_p) = p$.

Квантиль уровня p вычисляется следующим образом: $x_p = a + \sigma \cdot z_p$, где z_p - квантиль стандартной нормальной с.в. Для значений p , которых нет в таблице надо воспользоваться чётностью функции плотности $x_p = x_{1-p}$.

Пример. Найдите квантили стандартной нормальной величины $z_{0.62}$ и $z_{0.33}$.

Решение. Смотрим в таблицу и находим 0.6217, так как это ближайшее число к 0.62. $z_{0.62} \approx 0.31$.

Чтобы найти $z_{0.33}$, вспоминаем про чётность функции плотности, потому просто 0.33 в таблице не найти.

$$z_{0.33} = z_{1-0.33} = z_{0.67} \approx 0.44.$$

z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0,6554	0,6591	0,6628	0,6664	0,6700
0.5	0,6915	0,6950	0,6985	0,7019	0,7054
0.6	0,7257	0,7291	0,7324	0,7357	0,7389
0.7	0,7580	0,7611	0,7642	0,7673	0,7704

Пример. Случайная величина X имеет нормальное распределение со средним значением 10 и дисперсией 25. Чему равна вероятность того, что эта случайная величина примет значение большее 25? Вычислите квантиль уровня 0.99.

Решение. Для того, чтобы найти требуемую вероятность, перейдем к стандартной нормальной величине $Z = \frac{X-10}{\sqrt{25}}$ и воспользуемся таблицей для нормального распределения

$$\begin{aligned}
 P(X \geq 25) &= P\left(\frac{X-10}{5} \geq \frac{25-10}{5}\right) = P(Z \geq 3) = \\
 &= 1 - P(Z \leq 3) \approx 1 - 0.99865 = 0.00135.
 \end{aligned}$$

Чтобы найти квантиль $x_{0.99}$, сначала найдём по таблице квантиль $z_{0.99} \approx 2.33$. Теперь $x_{0.99} = a + \sigma \cdot z_{0.99} \approx 10 + 5 \cdot 2.33 = 21.65$.

В.2 Распределение хи-квадрат

Пусть ξ_1, \dots, ξ_k — совместно независимые стандартные нормальные случайные величины, то есть: $\xi_i \sim N(0, 1)$. Тогда случайная величина $\xi = \xi_1^2 + \dots + \xi_k^2$ имеет распределение хи-квадрат с k степенями свободы. Для этого распределения составлена таблица наиболее используемых значений. В приведённой таблице указаны некоторые значения для вероятностей вида $P(\chi^2(k) \geq x)$.

Пример. Для распределения χ^2 с семью степенями свободы найдите квантиль уровня 0.99.

Решение. Смотрим в строчку для семи степеней свободы, то есть $k = 7$, и столбец для $\alpha = 0.01$. Находим значение 18.475, это и есть $\chi_{0.99}^2(7) \approx 18.475$.

k	0.100	0.050	0.025	0.020	0.010	0.005	0.001
1	2.706	3.841	5.024	5.412	6.635	7.879	10.828
2	4.605	5.991	7.378	7.824	9.210	10.597	13.816
3	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	7.779	9.488	11.143	11.668	13.277	14.860	18.467
5	9.236	11.070	12.833	13.388	15.086	16.750	20.515
6	10.645	12.592	14.449	15.033	16.812	18.548	22.458
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	13.362	15.507	17.535	18.168	20.090	21.955	26.124

Пример. Для распределения χ^2 с двумя степенями свободы найдите квантиль уровня 0.95.

Решение. Смотрим в строчку для двух степеней свободы, то есть $k = 2$, и столбец для $\alpha = 0.05$. Находим значение 5.991, это и есть $\chi_{0.95}^2(2) \approx 5.991$.

В.3 Таблица распределения Стьюдента

Пусть $\xi_0, \xi_1, \dots, \xi_n$ — независимые стандартные нормальные случайные величины. Тогда распределение случайной величины $t = \frac{\xi_0}{\sqrt{\frac{1}{n} \sum_{i=1}^n \xi_i^2}}$,

называется распределением Стьюдента с n степенями свободы и пишут $t \sim t(n)$. Для этого распределения составлена таблица наиболее используемых значений. В нижеприведённой таблице указаны некоторые значения для вероятностей вида $P(t(n) \geq x)$ и $P(|t(n)| \geq x)$

Пример. Для распределения Стьюдента с тремя степенями свободы вычислите вероятности $P(t(3) < 5.841)$, $P(|t(3)| > 3.182)$ и найдите квантили уровня 0.95 и 0.9995.

Решение. Смотрим в строчку для трёх степеней свободы, то есть $k = 3$, и находим значение 5.841. Откуда $P(t(3) < 5.841) \approx 1 - 0.005$.

$$P(|t(3)| > 3.182) \approx 0.05.$$

$$t_{0.95}(3) \approx 2.353 \text{ и } t_{0.9995}(3) \approx -12.924.$$

Пример. Найдите с точностью до тысячных квантиль уровня 0.975 для случайной величины, имеющей распределение Стьюдента с 4 степенями свободы.

Решение. На пересечении строки соответствующей четырём степеням свободы и уровню значимости $1 - 0.975 = 0.025$ находим $t_{0.975}(4) \approx 2.776$.

<i>two – side</i>	0.100	0.050	0.025	0.010	0.005	0.001
<i>one – side</i>	0.050	0.0250	0.0125	0.0050	0.0025	0.0005
<i>k</i>						
1	6.314	12.706	25.452	63.657	127.321	636.619
2	2.920	4.303	6.205	9.925	14.089	31.599
3	2.353	3.182	4.177	5.841	7.453	12.924
4	2.132	2.776	3.495	4.604	5.598	8.610
5	2.015	2.571	3.163	4.032	4.773	6.869

Значения вероятностей зависят от числа степеней свободы, поэтому подробную таблицу теперь написать не получится, как для нормального распределения. Поэтому вероятности и квантили уже найти, как правило, не удастся. Надо использовать компьютер, например, EXCEL.

В.4 Распределение Фишера

Пусть ξ_1, ξ_2 — две независимые случайные величины, имеющие распределение хи-квадрат: $\xi_i \sim \chi^2(k_i)$, где $k_i \in \mathbb{N}$, $i = 1, 2$. Тогда распределение случайной величины $F = \frac{\xi_1/k_1}{\xi_2/k_2}$, называется распределением Фишера (распределением Снедекора) со степенями свободы k_1 и k_2 . Пишут $F \sim F(k_1, k_2)$.

Для распределения составлены таблицы наиболее используемых значений. В нижеприведённой таблице указаны значения для 5-процентных точек или квантилей уровня 0.95.

Пример. Найдите 5-процентную точку для $F(4; 8)$.

Решение. Смотрим в строчку для восьми степеней свободы, то есть $k_2 = 8$, и столбец для $k_1 = 4$. Находим значение 3.838.

Пример. Найдите 5-процентную точку для $F(8; 4)$.

Решение. Смотрим в строчку для четырёх степеней свободы, то есть $k_2 = 4$, и столбец для $k_1 = 8$. Находим значение 6.041.

$k_2 \setminus k_1$	1	2	3	4	5	6	7	8	9
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388

В.5 Использование компьютера

Приведем список функции MS Excel и OpenOffice для вычисления критических значений стандартных распределений с уровнем значимости α

Распределение	MS Excel (Рус)	MS Excel (Eng) OpenOffice
Двустороннее $\mathcal{N}(0, 1)$	НОРМСТОБР($1 - \alpha/2$)	NORMSINV($1 - \alpha/2$)
Одностороннее $\mathcal{N}(0, 1)$	НОРМСТОБР($1 - \alpha$)	NORMSINV($1 - \alpha$)
χ_k^2 (хи-квадрат)	ХИ2ОБР($\alpha; k$)	CHIINV($\alpha; k$)
Двустороннее t_k (Стьюдента)	СТЮДРАСПОБР($\alpha; k$)	TINV($\alpha; k$)
Одностороннее t_k (Стьюдента)	СТЮДРАСПОБР($2\alpha; k$)	TINV($2\alpha; k$)
Фишера F_{k_1, k_2}	ФРАСПОБР($\alpha; k_1; k_2$)	FINV ($\alpha; k_1; k_2$)

Литература

- [1] Айвазян С. А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. - М.: Юнити-Дана, 2001. - 656 с.
- [2] Артамонов Н.В. Введение в эконометрику.– 2-е изд., испр. и доп.– М.:МЦНМО, 2014. – 224 с.
- [3] Артамонов Н.В. Теория вероятностей и математическая статистика: углубленный курс / Н.В. Артамонов. – М.: МГИМО-Университет, 2008. – 98 с.
- [4] Ивашев-Мусатов О. С. Теория вероятностей и математическая статистика: учеб. пособие. - 2-е изд., перераб. и доп. - М.: ФИМА, 2003. - 224 с.
- [5] Фадеева Л. Н., Лебедев А. В., Теория вероятностей и математическая статистика: учебное пособие. - 2-е изд., перераб. и доп. - М.: Эксмо, 2010. - 496 с. – (Новое экономическое образование).
- [6] Тюрин Ю. Н., Макаров А.А., Анализ данных на компьютере: учебное пособие. - 4-е изд., перераб. - М.: ИД Форум, 2008. - 368 с., ил. - (Высшее образование).
- [7] Иванов О.В. Статистика. Учебный курс для социологов и менеджеров. Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода. - М. 2005. - 187 с.
- [8] Иванов О.В. Статистика. Учебный курс для социологов и менеджеров. Часть 2. Доверительные интервалы. Проверка гипотез. Методы и их применение. – М. 2005. – 220 с.
- [9] Лагутин М.Б. Наглядная математическая статистика: учебное пособие. - 2-е изд., перераб. - М.: БИНОМ. Лаборатория знаний, 2009. – 472 с.
- [10] Макаров А.А., Ивин Е.А., Курбацкий А.Н., Курс теории вероятностей в задачах и упражнениях: учебное пособие для социально-экономических специальностей. - М.: МАКС Пресс, 2014 - 116 с.

Научное издание

Ивин Евгений Александрович
Курбацкий Алексей Николаевич
Артамонов Дмитрий Вячеславович

МЕТОДИЧЕСКОЕ ПОСОБИЕ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

для социально-экономических специальностей

Редакционная подготовка	Т.В. Алешина
Оригинал-макет	И.В. Артамонов
Корректор	М.В. Чумаченко

Подписано в печать _ . _ .2016.
Формат бумаги _ × _ / _ . Печать цифровая. Бумага офсетная.
Усл. печ. л. _ , _ Печ. л. _ , _ . Тираж _ экз. Заказ № _

Федеральное государственное бюджетное учреждение науки
Институт социально-экономического развития территорий РАН
(ИСЭРТ РАН)
160014, г. Вологда, ул. Горького, 56а
Телефон (8172) 59-78-03, e-mail: common@vscc.ac.ru