

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ  
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ  
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

*Кафедра дискретной математики ФИВТ*

## **ВВЕДЕНИЕ В МАТЕМАТИЧЕСКУЮ СТАТИСТИКУ**

Учебное пособие

Составители *М.Е. Жуковский, И.В. Родионов, Д.А. Шабанов*

МОСКВА  
МФТИ  
2016

Рецензент  
В.И. Питербарг

**Введение в математическую статистику:** учебное пособие / сост. М.Е. Жуковский, И.В. Родионов, Д.А. Шабанов. – М.: МФТИ, 2016. – 107 с.

В этом учебном пособии отражено содержание курсов по математической статистике, которые авторы ведут у студентов факультета инноваций и высоких технологий Московского физико-технического института (государственного университета) в четвертом семестре. Каждая глава пособия содержит определенный теоретический материал, а также набор задач по соответствующей теме с решениями. В конце каждой главы приведен список упражнений. Сформулированные в пособии теоремы и теоретические утверждения необходимы для решения задач, их доказательства не приводятся. Кроме того, книга не содержит многих вспомогательных утверждений. Пособие охватывает семестровый курс математической статистики и рассчитано на студентов математических и физических специальностей высших учебных заведений.

- © Жуковский М.Е., Родионов И.В.,  
Шабанов Д.А., 2016
- © Федеральное государственное автономное  
образовательное учреждение  
высшего профессионального образования  
«Московский физико-технический институт  
(государственный университет)», 2016

# Содержание

<b>1. Виды сходимостей случайных векторов</b>	<b>5</b>
<b>2. Статистики и оценки</b>	<b>14</b>
<b>3. Методы нахождения оценок</b>	<b>21</b>
3.1. Метод моментов . . . . .	21
3.2. Метод максимального правдоподобия . . . . .	23
3.3. Метод выборочных квантилей . . . . .	25
<b>4. Сравнение оценок и эффективные оценки</b>	<b>30</b>
<b>5. Условное математическое ожидание</b>	<b>37</b>
<b>6. Условные распределения и подсчет условных математических ожиданий</b>	<b>44</b>
<b>7. Достаточные статистики и оптимальные оценки</b>	<b>50</b>
7.1. Достаточные статистики . . . . .	50
7.2. Полные статистики . . . . .	52
7.3. Оптимальные оценки . . . . .	54
<b>8. Доверительные интервалы</b>	<b>57</b>
8.1. Определение и методы построения доверительных интервалов . . . . .	57
8.2. Асимптотические доверительные интервалы . . .	59
<b>9. Линейная регрессия</b>	<b>63</b>
<b>10. Проверка статистических гипотез</b>	<b>70</b>
10.1. Равномерно наиболее мощные критерии . . . . .	71
10.2. Проверка простых гипотез . . . . .	72
10.3. Проверка сложных гипотез . . . . .	74
<b>11. Проверка линейных гипотез в гауссовской регрессионной модели</b>	<b>77</b>

<b>12.Критерии согласия</b>	<b>84</b>
12.1. Критерий согласия Колмогорова . . . . .	84
12.2. Критерий согласия хи-квадрат . . . . .	86
<b>13.Коэффициенты корреляции</b>	<b>92</b>
13.1. Коэффициент корреляции Пирсона . . . . .	92
13.2. Коэффициент корреляции Спирмэна . . . . .	95
13.3. Коэффициент корреляции Кэндалла . . . . .	97
<b>14.Байесовские оценки</b>	<b>102</b>
<b>Литература</b>	<b>107</b>

# 1. Виды сходимостей случайных векторов

Первая глава относится к курсу теории вероятностей и является вспомогательной. Мы включили ее в настоящее пособие, потому что, как правило, в стандартном курсе теории вероятностей математических и технических вузов (и, в частности, в предлагаемом нами пособии по теории вероятностей) вопрос сходимости случайных векторов не затрагивается подробно. При изучении же свойств оценок многомерных параметров различных распределений в курсе математической статистики такой вопрос возникает.

Определения сходимостей в многомерном случае аналогичны одномерному случаю.

**▲ Определение 1.1.** Пусть  $\{\xi_n\}_{n \in \mathbb{N}}$  — последовательность случайных векторов размерности  $m$ .

1) Последовательность  $\xi_n$  сходится *почти наверное* к случайному вектору  $\xi$  при  $n \rightarrow \infty$  (пишут  $\xi_n \xrightarrow{\text{п.н.}} \xi$ ), если

$$P\left(\omega : \lim_{n \rightarrow \infty} \xi_n(\omega) = \xi(\omega)\right) = 1.$$

2) Последовательность  $\xi_n$  сходится *по вероятности* к случайному вектору  $\xi$  при  $n \rightarrow \infty$  (пишут  $\xi_n \xrightarrow{P} \xi$ ), если для любого  $\varepsilon > 0$  выполнено

$$P(\|\xi_n - \xi\|_2 \geq \varepsilon) \rightarrow 0, \quad n \rightarrow \infty,$$

где  $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \dots + x_m^2}$  для  $\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}^m$ .

3) Последовательность  $\xi_n$  сходится в  $L^p$  к случайному вектору  $\xi$  при  $n \rightarrow \infty$  (пишут  $\xi_n \xrightarrow{L^p} \xi$ ), если

$$E(\|\xi_n - \xi\|_p)^p \rightarrow 0, \quad n \rightarrow \infty,$$

где  $\|\mathbf{x}\|_p = (|x_1|^p + \dots + |x_m|^p)^{1/p}$ .

4) Последовательность  $\xi_n$  сходится *по распределению* к случайному вектору  $\xi$  при  $n \rightarrow \infty$  (пишут  $\xi_n \xrightarrow{d} \xi$ ), если для любой ограниченной непрерывной функции  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  выполнено

$$Ef(\xi_n) \rightarrow Ef(\xi), \quad n \rightarrow \infty.$$

Имеют место следующие связи между сходимостями векторов и компонент этих векторов.

**Теорема 1.1** Пусть заданы случайные векторы

$$\xi = (\xi^{(1)}, \dots, \xi^{(m)}), \xi_n = (\xi_n^{(1)}, \dots, \xi_n^{(m)}),$$

где  $n \in \mathbb{N}$ . Тогда

$$\xi_n \xrightarrow{n.n.} \xi \iff \forall i = 1, \dots, m \quad \xi_n^{(i)} \xrightarrow{n.n.} \xi^{(i)},$$

$$\xi_n \xrightarrow{P} \xi \iff \forall i = 1, \dots, m \quad \xi_n^{(i)} \xrightarrow{P} \xi^{(i)},$$

$$\xi_n \xrightarrow{L^p} \xi \iff \forall i = 1, \dots, m \quad \xi_n^{(i)} \xrightarrow{L^p} \xi^{(i)}.$$

Заметим, что для сходимости по распределению теорема неверна. Из сходимости по распределению случайных векторов следует покомпонентная сходимость, однако обратное неверно, что показывает следующий пример.

Пусть  $\xi$  и  $\eta$  — независимые одинаково распределённые случайные величины. Пусть последовательности случайных величин  $\{\xi_n\}_{n \in \mathbb{N}}$  и  $\{\eta_n\}_{n \in \mathbb{N}}$  таковы, что  $\xi_n = \xi$  и  $\eta_n = \eta$  для любого  $n \in \mathbb{N}$ . Тогда, очевидно,  $\xi_n \xrightarrow{d} \xi$  и  $\eta_n \xrightarrow{d} \xi$ , так как  $\xi \stackrel{d}{=} \eta$  по условию. Но векторной сходимости нет:  $(\xi_n, \eta_n) \not\xrightarrow{d} (\xi, \xi)$ , ведь  $(\xi_n, \eta_n) \xrightarrow{d} (\xi, \eta)$  и распределения векторов  $(\xi, \eta)$  и  $(\xi, \xi)$  не совпадают, так как все значения второго вектора почти наверное лежат на прямой  $y = x$ , а для первого вектора это далеко не всегда так.

Заметим, наконец, что для случайных векторов взаимосвязь различных видов сходимостей та же, что и для случайных величин.

▲ **Задача 1.1.** Докажите, что из сходимости последовательности  $m$ -мерных случайных векторов  $\xi_n$  к константе  $C \in \mathbb{R}^m$  по распределению следует сходимость  $\xi_n \xrightarrow{P} C$ .

## Решение

Из теоремы 1.1 следует, что для любого  $i \in \{1, \dots, m\}$  выполнено  $\xi_n^i \xrightarrow{d} C^i$ , где  $\xi_n = (\xi_n^1, \dots, \xi_n^m)$ ,  $C = (C^1, \dots, C^m)$ . Из курса вероятности нам известно, что последняя сходимости влечет сходимость по вероятности:  $\xi_n^i \xrightarrow{P} C^i$  для любого  $i \in \{1, \dots, m\}$ . Но тогда из теоремы 1.1 имеем  $\xi_n \xrightarrow{P} C$ .

### Задача решена.

Следующим важным результатом, который нам понадобится, является теорема о наследовании сходимостей.

**Теорема 1.2** *Имеют место следующие утверждения.*

**(о наследовании сходимостей)**

1. Пусть  $\xi_n \xrightarrow{n.n.} \xi$  — случайные векторы размерности  $m$ . Пусть  $h : \mathbb{R}^m \rightarrow \mathbb{R}^s$  — функция, непрерывная почти всюду относительно распределения случайной величины  $\xi$  (т.е. существует такое множество  $B \in \mathcal{B}(\mathbb{R}^n)$ , что  $h$  непрерывна на  $B$  и  $P(\xi \in B) = 1$ ). Тогда  $h(\xi_n) \xrightarrow{n.n.} h(\xi)$ .
2. Пусть  $\xi_n \xrightarrow{P} \xi$  — случайные векторы размерности  $m$ . В тех же условиях, что и в пункте 1,  $h(\xi_n) \xrightarrow{P} h(\xi)$ .
3. Пусть  $\xi_n \xrightarrow{d} \xi$  — случайные векторы размерности  $m$ . Пусть  $h : \mathbb{R}^m \rightarrow \mathbb{R}^s$  — непрерывная функция (достаточно непрерывности всюду относительно распределения  $\xi$ ). Тогда  $h(\xi_n) \xrightarrow{d} h(\xi)$ .

В силу теорем 1.1 и 1.2, из сходимости последовательностей случайных величин  $\xi_n \xrightarrow{P} \xi$ ,  $\eta_n \xrightarrow{P} \eta$  следуют сходимости

$$\xi_n + \eta_n \xrightarrow{P} \xi + \eta, \quad \xi_n \eta_n \xrightarrow{P} \xi \eta.$$

Для сходимости по распределению аналогичного вывода сделать нельзя, ведь, как мы уже знаем, из покомпонентной сходимости по распределению не следует сходимость векторов по распределению. Однако если одна из компонент двумерного вектора сходится по распределению к константе, то можно брать от вектора непрерывные функции, и при этом сходимость по распределению будет сохраняться. В частности, справедливо следующее утверждение.

**Теорема 1.3 (лемма Слуцкого)** Пусть  $\xi_n \xrightarrow{d} \xi$  и  $\eta_n \xrightarrow{d} C = \text{const}$  — случайные величины. Тогда

$$\xi_n + \eta_n \xrightarrow{d} \xi + C,$$

$$\xi_n \eta_n \xrightarrow{d} \xi C.$$

Следствием теорем 1.2 и 1.3 является

**Теорема 1.4** Пусть  $\xi_n \xrightarrow{d} \xi$  — случайные векторы размерности  $m$ , а  $h(x) : \mathbb{R}^m \rightarrow \mathbb{R}$  — функция, дифференцируемая в точке  $a \in \mathbb{R}^m$ . Пусть  $b_n \rightarrow 0$ ,  $b_n \neq 0$ . Тогда

$$\frac{h(a + \xi_n b_n) - h(a)}{b_n} \xrightarrow{d} (\xi, \nabla h|_a),$$

где  $\nabla h|_a$  — градиент функции  $h(x)$ , взятый в точке  $a$ .

В частности, если  $\xi_n \xrightarrow{d} \xi$  — случайные величины,  $h(x) : \mathbb{R} \rightarrow \mathbb{R}$  — функция, дифференцируемая в точке  $a \in \mathbb{R}$ , то

$$\frac{h(a + \xi_n b_n) - h(a)}{b_n} \xrightarrow{d} \xi h'(a).$$



**▲ Задача 1.2.** Пусть  $\{\eta_n\}_{n \in \mathbb{N}}$  — последовательность независимых одинаково распределенных случайных величин с ненулевым математическим ожиданием  $a$  и дисперсией  $\sigma^2$ . Обозначим  $S_n = \eta_1 + \dots + \eta_n$  при всех  $n \in \mathbb{N}$ . Существует ли предел по распределению последовательности случайных величин  $\sqrt{n} \left( \frac{n}{S_n} - \frac{1}{a} \right)$ ? Если ответ положительный, то чему он равен?

### Решение

По центральной предельной теореме,

$$\sqrt{n} \left( \frac{S_n}{n} - a \right) \xrightarrow{d} \eta \sim \mathcal{N}(0, \sigma^2).$$

По закону больших чисел,  $\frac{S_n}{n} \xrightarrow{P} a$ . По теореме 1.2, имеем

$$\frac{n}{aS_n} \xrightarrow{P} \frac{1}{a^2}.$$

По лемме Слуцкого,

$$\sqrt{n} \left( \frac{n}{S_n} - \frac{1}{a} \right) = \frac{\sqrt{n} \left( a - \frac{S_n}{n} \right)}{a \frac{S_n}{n}} \xrightarrow{d} -\frac{1}{a^2} \eta \sim \mathcal{N} \left( 0, \frac{\sigma^2}{a^4} \right).$$

### Задача решена.

Задачу можно было решить и с помощью теоремы 1.4. Действительно, утверждение задачи незамедлительно следует из нее, если положить

$$\xi_n = \sqrt{n} \left( \frac{S_n}{n} - a \right), \quad h(x) = \frac{1}{x}, \quad b_n = \frac{1}{\sqrt{n}}, \quad \xi = \eta.$$

Для случайных векторов остаются верны предельные теоремы, аналогичные соответствующим теоремам для случайных величин: закон больших чисел, усиленный закон больших чисел, центральная предельная теорема. Приведем их формулировки. Далее для случайного вектора  $\xi$  мы будем писать  $E\xi$ , подразумевая вектор, составленный из математических ожиданий компонент вектора  $\xi$ .

**Теорема 1.5**  
(многомер-  
ный аналог  
закона  
больших  
чисел)

Пусть  $\{\xi_n\}_{n \in \mathbb{N}}$  — независимые случайные векторы, дисперсия каждой компоненты каждого случайного вектора ограничена константой  $C$ . Тогда

$$\frac{\xi_1 + \dots + \xi_n - \mathbb{E}(\xi_1 + \dots + \xi_n)}{n^{0.5+\delta}} \xrightarrow{P} 0,$$

где  $\delta$  — произвольное положительное число.

**Теорема 1.6**  
(многомер-  
ный аналог  
усиленно-  
го закона  
больших  
чисел)

Пусть  $\{\xi_n\}_{n \in \mathbb{N}}$  — независимые одинаково распределенные случайные векторы, математические ожидания всех компонент конечны и равны  $\mathbb{E}\xi_1 = a$ . Тогда

$$\frac{\xi_1 + \dots + \xi_n}{n} \xrightarrow{n.n.} a.$$

**Теорема 1.7**  
(многомер-  
ная цен-  
тральная  
предельная  
теорема)

Пусть  $\{\xi_n\}_{n \in \mathbb{N}}$  — независимые одинаково распределенные случайные векторы,  $\mathbb{E}\xi_1 = a$ , матрица ковариаций случайного вектора  $\xi_1$  равна  $D\xi_1 = \Sigma$ . Тогда

$$\sqrt{n} \left( \frac{\xi_1 + \dots + \xi_n}{n} - a \right) \xrightarrow{d} \eta \sim \mathcal{N}(0, \Sigma),$$

где  $\mathcal{N}(0, \Sigma)$  — многомерное нормальное распределение.

▲ **Задача 1.3.** Пусть  $X_1, \dots, X_n$  — независимые одинаково распределенные случайные величины с распределением  $\mathcal{N}(0, \sigma^2)$ . Рассмотрим  $Y = \frac{1}{n} \sum_{i=1}^n X_i^4$ ,  $Z = \frac{1}{n} \sum_{i=1}^n X_i^2$ . Используя многомерную центральную предельную теорему, найдите предел по распределению для выражения  $\sqrt{n}(T - \sigma)$ , где  $T = \sqrt{\frac{Y}{3Z}}$ .

## Решение

тесно связано с поиском асимптотически нормальных оценок для некоторых параметров, что мы будем проходить позднее. Поскольку для  $X_1$  выполнено  $\mathbf{E}X_1^{2n} = (2n - 1)!!\sigma^{2n}$ , то  $\mathbf{E}X_1^2 = \sigma^2$  и  $\mathbf{E}X_1^4 = 3\sigma^4$ , отсюда, пользуясь многомерной центральной предельной теоремой, получаем:

$$\begin{aligned} & \sqrt{n} \left( \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i^4 \\ \frac{1}{n} \sum_{i=1}^n X_i^2 \end{pmatrix} - \begin{pmatrix} \mathbf{E}X_1^4 \\ \mathbf{E}X_1^2 \end{pmatrix} \right) = \\ & = \sqrt{n} \left( \begin{pmatrix} Y \\ Z \end{pmatrix} - \begin{pmatrix} 3\sigma^4 \\ \sigma^2 \end{pmatrix} \right) \xrightarrow{d} \xi \sim \mathcal{N}(0, \Sigma), \end{aligned}$$

где матрица ковариаций  $\Sigma$  вектора  $(X_1^4, X_1^2)^T$  равна

$$\Sigma = \begin{pmatrix} 96\sigma^8 & 12\sigma^6 \\ 12\sigma^6 & 2\sigma^4 \end{pmatrix}.$$

Далее, используя теорему 1.4 для  $b_n = \frac{1}{\sqrt{n}}$ ,  $a = (EX_1^4, EX_1^2)^T = (3\sigma^4, \sigma^2)^T$ ,  $\xi_n = \sqrt{n}((Y, Z)^T - (EX_1^4, EX_1^2)^T)$  и  $h(x, y) = \sqrt{\frac{x}{3y}}$ , получаем следующее соотношение:

$$\sqrt{n}(h(Y, Z) - h(3\sigma^4, \sigma^2)) \xrightarrow{d} (\nabla h|_a, \xi) \sim \mathcal{N}(0, \nabla h|_a^T \Sigma \nabla h|_a),$$

где последнее равенство вытекает из свойств гауссовских векторов. Найдём число  $d^2 := \nabla h|_a^T \Sigma \nabla h|_a$  (а это действительно число, поскольку матрица размера  $2 \times 2$  умножается с двух сторон на вектор размерности 2). Находим

$$\nabla h = \left( \frac{1}{\sqrt{12xy}}, -\frac{\sqrt{x}}{\sqrt{12y^3}} \right)^T,$$

отсюда  $\nabla h|_{(3\sigma^4, \sigma^2)} = \left( \frac{1}{6\sigma^3}, -\frac{1}{2\sigma} \right)^T$  и  $d^2 = \frac{7}{6}\sigma^2$ . Окончательный ответ таков:

$$\sqrt{n}(T - \sigma) \xrightarrow{d} \eta \sim \mathcal{N}\left(0, \frac{7}{6}\sigma^2\right).$$

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Пусть  $X_1, \dots, X_n$  — независимые одинаково распределенные случайные величины с распределением  $\text{Exp}(\alpha)$ ,  $\alpha > 0$ . Рассмотрим статистику  $Y = \frac{1}{n} \sum_{i=1}^n X_i$ . Найдите такие константы  $a(\alpha)$  и  $\sigma^2(\alpha) > 0$ , что выполнено

$$\sqrt{n}(Y \sin Y - a(\alpha)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\alpha)) \text{ при } n \rightarrow \infty.$$

2. Пусть  $\{\xi_n\}_{n=1}^\infty$ ,  $\{\eta_n\}_{n=1}^\infty$  и  $\{\zeta_n\}_{n=1}^\infty$  — последовательности случайных величин. Докажите, что если  $\xi_n \xrightarrow{d} \xi$ ,  $|\xi_n - \eta_n| \leq \zeta_n |\xi_n|$  и  $\zeta_n \xrightarrow{P} 0$ , то  $\eta_n \xrightarrow{d} \xi$ .
3. Задана набор независимых одинаково распределенных случайных величин  $X_1, \dots, X_n$  с распределением  $\mathcal{N}(0, \sigma^2)$ . Рассмотрим статистики  $Y = \frac{1}{n} \sum_{i=1}^n |X_i|$ ,  $Z = \frac{1}{n} \sum_{i=1}^n X_i^2$  и  $T = \sqrt{\frac{2}{\pi}} Z/Y$ . Найдите предел сходимости по распределению выражения

$$\sqrt{n}(T - \sigma).$$

4. Пусть  $\{\xi_n\}_{n=1}^\infty$  и  $\{\eta_n\}_{n=1}^\infty$  — две последовательности случайных величин, причем для каждого  $n \geq 1$  величины  $\xi_n$  и  $\eta_n$  независимы. Пусть  $\xi_n \xrightarrow{P} \xi$ ,  $\eta_n \xrightarrow{P} \eta$ . Используя метод характеристических функций, докажите, что  $\xi$  и  $\eta$  — тоже независимы.
5. Пусть  $X_1, \dots, X_n$  — независимые случайные величины, имеющие распределение Лапласа с параметром  $\sigma$ , т.е. плотность равна

$$p(x) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}.$$

Рассмотрим  $Y = \frac{1}{n} \sum_{i=1}^n |X_i|$ ,  $Z = \frac{1}{n} \sum_{i=1}^n X_i^2$ . Используя многомерную центральную предельную теорему, найдите предел по распределению для выражения

$$\sqrt{n}(T - \sigma),$$

где  $T = Z^2/(4Y^3)$ .

6. Известно, что условие сходимости по распределению последовательности случайных величин  $\{\xi_n\}_{n=1}^\infty$  к случайной величине  $\xi$  можно ослабить:  $\xi_n \xrightarrow{d} \xi$ , если для любой ограниченной, *равномерно непрерывной* функции  $f(x)$  выполнено  $Ef(\xi_n) \rightarrow Ef(\xi)$  при  $n \rightarrow \infty$ . Докажите, используя это условие, обобщение леммы Слуцкого: пусть  $\xi_n \xrightarrow{d} \xi$  и  $\eta_n \xrightarrow{d} C = \text{const}$ . Тогда для любой непрерывной функции  $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$  выполнено

$$\varphi(\xi_n, \eta_n) \xrightarrow{d} \varphi(\xi, C).$$

7. Докажите теорему 1.4.
8. Докажите следующее обобщение теоремы 1.4 в одномерном случае. Пусть функция  $f(x)$  имеет производную в точке  $x = 0$ . Тогда если для последовательностей случайных величин  $\{\xi_n\}_{n=1}^\infty$ ,  $\{\eta_n\}_{n=1}^\infty$  выполнено  $\xi_n \eta_n \xrightarrow{d} \eta$  и  $\eta_n \xrightarrow{P} 0$ , то

$$\xi_n(f(\eta_n) - f(0)) \xrightarrow{d} f'(0)\eta.$$

## 2. Статистики и оценки

В начале этой главы стоит поговорить о задаче математической статистики. Вспомним, что в теории вероятностей мы предполагали, что нам известна природа некоторого явления (распределение случайной величины), и это знание позволяло нам изучать поведение наблюдаемых в экспериментах величин, связанных с упомянутым явлением. В математической статистике все наоборот — данными являются результаты экспериментов (значения случайной величины), а требуется понять природу изучаемого явления (распределение величины).

Пусть  $X$  — случайная величина (или случайный вектор), принимающая значения во множестве  $\mathcal{X}$  (всюду далее мы считаем, что это множество равно либо  $\mathbb{R}$ , либо  $\mathbb{R}^n$ ) и имеющая распределение  $\mathbf{P}$ . Без ограничения общности можно считать, что эта случайная величина задана на пространстве  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbf{P})$  правилом  $X(x) = x$  для любого  $x \in \mathcal{X}$ . Будем называть случайную величину  $X$  *наблюдением*. Задача математической статистики состоит в том, чтобы по набору наблюдений выяснить какую-либо информацию о распределении  $\mathbf{P}$ , которое предполагается неизвестным. Тем не менее, в таких задачах, как правило, известно, что распределение  $\mathbf{P}$  принадлежит некоторому заданному множеству распределений  $\mathcal{P}$ . Тройка  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathcal{P})$  называется *вероятностно-статистической моделью*.

Как правило, для выяснения природы изучаемого явления удастся провести *независимые* эксперименты. Говоря математическим языком, восстановить распределение нужно по набору независимых наблюдений, который называется *выборкой*. Иными словами, *выборка* — это вектор  $(X_1, \dots, X_n)$  (или просто  $X_1, \dots, X_n$ ) составленный из независимых случайных величин (векторов), распределенных так же, как и  $X$ . Такой вектор принимает значения во множестве  $\mathcal{X}^n$ , а его распределение  $\mathbf{P}^n$  задано на борелевской  $\sigma$ -алгебре  $\mathcal{B}(\mathcal{X}^n)$ . Можно доказать, что существует вероятностное пространство  $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X}^\infty), \mathbf{P}^\infty)$ , на котором можно задать бесконечную последовательность случайных величин (векторов)  $X_1, X_2, \dots$  (выборку бесконечного размера), распределенных так же, как и  $X$ . Тем самым, на этом пространстве задана выборка любого размера  $n$ . В дальнейшем мы для простоты обозначений будем писать  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbf{P})$  вместо

$(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X}^\infty), \mathbf{P}^\infty)$  соответственно. Более того, выборку мы тоже будем называть наблюдением.

**▲ Определение 2.1.** Пусть  $(\mathcal{Y}, \mathcal{E})$  — измеримое пространство (т.е.  $\mathcal{E}$  —  $\sigma$ -алгебра подмножеств  $\mathcal{Y}$ ). Любое  $(\mathcal{B}(\mathcal{X})|\mathcal{E})$ -измеримое отображение  $S : \mathcal{X} \rightarrow \mathcal{Y}$  (напомним, что отображение называется измеримым, если для любого множества  $E \in \mathcal{E}$  выполнено  $\{x : S(x) \in E\} \in \mathcal{B}(\mathcal{X})$ ) называется *статистикой*.

Если  $\mathcal{Y} = \Theta$  (где семейство распределений параметризовано следующим образом  $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ , т.е. мы имеем дело с *параметрической моделью*), то  $S$  называется *оценкой* параметра  $\theta$ .

Иными словами, если  $X = (X_1, \dots, X_n)$  — выборка, то статистикой  $S(X)$  называется измеримая функция от выборки, а оценкой — такая статистика, значением которой является параметр распределения (или функция  $\tau$  от параметра — в этом случае  $\mathcal{Y} = \tau(\Theta)$ ).

В случае параметрической модели задача отыскания “истинного” распределения сводится к нахождению “истинного” значения параметра. Для этого находится удачная (в каком-нибудь смысле) оценка параметра. В настоящей главе мы поговорим о том, какие оценки принято считать “удачными”.

### ▲ Примеры статистик:

1. Пусть  $X_1, \dots, X_n$  — выборка, а  $g(x)$  — борелевская функция. Величина

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

называется *выборочной характеристикой*  $g(x)$ . Например,

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

называется *выборочным средним*, а

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

— *выборочным  $k$ -тым моментом.*

2. Рассмотрим функции от выборочных характеристик:

$$S(X) = h\left(\overline{g_1(X)}, \dots, \overline{g_k(X)}\right),$$

где  $h$  — борелевская функция, заданная на  $\mathbb{R}^k$ . Примером может служить *выборочная дисперсия*

$$s^2 = \overline{X^2} - (\overline{X})^2.$$

3. Упорядочим значения выборки  $(X_1, \dots, X_n)$  по возрастанию:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

Полученная совокупность статистик называется *вариационным рядом*, а его член  $X_{(k)}$  —  *$k$ -той порядковой статистикой*.

Обратимся теперь к свойствам оценок.

Пусть  $X$  — наблюдение из неизвестного распределения  $P \in \{P_\theta, \theta \in \Theta\}$ .

**▲ Определение 2.2.** Оценка  $\theta^*(X)$  называется *несмещенной* оценкой параметра  $\tau(\theta)$ , если для любого  $\theta \in \Theta$  выполнено равенство  $E_\theta \theta^*(X) = \tau(\theta)$ .

Заметим, что в определении выше мы использовали запись  $E_\theta$  (сокращение от  $E_{P_\theta}$ ), которая означает математическое ожидание случайной величины с распределением  $P_\theta$ .

**▲ Задача 2.1.** Пусть  $X_1, \dots, X_n$  — выборка из неизвестного распределения  $P \in \{P_\theta, \theta \in \Theta\}$ , причем  $\Theta \subset \mathbb{R}$ , и для любого  $\theta \in \Theta$  выполнено  $E_\theta X_1 = \theta$ . Докажите, что  $\bar{X}$  — несмещенная оценка параметра  $\theta$ .



## Решение

В силу линейности математического ожидания,

$$\mathbb{E}_\theta \bar{X} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta X_i = \mathbb{E}_\theta X_1 = \theta.$$

**Задача решена.**

Далее рассмотрим три важных асимптотических свойства оценок.

**▲ Определение 2.3.** Пусть  $X = (X_1, \dots, X_n)$  — выборка. Оценка  $\theta^*(X)$  (точнее, последовательность оценок) называется *состоятельной* оценкой параметра  $\tau(\theta)$ , если для любого  $\theta \in \Theta$  выполнено  $\theta^*(X) \xrightarrow{P_\theta} \tau(\theta)$  при  $n \rightarrow \infty$ . Оценка  $\theta^*(X)$  называется *сильно состоятельной* оценкой  $\tau(\theta)$ , если для любого  $\theta \in \Theta$  выполнено  $\theta^*(X) \rightarrow \tau(\theta)$  ( $P_\theta$ -п.н.). Оценка  $\theta^*(X)$  называется *асимптотически нормальной* оценкой  $\tau(\theta)$ , если для любого  $\theta \in \Theta$

$$\sqrt{n}(\theta^*(X) - \tau(\theta)) \xrightarrow{d_\theta} \xi \sim \mathcal{N}(0, \sigma^2(\theta))$$

для некоторой функции  $\sigma^2(\theta)$ , которая называется *асимптотической дисперсией* оценки  $\theta^*(X)$ .

Заметим, что в определении выше мы использовали запись  $d_\theta$ , которая означает сходимость по распределению при условии, что элементы выборки имеют распределение  $P_\theta$ .

**▲ Задача 2.2.** Пусть задана выборка  $X_1, \dots, X_n$  из неизвестного распределения  $P \in \mathcal{P}$ . Пусть, кроме того,  $k \in \mathbb{N}$  и  $0 < \mathbb{E}_P |X_1|^{2k} < \infty$  для всех  $P \in \mathcal{P}$ . Доказать, что выборочный  $k$ -тый момент  $\bar{X}^k$  является несмещённой, сильно состоятельной и асимптотически нормальной оценкой  $\mathbb{E}_P X_1^k$ .

## Решение

Несмещенность, как и в задаче 1, следует из линейности математического ожидания. По закону больших чисел,  $\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} E_P X_1^k$ , что нам и требовалось для состоятельности. Наконец, асимптотическая нормальность следует из центральной предельной теоремы:

$$\sqrt{n} \left( \overline{X^k} - E_P X_1^k \right) \xrightarrow{d} \xi \sim \mathcal{N}(0, D_P X_1^k).$$

## Задача решена.

Основными инструментами доказательства того, что оценка является состоятельной или асимптотически нормальной, являются законы больших чисел, центральная предельная теорема, теорема о наследовании сходимостей, а также теорема о наследовании асимптотической нормальности, сформулированная ниже.

**Теорема 2.1 (о наследовании асимптотической нормальности)**

*Пусть  $\theta^*(X)$  — асимптотически нормальная оценка  $\theta$  с асимптотической дисперсией  $\sigma^2(\theta)$ , а  $\tau(\theta)$  — дифференцируемая функция на  $\Theta \subset \mathbb{R}$  (производная конечна и не равна нулю в каждой точке  $\theta \in \Theta$ ). Тогда  $\tau(\theta^*(X))$  — асимптотически нормальная оценка  $\tau(\theta)$  с асимптотической дисперсией  $\sigma^2(\theta)(\tau'(\theta))^2$ .*

▲ **Задача 2.3.** Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[1 + \theta, \theta(1 + \theta)]$ ,  $\theta > 1$ . Найдите асимптотически нормальную оценку параметра  $\theta$  и ее асимптотическую дисперсию.

## Решение

Положим  $m(\theta) = \frac{(\theta+1)^2}{2}$ . По центральной предельной теореме,  $\bar{X}$  — асимптотически нормальная оценка параметра  $m(\theta)$  с асимптотической дисперсией  $\sigma^2(\theta) = \frac{(\theta^2-1)^2}{12}$ . Заметим, что  $m(\theta)$  принимает значения во множестве  $(2, \infty)$ . Положим  $\tau(x) = \sqrt{2x} - 1$  (дифференцируемая на  $(2, \infty)$  функция). Тогда, по теореме о наследовании асимптотической нормальности,  $\tau(\bar{X})$  — асимптотически нормальная оценка  $\theta$  с асимптотической дисперсией

$$\sigma^2(\theta) \left( \tau' |_{m(\theta)} \right)^2 = \frac{(\theta^2 - 1)^2}{12} \frac{1}{2m(\theta)} = \frac{(\theta - 1)^2}{12}.$$

**Задача решена.**

## ▲ Задачи для самостоятельного решения

1. Пусть  $X_1, \dots, X_n$  — выборка из распределения  $R(0, \theta)$  (равномерного распределения на отрезке  $[0, \theta]$ ). Проверьте на несмещенность, состоятельность и сильную состоятельность следующие оценки параметра  $\theta$ :  $2\bar{X}$ ,  $\bar{X} + X_{(n)}/2$ ,  $(n+1)X_{(1)}$ ,  $X_{(1)} + X_{(n)}$ ,  $\frac{n+1}{n}X_{(n)}$ .
2. Пусть  $X_1, \dots, X_n$  — выборка из распределения  $\text{Bin}(1, \theta)$ . Для каких функций  $\tau(\theta)$  существуют несмещенные оценки?
3. Пусть  $\hat{\theta}_n(X)$  — асимптотически нормальная оценка параметра  $\theta$  с асимптотической дисперсией  $\sigma^2(\theta)$ . Докажите, что тогда  $\hat{\theta}_n(X)$  является состоятельной оценкой  $\theta$ .
4. Пусть  $X_1, \dots, X_n$  выборка из распределения с параметром  $\sigma^2$ . Пусть, кроме того,  $D_{\sigma^2}X_1 = \sigma^2$ . Докажите, что статистика  $s^2 = 1/n \sum_{i=1}^n (X_i - \bar{X})^2$  равна  $\bar{X}^2 - (\bar{X})^2$  и является состоятельной оценкой  $\sigma^2$ . Является ли она несмещенной оценкой того же параметра?
5. Задана выборка  $X_1, \dots, X_n$  из экспоненциального закона с параметром  $\lambda > 0$ . Для какой величины  $\tau(\theta)$  статистика

$\overline{X \ln X}$  является асимптотически нормальной оценкой? Вычислите асимптотическую дисперсию данной оценки.

6. Пусть  $X_1, \dots, X_n$  — выборка из экспоненциального распределения с параметром  $\theta$ . Покажите, что для любого  $k \in \mathbb{N}$  статистика  $\sqrt[k]{k!/\overline{X^k}}$  является асимптотически нормальной оценкой параметра  $\theta$ . Найдите ее асимптотическую дисперсию.
7. Пусть  $X_1, \dots, X_n$  — выборка из распределения  $R(0, \theta)$ . Найдите такое число  $\delta > 0$  и невырожденный закон распределения  $P_\theta$  (т.е. не являющийся законом распределения константы), что  $n^\delta(X_{(n)} - \theta) \xrightarrow{d} \xi \sim P_\theta$  при  $n \rightarrow \infty$ .
8. Постройте асимптотически нормальную оценку для  $b$  по выборке  $X_1, \dots, X_n$  из распределения

$$P(X_1 = -1) = 1 - a - a^2 - 4b,$$

$$P(X_1 = 2) = 5b,$$

$$P(X_1 = 5) = a + a^2 + b,$$

где  $a, b$  — неизвестные положительные параметры.

9. Задана выборка  $X_1, \dots, X_n$  из распределения  $R(0, \theta)$ , где  $\theta$  — неизвестный параметр. Рассмотрим статистики  $Y = \frac{1}{n} \sum_{i=1}^n X_i^2$ ,  $Z = \frac{1}{n} \sum_{i=1}^n X_i^3$ . Найдите асимптотически нормальную оценку  $\theta$  как функцию от  $Y/Z$ . Найдите ее асимптотическую дисперсию.
10. Задана выборка  $X_1, \dots, X_n$ ,  $X_1 = \xi + \eta$ ,  $\xi \sim R(0, \theta)$ ,  $\eta \sim R(\theta, 2\theta)$  — независимые случайные величины, где  $\theta > 0$  — неизвестный параметр. Найдите асимптотически нормальную оценку  $\theta$  и ее асимптотическую дисперсию.

### 3. Методы нахождения оценок

Во второй главе мы изучили некоторые свойства оценок. Возникает естественный вопрос: существует ли общий метод нахождения оценок, обладающих этими свойствами? Ответ на этот вопрос положительный. В настоящей главе мы рассмотрим два таких метода и поговорим о свойствах оценок, с помощью них полученных. В конце главы мы определим оценки, называемые выборочными квантилями, которые также обладают некоторыми из упомянутых свойств и могут быть использованы в случаях, когда нужную оценку не удастся построить ни одним из двух методов.

#### 3.1. Метод моментов

Пусть  $\Theta \subset \mathbb{R}^k$ . Рассмотрим некоторое семейство распределений  $\{\mathbf{P}_\theta, \theta \in \Theta\}$  и выборку  $X_1, \dots, X_n$  из неизвестного распределения  $\mathbf{P}_\theta$ .

Для построения оценки выберем такие борелевские функции  $g_1, \dots, g_k : \mathbb{R} \rightarrow \mathbb{R}$ , что для любого  $i \in \{1, \dots, k\}$  и любого  $\theta \in \Theta$  существует конечное  $\mathbf{E}_\theta g_i(X_1)$ . Рассмотрим функции

$$m_i(\theta) = \mathbf{E}_\theta g_i(X_1), \quad i \in \{1, \dots, k\}.$$

Положим  $m = (m_1, \dots, m_k)^T$ . Составим систему уравнений относительно  $\theta$ :

$$\begin{cases} m_1(\theta) = \overline{g_1(X)}, \\ \dots, \\ m_k(\theta) = \overline{g_k(X)}. \end{cases}$$

Предположим, что у этой системы существует единственное решение.

**▲ Определение 3.1.** Единственное решение такой системы  $\theta^*(X)$  называется *оценкой по методу моментов* с пробными функциями  $g_1, \dots, g_k$ . Функции  $g_i(x) = x^i$ ,  $i \in \{1, \dots, k\}$ , называются *стандартными пробными функциями*.

**Теорема 3.1** Если  $m : \Theta \rightarrow m(\Theta)$  — биекция и функцию  $m^{-1}$  можно доопределить до функции, заданной на всем множестве  $\mathbb{R}^k$  и непрерывной в каждой точке множества  $m(\Theta)$ , то оценка по методу моментов является сильно состоятельной.

Более того, утверждение этой теоремы можно дополнить следующим образом.

▲ **Задача 3.1.** Докажите, что в условиях теоремы 3.1 оценка по методу моментов является асимптотически нормальной, если функцию  $m^{-1}$  можно доопределить до функции, заданной на всем множестве  $\mathbb{R}^k$  и дифференцируемой в каждой точке множества  $m(\Theta)$ , и, кроме того, для любого  $i \in \{1, \dots, k\}$  и любого  $\theta \in \Theta$  выполнено неравенство  $0 < D_{\theta} g_i(X_1) < +\infty$ .

### Решение

Итак, оценка по методу моментов равна  $\theta^* = m^{-1}(\overline{g(X)})$ , где  $\overline{g(X)} = (\overline{g_1(X)}, \dots, \overline{g_k(X)})^T$ . По многомерной центральной предельной теореме, для любого  $\theta \in \Theta$  выполнено

$$\sqrt{n} \left( \overline{g(X)} - m(\theta) \right) \xrightarrow{d_q} \xi \sim \mathcal{N}(0, \Sigma(\theta)),$$

где  $\Sigma(\theta)$  — матрица ковариаций вектора  $(g_1(X_1), \dots, g_k(X_1))^T$ . По теореме 1.4, выполнено

$$\begin{aligned} \sqrt{n} \left( m^{-1} \left( \overline{g(X)} \right) - \theta \right) &\xrightarrow{d_q} (\xi, \nabla m^{-1}|_{m(\theta)}) \sim \\ &\sim \mathcal{N} \left( 0, (\nabla m^{-1}|_{m(\theta)})^T \Sigma(\theta) (\nabla m^{-1}|_{m(\theta)}) \right) \end{aligned}$$

(здесь мы положили  $h = m^{-1}$ ,  $b_n = 1/\sqrt{n}$ ,  $a = m(\theta)$ ,  $\xi_n = \sqrt{n}(\overline{g(X)} - m(\theta))$ ).

**Задача решена.**

▲ **Задача 3.2.** Найти оценки по методу моментов со стандартными пробными функциями для семейства распределений  $\mathcal{N}(a, \sigma^2)$ .

### Решение

Здесь  $\theta = (a, \sigma^2)$ ,  $\Theta = \mathbb{R} \times (0, \infty)$ . Выберем стандартные пробные функции  $g_1(x) = x$  и  $g_2(x) = x^2$ . Тогда

$$\mathbb{E}_\theta g_1(X_1) = \mathbb{E}_\theta X_1 = a,$$

$$\mathbb{E}_\theta g_2(X_1) = \mathbb{E}_\theta X_1^2 = D_\theta X_1 + (\mathbb{E}_\theta X_1)^2 = \sigma^2 + a^2.$$

Тем самым, получаем систему:

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

решая которую, получаем ответ  $\theta^* = (\overline{X}, s^2)$ .

**Задача решена.**

Заметим, что оценка по методу моментов *не обязательно является несмещённой*.

## 3.2. Метод максимального правдоподобия

Пусть  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  — параметрическое семейство распределений. Если все  $P_\theta$  — абсолютно непрерывные распределения, то положим  $p_\theta(x)$  равной плотности распределения  $P_\theta$ . Если же все  $\{P_\theta, \theta \in \Theta\}$  — дискретные распределения, то положим  $p_\theta(x) = P_\theta(X = x)$ . В обоих случаях будем называть семейство  $\mathcal{P}$  *доминируемым*.

▲ **Определение 3.2.** Пусть  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  — доминируемое семейство распределений,  $X_1, \dots, X_n$  — выборка, имеющая распределение из этого семейства. *Функцией правдоподобия* выборки  $X_1, \dots, X_n$  называется случайная величина

$$f_\theta(X_1, \dots, X_n) = p_\theta(X_1) \dots p_\theta(X_n).$$

Величина

$$L_\theta(X_1, \dots, X_n) = \ln f_\theta(X_1, \dots, X_n)$$

называется *логарифмической функцией правдоподобия*. Оценкой максимального правдоподобия (ОМП) параметра  $\theta$  называется

$$\hat{\theta}(X_1, \dots, X_n) = \arg \max_{\theta \in \Theta} f_{\theta}(X_1, \dots, X_n),$$

т.е. то значение  $\theta \in \Theta$ , при котором достигается максимум функции правдоподобия при фиксированных  $X_1, \dots, X_n$ .

**▲ Задача 3.3.** Пусть  $X_1, \dots, X_n$  — выборка из распределения  $\text{Bern}(\theta)$ . Найдите оценку максимального правдоподобия параметра  $\theta$ .

### Решение

Имеем

$$\begin{aligned} p_{\theta}(x) = \mathbf{P}_{\theta}(X = x) &= \begin{cases} \theta, & \text{если } x = 1, \\ 1 - \theta, & \text{если } x = 0 \end{cases} = \\ &= \theta^{I\{x=1\}}(1 - \theta)^{I\{x=0\}} = \theta^x(1 - \theta)^{1-x} I\{x \in \{0, 1\}\}. \end{aligned}$$

Следовательно, функция правдоподобия равна

$$f_{\theta}(X_1, \dots, X_n) = \theta^{\sum X_i} (1 - \theta)^{n - \sum X_i},$$

а логарифмическая функция правдоподобия (которую, как правило, легче дифференцировать, чтобы найти точку максимума) равна

$$L_{\theta}(X_1, \dots, X_n) = \sum X_i \ln \theta + \left(n - \sum X_i\right) \ln(1 - \theta).$$

Так как

$$\frac{\partial L_{\theta}(X_1, \dots, X_n)}{\partial \theta} = \frac{\sum X_i}{\theta} - \frac{n - \sum X_i}{1 - \theta},$$

то максимум  $f_{\theta}(X_1, \dots, X_n)$  достигается в точке  $\theta^* = \bar{X}$ , в которой производная  $\frac{\partial L_{\theta}(X_1, \dots, X_n)}{\partial \theta}$  обращается в ноль.

**Задача решена.**

Обратимся, наконец, к свойствам оценки максимального правдоподобия.



**Теорема 3.2** Пусть  $\Theta \subset \mathbb{R}$ , выполнены условия регулярности (см., например, [2], §3.2), функция правдоподобия имеет лишь один локальный максимум, лежащий внутри  $\Theta$  и совпадающий с оценкой максимального правдоподобия, и, наконец, функция правдоподобия трижды дифференцируема по  $\theta$  и при этом существует не зависящая от  $\theta$  функция  $M(x)$  такая, что для всех  $\theta \in \Theta$

$$\left| \frac{\partial^3 p_\theta(x)}{\partial \theta^3} \right| \leq M(x), \quad \mathbb{E}_\theta M(X_1) < \infty.$$

Тогда оценка максимального правдоподобия является асимптотически нормальной оценкой  $\theta$  с асимптотической дисперсией  $n \left( D_\theta \frac{\partial L_\theta}{\partial \theta} \right)^{-1}$ .

### 3.3. Метод выборочных квантилей

**▲ Определение 3.3.** Пусть  $F(x)$  — функция распределения на  $\mathbb{R}$ . Пусть, кроме того,  $p \in (0, 1)$ . Тогда  $p$ -квантилью функции распределения  $F(x)$  называют величину

$$z_p = \inf\{x : F(x) \geq p\}.$$

**▲ Определение 3.4.** Пусть  $X_1, \dots, X_n$  — выборка из неизвестного распределения  $P$ . Статистику

$$z_{n,p} = \begin{cases} X_{([np]+1)}, & \text{если } np \notin \mathbb{Z}; \\ X_{(np)}, & \text{если } np \in \mathbb{Z} \end{cases}$$

называют выборочной  $p$ -квантилью.

**Теорема 3.3** Пусть  $X_1, \dots, X_n$  — выборка из абсолютно непрерывного одномерного распределения  $P$  с плотностью  $f$ . Пусть  $z_p$  —  $p$ -квантиль распределения  $P$ , причём функция  $f$  непрерывно дифференцируема в некоторой окрестности  $z_p$  и  $f(z_p) > 0$ . Тогда (об асимптотической нормальности выборочной квантили)

$$\sqrt{n}(z_{n,p} - z_p) \xrightarrow{d} \xi \sim \mathcal{N}\left(0, \frac{1}{4f^2(z_p)}\right).$$

**▲ Определение 3.5.** Медианой распределения называют его  $\frac{1}{2}$ -квантиль.

**▲ Определение 3.6.** Выборочной медианой выборки  $X_1, \dots, X_n$  называют статистику

$$\hat{\mu} = \begin{cases} X_{(k+1)}, & \text{если } n = 2k + 1; \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & \text{если } n = 2k. \end{cases}$$

Асимптотическое поведение  $\hat{\mu}$  схоже с поведением  $z_{n,1/2}$  : в условиях теоремы о выборочной квантили

$$\sqrt{n}(\hat{\mu} - z_{1/2}) \xrightarrow{d} \xi \sim \mathcal{N}\left(0, \frac{1}{4f^2(z_{1/2})}\right)$$

(этот факт носит название теоремы о выборочной медиане).

**▲ Задача 3.4.** Пусть  $X_1, \dots, X_n$  — выборка из распределения Коши со сдвигом  $\theta$ . Найдите асимптотически нормальную оценку параметра  $\theta$  и ее асимптотическую дисперсию.

## Решение

Плотность распределения Коши имеет вид

$$p_{\theta}(x) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$

Так как у случайной величины с таким распределением нет конечного математического ожидания, то воспользоваться центральной предельной теоремой нам не удастся. Очевидно,  $z_{1/2} = \theta$ . Поэтому, по теореме об асимптотической нормальности выборочной медианы,  $\hat{\mu}$  — асимптотически нормальная оценка  $\theta$  с асимптотической дисперсией

$$\frac{1}{4p_{\theta}^2(\theta)} = \frac{\pi^2}{4}.$$

**Задача решена.**

## ▲ Задачи для самостоятельного решения

1. Найдите оценки по методу моментов для следующих распределений: а)  $\mathcal{N}(a, \sigma^2)$ , б)  $\Gamma(\alpha, \lambda)$ , в)  $R(a, b)$ , г)  $\text{Pois}(\lambda)$ , д)  $\text{Bin}(m, p)$ , е)  $\text{Geom}(p)$ , ж)  $\text{Beta}(\lambda_1, \lambda_2)$ .

*Замечание.* Напомним, что плотность распределения  $\Gamma(\alpha, \lambda)$  с параметрами  $\alpha > 0, \lambda > 0$  равна

$$p_{\alpha, \lambda}(x) = \frac{\alpha^{\lambda} x^{\lambda-1} e^{-\alpha x}}{\Gamma(\lambda)} I(x \geq 0),$$

где гамма-функция  $\Gamma(\lambda)$  определяется следующим образом:  $\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} e^{-x} dx$ . Стоит напомнить также важное свойство, которым обладает гамма-функция: для любого  $\lambda > 0$  выполнено  $\Gamma(\lambda + 1) = \lambda \Gamma(\lambda)$ . В этом смысле гамма-функция является обобщением факториала.

Плотность распределения  $\text{Beta}(\lambda_1, \lambda_2)$  с параметрами  $\lambda_1 > 0, \lambda_2 > 0$  равна

$$p_{\lambda_1, \lambda_2}(x) = \frac{x^{\lambda_1-1} (1-x)^{\lambda_2-1}}{B(\lambda_1, \lambda_2)} I(0 \leq x \leq 1),$$

где бета-функция  $B(\lambda_1, \lambda_2)$  определяется следующим образом:

$$B(\lambda_1, \lambda_2) = \int_0^1 x^{\lambda_1-1} (1-x)^{\lambda_2-1} dx = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1 + \lambda_2)}.$$

2. Найдите оценки по методу максимального правдоподобия для следующих распределений: а)  $\mathcal{N}(a, \sigma^2)$  в трех случаях: когда неизвестен только один из параметров и когда неизвестны оба параметра; б)  $\Gamma(\alpha, \lambda)$ , если параметр  $\lambda$  известен; в)  $R(a, b)$ ; г)  $\text{Pois}(\lambda)$ ; д)  $\text{Bin}(m, p)$ , если параметр  $m$  известен; е)  $\text{Geom}(p)$ , ж) для распределения с плотностью  $p(x) = \frac{2\theta^2}{x^3} I\{x > \theta\}$ .
3.  $X_1, \dots, X_n$  — выборка из распределения с плотностью

$$p_{\alpha, \beta}(x) = \frac{1}{\alpha} e^{(\beta-x)/\alpha} I_{[\beta, +\infty)}(x).$$

где  $\theta = (\alpha, \beta)$  — двумерный параметр. Найдите для  $\theta$  оценку максимального правдоподобия. Докажите, что полученная для  $\alpha$  оценка  $\hat{\alpha}_n$  является асимптотически нормальной, и найдите ее асимптотическую дисперсию.

4. Найдите оценку максимального правдоподобия для параметра сдвига в распределении Коши, т.е. плотность равна

$$p_{\theta}(x) = \frac{1}{\pi(1 + (x - \theta)^2)},$$

если выборка состоит из а) одного наблюдения, б) двух наблюдений (т.е.  $n = 1, 2$ ).

5. Предложите асимптотически нормальную оценку параметра  $\theta^2$  в модели распределения Коши со сдвигом (см. задачу 3.4), а также найдите её асимптотическую дисперсию.
6. Найдите оценки по методу моментов для следующих распределений: а) распределения Парето  $P(\gamma)$  с плотностью

$$p(x) = \gamma x^{-\gamma-1} I\{x > 1\},$$

б) распределение Коши  $C(\theta)$  с плотностью

$$\frac{\theta}{\pi(x^2 + \theta^2)}.$$

7. Пусть  $X_1 \sim R(0, \theta)$ . Найдите несмещённую оценку параметра  $1/\theta$ .
8. Найдите несмещённую оценку  $\lambda^3$  по выборке  $X_1, \dots, X_n$  из распределения  $\text{Pois}(\lambda)$ .
9. Пусть  $X_1, \dots, X_n$  — выборка,  $X_1 = e^\xi$ , где  $\xi$  имеет равномерное распределение на  $[0, \theta]$ . Найдите оценку максимального правдоподобия параметра  $\theta$  и проверьте ее на состоятельность.
10. Дана выборка  $X_1, \dots, X_n$  из распределения Лапласа с плотностью

$$p_\theta(x) = \frac{1}{2} e^{-|x-\theta|}.$$

Найдите оценку параметра  $\theta$  методом максимального правдоподобия.

## 4. Сравнение оценок и эффективные оценки

Во второй главе мы рассмотрели свойства оценок, а в третьей главе научились находить оценки с такими свойствами. Возникает естественный вопрос: какая из двух оценок, обладающих одинаковыми свойствами (например, состоятельностью, несмещенностью или асимптотической нормальностью), лучше? В этой главе мы поговорим о сравнении оценок.

Пусть  $\Theta \subset \mathbb{R}$ . Пусть, кроме того,  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  — некоторая функция (называемая *функцией потерь*, как правило, симметричная), принимающая только положительные значения (например,  $g(x, y) = |x - y|$  или  $g(x, y) = (x - y)^2$ ). Функция  $g(x, y) = (x - y)^2$  называется *квадратичной функцией потерь*.

**▲ Определение 4.1.** Пусть  $\theta^*$  — оценка параметра  $\theta$ . Функция  $R(\theta^*, \theta) = E_{\theta}g(\theta^*, \theta)$  называется *функцией риска* оценки  $\theta^*$ . Говорят, что оценка  $\theta^*$  лучше оценки  $\hat{\theta}$  в *равномерном подходе с функцией потерь  $g$* , если для любого  $\theta \in \Theta$  выполнено  $R(\theta^*, \theta) \leq R(\hat{\theta}, \theta)$  и для некоторого  $\theta \in \Theta$  неравенство строгое. Если  $\theta^*$  лучше всех других оценок в равномерном подходе с функцией потерь  $g$  в некотором классе оценок  $\mathcal{K}$  параметра  $\theta$  (например, в классе всех несмещенных оценок), то оценку  $\theta^*$  называют *наилучшей в классе  $\mathcal{K}$  в равномерном подходе с функцией потерь  $g$* . Равномерный подход с квадратичной функцией потерь называется *среднеквадратичным*.

Если  $\Theta \subset \mathbb{R}^k$ , где  $k > 1$ , то оценка  $\theta^*$  лучше оценки  $\hat{\theta}$  в среднеквадратичном подходе, если для любых  $\theta \in \Theta$  и  $a \in \mathbb{R}^k$  выполнено

$$E_{\theta}(\langle \theta^* - \theta, a \rangle)^2 \leq E_{\theta}(\langle \hat{\theta} - \theta, a \rangle)^2$$

и для некоторых  $\theta \in \Theta$  и  $a \in \mathbb{R}^k$  неравенство строгое.

**▲ Задача 4.1.** Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на  $[0, \theta]$ . Найдите наилучшую оценку в  $\mathcal{K} = \{cX_{(1)}, c \in \mathbb{R}\}$  в среднеквадратичном подходе.

## Решение

Напомним, что  $\mathbf{E}X_{(1)} = \frac{\theta}{n+1}$ . Действительно, плотность случайной величины  $X_{(1)}$  равна

$$p(x) = -(\mathbf{P}(X_{(1)} > x))' = n \frac{(\theta - x)^{n-1}}{\theta^n}.$$

Искомое значение математического ожидания получается интегрированием функции  $x p(x)$  на отрезке  $[0, \theta]$ .

Более того,

$$\begin{aligned} \mathbf{E}X_{(1)}^2 &= \int_0^\theta x^2 n \frac{(\theta - x)^{n-1}}{\theta^n} dx = \int_0^\theta (x - \theta)^2 n \frac{x^{n-1}}{\theta^n} dx = \\ &= \int_0^\theta \left( \frac{nx^{n+1}}{\theta^n} - 2 \frac{nx^n}{\theta^{n-1}} + \frac{nx^{n-1}}{\theta^{n-2}} \right) dx = \\ &= \frac{n\theta^2}{n+2} - \frac{2n\theta^2}{n+1} + \theta^2 = \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned}$$

Тогда наилучшая оценка равна  $cX_{(1)}$ , где

$$\begin{aligned} c &= \operatorname{argmin}(\mathbf{E}_\theta(cX_{(1)} - \theta)^2) = \\ &= \operatorname{argmin} \left( \frac{2\theta^2}{(n+1)(n+2)} c^2 - \frac{2\theta^2}{n+1} c + \theta^2 \right) = \frac{n+2}{2}. \end{aligned}$$

**Задача решена.**

В последней задаче, разумеется, лишь одна оценка в классе является несмещенной (при  $c = n+1$ ). Отдельной важной задачей является нахождение наилучшей оценки в классе несмещенных оценок. При некоторых условиях эту задачу удастся решить.

Итак, пусть  $X = (X_1, \dots, X_n)$  — выборка из неизвестного распределения  $\mathbf{P}_\theta$ ,  $\theta \in \Theta$ . Пусть, кроме того, существует функция правдоподобия  $f_\theta(X)$ . Будем считать, что для рассматриваемой параметрической модели выполнены условия регулярности (см., например, [2], §3.2).

▲ **Определение 4.2.** Величина

$$u_\theta(X) = \frac{\partial}{\partial \theta} L_\theta(X)$$

называется *вкладом* выборки  $X$ , а величина

$$I_X(\theta) = \mathbb{E}_\theta(u_\theta(X))^2$$

— *количеством информации (по Фишеру)*, содержащейся в выборке  $X$ .

Пусть  $i(\theta)$  — количество информации, содержащейся в одном элементе выборки  $X_1$ .

▲ **Задача 4.2.** Покажите, что  $I_X(\theta) = ni(\theta)$ .

**Решение**

По условию регулярности, интеграл от  $S(x)p_\theta(x)$  можно дифференцировать под знаком интеграла. В частности,

$$\int_{\mathcal{X}} \frac{\partial}{\partial \theta} p_\theta(x) \mu(dx) = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} p_\theta(x) \mu(dx) = 0.$$

Здесь и далее в подобных случаях мы обозначаем  $\mathbb{E}X =: \int_{\mathcal{X}} xp(x) \mu(dx)$ , имея в виду интеграл Римана в случае абсолютно непрерывного распределения случайной величины  $X$  с плотностью  $p$  и сумму (возможно, бесконечную) в случае дискретного распределения случайной величины  $X$  с  $\mathbb{P}(X = x) = p(x)$ .

Таким образом,

$$\mathbb{E}_\theta \frac{\partial}{\partial \theta} \ln p_\theta(X_1) = \int_{\mathcal{X}} \frac{\partial p_\theta(x)}{\partial \theta} \frac{p_\theta(x)}{p_\theta(x)} \mu(dx) = 0.$$



Следовательно,

$$\begin{aligned} I_X(\theta) &= D_\theta \left( \frac{\partial}{\partial \theta} L_\theta(X) \right) = D_\theta \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right) = \\ &= \sum_{i=1}^n D_\theta \left( \frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right) = ni(\theta). \end{aligned}$$

**Задача решена.**

Следующая теорема дает нижнюю оценку на значение функции риска (дисперсии) для наилучшей оценки в средне-квадратичном подходе в классе всех несмещенных оценок.

**Теорема 4.1** *Рассматривается выборка для параметрической модели, на которую наложены условия регулярности. Пусть  $\theta^*$  — несмещенная оценка  $\tau(\theta)$ . Тогда для любого  $\theta \in \Theta$*

$$D_\theta \theta^* \geq \frac{(\tau'(\theta))^2}{ni(\theta)}.$$

Оказывается, в некоторых случаях эта оценка является точной (т.е. для наилучшей оценки в неравенстве Рао-Крамера достигается равенство).

**Теорема 4.2** *Равенство в неравенстве Рао-Крамера достигается тогда и только тогда, когда  $\theta^* - \tau(\theta) = c(\theta)u_\theta(X)$  для некоторого  $c(\theta)$ . Более того, последнее равенство выполнено в том и только том случае, когда  $c(\theta) = \frac{\tau'(\theta)}{ni(\theta)}$ .*

Заметим, что равенство в неравенстве Рао-Крамера может быть выполнено только для одной несмещенной оценки, и эта

оценка является наилучшей в среднеквадратичном подходе в классе всех несмещенных оценок  $\tau(\theta)$ .

▲ **Определение 4.3.** Если в неравенстве Рао-Крамера достигается равенство, то  $\theta^*$  называется *эффективной оценкой*  $\tau(\theta)$ .

▲ **Задача 4.3.** Пусть  $X_1, \dots, X_n$  — выборка из распределения Бернулли с параметром  $\theta$ . Найдите эффективную оценку  $\theta$  и информацию  $i(\theta)$ .

### Решение

Для рассматриваемой в задаче параметрической модели выполнены условия регулярности. Функция правдоподобия равна

$$f_\theta(X) = \theta^{\sum X_i} (1 - \theta)^{n - \sum X_i}.$$

Тогда

$$\begin{aligned} u_\theta(X) &= \frac{\partial}{\partial \theta} \ln f_\theta(X) = \frac{\sum X_i}{\theta} - \frac{n - \sum X_i}{1 - \theta} = \\ &= \frac{\sum X_i - n\theta}{\theta(1 - \theta)} = \frac{n}{\theta(1 - \theta)} (\bar{X} - \theta). \end{aligned}$$

По критерию эффективности, получаем, что  $\bar{X}$  — эффективная оценка  $\theta$ . Кроме того,

$$\frac{\theta(1 - \theta)}{n} = \frac{1}{ni(\theta)}.$$

Следовательно,  $i(\theta) = \frac{1}{\theta(1 - \theta)}$ .

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[0, \theta]$ . Сравните следующие оценки параметра  $\theta$  в равномерном подходе с квадратичной функцией по-

теперь:  $2\bar{X}$ ,  $(n+1)X_{(1)}$ ,  $\frac{n+1}{n}X_{(n)}$ .

2. Пусть  $\theta_1^*(X)$  и  $\theta_2^*(X)$  — две наилучшие в среднеквадратичном подходе оценки параметра  $\theta$  в классе всех оценок с одним и тем же математическим ожиданием  $\tau(\theta)$ . Докажите, что тогда для любого  $\theta$  они совпадают почти наверное, т.е.  $\theta_1^*(X) = \theta_2^*(X)$  (Р $_{\theta}$ -п.н.).
3. Пусть  $X_1, \dots, X_n$  — выборка из биномиального распределения с параметрами  $(m, p)$ , причем  $m$  известно. Найдите информацию Фишера  $i(p)$  в данной модели, а также эффективную оценку параметра  $p$ .
4. Пусть  $X_1, \dots, X_n$  — выборка из экспоненциального распределения с параметром  $\theta$ . Для каких функций  $\tau(\theta)$  существует эффективная оценка? Вычислите информацию Фишера  $i(\theta)$  одного наблюдения в данной модели.
5. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(a, \sigma^2)$ . Найдите эффективную оценку
  - а) параметра  $a$ , если  $\sigma$  известно;
  - б) параметра  $\sigma^2$ , если  $a$  известно.Вычислите информацию Фишера одного наблюдения в обоих случаях.
6. Пусть  $X_1, \dots, X_n$  — выборка из логистического распределения со сдвигом  $\theta$ , имеющего плотность

$$p_{\theta}(x) = \frac{\exp\{\theta - x\}}{(1 + \exp\{\theta - x\})^2}.$$

Найдите информацию Фишера  $i(\theta)$  одного наблюдения в этой модели.

7. Пусть  $X_1, \dots, X_n$  — выборка из равномерного закона на отрезке  $[0, \theta]$ . Вычислите информацию Фишера  $i(\theta)$ , заключенную в статистике  $X_{(n)}$ .
8. Пусть  $X_1, \dots, X_n$  — выборка из распределения Гумбеля с функцией распределения

$$F_{\theta}(x) = e^{-e^{-\theta x}}, \quad \theta > 0.$$

Для каких функций  $\tau(\theta)$  существует эффективная оценка?

9.  $X_1, \dots, X_n$  — выборка из распределения  $\mathcal{N}(\theta, \theta^2)$ . Для каких функций  $\tau(\theta)$  существует эффективная оценка? Вычислите информацию Фишера  $i(\theta)$  одного элемента выборки.

## 5. Условное математическое ожидание

В этой главе мы второй раз отступим от решения статистических задач, чтобы определить еще один важный инструмент, относящийся к курсу теории вероятностей. В стандартном курсе теории вероятностей определяются понятия условной вероятности и (безусловного) математического ожидания. Напомним, что функция  $P(\cdot|B)$  является вероятностной мерой, и поэтому функция  $P(\xi = x|B)$  в случае дискретной случайной величины  $\xi$  и произвольного события  $B$  ненулевой вероятности является дискретным распределением вероятностей. Можно для такого “условного” распределения естественным образом определить “условное” математическое ожидание:  $E(\xi|B) = \sum_x xP(\xi = x|B)$ . Для произвольной дискретной случайной величины  $\eta$  рассмотрим событие  $B = \{\eta = y\}$  и определим  $E(\xi|\eta = y) := E(\xi|\{\eta = y\})$ . Последний объект играет огромную роль в теории оценивания неизвестных параметров и представляет из себя, грубо говоря, среднее значение случайной величины  $\xi$  при условии, что случайная величина  $\eta$  равна  $y$ . В общем случае (для случайных величин  $\xi, \eta$  с произвольными распределениями) оказывается проще определить сначала условное математическое ожидание, а потом, с помощью него, — условное распределение. Ниже мы дадим определение условного математического ожидания в общем случае и поговорим о его свойствах.

**▲ Определение 5.1.** Случайная величина  $\eta$  ( $= E(\xi|\mathcal{G})$ ) называется *условным математическим ожиданием* случайной величины  $\xi$  относительно сигма-алгебры  $\mathcal{G}$ , если

- 1) выполнено *свойство измеримости*:  $\eta$  является  $\mathcal{G}$ —измеримой случайной величиной;
- 2) выполнено *интегральное свойство*: для любого события  $A \in \mathcal{G}$  справедливо равенство

$$\int_A \xi(\omega)P(d\omega) = \int_A \eta(\omega)P(d\omega)$$

(или, другими словами,  $E(\xi I_A) = E(\eta I_A)$ ).

**Теорема 5.1** Если  $E|\xi| < \infty$ , то условное математическое ожидание  $E(\xi|\mathcal{G})$  существует и единственно (п.н.).

▲ **Определение 5.2.** Мера

$$P_\xi(B|\mathcal{G}) = E(I(\xi \in B)|\mathcal{G})$$

называется *условным распределением* случайной величины  $\xi$  относительно сигма-алгебры  $\mathcal{G}$ .

Кроме того, *условное математическое ожидание* случайной величины  $\xi$  относительно случайной величины  $\eta$  равно

$$E(\xi|\eta) = E(\xi|\sigma(\eta)),$$

где  $\sigma(\eta)$  —  $\sigma$ -алгебра, порожденная случайной величиной  $\eta$  (т.е.  $\sigma(\eta) = \{\eta^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ ).

Наконец, *условное распределение* случайной величины  $\xi$  относительно  $\eta$  на множестве  $B$  равно  $P_\xi(B|\sigma(\eta))$ .

Заметим, что для любого борелевского множества значение условного распределения на этом множестве является случайной величиной.

Теперь выясним, как выглядит условное математическое ожидание в простейшем (дискретном) случае, с помощью которого мы в начале главы мотивировали понятие условного математического ожидания. Пусть  $N$  — некоторое подмножество (возможно, конечное) множества натуральных чисел. *Разбиением* множества  $\Omega$  будем называть систему непересекающихся подмножеств  $\Omega$ , дизъюнктное объединение которых равно всему множеству  $\Omega$ .

▲ **Задача 5.1.** Пусть  $\mathcal{G}$  — сигма-алгебра, порождённая разбиением  $\{D_n, n \in N\}$  (наименьшая  $\sigma$ -алгебра, содержащая все множества  $D_n, n \in N$ ), причём для всех  $n \in N$  имеем  $P(D_n) > 0$ . Тогда

$$E(\xi|\mathcal{G}) = \sum_{n \in N} \frac{E(\xi I_{D_n})}{P(D_n)} I_{D_n},$$

т.е. условное математическое ожидание на произвольном  $D_n$  равно усреднению случайной величины  $\xi$  на этом множестве.

### Решение

По определению условного математического ожидания  $\eta = E(\xi|\mathcal{G})$ , выполнено свойство измеримости:  $\eta$  измерима относительно сигма-алгебры  $\mathcal{G}$ . Следовательно,

$$\eta = \sum_{n \in N} c_n I_{D_n}.$$

Действительно, если бы, скажем, на множестве  $D_1$  случайная величина  $\eta$  принимала два различных значения  $a \neq b$ , то, очевидно, множество  $\{\omega : \eta(\omega) = b\}$  не принадлежало бы  $\mathcal{G}$ , и тогда случайная величина  $\eta$  не была бы  $\mathcal{G}$ -измеримой.

Проверим теперь выполнение интегрального свойства, которое, в силу линейности интеграла Лебега, достаточно проверить только для множеств  $D_n$ ,  $n \in N$ :

$$\begin{aligned} E(\xi I_{D_n}) &= E(\eta I_{D_n}) = E\left(I_{D_n} \sum_{k \in N} c_k I_{D_k}\right) = \\ &= c_n E I_{D_n} = c_n P(D_n), \end{aligned}$$

откуда

$$c_n = \frac{E \xi I_{D_n}}{P(D_n)}.$$

**Задача решена.**

### ▲ Основные свойства условного математического ожидания:

1. Если  $\xi$  —  $\mathcal{G}$ -измеримая случайная величина, то  $E(\xi|\mathcal{G}) = \xi$  (п.н.).
2. (Линейность) Если  $a, b$  — постоянные и  $E|\xi| < \infty$ ,  $E|\eta| < \infty$ , то

$$E(a\xi + b\eta|\mathcal{G}) = aE(\xi|\mathcal{G}) + bE(\eta|\mathcal{G}) \quad (\text{п.н.}).$$

3. Если  $\xi \leq \eta$  (п.н.) и  $E|\xi| < \infty$ ,  $E|\eta| < \infty$ , то и  $E(\xi|\mathcal{G}) \leq E(\eta|\mathcal{G})$  (п.н.).

4. Если случайная величина  $\xi$  с конечным математическим ожиданием не зависит от сигма-алгебры  $\mathcal{G}$  (т.е. сигма-алгебры  $\sigma(\xi)$  и  $\mathcal{G}$  независимы), то  $E(\xi|\mathcal{G}) = E\xi$  (п.н.).

5. (Телескопическое свойство) Пусть  $\mathcal{G}_1 \subset \mathcal{G}_2$  и  $E|\xi| < \infty$  тогда

$$E(E(\xi|\mathcal{G}_1)|\mathcal{G}_2) = E(\xi|\mathcal{G}_1) \quad (\text{п.н.}),$$

$$E(E(\xi|\mathcal{G}_2)|\mathcal{G}_1) = E(\xi|\mathcal{G}_1) \quad (\text{п.н.}).$$

6. (Формула полной вероятности) Если  $E|\xi| < \infty$ , то

$$E(E(\xi|\mathcal{G})) = E\xi.$$

7. Если  $\xi$  —  $\mathcal{G}$ -измерима,  $E|\xi\eta| < \infty$ ,  $E|\eta| < \infty$ , то

$$E(\xi\eta|\mathcal{G}) = \xi E(\eta|\mathcal{G}) \quad (\text{п.н.}).$$

8. (Аналог теоремы Лебега) Пусть  $\xi_n \xrightarrow{\text{п.н.}} \xi$ ,  $|\xi_n| \leq \eta$  для всех  $n \in \mathbb{N}$  и  $E\eta < \infty$ . Тогда

$$E(\xi_n|\mathcal{G}) \xrightarrow{\text{п.н.}} E(\xi|\mathcal{G}),$$

$$E(|\xi_n - \xi| |\mathcal{G}) \xrightarrow{\text{п.н.}} 0.$$

**▲ Задача 5.2.** Пусть  $\xi, \eta$  — независимые случайные величины, распределённые по закону  $\mathcal{N}(3, 4)$ . Найдите  $E((\xi - 3\eta)^2|\eta)$ .



## Решение

В силу линейности условного математического ожидания,

$$\mathbb{E}((\xi - 3\eta)^2|\eta) = \mathbb{E}(\xi^2|\eta) - 6\mathbb{E}(\xi\eta|\eta) + 9\mathbb{E}(\eta^2|\eta) \quad (\text{п.н.}).$$

Поскольку  $\xi$  и  $\eta$  независимы, то  $\xi^2$  не зависит от  $\sigma(\eta)$  и, по свойству 4,

$$\mathbb{E}(\xi^2|\eta) = \mathbb{E}\xi^2 = D\xi + (\mathbb{E}\xi)^2 = 13 \quad (\text{п.н.}).$$

По свойствам 4 и 7,

$$\mathbb{E}(\xi\eta|\eta) = \eta\mathbb{E}(\xi|\eta) = \eta\mathbb{E}\xi = 3\eta \quad (\text{п.н.}),$$

далее, по свойству 1,

$$\mathbb{E}(\eta^2|\eta) = \eta^2 \quad (\text{п.н.}).$$

Окончательно получаем

$$\mathbb{E}((\xi - 3\eta)^2|\eta) = 9\eta^2 - 18\eta + 13 \quad (\text{п.н.}).$$

## Задача решена.

Заметим, что в ответе мы получили функцию от  $\eta$ , что неудивительно, так как  $\mathbb{E}(\cdot|\eta)$  является  $\eta$ -измеримой случайной величиной, а, как мы помним из курса теории вероятностей, случайная величина  $X$  является  $\eta$ -измеримой тогда и только тогда, когда существует такая борелевская функция  $f$ , что  $X = f(\eta)$ .

▲ **Задача 5.3.** Пусть  $(\xi, \eta)$  — гауссовский вектор с вектором математических ожиданий  $(1, -1)$  и матрицей ковариаций

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

Найдите  $\mathbb{E}(\xi - \eta|\xi + \eta)$ .

## Решение

Известно, что если две компоненты гауссовского вектора не коррелированы (ковариация равна нулю), то они независимы. Пусть требуется найти  $E(X|Y)$ , где  $X$  и  $Y$  — компоненты гауссовского вектора. Метод решения подобных задач заключается в том, чтобы представить  $X$  как  $\alpha Y + Z$ , где случайная величина  $Z$  распределена нормально и не зависит от  $Y$ . Тогда

$$E(X|Y) = E(\alpha Y + Z|Y) = \alpha Y + EZ \quad (\text{п.н.}).$$

Итак, найдём такое  $\alpha$ , что  $(\xi - \eta) - \alpha(\xi + \eta)$  не зависит от  $\xi + \eta$ , т.е.  $\text{cov}((\xi - \eta) - \alpha(\xi + \eta), \xi + \eta) = 0$ .

$$\begin{aligned} \text{cov}((\xi - \eta) - \alpha(\xi + \eta), \xi + \eta) &= \\ &= (1 - \alpha)\text{cov}(\xi, \xi) - (1 + \alpha)\text{cov}(\eta, \eta) - 2\alpha\text{cov}(\xi, \eta) = \\ &= 2(1 - \alpha) - 3(1 + \alpha) - 2\alpha = 0. \end{aligned}$$

Следовательно,  $\alpha = -\frac{1}{7}$ . Тогда легко записать ответ:

$$\begin{aligned} E(\xi - \eta|\xi + \eta) &= \\ &= -\frac{\xi + \eta}{7} + E((\xi - \eta) + \frac{1}{7}(\xi + \eta)) = 2 - \frac{\xi + \eta}{7} \quad (\text{п.н.}). \end{aligned}$$

**Задача решена.**

## ▲ Задачи для самостоятельного решения

1. Пусть  $X$  — число очков при случайном броске шестигранного кубика. Чему равняется  $E(X|(X - 2)^2)$ ?
2. Из урны, содержащей 2 черных, 4 красных и 2 белых шара вытаскиваются с возвращением шары по одному  $n$  раз. Пусть  $X$  — количество черных, а  $Y$  — количество красных шаров, появившихся в этих опытах. Найдите  $E(X|X + Y)$ .
3. Пусть  $(X, Y)$  — гауссовский вектор,  $(X, Y) \sim N(a, \Sigma)$ . Найдите  $E(X|Y)$  и  $E(X|X + Y)$ .

4. Случайная величина  $X$  имеет равномерное распределение на отрезке  $[-1, 1]$ . Вычислите  $E(X|X^2)$ .
5. Совместное распределение случайного вектора  $(\xi, \eta)$  таково:

$$P(\xi = k, \eta = l) = p^2(1 - p)^{k+l-2},$$

где  $k, l \in \mathbb{N}$ ,  $p \in (0, 1)$ . Чему равняется  $E(\xi|\eta)$ ?

6. Пусть  $\xi \sim \text{Exp}(\lambda)$ ,  $\lambda > 0$ , а сигма-алгебра  $\mathcal{C}$  порождена счётной системой событий  $\{B_n\}_{n \geq 1}$ , где  $B_n = \{n - 1 \leq \xi < n\}$ . Какое распределение имеет случайная величина  $\eta = E(\xi|\mathcal{C})$ ?
7. Найдите  $E((X - Y)^2|X + Y)$ , если случайный вектор  $(X, Y)$  имеет нормальное распределение со средним  $(1, -1)$  и матрицей ковариаций

$$\begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}.$$

8. Найдите  $E(X - Y - Z|X + 2Y + Z)$ , если случайный вектор  $(X, Y, Z)$  имеет нормальное распределение со средним  $(-1, 0, 2)$  и матрицей ковариаций

$$\begin{pmatrix} 3 & -2 & 2 \\ -2 & 4 & -1 \\ 2 & -1 & 3 \end{pmatrix}.$$

## 6. Условные распределения и подсчет условных математических ожиданий

В прошлой главе мы определили условное математическое ожидание и сформулировали ряд его свойств, которые позволили вычислять условное математическое ожидание в простейших случаях. Тем не менее, во многих ситуациях этих свойств недостаточно для вычисления условного математического ожидания. Заметим, что для двух дискретных случайных величин  $\xi, \eta$  вычислять условное математическое ожидание  $E(\xi|\eta)$  позволяет задача 5.1. Если же случайные величины являются абсолютно непрерывными, то для вычисления условного математического ожидания вводят понятие условной плотности (по аналогии с обычным математическим ожиданием). В этой главе мы введем это понятие и научимся с его помощью вычислять условное математическое ожидание и условные вероятности. Но прежде приведем пример ситуации, в которой вычисление условного математического ожидания возможно и без использования условной плотности.

▲ **Задача 6.1.** Пусть  $X$  и  $Y$  — независимые одинаково распределенные случайные величины. Найдите  $E(X|X+Y)$ .

Решение этой задачи опирается на следующее утверждение.

**Теорема 6.1** Пусть  $X_1, \dots, X_n$  — независимые одинаково распределенные случайные величины, а  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  — симметричная борелевская функция (для любых  $x_1, \dots, x_n$  и любой перестановки  $\sigma$  на  $\{1, \dots, n\}$  выполнено  $f(x_1, \dots, x_n) = f(x_{\sigma(1)}, \dots, x_{\sigma(n)})$ ). Тогда для любых различных  $i, j \in \{1, \dots, n\}$  справедливо равенство

$$E(X_i|f(X_1, \dots, X_n)) = E(X_j|f(X_1, \dots, X_n))$$

(п.н.).

### Доказательство теоремы 6.1

Пусть  $1 \leq i < j \leq n$ . Пусть, кроме того,  $A \in \sigma(f(X_1, \dots, X_n))$  — произвольное множество. Утверждение теоремы следует из интегрального свойства, которое мы и собираемся доказать:

$$\mathbf{E}(X_i I_A) = \mathbf{E}(\mathbf{E}(X_j | f(X_1, \dots, X_n)) I_A).$$

По определению условного математического ожидания, последнее равенство справедливо тогда и только тогда, когда  $\mathbf{E}(X_i I_A) = \mathbf{E}(X_j I_A)$ . Кроме того, по определению  $\sigma(f(X_1, \dots, X_n))$ , существует такое борелевское множество  $B$ , что  $A = \{f(X_1, \dots, X_n) \in B\}$ . В силу того, что случайные величины  $X_1, \dots, X_n$  независимы и одинаково распределены, случайные векторы  $X_{ij} = (X_1, \dots, X_i, \dots, X_j, \dots, X_n)$  и  $X_{ji} = (X_1, \dots, X_j, \dots, X_i, \dots, X_n)$  одинаково распределены. Поэтому одинаково распределены и случайные величины  $X_i I(f(X_{ij}) \in B)$  и  $X_j I(f(X_{ji}) \in B) = X_j I(f(X_{ij}) \in B)$  (последнее равенство выполнено в силу симметричности функции  $f$ ). Таким образом,

$$\begin{aligned} \mathbf{E}(X_i I_A) &= \mathbf{E}I(X_i(f(X_{ij}) \in B)) = \mathbf{E}I(X_j(f(X_{ij}) \in B)) = \\ &= \mathbf{E}(X_j I_A). \end{aligned}$$

**Теорема доказана.**

Обратимся теперь к решению задачи.

### Решение задачи 6.1

По теореме 6.1, имеем  $\mathbf{E}(X|X+Y) = \mathbf{E}(Y|X+Y)$ . Поэтому

$$\begin{aligned} \mathbf{E}(X|X+Y) &= \frac{1}{2}(\mathbf{E}(X|X+Y) + \mathbf{E}(Y|X+Y)) = \\ &= \frac{1}{2}\mathbf{E}(X+Y|X+Y) = \frac{X+Y}{2}. \end{aligned}$$

**Задача решена.**

Для введения понятия условной плотности нам потребуются следующие обозначения. Пусть  $E(\xi|\eta = y)$  — такая борелевская функция  $\varphi(y)$ , что для любого множества  $B \in \mathcal{B}(\mathbb{R})$  выполнено

$$E(\xi I\{\eta \in B\}) = \int_B \varphi(y) P_\eta(dy).$$

Условное распределение  $P_\xi(B|\eta = y) = P(\xi \in B|\eta = y)$  — это  $E(I\{\xi \in B\}|\eta = y)$ .

**▲ Определение 6.1.** Условной плотностью называется такая функция  $p_{\xi|\eta}(x|y)$ , что

$$P_\xi(B|\eta = y) = \int_B p_{\xi|\eta}(x|y) dx.$$

**Теорема 6.2** Если существует совместная плотность  $p_{(\xi,\eta)}(x,y)$ , то существует и условная плотность, причем она может быть вычислена по формуле

$$p_{\xi|\eta}(x|y) = \begin{cases} \frac{p_{(\xi,\eta)}(x,y)}{p_\eta(y)}, & \text{если } p_\eta(y) \neq 0, \\ 0, & \text{если } p_\eta(y) = 0, \end{cases}$$

где  $p_\eta$  — плотность случайной величины  $\eta$ .

**Теорема 6.3** Пусть  $f$  — такая борелевская функция, что  $E|f(\xi)| < +\infty$ . Пусть  $E(f(\xi)|\eta = y) = \varphi(y)$ . Тогда  $E(f(\xi)|\eta) = \varphi(\eta)$ . Кроме того,

$$\varphi(y) = \int_{\mathbb{R}} f(x) P_\xi(dx|\eta = y).$$

Если существует условная плотность  $p_{\xi|\eta}$ , то

$$E(f(\xi)|\eta = y) = \int_{\mathbb{R}} f(x) p_{\xi|\eta}(x|y) dx.$$

Разумеется, условное распределение можно вычислять, используя теорему 6.3. Итак, опишем схему вычисления условных математических ожиданий и вероятностей в случае существования совместной плотности.

**▲ Порядок вычисления условного математического ожидания и условной вероятности при заданной совместной плотности  $p_{(\xi, \eta)}(x, y)$ .**

1. Находим плотность случайной величины  $\eta$ :

$$p_{\eta}(y) = \int_{-\infty}^{+\infty} p_{(\xi, \eta)}(x, y) dx.$$

2. Вычисляем условную плотность  $p_{\xi|\eta}(x|y)$  с помощью теоремы 6.2.
3. Вычисляем  $\varphi(y) = E(f(\xi)|\eta = y)$  с помощью теоремы 6.3. В частности, если  $f(x) = I(x \in B)$ , то  $\varphi(y) = P_{\xi}(B|\eta = y)$ .
4. Подставляем  $\eta$  вместо  $y$  в формулу для  $\varphi(y)$  и получаем окончательный ответ  $E(f(\xi)|\eta) = \varphi(\eta)$ . В частности, если  $f(x) = I(x \in B)$ , то  $P_{\xi}(B|\eta) = \varphi(\eta)$ .

**▲ Задача 6.2.** Пусть  $\xi, \eta$  — независимые одинаково распределенные случайные величины с распределением  $\text{Exp}(1)$ . Найдите  $E(\xi^2|\xi + \eta)$ .

## Решение

Найдём совместную плотность случайных величин  $\xi$  и  $\xi + \eta$ . Для любых множеств  $B_1, B_2 \in \mathcal{B}(\mathbb{R})$  имеем

$$\begin{aligned} \int_{(x,y) \in B_1 \times B_2} p_{\xi, \xi+\eta}(x, y) dx dy &= P(\xi \in B_1, \xi + \eta \in B_2) = \\ &= \int_{u \in B_1, u+v \in B_2} e^{-(u+v)} I(u > 0, v > 0) du dv = \\ &= \int_{x \in B_1, y \in B_2} e^{-y} I(x > 0, y - x > 0) dx dy, \end{aligned}$$

где последний интеграл получен заменой  $u = x$ ,  $u + v = y$  (Якобиан этой замены равен

$$\frac{D(u, v)}{D(x, y)} = \left| \det \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \right| = 1$$

т.е. не меняет подынтегрального выражения). Поэтому  $p_{\xi, \xi+\eta}(x, y) = e^{-y} I(0 < x < y)$ . Плотность  $p_{\xi+\eta}(y)$  равна

$$\int_{-\infty}^{+\infty} p_{\xi, \xi+\eta}(x, y) dx = \int_0^y e^{-y} I\{y > 0\} dx = y e^{-y} I\{y > 0\}.$$

Таким образом,  $p_{\xi|\xi+\eta}(x|y) =$

$$= \begin{cases} \frac{p_{\xi, \xi+\eta}(x, y)}{p_{\xi+\eta}(y)}, & y > 0, \\ 0, & y \leq 0 \end{cases} = \begin{cases} \frac{1}{y} I\{0 < x < y\}, & y > 0, \\ 0, & y \leq 0. \end{cases}$$

Осталось найти условное математическое ожидание. Итак,

$$\begin{aligned} E(\xi^2 | \xi + \eta = y) &= \int_{\mathbb{R}} \frac{x^2}{y} I\{0 < x < y\} dx = \int_0^y \frac{x^2}{y} I(y > 0) dx = \\ &= \frac{y^2}{3} I(y > 0), \quad E(\xi^2 | \xi + \eta) = \varphi(\xi + \eta) = \frac{(\xi + \eta)^2}{3}. \end{aligned}$$

**Задача решена.**



### ▲ Задачи для самостоятельного решения

1. Пусть  $X, Y, Z$  — независимые одинаково распределённые случайные величины. Найдите  $E(4X - 3Y + Z | X + Y + Z)$ .
2. Пусть случайные величины  $X$  и  $Y$  независимы, плотность случайной величины  $X$  равна

$$\frac{1}{x} I_{[1,e]}(x),$$

плотность случайной величины  $Y$  равна

$$4x^3 I_{[0,1]}(x).$$

Вычислите  $E(\frac{1}{X} + \frac{1}{Y} | XY)$ .

3. Пусть  $X$  и  $Y$  — независимые случайные величины, равномерно распределённые на отрезке  $[0, 2]$ . Найдите  $E(Y^2 | X/Y)$ .
4. Пусть  $X$  и  $Y$  — независимые случайные величины,  $X$  имеет равномерное распределение на отрезке  $[0, 1]$ , а  $Y$  — экспоненциальное распределение с параметром 1. Найдите а)  $E(Y | X/Y)$ , б)  $E(Y^3/X^2 | X/Y)$ .
5. Пусть  $X_1, \dots, X_n$  — выборка из распределения  $R(0, 1)$ . Найдите а)  $E(X_1 | X_{(1)})$ , б)  $E(X_1 | X_{(n)})$ .
6. Найдите

$$E(3X^2 + 5X - 5XY - 2Y^2 - 10Y | 2Y - X)$$

при условии, что  $X$  и  $Y$  независимы,  $X$  имеет стандартное распределение Лапласа (с параметром  $\sigma = 1$ ),  $Y \sim R(-1, 2)$ .

7. Пусть  $\xi_1, \xi_2, \dots$  — независимые одинаково распределённые случайные величины с конечным математическим ожиданием,  $S_n = \xi_1 + \dots + \xi_n$ ,  $n \in \mathbb{N}$ . Докажите, что для всех  $n \in \mathbb{N}$  справедливо равенство

$$E(\xi_1 | S_n, S_{n+1}, \dots) = \frac{S_n}{n} \quad (\text{п.н.}).$$

## 7. Достаточные статистики и оптимальные оценки

В этой главе мы продолжим говорить о свойствах оценок. Нас по-прежнему будет интересовать вопрос существования оценки, наилучшей в среднеквадратичном подходе (т.е. “оптимальной” в смысле среднеквадратического отклонения) в классе всех несмещенных оценок. Разумеется, критерий эффективности не решает такую задачу полностью — его условия могут быть и не выполнены, в то время как наилучшая оценка может существовать. Прежде чем рассказать об алгоритме нахождения таких оценок, введем понятие достаточной статистики, а также понятие полной статистики.

### 7.1. Достаточные статистики

Пусть  $X$  — наблюдение с неизвестным распределением  $P \in \mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ .

**▲ Определение 7.1.** Статистика  $S(X)$  — *достаточная* для семейства распределений  $\mathcal{P}$ , если условное распределение  $P_\theta(X \in B | S(X) = x)$  не зависит от  $\theta$ .

Далее мы сформулируем критерий того, что статистика является достаточной, но прежде, чтобы проиллюстрировать определение этого понятия, решим задачу на проверку достаточности.

**▲ Задача 7.1.** Пусть  $X_1, \dots, X_n$  — выборка из распределения  $\text{Bern}(\theta)$ . Докажите, что  $\sum_{i=1}^n X_i$  — достаточная статистика.

## Решение

Проверим определение достаточной статистики:

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | S(X) = s) &= \\ &= \frac{P(X_1 = x_1, \dots, X_n = x_n, \sum X_i = s)}{P(\sum X_i = s)} = \\ &= I \left\{ \sum_{i=1}^n x_i = s \right\} \frac{\theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}}{C_n^s \theta^s (1 - \theta)^{n-s}} = \frac{1}{C_n^s} I \left\{ \sum x_i = s \right\}. \end{aligned}$$

Видим, что условное распределение не зависит от  $\theta$ , значит, статистика  $\sum_{i=1}^n X_i$  является достаточной для семейства распределений  $\text{Bern}(\theta)$ .

Задача решена.

**Теорема 7.1** (критерий факторизации Неймана-Фишера) Пусть  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  — доминируемое семейство распределений  $P_\theta$  с “обобщенной” плотностью  $p_\theta$  (в случае дискретного распределения  $p_\theta(x) = P_\theta(X = x)$ ),  $X$  — наблюдение с неизвестным распределением  $P_\theta \in \mathcal{P}$ . Тогда  $S(X)$  — достаточная статистика в том и только том случае, когда “обобщенная” плотность допускает представление

$$p_\theta(X) = \psi(S(X), \theta) h(X),$$

где функция  $h$  не зависит от параметра  $\theta$ .

▲ **Задача 7.2.** Пусть  $X_1, \dots, X_n$  — выборка из  $N(a, \sigma^2)$ ,  $\theta = (a, \sigma^2)$ . Найдите достаточную статистику.

## Решение

Найдем правдоподобие выборки из  $\mathcal{N}(a, \sigma^2)$  :

$$\begin{aligned} f_{\theta}(X_1, \dots, X_n) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum (X_i - a)^2 \right\} = \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum X_i^2 + \frac{a}{\sigma^2} \sum X_i - \frac{na^2}{2\sigma^2} \right\}. \end{aligned}$$

Положив  $h(x) = 1$ ,

$$\psi(s_1, s_2, \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} s_1 + \frac{a}{\sigma^2} s_2 - \frac{na^2}{2\sigma^2} \right\}$$

и применив критерий факторизации, получим, что статистика  $S(X) = (\sum X_i^2, \sum X_i)$  является достаточной.

**Задача решена.**

Для произвольного семейства распределений, параметризованного параметром  $\theta$ , и некоторой функции  $\tau(\theta)$  теорема Рао-Блэкуэлла-Колмогорова (см., например, [1]), казалось бы, утверждает, что оценка  $\theta^* = E_{\theta}(\hat{\theta}|S(X))$  является наилучшей в среднеквадратичном подходе при условии, что  $\hat{\theta}$  — несмещенная оценка  $\tau(\theta)$ , а  $S(X)$  — достаточная статистика (заметим, что условное математическое случайного вектора есть вектор, составленный из условных математических ожиданий компонент рассматриваемого случайного вектора). Тем не менее, это не так — эта теорема позволяет лишь улучшить несмещенную оценку  $\hat{\theta}$ , а наилучшую оценку она позволяет найти, если статистика  $S(X)$  является полной.

## 7.2. Полные статистики

**▲ Определение 7.2.** Статистика  $S(X)$  называется *полной* для  $\{P_{\theta}, \theta \in \Theta\}$ , если для всех функций  $f(x)$  со свойством

$$\forall \theta \in \Theta \quad E_{\theta} f(S(X)) = 0$$

выполнено  $f(S(X)) = 0$  ( $P_{\theta}$ -п.н.) для всех  $\theta \in \Theta$ .

▲ **Задача 7.3.** Докажите, что статистика  $\sum_{i=1}^n X_i$  является полной для семейства распределений  $\text{Bern}(\theta)$ ,  $\theta \in (0, 1)$ .

### Решение

Итак, нужно доказать, что если для всех  $\theta \in (0, 1)$  и какой-то функции  $f(x)$  выполнено  $E_\theta f(\sum X_i) = 0$ , то  $f(x) = 0$  для всех  $x \in \{0, 1, \dots, n\}$ . Имеем

$$E_\theta f\left(\sum X_i\right) = \sum_{k=0}^n f(k) C_n^k \theta^k (1-\theta)^{n-k}.$$

Мы получили в правой части многочлен не более чем  $n$ -ой степени от  $\theta$ . Если не все его коэффициенты равны 0, то он имеет не более чем  $n$  корней на  $(0, 1)$ . Но этот многочлен равен 0 для всех  $\theta \in (0, 1)$ , т.е. все  $\theta$  из интервала  $(0, 1)$  являются его корнями. Следовательно, все коэффициенты этого многочлена равны 0, а значит,  $f(k) = 0$  для всех  $k \in \{0, \dots, n\}$ .

**Задача решена.**

Разумеется, далеко не всегда статистика  $\sum_{i=1}^n X_i$  является полной. Как же найти полную статистику? Как мы увидим далее, для достаточно широкого класса распределений это очень просто.

▲ **Определение 7.3.** Семейство  $\{P_\theta, \theta \in \Theta\}$  называется *экспоненциальным*, если “обобщенная” плотность распределения  $P_\theta$  имеет вид

$$p_\theta(x) = h(x) \exp \left( \sum_{i=1}^k a_i(\theta) u_i(x) + v(\theta) \right). \quad (1)$$

Заметим, что нормальное, экспоненциальное, биномиальное, пуассоновское и многие другие семейства распределений являются экспоненциальными.

**Теорема 7.2** (об экспоненциальных семействах) Пусть  $\theta \in \Theta \subset \mathbb{R}^k$  и для семейства распределений  $\{P_\theta, \theta \in \Theta\}$  выполнено условие (1). Пусть, кроме того, множество всех значений вектора  $(a_1(\theta), \dots, a_k(\theta))$  при  $\theta \in \Theta$  содержит  $k$ -мерный параллелепипед. Тогда  $S(X) = (u_1(x), \dots, u_k(x))$  — полная достаточная статистика.

Стоит заметить, что условие теоремы будет выполнено, если множество  $\Theta$  “телесно”, т.е. содержит свои внутренние точки, и функции  $a_1(\theta), \dots, a_k(\theta)$  в окрестности какой-нибудь внутренней точки  $\theta_0 \in \Theta$  линейно независимы и гладкие.

### 7.3. Оптимальные оценки

В этом разделе мы, наконец, приведем алгоритм нахождения наилучших в упомянутом смысле оценок, которые принято называть оптимальными.

**▲ Определение 7.4.** Статистика  $\theta^*$  называется *оптимальной оценкой*  $\tau(\theta)$ , если  $\theta^*$  — наилучшая в среднеквадратичном подходе в классе несмещенных оценка  $\tau(\theta)$  (заметим, что  $\tau(\theta)$  может, вообще говоря, иметь размерность, большую 1).

**Теорема 7.3** (Лемана–Шеффе) Пусть  $S(X)$  — полная достаточная статистика,  $\hat{\theta}$  — несмещённая оценка  $\tau(\theta)$ , тогда  $\theta^* = E_\theta(\hat{\theta}|S(X))$  — оптимальная оценка  $\tau(\theta)$ .

#### ▲ Алгоритм нахождения оптимальной оценки.

1. Либо ищем достаточную статистику  $S(X)$  с помощью критерия факторизации и доказываем, что она является полной, либо сразу находим полную достаточную статистику, если данное семейство распределений является экспоненциальным.
2. Решаем относительно  $\varphi$  уравнение “несмещенности”  $E_\theta \varphi(S(X)) = \tau(\theta)$  (для любого  $\theta \in \Theta$ ). Его решение  $\varphi(S(X))$  является оптимальной оценкой  $\tau(\theta)$ .

▲ **Задача 7.4.** Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на  $[0, \theta]$ . Найдите оптимальную оценку параметра  $\theta$ .

### Решение

Функция правдоподобия равна  $\frac{1}{\theta^n} I\{X_{(n)} \leq \theta\}$ , поэтому, по критерию факторизации,  $X_{(n)}$  — достаточная статистика. Проверим, что эта статистика является полной:

$$E_{\theta} f(X_{(n)}) = \int_0^{\theta} f(x) n \frac{x^{n-1}}{\theta^n} dx.$$

По свойству интеграла Римана, последнее выражение равно 0 для любого  $\theta > 0$  тогда и только тогда, когда  $f(x)x^{n-1} = 0$  для почти всех  $x > 0$  (мера Лебега множества таких  $x$ , что последнее выражение отлично от 0, равна 0). Поэтому  $f(x) = 0$  ( $P_{\theta}$ -п.н.) для всех  $\theta \in \Theta$ .

Итак, мы знаем, что  $X_{(n)}$  — полная достаточная статистика. Осталось решить уравнение “несмещенности”  $E_{\theta} \varphi(X_{(n)}) = \theta$ . Как известно,

$$E_{\theta} X_{(n)} = \frac{n}{n+1} \theta.$$

Поэтому в качестве решения уравнения можно предложить функцию  $\varphi(x) = \frac{n+1}{n} x$ . Следовательно, оценка  $\frac{n+1}{n} X_{(n)}$  является оптимальной.

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Приведите пример такого параметрического семейства распределений  $\mathcal{P}$  и нетривиальной неполной достаточной статистики  $S(X_1, \dots, X_n)$ , где  $X_1, \dots, X_n$  — выборка из неизвестного распределения  $P \in \mathcal{P}$ , что размерность статистики  $S$  равна 1.
2. Докажите, что в модели  $N(\theta, \gamma\theta^2)$ ,  $(\theta, \gamma) \in \mathbb{R} \times (0, \infty)$ , ста-

тестистика  $(\bar{X}, s^2)$  является полной, но не является достаточной. *Указание:* для доказательства того, что статистика не является полной, выберите удачную функцию  $f$ , посчитав математические ожидания  $\bar{X}^2$  и  $s^2$ .

3. Найдите достаточные статистики для следующих параметрических распределений: а)  $\mathcal{N}(a, \sigma^2)$ , б)  $\Gamma(\alpha, \lambda)$ , в)  $R(a, b)$ , г)  $\text{Pois}(\lambda)$ , д)  $\text{Bin}(1, p)$ , е)  $\text{Geom}(p)$ .
4. Найдите оптимальную оценку параметра  $\theta > 0$  по выборке из распределения: а)  $\mathcal{N}(\theta, 1)$ , б)  $R(0, \theta)$ , в)  $\text{Pois}(\theta)$ , г)  $\text{Bin}(1, \theta)$  (здесь  $\theta \in (0, 1)$ ).
5. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(a, \sigma^2)$ ,  $a \in \mathbb{R}$ ,  $\sigma > 0$ . Найдите оптимальную оценку параметра  $\theta = (a, \sigma^2)$ .
6. Пусть  $X_1, \dots, X_n$  — выборка из экспоненциального распределения с параметром  $\theta > 0$ . Найдите оптимальные оценки для  $\theta$  и  $\tau(\theta) = \theta^{1/2}$ .
7. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(0, \theta^2)$ . Найдите оптимальную оценку для  $\theta$ .
8. Пусть  $X_1, \dots, X_n$  — выборка из пуассоновского распределения с параметром  $\theta > 0$ . Найдите  $\mathbb{E} \left( X_1^2 \left| \sum_{i=1}^n X_i \right. \right)$ .



## 8. Доверительные интервалы

В предыдущих главах мы изучали точечные оценки параметров. Разумеется, в большинстве случаев вероятность того, что истинное значение параметра совпадает с его оценкой, равна нулю. В этой связи удобнее в качестве оценки рассматривать целый интервал значений неизвестного параметра. При этом размер этого интервала, разумеется, должен быть как можно меньше, а вот вероятность попадания в него истинного значения параметра — наоборот, как можно больше. В этой главе мы и поговорим о таких интервалах.

### 8.1. Определение и методы построения доверительных интервалов

Пусть  $X$  — наблюдение с неизвестным распределением  $P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}$ .

**▲ Определение 8.1.** Пара статистик  $(T_1(X), T_2(X))$  (где  $X$  — наблюдение) называется *доверительным интервалом уровня доверия  $\gamma$*  для параметра  $\theta$ , если для любого  $\theta \in \Theta \subset \mathbb{R}$  выполнено

$$P_\theta(T_1(X) < \theta < T_2(X)) \geq \gamma.$$

Доверительный интервал называется *точным*, если для любого  $\theta \in \Theta$  вместо последнего неравенства выполнено равенство.

Для нахождения доверительного интервала можно, например, воспользоваться неравенством Маркова или неравенством Чебышева.

**▲ Задача 8.1.** Пусть  $X_1, \dots, X_n$  — выборка из пуассоновского распределения с параметром  $\lambda > 0$ . Постройте доверительный интервал для  $\lambda$  уровня доверия  $\gamma$ .

## Решение

В силу неравенства Чебышева,

$$\mathbf{P}\left(\left|\sum X_i - \lambda n\right| \geq \varepsilon n\right) < \frac{\lambda}{\varepsilon^2 n}.$$

Следовательно,

$$\mathbf{P}\left(\lambda - \varepsilon < \overline{X} < \lambda + \varepsilon\right) \geq 1 - \frac{\lambda}{\varepsilon^2 n} = \gamma.$$

Из последнего равенства находим  $\varepsilon = \sqrt{\lambda/((1-\gamma)n)}$ . Решим неравенство  $\overline{X} < \lambda + \sqrt{\lambda((1-\gamma)n)}$  относительно  $\lambda$ : для этого достаточно прибавить к левой и правой части неравенства  $\frac{1}{4(1-\gamma)n}$ , тем самым, в правой части образуется полный квадрат. Второе неравенство решается аналогично. Окончательно получаем

$$\mathbf{P}\left(\sqrt{\overline{X} + \frac{1}{4(1-\gamma)n}} - \sqrt{\frac{1}{2(1-\gamma)n}} < \lambda < \sqrt{\overline{X} - \frac{1}{4(1-\gamma)n}} + \sqrt{\frac{1}{2(1-\gamma)n}}\right) \geq \gamma.$$

**Задача решена.**

Для нахождения точных доверительных интервалов необходимо знание точного значения вероятности, а значит, необходимо воспользоваться некоторой функцией  $G(X, \theta)$ , распределение которой не зависит от  $\theta$ . Эта функция называется *центральной статистикой*, а описанный ниже метод построения точного доверительного интервала — *методом центральной статистики*.

Итак, пусть случайная величина  $G(X, \theta)$  такова, что её распределение известно и не зависит от  $\theta$ . Пусть также  $G(X, \theta)$  строго монотонна (например, возрастает) и непрерывна по  $\theta$ . Пусть  $0 \leq p_1 < p_2 \leq 1$  таковы, что  $p_2 - p_1 = \gamma$ . Для каждого  $i \in \{1, 2\}$  рассмотрим  $p_i$ -квантили  $z_{p_i}$  распределения  $G(X, \theta)$ . Пусть  $T_i(X)$

— решения уравнений  $G(X, T_i(X)) = z_{p_i}$ ,  $i \in \{1, 2\}$ . Тогда

$$P_\theta(T_1(X) < \theta < T_2(X)) = P_\theta(z_{p_1} < G(X, \theta) < z_{p_2}) = p_2 - p_1 = \gamma,$$

т.е.  $(T_1(X), T_2(X))$  — доверительный интервал уровня доверия  $\gamma$ .

Заметим, что для минимизации длины доверительного интервала  $p_1$  и  $p_2$  в методе центральной статистики выбираются часто симметричными относительно 0.5, так как у распределения  $G(X, \theta)$  могут быть тяжелые хвосты.

**▲ Задача 8.2.** Пусть  $X_1, \dots, X_n$  — выборка из распределения с плотностью

$$p_\theta(x) = e^{-(x-\theta)} I\{x \geq \theta\}$$

(экспоненциальное распределение со сдвигом  $\theta$ ). Построить доверительный интервал уровня доверия  $\gamma$ .

### Решение

Заметим, что  $X_i - \theta \sim \text{Exp}(1)$  (распределение не зависит от  $\theta$ ). Возьмем в качестве центральной статистики  $G(X, \theta) = \sum X_i - n\theta$ . Эта статистика имеет распределение  $\Gamma(1, n)$ , так как сумма независимых, одинаково распределенных экспоненциальных случайных величин имеет гамма-распределение. Рассмотрим  $\frac{1-\gamma}{2}$ - и  $\frac{1+\gamma}{2}$ -квантили  $u_{\frac{1-\gamma}{2}}$  и  $u_{\frac{1+\gamma}{2}}$  распределения  $\Gamma(1, n)$ . Имеем

$$\begin{aligned} \gamma &= \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = P_\theta \left( u_{\frac{1-\gamma}{2}} < \sum X_i - n\theta < u_{\frac{1+\gamma}{2}} \right) = \\ &= P_\theta \left( \bar{X} - \frac{u_{\frac{1+\gamma}{2}}}{n} < \theta < \bar{X} - \frac{u_{\frac{1-\gamma}{2}}}{n} \right). \end{aligned}$$

**Задача решена.**

## 8.2. Асимптотические доверительные интервалы

Доверительные интервалы малого размера построить удастся далеко не всегда. В случае выборки большого размера от ограничения размера выборки можно отойти и рассматривать асимптотическую задачу. При такой постановке доверительный

интервал малого размера найти, как правило, проще.

Итак, пусть  $X_1, \dots, X_n$  — выборка из  $P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}$ .

▲ **Определение 8.2.** Последовательность интервалов

$$\left( T_1^{(n)}(X_1, \dots, X_n), T_2^{(n)}(X_1, \dots, X_n) \right)$$

называется *асимптотическим доверительным интервалом уровня доверия  $\gamma$*  для  $\theta$ , если для любого  $\theta \in \Theta$  выполнено

$$\liminf_{n \rightarrow \infty} P_\theta \left( T_1^{(n)}(X_1, \dots, X_n) < \theta < T_2^{(n)}(X_1, \dots, X_n) \right) \geq \gamma.$$

Асимптотический доверительный интервал называется *точным*, если для любого  $\theta \in \Theta$

$$\lim_{n \rightarrow \infty} P_\theta \left( T_1^{(n)}(X_1, \dots, X_n) < \theta < T_2^{(n)}(X_1, \dots, X_n) \right) = \gamma.$$

Пусть  $\hat{\theta}_n(X_1, \dots, X_n)$  — асимптотически нормальная оценка  $\theta$  с асимптотической дисперсией  $\sigma^2(\theta)$ , т.е.

$$\sqrt{n} \frac{\hat{\theta}_n - \theta}{\sigma(\theta)} \xrightarrow{d_\theta} \xi \sim N(0, 1).$$

Рассмотрим  $\frac{1-\gamma}{2}$ - и  $\frac{1+\gamma}{2}$ -квантили  $u_{\frac{1-\gamma}{2}}$  и  $u_{\frac{1+\gamma}{2}}$  распределения  $N(0, 1)$ . Имеем

$$\lim_{n \rightarrow \infty} P \left( u_{\frac{1-\gamma}{2}} < \sqrt{n} \frac{\hat{\theta}_n - \theta}{\sigma(\theta)} < u_{\frac{1+\gamma}{2}} \right) = \gamma.$$

Если нам повезло с функцией  $\sigma(\theta)$ , то из последнего равенства можно извлечь точный асимптотический доверительный интервал. Если же нам повезло меньше — получить доверительный интервал не удастся, но если функция  $\sigma(\theta)$  является непрерывной, то, по теореме о наследовании сходимости по вероятности,  $\frac{\sigma(\theta)}{\sigma(\hat{\theta}_n)} \xrightarrow{P_\theta} 1$ . Тогда из леммы Slutsky,

$$\sqrt{n} \frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)} \xrightarrow{d_\theta} \xi,$$

откуда уже несложно получить точный асимптотический доверительный интервал для  $\theta$ .

**▲ Задача 8.3.** Пусть  $X_1, \dots, X_n$  — выборка из  $\text{Bern}(\theta)$ . Постройте асимптотический доверительный интервал уровня доверия  $\gamma$  для параметра  $\theta$ .

### Решение

По центральной предельной теореме,

$$\sqrt{n}(\bar{X} - \theta) \xrightarrow{d_\theta} \xi \sim \mathcal{N}(0, \theta(1 - \theta)).$$

Так как функция  $\sqrt{\theta(1 - \theta)}$  непрерывна по  $\theta$  на  $(0, 1)$ , то

$$\sqrt{n} \frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}} \xrightarrow{d_\theta} \xi.$$

Как обычно, рассмотрим  $\frac{1-\gamma}{2}$ - и  $\frac{1+\gamma}{2}$ -квантили  $u_{\frac{1-\gamma}{2}}$  и  $u_{\frac{1+\gamma}{2}}$  распределения  $\mathcal{N}(0, 1)$ . Положим

$$T_1^{(n)} = \hat{\theta}_n - \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} u_{\frac{1+\gamma}{2}}, \quad T_2^{(n)} = \hat{\theta}_n - \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} u_{\frac{1-\gamma}{2}}.$$

Тогда имеем  $\lim_{n \rightarrow \infty} \mathbb{P}(T_1^{(n)} < \theta < T_2^{(n)}) = \gamma$ . Следовательно,  $(T_1^{(n)}, T_2^{(n)})$  — асимптотический доверительный интервал уровня доверия  $\gamma$ .

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[0, \theta]$ ,  $\theta > 0$ . Постройте (неасимптотический) доверительный интервал для  $\theta$  уровня доверия  $\alpha$ , используя статистику а)  $\bar{X}$ , б)  $X_{(1)}$ , в)  $X_{(n)}$ .

2.  $X_1, \dots, X_n$  — выборка из распределения с плотностью

$$p_\theta(x) = \frac{3x^2}{8\theta^3} I\{x \in [0, 2\theta]\}.$$

С помощью статистики  $X_{(1)}$  постройте точный доверительный интервал уровня доверия  $\gamma$  для параметра  $\theta$ .

3.  $X_1, \dots, X_n$  — выборка,  $X_1 = \xi + \eta$ , где  $\xi, \eta$  — независимые случайные величины,  $\xi \sim R[0, \theta]$ ,  $\eta \sim Bin(1, \theta)$ . Постройте доверительный интервал для  $\theta$  уровня доверия  $1 - \alpha$  с помощью неравенства Чебышева.

4. Пусть  $X_1, \dots, X_n$  — выборка из распределения Коши со сдвигом, т.е.

$$p_\theta(x) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$

Постройте асимптотический доверительный интервал для  $\theta$  уровня доверия  $\alpha$ .

5. Пусть  $X_1, \dots, X_n$  — выборка из пуассоновского распределения с параметром  $\theta$ . Постройте асимптотический доверительный интервал для  $\theta$  уровня доверия  $\alpha$ .

6. Пусть  $X_1, \dots, X_n$  — выборка из гамма-распределения с параметрами  $(\theta, \lambda)$ . Постройте асимптотический доверительный интервал для  $\theta$  уровня доверия  $\alpha$ , если а)  $\lambda$  известно, б)  $\lambda$  неизвестно.

7.  $X_1, \dots, X_n$  — выборка,  $X_1 = \xi + \eta$ , где  $\xi, \eta$  — независимые случайные величины,  $\xi \sim N(\theta^2, \theta^2 + 1)$ ,  $P(\eta = 0) = P(\eta = 4\theta) = 1/2$ . Постройте доверительный интервал уровня доверия  $\gamma$  для параметра  $\theta$ .

8. Пусть  $X_1, \dots, X_n$  — выборка из распределения, имеющего плотность

$$p_\theta(x) = \frac{1}{2} \exp\{-||x| - 6\theta|\} \cdot I\{|x| > 6\theta\}.$$

Постройте доверительный интервал для  $\theta$  уровня доверия  $1 - \alpha$ .

## 9. Линейная регрессия

В этой главе мы рассмотрим достаточно специальную, но, в то же время, широко применимую задачу теории оценивания статистических параметров. Наблюдением в этой задаче уже будет являться не выборка (набор независимых случайных величин), а вектор, составленный из случайных величин, распределения которых различны. Рассмотрим известный пример с серией измерений масс, произведенных на одних и тех же весах. В этом примере имеется мешок монет масс  $l_1, \dots, l_k$ , различные комбинации которых взвешиваются суммарно  $n$  раз. Таким образом, истинные значения, которые должны наблюдаться в экспериментах (в случае отсутствия ошибок измерений), равны линейным комбинациям  $l_1, \dots, l_k$ . Но так как ошибки все же имеются, то наблюдаемая величина, полученная при  $i$ -ом взвешивании ( $i \in \{1, \dots, n\}$ ) равна  $z_{i,1}l_1 + \dots + z_{i,k}l_k + \varepsilon_i$ , где  $z_{i,j}$ ,  $j \in \{1, \dots, k\}$ , — количество монет веса  $l_j$ , участвующих в  $i$ -ом взвешивании,  $\varepsilon_i$  — ошибка измерения при  $i$ -ом взвешивании. Задача состоит в оценивании истинных значений  $l_1, \dots, l_k$  по наблюдению  $X_1, \dots, X_n$ .

Итак, пусть наблюдение  $X$  — случайный вектор из  $\mathbb{R}^n$ , причем  $X = l + \varepsilon$ , где  $l$  — фиксированный неизвестный вектор, а  $\varepsilon$  — случайный вектор. Про вектор  $\varepsilon$  известно, что  $E\varepsilon = 0$ ,  $D\varepsilon = \sigma^2 I_n$ , где  $D\varepsilon$  — ковариационная матрица вектора  $\varepsilon$ ,  $I_n$  — единичная диагональная матрица размера  $n \times n$ . Про вектор  $l$  известно, что  $l \in L$ , где  $L = \langle Z_1, \dots, Z_k \rangle$  — заданное линейное подпространство в  $\mathbb{R}^n$  размерности  $k < n$  ( $Z_1, \dots, Z_k$  — базисные вектор-столбцы пространства  $L$ ). Параметрами описанной *линейной регрессионной модели* являются  $l$  и  $\sigma^2$ , об оценивании которых речь пойдет ниже.

Введем матрицу  $Z = (Z_1, \dots, Z_k)$ . Тогда  $l = Z_1\theta_1 + \dots + Z_k\theta_k = Z\theta$ , где  $\theta = (\theta_1, \dots, \theta_k)^T$  — неизвестные координаты вектора  $l$  в базисе  $Z$ . Таким образом, задача оценивания  $l \in \mathbb{R}^n$  сведена к задаче оценивания  $\theta \in \mathbb{R}^k$ .

Поставленная задача оценивания вектора  $\theta$  решается с помощью метода наименьших квадратов. *Оценка по методу наи-*

меньших квадратов равна

$$\hat{\theta} = \arg \min_{\theta} \|X - Z\theta\|^2,$$

т.е.  $\hat{l} = Z\hat{\theta} = \text{proj}_L X$  (проекция  $X$  на  $L$ ).

### ▲ Свойства оценки по методу наименьших квадратов.

1. Для оценки существует явная формула:  $\hat{\theta} = (Z^T Z)^{-1} Z^T X$ .

Действительно,

$$\begin{aligned} \|X - Z\theta\|^2 &= (X - Z\theta)^T (X - Z\theta) = \\ &= X^T X - X^T Z\theta - \theta^T Z^T X + \theta^T Z^T Z\theta = \\ &= X^T X - 2X^T Z\theta + \theta^T (Z^T Z)\theta. \end{aligned}$$

Продифференцируем полученное выражение по  $\theta_i$  для каждого  $i \in \{1, \dots, k\}$  и приравняем эти производные к нулю, чтобы найти искомый аргумент, при котором достигается минимум:

$$\begin{aligned} -2(X^T Z)_i + 2(\theta^T Z^T Z)_i &= 0, \\ X^T Z - \theta^T Z^T Z &= 0, \\ \hat{\theta} &= (Z^T Z)^{-1} Z^T X. \end{aligned}$$

2. Из свойства 1 легко получить значения вектора математических ожиданий и ковариационную матрицу оценки:  $E_{\theta, \sigma^2} \hat{\theta} = \theta$ ,  $D_{\theta, \sigma^2} \hat{\theta} = \sigma^2 (Z^T Z)^{-1}$ . Таким образом, оценка по методу наименьших квадратов является несмещенной.
3. Зная ковариационную матрицу вектора  $\hat{\theta}$ , можно найти математическое ожидание расстояния от  $X$  до  $L$ :

$$E_{\theta, \sigma^2} (\|X - Z\hat{\theta}\|^2) = (n - k)\sigma^2.$$

Последнее равенство дает несмещенную оценку

$$\hat{\sigma}^2 = \frac{1}{n - k} \|X - Z\hat{\theta}\|^2 \quad (2)$$

параметра  $\sigma^2$ .



▲ **Задача 9.1.** Имеется 2 куска сыра с весами  $a$  и  $b$ . На одних и тех же весах взвесили первый, второй и потом оба куска вместе. Найдите оценки наименьших квадратов для  $a$  и  $b$ , а также несмещенную оценку дисперсии ошибки измерений.

### Решение

Пусть взвешивания показали результаты  $X_1, X_2, X_3$ . Положим  $X = (X_1, X_2, X_3)^T$ . Тогда  $X = l + \varepsilon$ , где  $l = (a, b, a + b)^T$ . Поэтому  $l = Z\theta$ , где  $\theta = (a, b)^T$  и

$$Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Тогда

$$\hat{\theta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (Z^T Z)^{-1} Z^T X = \begin{pmatrix} \frac{2}{3}X_1 - \frac{1}{3}X_2 + \frac{1}{3}X_3 \\ -\frac{1}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}X_3 \end{pmatrix}$$

Кроме того,

$$\begin{aligned} \hat{\sigma}^2 &= \|X - Z\hat{\theta}\|^2 = (X_1 - \hat{a})^2 + (X_2 - \hat{b})^2 + (X_3 - \hat{a} - \hat{b})^2 = \\ &= \frac{X_1^2 + X_2^2 + X_3^2}{3} - \frac{2(X_1X_2 + X_1X_3 + X_2X_3)}{9}. \end{aligned}$$

**Задача решена.**

Вторую половину настоящей главы мы посвятим исключительно важному для приложения частному случаю линейной регрессионной модели, а именно *гауссовской линейной модели*. Речь идет о модели линейной регрессии  $X = l + \varepsilon$ , в которой  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ .

**Теорема 9.1** *В гауссовской линейной модели  $(\text{proj}_L X, \|\text{proj}_{L^\perp} X\|^2)$  — полная достаточная статистика.*

В силу теоремы 9.1 и теоремы Лемана-Шеффе, в гауссовской линейной модели  $\hat{\theta}$  — оптимальная оценка параметра  $\theta$ ,

$\frac{1}{n-k}\|X - Z\hat{\theta}\|^2$  — оптимальная оценка параметра  $\sigma^2$ .

Итак, мы построили точечные несмещенные оценки параметров линейной модели, которые в гауссовском случае оказались оптимальными. Построим в этом случае доверительные интервалы для упомянутых параметров. Введем для этого новые классы распределений.

**▲ Определение 9.1.** Распределение  $\Gamma(\frac{1}{2}, \frac{n}{2})$  называют *хи-квадрат распределением с  $n$  степенями свободы* (и обозначают  $\chi_n^2$ ).

Пусть  $\xi_1, \xi_2$  — независимые случайные величины, распределенные по законам  $\chi_{n_1}^2, \chi_{n_2}^2$  соответственно. Тогда случайная величина

$$\frac{\xi_1/n_1}{\xi_2/n_2}$$

имеет *распределение Фишера* со степенями свободы  $n_1, n_2$  (обозначается  $F_{n_1, n_2}$ ).

Пусть  $\xi, \eta$  — независимые случайные величины, распределенные по законам  $\mathcal{N}(0, 1), \chi_n^2$  соответственно. Тогда случайная величина

$$\frac{\xi}{\sqrt{\eta/n}}$$

имеет *распределение Стьюдента* с  $n$  степенями свободы (обозначается  $T_n$ ).

**Теорема 9.2** Пусть  $X_1, \dots, X_n$  — независимые стандартные нормальные случайные величины. Тогда случайная величина  $X_1^2 + \dots + X_n^2$  распределена по закону  $\chi^2(n)$ .

**Теорема 9.3** Пусть  $X \sim \mathcal{N}(a, \sigma^2 I_n)$ ,  $L_1 \oplus \dots \oplus L_r$  — разложение  $\mathbb{R}^n$  в прямую сумму ортогональных подпространств. Для каждого  $j \in \{1, \dots, r\}$  положим  $Y_j = \text{proj}_{L_j} X$ . Тогда  $Y_1, \dots, Y_r$  — независимые в совокупности, причем

$$\frac{1}{\sigma^2} \|Y_j - \mathbb{E}Y_j\|^2 \sim \chi_{\dim L_j}^2.$$

Из теоремы об ортогональном разложении следует, что в гауссовской линейной модели случайная величина  $\frac{n-k}{\sigma^2} \hat{\sigma}^2$  (см. 2) не зависит от  $\hat{\theta}$  и имеет распределение  $\chi_{n-k}^2$ .

**▲ Задача 9.2.** В гауссовской линейной модели найдите доверительные интервалы для параметров  $\theta_i$ ,  $i \in \{1, \dots, k\}$ , и  $\sigma^2$  уровня доверия  $\gamma$ .

### Решение

Обозначим  $a_{i,i}$ ,  $i \in \{1, \dots, n\}$ , диагональный элемент матрицы  $(Z^T Z)^{-1}$ , стоящий на  $i$ -ом месте. Тогда по теореме об ортогональном разложении случайная величина

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{a_{i,i} \hat{\sigma}^2}}$$

имеет распределение  $T_{n-k}$ . Рассмотрим  $\frac{1-\gamma}{2}$ - и  $\frac{1+\gamma}{2}$ -квантили  $u_{\frac{1-\gamma}{2}}$  и  $u_{\frac{1+\gamma}{2}}$  этого распределения. Тогда

$$P_{\theta, \sigma^2} \left( u_{\frac{1-\gamma}{2}} < \frac{\hat{\theta}_i - \theta_i}{\sqrt{a_{i,i} \hat{\sigma}^2}} < u_{\frac{1+\gamma}{2}} \right) = \gamma,$$

откуда получаем, что доверительный интервал уровня доверия  $\gamma$  для  $\theta_i$  равен

$$\left( \hat{\theta}_i - \sqrt{a_{i,i} \hat{\sigma}^2} u_{\frac{1+\gamma}{2}}, \hat{\theta}_i - \sqrt{a_{i,i} \hat{\sigma}^2} u_{\frac{1-\gamma}{2}} \right).$$

Далее, пусть  $1 - \gamma$ -квантиль распределения  $\chi_{n-k}^2$  равна  $z_{1-\gamma}$ . Тогда

$$P_{\theta, \sigma^2} \left( \frac{(n-k)\hat{\sigma}^2}{\sigma^2} > z_{1-\gamma} \right) = \gamma.$$

Доверительный интервал уровня доверия  $\gamma$  для  $\sigma^2$  равен

$$\left( 0, \frac{(n-k)\hat{\sigma}^2}{z_{1-\gamma}} \right).$$

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. В четырехугольнике  $ABCD$  независимые равные по точности измерения углов  $ABD$ ,  $DBC$ ,  $ABC$ ,  $BCD$ ,  $CDB$ ,  $BDA$ ,  $CDA$ ,  $DAB$  (в градусах) дали результаты 50.78, 30.25, 78.29, 99.57, 50.42, 40.59, 88.87, 89.86 соответственно. Считая, что ошибки измерений распределены нормально по закону  $N(0, \sigma^2)$ , найдите оптимальные оценки углов  $\beta_1 = ABD$ ,  $\beta_2 = DBC$ ,  $\beta_3 = CDB$ ,  $\beta_4 = BDA$  и неизвестной дисперсии  $\sigma^2$ .

2. Пусть

$$X_i = \beta_1 + i\beta_2 + \varepsilon_0 + \dots + \varepsilon_i,$$

$i = 0, 1, \dots, n$ , где  $\beta_1$ ,  $\beta_2$  — неизвестные параметры, а  $\varepsilon_0, \dots, \varepsilon_n$  — независимые, распределенные по закону  $N(0, \sigma^2)$  случайные величины. Сведите задачу к линейной модели и найдите оценки наименьших квадратов для  $\beta_1$  и  $\beta_2$ , а также несмещенную оценку для  $\sigma^2$ .

3. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(a, \sigma^2)$ . Докажите, что статистики  $\bar{X}$  и

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

независимы и вычислите распределение статистики  $nS^2$ .

4. Пусть  $X_1, \dots, X_n$  — выборка из  $\mathcal{N}(a, \sigma^2)$  (оба параметра неизвестны). Постройте точные доверительные интервалы для каждого из параметров  $a, \sigma^2$ .
5. Взвешивание трех грузов массами  $a$  и  $b$  производится следующим образом:  $n_1$  раз взвешивается первый груз (все ошибки измерения имеют распределение  $\mathcal{N}(0, \sigma^2)$ ),  $n_2$  раза взвешивается второй груз на тех же самых весах, затем  $n_3$  раза на других весах взвешиваются первый и второй груз вместе, все ошибки измерения на которых имеют распределение  $\mathcal{N}(0, 3\sigma^2)$ . Сведите задачу к линейной модели и найдите оценки наименьших квадратов для  $a$  и  $b$ , а также оптимальную оценку для  $\sigma^2$ .
6. Пусть  $X_i, i \in \{1, \dots, n\}$  — независимые случайные величины, распределенные по нормальному закону с параметрами  $(a+bi, \sigma^2)$ . Постройте точные доверительные интервалы для параметров  $a, b, \sigma^2$ .
7. Пусть вектор  $X = (X_1, \dots, X_n)$  имеет распределение  $(ab, \sigma^2 \Sigma)$ , где  $b$  — известный вектор размерности  $n$ ,  $\Sigma$  — известная положительно определённая матрица,  $a$  и  $\sigma^2$  — неизвестные параметры. Сведите задачу к линейной гауссовской модели и найти оценки наименьших квадратов параметров  $a$  и  $\sigma^2$ .

## 10. Проверка статистических гипотез

Разумеется, если при проведении экспериментов исследователь не имеет никаких ожиданий относительно возможных значений параметров модели, то в предыдущих главах он сможет найти ответы на многие свои вопросы. Но часто исследователь ставит эксперименты на основе имеющихся предположений относительно возможных значений параметров. Иными словами, сначала исследователь выдвигает некоторую гипотезу, а затем экспериментально ее проверяет. В этой связи важной задачей математической статистики является проверка статистических гипотез.

Формально задачу можно поставить следующим образом. Пусть  $\mathcal{P}$  — некоторое семейство распределений,  $\mathcal{P}_0 \subset \mathcal{P}$ . *Статистическая гипотеза* — это предположение общего вида о неизвестном распределении

$$H_0 : P \in \mathcal{P}_0.$$

Если известно, что  $P$  принадлежит либо множеству  $\mathcal{P}_0$ , либо множеству  $\mathcal{P}_1 \subset \mathcal{P} \setminus \mathcal{P}_0$ , то рассматривается вторая гипотеза

$$H_1 : P \in \mathcal{P}_1,$$

называемая *альтернативой*. В этом случае гипотеза  $H_0$  называется *основной*. Задача состоит в опровержении (в пользу альтернативы  $H_1$ ) или подтверждении гипотезы  $H_0$  с помощью имеющегося наблюдения  $X \sim P$ . Иными словами, гипотеза отвергается, если  $X \in S$ , где  $S$  — некоторое подмножество выборочного пространства  $\mathcal{X}$ . Множество  $S$  называется *критерием* (или *критическим множеством*) для проверки  $H_0$  (против альтернативы  $H_1$ ), а  $\mathcal{X} \setminus S$  называется *областью принятия гипотезы*.

Если семейство  $\mathcal{P}$  параметризовано:  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ , то рассматриваются параметрические гипотезы:  $H_0 : \theta \in \Theta_0$ ,  $H_1 : \theta \in \Theta_1$ , где  $\Theta_0, \Theta_1 \subset \Theta$ ,  $\Theta_0 \cap \Theta_1 = \emptyset$ .

*Ошибкой первого рода* называется ситуация, в которой отвергается верная гипотеза  $H_0$ . *Ошибкой второго рода* называется ситуация, в которой принимается неверная гипотеза  $H_0$ . Принято ошибку первого рода считать более опасной, чем ошибку второго рода. В этой связи при выборе критерия заранее фиксируется

некоторый уровень, который не должна превышать вероятность ошибки первого рода. Среди всех таких критериев выбирается тот, вероятность ошибки второго рода которого наименьшая.

## 10.1. Равномерно наиболее мощные критерии

Пусть  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ ,  $X \sim P \in \mathcal{P}$ , а  $S$  — критерий для проверки  $H_0 : \theta \in \Theta_0$  против  $H_1 : \theta \in \Theta_1$ .

**▲ Определение 10.1.** *Функцией мощности критерия  $S$  называется функция  $\beta(\theta, S) = P_\theta(X \in S)$ . Величина  $\alpha$  называется уровнем значимости критерия  $S$ , если  $\alpha \geq \beta(\theta, S)$  для любого  $\theta \in \Theta_0$ . Минимальный уровень значимости*

$$\alpha_0 = \sup_{\theta \in \Theta_0} \beta(\theta, S)$$

называется *размером критерия  $S$* . Если  $S$  и  $R$  — два критерия уровня значимости  $\alpha$ , то  $S$  *мощнее*  $R$ , если для любого  $\theta \in \Theta_1$  выполнено  $\beta(\theta, S) \geq \beta(\theta, R)$ . Критерий  $S$  называется *равномерно наиболее мощным критерием (р.н.м.к.) уровня значимости  $\alpha$* , если он мощнее любого другого критерия уровня значимости  $\alpha$ .

Как уже было замечено выше, наша задача состоит в построении равномерно наиболее мощного критерия. Далее мы рассмотрим несколько ситуаций, в которых удастся это сделать.

**▲ Задача 10.1.** Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[0, \theta]$ ,  $\theta > 0$ . Постройте р.н.м.к. уровня значимости  $\alpha$  для проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  $H_1 : \theta < \theta_0$  в виде  $S(X_1, \dots, X_n) = \{X_{(n)} \leq c\theta_0\}$ .

## Решение

Докажем, что критерий  $S = \{X_{(n)} \leq c_0\theta_0\}$ , где  $P_{\theta_0}(X_{(n)} \leq c_0\theta_0) = \alpha$ , является равномерно наиболее мощным уровня значимости  $\alpha$ . Имеем

$$P_{\theta_0}(X_{(n)} \leq c_0\theta_0) = c_0^n = \alpha.$$

Следовательно,  $c_0 = \sqrt[n]{\alpha}$ . Пусть  $R$  — любой другой критерий уровня значимости  $\alpha$ . Тогда  $P_{\theta_0}((X_1, \dots, X_n) \in R) \leq \alpha$ . Если  $\theta \leq c_0\theta_0$ , то

$$1 = P_{\theta}((X_1, \dots, X_n) \in S) \geq P_{\theta}((X_1, \dots, X_n) \in R).$$

Кроме того, если  $\theta_0 > \theta > c_0\theta_0$ , то

$$P_{\theta}((X_1, \dots, X_n) \in S) = P_{\theta}(X_{(n)} \leq c_0\theta_0) = \alpha \left(\frac{\theta_0}{\theta}\right)^n;$$

$$\begin{aligned} P_{\theta}((X_1, \dots, X_n) \in R) &= \int_{[0, \theta]^n} \frac{1}{\theta^n} I((x_1, \dots, x_n) \in R) dx_1 \dots dx_n = \\ &= \frac{\theta_0^n}{\theta^n} \int_{[0, \theta]^n} \frac{1}{\theta_0^n} I((x_1, \dots, x_n) \in R) dx_1 \dots dx_n \leq \\ &\leq \frac{\theta_0^n}{\theta^n} \int_{[0, \theta_0]^n} \frac{1}{\theta_0^n} I((x_1, \dots, x_n) \in R) dx_1 \dots dx_n = \\ &= \frac{\theta_0^n}{\theta^n} P_{\theta_0}((X_1, \dots, X_n) \in R) \leq \alpha \frac{\theta_0^n}{\theta^n} = P_{\theta}((X_1, \dots, X_n) \in S). \end{aligned}$$

**Задача решена.**

## 10.2. Проверка простых гипотез

Пусть  $\{P_{\theta}, \theta \in \Theta\}$  — доминируемое семейство распределений (т.е. либо все распределения абсолютно непрерывные, либо все дискретные) с “обобщенными” плотностями  $p_{\theta}$  (в случае дискретного распределения  $p_{\theta}(x) = P_{\theta}(X = x)$ ). Пусть  $H_0 : \theta = \theta_0$  и  $H_1 : \theta = \theta_1$  (такие гипотезы называются *простыми*). Будем считать, что множество  $\{x : p_{\theta}(x) > 0\}$  не зависит от  $\theta \in \{\theta_0, \theta_1\}$  (это исключает тривиальные ситуации, когда при ровно одной из гипотез некоторые значения  $x$  запрещены, а значит, наблю-



дая такие значения в опыте, мы однозначно определяем, какая из гипотез верна).

**Лемма 10.1** *Если существует такое число  $c > 0$ , что*  
**(Неймана-Пирсона)** 
$$P_{\theta_0}(p_{\theta_1}(X) - cp_{\theta_0}(X) \geq 0) = \alpha,$$

*то*

$$S = \{x \in \mathcal{X} : p_{\theta_1}(x) - cp_{\theta_0}(x) \geq 0\}$$

*является р.н.м.к. уровня значимости  $\alpha$  для проверки  $H_0$  против  $H_1$ .*

**▲ Задача 10.2.** Пусть  $X_1, \dots, X_n$  — выборка из бернуллиевского закона с вероятностью успеха  $\theta$ . Постройте р.н.м.к. для проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  $H_1 : \theta = \theta_1$ , если а)  $\theta_0 < \theta_1$ , б)  $\theta_1 < \theta_0$ .

### Решение

Имеем  $p_{\theta}(X_1, \dots, X_n) = \theta^{\sum X_i} (1 - \theta)^{n - \sum X_i}$ . По лемме Неймана-Пирсона, р.н.м.к. равен

$$S = \left\{ \left( \frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right)^{\sum X_i} \geq c \right\},$$

причем вероятность последнего события равна  $\alpha$ . В случае а)  $\theta_0 < \theta_1$  имеем  $\alpha = P(\sum X_i \geq \tilde{c})$  для некоторой константы  $\tilde{c}$  и  $S = \{\sum X_i \geq \tilde{c}\}$ . Поэтому  $\tilde{c} = z_{1-\alpha} + 1$ , где  $z_{1-\alpha}$  —  $(1 - \alpha)$ -квантиль распределения  $\text{Bin}(n, \theta_0)$ . Заметим, что в этом случае число  $\alpha$  должно быть подобрано таким образом, что с  $1 - \alpha$  совпадает одно из значений функции распределения  $\text{Bin}(n, \theta)$ .

В случае б)  $\theta_1 < \theta_0$  имеем  $\alpha = P(\sum X_i \leq \tilde{c})$ ,  $S = \{\sum X_i \leq \tilde{c}\}$ , т.е.  $\tilde{c}$  —  $\alpha$ -квантиль распределения  $\text{Bin}(n, \theta_0)$ . Заметим, что в этом случае число  $\alpha$  должно быть подобрано таким образом, что с ним совпадает одно из значений функции распределения  $\text{Bin}(n, \theta)$ .

**Задача решена.**

### 10.3. Проверка сложных гипотез

Пусть  $\Theta \subset \mathbb{R}$ . Как и в случае простых гипотез, будем предполагать, что  $\{P_\theta, \theta \in \Theta\}$  — доминируемое семейство распределений с плотностями  $p_\theta$ , и считать, что множество  $\{x : p_\theta(x) > 0\}$  не зависит от  $\theta \in \Theta$ .

**▲ Определение 10.2.** Семейство  $\{P_\theta, \theta \in \Theta\}$  обладает *неубывающим (невозрастающим) отношением правдоподобия по статистике  $T(X)$*  (в обоих случаях говорят, что семейство обладает *монотонным отношением правдоподобия*), если функция

$$\frac{p_{\theta'}(X)}{p_{\theta''}(X)} = L_{\theta', \theta''}(T(X))$$

является неубывающей (невозрастающей) функцией от статистики  $T(X)$  для всех  $\theta'' < \theta'$ , принадлежащих  $\Theta$ .

**Теорема  
10.1 (о монотонном  
отношении  
правдоподобия)**

*Пусть  $\{P_\theta, \theta \in \Theta\}$  — семейство с неубывающим отношением правдоподобия по статистике  $T(X)$ , а  $\alpha < 1$  — некоторое положительное число. Пусть, кроме того,  $H_0 : \theta \leq \theta_0$  (или  $\theta = \theta_0$ ),  $H_1 : \theta > \theta_0$ . Если существует такое  $c$ , что  $P_{\theta_0}(T(X) \geq c) = \alpha$ , то  $S = \{T(X) \geq c\}$  есть р.н.м.к. уровня значимости  $\alpha$  для проверки  $H_0$  против  $H_1$ .*

**▲ Задача 10.3.** Пусть  $\{P_\theta, \theta \in \Theta\}$  — семейство с неубывающим отношением правдоподобия по статистике  $T(X)$ , а  $\alpha < 1$  — некоторое положительное число. Постройте р.н.м.к. уровня значимости  $\alpha$  для проверки гипотезы  $H_0$  против альтернативы  $H_1$ , где  $H_0 : \theta \geq \theta_0$  (или  $\theta = \theta_0$ ),  $H_1 : \theta < \theta_0$ .

## Решение

Рассмотрим семейство распределений  $\{\tilde{P}_{\tilde{\theta}}, \tilde{\theta} \in -\Theta\}$  (здесь  $-\Theta = \{-\theta : \theta \in \Theta\}$ ), где  $\tilde{P}_{\tilde{\theta}} = P_{-\tilde{\theta}}$  для любого  $\tilde{\theta} \in -\Theta$ . Обозначим  $\tilde{p}_{\tilde{\theta}}$  плотность распределения  $\tilde{P}_{\tilde{\theta}}$ . Разумеется, для любых  $\tilde{\theta}' > \tilde{\theta}''$  из  $-\Theta$  выполнено

$$\frac{\tilde{p}_{\tilde{\theta}'}(X)}{\tilde{p}_{\tilde{\theta}''}(X)} = \frac{p_{-\tilde{\theta}'}(X)}{p_{-\tilde{\theta}''}(X)}.$$

Последняя функция, разумеется, не возрастает по некоторой статистике  $T(X)$ , а, следовательно, не убывает по  $-T(X)$ . Кроме того, в новых обозначениях гипотезы  $H_0$  и  $H_1$  имеют вид  $\tilde{\theta} \leq \tilde{\theta}_0$  (или  $\tilde{\theta} = \tilde{\theta}_0$ ) и  $\tilde{\theta} > \tilde{\theta}_0$  соответственно. Следовательно, мы находимся в условиях теоремы 10.1. Поэтому р.н.м.к. уровня значимости  $\alpha$  равен  $S = \{-T(X) \geq c\} = \{T(X) \leq \tilde{c}\}$ , где  $P_{\theta_0}(T(X) \leq \tilde{c}) = \alpha$ .

**Задача решена.**

## ▲ Задачи для самостоятельного решения

1. Имеется  $X_1$  — выборка объема 1. Основная гипотеза  $H_0$  состоит в том, что  $X_1$  имеет равномерное распределение на отрезке  $[0, 1]$ , альтернатива — в том, что  $X_1$  имеет показательное распределение с параметром 1. Постройте наиболее мощный критерий уровня значимости  $\alpha$  для различения этих гипотез и вычислите его мощность.
2.  $X_1, \dots, X_n$  — выборка из экспоненциального распределения с параметром  $\theta$ . Постройте равномерно наиболее мощный критерий уровня значимости  $\alpha$  проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  
а)  $H_1 : \theta > \theta_0$ , б)  $H_1 : \theta < \theta_0$ .
3.  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(\theta, 1)$ . Постройте равномерно наиболее мощный критерий уровня значимости  $\alpha$  проверки  
а) гипотезы  $H_0 : \theta \geq \theta_0$  против альтернативы  $H_1 : \theta < \theta_0$ ,  
б) гипотезы  $H_0 : \theta \leq \theta_0$  против альтернативы  $H_1 : \theta > \theta_0$ .

4.  $X_1, \dots, X_n$  — выборка из распределения  $\text{Bern}(\theta)$ . Докажите, что не существует равномерного наиболее мощного критерия произвольного уровня значимости  $\alpha$  для проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  $H_1 : \theta \neq \theta_0$ .
5.  $X_1, \dots, X_n$  — выборка из распределения  $\mathcal{N}(0, \theta)$ . Постройте равномерно наиболее мощный критерий уровня значимости  $\alpha$  проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  $H_1 : \theta = \theta_1$  и найдите его мощность.
6. Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[0, \theta]$ ,  $\theta > 0$ . Постройте р.н.м.к. уровня значимости  $\alpha$  для проверки гипотезы  $H_0 : \theta = \theta_0$  против альтернативы  $H_1 : \theta \neq \theta_0$  в виде

$$S(X_1, \dots, X_n) = \{X_{(n)} \leq c\theta_0\} \cup \{X_{(n)} > \theta_0\}.$$

7. Пусть  $\{P_\theta, \theta \in \Theta\}$  — семейство с невозрастающим отношением правдоподобия по статистике  $T(X)$ , а  $\alpha < 1$  — некоторое положительное число. Постройте р.н.м.к. уровня значимости  $\alpha$  для проверки гипотезы  $H_0$  против альтернативы  $H_1$ , где а)  $H_0 : \theta \leq \theta_0$  (или  $\theta = \theta_0$ ),  $H_1 : \theta > \theta_0$ ; б)  $H_0 : \theta \geq \theta_0$  (или  $\theta = \theta_0$ ),  $H_1 : \theta < \theta_0$ .
8. Показать, что любой равномерно наиболее мощный несмещённый (т.е.  $\inf_{\theta \in \Theta_0} \beta(\theta, S) \geq \sup_{\theta \in \Theta_1} \beta(\theta, S)$ ) критерий  $S$  является допустимым, т.е. не существует другого критерия  $R$ , который был бы не менее мощен, чем  $S$ , при всех альтернативах и более мощен хотя бы при одной из альтернатив.

## 11. Проверка линейных гипотез в гауссовской регрессионной модели

В этой главе речь пойдет о проверке гипотез в гауссовской линейной модели. При применении этой модели на практике, разумеется, возникает необходимость проверять различные гипотезы о значениях вектора  $\theta$ . В нашем учебном пособии мы поговорим о проверке линейных гипотез.

Рассмотрим задачу оценивания параметров гауссовской линейной модели  $X = Z\theta + \varepsilon$ , где  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ ,  $\theta = (\theta_1, \dots, \theta_k)^T$ ,  $k \leq n$ . Мы хотим построить критерий для проверки линейной гипотезы следующего вида:  $H_0 : T\theta = \tau$ , где  $T$  — матрица размера  $m \times k$ ,  $m \leq k$ ,  $\text{rank}(T) = m$ ,  $\tau \in \mathbb{R}^m$ .

В главе 9 было установлено, что оценка

$$\hat{\theta} = (Z^T Z)^{-1} Z^T X$$

параметра  $\theta$ , полученная методом наименьших квадратов, является оптимальной, причем

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2 (Z^T Z)^{-1}).$$

Несложно показать, что поскольку  $T$  — это линейное преобразование, то  $\hat{t} = T\hat{\theta}$  — оптимальная оценка  $T\theta$ .

**▲ Задача 11.1.** Докажите, что  $\hat{t} = T\hat{\theta}$  — оптимальная оценка параметра  $T\theta$ .

## Решение

Несмещенность следует из свойства линейности математического ожидания. Проверим, что оценка  $\hat{t}$  является наилучшей. Пусть  $a \in \mathbb{R}^m$ . Тогда

$$\begin{aligned} \mathbb{E}_{\theta, \sigma^2} (\langle \hat{t} - T\theta, a \rangle)^2 &= \mathbb{E}_{\theta, \sigma^2} \left( (T(\hat{\theta} - \theta))^T a \right)^2 = \\ &= \mathbb{E}_{\theta, \sigma^2} \left( (\hat{\theta} - \theta)^T (T^T a) \right)^2 = \mathbb{E}_{\theta, \sigma^2} \left( \langle \hat{\theta} - \theta, T^T a \rangle \right)^2. \end{aligned}$$

Так как матрица  $T$  — полного ранга, то для любого вектора  $t^* \in \mathbb{R}^m$  существует такой вектор  $\theta^* \in \mathbb{R}^k$ , что  $t^* = T\theta^*$ . Пусть  $\mathbb{E}_{\theta, \sigma^2} t^* = T\theta$ . Разумеется, вектор  $\tilde{\theta} = \theta^* - \mathbb{E}_{\theta, \sigma^2} \theta^* + \theta$  является несмещенной оценкой параметра  $\theta$ , причем  $t^* = T\tilde{\theta}$ . Тогда

$$\begin{aligned} \mathbb{E}_{\theta, \sigma^2} (\langle \hat{t} - T\theta, a \rangle)^2 &= \mathbb{E}_{\theta, \sigma^2} \left( \langle \hat{\theta} - \theta, T^T a \rangle \right)^2 \leq \\ &\leq \mathbb{E}_{\theta, \sigma^2} \left( \langle \tilde{\theta} - \theta, T^T a \rangle \right)^2 = \mathbb{E}_{\theta, \sigma^2} (\langle t^* - T\theta, a \rangle)^2, \end{aligned}$$

причем хотя бы для одной пары  $(\theta, a)$  неравенство является строгим.

**Задача решена.**

В предположении верности гипотезы  $H_0$ ,

$$\hat{t} \sim \mathcal{N}(\tau, \sigma^2 T(Z^T Z)^{-1} T^T).$$

Обозначим  $B = T(Z^T Z)^{-1} T^T$ .

Так как матрицы  $T$  и  $Z$  имеют полный ранг, то, в силу определения гауссовского вектора, матрица  $B$  является симметричной и положительно определенной. Следовательно, она обратима, обратная матрица  $B^{-1}$  также симметрична и положительно определена, а, следовательно, имеет единственный положительно определенный симметричный квадратный корень  $\sqrt{B^{-1}}$ . Поэтому случайный вектор

$$\eta = \frac{1}{\sigma} \sqrt{B^{-1}} (\hat{t} - \tau)$$

имеет распределение  $\mathcal{N}(0, I_m)$ . Тогда

$$\frac{1}{\sigma^2}(\hat{t} - \tau)^T B^{-1}(\hat{t} - \tau) \sim \chi_m^2.$$

Последнее выражение является измеримой функцией от  $\hat{\theta}$ , а, следовательно, по теореме об ортогональном разложении гауссовского вектора, оно не зависит от  $X - Z\hat{\theta}$ . С другой стороны, как было замечено в главе 10,

$$\frac{1}{\sigma^2} \|X - Z\hat{\theta}\|^2 \sim \chi_{n-k}^2.$$

Рассмотрим отношение двух полученных статистик, нормированных числом степеней свободы соответствующих хи-квадрат распределений

$$F_T = \frac{(\hat{t} - \tau)^T B^{-1}(\hat{t} - \tau)}{\|X - Z\hat{\theta}\|^2} \frac{n - k}{m} \sim F_{m, n-k}.$$

Критерий для проверки гипотезы  $H_0$  мы будем строить с помощью статистики  $F_T$ .

Пусть  $u_{1-\alpha}$  —  $(1 - \alpha)$ -квантиль  $F_{m, n-k}$ . Тогда критерий для проверки гипотезы  $H_0$  (*F-критерий*) имеет вид

$$S = \{F_T > u_{1-\alpha}\}.$$

**▲ Задача 11.2.** Выпущена новая партия монет. Перед их использованием было решено провести исследование с целью проверки того, что монеты одинаковы. На весах с ошибкой измерений, распределенной по закону  $\mathcal{N}(0, \sigma^2)$ , были взвешены две монеты, причем первая была взвешена  $n$  раз, а вторая —  $m$  раз. Все измерения были произведены независимо. Постройте *F-критерий* для проверки гипотезы о том, что массы монет одинаковы.

## Решение

Пусть  $X_1, \dots, X_n$  и  $Y_1, \dots, Y_m$  — независимые выборки из распределений  $\mathcal{N}(a_1, \sigma^2)$  и  $\mathcal{N}(a_2, \sigma^2)$  соответственно. Рассматривается гипотеза  $H_0 : a_1 = a_2$ . Составим вектор  $W = (W_1, \dots, W_{n+m})^T = (X_1, \dots, X_n, Y_1, \dots, Y_m)^T$  и рассмотрим линейную модель  $W = Z\theta$ , где

$$\theta = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 \dots 1 & 0 \dots 0 \\ \underbrace{0 \dots 0}_n & \underbrace{1 \dots 1}_m \end{pmatrix}^T.$$

Предположение гипотезы можно представить в виде  $T\theta = \tau$ , где  $T = (1, -1)$ ,  $\tau = 0$ .

Имеем

$$B = T(Z^T Z)^{-1} T^T = (1 \quad -1) \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{m} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{n} + \frac{1}{m}.$$

Кроме того,

$$\hat{t} = T\hat{\theta} = T(Z^T Z)^{-1} Z^T W = \bar{X} - \bar{Y}.$$

Окончательно получаем

$$F_T = \frac{(\bar{X} - \bar{Y})^2 (\frac{1}{n} + \frac{1}{m})}{\|W - Z\hat{\theta}\|^2} (n + m - 2),$$

где

$$\|W - Z\hat{\theta}\|^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2 = ns_X^2 + ms_Y^2.$$

Отсюда получаем  $F$ -критерий:  $\{F_T > u_{1-\alpha}\}$ , где  $u_{1-\alpha}$  —  $(1 - \alpha)$ -квантиль  $F_{1, n+m-2}$ .

**Задача решена.**

В некоторых задачах подсчет статистики  $F_T$  затруднителен из-за того, что приходится вычислять элементы матрицы  $B^{-1}$ . В этих случаях следует воспользоваться следующей теоремой.



**Теорема  
11.1**

*В предположении верности гипотезы  $H_0$  имеем место равенство*

$$S_T = (\hat{t} - \tau)^T B^{-1}(\hat{t} - \tau) + \|X - Z\hat{\theta}\|^2,$$

где

$$S_T = \min_{\theta: T\theta = \tau} \|X - Z\theta\|^2.$$

Величина  $S_T$  называется *условной оценкой наименьших квадратов*.

**▲ Задача 11.3.** Про измеряемую величину  $x$  известна зависимость от температуры  $t$ :

$$x(t) = \beta_1 + \beta_2 t + \beta_3 t^2.$$

Произведена серия независимых экспериментов при значениях температуры, равных

$$t_1 = -1, \quad t_2 = 0, \quad t_3 = 1, \quad t_4 = 2, \quad t_5 = 3.$$

Получены соответствующие результаты:

$$X_1 = 1, \quad X_2 = 6, \quad X_3 = 10, \quad X_4 = 14, \quad X_5 = 19.$$

Предполагается, что ошибки измерений независимы и распределены по закону  $N(0, 1)$ . Кроме того, найдена оптимальная оценка  $\hat{\sigma}^2 = 0.4$  дисперсии ошибки  $\sigma^2$ . С помощью  $F$ -критерия проверьте гипотезу  $H_0$  о том, что  $\beta_1 = \beta_2$  и  $\beta_3 = 0$ , на уровне значимости  $\alpha = 0.1$ .

## Решение

Пусть  $X = (X_1, \dots, X_5)^T$ . Рассмотрим линейную модель  $X = Z\theta$ , где

$$\theta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 4 & 9 \end{pmatrix}^T.$$

Предположение гипотезы можно представить в виде  $T\theta = \tau$ , где

$$\tau = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Имеем

$$\begin{aligned} S_T &= \min_{\theta: \beta_1=\beta_2, \beta_3=0} \|X - Z\theta\|^2 = \min_{\beta} \sum_{i=1}^5 (X_i - \beta(1 + t_i))^2 = \\ &= \min_{\beta} (30\beta^2 - 2\beta(X_2 + 2X_3 + 3X_4 + 4X_5) + \sum_{i=1}^5 X_i^2) = \\ &= \min_{\beta} (30\beta^2 - 288\beta + 694) = 694 - 691.2 = 2.8. \end{aligned}$$

Кроме того, как известно из главы 9,

$$\hat{\sigma}^2 = \frac{1}{2} \|X - Z\hat{\theta}\|^2 = 0.4.$$

По теореме 11.1 получаем

$$F_T = \frac{2.8}{0.8} - 1 = 2.5.$$

Поэтому если  $2.5 > u_{1-\alpha}$ , то гипотеза отвергается, где  $u_{1-\alpha} - (1 - \alpha)$ -квантиль  $F_{2,2}$ . Так как  $u_{0.1} \approx 9$ , то на уровне значимости  $\alpha = 0.1$  гипотезу нужно принять.

**Задача решена.**

## ▲ Задачи для самостоятельного решения

1. Постройте  $F$ -критерий уровня значимости  $\alpha$  для проверки гипотезы  $H_0 : \beta_2 = \beta_1$  в задаче 2 из задач для самостоятельного решения раздела 9.
2.  $X_1, \dots, X_n$  — выборка из распределения  $\mathcal{N}(a_1, \sigma^2)$ ,  $Y_1, \dots, Y_m$  — выборка из распределения  $\mathcal{N}(a_2, \sigma^2)$ ,  $Z_1, \dots, Z_k$  — выборка из распределения  $\mathcal{N}(a_3, \sigma^2)$ . Постройте  $F$ -критерий размера  $\alpha$  для проверки гипотезы а)  $H_0 : a_1 = a_2$  и  $a_1 + a_2 = a_3$ , б)  $H_0 : a_1 = 2a_2$  и  $a_1 + 3a_2 = a_3$ .
3. Решите задачу 11.2 с помощью теоремы 11.1.
4. Пусть  $X_i \sim N(a, i\sigma^2)$ ,  $i = 0, 1, \dots, n$ ,  $Y_j \sim N(jb, \sigma^2)$ ,  $j = 1, \dots, m$ , — независимые случайные величины, где  $a, b, \sigma^2$  — неизвестные параметры. Сведите задачу к линейной модели и постройте  $F$ -критерий размера  $\alpha$  для проверки гипотезы  $H_0 : a + b = 1$ .
5. Используя метод линейной регрессии, постройте приближение функции  $f(x)$  многочленом третьей степени по следующим данным:

$f(x_i)$	3.9	5.0	5.7	6.5	7.1	7.6	7.8	8.1	8.4
$x_i$	4.0	5.2	6.1	7.0	7.9	8.6	8.9	9.5	9.9

6. Пусть  $X_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ ,  $i = 1, \dots, n_j$ ,  $j = 1, \dots, k$ , — независимые случайные величины, а  $\{\mu_j\}_{j=1}^k, \sigma^2$  — неизвестные параметры. Обозначим  $N = \sum_{j=1}^k n_j$ . Для проверки гипотезы однородности  $H_0 : \mu_1 = \dots = \mu_k$  используется  $F$ -критерий однофакторного дисперсионного анализа со следующей статистикой

$$R = \frac{\frac{1}{k-1} \sum_{j=1}^k n_j (X_{..j} - X_{..})^2}{\frac{1}{N-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - X_{.j})^2},$$

где

$$X_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \quad X_{..} = \frac{1}{k} \sum_{j=1}^k X_{.j}.$$

Докажите, что при верности  $H_0$  выполнено  $R \sim F_{k-1, N-k}$ .

## 12. Критерии согласия

Следующий класс гипотез, который мы рассмотрим в настоящем учебном пособии, это гипотезы о совпадении неизвестного распределения с известным. Иными словами, пусть  $X$  — наблюдение с неизвестным распределением, функция распределения которого равна  $F$ . Выдвигается гипотеза  $H_0 : F = F_0$ , где функция распределения  $F_0$  (соответствующее распределение мы обозначим  $P_0$ ) известна. Альтернативные распределения никак не конкретизируются, априори может быть ясен лишь тип распределения (например, дискретный или абсолютно непрерывный) и множество значений наблюдения  $X$ . Критерии проверки таких гипотез называются *критериями согласия*. Наиболее известными среди них являются критерий Колмогорова и критерий хи-квадрат, о которых мы и поговорим в этой главе.

В случае критериев согласия принято использовать те, которые обладают свойством состоятельности.

**▲ Определение 12.1.** Если  $X = (X_1, \dots, X_n)$  — выборка, то критерий  $S_n$  проверки простой гипотезы  $H_0 : P = P_0$  (фактически, последовательность критериев) называется *состоятельным*, если функция мощности  $\beta(P, S_n) = P(X \in S_n)$  стремится к 1 при  $n \rightarrow \infty$  для любого  $P \neq P_0$ .

### 12.1. Критерий согласия Колмогорова

Пусть имеется выборка  $X = (X_1, \dots, X_n)$  из неизвестного распределения на  $\mathbb{R}$  с непрерывной функцией распределения  $F$ . Построим критерий для проверки гипотезы  $H_0 : F = F_0$ .

**▲ Определение 12.2.** *Эмпирической функцией распределения* выборки  $X$  называется случайная величина

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}.$$

Обозначим

$$D_n = \sup_x |F_n^*(x) - F_0(x)|, \quad K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}.$$

Можно показать, что  $K(t)$  является функцией распределения (называется *функцией распределения Колмогорова*).

**Теорема 12.1 (А. Колмогорова)** *В предположении верности гипотезы  $H_0$  имеет место равенство*

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq t) = K(t).$$

▲ **Критерий Колмогорова** записывается следующим образом:  $\{\sqrt{n}D_n > K_{1-\alpha}\}$ , где  $K_{1-\alpha} - (1 - \alpha)$ -квантиль функции распределения Колмогорова.

Из теоремы 12.1 следует, что критерий Колмогорова является состоятельным.

Напоследок заметим, что критерий Колмогорова принято применять, если  $n \geq 20$ .

▲ **Задача 12.1.** Докажите, что

$$D_n = \max_{1 \leq k \leq n} \max \left\{ \frac{k}{n} - F(X_{(k)}), F(X_{(k)}) - \frac{k-1}{n} \right\}. \quad (3)$$

### Решение

Разумеется, для любого  $k \in \{1, \dots, n-1\}$  на интервале  $[X_{(k)}, X_{(k+1)})$  значение функции  $F_n^*$  постоянно и равно  $\frac{k}{n}$ , а функция  $F(x)$  не убывает, поэтому на этом интервале любое значение  $|F_n^*(x) - F(x)|$  не превосходит  $\max \left\{ \frac{k}{n} - F(X_{(k)}), F(X_{(k+1)}) - \frac{k+1}{n} \right\}$ . Кроме того, на интервале  $(-\infty, X_{(1)})$  значение функции  $F_n^*$  равно 0, а следовательно, на этом интервале  $|F_n^*(x) - F(x)| \leq F(X_{(1)})$ . Наконец, на интервале  $[X_{(n)}, \infty)$  значение функции  $F_n^*$  равно 1, а следовательно, на этом интервале  $|F_n^*(x) - F(x)| \leq 1 - F(X_{(n)})$ . Так как все значения из правой части равенства (3) достигаются функцией  $|F_n^*(x) - F(x)|$ , то справедливость этого равенства доказана.

**Задача решена.**

## 12.2. Критерий согласия хи-квадрат

Пусть  $X_1, \dots, X_n$  — выборка из схемы Бернулли с  $m \geq 2$  исходами, т.е.  $a_1, \dots, a_m$  — исходы и  $P(X_i = a_j) = p_j$  для  $j = 1, \dots, m$ . Построим критерий для проверки гипотезы

$$H_0 : p_j = p_j^0, \quad j \in \{1, \dots, m\}.$$

**▲ Определение 12.3.** *Статистикой хи-квадрат называется*

$$\hat{\chi} = \sum_{i=1}^m \frac{(\mu_j - np_j^0)^2}{np_j^0},$$

где

$$\mu_j = \sum_{i=1}^n I\{X_i = a_j\}.$$

**Теорема 12.2 (К. Пирсона)** *В предположении верности гипотезы  $H_0$  имеет место сходимость по распределению*

$$\hat{\chi} \xrightarrow[n \rightarrow \infty]{d} \chi_{m-1}^2.$$

**▲ Критерий хи-квадрат Пирсона** записывается следующим образом:  $\{\hat{\chi} > u_{1-\alpha}\}$ , где  $u_{1-\alpha}$  —  $(1 - \alpha)$ -квантиль распределения  $\chi_{m-1}^2$ .

Из теоремы 12.2 следует, что критерий хи-квадрат является состоятельным.

Напоследок заметим, что критерий хи-квадрат принято применять, если  $n \geq 50$  и  $\mu_j \geq 5$  для всех  $j \in \{1, \dots, m\}$ .

**▲ Задача 12.2.** По статистике, собранной в психиатрической больнице в течение года, количество пациентов, поступивших в отделение интенсивной терапии, имело следующее распределение по дням недели:

ПН — 36, ВТ — 53, СР — 35, ЧТ — 26, ПТ — 30, СБ — 44, ВС — 28.

Согласуются ли эти данные с гипотезой о том, что попадание в отделение не зависит от дня недели на уровне значимости 0.05

и 0.01?

### Решение

Положим  $m = 7$ ,  $p_j^0 = \frac{1}{7}$ ,  $j \in \{1, \dots, m\}$ . Общее количество наблюдений равно

$$n = 36 + 53 + 35 + 26 + 30 + 44 + 28 = 252.$$

Таким образом,  $np_j^0 = 36$  для всех  $j \in \{1, \dots, m\}$ . Имеем

$$\hat{\chi} = 0 + 8.03 + 0.03 + 2.78 + 1 + 1.78 + 1.78 = 15,4.$$

Квантили распределения хи-квадрат с 6 степенями свободы таковы:  $u_{0.95} \approx 12,6$ ,  $u_{0.99} \approx 16,8$ . Таким образом, на уровне значимости 0.05 мы гипотезу о равномерном поступлении пациентов отвергаем, а на уровне 0.01 — не отвергаем (но во втором случае вероятность ошибки второго рода, разумеется, существенно увеличивается, поэтому высокий уровень значимости не является целесообразным, если наблюдений не так много).

**Задача решена.**

Критерий хи-квадрат Пирсона можно применять и в случае сложных гипотез

$$H_0 : p_j = p_j^0(\theta), \quad j \in \{1, \dots, m\}, \quad \theta = (\theta_1, \dots, \theta_r) \in \Theta, \quad r < m - 1.$$

При проверке таких гипотез используется также статистика  $\hat{\chi} = \hat{\chi}(\theta)$ , при вычислении которой  $\theta$  заменяется на оценку максимального правдоподобия

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \prod_{j=1}^m (p_j^0(\theta))^{\mu_j}.$$

**Теорема  
12.3 (Р.  
Фишера)**

*Пусть выполнены следующие условия:*

1. гипотеза  $H_0$  верна;
2. для любого  $\theta \in \Theta$  справедливо равенство  $\sum_{j=1}^m p_j(\theta) = 1$ ;
3. существует такое положительное число  $c$ , что для всех  $j \in \{1, \dots, m\}$  и  $\theta \in \Theta$  выполнено  $p_j(\theta) \geq c$ ;
4. для любых  $j \in \{1, \dots, m\}$ ,  $k, l \in \{1, \dots, r\}$  существуют непрерывные производные  $\frac{\partial p_j^0(\theta)}{\partial \theta_k}$ ,  $\frac{\partial^2 p_j^0(\theta)}{\partial \theta_k \partial \theta_l}$ ;
5. матрица, составленная из частных производных  $\frac{\partial p_j^0(\theta)}{\partial \theta_k}$  ( $j \in \{1, \dots, m\}$ ,  $k \in \{1, \dots, r\}$ ), имеет ранг  $r$  для всех  $\theta \in \Theta$ .

*Тогда*

$$\hat{\chi}(\theta^*) \xrightarrow[n \rightarrow \infty]{d} \chi_{m-1-r}^2.$$

В случае сложных гипотез критерий хи-квадрат также применяется, если  $n \geq 50$  и  $\mu_j \geq 5$  для всех  $j \in \{1, \dots, m\}$ .

**▲ Задача 12.3.** Проверьте на уровне значимости  $\alpha$  гипотезу о том, что измеряемая величина  $X$  имеет биномиальное распределение с параметрами  $(2, \theta)$ , если при проведении  $n = 128$  измерений она  $\mu_1 = n/4$  раз приняла значение 0,  $\mu_2 = n/4$  раз — значение 1 и  $\mu_3 = n/2$  раз — значение 2.



## Решение

В задаче речь идет о проверке гипотезы

$$H_0 : p_1^0(\theta) = (1 - \theta)^2, p_2^0(\theta) = 2\theta(1 - \theta), p_3^0 = \theta^2.$$

Найдем оценку максимального правдоподобия:

$$\begin{aligned} \frac{\partial}{\partial \theta} \left( \ln \prod_{j=1}^3 (p_j^0(\theta))^{\mu_j} \right) &= -\frac{2\mu_1}{1 - \theta} + \frac{\mu_2}{\theta} - \frac{\mu_2}{1 - \theta} + \frac{2\mu_3}{\theta} = \\ &= \frac{2\mu_3 + \mu_2 - 2n\theta}{\theta(1 - \theta)}. \end{aligned}$$

Следовательно,  $\theta^* = \frac{2\mu_3 + \mu_2}{2n} = \frac{1}{2}$ . Окончательно получаем значение статистики

$$\begin{aligned} \hat{\chi}(\theta^*) &= \sum_{i=1}^3 \frac{(\mu_j - np_j^0(1/2))^2}{np_j^0(1/2)} = \\ &= \frac{(n/4 - n/2)^2}{n/2} + \frac{(n/2 - n/4)^2}{n/4} = \frac{n}{8} + \frac{n}{4} = \frac{3n}{8} = 48. \end{aligned}$$

Так как 0.99-квантиль распределения  $\chi_1^2$  с точностью до четвертого знака после запятой равна 6.6349, то гипотезу нельзя принять даже на уровне значимости 0.01.

**Задача решена.**

Критерий хи-квадрат, как и критерий согласия Колмогорова, применяется для проверки гипотезы о равенстве распределения, из которого берется наша выборка, какому-то определенному распределению  $P_0$ . В отличие от критерия Колмогорова, критерий хи-квадрат не требует больших вычислений, но является менее “точным”. Работать с критерием хи-квадрат в случае произвольного класса распределений можно следующим образом: область значений выборки разбивается на несколько интервалов, после чего вычисляется число членов выборки, попавших в каждый интервал. Полученные значения берутся в качестве  $\mu_j$ . В качестве  $p_j^0$  берутся вероятности попадания случайной величины с распределением  $P_0$  в  $j$ -ый интервал.

## ▲ Задачи для самостоятельного решения

1. Докажите, что в предположении гипотезы  $H_0 : F = F_0$  для любого  $x \in \mathbb{R}$  выполнено

$$F_n^*(x) \xrightarrow[n \rightarrow \infty]{\text{п.н.}} F_0(x).$$

2. С помощью теоремы 12.1 докажите состоятельность критерия Колмогорова.
3. Имеется выборка  $X_1, X_2, X_3$  объема 3. Для проверки гипотезы о том, что выборка взята из равномерного на отрезке  $[0,1]$  распределения, используется следующий вариант критерия Колмогорова: гипотеза о равномерности отвергается, если

$$\sup_{y \in [0,1]} |F_3^*(y) - y| > 1/3.$$

Чему равен размер этого критерия?

4. Докажите, что при условии  $0 \leq X_{(1)} \leq X_{(n)} \leq 1$  справедливо равенство

$$\int_0^1 (F_n^*(y) - y)^2 dy = \frac{1}{12n^2} + \frac{1}{n} \sum_{k=1}^n (X_{(k)} - (2k-1)/n)^2.$$

(с помощью этого представления часто вычисляется значение статистики  $\omega^2$ , которая используется в критерии Крамера–Мизеса–Смирнова).

5. С помощью теоремы 12.2 докажите состоятельность хи-квадрат критерия.
6. Цифры  $0, 1, 2, \dots, 9$  среди 800 первых десятичных знаков числа  $\pi$  появились

74, 92, 83, 79, 80, 73, 77, 75, 76, 91

раз соответственно. С помощью хи-квадрат критерия проверьте гипотезу о согласии этих данных с законом равномерного распределения на множестве  $\{0, 1, \dots, 9\}$  на уровне значимости а) 0.05, б) 0.5, в) 0.8.

7. Профессиональный дантист научился выбивать зубы мудрости кулаком. Известно, что 52 зуба мудрости он выбил с первой попытки, 31 – со второй, 3 – с третьей, на выбивание оставшихся 5 зубов ему потребовалось более 4 попыток. Проверить гипотезу о том, что дантист выбивает произвольный зуб мудрости с вероятностью  $2/3$ , на уровне значимости 0.1.
8. Среди 5000 семей, имеющих трех детей, есть ровно 1010 семей с тремя мальчиками, 2200 семей с двумя мальчиками и одной девочкой, 950 семей с одним мальчиком и двумя девочками (во всех остальных семьях все дети — девочки). Можно ли с уровнем значимости  $\alpha = 0.02$  считать, что количество мальчиков  $\xi$  в семье с тремя детьми имеет следующее распределение

$$P(\xi = 0) = \theta, P(\xi = 1) = \theta,$$

$$P(\xi = 2) = 2\theta, P(\xi = 3) = 1 - 4\theta,$$

где  $\theta \in (0, 1/4)$ ?

## 13. Коэффициенты корреляции

Последний класс гипотез, который мы рассмотрим в настоящем учебном пособии, это гипотезы о независимости имеющих наблюдения. Пусть  $X = (X_1, \dots, X_n)$  и  $Y = (Y_1, \dots, Y_m)$  — две выборки. Нас будет интересовать вопрос, являются ли они независимыми. Иными словами, выдвигается гипотеза

$$H_0 : F_{X,Y}(s, t) = F_X(s)F_Y(t) \text{ для всех } s, t,$$

где  $F_X$  — функция распределения  $X_1$ ,  $F_Y$  — функция распределения  $Y_1$ , а  $F_{X,Y}$  — функция распределения вектора  $(X_1, Y_1)$ . Альтернативой к гипотезе  $H_0$  выступает гипотеза  $H_1$  о зависимости выборок  $X$  и  $Y$ , т.е.

$$H_1 : F_{X,Y}(s, t) \neq F_X(s)F_Y(t) \text{ для некоторых } s, t.$$

Но проверить свойство независимости в общем случае проблематично, потому что оно должно быть выполнено для всех пар  $s, t$ , гораздо проще убедиться в том, что выборки являются зависимыми. Для решения подобных задач используются выборочные коэффициенты корреляции, и общее правило принятия решения в нашей задаче выглядит так: если коэффициент корреляции достаточно далеко отстоит от нуля, то отвергаем гипотезу о независимости.

Корреляционный анализ (а именно, анализ зависимостей между наблюдениями) играет важную роль в задачах регрессии, а также в теории временных рядов, в частности, анализе стационарных последовательностей и авторегрессионных моделей. В данной главе мы рассмотрим три самых используемых выборочных коэффициента корреляции: коэффициенты корреляции Пирсона, Спирмэна и Кэндалла.

### 13.1. Коэффициент корреляции Пирсона

Пусть  $X_1, \dots, X_n$  и  $Y_1, \dots, Y_n$  — две выборки с  $EX_1^2 < \infty$ ,  $EY_1^2 < \infty$ .

**▲ Определение 13.1.** Коэффициентом корреляции Пирсона, или обычным коэффициентом корреляции, называется следую-

щая статистика:

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

**▲ Свойства коэффициента корреляции Пирсона.**

1. Имеет место сходимость

$$\hat{\rho} \xrightarrow{P} \rho(X_1, Y_1) = \frac{\text{cov}(X_1, Y_1)}{\sqrt{D X_1 D Y_1}}, \quad n \rightarrow \infty$$

т.е., грубо говоря, коэффициент корреляции Пирсона соответствует стандартному коэффициенту корреляции двух случайных величин.

2. Если гипотеза  $H_0$  о независимости выборок верна и выборки имеют нормальное распределение, то

$$T = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim T_{n-2}.$$

Таким образом, критерий уровня значимости  $\alpha$  для проверки гипотезы  $H_0$  для нормальных выборок выглядит так: если  $T \notin (z_{\alpha/2}, z_{1-\alpha/2})$ , где  $z_{\alpha/2}$  и  $z_{1-\alpha/2}$  — квантили уровней  $\alpha/2$  и  $1 - \alpha/2$  соответственно из распределения Стьюдента с  $n - 2$  степенями свободы, то отвергаем гипотезу  $H_0$ .

**▲ Задача 13.1.** Доказать свойство 1 коэффициента корреляции Пирсона.

## Решение

Решение состоит в последовательном применении усиленного закона больших чисел (для независимых одинаково распределенных случайных величин с конечным математическим ожиданием) и теоремы о наследовании сходимостей.

Итак, по усиленному закону больших чисел,

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbf{E}X_i)(Y_i - \mathbf{E}Y_i) \xrightarrow{P} \mathbf{E}(X_1 - \mathbf{E}X_1)(Y_1 - \mathbf{E}Y_1),$$

а также  $\bar{X} \xrightarrow{P} \mathbf{E}X_1$  и  $\bar{Y} \xrightarrow{P} \mathbf{E}Y_1$  (разумеется, верны даже сходимости почти наверное). Следовательно, справедлива и сходимость векторов  $(\bar{X}, \bar{Y}) \xrightarrow{P} (\mathbf{E}X_1, \mathbf{E}Y_1)$ . Поэтому по теореме о наследовании сходимостей,

$$\left| \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) - \frac{1}{n} \sum_{i=1}^n (X_i - \mathbf{E}X_i)(Y_i - \mathbf{E}Y_i) \right| =$$
$$|\bar{X}\mathbf{E}Y_1 - \mathbf{E}X_1\mathbf{E}Y_1 + \bar{Y}\mathbf{E}X_1 - \bar{X} \cdot \bar{Y}| \xrightarrow{P} 0.$$

Отсюда,

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \xrightarrow{P} \mathbf{E}(X_1 - \mathbf{E}X_1)(Y_1 - \mathbf{E}Y_1) = \text{cov}(X_1, Y_1).$$

Как мы доказывали в предыдущих главах, выборочная дисперсия  $s_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  является состоятельной оценкой  $\mathbf{D}X_1$ , аналогичное утверждение верно и для выборки  $Y_1, \dots, Y_n$ . Тогда по теореме о наследовании сходимостей,

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \xrightarrow{P} \frac{\text{cov}(X_1, Y_1)}{\sqrt{\mathbf{D}X_1 \mathbf{D}Y_1}}.$$

**Задача решена.**

Вследствие того, что для нормальных выборок некоррелированность эквивалентна независимости, коэффициент корреляции Пирсона наиболее подходит для работы с нормальными выборками. Но, к сожалению, как и в случае многих параметрических методов, использование этого коэффициента корреляции приводит к получению неадекватных значений статистики (что означает неверность принимаемых статистических решений), так как коэффициент корреляции не является устойчивым к выбросам (при наличии в выборке далеко отстоящих данных, связанных, например, с ошибками измерений). Широко известен пример (так называемый “квартет Энскомба”), демонстрирующий, насколько подобные методы обработки статистических данных способны “врать”, даже если выброс всего один на 10 “обычных” результатов.

### 13.2. Коэффициент корреляции Спирмена

Пусть имеется выборка  $X_1, \dots, X_n$  из некоторого непрерывного распределения (в этом случае элементы выборки не совпадают почти наверное). Упорядочим элементы выборки по возрастанию (т.е. построим вариационный ряд выборки).

**▲ Определение 13.2.** Номера, которые получили элементы выборки при таком упорядочивании, называются их *рангами*.

Будем обозначать ранги выборки  $X_i$  как  $R(X_i)$  (получается, что ранг  $R(X_i)$  – это номер наблюдения  $X_i$  в вариационном ряде выборки  $X_1, \dots, X_n$ ).

Основное свойство рангов следующее:

$$P(R(X_1) = r_1, \dots, R(X_n) = r_n) = \frac{1}{n!},$$

где  $(r_1, \dots, r_n)$  – произвольная перестановка чисел  $(1, \dots, n)$  (действительно, для любых  $i < j$  выполнено  $P(X_i > X_j) = P(X_j > X_i)$ , так как компоненты выборки независимы и одинаково распределены). Из основного свойства рангов также следует, что их распределение не зависит от первоначального, неизвестного нам распределения, из которого бралась выборка.

Рассмотрим теперь, как и ранее, две выборки  $X_1, \dots, X_n$  с функцией распределения  $F_X$  и  $Y_1, \dots, Y_n$  с функцией распределения  $F_Y$ . Обозначим ранг наблюдения  $X_i$  в выборке  $(X_1, \dots, X_n)$

как  $R_i$  и ранг наблюдения  $Y_j$  в выборке  $(Y_1, \dots, Y_n)$  как  $S_j$ , где эти ранги, как легко видеть, могут принимать значения от 1 до  $n$ .

**▲ Определение 13.3.** Коэффициентом корреляции Спирмэна называется следующая статистика:

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}.$$

**▲ Свойства коэффициента корреляции Спирмэна.**

1.  $\rho_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2$ .
2. При верной гипотезе  $H_0$   $E\rho_S = 0$  и  $D\rho_S = \frac{1}{n-1}$ .
3.  $-1 \leq \rho_S \leq 1$ , причём обе границы достигаются, т.е. именование  $\rho_S$  коэффициентом корреляции оправданно.
4. При верной гипотезе  $H_0$

$$\frac{\rho_S}{\sqrt{D\rho_S}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Этим нормальным приближением для построения критерия можно пользоваться при  $n \geq 50$ , при меньших значениях  $n$  рекомендуется использовать исправленную статистику

$$\tilde{\rho}_S = \frac{1}{2}\rho_S \left( \sqrt{n-1} + \sqrt{\frac{n-2}{1-\rho_S^2}} \right).$$

Тогда критерий уровня значимости  $\alpha$  будет выглядеть так: отвергать гипотезу независимости  $H_0$ , если  $\tilde{\rho}_S \notin (z_{\alpha/2}, z_{1-\alpha/2})$ , где  $z_\gamma = \frac{1}{2}(x_\gamma + y_\gamma)$ ,  $x_\gamma$  —  $\gamma$ -квантиль  $\mathcal{N}(0, 1)$ , а  $y_\gamma$  —  $\gamma$ -квантиль распределения Стьюдента с  $n-2$  степенями свободы.

**▲ Задача 13.2.** Доказать свойство 1 коэффициента корреляции Спирмэна.



## Решение

Прежде всего, упростим  $\rho_S$ . Во-первых,

$$\bar{R} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2} = \bar{S}.$$

Далее, поскольку  $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$ , то

$$\sum_{i=1}^n (R_i - \bar{R})^2 = \sum_{i=1}^n (S_i - \bar{S})^2 = \sum_{i=1}^n (i - (n+1)/2)^2 = \frac{n^3 - n}{12}.$$

Определим статистику  $T_i$  следующим образом: если  $R_k = i$ , то  $T_i = S_k$ . Ясно, что набор значений  $\{T_i\}_{i=1}^n$  есть перестановка множества  $\{1, \dots, n\}$ . Имеем,

$$\begin{aligned} \rho_S &= \frac{12}{n^3 - n} \sum_{i=1}^n \left( R_i - \frac{n+1}{2} \right) \left( S_i - \frac{n+1}{2} \right) = \\ &= \frac{6}{n^3 - n} \left( n \frac{(n+1)^2}{2} - 2 \frac{n+1}{2} \sum_{i=1}^n (i + T_i) + 2 \sum_{i=1}^n iT_i \right) = \\ &= \frac{6}{n^3 - n} \left( 2 \sum_{i=1}^n iT_i - n \frac{(n+1)^2}{2} \right) = \\ &= \frac{6}{n^3 - n} \left( - \sum_{i=1}^n (i - T_i)^2 + 2 \frac{n(n+1)(2n+1)}{6} - n \frac{(n+1)^2}{2} \right) = \\ &= \frac{6}{n^3 - n} \left( \frac{n^3 - n}{6} - \sum_{i=1}^n (i - T_i)^2 \right) = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2. \end{aligned}$$

**Задача решена.**

### 13.3. Коэффициент корреляции Кэндалла

Ещё одним примером устойчивого к выбросам коэффициента корреляции, построенного с помощью ранговых методов, является коэффициент корреляции Кэндалла.

▲ **Определение 13.4.** Пары случайных величин  $(X_i, Y_i)$  и  $(X_j, Y_j)$  называются *согласованными*, если

$$\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = 1.$$

Пусть  $S$  — число согласованных пар,  $R$  — число несогласованных. По-прежнему считаем, что внутри выборок нет одинаковых элементов. Определим

$$T = S - R = \sum_{i < j} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j).$$

Легко видеть, что  $T$  может меняться от  $-\frac{n(n-1)}{2}$  до  $\frac{n(n-1)}{2}$ , так как  $S + R =$  количество всех пар  $\{i, j\}, i \neq j = \frac{n(n-1)}{2}$ .

▲ **Определение 13.5.** Коэффициентом корреляции Кэндалла называется статистика  $\tau = \frac{2}{n(n-1)}T$ .

▲ **Свойства коэффициента корреляции Кэндалла.**

1. Верно следующее представление:

$$\tau = 1 - \frac{4}{n(n-1)}R.$$

2.  $\tau$  является коэффициентом корреляции, т.е.  $-1 \leq \tau \leq 1$  и границы достигаются, и при верной гипотезе  $H_0$   $E\tau = 0$ . Кроме того, при верности той же гипотезы

$$D\tau = \frac{2(2n+5)}{9n(n-1)}.$$

3. При верной гипотезе  $H_0$

$$\frac{\tau}{\sqrt{D\tau}} \xrightarrow{d} \mathcal{N}(0, 1).$$

4. При верной гипотезе  $H_0$  коэффициенты корреляции Спирмена и Кэндалла сильно коррелированы:  $\rho(\rho_S, \tau) > 0,99$  при  $n > 5$  (но коэффициент корреляции Спирмена более чувствителен к количеству несогласованных пар).

▲ **Задача 13.3.** Было проведено исследование на выявление взаимосвязи между физическим весом и IQ первоклассников города Усть-Илимск.

вес	24	27	26	21	20	31	26	22	20	18	25	22
IQ	87	74	117	85	145	84	95	89	82	97	79	215

Проверить гипотезу о независимости веса от IQ на уровне значимости 0.1.

### Решение

Посчитаем коэффициент корреляции Спирмэна для наших данных. Заметим, что среди значений веса школьников встречаются одинаковые, поэтому было бы естественным положить их ранги одинаковыми. Это достигается за счёт использования так называемых *средних рангов*: ранги, соответствующие одинаковым значениям в выборке, заменяются на среднее арифметическое рангов по этим значениям. Тем самым, ранги значений данных нам выборок таковы (первая строка соответствует весу, вторая – IQ):

7	11	9.5	4	2.5	12	9.5	5.5	2.5	1	8	5.5
6	1	10	5	11	4	8	7	3	9	2	12

Получаем  $\rho_S = -0.348$ . Значение коэффициента корреляции довольно далеко отстоит от 0, что может свидетельствовать о зависимости выборок. Тем не менее, используем статистику  $\tilde{\rho}_S$ , чтобы корректно проверить гипотезу  $H_0$ . Итак, пользуясь формулой из свойства 4 коэффициента корреляции Спирмэна, имеем  $\tilde{\rho}_S = -1.17$ , тогда как квантиль  $z_{0.05} = -z_{0.95} = -1.73$ . Тем самым, несмотря на довольно большое абсолютное значение коэффициента корреляции, гипотезу о независимости выборок отвергнуть нельзя.

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Темпы роста ВВП России за 2006 – 2012 годах в процентах (всего и на душу населения, по отношению к 2005 году): 108.2 117.4 123.5 113.9 119.0 124.1 128.4 108.5 118.0 124.2 114.5 119.6 124.6 128.7. Те же цифры для Украины: 107 116 118 101 105 111 112 109 117 121 103 108 114 115. Проверьте гипотезу о независимости данных выборок на уровне значимости 0.05 (выберите наиболее удобный и применимый в данном случае критерий).

2. В институте проведено исследование на выявление зависимости между успеваемостью студента и средним количеством пончиков, съеденным им за день.

Понч.	9	1	3	7	3	5	1	0	60	63	2	9	0	0
Балл	4	9	8	7	3	8	7	5	8	3	6	5	7	8

Проверьте гипотезу о независимости успеваемости и количества съеденных пончиков на уровне значимости 0.05 (выберите наиболее удобный и применимый в данном случае критерий).

3. Докажите, что в случае верности гипотезы о независимости выборок  $D\rho_S = \frac{1}{n-1}$ , где  $\rho_S$  — коэффициент корреляции Спирмэна.
4. Докажите, используя введенные в задаче 13.2 статистики  $\{T_i\}$ , что

$$\rho_S = 1 - \frac{12}{n^3 - n} \sum_{i < j} (j - i) I\{T_i > T_j\},$$

$$\tau = 1 - \frac{4}{n^2 - n} \sum_{i < j} I\{T_i > T_j\}.$$

5. Даны выборки  $X = (X_1, \dots, X_n)$ ,  $Y = (Y_1, \dots, Y_n)$ . Рассмотрим произвольную функцию  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ . Для любых  $1 \leq i < j \leq n$  положим  $c_{ij}(X) = f(X_i, X_j)$ . Обобщённым коэффициентом корреляции называется следующая статистика:

$$\hat{r} = \frac{\sum_{i < j} c_{ij}(X)c_{ij}(Y)}{\sqrt{\sum_{i < j} c_{ij}(X)^2 \sum_{i < j} c_{ij}(Y)^2}}.$$

Докажите, что это действительно выборочный коэффициент корреляции, т.е. что  $|\hat{r}| \leq 1$  и значения  $\pm 1$  достигаются, а также что при верной гипотезе о независимости выборок  $E\hat{r} = 0$ .

6. Найдите коэффициенты корреляции Пирсона, Спирмэна и Кэндалла с помощью обобщённого коэффициента корреляции.
7. Коэффициент корреляции знаков Фехнера определяется формулой

$$I = (C - H)/(C + H),$$

где  $C$  — число пар, у которых знаки отклонений значений от их средних совпадают,  $H$  — число пар, у которых знаки отклонений значений от их средних не совпадают. Найдите распределение  $I$  при верности гипотезы о независимости выборок, если выборки сделаны из стандартного нормального распределения.

8. Можно ли с уровнем значимости 0.05 считать, что последовательность чисел 1.05, 1.12, 1.37, 1.50, 1.51, 1.73, 1.85, 1.98, 2.03, 2.17 является реализацией случайного вектора, все 10 компонент которого независимые одинаково распределённые случайные величины? Вывод сделайте на основе коэффициентов корреляции (выберите наиболее удобный и применимый в данном случае критерий).

## 14. Байесовские оценки

Напоследок мы поговорим об еще одном способе сравнении оценок. О сравнении оценок мы уже говорили в главе 4 и использовали для этого равномерный подход с квадратичной функцией потерь. При таком подходе одна оценка считается лучше другой, если дисперсия первой не больше дисперсии второй для любого значения параметра. Разумеется, если мы обладаем некоторыми априорными знаниями о значении параметра, то мы можем ослабить “требование наилучшести”, сравнив только усредненные значения дисперсий.

Итак, пусть  $\theta$  — случайная величина с известным законом распределения  $Q$  на множестве  $\Theta$ . Пусть, кроме того,  $X$  — наблюдение с неизвестным распределением  $P_\theta$  (будем считать, что  $\theta$  и  $X$  заданы на одном и том же вероятностном пространстве с вероятностной мерой  $P$ ). Всюду далее в этой главе мы будем считать, что  $\{P_\theta, \theta \in \Theta\}$  — доминируемое семейство распределений, причем плотность распределения  $P_\theta$  равна  $p_\theta$ . Пусть, кроме того,  $q$  — плотность (возможно, дискретного) распределения  $Q$  (такая плотность называется *априорной*).

**▲ Определение 14.1.** Говорят, что оценка  $\theta^*$  *лучше оценок*  $\hat{\theta}$  в байесовском подходе с функцией потерь  $g$ , если

$$\int_{\Theta} E_t g(\theta^*, t) dt < \int_{\Theta} E_t g(\hat{\theta}, t) dt.$$

Если  $\theta^*$  лучше всех других оценок в байесовском подходе с функцией потерь  $g$  в некотором классе оценок  $\mathcal{K}$ , то  $\theta^*$  называют *наилучшей в классе  $\mathcal{K}$  в байесовском подходе с функцией потерь  $g$* .

Итак, поговорим о методе нахождения наилучших в байесовском подходе оценок. Оценка  $\theta^* = E(\theta|X)$  называется *байесовской оценкой параметра  $\theta$* .

**Теорема  
14.1**

*Байесовская оценка является наилучшей оценкой  $\theta$  в байесовском подходе с квадратичной функцией потерь.*

Заметим, что условная плотность  $p_{\theta|X}(t|x)$  находится по формуле

$$p_{\theta|X}(t|x) = \frac{q(t)p_t(x)}{\int_{\Theta} q(u)p_u(x)du}$$

и называется *апостериорной плотностью*  $\theta$ .

▲ **Задача 14.1.** Найдите байесовскую оценку параметра  $\theta$  по выборке  $X_1, \dots, X_n$  из равномерного распределения на интервале  $[\theta, \theta + 1)$ , если  $\theta \sim \text{Pois}(1)$ . Является ли полученная оценка состоятельной оценкой параметра  $\theta$ ?

### Решение

Положим  $X = (X_1, \dots, X_n)$ . Так как  $P(0 \leq X_{(n)} - X_{(1)} < 1) = 1$ , то имеем

$$p_{\theta|X}(t|x) = \frac{\frac{\lambda^t e^{-\lambda}}{t!} I(t \leq x_{(1)} \leq x_{(n)} < t+1)}{\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} I(u \leq x_{(1)} \leq x_{(n)} < u+1)} dt =$$

$$\frac{(\lambda^t/t!) I(\lfloor x_{(1)} \rfloor = t)}{\lambda^{\lfloor x_{(1)} \rfloor} / (\lfloor x_{(1)} \rfloor!)} = I(\lfloor x_{(1)} \rfloor = t).$$

Теперь найдем байесовскую оценку:

$$\theta^* = E(\theta|X) = \sum_{t=0}^{\infty} t I(\lfloor X_{(1)} \rfloor = t) = \lfloor X_{(1)} \rfloor.$$

Проверим полученную оценку на состоятельность. Разумеется,  $P(\lfloor X_{(1)} \rfloor = \theta) = 1$ . Поэтому байесовская оценка является сильно состоятельной.

**Задача решена.**

В статистических задачах часто предполагается, что *априорное распределение*  $Q$  и *апостериорное распределение*  $P(\theta \in B|X)$  принадлежат одному и тому же семейству распределений. В этом случае семейство распределений, которому принадлежит  $Q$ , называется *сопряженным семейству распределений*  $\{P_{\theta}, \theta \in \Theta\}$ , распределение  $Q$  называется *сопряженным априорным распределением к семейству*  $\{P_{\theta}, \theta \in \Theta\}$ . Разумеется,

семейство всех распределений является сопряженным любому семейству распределений. Тем не менее, оказывается, что при рассмотрении более узких семейств распределений (обычных семейств, рассматриваемых в курсах теории вероятностей и математической статистики — нормального, экспоненциального, Бернулли и т.д.) вычисление байесовской оценки упрощается.

Прежде чем разобрать пример вычисления байесовской оценки в случае, когда априорное распределение является сопряженным, приведем примеры некоторых семейств сопряженных распределений.

### ▲ Примеры семейств сопряженных распределений:

1. Семейство  $\text{Beta}(\lambda_1, \lambda_2)$  является сопряженным к  $\text{Bern}(p)$ .

Действительно, найдем апостериорную плотность  $p_{\theta|X}$ , если

$$p_t(x) = p_t(x_1, \dots, x_n) = p_X(x_1, \dots, x_n) = t^{\sum x_i} (1-t)^{n-\sum x_i},$$

$$q(t) = p_{\theta}(t) = \frac{t^{\lambda_1-1}(1-t)^{\lambda_2-1}}{B(\lambda_1, \lambda_2)} I(t \in [0, 1]).$$

Имеем

$$p_{\theta|X}(t|x) = \frac{\frac{t^{\lambda_1-1}(1-t)^{\lambda_2-1}}{B(\lambda_1, \lambda_2)} t^{\sum x_i} (1-t)^{n-\sum x_i}}{\int_0^1 \frac{u^{\lambda_1-1}(1-u)^{\lambda_2-1}}{B(\lambda_1, \lambda_2)} u^{\sum x_i} (1-u)^{n-\sum x_i} du} =$$

$$\frac{t^{\sum x_i + \lambda_1 - 1} (1-t)^{n - \sum x_i + \lambda_2 - 1}}{B(\sum x_i + \lambda_1, n - \sum x_i + \lambda_2)}.$$

Следовательно, апостериорное распределение —  $\text{Beta}(\sum X_i + \lambda_1, n - \sum X_i + \lambda_2)$ .

2. Семейство  $\Gamma(\alpha, \beta)$  является сопряженным к  $\exp(\lambda)$ .

Действительно, найдем апостериорную плотность  $p_{\theta|X}$ , если

$$p_t(x) = e^{-t \sum x_i},$$

$$q(t) = \frac{\alpha^{\beta} t^{\beta-1} e^{-\alpha t}}{\Gamma(\beta)} I(t \geq 0).$$



Имеем

$$p_{\theta|X}(t|x) = \frac{\frac{\alpha^\beta t^{\beta-1} e^{-\alpha t}}{\Gamma(\beta)} e^{-t \sum x_i}}{\int_0^\infty \frac{\alpha^\beta u^{\beta-1} e^{-\alpha u}}{\Gamma(\beta)} e^{-u \sum x_i} du} = \frac{(\alpha + \sum x_i)^\beta t^{\beta-1} e^{-(\alpha + \sum x_i)t}}{\Gamma(\beta)}.$$

Следовательно, апостериорное распределение —  $\Gamma(\alpha + \sum X_i, \beta)$ .

▲ **Задача 14.2.** Найдите байесовскую оценку параметра  $\theta$  по выборке  $X_1, \dots, X_n$  из экспоненциального распределения с параметром  $\theta$ , имеющим сопряженное априорное распределение.

### Решение

Апостериорное распределение параметра  $\theta$  —  $\Gamma(\alpha + \sum X_i, \beta)$ , где  $\Gamma(\alpha, \beta)$  — априорное распределение  $\theta$ . Тогда байесовская оценка параметра  $\theta$  равна  $E(\theta|X)$ , где

$$\begin{aligned} E(\theta|X = (x_1, \dots, x_n)) &= \int_0^\infty t \frac{(\alpha + \sum x_i)^\beta t^{\beta-1} e^{-(\alpha + \sum x_i)t}}{\Gamma(\beta)} dt = \\ &= \frac{1}{\alpha + \sum x_i} \int_0^\infty \frac{(\alpha + \sum x_i)^{\beta+1} t^\beta e^{-(\alpha + \sum x_i)t}}{\Gamma(\beta)} dt = \\ &= \frac{1}{\alpha + \sum x_i} \frac{\Gamma(\beta + 1)}{\Gamma(\beta)} = \frac{\beta}{\alpha + \sum x_i}. \end{aligned}$$

Окончательно получаем ответ  $\theta^* = \frac{\beta}{\alpha + \sum X_i}$

**Задача решена.**

### ▲ Задачи для самостоятельного решения

1. Проверьте оценку, полученную в задаче 14.2, на состоятельность.
2. По выборке  $X_1, \dots, X_n$  из пуассоновского распределения с параметром  $\theta$ , где  $\theta \sim \Gamma(\alpha, \lambda)$ , постройте наилучшую оценку в байесовском подходе с квадратичной функцией потерь.

3. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(\theta, 1)$ . Найдите байесовскую оценку параметра  $\theta$ , если априорное распределение  $\theta$  есть  $\text{Bin}(1, p)$ . Будет ли полученная оценка состоятельной оценкой параметра  $\theta$ ?
4. Пусть  $X_1, \dots, X_n$  — выборка из равномерного распределения на отрезке  $[0, \theta]$ . Найдите байесовскую оценку параметра  $\theta$ , если  $\theta$  имеет априорное распределение а) равномерное на отрезке  $[0, 1]$ , б) с плотностью  $q(t) = 1/t^2$  при  $t \geq 1$ . Проверьте полученные оценки на состоятельность.
5. Пусть  $X_1, \dots, X_n$  — выборка из нормального распределения с параметрами  $(\theta, 1)$ . Найдите байесовскую оценку параметра  $\theta$ , если априорное распределение  $\theta$  есть  $\mathcal{N}(b, \sigma^2)$ .
6. Пусть  $X_1, \dots, X_n$  — выборка из распределения а)  $\mathcal{N}(\theta, 1)$ , б)  $\mathcal{N}(0, \theta)$ , в)  $\text{Bin}(n, \theta)$ . Подберите сопряженное распределение и найдите байесовскую оценку.

## Литература

1. *Боровков А.А.* Математическая статистика. 4-е изд. – Спб.: Лань, 2010.
2. *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику. – М.: ЛКИ, 2009.
3. *Севастьянов Б.А.* Курс теории вероятностей и математической статистики. – М.: Наука, 1982.
4. *Ширяев А.Н.* Вероятность. 4-е издание. – М.: МЦНМО, 2007.

Учебное издание

## ВВЕДЕНИЕ В МАТЕМАТИЧЕСКУЮ СТАТИСТИКУ

Учебное пособие

Составители

**Жуковский** Максим Евгеньевич  
**Родионов** Игорь Владимирович  
**Шабанов** Дмитрий Александрович

Редактор *О.П. Котова*. Корректор *Л.В. Себова*  
Компьютерная верстка *Е.А. Казеннова*

Подписано в печать 24.03.2015. Формат 60×82 <sup>1</sup>/<sub>16</sub>.  
Усл. печ. л. 5. Уч.-изд.л. 3,0. Тираж 200 экз. Заказ №84.

---

Федеральное государственное автономное образовательное  
учреждение высшего профессионального образования «Московский  
физико-технический институт(государственный университет)»  
141700, Московская обл., г. Долгопрудный, Институтский пер., 9  
E-mail: rio@mail.mipt.ru

---

Отдел оперативной полиграфии «Физтех-полиграф»  
141700, Московская обл., г. Долгопрудный, Институтский пер., 9  
E-mail: polygraph@mipt.ru