Казанский Федеральный Университет

Кафедра Математической статистики

Симушкин С.В., Заикин А.А., Кареев И.А., Салимов Р.Ф.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ВЫПОЛНЕНИЯ КУРСОВОЙ РАБОТЫ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Казань, 2019 г.

Содержание

Содержание **2**

Введение **7**

Обозначения **7**

Список источников **9**

I Базовые понятия и определения **10**

1 Выборка **10**
1.1 Вопросы и задания для самоконтроля 12

2 Функции распределения и квантили **12**
2.1 Распределение 12
2.2 Квантӣли — процентные точки 14
2.3 Вопросы и задания для самоконтроля 14

3 Проверка гипотез **15**
3.1 Гипотеза 15
3.2 Критерий и критическая область. Вероятности ошибок . . . 16
3.3 Тестовая статистика 19
3.4 p-значение (критический уровень значимости) 22
3.5 Общая схема построения статистических критериев 23
3.6 Вопросы и задания для самоконтроля 26

4 Точечное оценивание **26**
4.1 Несмещённость 27
4.2 Состоятельность 28
4.3 Оптимальность 29
4.4 Вопросы и задания для самоконтроля 30

5 Доверительное оценивание **31**
5.1 Определение 31
5.2 Интерпретация 32
5.3 Точность и надёжность интервала 33
5.4 Двухсторонний интервал через доверительные границы . . . 33
5.5 Связь с задачей проверки гипотез 33
5.6 Методы построения 33
5.7 Вопросы и задания для самоконтроля 34

6 Коэффициент корреляции и линейная регрессия 35
6.1 Вопросы и задания для самоконтроля 38

7 Справочник функций для программной реализации 38
7.1 Язык R. 39
7.2 MS Excel 41
7.3 Wolfram Mathematica 42

II Первичный статистический анализ 49

Задание 1. Выборочные характеристики 49
1 Постановка задачи 49
2 Теоретические основы 49
3 Вопросы и задания для самоконтроля 52

Задание 2. Гистограмма выборки 53
1 Постановка задачи 53
2 Теоретические основы 53
3 Вопросы и задания для самоконтроля 55

Задание 3. Эмпирическая функция распределения 55
1 Постановка задачи 55
2 Теоретические основы 56
3 Вопросы и задания для самоконтроля 57

III Проверка гипотезы о типе распределения 59

Задание 4. Критерий согласия хи-квадрат 59
1 Постановка задачи 59
2 Теоретические основы 60
3 Вопросы и задания для самоконтроля 62

Задание 5. Критерий согласия Колмогорова 63
1 Постановка задачи 63
2 Теоретические основы 63
3 Вопросы и задания для самоконтроля 64

IV Проверка гипотез однородности 65

Задание 6. Одновыборочный критерий Стьюдента **66**
1 Постановка задачи 66
2 Теоретические основы 66
3 Вопросы и задания для самоконтроля 68

Задание 7. Критерий знаков **69**
1 Постановка задачи 69
2 Теоретические основы 69
3 Вопросы и задания для самоконтроля 72

Задание 8. Двухвыборочный критерий Стьюдента **72**
1 Постановка задачи 72
2 Теоретические основы 72
3 Вопросы и задания для самоконтроля 73

Задание 9. Критерий Вилкоксона **74**
1 Постановка задачи 74
2 Теоретические основы 74
3 Вопросы и задания для самоконтроля 76

Задание 10. Критерий Фишера. Критерий сравнения дисперсий **77**
1 Постановка задачи 77
2 Теоретические основы 77
3 Вопросы и задания для самоконтроля 78

Задание 11. Критерий однородности хи-квадрат **79**
1 Постановка задачи 79
2 Теоретические основы 79
3 Больше двух выборок 80
4 Вопросы и задания для самоконтроля 81

Задание 12. Критерий однородности Смиронова **81**
1 Постановка задачи 81
2 Теоретические основы 82
3 Вопросы и задания для самоконтроля 82

V Точечное оценивание **84**

Задание 13. Метод моментов **84**
1 Постановка задачи 84
2 Теоретические основы 84
3 Пример 85

4 Вопросы и задания для самоконтроля 86

Задание 14. Метод максимального правдоподобия 86
1 Постановка задачи 86
2 Теоретические основы 86
3 Пример 88
4 Вопросы и задания для самоконтроля 89

VI Интервальные оценки 90

Задание 15. Интервальная оценка для неизвестного математического ожидания 90
1 Постановка задачи 90
2 Теоретические основы 90
3 Вопросы и задания для самоконтроля 91

Задание 16. Интервальная оценка для неизвестной дисперсии нормального распределения 91
1 Постановка задачи 91
2 Теоретические основы 92
3 Вопросы и задания для самоконтроля 92

Задание 17. Интервальная оценка для вероятности успеха 93
1 Постановка задачи 93
2 Теоретические основы 93
3 Вопросы и задания для самоконтроля 95

VII Исследование зависимости между двумя характеристиками 96

Задание 18. Проверить независимость двух характеристик по критерию сопряженности хи-квадрат 96
1 Постановка задачи 96
2 Теоретические основы 96
3 Вопросы и задания для самоконтроля 98

Задание 19. Проверка независимости двух нормальных выборок. Линейная регрессия 98
1 Постановка задачи 98
2 Теоретические основы 99
3 Вопросы и задания для самоконтроля 100

Задание 20. Ядерная оценка регрессии **101**
1 Постановка задачи 101
2 Теоретические основы 101
3 Рекомендации к программной реализации 103
4 Вопросы и задания для самоконтроля 104

Введение

При выполнении курсового проекта по математической статистике возникает много вопросов как по поводу теоретического обоснования применяемых процедур, так и по поводу их практической реализации. В данном пособии даются описания теоретических основ применения этих процедур. Кроме того, даются ссылки на полезный функционал в некоторых системах прикладного программного обеспечения.

Работу над курсовым проектом следует начать с изучения главы «Предварительные понятия и определения». Эта глава будет весьма полезна при подготовке ответов на контрольные вопросы. Выполнение каждого задания лучше всего начинать с изучения теоретического обоснования тех процедур, которые рассматриваются в этом задании. Причем желательно изучить весь материал заранее, до проведения соответствующего занятия в компьютерном классе.

Данное издание есть исправленное и дополненное пособие [1].

Обозначения

$X \sim F$ — случайная величина X имеет распределение F. В качестве F может выступать как символьное обозначение распределения (список приведён ниже), так и собственно функция распределения.

$X^{(n)} \sim F$ — каждый элемент выборки $X^{(n)}$ имеет распределение F.

A^{c} — дополнение к множеству A.

$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \in A^{\mathrm{c}} \end{cases}$ — индикатор множества A. В некоторых случаях используется запись $\mathbb{I}(A)$, в которой подразумевается, что A есть некоторое событие, которое наблюдается или не наблюдается в эксперименте.

Распределения:

- $\mathcal{N}(\mu, \sigma^2)$ — нормальное распределение со средним μ и дисперсией σ^2.
- $\mathrm{Fbin}(n, p)$ — биномиальное распределение с числом испытаний n и вероятностью успеха p.
- $\mathrm{Fbern}(p)$ — распределение Бернулли с вероятностью успеха p.
- $\mathrm{Fstud}(n)$ — распределение Стьюдента с n степенями свободы.
- $\mathrm{Ffish}(m, k)$ — распределение Фишера с параметрами m и k.
- $\mathrm{Fchisq}(n)$ — распределение Хи-квадрат с n степенями свободы.

- $\mathrm{Funif}(a, b)$ — равномерное распределение на отрезке $[a, b]$.
- Fkolm — распределение Колмогорова.

Функции распределения соответствуют обозначениям распределений с добавлением аргумента и с его расположением перед вертикальной чертой «$\mid$». Например, $\mathrm{Ffish}(x \mid m, k)$ — функция распределения Фишера с параметрами m и k, вычисленная в точке x. Исключение составляет нормальное распределение, для которого введена функция стандартного нормального $\mathcal{N}(0, 1)$ распределения, обозначаемое $\Phi(x)$.

Обратные функции обозначаются с верхним индексом $^{-1}$. Например, значение $x = \Phi^{-1}(y)$ есть решение уравнения $\Phi(x) = y$.

Список источников

[1] Симушкин С. В. Теоретические аспекты заданий курсового проекта по математической статистике. — Казань : Казанский государственный университет, 2004. — С. 68.

[2] Володин И. Н., Симушкин С. В. Лекции по теории вероятностей и математической статистике. — Казань : Издательство Казанского Университета, 2016. — С. 271.

[3] Большев Л. Н., Смирнов Н. В. Лекции по теории вероятностей и математической статистике. — М : Наука, 1983. — С. 416.

[4] Официальный сайт RStudio. — URL: https://www.rstudio.com/

[5] An Introduction to R. — URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf/

[6] Наглядная статистика. Используем R! / А. Б. Шипунов [и др.]. — М : ДМК Пресс, 2012. — С. 298. — URL: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf

[7] R Reference Card. — URL: https://cran.r-project.org/doc/contrib/Short-refcard.pdf

[8] Официальный сайт Wolfram Mathematica. — URL: http://www.wolfram.com/mathematica/

[9] Надарая Э. А. Об оценке регрессии // Теория вероятн. и ее примен. — 1964. — Т. 9, вып. 1. — С. 157—159.

Часть I
Базовые понятия и определения

1 Выборка

Предположим, что в эксперименте наблюдается случайная величина (с.в.) X. Функция распределения этой с.в.

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

неизвестна или известна с точностью до некоторого (возможно, векторного) параметра θ: $F(x) = F_\theta(x) = F(x \,|\, \theta)$, $\theta \in \Theta$, где Θ — пространство возможных значений θ (параметрическое пространство). Приведём несколько примеров.

1) Наблюдается пол новорожденного ребёнка и полагается $X = 0$ или $X = 1$, если родился мальчик или девочка соответственно. Функция $F(\cdot \,|\, \theta)$ — функция распределения Бернулли, параметр $\theta = \mathbf{P}(X = 1)$, $\Theta = [0, 1]$;

2) Измеряется коэффициент прочности X образца плавки металла. По соображениям физики процесса плавления и процесса измерения образцов можно предположить, что функция распределения $F(x \,|\, \theta) = \Phi((x - \mu)/\sigma)$, где Φ — функция распределения стандартного нормального закона, μ, σ — математическое ожидание и стандартное отклонение X; здесь двумерный параметр $\theta = (\mu, \sigma^2)$ (или $\theta = (\mu, \sigma)$), а параметрическое пространство $\Theta = \mathbb{R}^1 \times \mathbb{R}^1_+$;

3) Фиксируется время X безотказной работы электролампы. Если выход из строя лампы происходит в результате некоторых форс-мажорных обстоятельств, а не в результате «старения», то можно предположить, что функция распределения $F(x \,|\, \theta) = 1 - e^{-\theta x}$, где θ (> 0) — так называемая интенсивность отказов; пространство $\Theta = (0, \infty)$.

Выборкой объёма n называется вектор $X^{(n)} = (X_1, \ldots, X_n)$ независимых реализаций с.в. X. В дальнейшем тот факт, что элементы выборки $X^{(n)}$ имеют распределение F, будем обозначать как $X^{(n)} \sim F$. Более точно следует говорить о реализации независимых одинаково распределенных с.в. $X_1, \ldots, X_n$. Такая трактовка позволяет вычислять вероятности

тех или иных событий, связанных с выборкой. Тот факт, что эта вероятность (или соответствующие вероятностные характеристики) вычисляется при истинном значении параметра, равном θ, будет обозначаться значком θ у символов вероятности $\mathbf{P}_\theta$, мат.ожидания $\mathbf{E}_\theta$, дисперсии $\mathbf{D}_\theta$.

Задача статистического анализа состоит в принятии решений относительно распределения $F(x\,|\,\theta)$ наблюдаемой в эксперименте с.в. X. Чаще всего эта задача формулируется в терминах неизвестного значения параметра θ. Решение принимается на основе некоторой *статистики* $T = T(X^{(n)})$, которая представляет собой (не обязательно одномерную) функцию выборочных данных, независящую от неизвестных параметров вероятностной модели (неизвестного распределения).

Список определений:

- *Функция распределения* с.в. X — функция

$$F(x) = \mathbf{P}(X < x), \quad x \in \mathbb{R}^1.$$

  Если распределение с.в. X полностью характеризуется (возможно векторным) параметром $\theta \in \Theta$, то её функция распределения при конкретном значении параметра θ обозначается как

$$F(x\,|\,\theta) = \mathbf{P}_\theta(X < x),$$

  где $\mathbf{P}_\theta$ — вероятность, вычисляемая при значении параметра распределения, равном θ.

- *Случайная выборка* объема n (выборка из распределения F) — вектор $X^{(n)} = (X_1, \ldots, X_n)$ независимых одинаково распределённых случайных величин, имеющих одну и ту же функцию распределения F. Выборка $x^{(n)} = (x_1, \ldots, x_n)$ — наблюдённая или какая-либо возможная реализация случайной выборки $X^{(n)}$. Случайная выборка обозначается прописными символами $X^{(n)}$, выборка (конкретная реализация) $x^{(n)}$ — строчными.

- Задача *статистического анализа* — принятие решения относительно значений некоторых характеристик распределения на основе полученной из него выборки.

- *Статистикой* называют некоторую функцию $T(X^{(n)})$ выборочных данных, которая функционально не зависит от (неизвестных) параметров θ распределения наблюдаемой с.в. X.

1.1 Вопросы и задания для самоконтроля

1) Что такое выборка?

2) Выпишите совместную функцию распределения выборки $X^{(n)} = (X_1, \ldots, X_n)$ при условии, что известна функция распределения одной компоненты $F(x)$.

3) Что такое параметр распределения? Параметрическое пространство?

4) Что такое статистическая задача?

5) Что такое статистика?

6) Пусть дана выборка объема n из нормального $\mathcal{N}(\theta, 1)$ распределения, где $\theta \in \mathbb{R}$ — неизвестный параметр. Выпишите функцию распределения выборки, функцию плотности выборки. Какие статистики достаточно знать, чтобы вычислить функцию плотности (при известном θ)?

2 Функции распределения и квантили

2.1 Распределение

Функцией распределения случайной величины X называется функция

$$F(x) = \mathbf{P}(X < x), \quad x \in R^1,$$

т.е. вероятность попадания левее фиксированного значения x.

Функция распределения изменяется от 0 (при $x \to -\infty$) до 1 (при $x \to +\infty$), нигде не убывает и непрерывна слева. Разрыв функции распределения возможен лишь в точке, вероятность попадания в которую с.в. X не равна нулю (см. рис. 1). Величина скачка равна, как раз, вероятности попадания в эту точку.

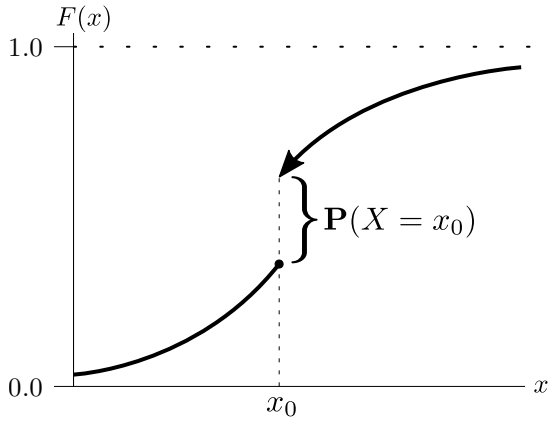

Рис. 1: Разрыв ф.р.

Функция

$$\overline{F}(x) = 1 - F(x) = \mathbf{P}(X \geqslant x),$$

т.е. вероятность попадания с.в. X правее значения x, называется функцией надёжности. Это определение связано с тем, что, если с.в. описывает время службы некоторого прибора, то $\overline{F}(x)$ есть вероятность того, что этот прибор прослужит дольше заданного времени x.

Если функция распределения имеет интегральное представление $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,\ x \in R^1$, то распределение называется *абсолютно непрерывным*, а функция f — *функцией плотности*. Для распределений, используемых на практике, в тех точках, где функция распределения дифференцируема, в качестве плотности можно взять производную функции распределения: $f(x) = F'(x)$; в остальных точках плотность можно выбрать произвольным образом, например положить $f(x) = 0$.

Функция плотности обладает следующими свойствами:

1) $f(x) \geqslant 0$, и полный интеграл

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

2) Если функция f чётна, то распределение симметрично около нуля, то есть с.в. X имеет такое же распределение, что и с.в. $-X$. График её функции распределения симметричен около точки $(0, 1/2) : F(-x) = 1 - F(x),\ \forall x \in R^1$. Таким образом, при построении таблиц симметричного распределения достаточно ограничиться лишь положительными значениями аргумента.

3) Значение функции распределения $F(x)$ равно площади под графиком функции плотности в интервале от $-\infty$ до x (см. рис. 2).

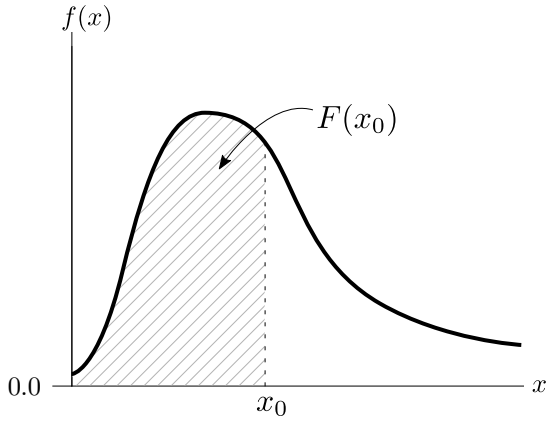

Рис. 2: Связь между функцией распределения $F(x)$ и функцией плотности $f(x)$

2.2 Квантѝли — процентные точки

Решение уравнения $F(t) = p$ (если оно существует) называют *p-квантѝлью* (или квантѝлью порядка p) распределения F; обозначают как $Q(p)$ (от английского Quantile). Если функция распределения имеет обратную, то

$$Q(p) = F^{-1}(p).$$

Если указанного решения не существует или таких решений много, то способ определения квантили можно легко понять из графического представления рис. 3.

Квантиль порядка $(1-p)$ иногда называют $p \cdot 100\%$*-ой точкой распределения*. Таким образом, скажем, 95%-ая точка есть такое значение, для которого вероятность попадания с.в. правее равна 0.05.

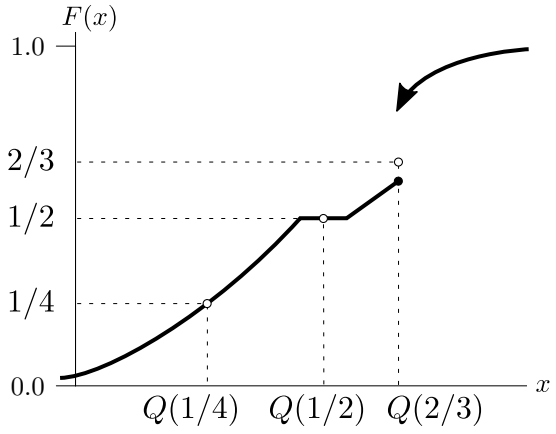

Рис. 3: p-квантиль

Замечание. В приложениях квантиль ещё называют нижней квантилью, дабы подчеркнуть, что это есть точка, вероятность попадания левее (ниже) которой равна p (левый хвост распределения). По тем же соображениям квантиль порядка $(1-p)$ называют верхней квантилью порядка p (правый хвост распределения). Вариант квантили, который даёт таблица или соответствующая функция используемого пакета программ, можно узнать, обратившись к справочнику.

2.3 Вопросы и задания для самоконтроля

1) Что такое функция распределения?

2) Пользуясь функцией распределения $F(x)$ с.в. X и предполагая её непрерывность, выпишите следующие вероятности: $\mathbf{P}(X \le x)$, $\mathbf{P}(X > x)$, $\mathbf{P}(|X| < x)$, $\mathbf{P}(y < X < x)$, $\mathbf{P}(X < x \cup X > > y)$, $\mathbf{P}(X = x)$, $\mathbf{P}(X \in [a,b])$, $\mathbf{P}(X \notin [a,b])$. Можно ли представить $\mathbf{P}(X < -x)$, $\mathbf{P}(|X| < x)$ в терминах $F(x)$ в дополнительных условиях симметричности распределения X?

3) Что такое функция плотности?

4) Что такое функция надёжности?

5) Что такое квантиль распределения?

6) Пользуясь книгами [2] и [3], выясните определение следующих распределений: равномерное, биномиальное, показательное, нормальное, Стьюдента, хи-квадрат, Фишера, Колмогорова. Как соотносятся между собой случайные величины, имеющие эти распределения?

7) Найдите q-квантиль с.в., имеющей экспоненциальное распределение с функцией плотности $f(x) = (\theta)^{-1}e^{-x/\theta}, x > 0$, где $\theta > 0$ — некоторый параметр.

3 Проверка гипотез

3.1 Гипотеза

Прежде всего выдвигается так называемая «нулевая гипотеза» H_0 о том, что распределение F удовлетворяет некоторому свойству. При этом, если есть возможность, желательно сразу конкретизировать альтернативное утверждение H_1 относительно F. Например,

1) $\mathrm{H}_0\colon F = F_0$ — распределение F в точности совпадает с некоторым известным распределением F_0. Так, вполне уместно рассмотреть гипотезу о том, что случайные значения, выдаваемые функцией Random или Rand некоторого языка программирования, действительно имеют равномерное на отрезке $[0, 1]$ распределение.

2) $\mathrm{H}_0\colon F \in \Psi$ — распределение F принадлежит некоторому известному семейству распределений Ψ. Весьма популярна на практике гипотеза о том, что наблюдаемая в эксперименте выборка имеет нормальное распределение с некоторыми неизвестными значениями среднего и дисперсии. Варианты альтернатив H_1, как в этом, так и в предыдущем примере, очень разнообразны и зависят от предпочтений исследователя. Наиболее общая альтернатива здесь — семейство всех возможных распределений на числовой прямой (в предыдущем примере — на отрезке $[0, 1]$). Отметим, что при такой альтернативе говорят не о проверке гипотезы H_0, а о проверке согласия данных с выдвинутой гипотезой H_0.

3) $\mathrm{H}_0\colon \theta \in \Theta_0$ — значение неизвестного параметра θ принадлежит некоторому подмножеству Θ_0. В качестве примера здесь можно привести задачу проверки гипотезы о том, что вероятность рождения мальчика больше $1/2$, при альтернативе — меньше или равна $1/2$. Другой пример, связанный с медициной, — исследование эффективности нового препарата при лечении пациентов с повышенным артериальным давлением в сравнении со стандартной методи-

кой лечения. Здесь в качестве нулевой гипотезы лучше выдвинуть утверждение, что новый препарат идентичен старому (или даже хуже), при альтернативе, что новый лучше старого.

Термин «нулевая» гипотеза связан с тем, что чаще всего к этой гипотезе относят предположения об отсутствии эффекта воздействия или отсутствии различий — лекарство не приводит к улучшению показателей здоровья, обработка не даёт дополнительного увеличения срока службы прибора, между признаками нет зависимости, вероятность рождения девочки равна $1/2$ и т.п. В общем случае в нулевую гипотезу включают значения параметра θ, удовлетворяющие условию $\theta = \theta_0$, где θ_0 — некоторая «норма». Отметим, что нулевая гипотеза не обязательно состоит из одной «нулевой» точки $\theta = \theta_0$, но обязательно эту точку содержит.

Список определений:

- *Статистической гипотезой* H называется утверждение относительно распределения наблюдаемой в эксперименте с.в.
- Если распределение наблюдаемой с.в. принадлежит параметрическому семейству $\{F(\theta),\ \theta \in \Theta\}$, то гипотеза формулируется в виде утверждения о параметре θ и представляет собой некоторое подмножество параметрического пространства Θ.
- В задачах проверки гипотезы обычно рассматривают пары гипотез $\mathrm{H}_0 : \theta \in \Theta_0$ (*нулевая гипотеза*) и $\mathrm{H}_1 : \theta \in \Theta_1$ (*альтернативная гипотеза*) таких, что $\Theta_0 \cap \Theta_1 = \varnothing$; часто $\Theta_1 = \Theta_0^{\mathrm{C}}$ (т.е. альтернатива утверждает нечто противоположное нулевой гипотезе).
- H — *простая гипотеза*, если она состоит из одного элемента (распределения, параметра); H — *сложная гипотеза*, если она состоит из нескольких элементов.

3.2 Критерий и критическая область. Вероятности ошибок

После выдвижения гипотезы и до проведения наблюдений строится критерий проверки этой гипотезы. *Критерий* — функция выборочных данных $\varphi(x^{(n)})$, принимающая значение 1, если данные таковы, что нулевую гипотезу следует отвергнуть, и значение 0, если нулевую гипотезу следует принять. Вместо критерия часто строят *критическую область* $A = \{x^{(n)} : \varphi(x^{(n)}) = 1\}$ — область, при попадании в которую вы-

борочных данных нулевая гипотеза отвергается. Ясно, что критерий φ есть индикаторная функция критической области: $\varphi(x^{(n)}) = \mathbb{I}_A(x^{(n)})$.

Обычно критерий строится с помощью некоторой тестовой статистики $T(x^{(n)})$, и в этом случае гипотеза отвергается, если значение статистики больше (или меньше) критической константы $C_{\text{крит}}$. Другими словами, критическая область

$$A = \{x^{(n)} \colon T > C_{\text{крит}}\} \quad \text{или} \quad A = \{x^{(n)} \colon T < C_{\text{крит}}\}.$$

Величина константы $C_{\text{крит}}$ выбирается (заранее до проведения наблюдений) в соответствии с требованиями на качество критерия.

Качество критерия (критической области) характеризуется двумя величинами — вероятностью ошибки первого рода и вероятностью ошибки второго рода.

Вероятность ошибки 1-го рода — вероятность отвергнуть нулевую гипотезу, если на самом деле она верна. *Вероятность ошибки 2-го рода* — вероятность принять нулевую гипотезу, если на самом деле верна альтернатива.

Эти определения вероятностей ошибок немного упрощённые и соответствуют рассматриваемой задаче только, когда нулевая гипотеза и её альтернатива простые, т.е. состоят только из одного распределения F_θ (одного параметра θ). Например, нулевая гипотеза — вероятность появления потомков с рецессивным признаком в опытах Менделя равна $\theta = 1/2$, альтернативная гипотеза — вероятность $\theta = 1/4$, или нулевая гипотеза — датчик случайных чисел выдаёт числа с равномерным распределением. Очень часто или нулевая гипотеза, или альтернатива содержат более одного неизвестного параметра — $\theta \in \Theta_0$ или $\theta \in \Theta_1$. Поэтому вероятности ошибок суть функции параметра $\theta \in \Theta$. Рассмотрим так называемую *функцию мощности*

$$\mathrm{m}_\varphi(\theta) = \mathbf{P}_\theta(X^{(n)} \in A) = \mathbf{E}_\theta\, \varphi(X^{(n)}), \quad \theta \in \Theta,$$

— вероятность отвержения нулевой гипотезы, когда истинное значение параметра равно θ. Тогда вероятность ошибки 1-го рода есть часть функции мощности при значениях параметра $\theta \in \Theta_0$; вероятность ошибки 2-го рода

$$\mathbf{P}_\theta(X^{(n)} \notin A) = 1 - \mathrm{m}_\varphi(\theta), \quad \theta \in \Theta_1.$$

Максимальная вероятность ошибки 1-го рода среди всех распределений, удовлетворяющих требованиям нулевой гипотезы,

$$\overline{\mathrm{m}}_\varphi = \sup_{\theta \in \Theta_0} \mathrm{m}_\varphi(\theta)$$

называется *размером* критерия φ.

Критерий, вероятность ошибки 1-го рода (размер) которого не превосходит некоторого наперёд заданного малого значения α, называется *критерием уровня* α, а заданная заранее константа α называется *уровнем значимости*. Таким образом, для критерия уровня α маловероятно (не больше α) отвержение справедливой нулевой гипотезы:

$$\mathbf{P}_F(\mathrm{H}_0 \text{ отвергается}) \leqslant \alpha \quad \text{для всех} \quad F \in \mathrm{H}_0.$$

Другими словами, нулевая гипотеза должна отвергаться только, если свидетельство экспериментальных данных против этой гипотезы будет достаточно существенным — значимым.

Поскольку чаще всего мы можем контролировать только ошибку первого рода, то обычно в качестве нулевой гипотезы выбирают утверждение, противоположное ожидаемому. Например, при исследовании нового лекарственного препарата следует проверять нулевую гипотезу о его полной неэффективности. В этом случае критерий заданного уровня будет обеспечивать низкую вероятность принятия неэффективного препарата.

Чем меньше уровень значимости, тем выше доверие к выводу в пользу альтернативы, например к выводу об эффективности нового лекарства. Наиболее популярное на практике значение $\alpha = 0.05$ (5%-ый уровень значимости); иногда рассматривают $\alpha = 0.10$ и $\alpha = 0.01$.

Все статистические критерии устроены так, что с уменьшением уровня значимости критическая область этих критериев уменьшается, что приводит к уменьшению всей функции мощности и, следовательно, к увеличению вероятности ошибки 2-го рода. Контроль обеих вероятностей ошибок возможен только при различении разделённых (не имеющих общих граничных точек) гипотез и при достаточно большом числе наблюдений.

Список определений:

- *Критерий* $\varphi(x^{(n)})$ — функция выборочных данных (статистика), принимающая значение 1 и 0, где 1 интерпретируется как решение об отвержении H_0, а 0 — о принятии H_0.

- *Критическая область* A *критерия* φ:

  $$A = \{x^{(n)} \colon \varphi(x^{(n)}) = 1\} \quad —$$

  область, при попадании в которую выборочных данных нулевая гипотеза отвергается.

- *Вероятность ошибки 1-го рода*:

  $$\mathbf{P}_\theta(X^{(n)} \in A) = \mathbf{E}_\theta\, \varphi(X^{(n)}), \quad \theta \in \mathrm{H}_0.$$

Вероятность ошибки 2-го рода:

$$\mathbf{P}_\theta(X^{(n)} \notin A) = 1 - \mathbf{E}_\theta\, \varphi(X^{(n)}), \quad \theta \in \mathrm{H}_1$$

- *Функция мощности критерия*:

$$\mathrm{m}_\varphi(\theta) = \mathbf{P}_\theta(X^{(n)} \in A) = \mathbf{E}_\theta\, \varphi(X^{(n)}), \quad \theta \in \Theta.$$

- *Размер критерия*: $\overline{\mathrm{m}}_\varphi(\varphi) = \sup_{\theta\in\Theta_0} \mathrm{m}(\theta)$.

- *Уровень значимости* — выбранная заранее верхняя граница на величину вероятности ошибки 1-го рода, которую исследователь считает приемлемой для рассматриваемой задачи.

- Если для заданного уровня значимости α размер критерия $\overline{\mathrm{m}}_\varphi \leqslant \alpha$, то критерий φ называется *критерием уровня* α.

3.3 Тестовая статистика

Для построения критерия (критической области) используют обычно какую-либо тестовую статистику $T = T(x^{(n)})$, которая чаще всего устроена так, что её большие значения в большей степени свидетельствуют в пользу альтернативы. Приведём два примера.

1) Фармакологическая фирма разработала вакцину от гриппа. До применения вакцины 63% всего населения в зимний сезон страдало от этого заболевания. Для проверки качества новой вакцины планируется проведение испытаний на контрольной группе пациентов. По результатам испытаний необходимо проверить нулевую гипотезу $\mathrm{H}_0 : \theta \geqslant \theta_0$ при альтернативе $\mathrm{H}_1 : \theta < \theta_0$, где θ — вероятность заболевания после применения вакцины, $\theta_0 = 0.63$. Ясно, что чем меньше заболевших среди испытуемых, тем больше уверенность в хорошем качестве вакцины. Поэтому выбираем тестовую статистику $T = \theta_0 - \widetilde{\theta}$, где $\widetilde{\theta}$ — относительная доля заболевших среди испытуемых. Заметим, что если нулевая гипотеза $\mathrm{H}_0 : \theta \leqslant \theta_0$ (скажем, для θ равного вероятности не заболеть и $\theta_0 = 0.37$), то тестовая статистика $T = \widetilde{\theta} - \theta_0$.

2) Если требуется проверить нулевую гипотезу о том, что значения, выдаваемые датчиком случайных чисел, имеют равномерное распределение на отрезке $[0, 1]$, то в качестве тестовой статистики можно рассмотреть $T = \max_x |F_n(x) - F_0(x)|$, где F_n — эмпирическая функция распределения, построенная по выборке объёма n, $F_0(x) = x$ — функция распределения равномерного закона.

Критическая область, основанная на тестовой статистике такого типа, будет иметь вид $A = \{x^{(n)} : T(x^{(n)}) > C_{\text{крит}}\}$. *Критическая константа* $C_{\text{крит}}$ выбирается из условия на вероятность ошибки 1-го рода:

$$\mathbf{P}_F(T(X^{(n)}) > C_{\text{крит}}) \leqslant \alpha, \tag{1}$$

где неравенство должно выполняться для всех распределений F, удовлетворяющих предположениям гипотезы H_0.

Замечание. Не зная $C_{\text{крит}}$, нельзя проверить гипотезу, как бы не было соблазнительно значение T, полученное в эксперименте. Например, если из двух баскетболистов один попал в 66% бросков с игры, а второй — только в 33%, то ещё нельзя сказать, что первый баскетболист «лучше» второго, так как для подобного вывода надо знать (чтобы вычислить $C_{\text{крит}}$) количество бросков каждого из игроков (сравните ситуацию, когда они сделали всего по 3 броска, с ситуацией, когда таких бросков было 1000).

Тестовые статистики, обычно используемые на практике, устроены так, что размер такого критерия (критической области) равен вероятности ошибки 1-го рода, вычисленной при некотором «нулевом» распределении F_0. Таким образом, для того, чтобы найти критическую константу, необходимо обладать информацией о точном или приближённом распределении тестовой статистики в предположении, что выборка получена из распределения F_0 (или из распределения с истинным значением параметра, равном θ_0).

Так, во втором из приведённых выше примеров распределение статистики T в предположении, что выборка поступает из равномерного распределения, может быть приближено при больших значениях $n\,(> 50)$ распределением Колмогорова, таблицы которого часто приводят в справочниках по математической статистике, но, по какой-то причине, не включают в пакеты статистических программ. Таким образом, значение критической константы приближённо совпадает с квантилью порядка $(1 - \alpha)$ распределения Колмогорова.

В первом примере удобнее перейти от тестовой статистики $T = \theta_0 - \widetilde{\theta}$ к статистике S равной количеству испытуемых, заболевших гриппом после вакцинации. Необходимость такого перехода объясняется тем, что статистика S имеет хорошо изученное биномиальное распределение. Относительно этой статистики критическая область будет выглядеть так:

$$A = \{x^{(n)} : S(x^{(n)}) < C\},$$

т.е. вакцинацию следует признавать полезной, если количество заболевших в контрольной группе мало. Критическая константа C для такого

критерия есть α-квантиль биномиального распределения с параметрами (n, θ_0).

Как показывает предыдущий пример, хотя тестовую статистику T можно выбрать так, что нулевая гипотеза будет отвергаться при $T > C_{\text{крит}}$, однако очень часто эта статистика связана с некоторой другой статистикой, для которой уже известна функция распределения. Поэтому критическую область удобнее представлять через эту известную статистику. Например, для статистики $T_1 = |\overline{x} - \theta_0|$ критическая область $A = \{|\overline{x} - \theta_0| > C\}$, где $\overline{x}$ — выборочное среднее арифметическое, θ_0 — истинное (с точки зрения нулевой гипотезы) значение математического ожидания, может быть представлена в виде

$$A = \{\overline{x} < \theta_0 - C\} \cup \{\overline{x} > \theta_0 + C\}$$

уже относительно статистики $T_2 = \overline{x}$.

Итак, если относительно некоторой статистики T нулевая гипотеза должна отвергаться:

- если $T < C_{\text{крит}}$, то критическая константа $C_{\text{крит}}$ находится как решение уравнения

  $$\mathbf{P}_0(T < C_{\text{крит}}) = \alpha,$$

  т.е. $C_{\text{крит}}$ представляет собой α-квантиль $Q(\alpha)$ распределения статистики T при «нулевом» распределении выборочных данных (см. пояснение выше о способе вычисления размера критерия);

- если $T > C_{\text{крит}}$, то критическая константа $C_{\text{крит}}$ находится как решение уравнения

  $$\mathbf{P}_0(T > C_{\text{крит}}) = \alpha, \tag{2}$$

  т.е. $C_{\text{крит}}$ представляет собой $(1-\alpha)$-квантиль $Q(1-\alpha)$ того же распределения статистики T;

- если $|T| > C_{\text{крит}}$, то критическая константа $C_{\text{крит}}$ находится как решение уравнения

  $$\mathbf{P}_0(T < -C_{\text{крит}}) + \mathbf{P}_0(T > +C_{\text{крит}}) = \alpha.$$

  Если «нулевое» распределение статистики T симметрично около точки 0, то $\mathbf{P}_0(T < -C_{\text{крит}}) = \mathbf{P}_0(T > C_{\text{крит}})$, следовательно, критическая константа находится из уравнения

  $$2\,\mathbf{P}_0(T > C_{\text{крит}}) = \alpha,$$

  т.е. $C_{\text{крит}}$ есть $(1-\alpha/2)$-квантиль $Q(1-\alpha/2)$ распределения статистики T. Для несимметричных распределений можно изменить критическую область к виду $A = \{T < C_1\} \cup \{T > C_2\}$ и выбрать C_1 как квантиль порядка $\alpha/2$, а C_2 — как квантиль порядка $(1-\alpha/2)$ (в сумме получим, как раз, α).

3.4 p-значение (критический уровень значимости)

Критическая константа критерия, основанного на тестовой статистике T, зависит от выбранного уровня значимости. Легко понять, что с увеличением уровня значимости критическая область такого критерия расширяется от пустого множества при $\alpha = 0$ до всего выборочного пространства при $\alpha = 1$. Следовательно, для вычисленного по экспериментальным данным значения тестовой статистики $T = t_{\text{эксп}}$ найдётся такое число $\mathsf{p} \in [0; 1]$, зависящее от $t_{\text{эксп}}$, что при уровне значимости $\alpha \geqslant \mathsf{p}$ нулевая гипотеза будет отвергаться (данные попадают в критическую область), а при $\alpha < \mathsf{p}$ — приниматься (данные не попадают в критическую область). Наименьший уровень значимости, при котором критерий, основанный на тестовой статистике T, отвергает нулевую гипотезу для полученного экспериментального значения $T = t_{\text{эксп}}$, называется *p-значением* (p-value; раньше в русскоязычной литературе использовались названия критический уровень значимости, достигнутый уровень значимости, фактический уровень значимости).

Если тестовая статистика показывает степень несогласованности данных и нулевой гипотезы, то p-значение можно интерпретировать как вероятность получения более значимых различий, чем те, что получены в текущем эксперименте.

Из определения понятно, что любой критерий можно описать эквивалентным образом через соответствующее этому критерию p-значение: критерий размера α отвергает нулевую гипотезу тогда и только тогда, когда p-значение $\mathsf{p} \leqslant \alpha$.

Способ вычисления p-значения зависит от вида критической области. В общем случае p-значение можно найти, слегка видоизменив уравнение (2), — константу $C_{\text{крит}}$ следует заменить на полученное по выборочным данным значение тестовой статистики: $C_{\text{крит}} \to t_{\text{эксп}}$, а уровень значимости на искомое p-значение: $\alpha \to \mathsf{p}$. Таким образом, если G_T — функция распределения статистики T, то

- для критической области вида $T < C$ p-значение
$$\mathsf{p} = G_T(t_{\text{эксп}});$$
- для критической области вида $T > C$ p-значение
$$\mathsf{p} = 1 - G_T(t_{\text{эксп}});$$
- для критической области вида $|T| > C$ p-значение
$$\mathsf{p} = G_T(-|t_{\text{эксп}}|) + 1 - G_T(|t_{\text{эксп}}|).$$

Если распределение статистики T симметрично около нуля, то

$$\mathsf{p} = 2\big(1 - G_T(|t_{\text{эксп}}|)\big).$$

Основное преимущество этого способа (кроме очевидного преимущества, связанного с вычислением прямой, а не обратной функции распределения) состоит в том, что мы можем не делать жёстких выводов типа «да, гипотеза верна» или «нет, гипотеза не верна», а принять более гибкое решение. Здесь можно предложить следующую градацию высказываний о справедливости нулевой гипотезы в зависимости от p-значения:

- $\mathsf{p} < 0.01$ — высоко значимое расхождение с гипотезой;
- $0.01 \leqslant \mathsf{p} \leqslant 0.05$ — значимое расхождение с гипотезой;
- $0.05 < \mathsf{p} \leqslant 0.10$ — слабо значимое расхождение с гипотезой;
- $0.10 < \mathsf{p} \leqslant 0.15$ — нет оснований отвергать нулевую гипотезу;
- $0.15 < \mathsf{p}$ — хорошее согласие с нулевой гипотезой.

3.5 Общая схема построения статистических критериев

Отчёт о применении того или иного критерия должен в обязательном порядке содержать следующие пункты.

1) Описание (физическое, экономическое, медицинское, биологическое и т.п.) объекта исследования и ожидания исследователя.

2) Описание наблюдаемой случайной величины и её распределения, а также ожидания исследователя относительно этого распределения (или его параметров).

3) Выдвигаемая нулевая гипотеза H_0 и выбранный уровень значимости α.

4) Применяемая тестовая статистика с рассуждениями о её значениях при справедливости той или иной гипотезы.

5) Вид критической области. Способ нахождения критической константы — из условия на вероятность ошибки первого рода.

6) Вид распределения тестовой статистики T при нахождении размера критерия.

7) Значение критической константы — квантиль распределения тестовой статистики соответствующего порядка. Критическая область критерия заданного уровня.

8) Вычисленное по данным значение T.

9) Вывод в пользу нулевой или альтернативной гипотезы.

10) Вычисление p-значения через распределение тестовой статистики

11) Подтверждение вывода пункта 9) на основе p-значения.

Приведём пример.

1) С целью изменения предела прочности металлических дисков предлагается подвергнуть металл, из которого изготовляются диски, специфической обработке. Ожидается, что после обработки прочность диска повысится.

2) Для анализа воздействия проводятся измерения предела прочности $n = 15$ образцов $(X_1, \ldots, X_n)$, изготовленных из необработанного металла, и $m = 21$ образцов $(Y_1, \ldots, Y_m)$, изготовленных из обработанного металла. Так как прочность металла пропорциональна количеству нарушений атомических связей в кристаллической решётке (каковое количество имеет биномиальное распределение), то в соответствии с теоремой Муавра–Лапласа можно предположить, что измерение прочности образцов будет иметь нормальное распределение: $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2), i = \overline{1,n}$ и $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2), i = \overline{1,m}$. Если измерение предела прочности и отливка образцов в обоих случаях производятся некоторым стандартным способом, то можно считать, что дисперсии наблюдений совпадают: $\sigma_1^2 = \sigma_2^2$. Таким образом, ожидания разработчиков нового способа изготовления дисков формализуются в виде неравенства $\mu_1 - \mu_2 < 0$ относительно математических ожиданий нормальных (гауссовских) наблюдений.

3) Требуется проверить нулевую гипотеза $\mathrm{H}_0 : \theta \geqslant 0$ при альтернативе $\mathrm{H}_1 : \theta < 0$, где $\theta = (\mu_1 - \mu_2)$ — изменение предела прочности металла после обработки. Проконсультировавшись с заказчиком данного исследования, решено выбрать уровень значимости $\alpha = 0.04$.

4) Поскольку данные имеют нормальное распределение с одинаковой для обеих групп дисперсией, то можно применить критерий Стьюдента. Тестовая статистика этого критерия имеет вид

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где $\overline{X}, \overline{Y}$ — выборочные средние, S_X^2, S_Y^2 — выборочные дисперсии (несмещённые оценки) соответствующих наблюдений. Так как $\overline{X}$

есть оценка математического ожидания μ_1, а $\overline{Y}$ — оценка математического ожидания μ_2, то при справедливости нулевой гипотезы ожидаются значения разности $\overline{X} - \overline{Y} \geqslant 0$.

5) Критическая область, т.е. область значений тестовой статистики, при которых нулевая гипотеза должна отвергаться и приниматься альтернатива, имеет вид

$$T < C_{\text{крит}},$$

где критическая константа C находится из условия на вероятность ошибки первого рода:

$$\mathbf{P}_\theta(T < C_{\text{крит}}) \leqslant 0.04, \quad \forall \theta \ (= \mu_1 - \mu_2) \geqslant 0.$$

6) Функция распределения статистики T убывает с ростом параметра θ. Следовательно, размер критерия Стьюдента равен вероятности ошибки 1-го рода, вычисленной при $\theta = 0$. Известно, что в этом случае статистика T имеет распределение Стьюдента с $(n + m - 2)$ степенями свободы.

7) Критическая константа есть квантиль порядка 0.04 распределения Стьюдента с 34 степенями свободы. Воспользовавшись функцией СТЬЮДЕНТ.ОБР(0.04; 34) пакета Excel (2010), находим, что $C_{\text{крит}} = -1.805$; критическая область: $T < -1.805$.

8) Статистический анализ полученных данных:

	без обработки	с обработкой
объём выборки	$n = 15$	$m = 21$
среднее значение	$\overline{X} = 65.01$	$\overline{Y} = 66.39$
дисперсия	$S_X^2 = 4.56$	$S_Y^2 = 4.04$
станд.ошибка среднего	$m_X = 0.55$	$m_Y = 0.44$
значение статистики Стьюдента:		$t = -1.98$
4%-я критическая область:		$T < -1.805$

9) Вывод: нулевая гипотеза отвергается.
Претензии разработчиков средства обработки оправданы.

10) p-значение находится как значение функции распределения Стьюдента с 34 степенями свободы при $t = -1.98$. С помощью встроенной функции СТЬЮДЕНТ.РАСП($-$ 1.98; 34; 1) пакета Excel получаем $\mathsf{p} = 0.028$.

11) Вывод: повышение предела прочности после обработки статистически значимо ($\mathsf{p} = 0.028$).

3.6 Вопросы и задания для самоконтроля

1) Что такое гипотеза?

2) В чём заключается задача проверки гипотез?

3) Как следует выбирать нулевую гипотезу?

4) Что такое критерий?

5) Что такое критическая область критерия?

6) Как определяется вероятность ошибки 1-го рода? Что такое размер критерия?

7) Что такое уровень значимости?

8) Какой уровень значимости лучше выбрать — 5% , 10% или 1%?

9) Является ли 5%-ая критическая область 10%-ой?

10) Что такое критическая константа критерия и как она связана с критической областью?

11) Как часто мы будем ошибаться, если будем применять критерий уровня $\alpha = 0.03$?

12) Что такое p-значение?

13) Как проверить гипотезу, основываясь на p-значении?

14) Можно ли признать новый метод лечения лучше старого, если при клинических испытаниях результативность нового метода составила 85%, а старого — 70%? Что ещё нужно знать, чтобы правильно ответить на этот вопрос?

4 Точечное оценивание

Пусть $X^{(n)} = (X_1, \ldots, X_n)$ — выборка из некоторого распределения $F(\theta)$ с параметром распределения θ (см. примеры ниже). *Оценочной функцией* (коротко, *оценкой*) параметра θ называется статистика $\hat{\theta} = \hat{\theta}(X^{(n)})$, принимающая значения в пространстве Θ возможных значений этого параметра. Среди свойств оценок выделяют обычно три основных: несмещённость, состоятельность и оптимальность.

4.1 Несмещённость

Оценочная функция $\hat{\theta}$ называется *несмещённой* (в среднем), если математическое ожидание

$$\mathbf{E}_\theta\,\hat{\theta} = \theta \quad \text{для всех} \quad \theta \in \Theta,$$

где, как обычно, индекс у знака математического ожидания означает, что вычисления производятся при истинном значении параметра, равном θ. Другими словами, несмещённая оценка в среднем совпадает с истинным значением параметра.

Равенство $\mathbf{E}_\theta(\hat{\theta} - \theta) = 0$ можно интерпретировать как отсутствие систематической ошибки при использовании несмещённой оценочной функции.

В смысле сренеквадратического отклонения несмещённая оценка находится ближе к истинному значению параметра, чем к какому-то другому: $\mathbf{E}_\theta(\hat{\theta} - \theta)^2 \leqslant \mathbf{E}_\theta(\hat{\theta} - \theta')^2$ для любого $\theta' \neq \theta$.

Рассмотрим примеры.

1) Выборочное среднее $\overline{X}$ есть несмёщенная оценка истинного математического ожидания распределения $\mu = \mathbf{E}\,X$:

$$\mathbf{E}_\mu\,\overline{X} = \mu.$$

2) Относительная частота осуществления некоторого события A (количество ν появлений события в выборке, делённое на общее число выборочных данных n) есть несмещённая оценка вероятности события $p = \mathbf{P}(A)$:

$$\mathbf{E}_p\,\frac{\nu}{n} = p.$$

В частности, эмпирическая функция распределения $F_n(x) = F_n\bigl(x \mid X^{(n)}\bigr)$ в каждой точке $x \in R^1$ есть несмещённая оценка истинной функции распределения $F(x)$: $\mathbf{E}_F\,F_n(x) = F(x)$.

3) Выборочная дисперсия S^2 — смещённая оценка истинной дисперсии $\sigma^2 = \mathbf{D}\,X = E(X-\mu)^2$:

$$\mathbf{E}_{\sigma^2}\,S^2 = \frac{n-1}{n}\sigma^2.$$

На практике чаще всего рассматривают так называемую исправленную на несмещённость выборочную дисперсию $\widetilde{S}^2 = \frac{n}{n-1}S^2$.

4) Только для симметричных вероятностных моделей выборочная медиана $M_n = \mathrm{median}(X^{(n)})$ будет несмещённой оценкой истинной медианы m. Про эту оценку также можно сказать, что она не имеет

систематической ошибки, т.к. для неё обе вероятности — и вероятность попадания левее m и вероятность попадания правее m, не превосходят $1/2$:

$$\mathbf{P}_m(M_n < m) \leqslant \frac{1}{2} \geqslant \mathbf{P}_m(M_n > m).$$

Оценки более сложных характеристик распределения в большинстве своём получаются смещёнными. В этом случае на применяемую оценку накладывают условие асимптотической несмещённости (по объёму выборки n):

$$\lim_{n\to\infty} \mathbf{E}_\theta \hat{\theta} = \theta.$$

Этому требованию удовлетворяют уже почти все используемые на практике оценки.

4.2 Состоятельность

Оценочная функция $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ называется *состоятельной* оценкой параметра $\theta \in \Theta$, если при увеличении объёма выборки $n \to \infty$ для любого $\theta \in \Theta$ имеет место сходимость по вероятности $\mathbf{P}_\theta$

$$\hat{\theta}_n \xrightarrow{\mathbf{P}_\theta} \theta.$$

Другими словами, состоятельная оценка с ростом объёма выборки n приближается к истинному значению оцениваемого параметра.

Замечание. Правильнее, конечно, говорить о состоятельности последовательности оценочных функций. Приведённый здесь сокращённый вариант определения воспринимается адекватно всеми специалистами математической статистики.

Большинство оценок вычисляются либо непосредственно как сумма некоторых функций от выборочных наблюдений, либо как непрерывная функция от таких сумм. Для исследования состоятельности подобных оценок полезна теорема, известная в теории вероятностей как Закон Больших Чисел, утверждающая, что среднее арифметическое с ростом числа слагаемых стремится по вероятности или почти наверное к теоретическому среднему (математическому ожиданию) своих слагаемых. Все рассмотренные выше оценки будут состоятельными именно в силу закона больших чисел. В качестве дополнительного примера можно ещё привести выборочное стандартное отклонение $S = \sqrt{S^2}$. Эта оценка будет состоятельной для истинного стандартного отклонения σ, так как она есть непрерывная функция состоятельной оценки S^2.

4.3 Оптимальность

Точность оценок характеризуется величиной средней ошибки. Теоретически наиболее полно исследовано сравнение оценок на основе среднеквадратической ошибки (среднеквадратического риска):

$$\mathcal{R}(\hat{\theta}, \theta) = \mathbf{E}_\theta(\hat{\theta} - \theta)^2.$$

Заметим, что здесь $\hat{\theta}$ — с.в., а θ — число. Для несмещённых оценок среднеквадратическая ошибка есть ни что иное, как дисперсия оценки.

Понятно, что предпочтительнее выбирать оценки, у которых средняя ошибка принимает наименьшее значение при любых значениях параметра θ среди всех допустимых оценок. К сожалению, это трудно достижимо, а точнее — никогда не достижимо. В качестве доказательства этого положения рассмотрим оценочную функцию $\hat{\theta} \equiv 7.5$, которая, конечно, имеет плохой риск при $\theta \neq 7.5$, однако в точке $\theta = 7.5$ её риск равен нулю, что невозможно улучшить. В связи с этим задачу построения оценки с минимальным риском рассматривают в классе оценочных функций, удовлетворяющих некоторому дополнительному свойству, например свойству несмещённости. Доказано, что для большинства практически полезных вероятностных моделей все рассмотренные выше несмещенные оценки будут иметь минимальную дисперсию.

Интересно отметить здесь, что с точки зрения среднеквадратического риска качество смещённой оценки дисперсии S^2 выше качества её несмещённого варианта $\hat{S}^2$. Так что, как поётся в известной песне, — «думайте сами, решайте сами иметь несмещённость или не иметь».

Список определений:

- Оценочная функция $\hat{\theta}(X^{(n)})$ называется *несмещённой* оценкой параметра θ, если для любого фиксированного θ (для любого истинного значения параметра)

  $$\mathbf{E}_\theta\, \hat{\theta}(X^{(n)}) = \theta.$$

  Оценка $\hat{\theta}(X^{(n)})$ называется асимптотически несмещённой, если:

  $$\lim_{n\to\infty} \mathbf{E}_\theta\, \hat{\theta}(X^{(n)}) = \theta.$$

- Оценочная функция $\hat{\theta}(X^{(n)})$ называется *состоятельной* оценкой параметра θ, если для любого истинного значения параметра θ при $n \to \infty$

  $$\hat{\theta}(X^{(n)}) \xrightarrow{\mathbf{P}_\theta} \theta, \quad \text{т.е.} \quad \lim_{n\to\infty} \mathbf{P}_\theta \left(|\hat{\theta}(X^{(n)}) - \theta| < \varepsilon \right) = 1, \quad \forall \varepsilon > 0.$$

- Функция *среднеквадратического риска*: $\mathcal{R}(\hat{\theta}, \theta) = \mathbf{E}_\theta(\hat{\theta} - \theta)^2$.

  Несмещённая оценка θ^* называется *оценкой с минимальной дисперсией*, если для любой другой несмещённой оценки $\hat{\theta}$ среднеквадратический риск (дисперсия) $\mathcal{R}(\theta^*, \theta) \leqslant \mathcal{R}(\hat{\theta}, \theta)$ при $\forall \theta \in \Theta$.

4.4 Вопросы и задания для самоконтроля

1) Что такое оценка?

2) Дайте определение состоятельности оценки и проинтерпретируйте смысл этого определения.

3) Можно ли сказать, что состоятельная оценка лучше не состоятельной оценки?

4) Дайте определение несмещённости оценки и проинтерпретируйте смысл этого определения.

5) Можно ли сказать, что несмещённая оценка лучше смещённой оценки?

6) Будет ли оценка $\hat{\theta} = \theta_0$ несмещённой оценкой θ? Состоятельной? Вычислите среднеквадратичный риск этой оценки.

7) Будет ли оценка $\hat{\theta} = X_1$ несмещённой оценкой среднего? Состоятельной? Вычислите среднеквадратичный риск этой оценки.

8) Как показать состоятельность выборочного среднего и выборочной дисперсии как оценок среднего и дисперсии соответственно?

9) Будет ли состоятельной поправленная на несмещённость оценка дисперсии $\widetilde{S}^2$?

10) Проверьте несмещённость оценки $\widetilde{S}^2$ для случая, когда выборка берётся из равномерного распределения $\mathrm{Funif}(0, 1)$, напрямую – с помощью метода Монте-Карло. Возьмите объём выборки n равным 10, а количество репликаций Монте-Карло равным 1000. Сравните с результатом для оценки S^2.

5 Доверительное оценивание

5.1 Определение

Для того чтобы учесть случайный характер выборочных данных и степень влияния этой случайности на точность оценки параметра модели, используют так называемые интервальные оценки (доверительные интервалы).

Пусть $X^{(n)} = (X_1, \ldots, X_n)$ — случайная выборка, распределение которой зависит от некоторого неизвестного действительного параметра $\theta \in \Theta \subset R^1$. Статистика $A(X^{(n)}) \subseteq \Theta$, принимающая в качестве своих значений подмножества параметрического пространства Θ, называется $(1-\alpha)$-*доверительным множеством* для параметра θ, если

$$\mathbf{P}_\theta\left(\theta \in A\right) \geqslant 1-\alpha \quad \text{для} \quad \forall \theta \in \Theta.$$

Интервал $[\underline{\theta}, \overline{\theta}]$ с границами $\underline{\theta}(X^{(n)})$, $\overline{\theta}(X^{(n)})$, зависящими от выборочных данных, называется $(1-\alpha)$-*доверительным интервалом* для параметра θ, если

$$\mathbf{P}_\theta\left(\underline{\theta} \leqslant \theta \leqslant \overline{\theta}\right) \geqslant 1-\alpha \quad \text{для} \quad \forall \theta \in \Theta.$$

Величина $(1-\alpha)$ называется *коэффициентом доверия*, а величина $Q = (1-\alpha)\cdot 100\%$ называется *надёжностью* интервала. Надёжность выбирается обычно в пределах от 90% до 99% (стандартное значение — 95%). Другими словами, с высокой (заданной) надёжностью доверительный интервал накроет истинное значение параметра.

Если предыдущее неравенство выполняется в пределе при объёме выборки $n \to \infty$, т.е.

$$\liminf_{n\to\infty} \mathbf{P}_\theta\left(\underline{\theta}_n \leqslant \theta \leqslant \overline{\theta}_n\right) \geqslant 1-\alpha \quad \text{для} \quad \forall \theta \in \Theta,$$

то последовательность таких интервалов $[\underline{\theta}_n, \overline{\theta}_n]$, $n \geqslant 1$, называется *асимптотически доверительной* с надёжностью $Q = (1-\alpha)\cdot 100\%$.

Наряду с двусторонними, рассматривают также и односторонние доверительные интервалы. Статистика $\overline{\theta} = \overline{\theta}(X^{(n)})$ называется верхней $(1-\alpha)$-доверительной границей для параметра θ, если

$$\mathbf{P}_\theta\left(\overline{\theta} \geqslant \theta\right) \geqslant 1-\alpha.$$

Понятие верхней доверительной границы $\overline{\theta}$ эквивалентно понятию одностороннего доверительного интервала вида $(-\infty, \overline{\theta})$.

Статистика $\underline{\theta}$ называется нижней $(1-\alpha)$-доверительной границей для параметра θ, если

$$\mathbf{P}_\theta\left(\underline{\theta} \leqslant \theta\right) \geqslant 1-\alpha.$$

Аналогично асимптотическим интервалам вводятся понятия асимптотических границ.

Легко понять, что если $\underline{\theta}$ — какая-либо нижняя $(1-\alpha)$-доверительная граница, $\overline{\theta}$ — верхняя $(1-\alpha)$-доверительная граница, причём всегда $\underline{\theta} < \overline{\theta}$, то двусторонний интервал $[\underline{\theta}, \overline{\theta}]$ будет также доверительным, но с меньшим коэффициентом доверия $1-2\alpha$.

Список определений:

- *$(1-\alpha)$-доверительным множеством* (*доверительным множеством с надёжностью* $1-\alpha$) называется статистика $A(X^{(n)}) \subseteq \Theta$, принимающая в качестве значений подмножества параметрического пространства Θ, для которой выполняется

$$\mathbf{P}_\theta(\theta \in A) \geqslant 1-\alpha. \tag{3}$$

- *$(1-\alpha)$-доверительным интервалом* (*доверительным интервалом с надёжностью* $1-\alpha$) называется образованный значениями пары статистик $\underline{\theta}$, $\overline{\theta}$ интервал $(\underline{\theta}, \overline{\theta})$, для которых выполняется

$$\mathbf{P}_\theta(\underline{\theta} \leqslant \theta \leqslant \overline{\theta}) \geqslant 1-\alpha. \tag{4}$$

- $\overline{\theta}$ называется *верхней $(1-\alpha)$-доверительной границей*, если

$$\mathbf{P}_\theta(\theta \leqslant \overline{\theta}) \geqslant 1-\alpha.$$

  Интервал $(-\infty, \overline{\theta})$ называется *верхним доверительным интервалом.*

- $\underline{\theta}$ называется нижней $(1-\alpha)$-доверительной границей, если

$$\mathbf{P}_\theta(\underline{\theta} \leqslant \theta) \geqslant 1-\alpha.$$

  Интервал $(\underline{\theta}, \infty)$ называется *нижним доверительным интервалом.*

5.2 Интерпретация

Смысл этих определений легко понять, если вспомнить, что индекс θ, стоящий у знака вероятности $\mathbf{P}_\theta$, указывает на истинное значение неизвестного параметра. Поэтому формула (4), например, означает, что с большой вероятностью доверительный интервал накроет истинное значение оцениваемого параметра. На практике обычно делается несколько вольный вывод, что с большой долей вероятности следует ожидать значение оцениваемого параметра, принадлежащее интервалу $(\underline{\theta}, \overline{\theta})$. В таком утверждении «скрытно» присутствует предположение

о случайности появления параметра θ. В действительности, оцениваемый параметр не случаен, а имеет некоторое фиксированное неизвестное значение.

5.3 Точность и надёжность интервала

Величина $Q = (1 - \alpha) \cdot 100\%$ называется надёжностью интервала и выбирается обычно в пределах от 90% до 99% (стандартное значение — 95%). На первый взгляд кажется, что чем выше значение надёжности, тем лучше будет построенный интервал. Однако здесь надо учитывать, что чем больше величина Q, тем шире получится доверительный интервал (в пределе при $\alpha = 0$ он будет совпадать с $\mathbb{R}$), то есть уменьшится его точность. Задача построения доверительного интервала с заданной точностью и надёжностью может быть решена только при достаточном объёме выборки.

5.4 Двухсторонний интервал через доверительные границы

Для построения $(1 - \alpha)$-доверительного интервала $(\underline{\theta}, \overline{\theta})$ можно построить отдельно верхнюю $\overline{\theta}$ и нижнюю $\underline{\theta}$ границы с надёжностью $(1 - \alpha/2) \cdot 100\%$.

5.5 Связь с задачей проверки гипотез

Пусть $B(X^{(n)})$ — некое $(1 - \alpha)$-доверительное множество. Тогда критерий, отвергающий гипотезу, если $B(X^{(n)})$ полностью попадает в область альтернативы, будет иметь уровень α. Так, при альтернативе $\mathrm{H}_1 \colon \theta > \theta_0$ гипотезу следует отвергать, если нижняя граница $\underline{\theta} > \theta_0$. Если же ошибочно принимать гипотезу, когда доверительное множество полностью попадает в область гипотезы (например, верхняя граница $\overline{\theta} \leqslant \theta_0$), то такой критерий будет иметь вовсе «неприемлемый» уровень $1 - \alpha$, вместо ожидаемого уровня α.

5.6 Методы построения

1) Метод опорной функции. Пусть для некоторой статистики $T = T(X^{(n)})$ существует монотонное по параметру θ преобразование $G(t, \theta)$ (так называемая опорная функция), для которого функция распределения $F(x) = \mathbf{P}_\theta(G(T, \theta) < x)$ не зависит от θ. Тогда, выбирая Δ из соотношения $F(\Delta) = 1 - \alpha$ и разрешая неравенство $G(t, \theta) < \Delta$ относительно θ

при полученном экспериментальном значении статистики $T = t$, получаем верхнюю или нижнюю (в зависимости от направления монотонности G) доверительную границу для θ.

Этот метод применяется при построении доверительных границ для среднего значения и дисперсии нормального распределения.

2) Метод, основанный на функции распределения оценки. Пусть функция распределение $F(t, \theta) = \mathbf{P}_\theta(T < t)$ статистики $T = T(X^{(n)})$ непрерывно и строго убывает с ростом параметра θ. Тогда, если экспериментальное значение статистики $T = t$, то значение нижней $(1 - \alpha)$-доверительной границы $\underline{\theta}$ можно получить как решение уравнения $F(t, \underline{\theta}) = 1 - \alpha$. Верхняя $(1 - \alpha)$-доверительная граница $\overline{\theta}$ получается как решение уравнения $\mathbf{P}_{\overline{\theta}}(T > t) = 1 - \alpha$.

Замечание. Часто функция распределения определяется как $F(t, \theta) = \mathbf{P}_\theta(T \leqslant t)$. Потому для распределений T, сосредоточенных на целых числах, нижняя граница есть решение (по $\underline{\theta}$) уравнения $F(t - 1, \underline{\theta}) = 1 - \alpha$, а верхняя — уравнения $F(t, \overline{\theta}) = \alpha$.

Этот метод применяется при построении доверительных границ для вероятности наблюдаемого события.

3) Метод, основанный на асимптотическом распределении оценок. Этот метод близок к первому методу. Если известно асимптотическое распределение некоторой статистики T, то, оценив мешающие параметры, можно построить опорную функцию, предельное распределение которой не будет зависеть от неизвестных параметров. Далее, поступая как и в методе 1), можно построить доверительную границу для неизвестного параметра. Надёжность такого доверительного утверждения с ростом объёма выборки будет приближаться к номинальной надёжности Q.

Этот метод также применяется при построении доверительных границ для вероятности наблюдаемого события.

5.7 Вопросы и задания для самоконтроля

1) Что такое доверительное множество?

2) Что такое нижняя доверительная граница и нижнее доверительное множество? Различаются ли эти два понятия?

3) Как построить двусторонний доверительный интервал, зная способы построения верхнего и нижнего доверительных интервалов?

4) Проинтерпретируйте смысл определений односторонних доверительных границ.

5) Можно ли утверждать, что чем выше надёжность, тем выше качество доверительного множества?

6) Докажите, что критерий, построенный по $(1-\alpha)$–доверительному множеству, имеет уровень α.

6 Коэффициент корреляции и линейная регрессия

Обозначим через X, Y наблюдаемые в эксперименте случайные величины. Попытаемся сначала спрогнозировать возможное значение характеристики Y, если известно наблюдённое значение характеристики $X = x$. Если прогноз осуществляется с помощью некоторой функции $h(x)$, то мерой качества такого прогноза служит среднеквадратическая ошибка $\mathbf{E}(Y - h(X))^2$. Функцию $h^*(x)$, для которой достигается минимум среднеквадратической ошибки, называют функцией регрессии Y на X. Теоретический вид этой функции весьма прост:

$$h^*(x) = \mathbf{E}(Y \mid X = x)$$

— условное среднее Y при фиксированном значении с.в. $X = x$.

На практике построение хорошей оценки для неё возможно лишь для дискретных с.в. X. Поэтому обычно рассматривают регрессию не среди всех возможных функций, а только среди линейных. Уравнение линейной среднеквадратической регрессии Y на X — линейной функции, минимизирующей среднеквадратическую ошибку, имеет вид

$$y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

где μ_X, σ_X^2 — среднее значение и дисперсия с.в. X, μ_Y, σ_Y^2 — среднее значение и дисперсия с.в. Y,

$$\rho = \frac{\mathbf{E}\{(X - \mu_X)(Y - \mu_Y)\}}{\sigma_X\sigma_Y}$$

— коэффициент корреляции между Y и X.

Коэффициент корреляции ρ часто называют коэффициентом линейной связности. Такая интерпретация обусловлена следующими свойствами этого коэффициента:

1) Он принимает значения от -1 до 1 и не зависит от масштаба измерений.

2) Если $\rho = \pm 1$, то между с.в. Y и X существует точная линейная связь, причем при $\rho = 1$ эта связь имеет положительную направленность (с ростом одной характеристики, растет и другая), а при $\rho = -1$ — отрицательную.

3) Если с.в. Y и X независимы, то $\rho = 0$.

4) Для нормального случайного вектора равенство нулю коэффициента корреляции эквивалентно независимости с.в. Y и X.

5) Минимальная среднеквадратическая ошибка линейного прогноза характеристики Y по значениям характеристики X равна $\sigma_Y^2(1-\rho^2)$.

6) Если $\rho = 0$, то наилучший прогноз с.в. Y — это её среднее μ_Y, а линия регрессии Y на X представляет собой прямую, параллельную оси OX.

Свойство 5) применяют при описании степени зависимости между случайными величинами. Считается, что разброс с.в. Y на $\rho^2 \cdot 100\%$ обусловлен влиянием на неё с.в. X и на $(1-\rho^2) \cdot 100\%$ внутренними факторами, присущими самой с.в. Y, или другими неучтёнными факторами.

Неизвестные параметры линейной регрессии легко оцениваются своими выборочными аналогами:

- $\overline{X}, S_X^2$ — средним значением и дисперсией выборки из X,
- $\overline{Y}, S_Y^2$ — средним значением и дисперсией выборки из Y,
- $r = \dfrac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{Y})}{S_X S_Y}$ — выборочным коэффициентом корреляции.

Заметим, что последний коэффициент является состоятельной, но смещённой оценкой истинного коэффициента корреляции ρ.

Таким образом, оценку уравнения регрессии Y на X можно записать в виде

$$y = \widehat{y}(x) = \overline{Y} + b_{Y/X}(x - \overline{X})$$

с коэффициентом регрессии $b_{Y/X} = r \cdot S_Y/S_X$. Для этой линии достигается минимум суммы расстояний по оси OY между выборочными точками и графиком линии регрессии, когда минимум ищется среди всех линейных функций:

$$\sum_{i=1}^n (Y_i - \widehat{y}(X_i))^2 = \min_{b,c} \sum_{i=1}^n (Y_i - (bX_i + c))^2.$$

При построении регрессии X на Y можно просто в уравнении регрессии произвести перестановку переменных $x \leftrightarrow y$:

$$x = \widehat{x}(y) = \overline{X} + b_{X/Y}(y - \overline{Y}), \quad \text{где} \quad b_{X/Y} = r \cdot S_X / S_Y.$$

Для этой линии достигается минимум суммы расстояний по оси OX между выборочными точками и графиком линии регрессии. Последнее уравнение используется при отыскании наилучшего прогноза характеристики X по наблюдённому значению Y. Для построения графика удобнее привести это уравнение к виду

$$y = \overline{Y} + \frac{1}{b_{X/Y}}(x - \overline{X}).$$

Таким образом, обе линии регрессии проходят через точку с координатами $(\overline{X}, \overline{Y})$ и отличаются лишь коэффициентом наклона. Можно показать, что в привычной системе координат (x — по оси абсцисс, y — по оси ординат) регрессия X на Y проходит круче регрессии Y на X. Схему построения регрессий Y на X и X на Y иллюстрирует рис. 4.

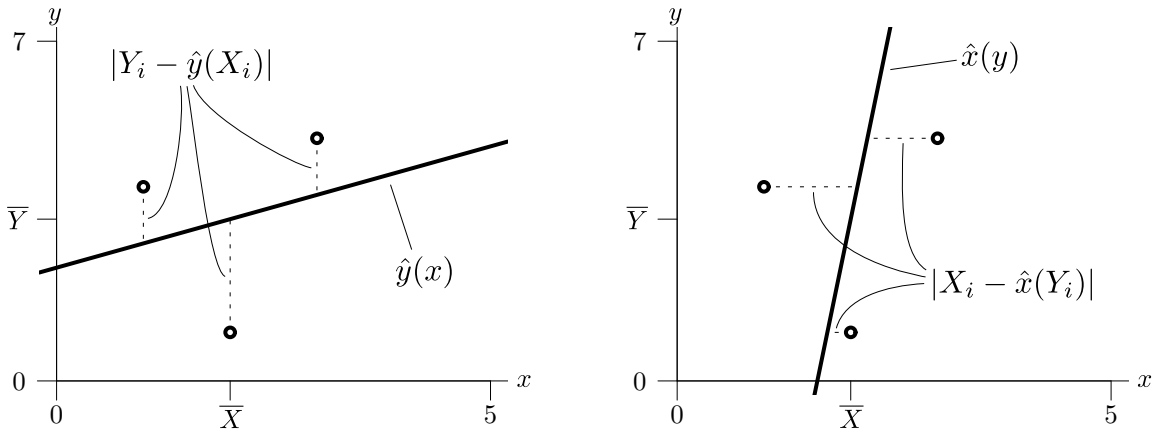

Рис. 4: Регрессия Y на X (слева) и X на Y (справа)

Если выборочный коэффициент корреляции r близок к нулю, то линии регрессии будут близки к взаимно перпендикулярным прямым, проходящим параллельно осям координат, причем регрессия Y на X будет параллельна оси OX. Если r по модулю близок к 1, то угол между линиями регрессии становится близким к нулю, и в пределе (при $|r| = 1$) обе линии совпадут.

На практике чаще всего свойство 4) коэффициента корреляции переносят на все вероятностные модели и интерпретируют равенство нулю коэффициента корреляции как независимость наблюдаемых характеристик, что не всегда верно. В общем случае независимость случайных величин X и Y означает, что

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A) \cdot \mathbf{P}(Y \in B) \tag{5}$$

для любых событий A и B из области значений с.в. X и Y. Если известны совместное распределение $F(x, y)$ вектора (X, Y) и частные распределения $F_X(x)$ и $F_Y(y)$ его компонентов, то для проверки независимости достаточно проверить выполнение равенства

$$F(x, y) = F_X(x) \cdot F_Y(y).$$

Если случайный вектор (X, Y) имеет не нормальное распределение, то для проверки гипотезы независимости его характеристик необходимо тем или иным способом проверить выполнение соотношения (5).

6.1 Вопросы и задания для самоконтроля

1) Что такое условное среднее $h^*(x) = \mathbf{E}(Y \mid X = x)$? Как его вычислить, зная совместную плотность $f(x, y)$ с.в. X и Y?

2) Дайте определение коэффициента корреляции. Как интерпретировать его значение?

3) Что такое функция регрессии?

4) Как определяется линейная регрессия Y на X?

5) Как оценить коэффициенты линейной регрессии по выборочным данным?

6) Что значит независимость двух случайных величин X и Y?

7) Что значит независимость двух выборок X и Y? Объясните в терминах совместной функции распределения выборки.

7 Справочник функций для программной реализации

Для выполнения заданий рекомендуется использовать языки программирования или специализированное программное обеспечение для статистических вычислений. В данном пособии будут даны краткие указания по средствам языка R, Microsoft Excel 2010 и Wolfram Mathematica, полезным при решении задач пособия.

Замечание. Большая часть процедур, оценок и критериев, рассматриваемых в пособии, уже реализована в рамках всех статистических пакетов. Тем не менее, в целях лучшего освоения механики работы со статистическими методами, рекомендуем выполнять программную часть заданий пособия полностью самостоятельно, без использования встроенных средств соответствующих программных пакетов.

7.1 Язык R.

Язык и система R — свободное и бесплатное программное обеспечение, предназначенное для статистической обработки и визуализации данных. Официальный сайт проекта: www.r-project.org. При работе с языком R рекомендуется дополнительно установить бесплатное IDE RStudio, официальный сайт: www.rstudio.com [4].

Командой разработчиков R создано краткое вводное пособие: «An Introduction to R» [5]. Другое пособие (на русском): «Наглядная статистика. Используем R!» [6]. Также могут быть полезны различные cheatsheets (краткие справочники) по языку R, например: R Reference Card [7]. Также для всех конструкций и функций языка можно вызвать справку командой «?», например: `?sum`

Язык R — интерпретируемый язык программирования с динамической типизацией. Основная особенность — фокус на работу с векторными и табличными данными. В связи с этим, в частности, все переменные языкая автоматически являются векторами (обычные числа — векторы длины 1), большинство операций и функций осуществляется поэлементно к своим векторнозначным аргументам. Как и все интерпретируемые языки, в R обычные циклы обрабатываются медленно; поэтому, рекомендуется (когда это возможно) производить преобразование и обработку данных, используя векторные функции и операции, и избегать циклов.

Основной принятый формат файла для хранения табличных данных — `.csv`. Для работы с `.xls`/`.xlsx` можно воспользоваться, например, пакетами `openxlsx` или `readxl`.

Пример 1. Загрузка данных из внешнего файла и вычисление дисперсии выборочных данных:

```
df = read.csv(''data.csv'')
x = df$x
m = sum(x) / length(x)
v = sum( (x - m)^2 ) / length(x)
cat(''Дисперсия:'', v, ''\n'')
```

Пример 2. Отрисовка графика функции cos и sin на отрезке $[0, \pi]$:

```
x = seq(0, pi, length.out = 100)
plot(x, cos(x), type = ''l'')
lines(x, sin(x), col = ''blue'')
```

Пример 3. Подсчёт количества элементов выборки, попавших в интервалы $[0, 2)$, $[2, 4)$, $[4, 6)$, $[6, 8)$, $[8, 10)$:

```
df = read.csv(''data.csv'')
x = df$x
y = cut(x, c(0,2,4,6,8,10), right = F)
z = table(y)
cat(''Частоты:\n'')
print(z)
```

Список функций. Служебные операции и загрузка данных:

- `read.csv` — загрузка таблицы из .csv файла
- `install.packages` — установка пакетов
- `library` — подключение пакета
- `help` — вызов справки по указанной команде

Преобразование данных и арифметические операции:

- `length` — количество элементов в векторе
- `sum` — сумма элементов вектора
- `mean` — среднее арифметическое элементов вектора
- `sort` — сортировка элементов вектора
- `seq` — генерация последовательностей чисел
- `floor` — целая часть числа (округление вниз)
- `ceiling` — целая часть числа $+1$ (округление вверх)
- `cut` — определение интервала, в который попадает каждый элемент вектора
- `table` — для каждого уникального значения в векторе подсчитывает количество элементов с соответствующим значеним. В связке с командой `cut` может использоваться для подсчёта количества попадений элементов вектора в интервалы заданного разбиения числовой прямой

Функции распределения и квантили:

- `pnorm` — функция распределения нормального закона
- `pt` — функция распределения Стьюдента
- `pchi` — функция распределения Хи-квадрат
- `pf` — функция распределения Фишера

- `pwilcox` — функция распредления Вилкоксона
- `qnorm` — квантиль нормального распределения
- `qt` — квантиль распределения Стьюдента
- `qchi` — квантиль распределения Хи-квадрат
- `qf` — квантиль распределения Фишера
- `qwilcox` — квантиль распредления Вилкоксона

Графические функции:

- `plot` — основная функция отрисовки графиков
- `lines` — добавление кривых к существующему графику
- `points` — добавление точек к существующему графику
- `abline` — отрисовка прямой по заданным коэффициенту наклона и константе
- `barplot` — ступенчетая кривая, полезна при отрисовке гистограммы

7.2 MS Excel

По состоянию на момент написания этого пособия, программа пользуется достаточно большой популярностью у практиков, не связанных с математикой, но которым необходимо ей как-то пользоваться. В первую очередь из-за того, что на Excel можно выполнять вычисления, не зная языков программирования. Это в значительной мере ограничивает обычный функционал программы, однако его достаточно для выполнения всех заданий этого учебного пособия.

Указанный функционал приведён для версии Excel 2010.

Список функций Функции для преобразования данных:

- `СУММ` — сумма ячеек
- `СРЗНАЧ` — среднее значение ячеек
- `ЧАСТОТА` — подсчёт количества элементов векторов, попадающие в интервалы разбиения числовой прямой

-

Функции распределения и квантили:

- `НОРМ.РАСП` — функция распределения нормального закона
- `СТЬЮДЕНТ.РАСП` — функция распределения Стьюдента

- `ХИ2.РАСП` — функция распределения Хи-квадрат
- `F.РАСП` — функция распределения Фишера
- `НОРМ.ОБР` — квантиль нормального распределения
- `СТЬЮДЕНТ.ОБР` — квантиль распределения Стьюдента
- `ХИ2.ОБР` — квантиль распределения Хи-квадрат
- `F.ОБР` — квантиль распределения Фишера

7.3 Wolfram Mathematica

Одна из ключевых особенностей системы Wolfram Mathematica (далее $\mathcal{WM}$) заключается в том, что она позволяет использовать символьные вычисления, например найти интеграл или производную. Кроме этого $\mathcal{WM}$ предоставляет для пользования полноценный язык программирования. Официальный сайт: www.wolfram.com/mathematica

Разработчики $\mathcal{WM}$ утверждают, что этот пакет программ самый мощный пакет, ориентированный на операции с векторами.

1) *Вектор* (или *список* — List на языке $\mathcal{WM}$) представляет собой набор любых допустимых элементов, заключённых в фигурные скобки и отделённых друг от друга запятой:

- `{1,5,2,0}` — список из четырёх чисел
- {$\frac{1}{2}$,{π,`{q,5,w}`}, `Plot[x,{x,0,1}]`} — список из 3-х элементов, где:
  - 1-й элемент — число $\frac{1}{2}$,
  - 2-й элемент — список из двух элементов: число π и трёхмерный вектор `{q,5,w}`,
  - 3-й элемент — график функции $y = x$ на отрезке $[0, 1]$;
- `matra={{11,12,13,14}, {21,22,23,24}, {31,32,33,34}}` — прямоугольная матрица из 3-х строк (элементы списка) и 4-х столбцов, при обращении к которой в дальнейшем можно использовать имя `matra`.

2) Доступ к элементам списка (вектора) осуществляется с помощью конструкции из спаренных квадратных скобок `[[...]]`:

- `matra[[3]]` — 3-й элемент матрицы `matra`, т.е. вектор `{31,32,33,34}`

- `{{a,7},{π,{q,5,w}}}[[2,2,1]]` — второй элемент `{π,{q,5,w}}` списка, из которого взят снова 2-й элемент `{q,5,w}`, из которого выбран 1-й элемент `q`;
- `matra[[{2,2,1,1}]]` — матрица 4×3, в которой 1-я и 2-я строки совпадают со 2-й строкой, а 3-я и 4-я строки — с 1-й строкой `matra`
- `matra[[{2,2,1,1},2]]` — все 2-ые элементы матрицы из предыдущего примера, т.е. вектор $\{22, 22, 12, 12\}$
- `matra[[All,2]]` — все 2-ые компоненты `{12, 22, 32, 42}` строк матрицы `matra` (т.е. столбец этой матрицы)

Столбец матрицы можно также получить, предварительно её транспонировав:

- `matra`$^{\mathtt{T}}$`[[2]]` — тот же вектор $\{12, 22, 32, 42\}$ (набирается Esc t r Esc).

Для выбора первого и последнего элементов списка можно использовать функции `First` и `Last`, соответственно.

3) Обращение к функциям $\mathcal{WM}$ осуществляется с помощью квадратных скобок: `Func[a,3/8,u,Log[3]]`. Если функция имеет только один аргумент, возможно векторный, то проще использовать так называемое постфиксное обращение вида `x // Func`, особенно, когда `x` представляет собой очень длинное выражение. Например, `Sin`$\left[\frac{\pi}{12}\right]$ даёт ответ $\frac{-1+\sqrt{3}}{2\sqrt{2}}$, который можно превратить в десятичную дробь с помощью функции `N`: `Sin`$\left[\frac{\pi}{12}\right]$ `// N` — результат 0.258819.

Замечание. Имена встроенных функций и констант $\mathcal{WM}$ всегда начинаются с прописных букв: `Length, Mean, Sin[Pi/2]` (результат, конечно, 1, так же как и `Log[E]`). Если имя состоит из нескольких слов, то (почти всегда) каждое слова также начинается с прописной буквы: `ChiSquareDistribution`, `NMinimize`.

4) Имена придуманных пользователем функций и констант сохраняются в текущей сессии $\mathcal{WM}$ вплоть до очередного присвоения. Выбор между строчными и прописными буквами полностью в его власти.

Замечание. Результат выполнения операции, завершающейся символом « ; » (точка с запятой), не выдаётся на экран. Дабы не потерять в недрах $\mathcal{WM}$ полученный результат, необходимо в этом случае присвоить операции какое-либо уникальное имя. Например, часто приходится строить по отдельности графики каких-либо сложных функций. После

того, как рисунки графиков полностью будут отработаны, можно каждому из графиков присвоить своё имя и изобразить их вместе в одной системе координат с помощью команды `Show`. При этом, если построение каждого отдельного графика будет завершаться символом « ; », то на экране не будет мельтешения ненужных картинок.

Круглые скобки `(...)` используются только для выделения блоков операций, т.е. для смыслового объединения операций, следующего за логикой построения. При этом, все операции, кроме, быть может, последней, должны заканчиваться символом « ; ».

5) Вероятностные распределения и их характеристики. Любой вероятностный закон имеет своё имя с параметрами

```
NameDistribution[Parameters]
```

и свои характеристики.

Имена распределений, чаще всего, «говорящие»:

- `NormalDistribution[`μ, σ`]` — нормальный закон с математическим ожиданием μ и дисперсией σ^2
- `ChiSquareDistribution[d]` — хи-квадрат распределение с d степенями свободы
- `StudentTDistribution[d]` — распределение Стьюдента с d степенями свободы
- `BinomialDistribution[n,p]` — биномиальное распределение с n испытаниями и вероятностью «успеха» p в одном испытании
- `ExponentialDistribution[`λ`]` — показательное (экспоненциальное) распределение с интенсивностью λ
- `FRatioDistribution[m,n]` — распределение Фишера с (m, n) степенями свободы

Вызов какой-либо характеристики `H` всегда осуществляется в виде

```
H[ NameDistribution[Parameters], ArgumentsH ],
```

где аргументы могут отсутствовать. Некоторые из характеристик:

- `Mean[ ExponentialDistribution[`λ`] ]` — математическое ожидание показательной случайной величины (результат $1/\lambda$)

  Аналогично находятся дисперсия (`Variance`), стандартное отклонение (`StandardDeviation`), медиана (`Median`), квантиль (`Quantile`) и все три квартили (`Quartiles`).

- `PDF[ Name[Parameters], x]` — значение в точке `x` плотности вероятностей (для абсолютно непрерывного) или вероятности (для дискретного) закона `Name` с параметрами `Parameters`
- `CDF[ Name[Parameters], x]` — значение в точке `x` функции распределения (т.е. вероятность попасть от $-\infty$ до x *включительно*) закона `Name` с параметрами `Parameters`
- `Quantile[ Name[Parameters], α ]` — квантиль порядка α вероятностного закона `Name[Parameters]`

Примеры:

- вероятность того, что биномиальная сл.в. с 11 испытаниями и вероятностью «успеха» 0.4 примет значение или 0, или 1, или 2:

```
CDF[ BinomialDistribution[11, 0.4], 2 ]
```

  (результат 0.118917);

- функция распределения нормального вероятностного закона со средним 3 и стандартным отклонением 7 в точке 3/2:

```
CDF[ NormalDistribution[3,7], 3/2 ];
```

- плотность стандартного нормального (0, 1) вероятностного закона:

```
PDF[ NormalDistribution[ ], x ];
```

- вероятность того, что нормальная сл.в. с математическим ожиданием 2 и дисперсией 16 попадёт в интервал от 0.5 до 1:

```
NRM = NormalDistribution[2, 4];

CDF[NRM, 1] - CDF[NRM, 0.5]
```

  (результат 0.0474634);

- верхняя 0.05-квантиль (5%-я точка) распределения Стьюдента с 17 степенями свободы :

```
Quantile[ StudentTDistribution[17], 0.95 ]
```

  (результат 1.73961)

6) Импорт данных из Excel-файла (и не только) осуществляется с помощью функции `Import`:

```
Dann = Import[''D:\\KursProject\\Zadan50.xls'']
```

где доступны также файлы с расширениями .xlsx и .csv.

Данные после импортирования имеют следующую структуру:

- `{` Лист1, Лист2, ... `}` — вектор (список) с элементами образованными листами файла, например, `Dann[[1]]` — список, содержащий данные 1-го листа.
- Структура листов: `{` Строка1, Строка2, ... `}` — матрица, строки которой совпадают со строками Excel-файла.

Пример выбора данных из 4-го столбца 3-го листа Excel-файла:

(`Dann[[3]]`)$^{\mathrm{T}}$`[[4]]` или `Dann[[ 3, All, 4 ]]`

Здесь $matrix^{\mathrm{T}}$ есть операция транспонирования матрицы (см. выше).

Пояснение. При импорте данных все листы получаются как прямоугольные матрицы. Так как не обязательно все строки имеют одинаковую длину, то $\mathcal{WM}$ устанавливает количество столбцов по самой длинной строке. Пустые места в остальных строках импортируются как пустые элементы. Например, могут получаться столбцы вида `Z = {Задан4, 1.1, 2.0, 3.2, , , , }`. Чтобы удалить эти пустоты и заодно мешающие символы, можно воспользоваться функцией `Select` с опцией выбора числовых элементов (работает только для одномерных списков): `Select[ Z , NumberQ ]` — результат `{1.1, 2.0, 3.2}`.

5) Функции графики.

- `Plot[F[z],{z,z0,z1}]` — график функции F в интервале $z \in [z0, z1]$;
- `Plot[{F[z],h[z]},{z,z0,z1}]` — график нескольких функций в интервале $z \in [z0, z1]$;
- `Graphics[{ Text[O1,{x1,y1}], Text[On,{xn,yn}] }]` — объекты O1, On, например числа, помещаются на плоскости так, что центр каждого объекта попадает в точку с соответствующими координатами;
- `Histogram[ dan, {gran}, ''PDF Options ]` — вариант гистограммы. Здесь

  dan – вектор числовых данных,

gran – список границ интервалов (данные должны полностью попасть от нижней границы и до верхней) – обратите внимание, что этот список должен быть взят в фигурные скобки,

"PDF- – тип гистограммы, когда высота ступенек равна относительной частоте попадания в соответствующий интервал, делённой на длину интервала.

Если опция Options = LabelingFunction –> Above, то выше ступенек (Above) будет выведено значение высоты ступеньки.

- `Show[graf1,graf2,Options]` — отображение нескольких заранее построенных графиков на общей системе координат с дополнительными опциями. Для лучшего визуального представления можно при построении каждого графика использовать свои стилевые опции (PlotStyle).

6) Вспомогательные функции.

- `Solve` — решение (системы) уравнений, для которых возможна запись результата через известные функции, с условием или без условия:

  Solve$[x^2 + y^2 == 5$ && $x + y == 1, \{x, y\}]$

  – результат $\{\{x \to 1, y \to 2\}, \{x \to 2, y \to -1\}\}$,

  Solve$[x^2 + y^2 == 5$ && $x + y == 1$ && $x > 0, \{x, y\}]$

  – результат $\{x \to 2, y \to -1\}$;

- `FindRoot` — решение (системы) уравнений итерационными численными методами с начальной точкой итераций:

  FindRoot$[\{x^2 + y^2 == 5, x + y == 1\}, \{x, 1\}, \{y, 0\}]$

  – результат $\{x \to 2, y \to -1\}$,

  FindRoot$[\{x^2 + y^2 == 5, x + y == 1\}, \{x, 0\}, \{y, 1\}]$

  – результат $\{x \to -1, y \to 2\}$.

  При неудачно заданной начальной точке решение может быть не найдено и $\mathcal{WM}$ предложит попытаться поискать решение с другой начальной точкой.

- `Minimize (Maximize - NMinimize, NMaximize)` — поиск (с условием или без условия) экстремума функции классическими методами математического анализа (вариант с префиксом N выдаёт результат в десятичной форме):

  NMinimize$[\{-3e^x + x^4 - 100Sin[x], -4 <= x < 0\}, x]$

– результат $\{-3., \{x \to 0.\}\}$, т.е. минимальное значение достигается в правой крайней точке $x = 0$;

NMinimize$[\{-3e^x + x^4 - 100Sin[x]\}, x]$

– результат $\{-108.348, \{x \to 1.56148\}\}$, т.е. глобальный минимум достигается в точке $x = 1.56148$;

- `FindMinimum (FindMaximum)` — поиск (с условием или без условия) одного локального экстремума функции:

  FindMinimum$[\{e^x - x^4 + 11Sin[x]\}, \{x, -2\}]$

  – результат $\{94.4854, \{x \to -2.89179\}\}$,

  FindMinimum$[\{e^x - x^4 + 11Sin[x]\}, \{x, 2\}]$

  – результат $\{-108.348, \{x \to 1.56148\}\}$, т.е. начиная из различных точек, можно нечаянно найти глобальный минимум,

Замечание. Удобство представления результата выполнения операций в виде подстановки $\{x \to x0, y \to y0\}$ особенно ярко проявляется при большом количестве аргументов:

otvet = NMinimize$[\{x^2 + y^2 + z^2 + a^2, 1 < x + 2y + z + 2a < 2\}, \{a, x, z, y\}]$;
$x + 5y - z + 5a$ /. Last[otvet]

– результат 2. Здесь otvet есть список $\{\ 0.1\ , \{a \to 0.2, x \to 0.1, z \to 0.1, y \to 0.2\}\}$, в котором первый элемент 0.1 равен минимальному значению функции, а второй элемент (он же последний – Last, он же (–1)-й), содержит список аргументов с указанием точек достижения минимума.

Если очень хочется узнать значение y (т.е. четвёртого элемента в списке аргументов при вызове функции NMinimize), то можно воспользоваться операцией [[...]] для выбора нужного элемента из списка: otvet [[–1, 4, 2]] – в последнем элементе списка (т.е. в (–1)-ом) выбрать 4-й элемент $\{y \to 0.2\}$, из которого взять 2-й элемент 0.2. Но можно и так: y /. Last[otvet], или так: y /. Otvet[[–1]].

Обращение otvet [[–1, All, 2]] выдаст вектор $\{0.2, 0.1, 0.1, 0.2\}$ значений аргументов в заявленном порядке $\{a, x, z, y\}$. Тот же результат даёт обращение $\{a, x, z, y\}$ /. Last[otvet]. Второй способ удобнее, т.к. не требует запоминания порядка элементов. Например, можно оформить обращение $\{x, z, a, y\}$ /. Last[otvet] – результат $\{0.1, 0.1, 0.2, 0.2\}$.

Часть II
Первичный статистический анализ

В этой главе приводятся описания первых трех заданий. Рассматриваемые здесь процедуры предваряют, как правило, любую статистическую обработку данных.

Теоретические основы:

- понятие выборки смотрите в п. 1, с. 10
- характеристики статистических оценок смотрите в п. 4, с. 26
- вероятностными модели смотрите в п. 2, с. 12

Задание 1. Выборочные характеристики

1. Постановка задачи

Дана выборка $X^{(n)} = (X_1, \ldots, X_n)$. Вычислить основные статистические характеристики и оценки параметров выборочных данных:

1) минимум, максимум, размах выборки
2) математическое ожидание
3) дисперсию
4) стандартное отклонение
5) коэффициент асимметрии
6) медиану
7) интерквартильную широту

2. Теоретические основы

Статистический анализ выборочных данных начинают обычно с вычисления выборочных моментов.

1) Размах. Размах выборки вычисляется как разность между максимальным и минимальным значениями:

$$X_{(n)} - X_{(1)},$$

где $X_{(k)}$ — k-ый элемент вариационного ряда, т.е. k-ый элемент упорядоченной по возрастанию выборки ($X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$).

2) Математическое ожидание. Выборочное математическое ожидание (выборочное среднее):

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

(читается «икс с чертой») — несмещённая и состоятельная оценка истинного математического ожидания (среднего значения) μ. Величина μ характеризует расположение наблюдаемой с.в. X. Очень часто сравнение различных совокупностей производят как раз по среднему. Однако, надо иметь в виду, что это имеет смысл только в случае, если остальные характеристики (см. ниже) приблизительно совпадают.

3) Дисперсия. Выборочная дисперсия:

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

— оценка истинной дисперсии σ^2. Дисперсия σ^2 служит мерой разброса с.в. около её среднего μ. По известному правилу трех сигм с вероятностью, большей 90%, следует ожидать значение с.в. в пределах $\mu \pm 3\sigma$. Интересно, что нормальная с.в. с вероятностью, большей 95%, принимает значения в интервале $\mu \pm 2\sigma$.

Дисперсия измерительного прибора (если этот прибор не имеет систематической ошибки, т.е. математическое ожидание ошибки равно нулю) характеризует его точность.

Часто используют поправленную на несмещённость оценку дисперсии

$$\widetilde{S}^2 = \frac{n}{n-1}S^2.$$

Именно эта оценка используется в большинстве пакетов в качестве стандартной оценки дисперсии. Обе оценки являются состоятельными.

Наряду с дисперсией, всегда вычисляется стандартное отклонение $S = \sqrt{S^2}$. Забегая вперед, скажем, что точность доверительного интервала для истинного значения среднего μ прямо пропорциональна S. Грубо говоря, истинное μ лежит где-то в пределах $\pm 2S/\sqrt{n}$ от выборочного среднего $\overline{X}$.

4) Коэффициент асимметрии.

Выборочный коэффициент асимметрии

$$g_1 = \frac{1}{nS^3}\sum_{i=1}^{n}(X_i - \overline{X})^3 = \frac{1}{S^3}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^3 - 3\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i^2 + 2\overline{X}^3\right)$$

— состоятельная, но смещённая оценка истинного коэффициента асимметрии

$$\gamma_1 = \mathbf{E}\left(\frac{X-\mu}{\sigma}\right)^3.$$

Несёт информацию о симметричности расположения данных относительно центра $\overline{x}$. При больших положительных значениях γ_1 распределение с.в. будет иметь более «тяжёлый» правый «хвост». Если график плотности такого распределения «насадить» на вертикальный штырь в точке достижения максимума, то график «упадёт» вправо.

Замечание. Выборочный коэффициент g_1 можно использовать для проверки согласия данных выборочного обследования с нормальной моделью распределения. Большие абсолютные значения g_1 будут свидетельствовать против гипотезы нормальности распределения, каковое, как известно, симметрично.

5) Медиана. Оценка медианы:

$$m = \begin{cases} X_{((n-1)/2+1)}, & \text{если } (n-1)/2 = [(n-1)/2]; \\ (X_{([(n-1)/2]+1)} + X_{([(n-1)/2]+2)})/2, & \text{если } (n-1)/2 > [(n-1)/2]; \end{cases}$$

где $X_{(k)}$ — k-ый элемент вариационного ряда, а $[x]$ — целая часть x. По-сути, для оценивания медианы достаточно упорядочить выборку и выбрать в качестве оценки центральный (по индексу) элемент последовательности (либо среднее арифметическое ближайших к центру элементов, если центрального нет).

Медиана — это точка, которая делит вероятностную массу пополам. Медиана характеризует некоторое центральное значение с.в. и близко с точки зрения интерпретации и применения к понятию математического ожидания. С точки зрения приложения, основное отличие медианы от математического ожидания — её робастность, отсутствие зависимости от крайних (максимальных и минимальных) значений в выборке. Желание уменьшить влияние крайних значений связано с тем, что в жизни они часто оказываются не «настоящими» наблюдениями с.в., а результатами каких-либо сбоев при проведении эксперимента, либо чрезвычайными, несвойственными для явления значениями.

Отличие медианы от математического ожидания может свидетельствовать о несимметричном расположении данных. Простым примером

может служить сравнение средней зарплаты по стране с медианной зарплатой. Из-за наличия немногочисленных работников с огромной зарплатой среднее значение будет «тянуться» за ними. В то же время медиана показывает, какую зарплату действительно получает «средний» человек.

Замечание. Существуют несколько незначительно отличающихся вариантов задания оценки медианы, и здесь приведён лишь самый распространённый (и, почти всегда, достаточный). В общем случае оценку медианы можно задать как такое число m, для которого $\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i < m) \le \frac{1}{2} \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \le m)$.

6) Интерквартильная широта Интерквартильная широта — это $Q(3/4) - Q(1/4)$. Оценка для квантили ($q \in [0,1]$):

$$\hat{Q}(q) = \begin{cases} X_{((n-1)q)+1}, & \text{если } (n-1)q = [(n-1)q]; \\ (X_{([(n-1)q]+1)} + X_{([(n-1)q]+2)})/2, & \text{если } (n-1)q > [(n-1)q]; \end{cases}$$

где $X_{(k)}$ — k-ый элемент вариационного ряда, а $[x]$ — целая часть x. Отсюда, оценка для интерквартильной широты:

$$\hat{Q}(3/4) - \hat{Q}(1/4).$$

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Является ли выборочное среднее (дисперсия, стандартное отклонение, коэффициент асимметрии, эксцесс) несмещённой оценкой?

3) Является ли выборочное среднее (дисперсия, стандартное отклонение, коэффициент асимметрии, эксцесс) состоятельной оценкой?

4) Что такое состоятельность и несмещённость?

5) Как можно исправить смещение выборочной дисперсии? Будет ли после такого исправления несмещённым стандартное отклонение? Будет ли состоятельной несмещённая оценка дисперсии?

6) Какую информацию несет коэффициент эксцесса (среднее значение, дисперсия, асимметрия, медиана)?

7) По какой формуле вычисляется дисперсия (среднее, асимметрия, эксцесс)?

Задание 2. Гистограмма выборки

1. Постановка задачи

Построить график гистограммы выборки.

2. Теоретические основы

Гистограмма — ступенчатая кривая, высота ступенек которой пропорциональна количеству выборочных данных, попавших в заданные интервалы числовой прямой. Гистограмма используется для геометрического представления данных.

Различают два варианта гистограмм:

- Частотная гистограмма — высота столбцов в точности равна количеству выборочных данных, попавших в интервал
- Вероятностная гистограмма — высота столбцов дополнительно нормируется таким образом, чтобы полученная ступенчатая кривая была сопоставима с функцией плотности.

В большинстве случаев разница между этими вариантами гистограммы сводится лишь к разному масштабу y-оси — в остальном полученные графики совпадают. В дальнейшем мы будем рассматривать только вероятностные гистограммы.

Гистограмму можно считать оценкой функции плотности. Проиллюстрируем связь между гистограммой выборки из с.в. и её истинной функцией плотности. Вероятность попадания в выбранный интервал может быть оценена относительной частотой попадания в этот интервал. Поэтому (и в силу теоремы о среднем значении) относительная частота, деленная на длину интервала, является оценкой функции плотности в некоторой средней точке этого интервала:

$$\frac{\nu}{n\Delta} \approx \frac{1}{\Delta}\int_a^{a+\Delta} f(x)\,\mathrm{d}x = f(x_{\text{средн}}), \quad x_{\text{средн}} \in (a; a+\Delta).$$

Для построения гистограммы необходимо

1) разбить область значений выборки на заданное число k интервалов (считая оба бесконечных крайних интервала);
2) для каждого $j = 1, \ldots, k$ подсчитать количество ν_j выборочных данных, попавших в j-ый интервал;
3) построить график ступенчатой кривой, у которой высота ступеньки над j-ым интервалом пропорциональна ν_j;

4) наложить на график гистограммы график функции плотности предполагаемого распределения (например, нормального).

Разберем по порядку каждый из этих пунктов.

1) Чаще всего все внутренние конечные интервалы выбираются одинаковой длины. Поэтому для построения разбиения достаточно сначала выбрать правую границу 1-го интервала a_1 и левую границу a_{k-1} последнего k-го интервала. Остальные границы вычисляются по формуле

$$a_j = a_1 + \Delta \cdot (j-1), \quad j = 2, \ldots, k-1,$$

с шагом $\Delta = (a_{k-1} - a_1)/(k-2)$. Левая граница 1-го интервала равна $a_0 = -\infty$, правая граница последнего интервала $a_k = +\infty$. По поводу выбора количества интервалов k и первой границы a_1 существует множество различных мнений. Здесь необходимо учитывать как качество визуального представления, так и дальнейшее использование этой гистограммы для проверки гипотез и оценки вероятностей попадания наблюдаемой случайной величины в различные области.

Если гистограмма используется только для визуального сравнения с функцией плотности, то, как вариант, можно взять число интервалов k равным приблизительно десятой части выборки, первую границу $a_1 = x_{\min} + \frac{\Delta}{2}$, а последнюю границу $a_{k-1} = x_{\max} - \frac{\Delta}{2}$, где $x_{\min}$ — минимальное, $x_{\max}$ — максимальное значения выборки, длина внутренних интервалов $\Delta = (x_{\max} - x_{\min})/(k-1)$.

Существуют и другие формулы для выбора k, например, формула Стёрджеса $k = 1 + \log_2 n$. Все они носят, так или иначе, эвристический характер и подразумевают дальнейшую «корректировку» под конкретные нужды исследователя.

2) Подсчет количества попаданий в каждый интервал можно осуществить без использования компьютера. Для этого необходимо на листе бумаги начертить схему расположения интервалов (можно без соблюдения масштаба) и последовательно просмотреть все данные. При попадании очередного числа в j-ый интервал нужно поставить над этим интервалом точку. Количество точек над каждым из интервалов по окончании просмотра и будет равно искомой частоте.

3) Гистограммы бывают двух типов: частотная (высота столбца ν_j) и вероятностная ($\nu_j/(n\Delta)$). Во втором случае график гистограммы будет соизмерим с графиком функции плотности. При выполнении заданий необходимо ориентироваться только на вероятностный тип.

4) Как уже было сказано, гистограмма представляет собой некую оценку функции плотности. Поэтому естественно попытаться сопоставить график гистограммы с графиком ожидаемой плотности $f(x)$. В идеале, если данные получены из распределения с плотностью $f(x)$, то график функции $f(x)$ должен пересечь каждую ступеньку графика гистограммы.

Замечание. При одновременном прорисовывании графика гистограммы и функции плотности следует помнить о выбранном масштабе представления. Другими словами, если используется частотный вариант гистограммы, то каждое значение функции плотности: $n\Delta \cdot f(x)$ придется умножить на константу $n\Delta$. Чтобы построить по возможности более точный график плотности, значения функции плотности необходимо было бы вычислить в большом числе точек.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Как строится гистограмма?

3) Как связаны значения гистограммы и функции плотности?

4) Оцените вероятность попадания в интервал.

5) Почему следует сравнивать гистограмму с нормальной плотностью?

6) Выпишите формулу плотности нормального закона (равномерного, экспоненциального)?

7) Чему полагаются равными параметры нормального закона (равномерного, экспоненциального) при отрисовке функции плотности?

Задание 3. Эмпирическая функция распределения

1. Постановка задачи

Построить график эмпирической функции распределения выборки.

2. Теоретические основы

Эмпирическая функция распределения (ЭФР):

$$F_n(x) = \frac{\text{число выборочных данных } X_i, \text{ для которых } X_i < x}{n} = \\ = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i < x),$$

где $\mathbb{I}(A)$ — индикаторная функция (события A), т.е. функция, принимающая значение 1, когда событие случилось, и 0 — когда не случилось.

Она служит оценкой истинной функции распределения и представляет собой возрастающую от 0 до 1 ступенчатую функцию. Изменения $F_n(x)$ происходят скачком в точках x, совпадающих с каким-либо выборочным значением X_i. Высота этого скачка равна числу выборочных данных, равных X_i, поделённому на общий объем выборки n. Внутри любого интервала значений x, не содержащего выборочных данных, функция $F_n(x)$ остаётся неизменной.

В отличие от гистограммы, ЭФР является достаточной статистикой, то есть сохраняет всю полноту информации выборки. Кроме того, при увеличении объёма выборки она сходится к истинной функции распределения $F(x)$:

$$D_n = \sup_x |F_n(x) - F(x)| \xrightarrow{\mathbf{P}} 0, \quad n \to \infty. \tag{6}$$

Другими словами, она является состоятельной оценкой $F(x)$. Легко показать, что ЭФР также и несмещённая оценка $F(x)$.

Для построения ЭФР необходимо сначала построить так называемый вариационный ряд — ряд упорядоченных выборочных данных $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$. Обратим внимание здесь на различие в написании индексов в исходной выборке X_i и в вариационном ряду $X_{(j)}$. В первом случае индекс указывает на номер в порядке поступления выборочного значения, а во втором случае — на его ранг, то есть на место, которое это значение занимает в ранжированном по возрастанию ряду выборочных данных. Следовательно, всегда $X_{(1)}$ —

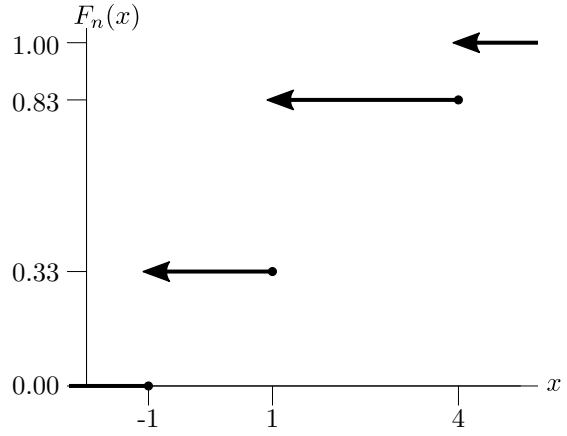

Рис. 5: Пример графика ЭФР, построенной по выборке $(-1, -1, 1, 1, 1, 4)$

минимальное значение выборки, $X_{(n)}$ — её максимальное значение при объеме выборки n. Для значений x, попадающих в интервал между k-ым и $(k+1)$-ым значениями вариационного ряда $(X_{(k)} < x \leqslant X_{(k+1)})$ ЭФР $F_n(x) = k/n$. В частности, заметим, что если $X_{(k-1)} < X_{(k)}$, то $F_n(X_{(k)}) = (k-1)/n$.

В качестве примера рассмотрим рис. 5, на котором приведён график ЭФР, построенной по 6 данным, среди которых -1 встречается два раза, 1 — три раза, а 4 — один раз.

Как и при построении гистограммы, график ЭФР полезно сравнить с графиком предполагаемого распределения (например, нормального). При этом некоторую информацию о степени достоверности этого распределения — правильности выдвинутого предположения о виде распределения — будет нести величина расхождения D, вычисленная по формуле (6). Неизвестные параметры модели можно оценить по выборке. В нормальной модели среднее μ оценивается выборочным средним $\overline{X}$, а дисперсия σ^2 — выборочной дисперсией S^2. Для показательного закона интенсивность отказа λ также может быть оценена посредством $\overline{x}$.

Если бы предполагаемое распределение было известно полностью, и не надо было оценивать неизвестные параметры, то на основе значений D_n можно было бы построить критерий проверки адекватности этого распределения выборочным данным — так называемый критерий Смирнова (см., например, сборник таблиц [3]). Применять этот критерий к моделям с неизвестными параметрами нельзя, поскольку вероятность ошибки 1-го рода такого критерия будет зависеть от неизвестных параметров.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Что такое вариационный ряд?

3) Дайте определение ЭФР?

4) Почему некоторые ступеньки ЭФР высокие, а некоторые низкие?

5) Почему одни ступеньки ЭФР длинные, а другие короткие?

6) Постройте ЭФР по следующим данным: 1; 2; 1; 3; 1; 5; 1; 3.

7) Выпишите формулу для функции распределения нормального закона (равномерного, экспоненциального).

8) Можно ли утверждать, что ЭФР является состоятельной оценкой истинной функции распределения? Что сие означает?

9) Можно ли утверждать, что ЭФР является несмещённой оценкой истинной функции распределения? Что сие означает?

10) Проверьте состоятельность и несмещённость ЭФР.

Часть III
Проверка гипотезы о типе распределения

Здесь описывается наиболее популярный метод проверки согласия выборочных данных с гипотезой о типе распределения.

Пусть $X \sim F$. В рамках задач на проверку гипотез о типе распределений рассматривают две основные разновидности гипотез:

1) $\mathrm{H}_0\colon F = F_0$ — распределение с.в. X в точности совпадает с F_0 (вплоть до значений параметров распределения).

2) $\mathrm{H}_0\colon F \in \Psi$ — распределение с.в. X принадлежит семейству распределений Ψ. Обычно, это означает, что мы проверяем принадлежность X некоторому типу распределения с точностью до одного или нескольких неизвестных параметров. Например, нормальному закону при неизвестной дисперсии $\Psi = \{\mathcal{N}(\mu, \sigma)\colon \sigma \geqslant 0\}$, либо нормальному закону при неизвестных среднем и дисперсии $\Psi = \{\mathcal{N}(\mu, \sigma)\colon \mu \in \mathbb{R},\, \sigma \geqslant 0\}$.

На практике, обычно, проверяется второй тип гипотезы согласия о принадлежности семейству. Однако при проверке сложных гипотез возникают некоторые теоретические сложности при обосновании надёжности получаемого статистического вывода. В случае проверки гипотезы о точном совпадении распределения заданному таких сложностей не возникает; однако и на практике такой задачи почти никогда не возникает.

Теоретические основы:

- критерии проверки гипотез смотрите в п. 3, с. 15.

Задание 4. Критерий согласия хи-квадрат

1. Постановка задачи

Требуется проверить гипотезу $\mathrm{H}_0\colon F \in \Psi$ о том, что функция распределения выборочных данных F принадлежит заданному семейству распределений Ψ (нормальному, экспоненциальному или равномерному).

2. Теоретические основы

В качестве критерия проверки такой гипотезы чаще всего выбирают критерий согласия хи-квадрат. Для принятия решения в соответствии с этим критерием необходимо:

1) Выдвинуть гипотезу $\mathrm{H}_0\colon F \in \Psi$ о виде распределения выборочных данных F.

2) Разбить область значений наблюдаемой с.в. на r интервалов.

3) По n выборочным данным подсчитать таблицу частот ν_k, аналогичную гистограммной таблице.

4) Для каждого интервала вычислить теоретические вероятности p_i попадания в этот интервал.

5) Вычислить тестовую статистику

$$T = \sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k},$$

представляющую собой некую меру расхождения между ожидаемыми(теоретическими) частотами np_k и выборочными (полученными в эксперименте) частотами ν_k.

6) Вычислить p-значение/$C_{\text{крит}}$ критерия.

7) Принять решение о статистической значимости проверяемой гипотезы.

Разберем каждый пункт на примере нормального закона с неизвестными параметрами распределения.

1) В этом случае в качестве Ψ выступает семейство распределений, образованное всевозможными комбинациями значений неизвестных параметров μ и σ^2:

$$\Psi = \{\mathcal{N}(\mu, \sigma)\colon \mu \in \mathbb{R},\ \sigma \geqslant 0\}.$$

2) Для теоретического обоснования надёжности критерия, разбиение числовой прямой на интервалы (как и число этих интервалов) необходимо выбирать независимо от наблюдённых в эксперименте значений. Однако практически это требование можно полноценно реализовать, только если гипотетическое распределение известно полностью (т.е. при рассмотрении простой гипотезы о полном совпадении распределения заданному). В противном случае очень часто будет наблюдаться ситуация, когда большинство выборочных данных попадут в один интервал.

Поэтому на практике интервалы строят в соответствии с выборочными данными. В качестве одного из возможных способов такого построения предлагается просто разбить размах выборочных данных (от $X_{\min}$ до $X_{\max}$) на r равных интервалов с последующим расширением двух крайних интервалов до $\pm\infty$. В итоге будет получено некоторое разбиение числовой пример на интервалы $I_1, \ldots, I_r$.

Один из вариантов выбора r — взять $r = n/10$. Другое довольно известное правило Стёрджеса предлагает использовать $r = 1 + \log_2 n$. Есть также предложение положить ширину интервалов равной $3.5Sn^{-1/3}$, где S — стандартное отклонение. Надо понимать, что все этим варианты имеют под собой довольно зыбкое основание и не могут быть использованы как универсальное правило выбора разбиения.

3) Пусть I_k — некоторый интервал разбиения, тогда

$$\nu_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in I_k).$$

4) Теоретическая вероятность попадания в интервал $I_k = [a_{k-1}, a_k)$ равна

$$p_k = F_0(a_k) - F_0(a_{k-1}),$$

где функция распределения F_0 либо совпадает с теоретической, если проверяется простая гипотеза, либо для её вычисления в гипотетическом распределении неизвестные параметры заменяются оценками.

В качестве оценок параметров нормального распределения можно взять выборочное среднее $\overline{X}$ и выборочную дисперсию S^2. В этом случае

$$F_0(x) = \Phi\left(\frac{x - \overline{X}}{S}\right).$$

Для двух крайних интервалов $(-\infty, a_1)$ и $[a_{r-1}, \infty)$ вероятности равны $p_1 = F_0(a_1)$ и $p_r = 1 - F_0(a_{r-1})$, соответственно.

6) Для вычисления p-значения и $C_{\text{крит}}$ необходимо знать функцию распределения $G(x)$ тестовой статистики T. Если бы гипотетическое распределение было известно точно, то есть были бы известны все параметры проверяемой модели (в данном случае — точное значение μ и σ^2), то при большом объёме выборки функцию $G(x)$ можно было бы аппроксимировать хи-квадрат распределением $\mathrm{Fchisq}(x \,|\, r-1)$ с $r-1$-ой степенью свободы. В этом случае критический уровень значимости

$$\mathsf{p} \approx 1 - \mathrm{Fchisq}(T \,|\, r-1).$$

Если параметры модели оцениваются по выборке, то функция $G(x)$ начинает зависеть от этих параметров. Известно, однако, что если m

неизвестных параметров оцениваются по методу максимального правдоподобия, то при $n \to \infty$ справедливо неравенство

$$\mathrm{Fchisq}(x \,|\, r-1) \leqslant G(x) \leqslant \mathrm{Fchisq}(x \,|\, r-1-m)$$

Поэтому, например, при проверке гипотезы нормальности (модель с двумя неизвестными параметрами) необходимо вычислить два значения:

$$\mathsf{p}_{r-1} = 1 - \mathrm{Fchisq}(T \,|\, r-1) \quad \text{и} \quad \mathsf{p}_{r-3} = 1 - \mathrm{Fchisq}(T \,|\, r-3)$$

Замечание. Кстати, если границы интервалов выбирать в зависимости от данных (что и происходит на практике), то поведение функции распределения $G(x)$ становится ещё более непредсказуемым.

7) Выводы о справедливости гипотезы делаются в зависимости от расположения p_{r-m-1}, p_{r-1} относительно выбранного уровня значимости α. Если

- $\mathsf{p}_{r-m-1} > \alpha$ — гипотеза принимается с $\mathsf{p} = \mathsf{p}_{r-m-1}$.
- $\mathsf{p}_{r-1} < \alpha$ — гипотеза отвергается с $\mathsf{p} = \mathsf{p}_{r-1}$.
- $\mathsf{p}_{r-m-1} \leq \alpha \leq \mathsf{p}_{r-1}$ — нет достаточных оснований для принятия какого-либо решения ($\mathsf{p} = \alpha$).

Замечание. Критерий хи-квадрат носит название критерия согласия, поскольку при его применении не учитывается вид альтернативы и вывод, который при этом делается, означает только, что или мы гипотезу принимаем — данные согласуются с гипотезой, или отвергаем — данные не согласуются с гипотезой.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Как строится критерий согласия хи-квадрат?

3) Почему критерий называется критерием согласия?

4) Выпишите формулу тестовой статистики критерия согласия хи-квадрат. Почему эту статистику можно считать мерой близости выборочных данных к выдвинутой гипотезе?

5) Какое распределение имеет статистика критерия хи-квадрат?

6) Почему иногда приходится вычислять два критических уровня значимости?

7) Чему равен критический уровень значимости при проверке гипотезы о равномерном (нормальном, экспоненциальном) распределении?

8) Почему при построении критерия хи-квадрат нельзя выбирать интервалы группировки в зависимости от выборочных данных?

9) Докажите, что $\mathsf{p}_{r-3} < \mathsf{p}_{r-1}$.

Задание 5. Критерий согласия Колмогорова

1. Постановка задачи

Для выборки $X^{(n)} \sim F$ требуется проверить гипотезу $\mathrm{H}_0\colon F = G$ о том, что функция распределения выборочных данных совпадает с заданным распределением G.

Замечание. Критерий согласия Колмогорова можно использовать только, когда все параметры распределения G заранее заданы, т.е. не оцениваются исходя из выборочных данных. В обратном случае ошибки 1-го и 2-го рода критерия могут оказаться непредсказуемо высокими.

2. Теоретические основы

Идея критерия Колмогорова заключается в сравнении ЭФР F_n и функции распределения G. Критерий Колмогорова отклоняет H_0, если функции слишком сильно расходятся. Реализуется эта идея в терминах статистики

$$D_n = \sup_x |F_n(x) - G(x)|.$$

Известно, что, в случае верности H_0, $\forall t > 0$

$$\lim_{n\to\infty} \mathbf{P}(\sqrt{n} D_n < t) = \mathrm{Fkolm}(t),$$

т.е. статистика $\sqrt{n} D_n$ асимптотически имеет распределение Колмогорова $\mathrm{Fkolm}(t)$.

Критерий Колмогорова с уровнем значимости α отклоняет H_0, если

$$\sqrt{n} D_n > \mathrm{Fkolm}^{-1}(1-\alpha),$$

где $\mathrm{Fkolm}^{-1}(1-\alpha)$ — $(1-\alpha)$-квантиль распределения Колмогорова.

Замечание. В большинстве пакетов вычислительной статистики нет методов для вычисления функции распределения Колмогорова и её квантилей.

Для вычисления значения функции распределения при больших значениях t можно воспользоваться аппроксимацией

$$\mathrm{Fkolm}(t) \approx 1 - 2\mathrm{e}^{-2t^2}.$$

Для вычисления квантили при малых α можно воспользоваться аппроксимацией:

$$\mathrm{Fkolm}^{-1}(1-\alpha) \approx \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}.$$

Замечание. Иногда этот критерий называют критерием Колмогорова-Смирнова.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Как строится критерий Колмогорова?

3) Почему большие значения статистики D_n свидетельствуют о расхождении с гипотезой?

4) Какое распределение имеет статистика $\sqrt{n}D_n$?

5) Как вычислить супремум $\sup_x |F_n(x) - G(x)|$? Всегда ли этот супремум достигается в некоторой точке?

6) Как вычислить p-значение критерия Колмогорова?

7) Можно ли проверять гипотезу нормальности выборочных данных с помощью критерия Колмогорова?

8) Можно ли проверять гипотезу о том, что выборочные данные происходят из нормального $\mathcal{N}(0,1)$ распределения?

Часть IV
Проверка гипотез однородности

В практике применения методов статистического анализа часто возникает задача сравнения различных совокупностей выборочных данных с целью выяснения их «однородности». Например, при исследовании лечебных свойств нового препарата требуется

- сравнить воздействие этого препарата на некоторую характеристику здоровья в одной группе пациентов (— первая выборка) с воздействием старого препарата на ту же характеристику в другой группе пациентов (— вторая выборка)

- сравнить характеристики здоровья у пациентов одной группы до лечения (— первая выборка) и после лечения препаратом (— вторая выборка)

- сравнить долю выздоровевших пациентов при различных способах лечения.

В общем виде задачу можно сформулировать следующим образом: имеются две совокупности выборочных данных; требуется определить, есть ли значимые отличия между ними. В терминах гипотез эта задача может быть выражена как нулевая гипотеза о совпадении распределений выборок в рамках некоторых предположений об их природе.

Заметим, что каждый из рассмотренных ниже критериев фактически проверяет совпадение распределений лишь в терминах какого-либо преобразования, или в рамках неких предположений. Поэтому вывод об «однородности» двух выборок является лишь некоторой интерпретацией результатов применения критерия, и ответственность за правильность этой интерпретации целиком ложится на плечи исследователя.

Выбор того или иного критерия для проведения исследования в некой текущей задаче зависит от наличия у исследователя тех или иных априорных знаний о природе наблюдаемых характеристик, или возможности сделать о ней соответствующие предположения. Например, исходя из ответов на вопросы:

1) Имеют ли выборки нормальное распределение?

2) Можно ли считать наблюдения в различных выборках независимыми?

Теоретические основы:

- критерии проверки гипотез смотрите в п. 3, с. 15.

Задание 6. Одновыборочный критерий Стьюдента

1. Постановка задачи

Одновыборочный вариант. Имеется одна выборка из нормального распределения. Требуется проверить гипотезу о том, что среднее значение этого распределения не превосходит заданной величины $C_{\text{норм}}$.

Двухвыборочный вариант. Имеются две выборки $X^{(n)}$, $Y^{(n)}$ одинакового объёма. Известно, что для каждого i наблюдения X_i и Y_i являются наблюдениями одного и того же объекта, но сделанные в разных условиях (например, до и после применения лекарства). Предполагается нормальное распределение этих выборок.

Требуется проверить гипотезу однородности выборок из X и Y.

Замечание. Фактически двухвыборочный вариант критерия реализуется сведением задачи к одновыборочному варианту критерия — вместо исходных выборок из X и Y рассматривается их разность $U_i = X_i - Y_i$, относительно которой вводится нулевая гипотеза о равенстве нулю её среднего.

2. Теоретические основы

Одновыборочный вариант. Пусть $X^{(n)}$ — выборка из $X \sim \mathcal{N}(\mu, \sigma^2)$ с неизвестными μ и σ^2. Пусть для некоторой константы $m_0 \in \mathbb{R}$ рассматривается гипотеза

$$\mathrm{H}_0 \colon \mu = m_0,$$

В качестве альтернативы, обычно, рассматриваться гипотезы вида

$$\mathrm{H}_1 \colon \mu \neq m_0, \quad \mathrm{H}_1 \colon \mu < m_0 \quad \text{или} \quad \mathrm{H}_1 \colon \mu > m_0.$$

Выбор более «узкого» варианта альтернативы может увеличить мощность итогового критерия, т.е. вероятность правильного обнаружения случая верности альтернативы.

Замечание. При проверке гипотез рассматриваемыми критериями, по сути, единственное статистически значимое решение (т.е. про которое можно сказать, что оно верно с высоким уровнем уверенности)

— это принятие по результатам эксперимента альтернативной гипотезы H_1. Связано это с тем, что такие критерии контролируют только ошибку 1-го рода (т.е. гарантируют её малость) — вероятность принять H_1, когда в действительности верна H_0. В связи с этим возникает ''парадоксальная'' ситуация — хотя формально критерии проверяют H_0, но фактически именно выбор H_1 соответствует тому явлению, которое хочет обнаружить исследователь. Нулевая же гипотеза H_0 почти всегда соответствует некому «нейтральному» или «плохому» явлению, и определяется таким образом, чтобы упростить построение и вычисление критерия.

Статистика одновыборочного критерия Стьюдента равна

$$T = \frac{\overline{X} - m_0}{S_X}\sqrt{n-1}, \tag{7}$$

где $\overline{X}$ — выборочное среднее, а S_X^2 — смещённая выборочная дисперсия, вычисленные по выборке $X^{(n)}$. Если справедливо предположение о нормальности распределения X и верна нулевая гипотеза H_0, то статистика T имеет распределение Стьюдента $\mathrm{Fstud}(t \,|\, n-1)$ с $(n-1)$-ой степенью свободы.

Из выражения (7) для T можно увидеть, что большим значениям μ соответствуют, в среднем, большие значения T. Исходя из этого, разным альтернативам H_1 соответствуют следующие критическая область критерия, $C_{\text{крит}}$ и выражение для вычисления p-значения:

H_1	вид крит. обл.	$C_{\text{крит}}$	p-значение
$\mu > m_0$	$\{T\colon T > C_{\text{крит}}\}$	$\mathrm{Fstud}^{-1}(1-\alpha \,|\, n-1)$	$1 - \mathrm{Fstud}(t \,|\, n-1)$
$\mu < m_0$	$\{T\colon T < C_{\text{крит}}\}$	$\mathrm{Fstud}^{-1}(\alpha \,|\, n-1)$	$\mathrm{Fstud}(t \,|\, n-1)$
$\mu \neq m_0$	$\{T\colon \lvert T\rvert > C_{\text{крит}}\}$	$\mathrm{Fstud}^{-1}(1-\alpha/2 \,|\, n-1)$	$2(1 - \mathrm{Fstud}(t \,|\, n-1))$

Двухвыборочный вариант. Пусть $X^{(n)}, Y^{(n)}$ — это выборки из с.в. X и Y, соответственно. Предполагается, что i-ые наблюдения X_i, Y_i в выборках являются наблюдениями характеристик одного и того же i-го объекта (по-другому — наблюдаются значения двухмерного случайного вектора (X, Y)). Например, в задаче клинических испытаний медицинских препаратов X_i может быть значением артериального давления пациента i до принятия препарата, а Y_i — после принятия препарата. Пусть

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2).$$

Введём с.в. $U = X - Y$; соответственно, наблюдения $U_i = X_i - Y_i$. Тогда $U \sim \mathcal{N}(\mu, \sigma^2)$ с $\mu = \mu_X - \mu_Y$. Рассмотрим гипотезу

$$\mathrm{H}_0\colon \mu_X = \mu_Y,$$

и одну из альтернатив

$$\mathrm{H}_1\colon \mu_X \neq \mu_Y, \quad \mathrm{H}_1\colon \mu_X < \mu_Y \quad \text{или} \quad \mathrm{H}_1\colon \mu_Y > \mu_Y.$$

Очевидно, эту гипотезу/альтернативу можно проверить, если применить к выборке $U^{(n)}$ ранее рассмотренного одновыборочного варианта одновыборочного критерия Стьюдента для проверки гипотезы/альтернативы о значении μ относительно нуля. Таким образом, двухвыборочный вариант критерия сводится к применению одновыборочного варианта к разности выборок и введением соответствующей пары гипотезы/альтернативы.

Замечание. Одновыборочный критерий Стьюдента фактически проверяет лишь равенство математических ожиданий выборок. Однако в приложениях часто можно заранее предполагать равенство дисперсий выборок — тогда совпадение математических ожиданий вкупе с предположением о нормальности распределений будет означать полное совпадение распределений выборок, однородность.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу

2) Как вычисляется статистика одновыборочного критерия Стьюдента?

3) Когда следует применять критерий Стьюдента, а когда критерий знаков?

4) Чему равен критический уровень значимости для критерия Стьюдента при двухсторонней альтернативе?

5) Можно ли к рассматриваемым данным применить критерий однородности хи-квадрат?

6) Обязаны ли с.в. X и Y для двувыборочного варианта критерия быть независимыми?

7) Если с.в. X и Y независимы в двувыборочном варианте, то чему равна дисперсия σ^2 распределения $U = X - Y$?

8) Можно ли однозначно утверждать, что пребывание в спортивном летнем лагере повышает спортивную форму, если средний вес случайно отобранной части студентов после пребывания в лагере уменьшился на 7 кг?

Задание 7. Критерий знаков

1. Постановка задачи

Одновыборочный вариант. Имеется выборка $X^{(n)}$ из с.в. $X = \mathbb{I}(A)$, где A — некоторое событие. Требуется проверить гипотезу, что событие A происходит чаще, чем противоположное к этому событию утверждение (например, лечение чаще приводит к выздоровлению).

Двухвыборочный вариант. Имеются две выборки $X^{(n)}$, $Y^{(n)}$ одинакового объема. Известно, что каждая пара i-х наблюдений X_i, Y_i — наблюдения некоторых характеристик (с.в. X и Y) у одного и того же i-го объекта. Распределение X и Y неизвестно. Требуется проверить гипотезу однородности выборок.

Замечание. Фактически двухвыборочный вариант критерия реализуется сведением задачи к одновыборочному варианту — вместо исходных выборок из X и Y рассматривается преобразование вида $U_i = \mathbb{I}(X_i < Y_i)$. Далее относительно распределения выборки $U^{(n)}$ рассматривается гипотеза о том, что вероятность успеха равна 0.5.

2. Теоретические основы

Одновыборочный вариант. Так как X определяется как индикаторная функция события, то $X \sim \mathrm{Fbern}(p)$, где $p = \mathbf{P}(A)$. В рамках критерия знаков гипотезы формулируются в терминах параметра вероятности успеха p:

$$\mathrm{H}_0 \colon p = p_0,$$

$$\mathrm{H}_1 \colon p < p_0, \quad \mathrm{H}_1 \colon p \neq p_0, \quad \mathrm{H}_1 \colon p > p_0,$$

где $p_0 \in [0, 1]$ — заданный уровень, относительно которого сравнивается вероятность p возникновения события A. При постановке задачи в виде «проверить гипотезу, что событие A происходит чаще противоположного» полагают $p_0 = 0.5$.

Статистика критерия знака имеет вид

$$M = \sum_i X_i \sim \mathrm{Fbin}(n, p).$$

Большим значениям параметра p в среднем соответствуют большие значения статистики M. Заметим, что в силу дискретности биномиального распределения, не всегда возможно подобрать критическую область так, чтобы ошибка 1-го рода получаемого критерия в точности равнялась заданному уровню значимости α. Исходя из конкретного вида H_1, критическая область может принимать вид:

1) $\mathrm{H}_1\colon p > p_0$. Вид критической области: $\{M\colon M > C_{\text{крит}}\}$.
   В силу дискретности биномиального закона, функция $\mathrm{Fbin}^{-1}(1 - \alpha \,|\, n, p_0)$ может быть не определена в точке $1-\alpha$. Поэтому критическая константа определяется как

$$C_{\text{крит}} = \min\{C \in \mathbb{Z}\colon \mathrm{Fbin}(C+1 \,|\, n, p_0) > 1-\alpha\}.$$

   p-значение: $\mathsf{p} = 1 - \mathrm{Fbin}(M+1 \,|\, n, p_0)$.

2) $\mathrm{H}_1\colon p < p_0$. Вид критической области: $\{M\colon M < C_{\text{крит}}\}$.
   Критическая константа:

$$C_{\text{крит}} = \max\{C \in \mathbb{Z}\colon \mathrm{Fbin}(C \,|\, n, p_0) \leqslant \alpha\}.$$

   p-значение: $\mathsf{p} = \mathrm{Fbin}(M \,|\, n, p_0)$.

3) $\mathrm{H}_1\colon p \neq p_0$. В силу симметричности биномиального распределения при $p = 0.5$, критическую область можно строить в виде:

$$\{M\colon M < C_{\text{крит}} \quad \text{или} \quad M > n - C_{\text{крит}}\}.$$

   Критическая константа:

$$C_{\text{крит}} = \max\left\{C \in \mathbb{Z}\colon \mathrm{Fbin}(C \,|\, n, p_0) \leqslant \frac{\alpha}{2}\right\}.$$

   p-значение: $\mathsf{p} = \begin{cases} \mathrm{Fbin}(M \,|\, n, p_0)/2, & \text{если } M \leqslant n/2, \\ \mathrm{Fbin}(n - M \,|\, n, p_0)/2, & \text{если } M > n/2. \end{cases}$

   Замечание Обойтись лишь одной критической константой при построении критерия для $\mathrm{H}_1\colon p \neq p_0$ можно только при $p_0 = 0.5$. Для других случаев, в общем случае, приходится иметь дело с парой критических констант C_1, C_2 и критической областью вида $[0, C_1) \cup (C_2, n]$.

Пример 1. Описанную схему можно применять также для проверки гипотезы о вероятности «успеха» при биномиальных испытаниях — одновыборочный вариант критерия. В качестве примера рассмотрим ситуацию, когда при составлении договора купли-продажи заказчиком была оговорена нижняя граница в 92% для доли доброкачественной продукции. При поступлении товара заказчик проводит контрольные измерения n единиц продукции. По результатам испытаний, основываясь только на количестве кондиционной продукции, требуется проверить гипотезу $\mathrm{H}_0\colon p \leqslant 0.92$ (опять же гипотеза противоположна ожиданиям).

Пример 2. Другой пример. С целью прогнозирования результатов будущих выборов было опрошено 1000 респондентов. Среди них оказалось 35 сторонников партии «Будет ещё хуже!». Можно ли утверждать, что эта партия не попадет в думу?

Если считать 1000 респондентов каплей в море всех избирателей и отбор респондентов производился абсолютно случайно, то можно описать наши наблюдения как выборку из распределения Бернулли с вероятностью успеха p, равной доле всех сторонников указанной партии. Если мы находимся на позициях противников партии «Будет ещё хуже!», мы хотели бы, что бы эта доля была меньше 0.05. Поэтому в качестве альтернативы мы должны выбрать утверждение $\mathrm{H}_1\colon p \leqslant 0.05$.

Итак, $n = 1000$, $m = 35$, граничное значение $p_0 = 0.05$. Критический уровень значимости равен $\mathsf{p} = \mathbf{P}(M \leqslant m \,|\, p = 0.05) = 0.014$. Вывод: скорее всего (с надежностью 98.6%), партия «Будет ещё хуже!» не пройдет в Думу.

Двухвыборочный вариант. Если исследователя интересует лишь факт наличия эффекта воздействия и нет оснований предполагать какое-либо распределение у X и Y, то можно каждую пару исходных данных (X_i, Y_i) заменить величиной Z_i, принимающей всего два значения: $Z_i = 1$, если эффект есть, и $Z_i = 0$, если эффекта нет. Под эффектом может пониматься, например, уменьшение артериального давления после лечения или увеличение доли полезных веществ в пищевом продукте после стерилизации. Если предположить, что воздействие не обладает никаким эффектом, то $Z_i \sim \mathrm{Fbern}(p)$ с вероятностью успеха $p = 0.5$ (— приблизительно в 50% случаев должен наблюдаться эффект).

Исследователю было бы интересно проверить гипотезу

$$\mathrm{H}_0\colon p \leqslant 0.5$$

— эффект отсутствует или направлен противоположно, при альтернативе $\mathrm{H}_1\colon p > 0.5$.

В такой постановке эта гипотеза/альтернатива относительно $Z^{(n)}$ может быть проверена с использованием критерия знаков согласно описанной ранее процедуры.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Чему равна статистика критерия знаков?

3) Чему равен критический уровень значимости критерия знаков?

4) Когда следует применять критерий Стьюдента, а когда критерий знаков?

5) Чему равен размер приведённого здесь критерия? Как он соотносится с уровнем значимости?

6) Вычислите значение критического уровня значимости, если число успехов равно 6 при 9 испытаниях.

7) Проверьте гипотезу о том, что вероятность рождения мальчика равна 0.515, если среди 1000 новорожденных детей 509 оказались мальчики.

Задание 8. Двухвыборочный критерий Стьюдента

1. Постановка задачи

Имеются две выборки $X^{(n_1)}$, $Y^{(n_2)}$, относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону с одинаковыми для обеих выборок дисперсиями. Требуется проверить гипотезу однородности выборок.

2. Теоретические основы

Однородность выборок в предположении нормальности их распределения эквивалентна совпадению математических ожиданий и дисперсий. Критерий Стьюдента применяется в том случае, если можно априори, по тем или иным соображениям, предположить, что дисперсии одинаковы. Например, одна и та же характеристика измеряется одним и

тем же прибором, и разброс данных обусловлен исключительно ошибками этого прибора. Другой пример, обычный для медицинской практики, — сравнение показателей здоровья в двух группах пациентов, подвергнутых двум различным методам лечения, причём группы сформированы одинаковым образом. В этом случае гипотеза однородности эквивалентна равенству математических ожиданий $\mathrm{H}_0\colon \mu_1 = \mu_2$.

Статистика двухвыборочного критерия Стьюдента равна

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{n_1 S_X^2 + n_2 S_Y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

где $\overline{X}, \overline{Y}$ — выборочные средние, а S_X^2, S_Y^2 — выборочные дисперсии (смещённые оценки) первой и второй выборки, соответственно. Если выборки независимы, и происходят из нормального распределения с одинаковыми параметрами, то статистика Стьюдента T имеет распределение Стьюдента $\mathrm{Fstud}(t \,|\, n_1 + n_2 - 2)$ с $(n_1 + n_2 - 2)$ степенями свободы.

Поэтому критический уровень значимости при односторонней альтернативе $\mathrm{H}_1\colon \mu_1 < \mu_2$ равен

$$\mathsf{p} = \mathbf{P}(T \leqslant t) = \mathrm{Fstud}(t \,|\, n_1 + n_2 - 2).$$

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Какие предположения лежат в основе применения двувыборочного критерия Стьюдента?

3) Как вычисляется статистика двухвыборочного критерия Стьюдента?

4) В каких случаях следует применять одновыборочный критерий Стьюдента, а в каких — двувыборочный?

5) Когда следует применять критерий Стьюдента, а когда критерий Вилкоксона?

6) Чему равен критический уровень значимости для критерия Стьюдента при двухсторонней альтернативе; при односторонней альтернативе типа – «в первой группе больше»?

7) Можно ли к рассматриваемым данным применить критерий однородности хи-квадрат?

8) Что такое ошибка среднего? Какую смысловую нагрузку она несёт применительно к рассматриваемому критерию?

Задание 9. Критерий Вилкоксона

1. Постановка задачи

Имеются две выборки $X^{(n_1)}$, $Y^{(n_2)}$, относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда в качестве альтернативы ожидается, что значения в 1-й выборке будут ''равномерно'' меньше значений во второй выборке.

2. Теоретические основы

Пусть $X^{(n_1)}$ — выборка из X, $Y^{(n_2)}$ — выборка из Y. Описываемый здесь критерий применяется в том случае, когда

1) распределение выборки неизвестно, однако предполагается непрерывность распределений,

2) в качестве альтернативы однородности выборок выдвигается гипотеза
$$\mathrm{H}_1 \colon F_Y(x) = F_X(x - \Delta), \quad \forall x,\ \Delta > 0.$$

Другими словами, в предположении альтернативы распределение первой выборки (X-ов) сдвинуто влево относительно распределения второй выборки (Y-ов), то есть ожидаемые значения X-ов должны быть меньше значений Y-ов. Идею критерия Вилкоксона можно проиллюстрировать на следующем «крайнем» примере. Предположим, что все выборочные значения из 2-ой выборки больше всех значений из 1-ой выборки. Такая ситуация вполне ожидаема, если верна альтернатива. В этом случае, расположив обе выборки в один общий ряд, мы увидим, что 1-ая выборка занимает меньшие по порядку места (ранги), чем 2-ая выборка. Если же обе выборки равномерно перемешаны (что естественно, если верна гипотеза), то средние ранги для обеих выборок должны быть приблизительно равны. Таким образом, малые значения средних рангов 1-ой выборки будут свидетельствовать в пользу альтернативы.

Для построения критерия Вилкоксона необходимо обе выборки расположить в один общий ряд, упорядоченный по возрастанию. При этом информация о принадлежности каждого значения к той или иной выборке не должна быть утеряна. Статистика Вилкоксона равна

$$W = \sum_{i=1}^{n_1} r_i,$$

где $r_1, \ldots, r_{n_1}$ — ранги всех значений 1-ой выборки (— той, для которой альтернатива предполагает сдвиг влево). Известно, что при нулевой гипотезе статистика W имеет так называемое распределение Вилкоксона (Wilcoxon) с параметрами (n_1, n_2). Нулевая гипотеза отвергается, если

$$W < C_{\text{крит}}$$

Критическая константа $C_{\text{крит}}$, как всегда, находится из условия

$$\mathbf{P}_0(W < C_{\text{крит}}) \leqslant \alpha,$$

где $\mathbf{P}_0$ означает, что вероятность события вычисляется в предположении верности нулевой гипотезы.

Другая форма критерия, как обычно, связана p-значением

$$\mathsf{p} = \mathbf{P}_0(W \leqslant w).$$

Для вычисления функции распределения Вилкоксона можно воспользоваться асимптотической формулой. Известно, что статистика W асимптотически нормальна с

- математическим ожиданием $\mu_W = \dfrac{n_1(n_1 + n_2 + 1)}{2} + 0.5$,
- дисперсией $\sigma_W^2 = \dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}$.

То есть, если w — выборочное значение статистики W, полученное по рангам 1-ой выборки, то для альтернативы H_1: «1-ая выборка сдвинута влево» p-значение равно

$$\mathsf{p} = \mathbf{P}_0(W \leqslant w) \approx \Phi\Big(\frac{w - \mu_W}{\sqrt{\sigma_W^2}}\Big),$$

где Φ — стандартная нормальная функция распределения.

Другая проблема, связанная с вычислением статистики Вилкоксона — совпадающие наблюдения. Если два наблюдения имеют одинаковые значения и принадлежат к одной группе, то статистика Вилкоксона не будет изменяться при случайных перестановках этих наблюдений в общем ряду данных. Если же совпадающие наблюдения принадлежат разным группам, то встаёт вопрос, какое из этих наблюдений следует поставить раньше? Чтобы избежать ненужного здесь «волюнтаризма», можно всем совпадающим значениям присвоить один и тот же ранг, равный среднему арифметическому мест, на которых эти значения находятся. Например, если четыре совпадающих значения занимают места с 9-го по 12-е, то все четыре значения получают ранг 10.5; последующим значениям выборки присваиваются ранги, начиная с 13. Распределение изменённой таким образом статистики Вилкоксона очень тяжело вычислить. В качестве «первого приближения» можно воспользоваться описанной выше методикой нахождения критической константы

по таблицам или методикой вычисления асимптотического уровня значимости.

Заметим, как удивительно точно «работает» асимптотическая формула для критического уровня значимости. Приведем фрагмент таблицы [3] с точными значениями квантилей статистики Вилкоксона в сравнении с приближёнными значениями:

n_1	n_2	α					
		0.001	0.005	0.01	0.025	0.05	0.10
12	15	106	115	120	127	133	141
		приближённые значения					
		105.2	115.7	120.8	128.3	134.8	142.2

Замечание. Мы рекомендуем пользоваться асимптотической формулой для распределения статистики Вилкоксона как при отыскании p-значения, так и критической константы $C_{\text{крит}}$.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Есть ли какие-либо предопложения о распределении выборок?

3) Как вычисляется статистика критерия Вилкоксона?

4) При каких альтернативах следует прибегать к критерию Вилкоксона?

5) Как присваивать ранги совпадающим значениям?

6) Чему равен критический уровень значимости критерия Вилкоксона?

7) Как вычислить критическую константу критерия Вилкоксона с использованием нормальной асимптотики?

Задание 10. Критерий Фишера. Критерий сравнения дисперсий

1. Постановка задачи

Имеются две выборки $X^{(n_1)}$, $Y^{(n_2)}$, относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону. Требуется сравнить дисперсии наблюдений в этих групп.

2. Теоретические основы

Пусть $X^{(n_1)}$ — выборка из X, $Y^{(n_2)}$ — выборка из Y. Пусть σ_X^2 и σ_Y^2 — дисперсии X и Y, соответственно. В рамках критерия Фишера рассматриваются гипотеза

$$\mathrm{H}_0\colon \sigma_X^2 = \sigma_Y^2$$

и альтернативы вида

$$\mathrm{H}_1\colon \sigma_X^2 < \sigma_Y^2, \quad \mathrm{H}_1\colon \sigma_X^2 > \sigma_Y^2, \quad \mathrm{H}_1\colon \sigma_X^2 \neq \sigma_Y^2.$$

Тестовая статистика критерия Фишера:

$$\mathcal{F} = \frac{\tilde{S}_X^2}{\tilde{S}_Y^2},$$

где $\tilde{S}_X^2$, $\tilde{S}_Y^2$ — несмещённые оценки дисперсий в соответствующих выборках. В предположении нормальности данных и при совпадении теоретических дисперсий (верности нулевой гипотезы) статистика $\mathcal{F}$ имеет распределение Фишера $\mathrm{Ffish}(x \,|\, k, m)$ с параметрами $k = n_1 - 1$ и $m = n_2 - 1$. Таблицы этого распределения имеются в большинстве справочников по математической статистике и в пакетах программ математического характера.

Замечание. В случае двусторонней альтернативы $\mathrm{H}_1 : \sigma_X^2/\sigma_Y^2 \neq 1$ область принятия нулевой гипотезы $\mathrm{H}_0 : \sigma_X^2/\sigma_Y^2 = 1$ можно выбрать в виде

$$\frac{1}{C_{\text{крит}}} \leqslant \mathcal{F} \leqslant C_{\text{крит}}.$$

При отыскании критической константы $C_{\text{крит}}$ и p-значения следует воспользоваться очевидным равенством для функции распределения Фишера: $\mathrm{Ffish}(x \,|\, k, m) = 1 - \mathrm{Ffish}(1/x \,|\, m, k)$. Таким образом, критическая константа $C_{\text{крит}}$ есть решение уравнения $2 - \mathrm{Ffish}(C \,|\, k, m) - \mathrm{Ffish}(C \,|\, m, k) = \alpha$. Как в этом случае вычислить p-значение?

Замечание. Во многих учебниках по математической статистике рекомендуют предварить применение двухвыборочного критерия Стьюдента проверкой гипотезы о равенстве дисперсий в группах. В соответствии с такой рекомендацией две выборки будут считаться не однородными, если или критерий Фишера отвергнет равенство дисперсий (обозначим такое событие через F), или критерий Фишера признает дисперсии равными (событие F^c), но критерий Стьюдента отвергнет гипотезу равенства средних значений (событие S). Другими словами, гипотеза однородности будет отвергаться, если произойдет событие $F + F^c S$. Размер такого критерия равен вероятности

$$\mathbf{P}\left\{F + F^{\mathrm{c}}\, S\right\} = \mathbf{P}\left\{F\right\} + \mathbf{P}\left\{F^{\mathrm{c}}\, S\right\}.$$

Если размер критерия Фишера равен α_1, а критерия Стьюдента — α_2, то первое слагаемое в последнем равенстве равно α_1. Про второе слагаемое можно сказать только, что оно не больше α_2. Таким образом, уровень значимости составного критерия $\alpha_1 + \alpha_2$. Достичь желаемого уровня в 5% (вернее, меньше, чем 5%) в этом случае можно, если положить $\alpha_1 = \alpha_2 = 0.025$.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Чему равна статистика Фишера?

3) Какое распределение имеет статистика Фишера?

4) Чему равен критический уровень значимости при односторонней альтернативе : «дисперсия 1-ой выборки меньше дисперсии 2-ой выборки»?

5) Можно ли применить критерий Фишера к проверке гипотезы о равенстве дисперсий для данных из задания 6?

6) Можно ли применять критерий Фишера, как предварительный тест для проверки условий применимости двухвыборочного критерия Стьюдента?

Задание 11. Критерий однородности хи-квадрат

1. Постановка задачи

Имеются две выборки $X^{(n_1)}$, $Y^{(n_2)}$, относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда неизвестна модель распределения выборок, и нет никакой информации о соотношении между этими выборками.

2. Теоретические основы

Рассмотрим ситуацию, когда нет никакой информации ни о нормальности распределения данных, ни о соотношении между группами типа «левее – правее». В этом случае можно воспользоваться идеей гистограммного представления данных и сравнить частоты попадания результатов измерений в различных группах в одни и те же интервалы числовой прямой.

Итак, пусть имеются две группы измерений объемов n_1 и n_2, для которых подсчитаны частоты ν_{i1}, $i = 1, \ldots, r$, и ν_{i2}, $i = 1, \ldots, r$, попадания данных в r одних и тех же интервалов. Если гипотеза однородности выборок верна, то относительные частоты ν_{i1}/n_1 и ν_{i2}/n_2 должны быть близки друг к другу. Это соображение приводит нас к следующей тестовой статистике.

Для каждого из интервалов $i = 1, \ldots, r$, подсчитаем общее число данных $\nu_{i\bullet} = \nu_{i1} + \nu_{i2}$, попавших в этот интервал. Статистика критерия однородности хи-квадрат равна

$$T = n_1 n_2 \sum_{i=1}^{r} \frac{1}{\nu_{i\bullet}} \Big(\frac{\nu_{i1}}{n_1} - \frac{\nu_{i2}}{n_2} \Big)^2 .$$

При справедливости гипотезы однородности распределение статистики T можно аппроксимировать распределением хи-квадрат с $(r-1)$-ой степенью свободы:

$$\mathbf{P}(T < t) \approx \mathrm{Fchisq}(t \,|\, r-1) \quad (n_1, n_2 \to \infty).$$

Ясно, что при справедливости гипотезы однородности статистика T будет принимать «малые» значения. p-значение критерия хи-квадрат равняется

$$\mathsf{p} = \mathbf{P}(T > t^*) \approx 1 - \mathrm{Fchisq}(t^* \,|\, r-1),$$

где через t^* обозначается значение статистики в текущем эксперименте.

Замечание 1. Кроме вывода об однородности или неоднородности групп, здесь полезно визуально сравнить распределения в группах. Для этого можно совместить гистограммы обеих выборок. Следует только помнить, что, поскольку объемы выборок могут быть различны, гистограммы должны быть построены по относительным (делённым на соответствующие объёмы выборок) частотам.

Замечание 2. Построенный критерий не зависит от способа, каким были получены частоты. Этот критерий можно использовать и для проверки однородности двух выборок, когда частоты представляют собой количества выборочных данных, удовлетворяющих произвольным взаимоисключающим условиям. Например, можно сравнить успеваемость по курсу «Математический анализ» в двух различных вузах по результатам тестирования части студентов (здесь r — количество градаций оценки при тестировании). Другой пример. В медицинской практике очень часто требуется сравнить новую методику лечения со старой методикой по результатам клинических наблюдений. При этом пациентов, прошедших курс лечения подразделяют, скажем, на $r = 3$ группы — а) не выздоровели, б) выздоровели, но через год болезнь повторилась, и в) выздоровели без последующего рецидива.

3. Больше двух выборок

Критерий однородности хи-квадрат может быть применён и в случае, если число выборок больше двух. Пусть ν_{ij} — число исходов в j-ой выборке $(j = 1, \ldots, s)$, попавших в i-ый интервал группировки $(i = 1, \ldots, r)$. Таким образом, данные могут быть представлены в виде таблицы, где, как и раньше, точка на месте одного из индексов означает суммирование данных по этому индексу при фиксированном другом индексе. Так, например, общий объем выборок равен $n = \nu_{\bullet\bullet}$.

Выборка / Интервал	1	...	s	Всего
1	ν_{11}	...	ν_{1s}	$\nu_{1\bullet}$
...	...	...	...	...
r	ν_{r1}	...	ν_{rs}	$\nu_{r\bullet}$
Всего	$\nu_{\bullet 1}$	...	$\nu_{\bullet s}$	$\nu_{\bullet\bullet} = n$

В предположениях гипотезы однородности статистика

$$T = n \left(\sum_{j=1}^{s} \sum_{i=1}^{r} \frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}} - 1 \right)$$

имеет асимптотическое хи-квадрат распределение с $m = (r-1)(s-1)$ степенями свободы:

$$\mathbf{P}(T < t) \approx \mathrm{Fchisq}(t \,|\, m), \quad n \to \infty.$$

Поэтому гипотеза однородности всех s выборок должна отвергаться, если p-значение

$$\mathsf{p} \approx 1 - \mathrm{Fchisq}(T \,|\, m)$$

меньше выбранного уровня значимости α.

4. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Что означает (в вероятностном смысле) однородность выборок?

3) Чему равна статистика критерия однородности хи-квадрат? Почему эта статистика может трактоваться как мера близости к гипотезе?

4) Чему равен критический уровень значимости критерия однородности?

5) Можно ли с помощью этого критерия проверить гипотезу о том, что 1-ая половина данных из задания 1 имеет такое же распределение, как и 2-ая половина данных? Как в этом случае следует строить таблицу частот?

6) При клинических испытаниях из 80 пациентов, лечившихся по новой методике, 85% полностью выздоровели. Можно ли сказать, что новая методика лучше старой, если из 50 пациентов, лечившихся по старой методике, 35 пациентов полностью выздоровели?

Задание 12. Критерий однородности Смирнова

1. Постановка задачи

Пусть $X^{(n)} \sim F$ и $Y^{(m)} \sim G$ — две независимые выборки. Требуется проверить гипотезу $\mathrm{H}_0\colon F = G$ о совпадении функции распределения выборки $X^{(n)}$ с функцией распределения выборки $Y^{(m)}$.

2. Теоретические основы

Концептуально критерий Смирнова схож с критерием Колмогорова (см. зад. 5, с. 63). Идея критерия заключается в сравнении ЭФР F_n и ЭФР G_m. Критерий Смирнова отклоняет H_0, если F_n и G_m слишком сильно расходятся. Реализуется эта идея в терминах статистики

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|.$$

Известно, что в случае верности H_0, $\forall t > 0$

$$\lim_{n,m\to\infty} \mathbf{P}\left(\sqrt{\frac{nm}{n+m}} D_{n,m} < t\right) = \mathrm{Fkolm}(t),$$

т.е. $\sqrt{nm/(n+m)}D_{n,m}$ асимптотически имеет распределение Колмогорова $\mathrm{Fkolm}(t)$.

Критерий Смирнова с уровнем значимости α отклоняет H_0, если

$$\sqrt{\frac{nm}{n+m}} D_{n,m} > \mathrm{Fkolm}^{-1}(1-\alpha),$$

где $\mathrm{Fkolm}^{-1}(1-\alpha)$ — $(1-\alpha)$-квантиль распределения Колмогорова.

Замечание. В большинстве пакетов вычислительной статистики нет методов для вычисления функции распределения Колмогорова и её квантилей.

Для вычисления значения функции распределения при больших значениях t можно воспользоваться аппроксимацией

$$\mathrm{Fkolm}(t) \approx 1 - 2\mathrm{e}^{-2t^2}.$$

Для вычисления квантили при малых α можно воспользоваться аппроксимацией:

$$\mathrm{Fkolm}^{-1}(1-\alpha) \approx \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}.$$

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Можно ли применять критерий Смирнова для сравнения двух нормальных выборок с одинаковой дисперсией?

3) Можно ли применять критерий Смирнова для проверки гипотезы об отсутствии эффекта лечения на выборках $X^{(n)}$ — артериальное давление пациентов до лечения, и $Y^{(n)}$ — артериальное давление этих же пациентов после лечения?

4) Как вычислить p-значение критерия Смирнова?

5) Как вычислить $\sup_x |F_n(x) - G_m(x)|$? Достигается ли в каких-либо точках этот супремум?

6) Как можно аппроксимировать функцию распределения и квантиль распределения Колмогорова?

Часть V
Точечное оценивание

Теоретические основы:

- характеристики статистических оценок смотрите в п. 4, с. 26

Задание 13. Метод моментов

1. Постановка задачи

Пусть наблюдается выборка $X^{(n)} = (X_1, \ldots, X_n)$ из распределения $\mathbf{P}_\theta$, $\theta \in \Theta$. Известна функция плотности наблюдения $f(x \,|\, \theta)$. Требуется построить оценку по методу моментов для параметра θ.

2. Теоретические основы

Идея метода моментов заключается в достаточно простом факте: если для некоторой функции $g(x)$ существует математическое ожидание $\mathbf{E}_\theta\, g(X_1) = a(\theta)$, то из закона больших чисел следует, что эмпирические (выборочные) моменты сходятся к теоретическим:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathbf{P}_\theta} a(\theta). \tag{8}$$

Предположим для простоты, что параметр является скаляром: $\Theta \subset \mathbb{R}$. Приравняв левую и правую часть (8) и решив уравнение относительно θ, можно надеяться, что полученная таким образом оценка будет обладать некоторыми оптимальными свойствами. В частности, при непрерывной и монотонной функции $a(\theta)$ состоятельность такой оценки следует из теоремы о непрерывном отображении.

Вообще говоря, параметрическое пространство может быть многомерным: $\Theta \subset \mathbb{R}^M$, $M \geq 1$. В методе моментов мы должны задать набор функций $g_1, \ldots, g_M$ таких, что при всех $\theta \in \Theta$ существуют математические ожидания $\mathbf{E}_\theta\, g_m(X_1) = a_m(\theta)$, $m = 1, \ldots, M$. Из этих функций затем составляется система уравнений

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} g_1(X_i) = a_1(\widehat{\theta}), \\ \qquad \ldots \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} g_M(X_i) = a_M(\widehat{\theta}). \end{cases} \tag{9}$$

Значение $\widehat{\theta}$, которое удовлетворяет этой системе и является оценкой по методу моментов.

Следует выбирать такие функции $g_1, \ldots, g_m$, что решение системы (9) существует и единственно. Обычно полагают $g_m(x) = x^m$, и, таким образом, $a_m(\theta)$ является моментом порядка m. Отсюда и название: «метод моментов». Другим примером выбора g может являться индикаторная функция $g(x) = I(x \in A)$ для некоторого фиксированного события A. В таком случае $a(\theta) = \mathbf{E}_\theta\, I(X_1 \in A) = \mathbf{P}_\theta(X_1 \in A)$.

В левой части выражения (8), и соответственно, в самом методе моментов, возможно также использование любых выборочных характеристик, имеющих в качестве предела некую величину, зависящую только от параметра θ. Примером могут служить выборочная дисперсия или выборочная медиана.

Заметим, что от выбора функций g_m существенно зависит качество оценки по методу моментов. Например, для распределения Пуассона $X \sim P(\lambda)$ и математическое ожидание, и дисперсия равны λ. Поэтому для оценки параметра λ можно использовать выборочное среднее или выборочную дисперсию. Однако, можно убедиться, что выборочное среднее имеет значительно меньший среднеквадратичный риск, нежели выборочная дисперсия при любом значении параметра λ.

Наконец, следует иметь ввиду, что решение системы уравнений (9) может не принадлежать Θ. Например, для биномиального распределения $\mathrm{Fbin}(p, k)$ с неизвестными параметрами p, k оценка для p, построенная с помощью среднего и дисперсии, может выходить за рамки отрезка $[0, 1]$, а оценка для k принимать нецелые значения. При этом нужно с осторожностью подходить к поправкам таких значений, поскольку это может привести к трудно предсказуемым последствиям.

3. Пример

Рассмотрим метод моментов на примере нормального распределения. Пусть наблюдения происходят из нормального распределения с неизвестными средним и дисперсией. Тогда параметр является двумерным: $\theta = (\mu, \sigma^2)$, и параметрическое пространство $\Theta = (-\infty, \infty) \times (0, \infty)$.

Положим $g_1(x) = x, g_2(x) = x^2$. Тогда, как известно,

$$a_1(\theta) = \mathbf{E}_\theta\, g_1(X_1) = \mu,$$

$$a_2(\theta) = \mathbf{E}_\theta\, g_2(X_1) = \sigma^2 + \mu^2.$$

Приравняем теоретические моменты выборочным:

$$a_1(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad a_2(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Отсюда имеем

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^2 = S^2.$$

4. Вопросы и задания для самоконтроля

1) Где в методе моментов используется функция плотности $f(x \,|\, \theta)$, указанная в задании?

2) Можно ли в методе моментов использовать несмещённую оценку дисперсии вместо обычной выборочной дисперсии?

3) Найдите оценку по методу моментов для нормального распределения с неизвестными средним и дисперсией в случае, если выбраны индикаторные функции $g_1(x) = \mathbb{I}(x < 1)$, $g_2(x) = \mathbb{I}(x < -1)$. Как можно сравнить эти оценки с $\widehat{\mu}$ и $\widehat{\sigma}^2$, полученными в примере?

4) Найдите оценку по методу моментов для биномиального распределения $\mathrm{Fbin}(p, k)$ с неизвестными параметрами p и k на основе среднего и дисперсии. Убедитесь, что оценка для p может выходить за рамки отрезка $[0, 1]$, а оценка для k принимать нецелые значения.

Задание 14. Метод максимального правдоподобия

1. Постановка задачи

Пусть наблюдается выборка $X^{(n)} = (X_1, \ldots, X_n)$ из распределения $\mathbf{P}_\theta$, $\theta \in \Theta$. Известна функция плотности наблюдения $f(x \,|\, \theta)$. Требуется построить оценку по методу максимального правдоподобия для параметра θ.

2. Теоретические основы

Оценкой максимального правдоподобия называется точка достижения максимума у функции плотности случайной выборки:

$$\widehat{\theta}(X^{(n)}) = \arg\max_{\theta \in \Theta} f(X_1, \ldots, X_n \,|\, \theta). \tag{10}$$

Выпишем функцию плотности выборки. Так как из определения выборки наблюдения независимы и одинаково распределены, совместная

функция плотности наблюдений распадается на произведение маргинальных плотностей:

$$f(X_1, \ldots, X_n \,|\, \theta) = \prod_{i=1}^{n} f(X_i \,|\, \theta). \tag{11}$$

Функция (11) называется функцией правдоподобия и обозначается $L(\theta \,|\, X^{(n)})$. Отметим, что функция плотности тогда «становится» функцией правдоподобия, когда интерпретируется как случайная величина (статистика), то есть когда аргументы совместной функции плотности есть наблюдения, полученные в эксперименте.

Обычно для максимизации удобнее работать с логарифмом правдоподобия. Эта функция разбивается в сумму отдельных компонент:

$$\ln L(\theta \,|\, X^{(n)}) = \sum_{i=1}^{n} \ln f(X_i \,|\, \theta).$$

Так как преобразование логарифма монотонно, эта функция имеет тот же максимум, что и (11). Если функция $\ln f(x \,|\, \theta)$ непрерывно дифференцируема во внутренних точках $\Theta \subset \mathbb{R}^M$, то максимизацию можно совершать, приравнивая градиент логарифма правдоподобия относительно $\theta = (\theta_1, \ldots, \theta_M)$ к нулю:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \ln f(X_i \,|\, \theta) = 0, \quad k = 1, \ldots, M.$$

Решение этой системы есть точка экстремума функции правдоподобия, и необходимо проверить этот экстремум на локальный максимум. Нельзя также забывать, что кроме нулей градиента, максимум может достигаться на границе параметрического пространства Θ.

Оценка максимального правдоподобия по построению обязана принадлежать Θ.

При выполнении достаточно слабых условий оценка максимального правдоподобия обладает некотрыми оптимальными свойствами. В частности, она состоятельна и асимптотически нормальна. Это одна из причин почему эта оценка очень популярна на практике.

В некоторых случаях оценка максимального правдоподобия не существует. Например, если наблюдается одно наблюдение из нормального распределения с неизвестными средним и дисперсией.

3. Пример

Пусть наблюдения выборки имеют нормальное распределение с неизвестными средним и дисперсией. Как и в примере для метода моментов, $\theta = (\mu, \sigma^2)$, и параметрическое пространство $\Theta = (-\infty, \infty) \times (0, \infty)$.

Функция плотности наблюдения равна

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Логарифм функции правдоподобия, таким образом, равен

$$\ln L(\theta \mid X^{(n)}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}.$$

Эта функция непрерывно дифференцируема по θ в области Θ. Вычислим частные производные этой функции и приравняем их нулю:

$$\frac{\partial \ln L(\theta \mid X^{(n)})}{\partial \mu} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma^2} = 0$$

$$\frac{\partial \ln L(\theta \mid X^{(n)})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^4} = 0.$$

Решением этой системы является, как нетрудно убедиться, оценка

$$\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2), \quad \text{где} \quad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = S_n^2.$$

Оценка $\widehat{\theta} \in \Theta$ при любых значениях $X^{(n)}$. То, что это действительно точка максимума, можно проверить по матрице вторых производных. В точке максимума правдоподобия она равна

$$\begin{pmatrix} -n\widehat{\sigma}^{-2} & 0 \\ 0 & -n(2\widehat{\sigma}^4)^{-1} \end{pmatrix}$$

Такая матрица, очевидно, отрицательно определена.

Другие примеры нахождения оценок максимального правдоподобия можно посмотреть в учебнике [2].

4. Вопросы и задания для самоконтроля

1) Как изменится метод максимального правдоподобия, когда наблюдения являются дискретными случайными величинами?

2) Как искать оценку максимального правдоподобия для конечного множества Θ? Для примера можно рассмотреть биномиальное распределение $\mathrm{Fbin}(p, m)$ с неизвестным параметром m.

3) Убедитесь в том, что если выборка состоит из единственного наблюдения с нормальным распределением с неизвестными средним и дисперсией, то оценки максимального правдоподобия не существует.

4) Является ли состоятельной оценка максимального правдоподобия из Примера? Будет ли эта оценка несмещённой?

5) Предположим, что в условиях Примера стало известно, что среднее значение μ положительно. Как от этого изменится оценка максимального правдоподобия?

Часть VI
Интервальные оценки

Теоретические основы:

- интервальные оценки смотрите в п. 5, с. 31

Задание 15. Интервальная оценка для неизвестного математического ожидания

1. Постановка задачи

В эксперименте наблюдается случайная величина X. Предполагая, что распределение X нормально, необходимо построить доверительный интервал (верхнюю или нижнюю границу) для неизвестного математического ожидания $\mu = \mathbf{E}\,X$.

2. Теоретические основы

Пусть $\overline{X}$ — выборочное среднее, S^2 — выборочная дисперсия, вычисленные по выборке из нормального распределения со средним μ и дисперсией σ^2. Функция

$$G = \frac{\overline{X} - \mu}{S}\sqrt{n-1}$$

монотонно убывает по μ и её распределение не зависит от параметров μ и σ (докажите это!), т.е. G есть опорная функция относительно μ. Известно, что G имеет распределение Стьюдента $\mathrm{Fstud}(t \,|\, n-1)$ с $(n-1)$-ой степенью свободы. Пусть $t^{(\alpha)} = t^{(\alpha)}(n-1) = \mathrm{Fstud}^{-1}(1-\alpha \,|\, n-1)$ — $(1-\alpha)$-квантиль распределения $\mathrm{Fstud}(t \,|\, n-1)$, тогда с надёжностью $(1-\alpha)\cdot 100\%$:

1) $\underline{\mu} = \overline{X} - \dfrac{S}{\sqrt{n-1}}t^{(\alpha)}$ — нижняя доверительная граница;

2) $\overline{\mu} = \overline{X} + \dfrac{S}{\sqrt{n-1}}t^{(\alpha)}$ — верхняя доверительная граница.

Как видно из этих формул, ширина двустороннего доверительного интервала пропорциональна отношению $S/\sqrt{n-1}$. Это отношение называется стандартной ошибкой среднего, обозначается обычно буквой m и

весьма оригинально читается — «эм малое». В отчёт о проведённых исследованиях результат часто записывают в виде 4.366 ± 0.107, где первое слагаемое есть среднее арифметическое $\overline{X}$, а второе слагаемое — ошибка среднего m.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Дайте интерпретацию определения нижней (верхней) доверительной границы.

3) Какую надёжность будет иметь двухсторонний доверительный интервал, если он построен на основе 90%-ой нижней границы и 95%-ой верхней границы?

4) Можно ли с помощью двухстороннего доверительного интервала проверить гипотезу о том, что истинное значение оцениваемого параметра будет больше 18?

5) Приведите формулы доверительных границ (доверительного интервала) для среднего значения нормального распределения.

6) Что такое стандартная ошибка среднего?

7) Можно ли, основываясь на записи вида , построить доверительный интервал для среднего значения?

Задание 16. Интервальная оценка для неизвестной дисперсии нормального распределения

1. Постановка задачи

В эксперименте наблюдается случайная величина X. Предполагая, что распределение X нормально, необходимо построить доверительный интервал (верхнюю или нижнюю границу) для неизвестной дисперсии $\sigma^2 = \mathbf{D}X$ и стандартного отклонения σ.

2. Теоретические основы

Пусть S^2 — выборочная дисперсия, построенная по выборке из нормального распределения с неизвестной дисперсией σ^2, тогда функция

$$G = \frac{nS^2}{\sigma^2}$$

есть опорная функция относительно σ^2. По знаменитой Лемме Фишера её распределение совпадает с распределением хи-квадрат $\mathrm{Fchisq}(x \,|\, n - 1)$ с $(n-1)$ степенью свободы. Таким образом,

1) $\underline{\sigma}^2 = \dfrac{n}{\chi^{(\alpha)}} S^2$ – нижняя $(1-\alpha)$-доверительная граница для σ^2;

2) $\overline{\sigma}^2 = \dfrac{n}{\chi^{(1-\alpha)}} S^2$ – верхняя $(1-\alpha)$-доверительная граница для σ^2;

где $\chi^{(p)} = \mathrm{Fchisq}^{-1}(1-p \,|\, n-1)$ — квантиль порядка $(1-p)$ распределения хи-квадрат. Например, при 19 степенях свободы $\chi^{(0.025)} = 32.852$, $\chi^{(0.975)} = 8.907$ и, следовательно, 95%-й доверительный интервал можно представить как $[0.578 \cdot S^2, 2.133 \cdot S^2]$ (при $n = 100$ интервал сужается до $[0.771 \cdot S^2, 1.349 \cdot S^2]$).

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Приведите формулы доверительных границ (доверительного интервала) для дисперсии нормального распределения.

3) Приведите формулы доверительных границ (доверительного интервала) для стандартного отклонения нормального распределения. Покажите, что интервал, представленный этой формулой, является доверительным с заданным уровнем надёжности.

4) Найдите по таблице 5%-ю и 90%-ю точки хи-квадрат распределения для объема выборки $n = 42$.

5) Проверьте гипотезу $\mathrm{H}_0\colon \sigma^2 = 0.55$ на уровне 2.5% о значении истинной дисперсии при альтернативе $\mathrm{H}_1\colon \sigma^2 < 0.55$, если выборочная дисперсия по 20 наблюдениям оказалась равной 0.5.

6) Как построить двусторонний доверительный интервал с надёжностью $1-\alpha$?

Задание 17. Интервальная оценка для вероятности успеха

1. Постановка задачи

В эксперименте подсчитывается число успешных реализаций некоторого события (например, число доброкачественных изделий). Требуется построить доверительный интервал (верхнюю или нижнюю границу) для истинной вероятности p этого события.

2. Теоретические основы

Начнём с наиболее простого асимптотического метода построения границ (метод 3). Обозначим через T случайную величину, равную числу успехов в n независимых наблюдениях. По известной теореме Муавра–Лапласа статистика T в пределе при $n \to \infty$ имеет нормальное распределение со средним np и дисперсией $np(1-p)$, т.е.

$$\mathbf{P}_p\left(\frac{T-np}{\sqrt{np(1-p)}} < x\right) \to \Phi(x). \tag{12}$$

Кроме того, легко видеть, что функция $H(p) = (T-np)/\sqrt{np(1-p)}$ убывает по параметру $p \in (0,1)$. Таким образом, в асимптотическом смысле H есть опорная функция относительного p.

Можно было бы теперь, положив $t^{(\alpha)} = \Phi^{-1}(1-\alpha)$ и решив уравнение $H(p) = t^{(\alpha)}$ по параметру p, найти нижнюю границу для p. Однако, гораздо проще приближённые доверительные границы для вероятности события находятся, если воспользоваться тем, что соотношение (12) имеем место и после замены неизвестной вероятности p в знаменателе на состоятельную оценку $\widetilde{p} = T/n$. В этом случае вид доверительных границ для p

$$\widetilde{p} \pm m\, t^{(\alpha)}, \quad m = \frac{\sqrt{\widetilde{p}(1-\widetilde{p})}}{\sqrt{n}},$$

совпадает с видом доверительных границ для математического ожидания нормального закона. По аналогии с нормальным законом здесь также статистику m называют стандартной ошибкой среднего.

Надёжность приближённых границ при малом объёме выборки вызывает некоторое сомнение. К счастью, в данном случае можно применить точный метод 2, который даёт доверительные границы с коэффициентом доверия, чуть большим (ввиду дискретности наблюдений) ожидаемого уровня $(1-\alpha)$.

Пусть T — число успешных реализаций исследуемого события в n экспериментах. Известно, что T распределено по биномиальному закону с параметрами (n,p): $\mathbf{P}_p(T \leqslant t) = \mathrm{Fbin}(t \,|\, n,p)$, где p — вероятность

этого события при однократном наблюдении. Функция распределения $\mathrm{Fbin}(t\,|\,n,p)$ убывает с ростом p. Поэтому, если t — полученное в эксперименте значение статистики T, то

1) нижняя $(1-\alpha)$-доверительная граница $\underline{p}$ для вероятности успеха p есть решение уравнения $\mathrm{Fbin}(t-1\,|\,n,\underline{p}) = 1-\alpha$ (если $t=0$ полагаем $\underline{p}=0$);

2) верхняя $(1-\alpha)$-доверительная граница $\overline{p}$ для вероятности успеха p есть решение уравнения $\mathrm{Fbin}(t\,|\,n,\overline{p}) = \alpha$ (если $t=1$ полагаем $\overline{p}=1$).

Доказательство корректности метода 2. Пусть $\underline{\theta}\ (=\underline{\theta}(t))$ — единственное решение уравнения $\mathbf{P}_{\underline{\theta}}(T<t) = 1-\alpha$, тогда при любом $\theta > \underline{\theta}$ в силу строгой монотонности функции распределения $\mathbf{P}_{\underline{\theta}}(T<t) < 1-\alpha$.

Определим также константу (квантиль порядка $(1-\alpha)$)

$$\widetilde{t}(\theta) = \sup\big\{t : \mathbf{P}_\theta(T<t) \leqslant 1-\alpha\big\},$$

В силу непрерывности слева функции распределения в точке $\widetilde{t}$ также выполняется неравенство $\mathbf{P}_\theta\big(T<\widetilde{t}(\theta)\big) \leqslant 1-\alpha$. С другой стороны, т.к. при любых $t>\widetilde{t}$ имеем $\mathbf{P}_\theta\big(T\leqslant t\big) \geqslant \mathbf{P}_\theta\big(T<t\big) > 1-\alpha$, то для непрерывной справа функции $\mathbf{P}_\theta\big(T\leqslant t\big)$ справедливо обратное неравенство

$$\mathbf{P}_\theta\big(\,T\leqslant\widetilde{t}(\theta)\,\big) \geqslant 1-\alpha. \tag{13}$$

Таким образом, справедлива следующая цепочка эквивалентностей:

$$\Big[\ \underline{\theta}(t)\leqslant\theta\ \Big] \iff \Big[\ \mathbf{P}_\theta\big(T<t\big)\leqslant 1-\alpha\ \Big] \iff \Big[\ t\leqslant\widetilde{t}(\theta)\ \Big].$$

Соотношение (13) гарантирует теперь, что

$$\mathbf{P}_\theta\big(\,\underline{\theta}(T)\leqslant\theta\,\big) = \mathbf{P}_\theta\big(T\leqslant\widetilde{t}(\theta)\big) \geqslant 1-\alpha,$$

что и требовалось доказать. Аналогично доказывается утверждение для верхней границы.

Заметим, что при дискретном распределении статистики T неравенство (13) может быть строгим, т.е. построенные таким образом границы имеют надёжность, превышающую номинальный уровень $(1-\alpha)$.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Приведите формулы асимптотических границ для вероятности успеха.

3) Как можно построить точную границу для вероятности успеха?

4) Построив предварительно соответствующую доверительную границу, проверьте гипотезу о том, что вероятность рождения девочки меньше 0.5, если в 50 случаях наблюдалось 28 рождений мальчиков.

5) Постройте точные и приближённые доверительные границы интенсивности λ по n реализациям пуассоновской случайной величины.

Часть VII
Исследование зависимости между двумя характеристиками

Очень часто в эксперименте наблюдается не одна, а несколько характеристик одного и того же объекта (например, рост и вес человека, урожайность зерна и количество внесенных удобрений и т.п.). Предполагается, что значения характеристик изменяются от объекта к объекту случайным образом. Требуется

1) выяснить, имеется ли зависимость между исследуемыми характеристиками;

2) построить уравнение наилучшего прогноза одной характеристики по значениям другой.

Теоретические основы

- критерии проверки гипотез смотрите в п. 3, с. 15
- коэффициент корреляции и построение линейной регрессии смотрите в п. 6, с. 35

Задание 18. Проверить независимость двух характеристик по критерию сопряженности хи-квадрат

1. Постановка задачи

По выборке $(X_1, Y_1), \ldots, (X_n, Y_n)$ из двумерного распределения (не обязательно нормального) проверить гипотезу независимости компонент наблюдаемого случайного вектора (X, Y).

2. Теоретические основы

При отсутствии нормальности распределения вектора (X, Y) для проверки независимости его компонентов применяется критерий сопряженности хи-квадрат. Для построения этого критерия необходимо

1) область значений признака X разбить на r интервалов $A_1^x, A_2^x, \ldots, A_r^x$, а область значений признака Y на s интервалов $B_1^y, B_2^y, \ldots, B_s^y$.

2) Для каждого сочетания (i, j) подсчитать количество n_{ij} выборочных данных, для которых, одновременно, признак X попадает в i-ый интервал A_i^x, а признак Y — в m-ый интервал B_m^y. Результаты подсчёта свести в таблицу сопряженности признаков.

X \ Y	1-й	...	s-й	Всего
1-й	n_{11}	...	n_{1s}	$n_{1\bullet}$
...	...	...	...	...
r-й	n_{r1}	...	n_{rs}	$n_{r\bullet}$
Всего	$n_{\bullet 1}$	...	$n_{\bullet s}$	$n_{\bullet\bullet} = n$

где, как обычно, точка $\bullet$ на месте одного из индексов означает сумму всех чисел по этому индексу с фиксированным значением второго индекса. Проще говоря, нужно просуммировать значения по всем столбцам и строкам таблицы (столбец и строка «Всего»). Число в правой крайней нижней ячейке должно равняться общему объему выборки n.

1) Вычислить статистику критерия сопряженности хи-квадрат

$$T = \sum_{i=1}^{r} \sum_{m=1}^{s} \frac{(n \cdot n_{im} - n_{i\bullet} n_{\bullet m})^2}{n \cdot n_{i\bullet} n_{\bullet m}}.$$

При справедливости гипотезы независимости распределение статистики T может быть аппроксимировано распределением хи-квадрат $\mathrm{Fchisq}(x \,|\, \nu)$ с $\nu = (r-1)(s-1)$ степенями свободы:

$$\mathbf{P}(T < t) \approx \mathrm{Fchisq}(t \,|\, \nu) \quad (n \to \infty).$$

Гипотеза независимости отвергается при больших значениях статистики T. p-значение критерия приближённо вычисляется по формуле

$$\mathsf{p} \approx 1 - \mathrm{Fchisq}(T \,|\, \nu).$$

2) Признаки следует признать независимыми, если $\mathsf{p} > \alpha$.

Идея критерия сопряженности основана на том, что по закону больших чисел относительная частота n_{im}/n — есть состоятельная оценка вероятности $\mathbf{P}(X \in A_i^x, Y \in B_m^y)$, а частоты $n_{i\bullet}/n$, $n_{\bullet m}/n$ — состоятельные оценки вероятностей $\mathbf{P}(X \in A_i^x)$, $\mathbf{P}(Y \in B_m^y)$.

Поэтому можно ожидать, что для независимых признаков

$$\frac{n_{im}}{n} \approx \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet m}}{n},$$

и поэтому значение статистики T будет «не слишком» большим.

Замечание. Критерий «безразличен» к способу получения таблицы сопряженности. Очень часто данные сразу имеют вид такой таблицы. Например, критерием можно решать задачу проверки гипотезы независимости уровня образования от количества детей в семье. Данные получены путем обследования некоторой совокупности семей, сгруппированной по двум признакам: уровень образования (две градации, $r = 2$) и число детей (четыре градации, $s = 4$).

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Что такое независимость случайных величин?

3) Выпишите формулу для вычисления статистики критерия сопряженности хи-квадрат.

4) Почему эта статистика может служить мерой близости данных к гипотезе независимости?

5) Чему равен критический уровень значимости критерия сопряженности признаков?

6) Каким еще критерием (и в каком случае) можно проверить гипотезу независимости двух наблюдаемых характеристик?

Задание 19. Проверка независимости двух нормальных выборок. Линейная регрессия

1. Постановка задачи

По выборке $(X_1, Y_1), \ldots, (X_n, Y_n)$ из двумерного нормального распределения проверить гипотезу независимости компонентов наблюдаемого случайного вектора (X, Y). Построить линии регрессии одного из признаков по другому признаку. Найти наилучший прогноз признака Y при фиксированном значении признака $X = 120$.

2. Теоретические основы

Если вектор (X, Y) имеет нормальное распределение, то независимость его компонентов эквивалентна равенству нулю коэффициента корреляции. Поэтому для проверки гипотезы независимости можно проверить гипотезу $\mathrm{H}_0\colon \rho = 0$ о коэффициенте корреляции ρ.

Преобразование Стьюдента для выборочной корреляции r имеет вид

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}.$$

При выборе из двумерного нормального распределения и в условиях гипотезы H_0 статистика T имеет распределение Стьюдента $\mathrm{Fstud}(t \,|\, n-2)$ с $(n-2)$-мя степенями свободы. Таким образом, если нулевую гипотезу отвергать при значениях статистики Стьюдента, в ту или иную сторону (в зависимости от альтернативы) отличающихся от нуля, то критический уровень значимости может быть найден по следующей схеме:

Альтернатива	p-значение	пояснение
$\mathrm{H}_1\colon \rho \neq 0$	$2(1 - \mathrm{Fstud}(t \,|\, n-2))$	$\mathbf{P}\{\lvert T\rvert > t\}$
$\mathrm{H}_1\colon \rho < 0$	$1 - \mathrm{Fstud}(t \,|\, n-2)$	$\mathbf{P}\{T > t\}$
$\mathrm{H}_1\colon \rho > 0$	$\mathrm{Fstud}(t \,|\, n-2)$	$\mathbf{P}\{T < t\}$.

Замечание 2. Здесь следует подчеркнуть различие между статистической и практической значимостью коэффициента корреляции. Статистическая значимость коэффициента корреляции означает лишь, что наших данных достаточно для подтверждения зависимости между исследуемыми признаками. Практическая значимость при этом будет означать, что эти признаки могут быть достаточно точно спрогнозированы один по другому. Таким образом, если для практической значимости необходимо, чтобы истинный коэффициент корреляции был очень большим ($\pm\, 0.7$ и выше), то для статистической значимости может оказаться достаточным проведение большого числа наблюдений при очень маленьком коэффициенте корреляции.

Проиллюстрируем эти положения. По некоторым данным получено значение $r = -0.5$ при объеме выборки $n = 101$. Таким образом, критический уровень значимости $\mathsf{p} < 0.001$, что свидетельствует об очень высокой статистической значимости, однако в этом случае только 25% изменчивости каждого из признаков (см. свойство 5) коэффициента корреляции) можно объяснить влиянием на него другого признака. Другими словами, хотя зависимость между признаками и есть, однако она имеет низкую практическую значимость.

Графически это можно проиллюстрировать следующим образом. Проведём на графике линий регрессии вертикальную линию из точки $x = 120$ (см. рис. 6). Точка пересечения этой линии с прямой регрессии Y на X даст наилучший прогноз значения признака Y при значении признака $X = 120$: $Y = -0.3 \cdot 120 + 170.817 = 134.8$.

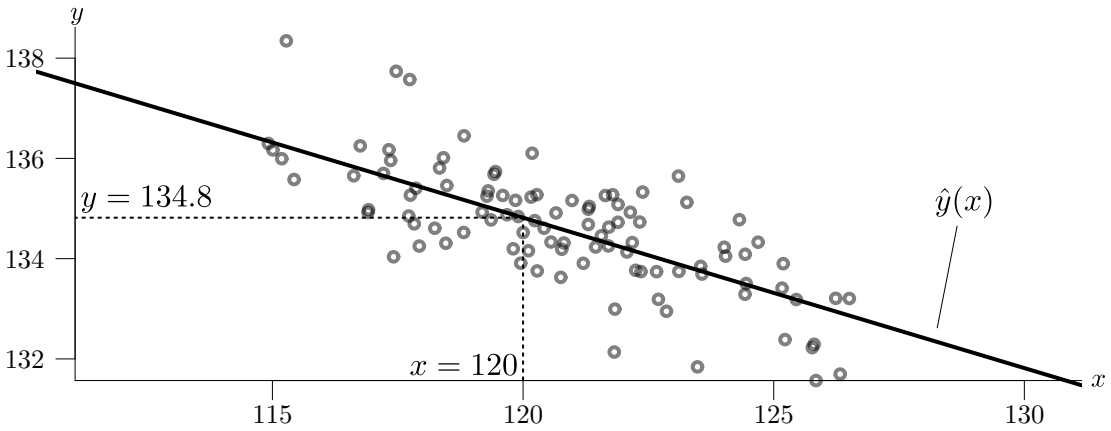

Рис. 6: Прогноз значения Y при $X = 120$ по графику линейной регрессии

Реальные значения, близкие к вертикальной линии $x = 120$, имеют относительно небольшой разброс по вертикали в сравнении с полным размахом данных вдоль оси y, что говорит о хорошем качестве прогноза.

3. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Что такое коэффициент корреляции и как его интерпретировать?

3) Выпишите формулу для выборочного коэффициента корреляции.

4) Не вычисляя, скажите, чему равен выборочный коэффициент корреляции для следующих данных: $(1, 2), (2, 4), (3, 6), (7, 14)$.

5) Является ли выборочный коэффициент корреляции несмещённой оценкой для истинного коэффициента корреляции? Состоятельной?

6) Как изменится значение коэффициента корреляции между ростом и весом человека, если значение веса сначала измерять сначала в килограммах, а потом в граммах?

7) Почему и когда надо проверять гипотезу независимости, основываясь на коэффициенте корреляции?

8) Как преобразование Стьюдента для выборочного коэффициента корреляции свидетельствует о справедливости гипотезы?

9) Что такое линейная регрессия?

10) Выпишите уравнение линейной регрессии Y на X. Можно ли по этому уравнению вычислить приближённое значение X, если задано значение признака Y?

11) В каком случае обе регрессионные линии (Y на X и X на Y) совпадут?

12) Как будут располагаться линии регрессии, если коэффициент корреляции близок к 0?

13) Как будут располагаться линии регрессии, если коэффициент корреляции близок к 1?

14) Известно, что высота h, с которой падает предмет, и время его падения t удовлетворяет соотношению $h = gt^2/2$, где g — ускорение свободного падения. Как с помощью методов регрессионного анализа оценить величину g по ряду связанных замеров h и t?

Задание 20. Ядерная оценка регрессии

1. Постановка задачи

Пусть наблюдается выборка $(X_1, Y_1), \ldots, (X_n, Y_n)$ из независимых копий случайного вектора (X, Y). Построить оценку регрессии одного из признаков по другому признаку с помощью ядерной оценки. Найти прогноз признака Y при фиксированном значении признака $X = 120$.

2. Теоретические основы

В случае, когда нельзя предполагать нормальность наблюдаемых данных, использование линейной регрессии как оценки условного математического ожидания $h(x) = \mathbf{E}(Y \mid X = x)$ не оправдано. Это не мешает, однако, её повсеместному использованию в приложениях вследствие простоты и статистической стабильности получаемых результатов.

В случаях, когда число наблюдений n довольно велико, возможно использование довольно простого метода оценки условного среднего в точке x. Если мы предполагаем о некоторой непрерывности функции $h(x)$, то точки x', близкие к x должны иметь довольно малое отклонение $|h(x') - h(x)|$. Тогда близкие к x наблюдения должны иметь некий вклад в оценку регрессии. Для оценки $h(x)$ возьмём все наблюдения X_i,

которые находятся на расстоянии не больше λ от x, и вычислим среднее значение соответствующих значений Y_i. Вполне естественно, что более далёкие от x наблюдения должны иметь меньше влияния, чем более близкие. Это учитывается путём присвоения весов наблюдений, которые зависят от расстояния до x.

Рассмотрим данный процесс подробнее.

Сначала вводится некая функция $K_\lambda(t) > 0$, которую обычно называют *ядерной функцией* (*ядром*). Эта функция представляет собой величину влияния (вес) наблюдения, которое находится на расстоянии t от точки x. Оценка функции регрессии в точке x вычисляется следующим образом:

$$\widehat{h}(x) = \left(\sum_{i=1}^{n} K_\lambda(x - X_i) \right)^{-1} \sum_{i=1}^{n} K_\lambda(x - X_i) Y_i. \tag{14}$$

Оценка (14) состоятельна [9] при существовании второго момента у случайной величины Y в любой точке непрерывности функции плотности случайной величины X и некоторых ограничениях на ядерную функцию $K_\lambda(t)$. Для состоятельности также необходимо, чтобы параметр λ зависел от n, и $\lambda(n) \to 0, n(\lambda(n))^2 \to \infty$.

Самый простой пример выбора ядерной функции — это равномерное ядро $K_\lambda(t) = \mathbb{I}_{[-\lambda;\lambda]}(t)$. В таком случае $\widehat{h}(x)$ есть упомянутое выше среднее значение достаточно близких наблюдений. Другие примеры, распределяющие веса в зависимости от расстояния до x, приведены ниже (треугольная и квадратичная ядерные функции):

$$K_\lambda(t) = \begin{cases} t/\lambda + 1, & -\lambda \le t < 0, \\ 1 - t/\lambda, & 0 \le t \le \lambda, \\ 0, & \text{иначе} \end{cases} \qquad K_\lambda(t) = \begin{cases} \frac{3}{4}\left(1 - (t/\lambda)^2\right), & -\lambda \le t \le \lambda, \\ 0, & \text{иначе.} \end{cases}$$

Во всех случаях параметр λ контролирует ширину окна. Графики трёх ядерных функций приведены на рисунке 7.

При выборе ядерной функции и величины параметра λ, как правило, исходят из методов обычной или перекрёстной проверки. Описание этих методов мы здесь опустим, однако заметим, что выбор большого значения λ стабилизирует оценку (уменьшает дисперсию), а выбор малого λ уменьшает смещение (приближает $\mathbf{E}\{\widehat{h}(x) \,|\, X = x\}$ к $h(x)$). Пример применения различных ядерных функций показан на рисунке 8.

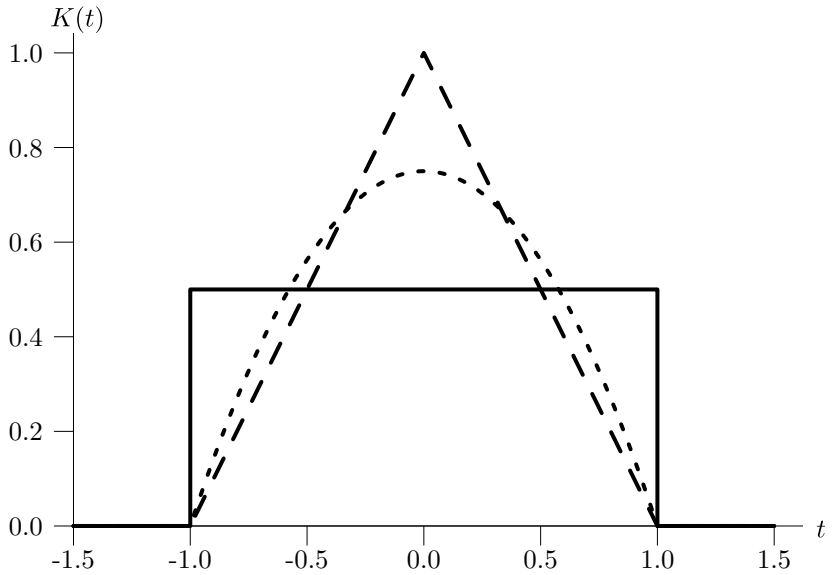

Рис. 7: Ядерные функции при $\lambda = 1$: равномерное (сплошная линия), треугольное (штрихованная линия), квадратичное (точечная линия)

Оценка функции плотности. В качестве дополнительного замечания отметим, что таким же нехитрым образом можно оценивать функцию плотности $f(x)$ случайной величины X. Ядерной оценкой функции плотности называют следующую величину:

$$\widehat{f}(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K_\lambda(x - X_i).$$

Иногда её ещё называют Парзеновской оценкой или методом Парзеновского окна. В данном случае предполагается, что $\int_{\mathbb{R}} K_\lambda(t)\,\mathrm{d}t = \lambda$ (все приведённые выше ядерные функции обладают этим свойством). Гистограмма тоже в некотором смысле является ядерной оценкой функции плотности, у которой ядро $K_\lambda(t, x)$ зависит также от точки x. Сравнение гистограммы и ядерных оценок приведены на рисунке 9.

3. Рекомендации к программной реализации

В системах, поддерживающих написание циклов, вычисление ядерных оценок не должно вызвать затруднений. Таковыми являются R, Wolfram Mathematica и встроенный в Excel Visual Basic. В нативном Excel для этого придётся завести столько столбцов, сколько наблюдений содержится в выборке, и в каждом столбце вычислять $K_\lambda(x - X_i)$ для каждого интересующего нас значения x (предварительно расположив выборку в строку).

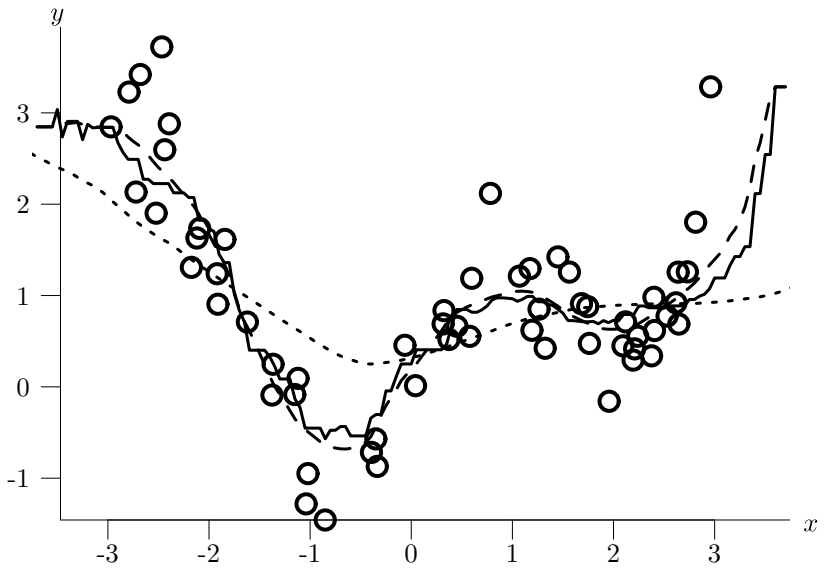

Рис. 8: Ядерная оценка регрессии с различными ядерными функциями: равномерное ядро с $\lambda = 1$ (сплошная линия), квадратичное ядро с $\lambda = 1$ (штрихованная линия), квадратичное ядро с $\lambda = 2$ (точечная линия)

4. Вопросы и задания для самоконтроля

1) Сформулируйте статистическую задачу.

2) Почему мы хотим вычислить условное математическое ожидание?

3) Почему точки x', близкие к x должны иметь малое значение отклонения $|h(x') - h(x)|$.

4) Как будет выглядеть ядерная оценка, если устремить λ к нулю? Рассмотрите задачи построения регрессии и функции плотности.

5) Как будет выглядеть ядерная оценка, если устремить λ к бесконечности? Рассмотрите задачи построения регрессии и функции плотности.

6) Попытайтесь объяснить, почему большое значение λ будет уменьшать дисперсию оценки $\widehat{h}(x)$, а малое значение λ уменьшать смещение оценки $\widehat{h}(x)$.

7) Что означает выражение $\mathbf{E}\{\widehat{h}(x) \,|\, X = x\}$ и чем оно отличается от $\mathbf{E}\,\widehat{h}(x)$?

8) Зачем для оценки плотности требуется, чтобы $\int_{\mathbb{R}} K_\lambda(t)\,\mathrm{d}t = \lambda$?

9) Приведите вид функции $K_\lambda(t, x)$, для которого оценка

$$\widehat{f}(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K_\lambda(x - X_i, x)$$

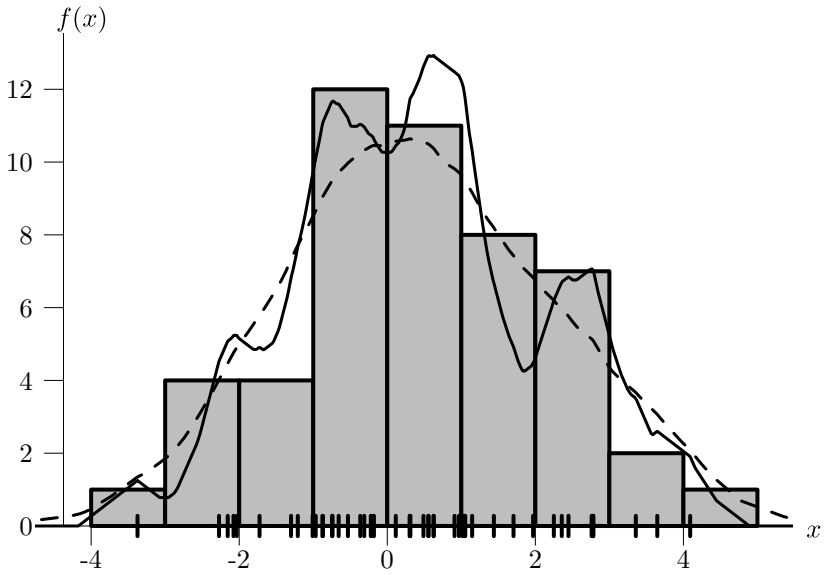

Рис. 9: Оценки функции плотности: гистограмма, ядерная оценка с треугольным ядром и $\lambda = 1$ (сплошная линия), ядерная оценка с квадратичным ядром и $\lambda = 2$ (штрихованная линия). Данные сгенерированы из нормального распределения $\mathcal{N}(1, 2)$

равна гистограммной оценке плотности.