

## Урок 2

# Доверительные интервалы

### 2.1. Интервальные оценки с помощью квантилей

В этой части речь пойдёт о построении интервальных оценок. Об этом говорилось в первом курсе специализации: разбирались некоторые частные случаи построения доверительных интервалов, в частности, использование правила двух сигм.

#### 2.1.1. Правило двух сигм

Необходимо вспомнить, как выглядит правило двух сигм. Если случайная величина имеет нормальное распределение с математическим ожиданием  $\mu$  и дисперсией  $\sigma^2$  ( $X \sim N(\mu, \sigma^2)$ ), то с вероятностью примерно 95 % она принимает значение из интервала  $\mu \pm 2\sigma$  (рисунок 2.1):

$$\mathbf{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95.$$

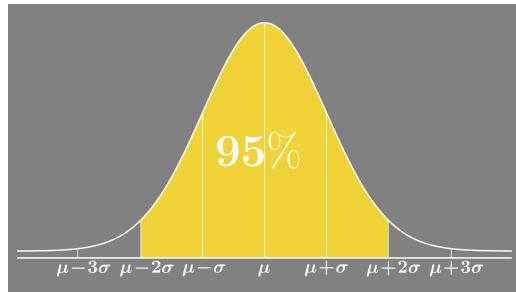


Рис. 2.1: Правило двух сигм.

При решении статистических задач правила двух сигм недостаточно: во-первых, эта оценка неточная, во-вторых, хочется строить такие оценки не только для вероятности 0.95, но и для любой другой.

#### 2.1.2. Уточнение правила двух сигм

Пусть задано число  $\alpha \in (0, 1)$ . Тогда квантилем порядка  $\alpha$  случайной величины  $X$  называется такая величина  $X_\alpha$ , что:

$$\mathbf{P}(X \leq X_\alpha) \geq \alpha, \quad \mathbf{P}(X \geq X_\alpha) \geq 1 - \alpha.$$

Существуют другие эквивалентные определения квантиля. В частности, если случайная величина  $X$  задана функцией распределения  $F(x)$ :

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

то

$$X_\alpha = F^{-1}(\alpha) = \inf\{x: F(x) \geq \alpha\},$$

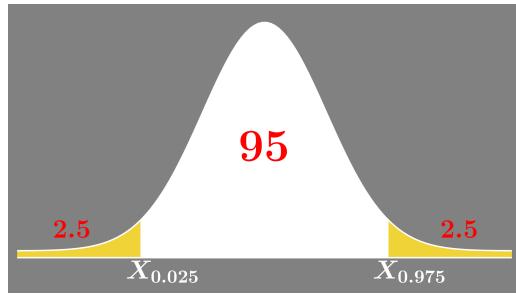


Рис. 2.2: Плотность вероятности нормально распределённой случайной величины.

то есть наименьшее  $x$ , для которого функция распределения  $F(x) \geq \alpha$ .

Определение квантиля можно использовать для уточнения правила двух сигм. Задача ставится следующим образом: требуется найти такие границы отрезка, что случайная величина  $X$  лежит внутри него с вероятностью ровно 95%.

На рисунке 2.2 показана плотность вероятности нормально распределённой случайной величины (плотность — это функция, интеграл от которой по всей числовой прямой равен 1, а по любому отрезку — вероятности попадания случайной величины в этот отрезок; интеграл — это площадь под кривой). У плотности можно выделить левый и правый "хвосты", так, чтобы их площади были равны 2.5%. Тогда площадь под центральной частью графика будет равна 95% (0.95). По определению, границы таких хвостов задаются квантилями  $X_{0.025}$  и  $X_{0.975}$ . Искомый интервал найден:

$$\mathbf{P}(X_{0.025} \leq X \leq X_{0.975}) = 0.95.$$

### 2.1.3. Предсказательный интервал

Такой интервал можно найти для произвольно распределённой случайной величины. Если случайная величина задаётся функцией распределения  $F(x)$ , то

$$\mathbf{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Отрезок  $[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}]$  называется предсказательным интервалом порядка  $1-\alpha$  для случайной величины  $X$ .

Если случайная величина  $X$  распределена нормально ( $X \sim N(\mu, \sigma^2)$ ), то её квантили можно выразить через параметры  $\mu$  и  $\sigma$ , а также квантили  $z_\alpha$  стандартного нормального распределения  $N(0, 1)$ :

$$\mathbf{P}\left(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma\right) = 1 - \alpha.$$

Нормальное распределение симметрично, поэтому  $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ .

При  $\alpha = 0.05$  квантиль стандартного нормального распределения  $z_{1-\frac{\alpha}{2}}$  равен

$$z_{0.975} \approx 1.95996 \approx 2.$$

Именно отсюда следует правило двух сигм.

## 2.2. Доверительные интервалы с помощью квантилей

В этой части будет рассказано о доверительных интервалах, о том, как их строить, и их отличиях от предсказательных интервалов.

### 2.2.1. Точечные оценки

Пусть имеется некоторая случайная величина  $X$ , функция распределения которой зависит от неизвестного параметра  $\theta$ :

$$X \sim F(x, \theta).$$

Чтобы высказать предположение о значении параметра  $\theta$ , можно собрать выборку

$$X^n = (X_1, \dots, X_n),$$

и по этой выборке подсчитать значение некоторой статистики  $\hat{\theta}$ . Если статистика подобрана хорошо, то она может служить оценкой для неизвестного параметра  $\theta$ . Например, если  $\theta$  — это математическое ожидание  $X$ , то выборочное среднее

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

будет хорошей оценкой этого параметра.

### 2.2.2. Доверительные интервалы

Помимо точечных, интерес представляют интервальные оценки, то есть доверительные интервалы. Доверительный интервал для параметра  $\theta$  задаётся парой статистик  $C_L, C_U$ :

$$\mathbf{P}(C_L \leq \theta \leq C_U) \geq 1 - \alpha,$$

где  $1 - \alpha$  — это уровень доверия интервала. Осталось понять, как  $C_L$  и  $C_U$  (нижние и верхние доверительные пределы) оценивать по выборке.

Если  $\hat{\theta}$  — оценка параметра  $\theta$  и известно её распределение  $F_{\hat{\theta}}(x)$ , то доверительные пределы можно выразить через квантили этого распределения:

$$\mathbf{P}\left(F_{\hat{\theta}}^{-1}\left(\frac{\alpha}{2}\right) \leq \theta \leq F_{\hat{\theta}}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Эти квантили задают доверительный интервал с уровнем доверия  $1 - \alpha$ .

#### Нормальное распределение

По выборке  $X^n = (X_1, \dots, X_n)$  можно построить доверительный интервал для математического ожидания нормально распределенной случайной величины  $X \sim N(\mu, \sigma^2)$ . Предположим, что дисперсия известна. Оценкой для параметра  $\mathbb{E}X = \mu$  является выборочное среднее  $\bar{X}_n$ . Выборка взята из нормального распределения, оно замкнуто относительно суммирования, значит, выборочное среднее — это нормально распределённая случайная величина:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Таким образом, для выборочного среднего известно распределение, а, значит, можно построить предсказательный интервал:

$$\mathbf{P}\left(\mu - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

В таком интервале выборочное среднее лежит с вероятностью  $1 - \alpha$ .

Осталось перегруппировать  $\mu$  и  $\bar{X}_n$  в неравенствах, которые стоят под знаком вероятности. Получается доверительный интервал для  $\mu$ :

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

#### Отличия предсказательного и доверительного интервалов

Стоит отметить важные различия между предсказательным и доверительным интервалами. У предсказательного интервала границы не случайны, случайно то, что стоит между этих границ (в рассмотренном выше примере — выборочное среднее). В доверительном интервале все ровно наоборот: то, что стоит в середине — это не случайный параметр. Параметр  $\mu$  — это неизвестная фиксированная константа, а случайными являются границы интервала.

Для нормально распределенной случайной величины  $X \sim N(\mu, \sigma^2)$  предсказательный интервал имеет вид

$$\mathbf{P}\left(\mu - z_{1-\frac{\alpha}{2}} \sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}} \sigma\right) = 1 - \alpha.$$

Если требуется оценить этот предсказательный интервал по выборке, то нужно избавиться от  $\mu$  в его границах, потому что значение  $\mu$  неизвестно. Единственное (и лучшее), что можно сделать, — это заменить  $\mu$  на выборочное среднее:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \sigma \leq X \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \sigma\right) \approx 1 - \alpha$$

В свою очередь, доверительный интервал для  $\mu$ , который можно построить по той же самой выборке, имеет вид:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Доверительный интервал получился в  $\sqrt{n}$  раз уже предсказательного интервала. Это неудивительно, поскольку предсказательный интервал оценивает диапазон, в котором меняется случайная величина, а доверительный интервал для среднего показывает, в каком диапазоне, скорее всего, лежит среднее этой случайной величины.

## Другие распределения

Вообще говоря, этой техникой можно пользоваться для построения доверительных интервалов математического ожидания не только нормально распределенных случайных величин, но и практически любых других. Пусть  $X \sim F(x)$ ,  $\bar{X}_n$  — оценка  $\mathbb{E}X$  по выборке  $X^n = (X_1, \dots, X_n)$ .

Используем центральную предельную теорему. В ней утверждается, что распределение выборочного среднего по достаточно большой выборке (если распределение исходной случайной величины не слишком скошено) может быть аппроксимировано нормальным:

$$\bar{X}_n \approx N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$$

Таким образом, доверительный интервал для математического ожидания исходной случайной величины имеет вид:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\mathbb{D}X}{n}} \leq \mathbb{E}X \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\mathbb{D}X}{n}}\right) \approx 1 - \alpha.$$

## 2.3. Распределения, производные от нормального

### 2.3.1. Нормальное распределение

Прежде чем говорить о распределениях, производных от нормального, полезно вспомнить, что из себя представляет нормальное распределение. Оно задаётся двумя параметрами:

$$X \sim N(\mu, \sigma^2).$$

Параметр  $\mu$  — это математическое ожидание,  $\sigma^2$  — дисперсия. Плотность вероятности этой случайной величины:

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

а функция распределения:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Стоит отметить, что функция распределения не выражается аналитически, а график плотности распределения похож на «шляпу» (рисунок 2.3).

### 2.3.2. Распределение $\chi^2$

Пусть есть  $k$  независимых одинаково распределенных нормальных случайных величин:

$$X_1, X_2, \dots, X_k \sim N(0, 1).$$

Определим новую случайную величину  $X$ :

$$X = \sum_{i=1}^k X_i^2 \sim \chi_k^2.$$

Распределение такой случайной величины называется распределением хи-квадрат с  $k$  степенями свободы.

При  $k = 1, 2$  плотность распределения  $\chi^2$  — монотонно убывающая функция, максимум которой находится в точке  $x = 0$  (рисунок 2.4). При  $k > 3$  плотность перестаёт монотонно убывать, и с ростом  $k$  её максимум постепенно смещается вправо по числовой оси.

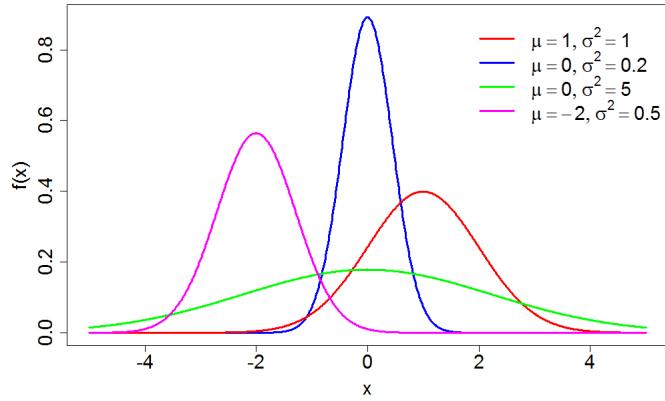


Рис. 2.3: Плотность вероятности нормального распределения с различными параметрами

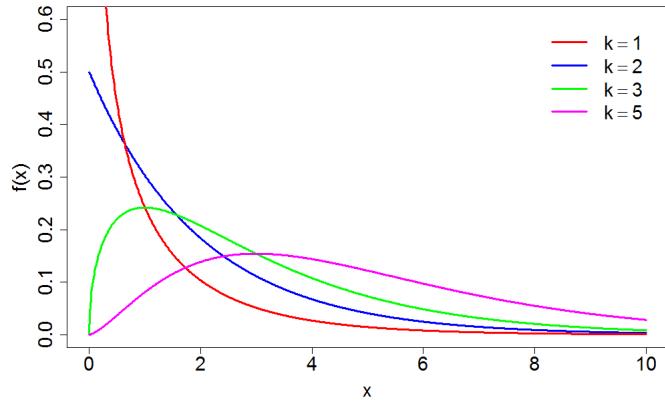


Рис. 2.4: Плотности распределений  $\chi_k^2$  с различными  $k$

### 2.3.3. Распределение Стьюдента

Пусть теперь имеются две независимые случайные величины:

$$X_1 \sim N(0, 1), \quad X_2 \sim \chi_\nu^2.$$

Новая случайная величина

$$X = \frac{X_1}{\sqrt{X_2/\nu}} \sim St(\nu),$$

будет иметь распределение Стьюдента с числом степеней свободы  $\nu$ .

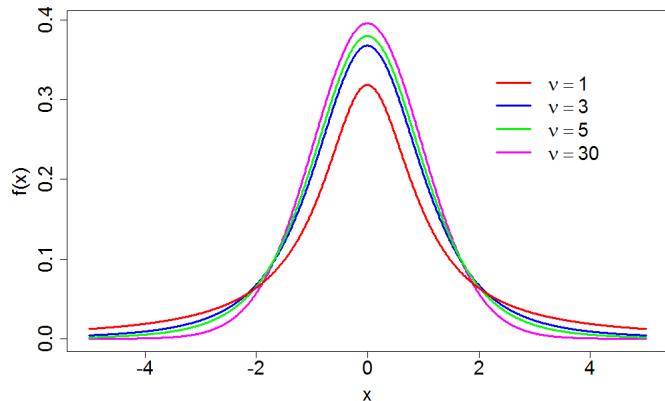


Рис. 2.5: Плотность вероятности распределения Стьюдента

На рисунке 2.5 изображены плотности вероятности распределения Стьюдента при разных значениях параметра  $\nu$ . На первый взгляд они кажутся похожими на плотности нормального распределения, однако у этих распределений есть несколько отличий. Во-первых, распределение всегда центрировано в точке  $x = 0$ , и не может сдвигаться по числовой оси. Кроме того, у распределения Стьюдента более тяжелые хвосты, то есть для такой случайной величины большие по модулю значения более вероятны, чем в нормальном распределении. Однако чем больше значение параметра  $\nu$ , тем меньше распределение Стьюдента отличается от нормального. При  $\nu > 30$  становится практически невозможно визуально различить эти распределения.

### 2.3.4. Распределение Фишера

Пусть теперь определены две независимые случайные величины  $X_1$  и  $X_2$ , принадлежащие распределению  $\chi^2$ :

$$X_1 \sim \chi_{d_1}^2, \quad X_2 \sim \chi_{d_2}^2.$$

Распределение случайной величины

$$X = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$$

называется распределением Фишера с числом степеней свободы  $d_1$  и  $d_2$ . Графики плотностей распределения Фишера выглядят очень по-разному в зависимости от значений параметров  $d_1$  и  $d_2$  (рисунок 2.6).

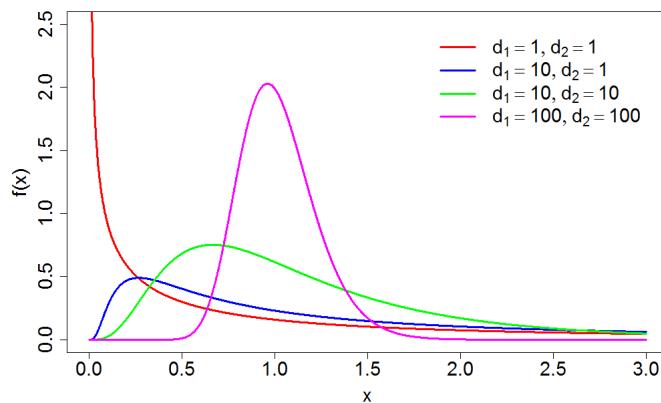


Рис. 2.6: Плотность вероятности распределения Фишера

### 2.3.5. Пример случайных величин из описанных распределений

Чтобы разобраться, зачем нужны описанные выше распределения, рассмотрим случаи, когда они встречаются на практике.

Пусть задана выборка из нормального распределения:

$$X \sim N(\mu, \sigma^2), \quad X^n = (X_1, \dots, X_n).$$

Мы знаем, что выборочное среднее для такой выборки также имеет нормальное распределение:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Что же касается выборочной дисперсии

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

то из формулы видно, что это сумма квадратов независимых одинаково распределенных нормальных случайных величин. Можно показать, что специальным образом нормированная выборочная дисперсия имеет распределение  $\chi^2$  с числом степеней свободы  $n-1$ :

$$(n-1) \frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

В свою очередь, так называемая  $T$ -статистика, активно применяющаяся в проверке гипотез и задаваемая выражением

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim St(n-1)$$

имеет распределение Стьюдента с числом степеней свободы  $n-1$ .

Наконец, пусть заданы две выборки разного размера из нормального распределения с разными параметрами:

$$\begin{aligned} X_1 &\sim N(\mu_1, \sigma_1^2), \quad X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), \\ X_2 &\sim N(\mu_2, \sigma_2^2), \quad X_2^{n_2} = (X_{21}, \dots, X_{2n_2}). \end{aligned}$$

Нормированное отношение выборочных дисперсий этих выборок имеет распределение Фишера с числом степеней свободы  $n_1-1, n_2-1$ :

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

## 2.4. Построение доверительных интервалов для среднего

Часто недостаточно построить точечную оценку среднего по выборке (выборочное среднее), и хочется понять, в каком диапазоне может меняться среднее. Именно в таких случаях используют доверительные интервалы для среднего. Далее будут рассмотрены два способа построения доверительных интервалов: с помощью z-интервала и t-интервала.

### 2.4.1. z-интервал

Для построения z-интервала необходимо знать дисперсию выборки или выдвинуть какое-то предположение о её значении:

$$\bar{X}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Случай, когда известна дисперсия, очень редки, на практике её значение практически никогда неизвестно. Пример случая, когда можно использовать z-интервал, — оценка работы некоторого прибора, в таких случаях обычно известна погрешность, а значит, и дисперсия.

### 2.4.2. t-интервал

В случаях, когда дисперсия неизвестна, лучше не делать ничем не подкреплённых предположений о её значении, а использовать t-интервал. Вместо гипотетической дисперсии в этом методе используется выборочная дисперсия  $S^2$ :

$$\bar{X}_n \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}.$$

## 2.5. Построение доверительных интервалов для доли

В этой части будут описаны методы построения доверительных интервалов для доли. Работа в таких случаях ведётся с генеральной совокупностью, состоящей из бинарных событий. Это такие события, каждое из которых можно описать 0 или 1, или по-другому, связать с успехом или с неудачей. В жизни довольно много примеров таких событий: проигрыш или выигрыш в лотерею, покупка или не покупка товара, клик или не клик на рекомендацию.

Доверительный интервал для доли можно строить на основе нормального распределения с использованием центральной предельной теоремы. Формула для такого интервала:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Следующий метод, который очень часто используют, — это доверительный интервал Уилсона. Этот метод является некоторым улучшением предыдущего метода, которое позволяет получать качественные оценки в крайних случаях (то есть когда доля близка к 0 или 1). Формула для расчета:

$$\frac{1}{1 + \frac{z^2}{n}} \left( \hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}} \right), \quad z \equiv z_{1 - \frac{\alpha}{2}}.$$

## 2.6. Построение доверительных интервалов для двух долей

Пусть существует некоторая услуга, которую необходимо рекламировать, и для этих целей используется рекламный баннер. Если появляется новый баннер, который кажется более красивым, то возникает необходимость проверить, какой же из двух баннеров лучше. Для этого можно поступить следующим образом: создать веб-форму, загрузить туда два баннера и попросить некоторое количество людей (например, 1000) посмотреть на эти баннеры и нажать на кнопку «лайк», если баннер им понравился. Таким образом, нужно будет сравнить доли «лайков» каждого из баннеров. В случаях, например, когда обе доли имеют близкое к нулю значение, имеет смысл построить доверительные интервалы.

Если просто построить два доверительных интервала, то какие-то выводы из этой информации можно сделать только если они не пересекаются.

	$X_1$	$X_2$
1	$a$	$b$
0	$c$	$d$
$\sum$	$n_1$	$n_2$

Таблица 2.1: Таблица для построения доверительного интервала для разности долей

Для того, чтобы сравнивать пересекающиеся интервалы, можно построить доверительный интервал для двух долей. Если выборки независимы (например, каждый баннер смотрели разные люди), нужно построить таблицу, в которой суммируется информация о «лайках» для каждого баннера (2.1). На основании этой таблицы вычисляются статистики  $\hat{p}_1$  и  $\hat{p}_2$ :

$$\hat{p}_1 = \frac{a}{n_1}, \quad \hat{p}_2 = \frac{b}{n_2}.$$

Доверительный интервал для разности долей  $p_1 - p_2$  оценивается по следующей формуле:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1 - \frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

$X_1 \backslash X_2$	1	0	$\sum$
1	$e$	$f$	$e + f$
0	$g$	$h$	$g + h$
$\sum$	$e + g$	$f + h$	$n$

Таблица 2.2: Таблица сопряжённости

Если выборки связанные (например, два баннера оценивали одни и те же люди), то используется другая оценка разности долей. Для этого нужно построить таблицу сопряжённости (2.2) и вычислить следующие статистики:

$$\hat{p}_1 = \frac{e + f}{n}, \quad \hat{p}_2 = \frac{e + g}{n}, \quad \hat{p}_1 - \hat{p}_2 = \frac{f - g}{n}.$$

Доверительный интервал для разности долей в двух связанных выборках вычисляется по следующей формуле:

$$\frac{f - g}{n} \pm z_{1 - \frac{\alpha}{2}} \sqrt{\frac{f + g}{n^2} - \frac{(f - g)^2}{n^3}}.$$

## 2.7. Построение доверительных интервалов на основе бутстрепа

Часто возникает необходимость построить интервальную оценку для некоторой не очень удобной статистики, про распределение которой ничего не известно. Это могут быть квантили (например, медиана) или сочетание известных статистик (например, отношение долей).

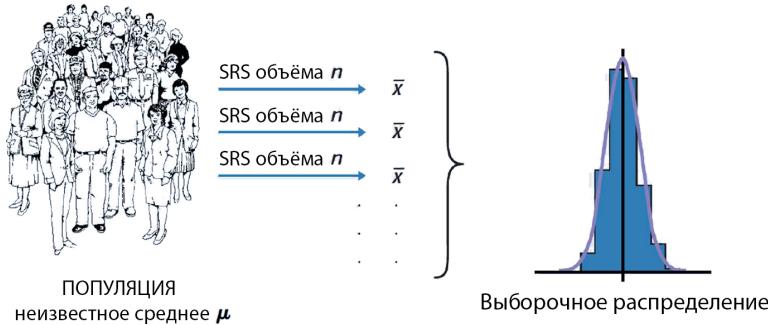


Рис. 2.7: Наивный метод построения выборочного распределения статистики

Чтобы построить доверительный интервал для статистики  $T_n = T(X^n)$ , необходимо знать её выборочное распределение  $F_{T_n(x)}$ . Нужно придумать, как это распределение получить. Первым приходит в голову наивный метод (рисунок 2.7): из генеральной совокупности извлечь  $N$  выборок размера  $n$  и оценить выборочное распределение  $T_n$  эмпирически. Однако этот метод применим скорее в теории, чем на практике: если не предстает сложности неограниченно генерировать выборки из генеральной совокупности, то можно и саму статистику вычислить на генеральной совокупности, а значит, интервальная оценка не нужна, поскольку известно настоящее значения статистики.

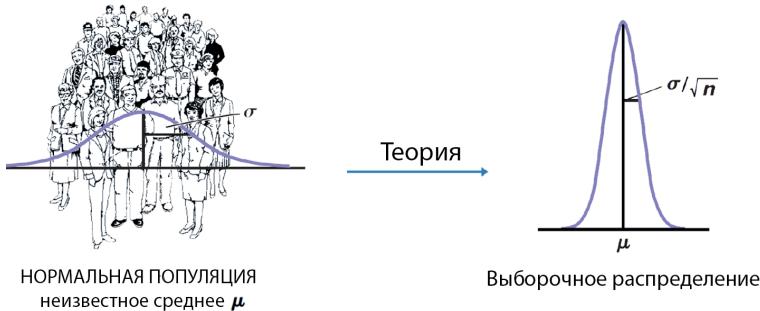


Рис. 2.8: Параметрический подход к построению выборочного распределения статистики

Другой подход — параметрический (рисунок 2.8). Предполагается, что известно распределение  $F_X(x)$  случайной величины  $X$ , и из него можно получить распределение статистики  $T_n$ , а затем параметры этого распределения оцениваются по выборке. Это тоже не самый лучший способ, поскольку непонятно, из каких соображений выбирать семейство распределений: про данные ничего не известно, всё, что доступно, — это выборка.

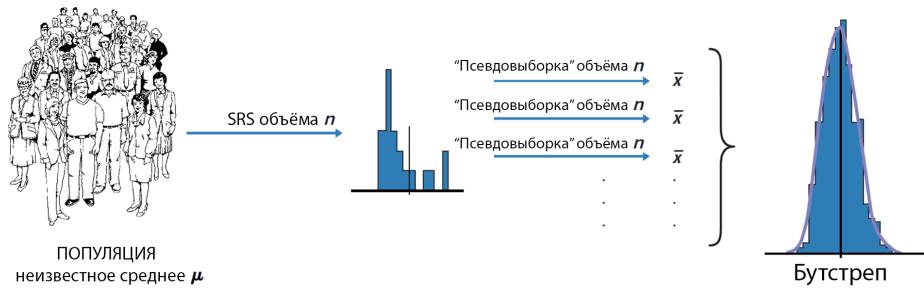


Рис. 2.9: Использование бутстрепа для построения выборочного распределения статистики

Вышеизложенные подходы подводят к идеи бутстрепа. Извлечение выборок из генеральной совокупности — это сэмплирование из неизвестного распределения  $F_X(x)$ . Лучшая оценка этого распределения, которая имеется в распоряжении, — это  $F_{X^n}(x)$ . Можно сэмплировать из этого распределения: из  $X^n$  извлекать с возвращением выборки объёма  $n$  (рисунок 2.9). Далее на каждой из этих выборок можно вычислить нужную статистику, и таким образом оценить эмпирическую функцию распределения. В этом и заключается идея бутстрепа.

# Урок 3

## Проверка гипотез

### 3.1. Проверка гипотез: начало

Проверка статистических гипотез — это важнейший инструмент, которым необходимо владеть в совершенстве, чтобы заниматься анализом данных. В этом уроке будут разобраны все компоненты, из которых состоит этот инструмент.

#### 3.1.1. Предсказание будущего

Можно представить себе человека, который утверждает, что он может предсказывать будущее. Не так важно, как он это делает: он может использовать гадание на кофейной гуще или делать свои предсказания на основании исторической информации, применяя обучение с учителем с хорошо измеренными признаками. Чтобы проверить его утверждение о способности предсказывать будущее, нужно провести эксперимент: записать все предсказания, сгенерировать соответствующие им события или подождать, сбудутся они или нет, а затем проверить правильность предсказаний.

Этот эксперимент порождает выборку  $X^n = (X_1, \dots, X_n)$ , которая может состоять, например, из 0 и 1:  $X = 1$  соответствует сбывшемуся предсказанию, а  $X = 0$  — несбывшемуся. Также она может состоять из точностей предсказания, то есть разностей между фактом и прогнозом.

Предсказатель полезен, если он предсказывает лучше, чем генератор случайных чисел. Можно рассмотреть гипотезу, что предсказатель — это и есть генератор случайных чисел. Для этого нужно посмотреть на данные и подумать, свидетельствуют ли они против такого предположения. Примерно так и используется проверка гипотез.

#### 3.1.2. Проверка гипотез: формальное определение

Теперь введём все необходимые компоненты механизма проверки гипотез формально (таблица 3.1).

выборка:	$X^n = (X_1, \dots, X_n)$ , $X \sim \mathbf{P}$ ;
нулевая гипотеза:	$H_0: \mathbf{P} \in \omega$ ;
альтернатива:	$H_1: \mathbf{P} \notin \omega$ ;
статистика:	$T(X^n)$ , $T(X^n) \sim F(x)$ при $H_0$ ; $T(X^n) \not\sim F(x)$ при $H_1$ .

Таблица 3.1: Проверка гипотез

Итак, имеется некоторая выборка из случайной величины  $X$ , которая имеет неизвестное распределение  $\mathbf{P}$ . Кроме того, выдвинута нулевая гипотеза об этом распределении (например, "Р принадлежит некоторому семейству распределений  $\omega$ ") и альтернативная гипотеза.

Требуется проверить, глядя на имеющиеся данные, какая из двух гипотез, нулевая или альтернативная, более вероятна. Для этого используется некоторая статистика  $T$ , которая обладает очень важным свойством: если нулевая гипотеза справедлива, то точно известно, какое у статистики распределение, а если справедлива

альтернатива, то распределение статистики — какое-то другое. Распределение  $F(x)$  называется нулевым распределением статистики, а пара, состоящая из статистики и нулевого распределения, образует статистический критерий для проверки нулевой гипотезы против альтернативы.

### 3.1.3. Нулевое распределение

Итак, пусть выборка собрана и подсчитано значение статистики на этой выборке:

$$T(X) = t.$$

Осталось понять, какова вероятность получить именно такое значение статистики при условии справедливости нулевой гипотезы. Вообще говоря, если распределение нулевой статистики непрерывно, то каждому конкретному значению соответствует нулевая вероятность, поэтому такая постановка задачи некорректна. Чтобы её переформулировать, нужно понять, какие значения статистики соответствуют альтернативной гипотезе.

Пусть, например, при справедливости альтернативы более вероятны большие значения статистики. Теперь возникает вопрос, с какой вероятностью можно получить значение статистики  $T(X) \geq t$  при справедливости нулевой гипотезы. Эта вероятность является ключевым компонентом механизма проверки гипотез и называется достижимым уровнем значимости, или p-value.

Достигаемый уровень значимости — это вероятность получить такое же значение статистики, как в эксперименте, или еще более экстремальное, при справедливости нулевой гипотезы. То, какие значения считаются экстремальными, определяется относительно альтернативной гипотезы, то есть, с учетом того, какие значения статистики более вероятны при альтернативе.

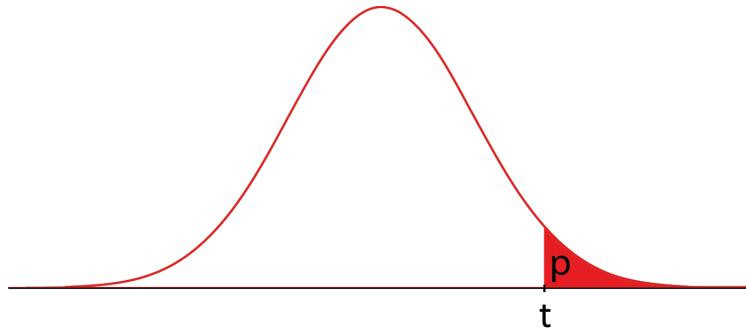


Рис. 3.1: Нулевое распределение

Зная нулевое распределение статистики и значение статистики, реализованное в эксперименте, можно вычислить p-value. В случае, когда критическими, то есть, более вероятными при альтернативе, являются большие значения статистики, p-value — это интеграл от плотности нулевого распределения по правому хвосту начиная с  $t$  и до  $\infty$  (рисунок 3.1).

Если полученное значение p-value мало, это значит, что данные свидетельствуют против нулевой гипотезы в пользу альтернативной, поскольку вероятность получить такие данные при условии, что нулевая гипотеза справедлива, мала. Обычно p-value сравнивают с порогом  $\alpha$ , который называется уровнем значимости. Чаще всего  $\alpha = 0.05$ . Если  $p \leq \alpha$ , то нулевая гипотеза отвергается в пользу альтернативы. Если  $p > \alpha$ , то нулевая гипотеза не отвергается.

## 3.2. Ошибки I и II рода

Существенная особенность механизма проверки гипотез — его несимметричность относительно пары нулевая гипотеза — альтернатива. Эта особенность тесно связана с понятиями ошибок первого и второго рода.

Нулевая гипотеза может быть либо верна, либо неверна. В результате проверки гипотезы её можно либо принять, либо отвергнуть. Из этих соображений составлена таблица 3.2. На главной диагонали находятся верные решения: либо принимается верная нулевая гипотеза, либо отвергается неверная нулевая гипотеза. А вот на побочной диагонали располагаются ошибки. Совершить ошибку первого рода — значит отвергнуть верную нулевую гипотезу. Если же принимается неверная нулевая гипотеза, то это ошибка второго рода.

	$H_0$ верна	$H_0$ неверна
$H_0$ принимается	$H_0$ верно принята	Ошибка II рода
$H_0$ отвергается	Ошибка I рода	$H_0$ верно отвергнута

Таблица 3.2: Ошибки I и II рода

В механизме проверки гипотез ошибки первого и второго рода неравнозначны. Ошибка первого рода критичнее, вероятность отвержение нулевой гипотезы в случае, когда она верна, жестко ограничивается. Если нулевая гипотеза отвергается при значении уровня значимости  $p \leq \alpha$ , то вероятность ошибки первого рода получается ограниченной сверху:

$$\mathbf{P}(H_0 \text{ отвергнута} | H_0 \text{ верна}) = \mathbf{P}(p \leq \alpha | H_0) \leq \alpha.$$

Таким образом, любой корректный, хорошо построенный критерий имеет вероятность ошибки первого рода не больше, чем  $\alpha$ .

Что касается ошибки второго рода, то она минимизируется по остаточному принципу. Понятие ошибки второго рода связано с понятием мощности статистического критерия. Мощность — это вероятность отвергнуть неверную нулевую гипотезу:

$$\text{pow} = \mathbf{P}(\text{отвергаем } H_0 | H_1) = 1 - \mathbf{P}(\text{принимаем } H_0 | H_1).$$

Чтобы найти идеальный критерий для проверки пары нулевая гипотеза – альтернатива, нужно среди всех корректных критериев выбрать критерий с максимальной мощностью.

Неравнозначность нулевой и альтернативной гипотезы видна уже на уровне терминологии. Если достигаемый уровень значимости  $p \leq \alpha$ , то говорят, что нулевая гипотеза отвергается в пользу альтернативы. Если достигаемый уровень значимости  $p > \alpha$ , то нулевая гипотеза не отвергается. Когда гипотеза не отвергается, это значит только то, что нет доказательств того что она неверна. Но отсутствие доказательств не является доказательством ее верности!

Это можно лучше понять на примере судебного процесса. Основное положение — презумпции невиновности: подсудимый по умолчанию невиновен (это нулевая гипотеза), и, если доказательств обратному нет, нельзя утверждать, что он преступник, даже если он на самом деле совершил преступление.

### 3.3. Достигаемый уровень значимости

Достигаемый уровень значимости — это достаточно сложная теоретическая концепция, которую часто понимают неправильно даже те, кто регулярно пользуется статистикой и проверкой гипотез.

Достигаемый уровень значимости — это вероятность при справедливости нулевой гипотезы получить такое же значение статистики, как в эксперименте, или еще более экстремальное:

$$p = \mathbf{P}(T \geq t | H_0)$$

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Проблема определения p-value в том, что оно длинное, но из него ничего нельзя выбросить так, чтобы оно не стало неправильным. Например, часто хочется думать, что p-value — это просто вероятность справедливости нулевой гипотезы, или вероятность справедливости нулевой гипотезы при условии полученных данных, но это не так!

$$\begin{aligned} p &= \mathbf{P}(T \geq t | H_0) \neq \mathbf{P}(H_0) \\ &\neq \mathbf{P}(H_0 | T \geq t) \end{aligned}$$

Это хорошо понятно на следующем примере. В 2010 году осьминог Поль угадывал результаты матчей чемпионата мира по футболу с участием сборной Германии, выбирая из двух кормушек ту, на которой был

изображён флаг страны-победителя. Из 13 матчей, в которых он пробовал свои силы, результаты 11 ему удалось угадать. Используя эти данные как выборку, можно проверить нулевую гипотезу о том, что он выбирает кормушку наугад против альтернативы о том, что осьминог есть сверхспособности к предсказанию результатов матчей. Критерий, которым проверяется эта нулевая гипотеза, будет разобран позже. Но если его применить, получится достигаемый уровень значимости  $p = 0.0112$ . Это значение — не вероятность, что осьминог выбирает кормушку наугад. Вероятность того, что осьминог выбирает кормушку наугад, равна единице!  $p = 0.0112$  — это именно вероятность получить такие или ещё более экстремальные данные при условии справедливости нулевой гипотезы. Эта вероятность достаточно мала, но редкие события тоже происходят. И, как правило, именно о них пишут в газетах.

## 3.4. Статистическая и практическая значимость

### 3.4.1. Размер эффекта

На самом деле эксперименты проводятся не для того, чтобы получить значение  $p$ -value. Как правило, исследователя интересует размер эффекта, то есть степень отклонения данных от нулевой гипотезы. Например, если эксперимент связан с проверкой способностей предсказателя будущего, то размер эффекта — это вероятность верного предсказания. Если проверяется эффективность лекарства, то размер эффекта — это вероятность выздоровления пациента, который это лекарство принимает, за вычетом эффекта плацебо. При запуске программы лояльности для пользователей интернет-магазина размер эффекта — это последующее увеличение среднего чека.

Размер эффекта — это величина, определенная на генеральной совокупности. Но, как правило, у исследователя есть только небольшая выборка из нее, а оценка размера эффекта по выборке — это случайная величина. Маленький достигаемый уровень значимости является показателем того, что такую оценку размера эффекта, какая получена по выборке, с маленькой вероятностью можно было получить случайно.

Достигаемый уровень значимости зависит не только от размера эффекта, но и от объема выборки, по которой оценивается эффект. Если выборка небольшая, скорее всего, нулевая гипотеза на ней не отвергается (если только она не слишком дикая). Однако с ростом объема выборки начинают проявляться все более тонкие отклонения данных от нулевой гипотезы. Велика вероятность, что на достаточно большой выборке значительная часть разумных нулевых гипотез будет отвергнута. Именно поэтому, даже если нулевая гипотеза отвергнута, это еще не значит, что полученный эффект имеет какую-то практическую значимость, её нужно оценивать отдельно. Чтобы лучше это понять, давайте рассмотрим несколько примеров.

### 3.4.2. Статистически значимо, практически незначимо

Первый пример связан с большим исследованием, в рамках которого на протяжении трех лет у большой выборки женщин измеряли вес, а также оценивали, насколько активно они занимаются спортом. По итогам исследования выяснилось, что женщины, которые в течение этого времени упражнялись не меньше часа в день, набрали значительно меньше веса, чем женщины, которые упражнялись менее 20 минут в день. Статистическая значимость этого результата достаточно высока:  $p < 0.001$ . Проблема в размере эффекта: разница в набранном весе между двумя исследуемыми группами женщин составила всего 150 граммов. 150 граммов за 3 года — это не очень много. Крайне сомнительно, что этот эффект имеет какую-то практическую значимость.

Еще один пример связан с клиническими испытаниями гормонального препарата «Премарин», который облегчает симптомы менопаузы. В 2002 году эти испытания были прерваны досрочно, поскольку было обнаружено, что прием препарата ведет к значимому увеличению риска развития рака груди (на 0.08%), инсульта (на 0.08%) и инфаркта (на 0.07%). Этот эффект статистически значим; при этом на первый взгляд кажется, что размеры эффектов ничтожны. Например, если кому-то сказать, что его любимые конфеты повышают риск возникновения инфаркта на 0.07%, вряд ли это заставит человека отказаться от этих конфет. Тем не менее, если пересчитать размеры эффектов на всю популяцию людей, которым этот препарат может быть потенциально приписан, результатом будут тысячи дополнительных смертей. Разработчики препарата не могут взять на себя эту ответственность, поэтому такой препарат немедленно запрещают и снимают с рынка.

Этот пример показывает, что практическую значимость результата нельзя определить на глаз. В идеале она должна определяться человеком, который поставил задачу и понимает предметную область.

### 3.4.3. Статистически незначимо, практически значимо

Еще один пример — это испытание лекарства, которое замедляет ослабление интеллекта у людей, страдающих болезнью Альцгеймера. В этом исследовании очень сложно измерить размер эффекта. В течение эксперимента одна часть испытуемых должна принимать лекарство, а другая — плацебо. Только по прошествии нескольких лет можно будет сравнить эти две группы. Поэтому такое исследование длится долгое время и дорогое стоит. Если при испытании оказывается, что разница между снижением IQ в контрольной группе, где люди принимали плацебо, и тестовой группе, где люди принимали препарат, составляет 13 пунктов, это различие очень большое, и на практике этот эффект крайне значим. При этом может оказаться, что статистическая значимость не была достигнута, то есть  $p > \alpha$ , и формально нулевую гипотезу об отсутствии эффекта лекарства нельзя отвергнуть. Если предмет исследования очень важен, то, оказавшись в подобной ситуации, возможно, стоит продолжать исследования: набрать еще выборку, уменьшить дисперсию оценки размера эффекта и убедиться в том, что важное открытие не упущено.

## 3.5. Биномиальный критерий для доли

Джеймс Бонд утверждает, что он предпочитает пить мартини взболтанным, но не смешанным. Чтобы проверить это на практике, можно предложить Джеймсу Бонду пройти так называемый blind test, или слепое тестирование. Можно было бы завязать ему глаза, несколько раз предложить на выбор взболтанный и смешанный мартини, а после этого спросить, какой напиток он предпочитает. В данном случае если бы Джеймс Бонд выбирал взболтанный напиток, это считалось бы успехом, потому что его выбор соответствует его утверждению. В противном случае считалось бы, что произошла неудача, так как выбор утверждению не соответствует.

В данном случае необходимо проверить нулевую гипотезу  $H_0$ : Джеймс Бонд не различает два вида напитков и выбирает наугад, против некоторой альтернативы. Но альтернатива, вообще говоря, могла бы быть разной. С одной стороны, можно рассматривать двустороннюю альтернативу (Джеймс Бонд отличает два вида напитков, и у него есть некоторые предпочтения) или одну из односторонних (Джеймс Бонд предпочитает взболтанный мартини, так, как он утверждает, или Джеймс Бонд предпочитает смешанный). Такой эксперимент нужно провести  $n$  раз и в качестве Т-статистики использовать количество единиц выборки или сумму элементов выборки. Если нулевая гипотеза справедлива, то есть Джеймс Бонд выбирает напиток наугад, то можно было бы равновероятно получить любую комбинацию из нулей и единиц. Таких комбинаций ровно  $2^n$ , поэтому для того, чтобы получить нулевое распределение, можно было бы сгенерировать все эти наборы данных, на каждом посчитать значение статистики и таким образом получить распределение. На самом деле, в данном случае этот шаг можно пропустить, потому что исследуемая выборка состоит из нулей и единиц и взята из распределения Бернулли с вероятностью успеха  $p$ . В данном случае вероятность успеха  $p = 0.5$ , потому что если нулевая гипотеза справедлива, то успех и неудачи происходят равновероятно. Соответственно выборка представляет из себя сумму  $n$  независимых одинаково распределенных величин из распределения Бернулли. Значит, нулевое распределение статистики — это биномиальное распределение с параметрами  $n$  (количество экспериментов) и  $p$  (вероятность успеха).

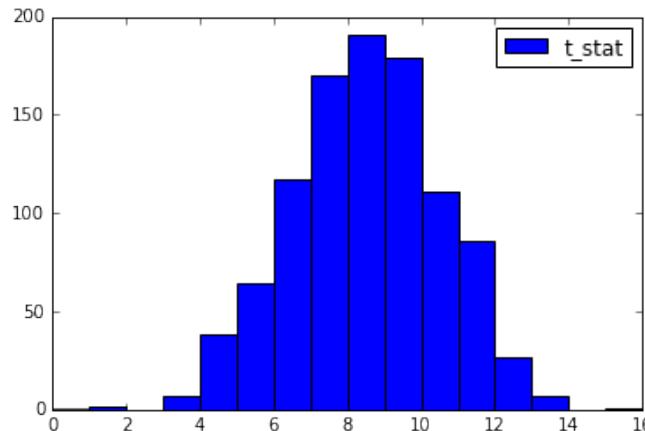


Рис. 3.2: Биномиальное распределение с параметрами  $n = 16$  и  $p = 0.5$

Нулевое распределение при параметрах  $n = 16$  и  $p = 0.5$  показано на рисунке 3.2. Оно выглядит так, как и ожидалось: пик находится в центре.

Итак, сначала можно протестировать нулевую гипотезу против односторонней альтернативы  $H_1$ : Джеймс Бонд предпочитает взболтанный мартини. При такой альтернативе более вероятно попасть в правый конец распределения, то есть получить много единиц в выборке. Пусть было проведено 16 испытаний, и при этом в 12 из них Джеймс Бонд выбрал взболтанный мартини, то есть произошел успех. Если построить соответствующее нулевое распределение, то в данном случае Т-статистика была бы равна 12 и интерес представлял бы правый «хвост» распределения. В данном случае требуется просуммировать высоту столбцов, начиная со столбца, соответствующего 12, и правее, то есть правый «хвост» распределения, полученное значение — это достигаемый уровень значимости. В данном случае получается  $p$ -value 0.038, это говорит о том, что на уровне значимости 0.05 нулевая гипотеза отвергается. То есть если успех происходит 12 раз из 16, то можно сделать вывод, что Джеймс Бонд предпочитает взболтанный мартини. Если бы успехов было немного меньше, например, 11, то значение  $p$ -value стало бы больше:  $p = 0.105$ , то есть на уровне значимости 0.05 уже нельзя отвергнуть нулевую гипотезу.

В случае двусторонней альтернативы гипотеза  $H_1$  переформулируется следующим образом: Джеймс Бонд предпочитает какой-то один определенный вид мартини, не требуется выбирать, какой именно. При такой альтернативе будут очень вероятны либо большие значения статистики, либо очень маленькие. При расчете достигаемого уровня значимости будут учитываться как правый, так и левый концы распределения. Если предположить, что произошло 12 успехов, то есть 12 раз Джеймс Бонд выбрал взболтанный мартини, то необходимо просуммировать тот же самый правый конец, но теперь к нему добавляется и левый. Значение достигаемого уровня значимости  $p = 0.077$ , это больше, чем при проверке нулевой гипотезы против односторонней альтернативы. Соответственно, в данном случае нельзя отвергнуть гипотезу на уровне значимости 0.05, однако можно отвергнуть нулевую гипотезу на уровне значимости 0.1. Можно посмотреть, достаточно ли 13 успехов, чтобы отвергнуть нулевую гипотезу на уровне 0.05. В данном случае  $p$ -value  $p = 0.021$ , отвергнуть нулевую гипотезу на уровне значимости 0.05 можно.

### 3.5.1. Критерий согласия Пирсона (хи-квадрат)

Критерий согласия Пирсона (или критерий хи-квадрат) используется для проверки того, что некоторая наблюдаемая случайная величина подчиняется тому или иному теоретическому закону распределения.

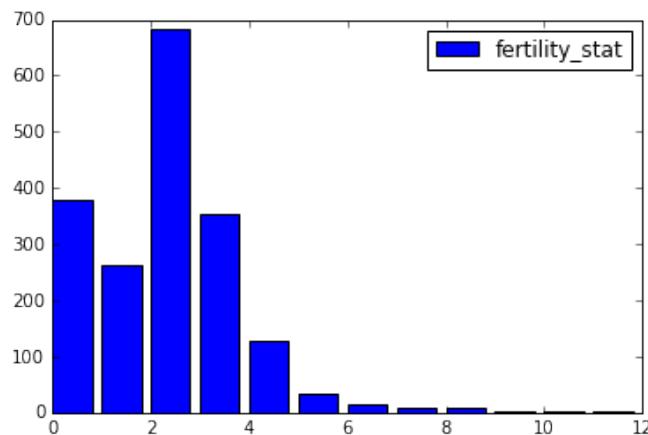


Рис. 3.3: Данные о количестве детей у женщин старше 45 лет

В качестве примера используются данные об исчерпанной рождаемости (рисунок 3.3). Этот признак связан с количеством детей, родившихся у женщины на момент окончания репродуктивного возраста (приблизительно 45 лет). Для 1878 женщин старше 45, участвующих в социологическом опросе жителей Швейцарии, известно количество детей. Этот признак — типичный счётчик, поэтому его можно попробовать оценить с помощью распределения Пуассона. В данном случае выборка — это целочисленный вектор длины  $n$  (в данном случае  $n = 1878$ ), где каждая компонента вектора — это количество детей, рожденных у женщины. В данном случае гипотеза  $H_0$ : наблюдаемая величина имеет распределение Пуассона.

Из распределения данных (рисунок 3.3) видно, что количество детей меняется от 0 до 11. Чаще всего у женщины не более четырёх детей. Наиболее часто встречающееся количество детей — это два ребёнка.

Кажется, что такие данные должны хорошо описываться распределением Пуассона. В предыдущих курсах было показано, что лучшая оценка на параметр  $\lambda$  для распределения Пуассона, — это просто выборочное среднее. Если его вычислить, получается  $\lambda = 1.937$ .

Необходимо проверить следующую гипотезу  $H_0$ : наблюдаемая случайная величина имеет распределение Пуассона с параметром  $\lambda = 2$ . Это можно делать с помощью критерия согласия Пирсона. Для этого нужно подготовить данные. Первое, что представляет интерес, — это наблюдаемые частоты. Известно, сколько раз встретилось каждое количество детей, так что интересующую величину несложно получить. Тогда элемент результирующего вектора 0 говорит о том, сколько раз в нашей выборке встретилось количество детей, равное 0 (в данном случае это 379), и последний 11 элемент означает, что 11 детей встретилось всего лишь 1 раз. Это наблюдаемые частоты.

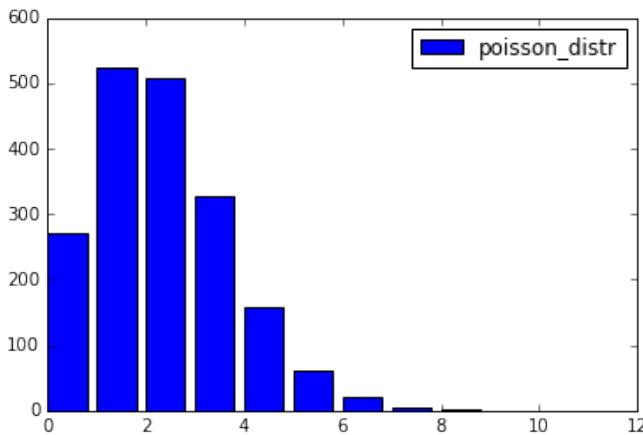


Рис. 3.4: Ожидаемые частоты для распределения Пуассона с параметром  $\lambda = 2$

Теперь нужно построить так называемые ожидаемые частоты. Это те частоты, которые бы наблюдались, если бы данные имели распределение Пуассона с параметром  $\lambda = 2$ , и размер выборки был бы таким же. Результирующие ожидаемые частоты показаны на рисунке 3.4. Видно, что наблюдаемые частоты отличаются от ожидаемых. Страго оценить это различие можно с помощью критерия хи-квадрат. Статистика этого критерия:

$$\chi^2 = \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i}.$$

При справедливости нулевой гипотезы статистика имеет распределение хи-квадрат с числом степеней свободы  $K - 1 - m$ , где  $m$  — число параметров распределения, оцененных по выборке.

Достигаемый уровень значимости, полученный с помощью критерия хи-квадрат:  $p = 1.77 \times 10^{-86}$ . Это значение очень близко к нулю, значит, можно смело отвергнуть нулевую гипотезу о том, что данные имеют распределение Пуассона с параметром  $\lambda = 2$ .

## 3.6. Связь между проверкой гипотез и доверительными интервалами

### 3.6.1. Проверка гипотез при помощи построения доверительных интервалов

Пусть требуется оценить качество предсказаний бинарного классификатора на тестовой выборке из 100 объектов. Этот классификатор верно предсказывает метку класса на 60 из 100 объектов. С одной стороны кажется, что 60 из 100 — это не очень много. С другой стороны, может быть, эта задача достаточно сложная, и предсказать лучше нельзя. Чтобы определить качество работы классификатора, его нужно сравнить с самым бесполезным классификатором — генератором случайных чисел. Если классы в задаче сбалансированы, то генератор случайных чисел в среднем будет угадывать метку у 50 объектов из 100, и вероятность угадать составляет 0.5. Можно ли считать, что классификатор, который угадывает классы 60 из 100 объектов, лучше, чем генератор случайных чисел?

Чтобы ответить на этот вопрос, можно построить доверительный интервал для доли верно предсказанных меток. Результат при уровне доверия 0.95: [0.504, 0.696]. Соответствующая генератору случайных чисел вероятность 0.5 не содержится в этом интервале. Из этого следует, что исследуемый классификатор значимо лучше, чем генератор случайных чисел.

В общем случае, если проверяется точечная нулевая гипотеза против двусторонней альтернативы:

$$H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0,$$

то производить эту проверку можно путём построения доверительного интервала, как было показано выше. Нулевая гипотеза отвергается на уровне значимости  $\alpha$ , если доверительный интервал для  $\theta$  с уровнем доверия  $1-\alpha$  не содержит  $\theta_0$ .

Также с помощью доверительного интервала можно вычислить достигаемый уровень значимости  $p$ . Это наибольшее значение  $\alpha$ , при котором доверительный интервал с уровнем доверия  $1-\alpha$  содержит  $\theta_0$ . Таким образом, перебирая разные значения  $\alpha$ , можно численно найти достигаемый уровень значимости.

Многие статистические критерии эквивалентны построению доверительных интервалов. Например, нормальные доверительные интервалы для доли эквивалентны z-критерию для той же самой доли (этот критерий будет описан позднее). Для таких пар нет необходимости в численном подсчёте достигаемого уровня значимости, это можно сделать аналитически.

Однако, например, для метода построения доверительных интервалов Уилсона нельзя явно записать выражение для статистики соответствующего критерия. Для подобных методов численный поиск достижимых уровней значимости оказывается очень полезным. Например, в задаче с бинарным классификатором 95% доверительный интервал Уилсона для доли верно предсказанных меток: [0.502, 0.691]. Полученный интервал похож на нормальный доверительный интервал, однако в других случаях, особенно когда значение  $p$  близко к 0 или 1, доверительный интервал Уилсона может существенно отличаться (как и достижимые уровни значимости), и лучше пользоваться именно им, поскольку он точнее. В данном случае достигаемый уровень значимости  $p = 0.045$ . То есть гипотеза о том, что рассматриваемый классификатор не лучше, чем генератор случайных чисел, может быть отвергнута на уровне значимости 0.05.

### 3.6.2. Построение доверительных интервалов с помощью критерииев проверки гипотез

Выше описан метод проверки гипотез с использованием доверительных интервалов. Однако можно делать и наоборот: строить доверительные интервалы с помощью критерия, проверяющего гипотезу.

Пусть снова заданы точечная гипотеза относительно параметра  $\theta$  и двусторонняя альтернатива:

$$H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0.$$

В таком случае доверительный интервал с уровнем доверия  $1 - \alpha$  будет состоять из всех значений  $\theta_0$ , для которых такая нулевая гипотеза не отвергается на уровне значимости  $\alpha$  против двусторонней альтернативы.

Этот метод построения доверительных интервалов не слишком конструктивный, но иногда его можно применять, если под рукой нет никакого метода получше.

### 3.6.3. Проверка гипотез и построение доверительных интервалов при сравнении двух классификаторов

Пусть теперь помимо описанного ранее бинарного классификатора имеется второй классификатор, который на той же самой тестовой выборке верно предсказывает метки для 75 объектов из 100. Требуется определить, какой из двух классификаторов лучше. С одной стороны, 75 больше, чем 60. Но с другой стороны, выборка из 100 объектов не очень большая, и такая разница может возникнуть и случайно.

Учесть влияние случайности можно с помощью построения доверительных интервалов. Для первого классификатора доверительный интервал Уилсона для доли верных предсказаний: [0.502, 0.691]. Для второго классификатора такой же доверительный интервал: [0.657, 0.825]. Эти доверительные интервалы пересекаются по отрезку [0.657, 0.691]. Но пересечение доверительных интервалов не означает, что классификаторы нельзя различить по качеству. В данном случае выдвинута точечная нулевая гипотеза относительно двух параметров,  $\theta_1$  и  $\theta_2$ , и необходимо проверить её против двусторонней альтернативы:

$$H_0: \theta_1 = \theta_2, \quad H_1: \theta_1 \neq \theta_2.$$

Правильным решением будет построить доверительный интервал для разности параметров  $\theta_1$  и  $\theta_2$ , именно она полностью соответствует выдвинутой нулевой гипотезе (если  $\theta_1 = \theta_2$ , значит, их разность равна нулю). 95% доверительный интервал для разности долей в данной задаче: [0.022, 0.278]. Этот доверительный интервал не содержит ноль, значит, можно утверждать, что второй классификатор значимо лучше.

Если инвертировать этот доверительный интервал описанным ранее способом и выбрать наибольшее значение  $\alpha$ , при котором ноль попадает в доверительный интервал, получится уровень значимости  $p = 0.022$ . То есть, на уровне значимости 0.05 отвергается нулевая гипотеза о том, что два классификатора по качеству одинаковы.

		II	+	-	$\Sigma$
		I			
		+	55	5	60
		-	20	20	40
		$\Sigma$	75	25	100

Таблица 3.3: Таблица сопряжённости

Ранее не было учтено, что качество классификаторов определяется на одной и той же обучающей выборке, а значит, выборки в этой задаче — связанные. В такой ситуации доверительный интервал правильнее строить другим методом. Для этого используется таблица сопряжённости 3.3, и учитывается не количество ошибок каждого классификатора отдельно, а количество объектов, на которых классификаторы дали разные ответы (20 и 5 в этой таблице). Полученный 95% доверительный интервал для разности долей в связанных выборках равен [0.06, 0.243]. Обратите внимание, что этот доверительный интервал уже, и его левая граница дальше отстоит от нуля, то есть, при учёте связанныности увеличивается уверенность в том, что классификаторы отличаются. Для данного интервала достигаемый уровень значимости  $p = 0.002$ . Он почти в десять раз меньше, чем при построении доверительного интервала без учёта связаннысти выборок.

# **Урок 4**

## **Введение в АБ-тесты**

### **4.1. Что такое АБ-тестирование**

#### **4.1.1. Тестирование идей**

Во всех сферах бизнеса постоянно требуется улучшать его ключевые показатели. Идеи возможных улучшений могут возникать, с одной стороны, исходя из анализа рынка, поведения пользователей, их потребностей, а с другой — из знания об устройстве бизнеса, желаемом направлении развития.

Хочется уметь как-то проверять такие идеи: их много, а применять нужно только самые успешные. Существует много различных методов проверки. Прежде всего, идея должна соответствовать здравому смыслу. Как правило, это устанавливается в разговоре с коллегами, и принимается решение: стоит ли дальше заниматься идеей или от неё можно отказаться уже сейчас.

Затем бывает полезно протестировать идею на реальных пользователях. Как правило, такие проверки стоят дорого из-за привлечения людей, которые тратят на проверку свое время. Однако такой способ позволяет получить более релевантную оценку, потому что тестирование происходит на тех же пользователях, что и будут потом пользоваться услугой.

Существует ряд требований, предъявляемых к таким проверкам. С одной стороны, хочется, чтобы они были достоверными, то есть условия, в которых проверяется идея, были максимально приближены к реальным. С другой стороны, не хочется тратить на это очень много времени и денег.

В качестве проверки качества идеи может выступать опрос пользователей: заранее подготавливаются какие-то вопросы, а затем людей просят на них ответить.

#### **4.1.2. Фокус-группы**

«Фокус-группы» — это ещё один метод проверки идеи. Набирается небольшая группа пользователей, и от каждого из них поступает очень много информации: что он чувствует по отношению к новому баннеру; как именно на него кликает; производятся наблюдения, как долго его взгляд задерживается на каких-то деталях сайта. По сравнению с опросами, данные, полученные с помощью фокус-групп, глубже, однако такие исследования нельзя провести на большом количестве людей, поскольку это дорого.

Казалось бы, фокус-группы — почти идеальный механизм проверки бизнес-идей. Однако есть следующий нюанс: не всегда удается с точностью воссоздать те самые условия, в которых пользователи будут использовать продукт. Известный пример — тестирование количества сахара в напитке Кока-Кола. В рамках исследования была собрана фокус-группа, которой предложили на выбор два напитка: со стандартным содержанием сахара и увеличенным. Людям нужно было попробовать оба и выбрать тот, который им больше понравился. В результате этого исследования выяснилось, что большее количество людей предпочитает Кока-Колу с увеличенным количеством сахара. По результатам исследования количество сахара в напитке увеличили и предложили его более широкой аудитории. Неожиданно продажи упали. Эксперты стали разбираться, почему так произошло, ведь в рамках фокус-группы было показано, что больше сахара — это вкуснее. Выяснилось, что это исследование проходило не в тех самых условиях, в которых люди обычно пьют Кока-Колу. Если речь идёт всего лишь об одном стакане, то, действительно, людям нравится большее количество сахара. Однако если напиток употребляется постоянно в больших количествах, то больше сахара — хуже. Кажется, что это логично: сложно выпить большое количество очень сладкого напитка. Поэтому, когда проводятся исследования на фокус-группах, нужно следить за тем, чтобы условия были максимально приближены к настоящим.

Однако это не всегда возможно.

### 4.1.3. АБ-тестирование

Проблема приближения условий к настоящим решается полностью в рамках другого способа оценки идей: «A/B testing». «A/B testing» — это способ проверки идей непосредственно в боевых условиях. Пользователям предлагают новую функциональность или новый элемент дизайна именно в тех условиях, в которых они взаимодействуют с продуктом.

Очевидный плюс А/Б-теста в том, что, с одной стороны, условия больше похожи на настоящие. С другой стороны, финансовые риски максимально снижены: А/Б-тестирование проводится на небольшой группе пользователей, поэтому максимальные потери можно заранее спрогнозировать.

## 4.2. Где используется АБ-тестирование

В области ИТ А/Б-тесты используются практически повсеместно. Любые сайты, независимо от их профиля, используют механизм А/Б-тестирования для принятия решения о внесении каких-то изменений. Кроме того, такой метод используется в приложениях, играх и вообще во всем, что взаимодействует с конечным пользователем.

Во всех этих сферах нужно тестировать такие вещи, как изменения в дизайне, изменения функциональности, новые возможности для пользователя или изменения в алгоритмах.

С одной стороны, хочется как можно меньше времени тратить на эксперименты и проверять множество идей сразу. С другой стороны, изменения разных типов плохо тестируются одновременно. Может оказаться, что разные изменения просто технически несовместимы друг с другом. Или, например, одно из изменений действует на бизнес-показатель положительно, а другое — отрицательно, и, применяя их одновременно, не получится разделить эти два эффекта.

Такие техники тестирования идей применяются не только в ИТ, но и в довольно неожиданных местах, например, для оптимизации работы государственных органов. При правительствах США и Великобритании есть небольшие группы, которые совместно с психологами-бихевиористами выдвигают гипотезы о том, как люди взаимодействуют с государством, и на основании таких гипотез проводятся эксперименты: например, небольшие изменения в дизайне налоговой формы, или изменения в способе записи на донорство или трансплантацию органов. Такие вещи, которые легко и очень дешево можно поменять, оказывается, приводят к тому, что государство может сэкономить миллионы.

В целом, процесс А/Б-тестирования распадается на две большие части. Первая часть — это планирование эксперимента: как именно будет выглядеть А/Б-тест, как пользователей будут делить на группы, сколько будет длиться тест и многие другие вопросы, связанные с этим. Вторая часть — это непосредственно проверка гипотез, принятие решения о том, положительно или отрицательно влияют изменения на бизнес в целом.

## 4.3. Метрики

### 4.3.1. Что такое метрики

Итак, для того, чтобы показать эффективность идеи, нужно провести эксперимент, в котором она применяется, и при этом получить улучшение некого показателя. Этот показатель необходимо выбрать перед тем как проводить эксперимент. Речь идёт о выборе метрик. Часто для того, чтобы показать, что состояние бизнеса улучшилось, а именно, что его ключевые показатели изменились в ожидаемом направлении, нужно выбрать некоторые метрики, связанные напрямую с состоянием бизнеса.

Чаще всего это метрики, связанные непосредственно с деньгами, или аудиторные метрики. Однако при их применении возникают проблемы. Во-первых, часто их сложно измерить, во-вторых, они бывают достаточно грубыми и практически не реагируют на небольшие изменения в функциональности или дизайне. Кроме того, во многих случаях требуется очень много времени, чтобы измерить интересующие метрики. Например, после внесения изменения в сайт с арендой квартир, требуется узнать, как увеличилось количество людей, рекомендующих этот сервис своим знакомым. Чтобы измерить этот показатель, может понадобиться целый год, потому что люди переезжают не так уж часто.

### **4.3.2. Промежуточные метрики**

Это важное замечание приводит к идеи использования так называемых прокси-, или промежуточных метрик. Это такие метрики, которые, с одной стороны, достаточно чувствительны, чтобы измерять их в рамках А/Б-тестирования, а с другой стороны, хорошо согласуются с теми бизнес-показателями, которые в реальности требуется измерить. Например, при внесении изменений в сайт с арендой квартир в качестве такой метрики может быть использована метрика «среднее количество визитов на сайт в день», «среднее количество уникальных пользователей» или «количество шеров сайта в социальных сетях». Эти метрики обладают требуемыми свойствами промежуточных метрик.

### **4.3.3. Оффлайн-тестирование**

Важным этапом при принятии решения о том, следует ли проверять изменение в А/В-тестинге, является оффлайн-тестирование. В рамках оффлайн-тестирования можно по историческим данным проверить, как определённые изменения сказываются на поведении пользователей.

Допустим, изменён алгоритм ранжирования результатов по запросам пользователей при поиске квартир на сайте недвижимости. В этом случае можно поступить следующим образом: проанализировать запросы пользователей в прошлом (если известно, что пользователи искали на сайте ранее), и эмулировать выдачу новым алгоритмом.

Таким образом, с одной стороны, доступна информация о том, что пользователи искали и на какие позиции они кликали, какие ответы показались им релевантными. С другой стороны, теперь имеется новое ранжирование, новые результаты поиска. Совместив эти данные, можно проверить, на какие позиции теперь приходятся клики пользователей. В результате можно получить такие метрики, как «средняя позиция клика». Если значение этой метрики уменьшается, то, наверное, внесённое изменение хорошее. Значит, имеет смысл протестировать его в онлайне. А если, например, новый алгоритм ранжирования совсем не находит те результаты, которые пользователи посчитали релевантными, то, возможно, этот алгоритм не стоит тестировать на реальных пользователях.

Возникает следующая иерархия метрик: предварительные метрики, измеряемые до начала эксперимента, экспериментальные, на основании которых принимается решение о том, хорошее изменение или плохое, и бизнес-метрики, которые измеряются в самом конце и на изменение которых направлены все действия.

## **4.4. Дизайн эксперимента**

### **4.4.1. Стратификация и рандомизация**

Для проведения эксперимента требуется небольшая группа пользователей, которой будут предъявлены изменения. Для того, чтобы результаты, полученные на этой небольшой группе можно было обобщать на всех пользователей, группа должна быть репрезентативной. Это значит, что её структура должна совпадать со структурой набора всех пользователей. Например, если известно, что 2/3 пользователей продукта — женщины, то в экспериментальной группе должно быть 2/3 женщин.

Таким образом, при построении экспериментальной группы можно выделять какие-то важные свойства пользователей: например, возраст, или другие интересующие характеристики, — а затем искусственно делать так, чтобы в экспериментальной группе были ровно такие же доли по разным подгруппам, как и среди пользователей в целом. Такой подход называется стратификацией.

Другой подход, в каком-то смысле противоположный ему, — это рандомизация. Если набирать пользователей в экспериментальную группу абсолютно случайно, то в среднем она получится такого же состава, как и вся генеральная совокупность пользователей. Дополнительный плюс рандомизации заключается в том, что при этом экспериментальная группа пользователей выравнивается со генеральной совокупностью по всем возможным показателям, а не только по тем, которые оказались важными.

### **4.4.2. Связанные выборки**

В некоторых случаях оказывается важным измерить, как на пользователя влияет несколько воздействий сразу, например, как он реагирует на сайт без изменений и на сайт с изменениями. Такой дизайн эксперимента называется парным, или связанным. Выборки результатов получаются не зависимые, а связанные, и это очень выгодно в ситуациях, когда измеряемый показатель имеет большую индивидуальную дисперсию (то есть пользователи очень сильно отличаются по этому показателю).

При связанном дизайне эксперимента зачастую оказывается важным, в каком порядке пользователю предъявляются разные варианты. Для того, чтобы снять влияние порядка, можно использовать дизайн крест-накрест: половине пользователей показать сначала новый вариант, потом — старый, а другой половине — наоборот.

#### 4.4.3. Проведение нескольких экспериментов сразу

Одновременно можно проводить большое количество экспериментов. Но в этой ситуации не возникает никаких проблем только до тех пор, пока каждый пользователь участвует в одном эксперименте. Если существует вероятность, что каждый пользователь попадает сразу в несколько экспериментальных групп, нужно внимательно следить за тем, чтобы эти эксперименты друг другу не противоречили.

Например, известна история о том, как Google тестировал 41 оттенок синего в цвете ссылок в поисковой выдаче. Если допустить, что одновременно еще проводился бы эксперимент о том, как выбрать цвет страницы или цвет, на фоне которого показываются эти ссылки, очевидно, что они не должны быть тех же самых цветов, что и текст ссылок, иначе пользователь просто не сможет ничего прочитать. То есть пример подбора одновременно цвета текста и цвета фона, на котором он показывается, — это пример экспериментов, которые друг с другом не сочетаются.

### 4.5. Устойчивость

Одно из важнейших требований к А/Б-тестированию, которое обязательно должно быть заложено в дизайн эксперимента, — это требование устойчивости. В данном случае под устойчивостью понимается следующее: во-первых, хочется не видеть значимых изменений там, где их на самом деле нет, во вторых, если какие-то значимые изменения есть, то хочется, чтобы они отражались на метриках.

Это очень просто понять на примере. Пусть в эксперименте участвуют две одинаковые версии сервиса. В данном случае, конечно же, на всех метриках хочется видеть одинаковый результат. С другой стороны, если в одну из версий внесены некоторые значимые изменения, то, конечно, хочется увидеть это на метриках и убедиться, что по всем метрикам есть значимый прирост.

#### 4.5.1. Обратный эксперимент

Казалось бы, требования устойчивости очень простые, логичные и должны всегда выполняться. Однако на практике это часто не так. Например, есть некоторый сайт, который позволяет производить поиск по некоторому специальному контенту, например, по объявлениям о продаже/аренде недвижимости. Можно изменить дизайн и проверить, как это повлияло на поведение пользователей: правда ли, что новый дизайн им нравится больше, и они начинают более активно пользоваться сервисом. Можно сделать очень простое изменение, например, перекрасить кнопку поиска из синего цвета в зеленый. В данном случае легко понять, как будет выглядеть А/Б-тестирование. Пользователей разбирают на тестовую и контрольную группу. Одной группе будут показывать кнопку синего цвета (старый дизайн), а другой группе пользователей — кнопку зеленого цвета (новый дизайн). Далее можно подсчитать онлайн-метрику (количество нажатий на эту кнопку), и посмотреть, как она изменилась. Часто в таких экспериментах можно наблюдать следующий эффект: количество кликов будет больше в контрольной группе (с новым дизайном). Трактовка этого может быть двойкой. Те пользователи, которые часто пользуются сайтом и уже привыкли к тому, что кнопка имеет синий цвет, могут удивиться, что что-то изменилось, и захотеть проверить, изменился ли только дизайн или, может быть, и поведение. Соответственно, они могут начать чаще нажимать на кнопку просто из любопытства. В данном случае важно убедиться, что наблюдаемые метрики не учитывают это изменение как значимое.

Для того, чтобы обезопасить себя от ситуации, в которой незначимые изменения принимаются за значимые, можно поступить следующим образом. Пусть классический А/В-тест показал, что новый дизайн лучше, то есть кнопка зеленого цвета больше нравится пользователям. По измеряемым метрикам (например, доле кликов или длине сессий) наблюдаются значимые улучшения. Логично сделать следующее: выбрать новый дизайн и применить его для всех пользователей, то есть показывать всем пользователям кнопку зеленого цвета. Однако можно поступить несколько хитрее: выбрать небольшую группу пользователей (например, меньше 1 %) и продолжать показывать им старый дизайн после выкатки нового. То есть будет отдельно существовать некоторая группа пользователей, которая в течение какого-то существенного промежутка времени будет видеть старый дизайн. Это позволит в течение большего срока рассчитывать те же самые метрики, что и в процессе А/Б-тестирования. После этого можно снова сравнить поведение пользователей, которые

видят новый дизайн, с теми, кто видит старый дизайн. В данном случае, если значимые изменения не будут наблюдаться, то можно сделать вывод о том, что, во-первых, дизайн эксперимента не позволяет отличать незначимые изменения от значимых (иначе результат первоначального А/Б-тестирования был бы таким же), а во-вторых, можно оставить любой дизайн, потому что пользователи их не отличают.

Такая техника называется обратным экспериментом. Ее идея заключается в том, что после классического А/Б-тестирования эксперимент продолжается: выделяется некоторая маленькая группа пользователей, которые продолжают видеть старое решение. Такой подход предоставляет возможность убедиться в том, что изменения действительно значимы и приводят к ожидаемому эффекту.

#### 4.5.2. А/А-тестирование

Казалось бы, технология обратного эксперимента способна решить все проблемы и помочь очевидным образом отличать значимые изменения от незначимых.

Однако пусть А/Б-тестирование проводится часто и каждый раз демонстрирует значимые изменения метрик на целевой и контрольной группе. Тестируемые нововведения запускаются в технологию «обратный эксперимент», и оказывается, что на самом деле изменений нет. Конечно же, эта ситуация является крайне нежелательной, потому что на проведение А/Б-теста и обратного эксперимента тратится много времени. Возникает вопрос: можно ли заранее убедиться в том, что дизайн эксперимента (в частности, размер контрольной и целевой групп, а также длительности эксперимента) позволяет отличать значимые изменения от незначимых.

Для того, чтобы эту задачу решить, применяется технология А/А-тестирования. Она работает следующим образом. Пусть принято решение о проведении А/Б-тестирования нового алгоритма (например поиска или рекомендаций). В таком случае классический А/Б-тест выглядел бы следующим образом: пользователей разделили бы на контрольную и тестовую группу и показывали бы разные алгоритмы в разных группах, после чего сравнили бы метрики, рассчитанные на разных группах.

Перед тем, как запускать классический А/Б тестинг, можно поделить пользователей на группы точно так же, как это сделали бы в рамках А/Б тестинга, но обеим группам демонстрировать один и тот же алгоритм. Этот эксперимент должен длиться ровно столько же, сколько бы длился А/Б тестинг. В результате нужно определить, видны ли значимые изменения на интересуемых метриках. Если значимые изменения не видны, то это хорошо, потому что эксперимент не показывает значимые изменения там, где их нет. В противоположной ситуации, если вдруг появятся значимые изменения, это должно наводить на мысль, что в дизайне эксперимента что-то не так: например, эксперимент длится недостаточно долго, или пользователи неправильно разбиты на группы. В любом случае, это повод задуматься об ошибках в дизайне эксперимента.

#### 4.5.3. Размер выборки

Итак, метрика, дизайн эксперимента, метод его анализа выбраны, вся экспериментальная инфраструктура в достаточной степени устойчива и практически всё готово к запуску эксперимента. Единственный вопрос, на которой остается ответить, — это как долго эксперимент должен длиться, и сколько пользователей должно быть в тестовой выборке, чтобы можно было с уверенностью ответить на поставленные вопросы.

Задача определения необходимого объема выборки тесно связана с тем, какой именно статистический инструмент будет использоваться для ее анализа. Для каждого конкретного критерия подбор необходимого объема выборки делается своим способом.

Для того, чтобы понять, какой объем выборки необходим, нужно зафиксировать некоторые параметры. Во-первых, минимальный размер эффекта, который хочется измерить. То есть, насколько большие отклонения от значения по умолчанию (показатель, который сохраняется, если изменения никак не влияют на пользователей) хочется наблюдать в эксперименте.

Следующий показатель, который необходимо зафиксировать, — это допустимые вероятности ошибок первого и второго рода. В А/Б-тестах, как правило, выдвигается нулевая гипотеза, что никакие примененные изменения не повлияли на пользователей, и она проверяется против альтернативы, что изменения как-то повлияли. Ошибкой первого рода в этой ситуации будет отвержение неверной нулевой гипотезы, то есть принятие изменений, которые на самом деле не влияют на пользователей. Ошибка второго рода — это, наоборот, отклонение действительно хороших и влияющих на пользователей изменений. В статистике, как правило, вероятность ошибки первого рода — 0.05, а вероятность ошибки второго рода — 0.2. В конкретном эксперименте стоимости ошибок первого и второго рода могут быть существенно разными, поэтому часто может оказаться выгодно вручную выбрать эти пороги.

Наконец, когда размер эффекта и допустимые вероятности ошибок зафиксированы, можно выбрать статистический критерий и использовать калькулятор мощности этого критерия. Вообще, для всех статистических критериев между собой связаны несколько величин: тип альтернативы, размер эффекта, размер выборки и допустимые вероятности ошибок первого и второго рода. Если зафиксировать какие-то из этих величин, то можно рассчитать оставшиеся, используя калькулятор мощности.

# Урок 5

## Параметрические критерии

Этот урок посвящен параметрическим критериям проверки гипотез. Эти критерии называются параметрическими потому, что в проверяемых ими гипотезах высказывается предположение о значении параметра распределений, из которых предположительно взята выборка.

### 5.1. Одновыборочные критерии Стьюдента

Семейство критериев Стьюдента позволяет проверять гипотезы о математических ожиданиях нормальных распределений.

**Пример: средний вес детей при рождении.** Средний вес детей при рождении составляет 3300 г. В то же время, если мать ребёнка живёт за чертой бедности, то средний вес таких детей — 2800 г. Вес при рождении — это очень важный показатель здоровья ребенка. Так, только 7% детей рождаются с весом меньше 2.5 кг, однако на них приходится 70% детских смертей.

С целью увеличить вес тех детей, чьи матери живут за чертой бедности, разработана экспериментальная программа ведения беременности. Чтобы проверить ее эффективность, проводится эксперимент. В нем принимают участие 25 женщин, живущих за чертой бедности. У всех них рождаются дети, и их средний вес составляет 3075 г.

Для того, чтобы ответить на вопрос, эффективна ли программа, используется критерий Стьюдента.

#### 5.1.1. Z-критерий

Информация о критерии суммирована в таблице 5.1, нулевое распределение показано на рисунке 5.1. Этот критерий называется Z-критерием (как и большинство критериев, статистики которых имеют стандартное нормальное нулевое распределение).

выборка:	$X^n = (X_1, \dots, X_n),$ $X \sim N(\mu, \sigma^2), \sigma$ известна;
нулевая гипотеза:	$H_0: \mu = \mu_0;$
альтернатива:	$H_1: \mu < \neq > \mu_0;$
статистика:	$Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}};$
нулевое распределение:	$Z(X^n) \sim N(0, 1).$

Таблица 5.1: Описание Z-критерия

Способ подсчёта достигаемого уровня значимости такого критерия зависит от используемого типа альтернативы. Если альтернатива односторонняя:

$$H_1: \mu < \mu_0,$$

то, если она справедлива, более вероятными являются маленькие значения Z-статистики, то есть, левый хвост нулевого распределения (рисунок 5.2). Таким образом, чтобы посчитать достигаемый уровень значимости, нужно взять интеграл плотности стандартного нормального распределения от  $-\infty$  до значения статистики  $Z$ ,

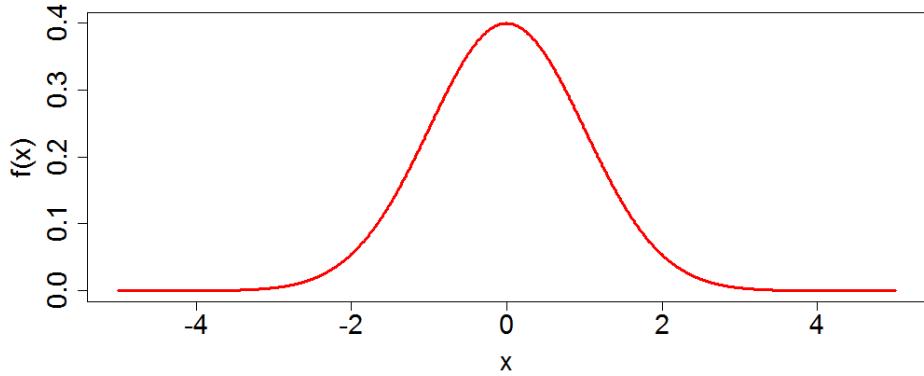


Рис. 5.1: Стандартное нормальное распределение

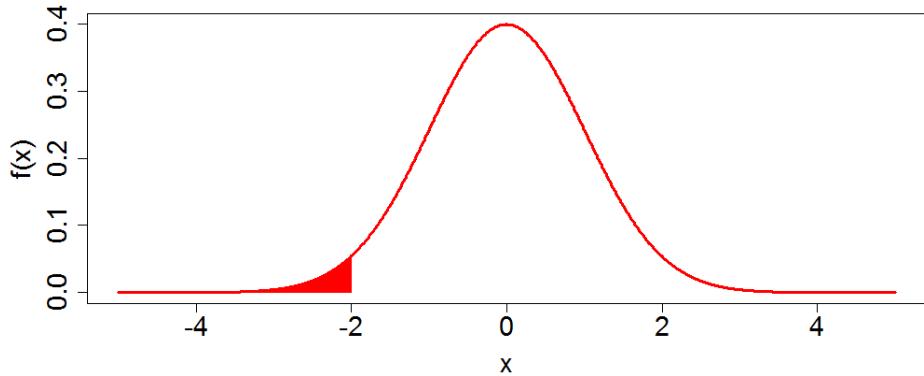


Рис. 5.2: Более вероятные значения статистики при использовании альтернативы  $H_1: \mu < \mu_0$

реализовавшегося в эксперименте. Например, если в эксперименте получено значение  $Z = -2$ , то достигаемый уровень значимости — это интеграл плотности стандартного нормального распределения от  $-\infty$  до  $-2$ . На самом деле, значение интеграла вычислять не нужно, потому что оно в точности равно значению функции стандартного нормального распределения в точке  $Z$ :

$$p = F_{N(0,1)}(z).$$

Если используется противоположная односторонняя альтернатива вида

$$H_1: \mu > \mu_0,$$

то более вероятными являются большие значения статистики (рис. 5.3), и, чтобы посчитать достигаемый уровень значимости, нужно взять интеграл по правому хвосту плотности нулевого распределения. Этот интеграл, в свою очередь, равен

$$p = 1 - F_{N(0,1)}(z).$$

Если же используется двусторонняя альтернатива

$$H_1: \mu \neq \mu_0,$$

то при ее справедливости более вероятными будут большие по модулю значения статистики  $Z$  (рис. 5.4). Поэтому при подсчете достигаемого уровня значимости представляют интерес и левый, и правый хвосты нулевого распределения. Если получено значение статистики  $Z = -2$ , то достигаемый уровень значимости равен сумме интегралов от  $-\infty$  до  $-2$  и от  $2$  до  $\infty$ . Чтобы не считать эти два интеграла, можно снова использовать функцию стандартного нормального распределения:

$$p = 2(1 - F_{N(0,1)}(|z|)).$$

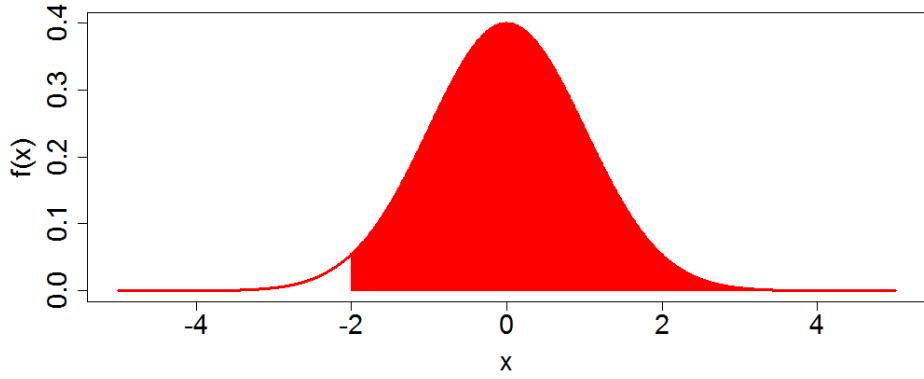


Рис. 5.3: Более вероятные значения статистики при использовании альтернативы  $H_1: \mu > \mu_0$

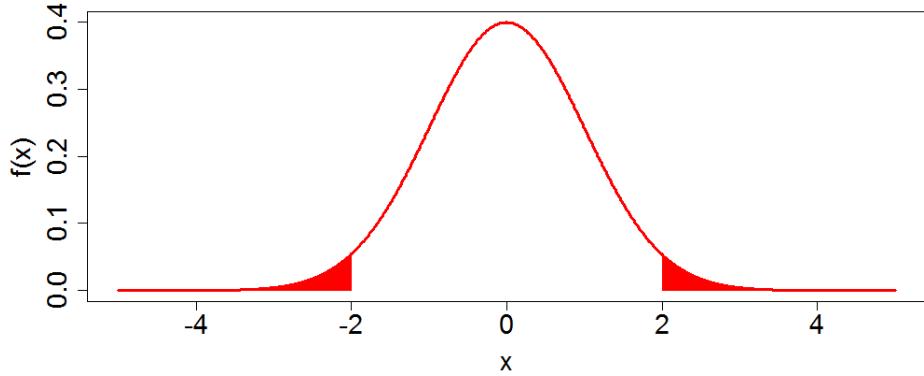


Рис. 5.4: Более вероятные значения статистики при использовании альтернативы  $H_1: \mu \neq \mu_0$

### 5.1.2. t-критерий

Если дисперсия выборки неизвестна, вместо Z-критерия Стьюдента нужно применять t-критерий Стьюдента (таблица 5.2). Он основан на следующей идее: поскольку  $\sigma$  неизвестна, то нужно там, где используется  $\sigma$  (в формуле статистики), заменить  $\sigma$  на  $S$  (выборочное стандартное отклонение). Такая статистика имеет уже не стандартное нормальное нулевое распределение, а распределение Стьюдента с числом степеней свободы  $n-1$  (рис. 5.5).

выборка:	$X^n = (X_1, \dots, X_n)$ ,
нулевая гипотеза:	$X \sim N(\mu, \sigma^2)$ , $\sigma$ неизвестна;
альтернатива:	$H_0: \mu = \mu_0$ ;
статистика:	$H_1: \mu < \neq \mu_0$ ;
нулевое распределение:	$T(X^n) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ ;
	$T(X^n) \sim St(n-1)$ .

Таблица 5.2: Описание t-критерия

Достигаемый уровень значимости для t-критерия Стьюдента считается абсолютно так же, как и для Z-критерия. В зависимости от типа альтернативы нужно выбрать одно из трех выражений для достигаемого уровня значимости:

$$p = \begin{cases} F_{St(n-1)}(t), & H_1: \mu < \mu_0, \\ 1 - F_{St(n-1)}(t), & H_1: \mu > \mu_0, \\ 2(1 - F_{St(n-1)}(|t|)), & H_1: \mu \neq \mu_0. \end{cases}$$

Единственное отличие от Z-критерия заключается в том, что вместо функции стандартного нормального распределения используется функция распределения Стьюдента с числом степеней свободы  $n-1$ .

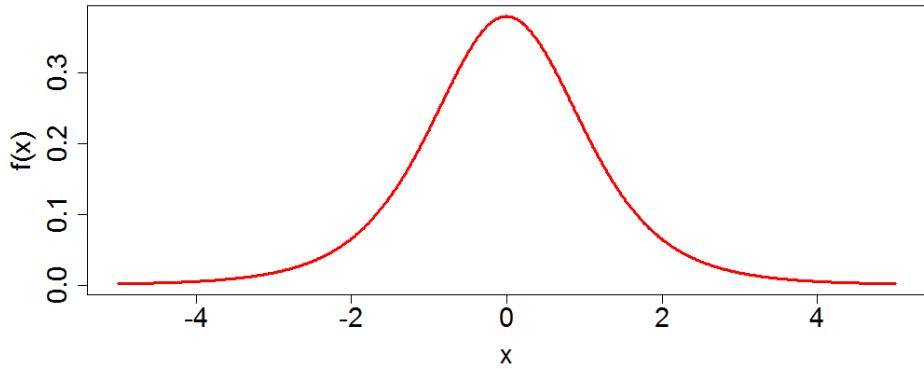


Рис. 5.5: Распределение Стьюдента

Чем больше объем выборки в задаче, тем меньше различий между t-критерием и Z-критерием. Это происходит по двум причинам: во-первых, чем больше  $n$ , тем точнее выборочная дисперсия  $S^2$  оценивает теоретическую дисперсию  $\sigma^2$ . Во-вторых, с ростом  $n$  увеличивается число степеней свободы у нулевого распределения t-критерия, а чем больше степеней свободы у распределения Стьюдента, тем больше оно похоже на стандартное нормальное. Ранее уже упоминалось, что, начиная с 30 степеней свободы, распределение Стьюдента визуально практически неотличимо от стандартного нормального. Благодаря этим двум фактам для достаточно больших выборок знание истинного значения дисперсии не оказывает большое влияние на результат.

### 5.1.3. Математическая формулировка задачи о весе детей при рождении

**Пример: вес детей при рождении (продолжение)** С использованием описанных критериев можно математически сформулировать задачу оценки эффективности экспериментальной программы, о которой шла речь ранее.

Выдвигается нулевая гипотеза о том, что программа неэффективна:

$$H_0: \mu = 2800,$$

то есть средний вес детей, прошедших экспериментальную программу, такой же, как и в целом у детей, живущих за чертой бедности. Этую нулевую гипотезу необходимо проверить против двусторонней альтернативы (программа как-то влияет на вес детей):

$$H_0: \mu \neq 2800.$$

Поскольку теоретическая дисперсия  $\sigma^2$  неизвестна, а известна только ее выборочная оценка  $S^2$ , то нужно использовать t-критерий Стьюдента. С его помощью для такой пары гипотеза-альтернатива получается достигаемый уровень значимости  $p = 0.0111$ , то есть нулевая гипотеза отклоняется. Точечная оценка для прироста среднего веса детей в результате экспериментальной программы — это  $\mu - \mu_0 = 275$  г. 95% доверительный интервал для этой величины получается из применённого критерия Стьюдента, и он составляет [233.7, 316.3] г.

Вообще говоря, в этой задаче нулевую гипотезу можно проверять против не двусторонней, а односторонней альтернативы. Тогда нулевая гипотеза остаётся той же (программа неэффективна):

$$H_0: \mu = 2800,$$

а альтернатива — «программа эффективна», то есть средний вес детей в результате программы повышается:

$$H_0: \mu > 2800.$$

Для такой пары гипотеза-альтернатива t-критерий дает достигаемый уровень значимости ровно в 2 раза меньше:  $p = 0.0056$ . Точечная оценка для прироста среднего веса не меняется и составляет все еще 375 граммов. А вот доверительный интервал становится односторонним, то есть утверждается, что на уровне доверия 95% средний вес детей увеличивается не меньше, чем на 241 грамм.

### 5.1.4. Выбор альтернативы

Рассмотренный пример показывает, что, используя одностороннюю альтернативу вместо двусторонней, можно получить достигаемый уровень значимости в два раза меньше. В данной задаче это было не критично, поскольку оба достигаемых уровня значимости маленькие. Но иногда может оказаться, что  $p$ -value при двусторонней альтернативе больше магического порога в 0.05, а при односторонней альтернативе — меньше. То есть, используя одностороннюю альтернативу вместо двухсторонней, можно отвергнуть нулевую гипотезу. В таком случае кажется, что можно всегда использовать одностороннюю альтернативу, однако это нечестно. Альтернатива может быть односторонней только в некоторых случаях. Во-первых, если среднее должно изменится в какую-то определенную сторону и изменение в противоположную сторону невероятно. Во-вторых, направление изменения нужно определить до получения данных. Если односторонняя альтернатива выбирается после того, как данные получены, и она выбирается так, что ее знак соответствует знаку изменения выборочного среднего относительно  $\mu_0$ , то это — переобучение, и так делать нельзя.

## 5.2. Двухвыборочные критерии Стьюдента, независимые выборки

С помощью двухвыборочных критериев Стьюдента можно сравнивать среднее значение двух выборок из нормального распределения. Использование этих критериев будет продемонстрировано на данных General Social Survey. Это социологический опрос, который проводится на достаточно больших выборках в США уже больше 40 лет. В этом опросе очень много вопросов, которые задают респондентам, здесь будет рассматриваться только один.

В 1974 году число респондентов, работающих неполный рабочий день, составляло 108. В 2014 году — 196. Для каждого из опрошенных известно количество рабочих часов за неделю, предшествующую опросу. Используя эти данные, требуется понять, изменилось ли за прошедшие 40 лет среднее время работы у работающих неполный день.

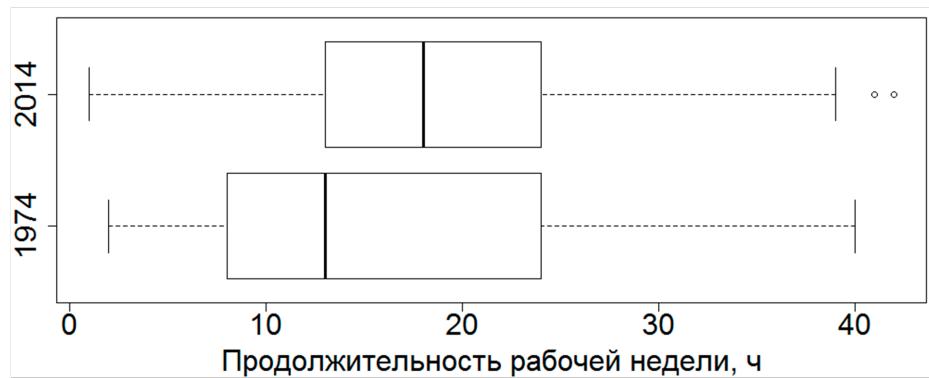


Рис. 5.6: Количество рабочих часов у работающих неполный день в 1974 году и в 2014 году

Данные, которые требуется проанализировать, показаны на рисунке 5.6. Изображённый график называется boxplot, или ящик с усами. Это способ визуализации основных характеристик распределения, принцип построения такого графика показан на рисунке 5.7. Boxplot состоит из прямоугольника — ящика — и торчащих из него усов. Чертёж в середине прямоугольника соответствует выборочной медиане выборки. Ширина ящика равна интерквартильному размаху, то есть, его нижняя граница — это 25%-й квантиль, а верхняя — 75%-й квантиль. Длина усов составляет 1.5 интерквартильных размаха, однако в разных реализациях кончик уса может рисоваться в разных местах. Так, на рисунке 5.7 усы обрезаются так, что их конец соответствует последнему элементу выборки в этом направлении. Два кружочка на верхнем графике на рисунке 5.6 — это объекты выборки, не попавшие в диапазон 1.5 интерквартильных размаха.

Из рисунка 5.6 видно, что выборочные медианы выборок, соответствующих 1974 и 2014 году, отличаются. В 2014 году люди работали в среднем больше. Для того, чтобы проверить, значимо ли это различие,

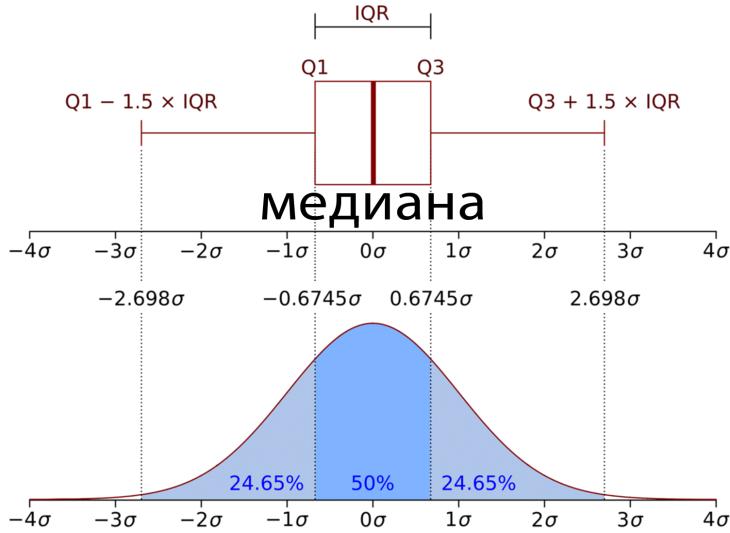


Рис. 5.7: Структура ”ящика с усами”

необходимо использовать статистический критерий.

### 5.2.1. Z-критерий

Если имеются две выборки из нормального распределения с различными параметрами, дисперсии для которых известны, то используется Z-критерий (таблица 5.3). Статистика этого критерия имеет стандартное нормальное распределение (рисунок 5.1).

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$ $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}),$ $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2),$ $\sigma_1, \sigma_2$ известны;
нулевая гипотеза:	$H_0: \mu_1 = \mu_2;$
альтернатива:	$H_1: \mu_1 \neq \mu_2;$
статистика:	$Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}};$
нулевое распределение:	$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0, 1).$

Таблица 5.3: Описание двухвыборочного Z-критерия

### 5.2.2. t-критерий

В более сложном случае дисперсии выборок неизвестны. Можно действовать по аналогии с одновыборочным случаем: в формуле для Z-критерия заменить все неизвестные  $\sigma$  на их выборочные оценки  $S_1$  и  $S_2$ . Получится t-статистика (таблица 5.4). При выполнении нулевой гипотезы она будет распределена по Стьюденту (рисунок 5.5).

У этой задачи есть две проблемы. Во-первых, число степеней свободы  $\nu$  у этого нулевого распределения Стьюдента вычисляется по достаточно сложной формуле:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Во-вторых, нулевое распределение t-статистики не точное, а приближенное. Точного решения (то есть точного нулевого распределения для такой статистики) не существует. Эта проблема называется проблемой Беренца-Фишера: невозможно точно сравнить средние значения в двух выборках, дисперсии которых неизвестны.

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$ $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}),$ $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2),$ $\sigma_1, \sigma_2$ неизвестны;
нулевая гипотеза:	$H_0: \mu_1 = \mu_2;$
альтернатива:	$H_1: \mu_1 < \neq > \mu_2;$
статистика:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$
нулевое распределение:	$T(X_1^{n_1}, X_2^{n_2}) \approx St(\nu).$

Таблица 5.4: Двухвыборочный t-критерий

Однако рассмотренная аппроксимация достаточно точна в двух ситуациях. Во-первых, если выборки  $X_1$  и  $X_2$  одинакового объема, то есть  $n_1 = n_2$ . Во-вторых, если знак неравенства между  $n_1$  и  $n_2$  такой же, как между  $\sigma_1$  и  $\sigma_2$ , то есть выборка с большей дисперсией по размеру не может быть меньше другой выборки. Если это условие выполнено, то можно использовать t-критерий Стьюдента и не переживать о точности аппроксимации. Если это не так, то возникает проблема: критерий Стьюдента перестает правильно работать, и вероятность ошибок первого рода начинает превышать уровень значимости  $\alpha$ . Это проблема не только критерия Стьюдента, она возникает при проверке любым способом гипотезы о равенстве средних в двух выборках с разной дисперсией. Поэтому, сравнивая средние значения в двух выборках, важно всегда следить за тем, чтобы выборка с большей дисперсией всегда была не меньшего объема, чем вторая выборка.

Итак, необходимо проверить нулевую гипотезу о том, что средняя продолжительность рабочей недели у людей, которые работают не полный рабочий день, не изменилась за прошедшие 40 лет:

$$H_0: \mu_1 = \mu_2.$$

Альтернативная гипотеза двусторонняя, среднее время работы изменилось:

$$H_0: \mu_1 \neq \mu_2.$$

Можно было бы использовать и одностороннюю альтернативу. Однако сложно заранее предугадать знак сравнения, и данные уже известны, так что выбирать одностороннюю альтернативу нечестно.

Критерий Стьюдента в этой задаче дает достигаемый уровень значимости  $p = 0.02707$ . То есть, гипотеза о равенстве средних отвергается на уровне значимости 0.05. Точечная оценка для прироста средней продолжительности рабочей недели составляет 2.57 часов. 95%-й доверительный интервал для нее: [0.29, 4.85] ч. То есть, люди в среднем стали работать больше, и доверительный интервал прироста этого времени составляет от получаса до 5 часов.

## 5.3. Двухвыборочные критерии Стьюдента, связанные выборки

### 5.3.1. Лечение СДВГ

Проводится исследование метода лечения синдрома дефицита внимания и гиперактивности (СДВГ) у умственно отсталых детей. В эксперименте участвуют 24 ребенка. Каждый из них неделю принимает плацебо, а неделю препарат метилфенидат. По окончании каждой недели каждый ребенок проходит тест на способность к подавлению импульсивных поведенческих реакций. Анализируемые данные показаны на диаграмме рассеяния, (см. рисунок 5.8). По горизонтальной оси отложена способность к подавлению импульсивных поведенческих реакций после недели приема плацебо, по вертикальной — после недели приема препарата. Каждая точка соответствует одному ребенку. Таким образом, несмотря на то, что имеются две выборки, они не являются независимыми, поскольку значения здесь измерены на одних и тех же объектах. Такие выборки называются связанными.

Хочется понять, эффективно ли лечение с помощью метилфенидата. Большая часть точек на этом графике лежит выше диагонали. Это значит, что после приема метилфенидата у большинства детей способность к подавлению импульсивных поведенческих реакций увеличилась. Для того, чтобы определить, значимо ли это изменение, необходимо использовать статистический критерий.

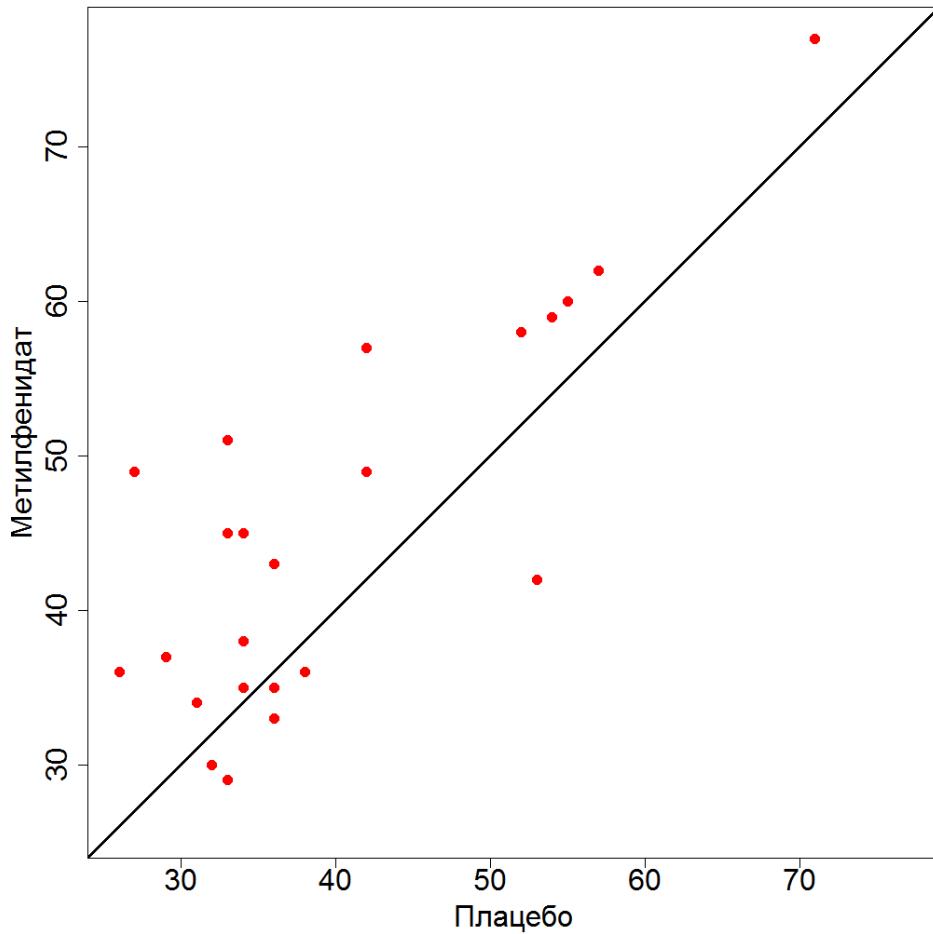


Рис. 5.8: Результаты эксперимента по сравнению действия плацебо и препарата метилфенидат на умственно-отсталых детей с синдромом дефицита внимания и гиперактивности

### 5.3.2. t-критерий Стьюдента для связанных выборок

Для проверки равенства математических ожиданий двух выборок одинакового объёма из нормальных распределений используется t-критерий Стьюдента (таблица 5.5).

выборки:	$X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim N(\mu_1, \sigma_1^2),$
нулевая гипотеза:	$X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim N(\mu_2, \sigma_2^2),$
альтернатива:	$H_0: \mu_1 = \mu_2;$
статистика:	$H_1: \mu_1 < \neq > \mu_2;$
	$T(X_1^n, X_2^n) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$
	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2, D_i = X_{1i} - X_{2i};$
нулевое распределение:	$T(X_1^n, X_2^n) \sim St(n-1).$

Таблица 5.5: Описание t-критерия Стьюдента для связанных выборок

В числителе Т-статистики стоит разность  $\bar{X}_1 - \bar{X}_2$ . Это то же самое, что  $\overline{X_1 - X_2}$ . Таким образом, t-критерий для двух связанных выборок эквивалентен одновыборочному t-критерию, примененному к выборке попарных разностей.

### 5.3.3. Применение t-критерия Стьюдента к задаче лечения СДВГ

В описанной ранее задаче нулевая гипотеза — это неэффективность лечения (способность к подавлению импульсивных поведенческих реакций не изменилась):

$$H_0: \mu_1 = \mu_2.$$

Проверить эту гипотезу нужно против двусторонней альтернативы поскольку нельзя исключать, что способность к подавлению импульсивных поведенческих реакций в результате применения препарата может уменьшиться:

$$H_1: \mu_1 \neq \mu_2.$$

t-критерий Стьюдента для связанных выборок дает значение достигаемого уровня значимости  $p = 0.00377$ . Нулевая гипотеза о том, что средняя способность к подавлению импульсивных поведенческих реакций не изменилась, отвергается на уровне значимости 0.05. Точечная оценка изменения признака в результате применения препарата (разность выборочных средних) — 4.95 пунктов. 95% доверительный интервал для этой величины, построенный с помощью распределения Стьюдента: [1.78, 8.14] пунктов.

## 5.4. Нормальность выборок

### 5.4.1. Критерий хи-квадрат

Критерии Стьюдента проверяют гипотезы о средних значениях выборок в предположении, что эти выборки взяты из нормального распределения. Нормальность можно проверять с помощью критерия согласия Пирсона, или критерия хи-квадрат (таблица 5.6).

выборка: нулевая гипотеза: альтернатива: статистика: нулевое распределение:	$X^n = (X_1, \dots, X_n);$ $H_0: X \sim N(\mu, \sigma^2);$ $H_1: H_0$ неверна; $\chi^2(X^n) = \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i};$ $\chi^2(X^n) \sim \begin{cases} \chi^2_{K-1}, & \mu, \sigma \text{ заданы,} \\ \chi^2_{K-3}, & \mu, \sigma \text{ оцениваются;} \end{cases}$ $n_i$ — число элементов выборки в $[a_i, a_{i+1}]$ , $p_i = F_{N(\mu, \sigma^2)}(a_{i+1}) - F_{N(\mu, \sigma^2)}(a_i)$ .
---	--

Таблица 5.6: Описание критерия хи-квадрат

Статистика критерия конструируется следующим образом: область изменения случайной величины разбивается на  $K$  интервалов (карманов). Границы этих интервалов задаются величинами  $a_i$ . Для каждого интервала  $[a_i, a_{i+1}]$  вычисляются две величины. Во-первых,  $n_i$  — число элементов выборки, которое попало в интервал. Во-вторых,  $p_i$  — теоретическая вероятность попадания в этот интервал при условии справедливости нулевой гипотезы. В данном случае это разность функций нормального распределения в точках  $a_{i+1}$  и  $a_i$ :

$$p_i = F_{N(\mu, \sigma^2)}(a_{i+1}) - F_{N(\mu, \sigma^2)}(a_i).$$

Значение статистики выглядит следующим образом:

$$\chi^2(X^n) = \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i}.$$

Если нулевая гипотеза справедлива, то такая статистика имеет распределение хи-квадрат (рисунок 5.9).

Чтобы вычислить достигаемый уровень значимости, необходимо взять интеграл от распределения хи-квадрат, начиная от значения статистики, которое реализуется в данных, до бесконечности.

Критерий хи-квадрат обладает несколькими очевидными недостатками. Во-первых, разбиение на интервалы в нём никак не зафиксировано, и, выбирая различные интервалы, можно получать разные результаты. Кроме того, для того, чтобы его использовать, необходимо иметь достаточно большую выборку: ожидаемое количество объектов выборки  $np_i$  в каждом интервале должно превышать 5 как минимум для 80% ячеек.

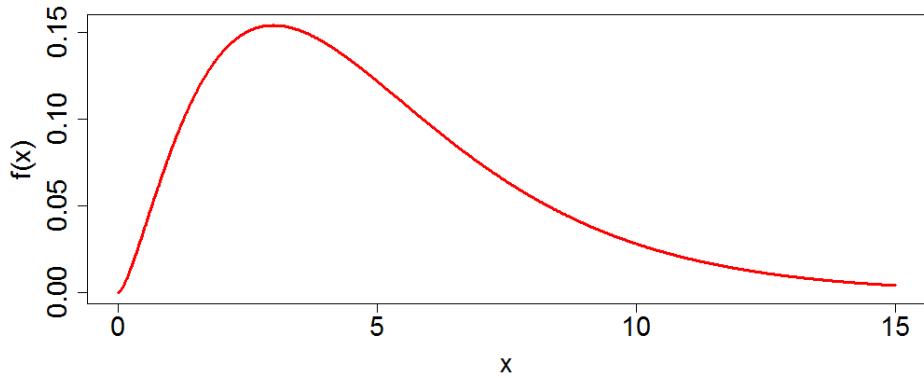


Рис. 5.9: Распределение хи-квадрат

#### 5.4.2. Q-Q график

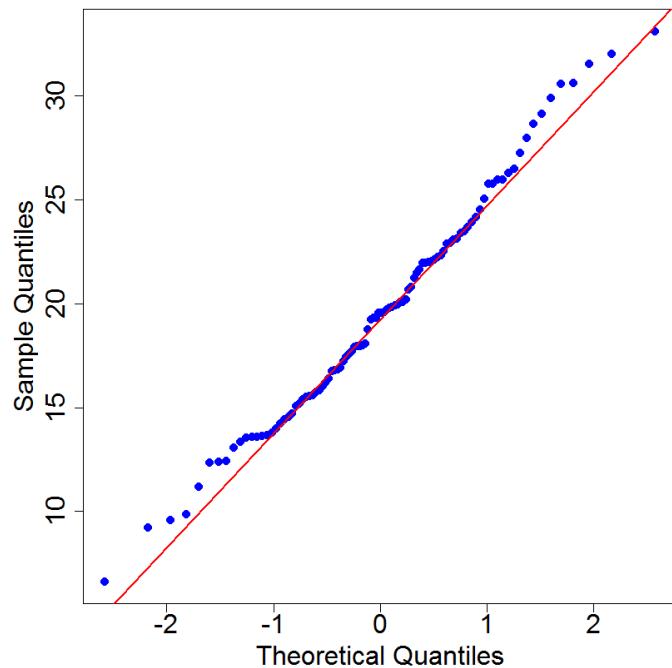


Рис. 5.10: Q-Q график

Очень удобный способ визуальной проверки предположения нормальности — Q-Q график (рисунок 5.10). Чтобы построить такой график, выборку нужно превратить в вариационный ряд, то есть отсортировать по неубыванию, а дальше каждому объекту выборки сопоставить точку на графике. Значение по вертикальной оси соответствует значению  $X$ , а значение по горизонтальной оси — математическому ожиданию квантиля стандартного нормального распределения, посчитанного по выборке такого объема.

Чтобы это лучше понять, можно посмотреть на точку в нижнем левом углу. Эта точка соответствует наименьшему значению в выборке. Пусть объем выборки — 100. Таким образом, эта точка — это минимум из всех 100 элементов. Значение этого минимума отложено по вертикальной оси. По горизонтальной оси отложено математическое ожидание минимума из 100 независимых одинаково распределенных случайных величин из стандартного нормального распределения. Если выборка взята из нормального распределения, точки на Q-Q графике должны лежать примерно на прямой. Если точки лучше описываются нелинейной кривой или какие-то из точек лежат от прямой очень далеко, скорее всего, распределение отличается от нормального.

### 5.4.3. Критерий Шапиро-Уилка

Критерий Шапиро-Уилка (таблица 5.7) — это ещё один способ формально проверить соответствие распределения выборки нормальному. Этот критерий основан на Q-Q графике. Фактически он проверяет, насколько сильно точки на Q-Q графике отклоняются от прямой.

выборка:	$X^n = (X_1, \dots, X_n)$ ;
нулевая гипотеза:	$H_0: X \sim N(\mu, \sigma^2)$ ;
альтернатива:	$H_1: H_0$ неверна;
статистика:	$W(X^n) = \frac{\left( \sum_{i=1}^n a_i X_{(i)} \right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$
нулевое распределение:	табличное.

Таблица 5.7: Описание критерия Шапиро-Уилка

Статистика  $W$  рассчитывается на основании вариационного ряда, полученного из выборки, и некоторых величин  $a$ . Эти величины основаны на математических ожиданиях порядковых статистик из стандартного нормального распределения, они табулированы, для них не существует аналитических выражений. Кроме того, табулировано и нулевое распределение статистики критерия Шапиро-Уилка, то есть его невозможно записать аналитически. Используя таблицы этих величин, можно вычислить достигаемый уровень значимости.

### 5.4.4. Зачем проверять нормальность?

Для проверки гипотезы нормальности существуют еще десятки других критериев: критерий Харке-Бера, Колмогорова, он же Лиллиефорса, Крамера-фон Мизеса, Андерсона-Дарлинга, и т. д. Чтобы понять, какие из этих критериев лучше использовать, необходимо вернуться на шаг назад и вспомнить, зачем нужно формально проверять нормальность.

Дело в том, что проверка гипотезы нормальности наследует плохие свойства всего аппарата проверки гипотез: на маленьких выборках нулевая гипотеза, как правило, не отклоняется, а на выборках огромного размера — практически наверняка отклоняется. То есть, если выборка маленькая, то, формально проверяя гипотезу о нормальности, её не получается отклонить, а если выборка огромна, то гипотеза отклоняется, даже если распределение отличается от нормального совсем чуть-чуть.

Многие методы, предполагающие нормальность, в том числе критерии Стьюдента, нечувствительны к небольшим отклонениям от нормальности, то есть истинное распределение выборки может слегка отличаться от нормального, и  $t$ -критерий будет всё еще правильно работать. Нормальное распределение — это математический конструкт. Никаких нормальных выборок в природе не существует. Однако, как говорил Джордж Бокс: «Все модели неверны, а некоторые полезны» — а нормальные модели очень полезны, поэтому их имеет смысл использовать.

### 5.4.5. Как проверять нормальность?

В итоге предлагается использовать следующий алгоритм. Если анализируемые данные имеют распределение, явно отличающееся от нормального (например, выборка бинарна или измеряемый признак — категориальный), не нужно применять метод, предполагающий нормальность. Лучше использовать метод, специально разработанный для такого распределения. Если исследуемый признак, по крайней мере, измерен в непрерывной шкале, можно построить Q-Q график. Если на этом графике не видно существенных отклонений от нормальности (точки лежат примерно на прямой), можно использовать методы, устойчивые к небольшим отклонениям от нормальности, например, критерии Стьюдента. Если используемый метод чувствителен к отклонениям от нормальности, необходимо формально проверить нормальность, и рекомендуется это делать с помощью метода Шапиро-Уилка. Показано, что критерий Шапиро-Уилка обладает достаточно хорошей мощностью для разных классов альтернатив. Если критерий Шапиро-Уилка отвергает нормальность, не нужно использовать методы, чувствительные к отклонениям от нормальности.

## 5.5. Гипотезы оолях

Ещё одно семейство параметрических критериев — это критерии, которые работают с распределениями Бернулли. Они принимают на вход выборки из нулей и единиц и проверяют гипотезы о параметрах  $p$  этих распределений (вероятность появления единицы в выборке). С распределением Бернулли работать удобно потому, что, в отличие от нормального распределения, не нужно применять никаких методов, чтобы доказать, что выборка взята именно из этого распределения. Если в выборке присутствуют только 2 значения, то она взята из распределения Бернулли.

Далее будут рассмотрены критерии, решающие три задачи: одновыборочную, двухвыборочную с независимыми выборками и двухвыборочную со связанными выборками.

### 5.5.1. Задача о присяжных

В 70-х годах известный педиатр и автор книг по воспитанию детей Бенджамин Спок был арестован за участие в антивоенной демонстрации в Бостоне. Его дело должен был рассматривать суд присяжных. Отбор присяжных — это сложная многоступенчатая процедура. На очередном этапе остаётся 300 человек, из которых отбираются финальные 12. В процессе Бенджамина Спока среди этих 300 только 90 были женщинами, и адвокаты подали протест. Поскольку в те времена воспитанием детей занимались в основном женщины, Бенджамин Спок среди них был более популярен, поэтому адвокаты заподозрили, что обвинение специально пытается сделать финальный состав присяжных менее благосклонным к подсудимому.

### 5.5.2. Z-критерий для доли

Чтобы по описанным выше данным проверить, был ли отбор беспристрастным, нужно использовать статистический критерий, например, Z-критерий для доли (таблица 5.8).

выборка:	$X^n = (X_1, \dots, X_n),$ $X \sim Ber(p);$
нулевая гипотеза:	$H_0: p = p_0;$
альтернатива:	$H_1: p < \neq p_0;$
статистика:	$Z(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \hat{p} = \bar{X}_n;$
нулевое распределение:	$Z(X^n) \sim N(0, 1).$

Таблица 5.8: Описание Z-критерия для доли

В задаче про отбор присяжных нулевая гипотеза состоит в том, что процедура отбора беспристрастна, женщины попадают в выборку с вероятностью 0.5; альтернатива — двусторонняя. Эта нулевая гипотеза отвергается: достигаемый уровень значимости  $p = 4.6 \times 10^{-12}$ . Точечная оценка вероятности попадания женщин в выборку составляет 0.3. 95% интервал для этой вероятности: [0.248, 0.352]. Выборка достаточно большая, поэтому неизвестную долю можно оценить с погрешностью порядка 10 %.

### 5.5.3. Рейтинг премьер-министра

1600 гражданам Великобритании с правом голоса задают вопрос: одобряют ли они деятельность премьер-министра. 944 человека говорят, что одобряют. Через 6 месяцев опрос повторяется. На этот раз из 1600 опрошенных 880 говорят, что поддерживают премьер-министра. Чтобы понять, изменился ли рейтинг премьер-министра, нужно использовать статистический критерий.

### 5.5.4. Z-критерий для доли для двух независимых выборок

Для решения предыдущей задачи можно использовать Z-критерий для двух долей (таблица 5.9).

Данные в подобных задачах можно записать при помощи таблицы сопряженности  $2 \times 2$  (таблица 5.10). В ней в столбцах расположены выборки, а в строках — исходы.

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim Ber(p_1);$
нулевая гипотеза:	$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim Ber(p_2),$
альтернатива:	$H_0: p_1 = p_2;$
статистика:	$Z(X_1^{n_1}, X_2^{n_2}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$
нулевое распределение:	$P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2};$ $Z(X_1^{n_1}, X_2^{n_2}) \sim N(0, 1).$

Таблица 5.9: Описание Z-критерия для доли для двух независимых выборок

Исход	Выборка	
	$X_1$	$X_2$
1	$a$	$b$
0	$c$	$d$
$\sum$	$n_1$	$n_2$

Таблица 5.10: Таблица сопряжённости

Если выборки независимы, то из четырёх чисел  $a, b, c, d$  в таблице сопряжённости Z-критерий использует только два, стоящие в первой строчке (количество единиц в первой и во второй выборках):

$$\hat{p}_1 = \frac{a}{n_1}, \quad \hat{p}_2 = \frac{b}{n_2}.$$

Результат	Опрос	
	I	II
+	$a = 944$	$b = 880$
-	$c = 656$	$d = 720$
$\sum$	$n_1 = 1600$	$n_2 = 1600$

Таблица 5.11: Таблица сопряжённости для задачи о рейтинге премьер-министра

В задаче оценки изменения рейтинга премьер-министра (таблица 5.11) нулевая гипотеза о том, что рейтинг не изменился против двусторонней альтернативы отвергается Z-критерием с достигаемым уровнем значимости  $p = 0.022$ . Рейтинг упал на 4 %, 95% доверительный интервал — [0.6, 7.4]%

### 5.5.5. Z-критерий для доли для двух связанных выборок

На самом деле, в двух рассматриваемых опросах участвовали одни и те же люди, то есть, выборки являются связанными, поскольку значения признаков измерены на одних и тех же объектах. Таблица  $2 \times 2$ , с помощью которой записываются данные, слегка меняет свой вид (таблица 5.12).

I	II		$\sum$
	+	-	
+	794	150	944
-	86	570	656
$\sum$	880	720	1600

Таблица 5.12: Таблица сопряженности для случая двух связанных выборок в задаче о рейтинге премьер-министра

Теперь в строках таблицы находятся результаты первого опроса, в столбцах — результаты второго, а в каждой ячейке — количество людей, которые в первом и втором опросе ответили именно так (например, 794 человека поддерживали премьер-министра в обоих опросах, 150 — только в первом и т.д.).

Чтобы использовать новые данные для проверки гипотезы о том, что рейтинг не изменился, нужно применить модифицированную версию Z-критерия для связанных выборок (таблица 5.13). Новые обозначения приведены в таблице 5.14.

выборки:	$X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim Ber(p_1);$ $X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim Ber(p_2),$ выборки связанные;
нулевая гипотеза:	$H_0: p_1 = p_2;$
альтернатива:	$H_1: p_1 < \neq > p_2;$
статистика:	$Z(X_1^n, X_2^n) = \frac{f-g}{\sqrt{f+g-\frac{(f-g)^2}{n}}};$
нулевое распределение:	$Z(X_1^n, X_2^n) \sim N(0, 1).$

Таблица 5.13: Описание Z-критерия для связанных выборок

$X_1 \backslash X_2$	1	0	$\Sigma$
1	$e$	$f$	$e + f$
0	$g$	$h$	$g + h$
$\Sigma$	$e + g$	$f + h$	$n$

Таблица 5.14: Таблица сопряженности для случая двух связанных выборок

В Z-критерии для связанных выборок используется статистика, в которую входят только недиагональные элементы  $f, g$  таблицы  $2 \times 2$ , то есть только те объекты, на которых значения двух признаков отличаются. Объекты  $e$  и  $h$ , на которых значения признаков совпадают, в критерии не используются.

В задаче о рейтинге премьер-министра Z-критерий для связанных выборок уверенно отвергает нулевую гипотезу о том, что рейтинг не изменился, против двусторонней альтернативы. Достигаемый уровень значимости  $p = 2.8 \times 10^{-5}$ , что существенно меньше, чем без учёта связаннысти выборок. Точечная оценка не меняется: рейтинг упал на 4%. А вот 95%-доверительный интервал для изменения уже другой: [2.1, 5.8]%. Этот интервал уже и его край дальше отстоит от 0, значения, соответствующего нулевой гипотезе.

## Урок 6

# Непараметрические критерии

### 6.1. Как работают непараметрические критерии?

В прошлом уроке шла речь о параметрических критериях — критериях, которые предполагают, что поступающая на вход выборка взята из какого-то распределения, и проверяют гипотезы о значениях параметров этого распределения. В этом уроке все будет иначе. Непараметрические критерии применяются в задачах следующего типа. Есть выборка объема  $n$  из какого-то распределения  $F(x)$ :

$$X^n = (X_1, \dots, X_n), X \sim F(x).$$

Проверяется гипотеза о равенстве нулю среднего значения случайной величины, из которой взята эта выборка.

Чтобы проверять любую гипотезу, нужна  $T$ -статистика, для которой должно быть известно нулевое распределение, то есть распределение при условии справедливости нулевой гипотезы. Если про исходное распределение  $F(x)$  что-то известно и  $T$ -статистика выбрана удачно, то нулевое распределение статистики может быть выражено аналитически. Однако распределение  $F(x)$  может быть нестандартным, и о нём может быть ничего не известно.

Гипотезы про среднее значение можно проверять с использованием центральной предельной теоремы (фактически, с помощью  $Z$ -критерия). Но центральная предельная теорема не всегда применима: иногда распределения бывают слишком скошенные, иногда выборка недостаточно большая, чтобы распределение ее выборочного среднего можно было считать нормальным.

В таких ситуациях существует два варианта действий. Во-первых, имеющуюся выборку из неизвестного распределения можно преобразовать так, что о её распределении будет больше информации. Во-вторых, можно сделать какие-то предположения о функции распределения исходной выборки  $F(x)$ , и на основании этих предположений построить статистику, нулевое распределение которой можно оценить.

В методах, которые будут рассмотрены далее в этом уроке, в разных комбинациях используются эти два способа работы с выборками.

### 6.2. Критерии знаков

Критерии знаков — это одно из семейств непараметрических критериев. Эти критерии обладают невысокой мощностью, но они крайне универсальны и практически ничего не требуют от данных, поэтому они очень полезны на практике.

### 6.2.1. Критерий знаков для одной выборки

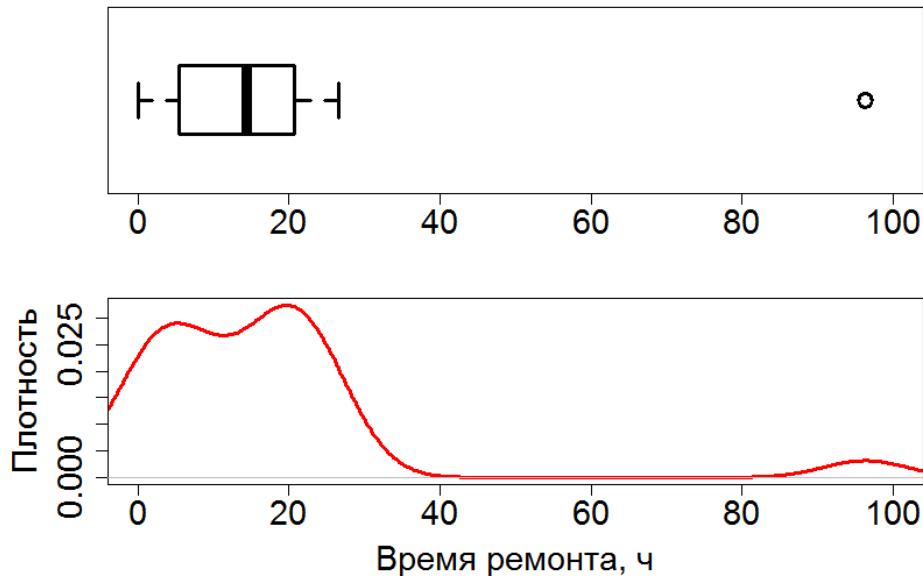


Рис. 6.1: Данные о времени ремонта интернет-оборудования клиентов провайдера Verizon

Для примера можно рассмотреть данные о времени ремонта интернет-оборудования клиентов провайдера Verizon. Выборка состоит из 23 наблюдений, и, как видно по графикам на рисунке 6.1, распределение признака не похоже на нормальное. Хочется понять, позволяют ли собранные данные утверждать, что среднее время ремонта составляет больше восьми часов. Использовать для этого параметрические критерии (например, критерий Стьюдента) не стоит, поскольку у распределения признака тяжелый правый хвост. Кроме того, объем выборки достаточно маленький, поэтому не получается воспользоваться центральной предельной теоремой.

Решить эту задачу можно с помощью критерия знаков.

выборка:	$X^n = (X_1, \dots, X_n), X_i \neq m_0;$
нулевая гипотеза:	$H_0: \text{med } X = m_0;$
альтернатива:	$H_1: \text{med } X < \neq > m_0;$
статистика:	$T(X^n) = \sum_{i=1}^n [X_i > m_0];$
нулевое распределение:	$T(X^n) \sim \text{Bin}(n, \frac{1}{2}).$

Таблица 6.1: Описание одновыборочного критерия знаков

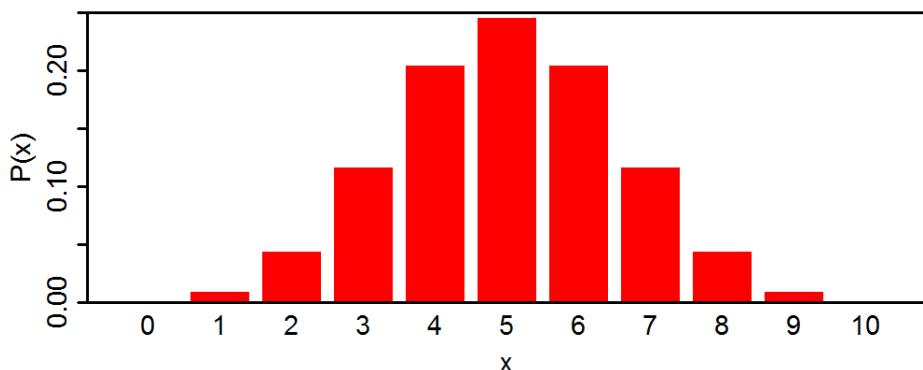


Рис. 6.2: Биномиальное распределение  $\text{Bin}(10, 0.5)$

Единственное требование при применении этого критерий — в выборке не должно быть ни одного объекта,

значение признака которого в точности совпадает с  $m_0$ . Если нулевая гипотеза справедлива, то статистика этого критерия имеет биномиальное распределение (рисунок 6.2).

Итак, в задаче о времени ремонта интернет-оборудования проверяется нулевая гипотеза о том, что медиана времени ремонта составляет 8 часов:

$$H_0: \text{med } X = 8.$$

Односторонняя альтернатива утверждает, что ремонт в среднем длится дольше 8 часов:

$$H_1: \text{med } X > 8.$$

В имеющейся выборке ремонт занял больше 8 часов в 15 случаях из 23. Критерий знаков утверждает, что это недостаточно много. Его достигаемый уровень значимости (вероятность получить 15 из 23 в условиях справедливости нулевой гипотезы)  $p = 0.105$ . Нулевая гипотеза не отвергается, данные не позволяют утверждать, что ремонт в среднем длится дольше 8 часов.

### 6.2.2. Цензурированная выборка

Критерий знаков настолько нетребователен к данным, что его можно использовать даже на цензурированных выборках.

**Пример.** Наблюдаются пациенты с лимфоцитарной лимфомой, измеряемый признак — это время их жизни в неделях после того, как был поставлен диагноз. Исследование длится семь лет. В выборке есть один пациент, который после семи лет (362 недель наблюдений) остался жив. Поскольку исследование закончилось, неизвестно, сколько еще он прожил после этого. Такая выборка называется цензурированной сверху, поскольку на части объектов известна только нижняя граница значения признака.

Если требуется проверить гипотезу о том, что среднее время дожития составляет 200 недель, против односторонней альтернативы, что оно больше 200 недель, для этой выборки можно без проблем использовать критерий знаков. Его достигаемый уровень значимости  $p = 0.9453$ . То есть нулевую гипотезу нельзя отклонить против односторонней альтернативы.

### 6.2.3. Критерий знаков для связанных выборок

	AUC <sub>C4.5</sub>	AUC <sub>C4.5+m</sub>
adult (sample)	0.763	<b>0.768</b>
breast cancer	<b>0.599</b>	0.591
breast cancer wisconsin	0.954	<b>0.971</b>
cmc	0.628	<b>0.661</b>
ionosphere	0.882	<b>0.888</b>
iris	<b>0.936</b>	0.931
liver disorders	0.661	<b>0.668</b>
lung cancer	0.583	0.583
lymphography	0.775	<b>0.838</b>
mushroom	1.000	1.000
primary tumor	0.940	<b>0.962</b>
rheum	0.619	<b>0.666</b>
voting	0.972	<b>0.981</b>
wine	0.957	<b>0.978</b>

Таблица 6.2: Данные о качестве классификаторов

В этом примере рассматривается классификатор C4.5 (один из способов построения деревьев решений). Для этого классификатора на 14 стандартных наборах данных посчитали площадь под ROC-кривой на тестовой выборке. Эти данные находятся в первом столбце таблицы 6.2. Далее этот классификатор модифицировали — изменили один из его гиперпараметров, минимальное количество объектов в листе  $t$ . Для нового классификатора на тех же 14 наборах данных посчитали площадь под ROC-кривой на тестовой выборке, результат — второй столбец таблицы 6.2.

Используя полученные данные, нужно определить, какая из этих двух версий классификатора лучше. Во-первых, можно заметить, что из 14 датасетов на 10 площадь под ROC-кривой больше у второй версии

классификатора. На двух наборах данных показывает лучший результат первая версия классификатора, и еще на двух — ничья.

Чтобы по этим цифрам посчитать статистическую значимость, можно использовать критерий знаков для связанных выборок (таблица 6.3).

выборки:	$X_1^n = (X_{11}, \dots, X_{1n}),$ $X_2^n = (X_{21}, \dots, X_{2n}),$ $X_{1i} \neq X_{2i}$ , выборки связанные;
нулевая гипотеза:	$H_0: \mathbf{P}(X_1 > X_2) = \frac{1}{2};$
альтернатива:	$H_1: \mathbf{P}(X_1 > X_2) < \neq > \frac{1}{2};$
статистика:	$T(X_1^n, X_2^n) = \sum_{i=1}^n [X_{1i} > X_{2i}];$
нулевое распределение:	$T(X_1^n, X_2^n) \sim \text{Bin}(n, \frac{1}{2}).$

Таблица 6.3: Описание критерия знаков для связанных выборок

В этом уроке идёт речь о непараметрических критериях, которые проверяют гипотезы о средних, но под «средними» они часто понимают совершенно разные вещи. Так, одновыборочный критерий знаков под средним понимает медиану. Двухвыборочный критерий знаков гипотезу о средних формулирует в представленном выше экзотическом виде. Другие критерии могут использовать другие варианты нулевых гипотез, но, тем не менее, всё это — в каком-то виде утверждение о средних.

Статистика двухвыборочного критерия знаков — это сумма индикаторов того, что элемент первой выборки больше, чем соответствующий элемент второй выборки:

$$T(X_1^n, X_2^n) = \sum_{i=1}^n [X_{1i} > X_{2i}].$$

Если нулевая гипотеза справедлива, эта статистика, так же, как и в случае одновыборочного критерия, имеет биномиальное распределение (рисунок 6.2) с параметрами  $n, \frac{1}{2}$ :

$$T(X_1^n, X_2^n) \sim \text{Bin}(n, \frac{1}{2}).$$

В задаче о качестве классификаторов требуется проверить нулевую гипотезу о том, что их среднее качество одинаково:

$$H_0: \mathbf{P}(\text{AUC}_{C4.5+m} > \text{AUC}_{C4.5}) = \frac{1}{2}.$$

Эта гипотеза проверяется против односторонней альтернативы о том, что качество модифицированного классификатора выше:

$$H_1: \mathbf{P}(\text{AUC}_{C4.5+m} > \text{AUC}_{C4.5}) > \frac{1}{2}$$

Странно предполагать, что при настройке какого-то гиперпараметра получится в среднем падение качества классификатора, поэтому используется именно односторонняя альтернатива. Критерий знаков дает достигаемый уровень значимости  $p = 0.019$ . На уровне значимости 0.05 отвергается нулевая гипотеза о том, что у этих классификаторов качество одинаковое, против альтернативы о том, что второй классификатор лучше. Модифицированный алгоритм лучше на 83% датасетов. 95% нижний доверительный предел для доли датасетов, на которых модифицированный классификатор лучше, — 56.2%.

### 6.3. Ранговые критерии

Для проверки гипотез о средних критерии знаков выбрасывают большую часть информации, содержащуюся в выборке. Вместо исходных значений признака используется бинарный вектор. Ранговые критерии позволяют сохранить большую информацию.

Выборку

$$X_1, \dots, X_n$$

всегда можно превратить в вариационный ряд, то есть упорядочить её по неубыванию:

$$X_{(1)} \leq \dots < \underbrace{X_{(k_1)} = \dots = X_{(k_2)}}_{\text{связка размёра } k_2 - k_1 + 1} < \dots \leq X_{(n)}.$$

Если при этом есть какие-то части вариационного ряда, в которых элементы полностью совпадают, эти части называются «связками».

Рангом наблюдения  $X_i$  называется его позиция в вариационном ряду. Если  $X_i$  не попадает в связку, то

$$\text{rank}(X_i) = r: X_i = X_{(r)},$$

а если  $X_i$  оказывается в связке  $X_{(k_1)}, \dots, X_{(k_2)}$ , то

$$\text{rank}(X_i) = \frac{k_1 + k_2}{2},$$

то есть в связке все объекты получают одинаковый средний ранг.

### 6.3.1. Критерий знаковых рангов Уилкоксона

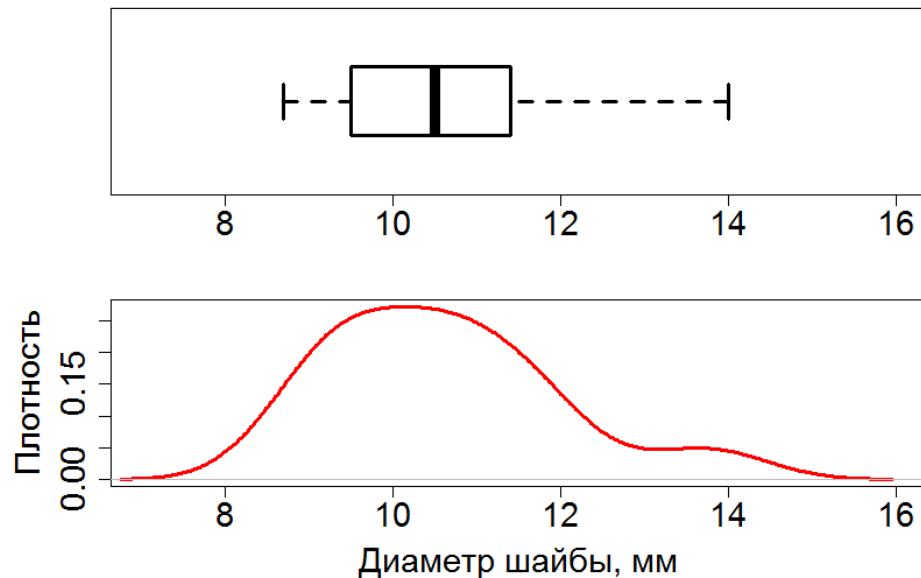


Рис. 6.3: Данные в задаче о размере шайбы

Использовать ранги можно для решения следующей задачи: имеются 24 шайбы, произведённые на одном и том же конвейере, для которых измерены диаметры. По этой выборке требуется понять, соответствует ли диаметр шайбы стандартному размеру в  $m_0 = 10$  мм. Для этого будет использоваться критерий знаковых рангов, или, как его иногда называют, критерий знаковых рангов Уилкоксона (таблица 6.4).

выборка:	$X^n = (X_1, \dots, X_n), X_i \neq m_0,$
	$F_X$ симметрично относительно медианы;
нулевая гипотеза:	$H_0: \text{med } X = m_0;$
альтернатива:	$H_1: \text{med } X < \neq m_0;$
статистика:	$W(X^n) = \sum_{i=1}^n \text{rank}( X_i - m_0 ) \cdot \text{sign}(X_i - m_0);$
нулевое распределение:	табличное.

Таблица 6.4: Описание критерия знаковых рангов Уилкоксона

1	2	3	4	5	$W$
-	-	-	-	-	-15
+	-	-	-	-	-13
-	+	-	-	-	-11
+	+	-	-	-	-9
-	-	+	-	-	-9
...	...	...	...	...	...
+	+	-	+	+	9
-	-	+	+	+	9
+	-	+	+	+	11
-	+	+	+	+	13
+	+	+	+	+	15

Таблица 6.5: Возможные реализации знаков рангов и соответствующие им значения статистики

При справедливости нулевой гипотезы каждый из рангов в выборке мог с одинаковой вероятностью реализоваться с любым знаком ( $\text{sign}(X_i - m_0)$ ): и с «+», и с «-». Таким образом, получается  $2^n$  вариантов распределения знаков по рангам. Перебирая все эти варианты, для каждого из них можно вычислить значение статистики, пример перебора показан в таблице 6.5. Именно так строится нулевое распределение критерия знаковых рангов.

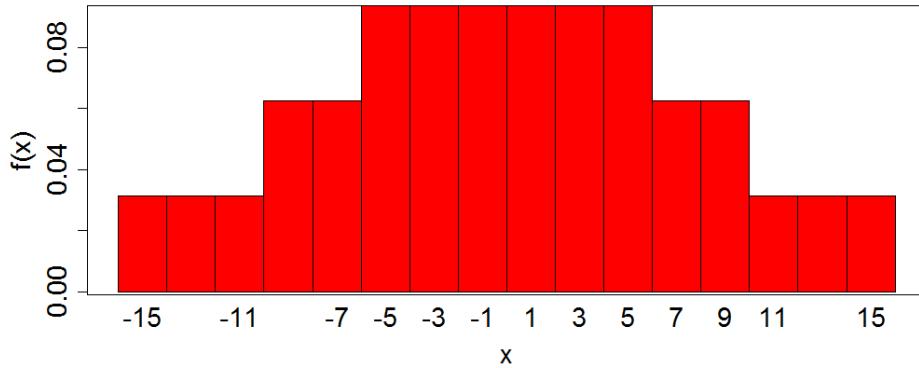


Рис. 6.4: Нулевое распределение при размере выборки  $n = 5$

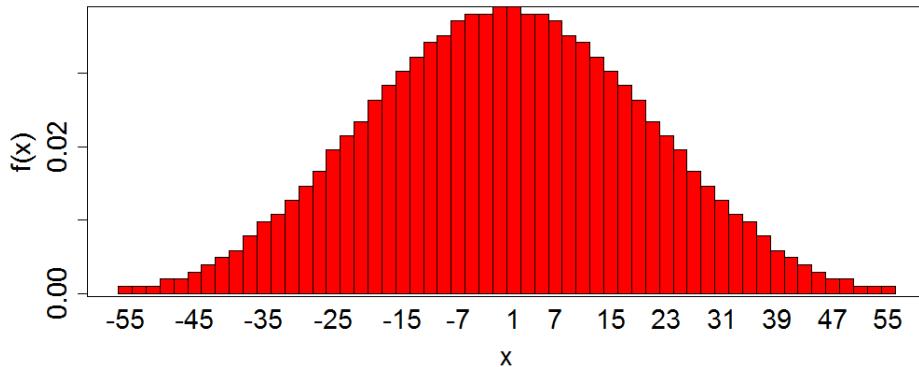


Рис. 6.5: Нулевое распределение при размере выборки  $n = 10$

Нулевое распределение статистики при размерах выборки  $n = 5, 10, 15$  показано на рисунках 6.4, 6.5, 6.6. Из них видно, что с ростом объёма выборки нулевое распределение становится похожим на нормальное. При размере выборки  $n > 20$  можно использовать следующую нормальную аппроксимацию:

$$W \approx \sim N \left( 0, \frac{n(n+1)(2n+1)}{6} \right).$$

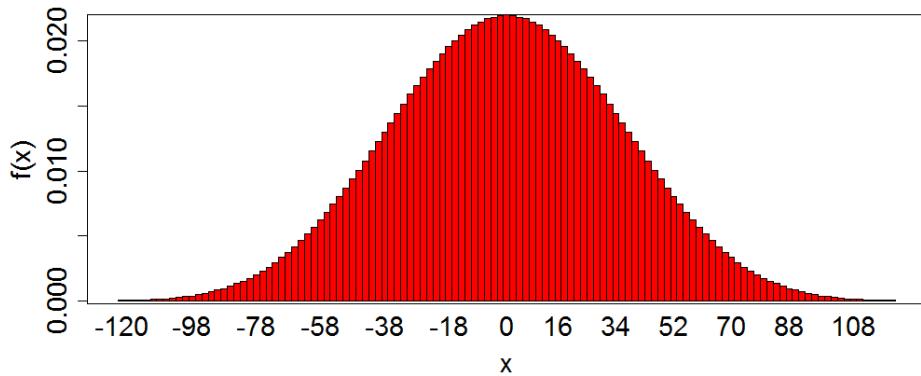


Рис. 6.6: Нулевое распределение при размере выборки  $n = 15$

В задаче о диаметре шайбы проверяется нулевая гипотеза о том, что средний размер шайбы составляет 10 миллиметров:

$$H_0: \text{med } X = 10$$

против двусторонней альтернативы:

$$H_1: \text{med } X \neq 10$$

Критерий знаковых рангов даёт достигаемый уровень значимости  $p = 0.0673$ , нулевая гипотеза не отвергается. Выборочная медиана диаметра составляет 10.5 мм, 95% доверительный интервал: [9.95, 11.15] мм. Доверительной интервал содержит целевое значение  $m_0 = 10$ . Так и должно быть, когда достигаемый уровень значимости выше порога.

### 6.3.2. Двухвыборочная задача со связанными выборками

Как и до этого в курсе, двухвыборочная задача со связанными выборками решается с использованием того же самого критерия, что и одновыборочная. Версия критерия знаков для двух связанных выборок показана в таблице 6.6.

выборки:	$X_1^n = (X_{11}, \dots, X_{1n}),$ $X_2^n = (X_{21}, \dots, X_{2n}),$ $X_{1i} \neq X_{2i}$ , выборки связанные;
нулевая гипотеза:	$H_0: \text{med}(X_1 - X_2) = 0;$
альтернатива:	$H_1: \text{med}(X_1 - X_2) < \neq > 0;$
статистика:	$W(X_1^n, X_2^n) = \sum_{i=1}^n \text{rank}( X_{1i} - X_{2i} ) \cdot \text{sign}(X_{1i} - X_{2i});$
нулевое распределение:	табличное.

Таблица 6.6: Описание критерия знаковых рангов для связанных выборок

На рисунке 6.7 показан график, отражающий данные о депрессивности 9 пациентов, измеренной по шкале Гамильтона до и после первого приёма транквилизатора. Хочется понять, действует ли транквилизатор, то есть снижается ли у этих пациентов депрессивность. Формально проверяется нулевая гипотеза о равенстве нулю медианы попарных разностей депрессивности до и после приёма транквилизаторов:

$$H_0: \text{med}(X_2 - X_1) = 0.$$

Альтернативная односторонняя гипотеза — депрессивность снизилась:

$$H_1: \text{med}(X_2 - X_1) < 0$$

Критерий знаковых рангов даёт достигаемый уровень значимости  $p = 0.019$ , то есть нулевая гипотеза отвергается в пользу односторонней альтернативы. Медиана снижения составляет 0.49 пунктов. 95% нижний доверительный предел для снижения: 0.175 пунктов.

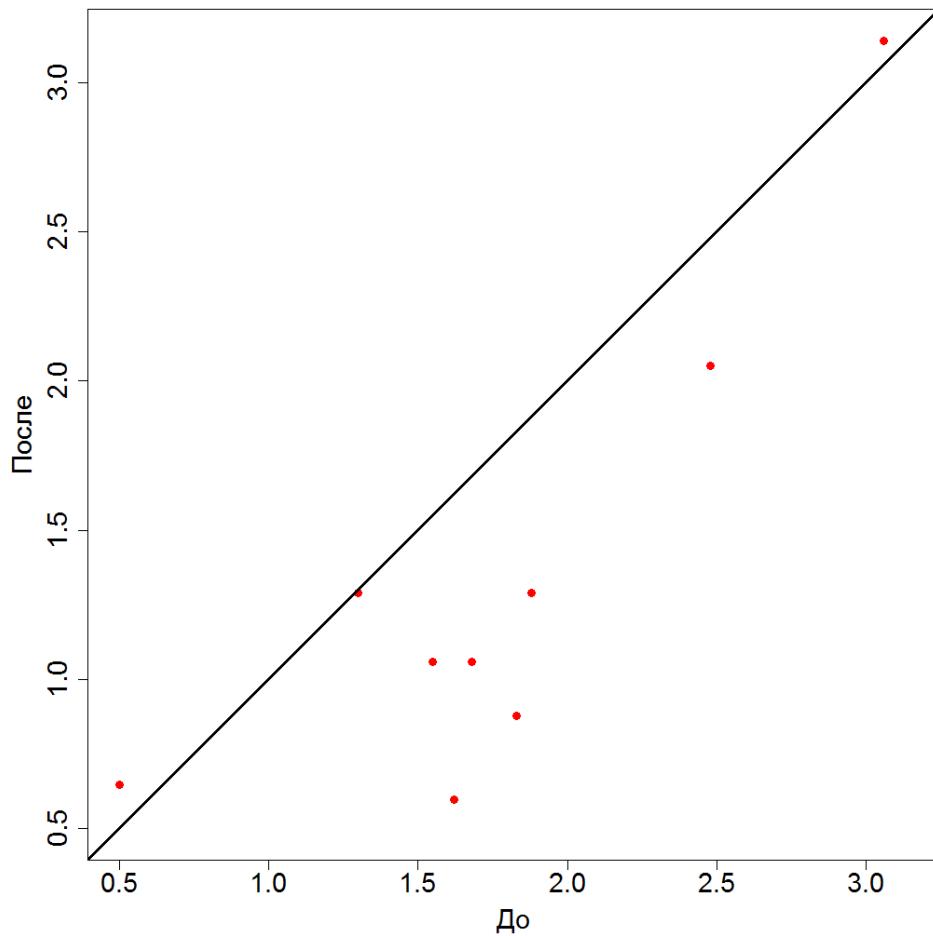


Рис. 6.7: Данные о депрессивности пациентов до и после приёма транквилизатора

### 6.3.3. Двухвыборочная задача с независимыми выборками

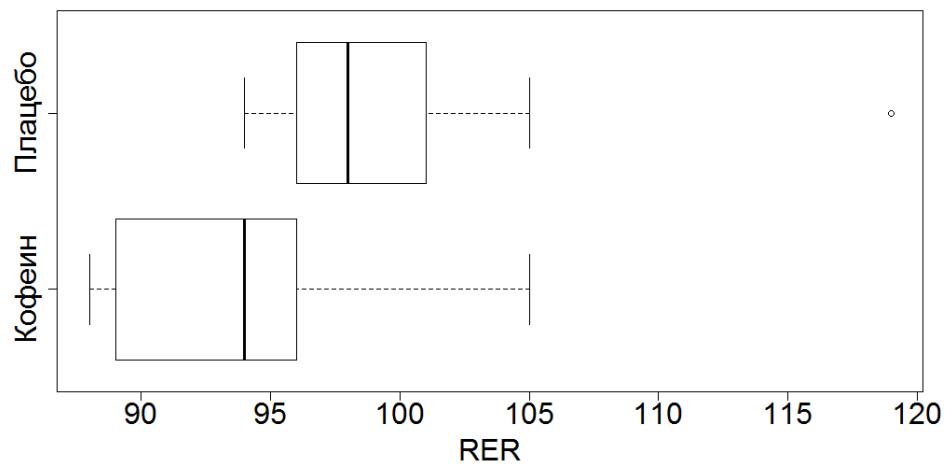


Рис. 6.8: Данные о респираторном обмене испытуемых после принятия кофеина или плацебо

В этой задаче измеряемый признак — это респираторный обмен, соотношение числа молекул углекислого газа и кислорода в выдыхаемом воздухе. Респираторный обмен является косвенным признаком того, из чего в данный момент мышцы вырабатывают энергию, из жиров или углеводов. В эксперименте измеряется респираторный обмен у 18 испытуемых в процессе физических упражнений (рисунок 6.8). За час до этого 9 из

них получили таблетку кофеина, а оставшиеся 9 — таблетку плацебо. Хочется понять, повлиял ли кофеин на среднее значение показателей респираторного обмена.

Эту задачу можно решить с помощью критерия Манна-Уитни, который иногда называют критерием Уилкоксона-Манна-Уитни (таблица 6.7).

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$
нулевая гипотеза:	$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}),$
альтернатива:	$H_0: F_{X_1}(x) = F_{X_2}(x);$
статистика:	$H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0;$
	$X_{(1)} \leq \dots \leq X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \cup X_2^{n_2}$ ,
	$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \text{rank}(X_{1i});$
нулевое распределение:	табличное.

Таблица 6.7: Описание критерия Манна-Уитни

Относительно параметра  $\Delta$  альтернатива в этом критерии может быть односторонней или двусторонней. Если справедлива альтернативная гипотеза и между распределениями действительно есть сдвиг, то средние значения признаков в выборках будут различаться. Поэтому это тоже в каком-то виде гипотеза о средних.

Для того чтобы построить статистику критерия Манни-Уитни, для объединенной выборки  $X = X_1^{n_1} \cup X_2^{n_2}$  строится вариационный ряд

$$X_{(1)} \leq \dots \leq X_{(n_1+n_2)},$$

и подсчитываются ранги

$$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \text{rank}(X_{1i}).$$

$X_1$	$X_2$	$R_1$
{1,2,3}	{4,5,6,7}	6
{1,2,4}	{3,5,6,7}	7
{1,2,5}	{3,4,6,7}	8
{1,2,6}	{3,4,5,7}	9
{1,2,7}	{3,4,5,6}	10
{1,3,4}	{2,5,6,7}	8
...	...	...
{3,5,7}	{1,2,4,6}	15
{3,6,7}	{1,2,4,5}	16
{4,5,6}	{1,2,3,7}	15
{4,5,7}	{1,2,3,6}	16
{4,6,7}	{1,2,3,5}	17
{5,6,7}	{1,2,3,4}	18

Таблица 6.8: Возможные распределения рангов между выборками

Статистикой будет сумма рангов элементов первой выборки в объединенном вариационном ряду. Нулевое распределение этой статистики, как и в предыдущем случае, табличное. Оно получается следующим образом. Если нулевая гипотеза справедлива, то каждый из рангов с одинаковой вероятностью мог реализоваться как в выборке  $X_1$ , так и в выборке  $X_2$ . Необходимо перебрать все возможные варианты того, как это могло произойти (таблица 6.8), всего таких вариантов  $C_{n_1+n_2}^{n_1}$ . На каждом из этих вариантов нужно вычислить значение статистики критерия Манни-Уитни, так и получается нулевое распределение.

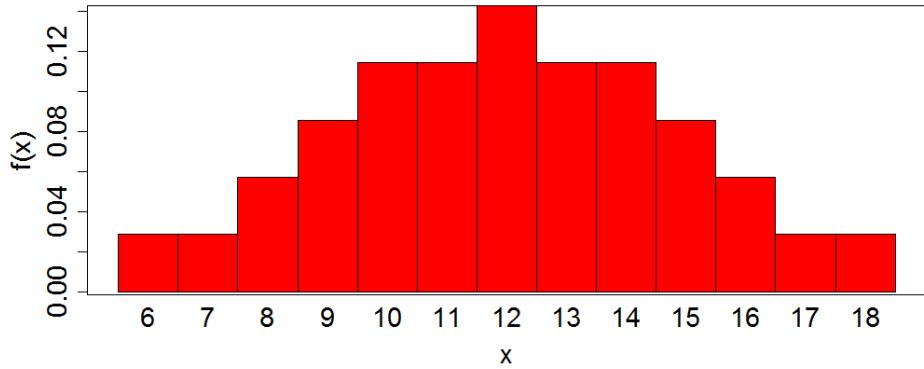


Рис. 6.9: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок  $n_1 = 3$ ,  $n_2 = 4$

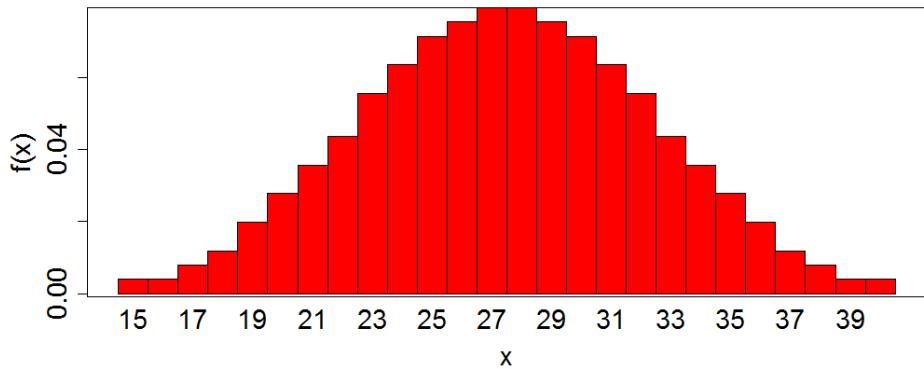


Рис. 6.10: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок  $n_1 = n_2 = 5$

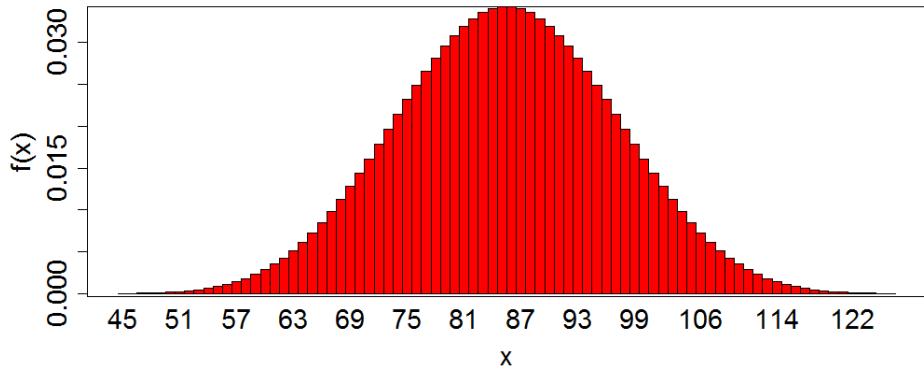


Рис. 6.11: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок  $n_1 = n_2 = 10$

Нулевые распределения статистики критерия Манна-Уитни для различных размеров выборок показаны на рисунках 6.9, 6.10, 6.11. Как и в предыдущем случае, при увеличении объёма выборок нулевое распределение стремится к нормальному. Для критерия Манни-Уитни также можно использовать нормальную аппроксимацию нулевого распределения, если в каждой из выборок есть по меньшей мере десять объектов:

$$R_1 \sim N \left( \frac{n_1(n_1 + n_2 + 1)}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \right).$$

В задаче о кофеине и респираторном обмене проверяется нулевая гипотеза  $H_0$ : среднее значение показателей в двух группах не отличаются, против двухсторонней альтернативы  $H_1$ : среднее значение показателя респираторного обмена отличается в двух группах. Критерий Манни-Уитни даёт достигаемый уровень значимости  $p = 0.0521$ , это совсем немного больше стандартного уровня значимости в 0.05. Сдвиг между

средними значениями показателей в двух выборках составляет 6 пунктов, 95% доверительный интервал —  $[-0.00005, 12]$  пт, 0 в него всё-таки попадает, отвергнуть нулевую гипотезу нельзя.

## 6.4. Перестановочные критерии

При использовании ранговых критериев выборки превращают в ранги, затем делается какое-то дополнительное предположение, и на основании этого предположения получается, что разные конфигурации этих рангов при справедливости нулевой гипотезы могут реализоваться с равной вероятностью. Далее необходимо перебрать все конфигурации, и на каждой посчитать значение статистики — таким образом оценивается ее нулевое распределение.

Если в этом алгоритме пропустить первый пункт (не превращать наблюдения в ранги), а остальное делать точно так же, то получится алгоритм работы перестановочных критериев.

### 6.4.1. Одновыборочный перестановочный критерий

выборка:	$X_1^n = (X_1, \dots, X_n)$ ,
нулевая гипотеза:	$F(X)$ симметрично относительно матожидания;
альтернатива:	$H_0: \mathbb{E}X = m_0$ ;
статистика:	$H_1: \mathbb{E}X <\neq> m_0$ ;
нулевое распределение:	$T(X^n) = \sum_{i=1}^n (X_i - m_0)$ , порождается перебором $2^n$ знаков перед слагаемыми $X_i - m_0$ .

Таблица 6.9: Описание одновыборочного перестановочного критерия

Имеется выборка размера  $n$ : и делается предположение, что функция распределения  $F(x)$  симметрична относительно математического ожидания. Одновыборочный перестановочный критерий (таблица 6.9) проверяет нулевую гипотезу о значении математического ожидания случайной величины, из которой взята выборка.

Если нулевая гипотеза этого критерия справедлива, каждый из объектов выборки мог с одинаковой вероятностью реализоваться слева и справа от математического ожидания. Поэтому нужно перебрать все  $2^n$  знаков, которые могут стоять в выражении для статистики перед разностью  $x_i - m_0$ . На основании этого перебора и будет восстановлено нулевое распределение статистики.

В качестве примера использования одновыборочного перестановочного можно вспомнить задачу анализа диаметра шайб (рисунок 6.3): по выборке из 24 элементов требуется понять, соответствует ли средний диаметр шайбы стандарту — 10 миллиметров:

$$H_0: \mathbb{E}X = 10.$$

Эта нулевая гипотеза проверяется против двусторонней альтернативы о том, что средний диаметр шайбы не соответствует стандарту:

$$H_1: \mathbb{E}X \neq 10.$$

Критерий знаковых рангов в этом случае давал достигаемый уровень значимости  $p = 0.0673$ , нулевое распределение показано на рисунке 6.12.

Нулевое распределение, полученное при использовании перестановочного критерия, показано на рисунке 6.13. Значение статистики, реализовавшейся в эксперименте:  $T = 14.6$ .

Чтобы вычислить достигаемый уровень значимости, нужно просуммировать высоты всех столбцов в распределении статистики, начиная от значения 14.6 и больше, а также от  $-14.6$  и меньше (поскольку альтернатива двухсторонняя). В результате получается достигаемый уровень значимости  $p = 0.1026$ , то есть нулевая гипотеза не отвергается.

Фактически достигаемый уровень значимости перестановочного критерия — это доля перебираемых перестановок, на которых получается такое же или еще более экстремальное значение статистики.

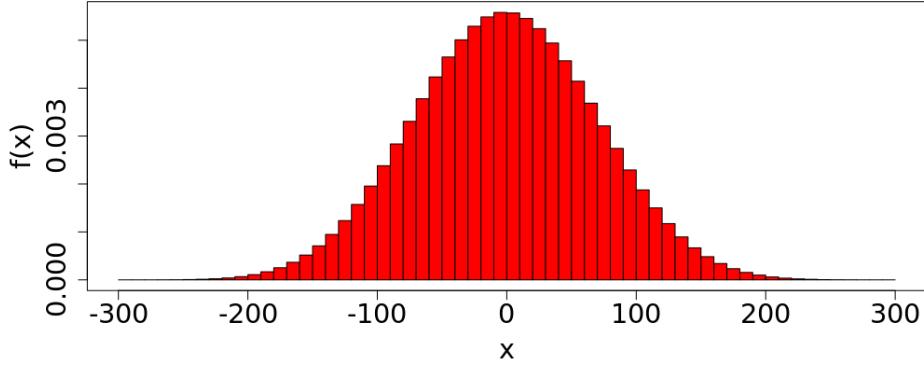


Рис. 6.12: Распределение тестовой статистики в решении задачи о размере шайб с использованием критерия знаковых рангов

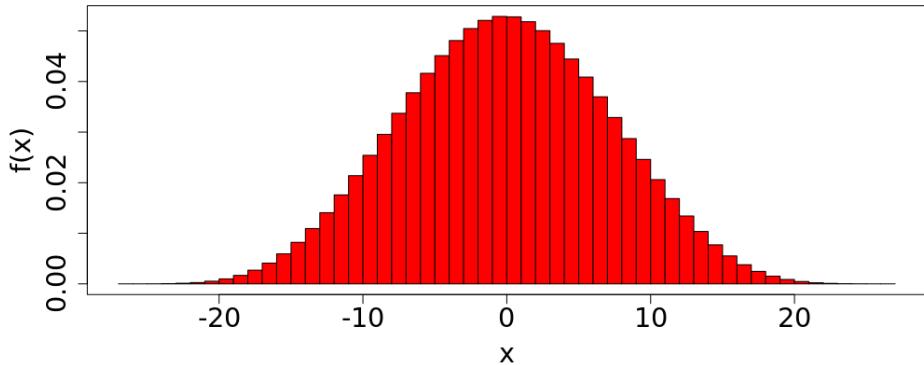


Рис. 6.13: Распределение тестовой статистики в решении задачи о размере шайб с использованием перестановочного критерия

#### 6.4.2. Двухвыборочный критерий для связанных выборок

Двухвыборочная задача со связанными выборками решается с использованием абсолютно такого же критерия: от двух связанных выборок происходит переход к одной выборке соответствующих попарных разностей (таблица 6.10).

выборки:	$X_1^n = (X_{11}, \dots, X_{1n}),$ $X_2^n = (X_{21}, \dots, X_{2n}),$ выборки связанные;
нулевая гипотеза:	$H_0: \mathbb{E}(X_1 - X_2) = 0;$
альтернатива:	$H_1: \mathbb{E}(X_1 - X_2) < \neq > 0;$
статистика:	$D^n = (X_{1i} - X_{2i}),$ $T(X_1^n, X_2^n) = T(D^n) = \sum_{i=1}^n D_i,$ порождается перебором $2^n$ знаков перед слагаемыми $D_i$ .
нулевое распределение:	

Таблица 6.10: Описание двухвыборочного перестановочного критерия для связанных выборок

В качестве примера можно вспомнить задачу об эффективности транквилизатора. У девяти пациентов, до и после приема транквилизатора, была измерена депрессивность по шкале Гамильтона (рисунок 6.7). Требуется проверить нулевую гипотезу о том, что депрессивность не изменилась:

$$H_0: \mathbb{E}(X_1 - X_2) = 0,$$

против односторонней альтернативы о том, что транквилизатор подействовал, то есть депрессивность снизи-

лась:

$$H_1: \mathbb{E}(X_1 - X_2) > 0$$

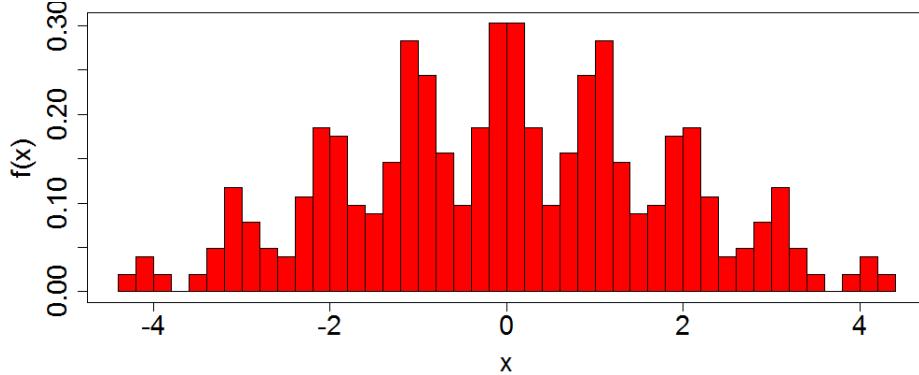


Рис. 6.14: Распределение порядковой статистики при использовании перестановочного критерия в задаче об эффективности транквилизатора

Критерий знаковых рангов давал достигаемый уровень значимости  $p = 0.019$ .

Нулевое распределение перестановочного критерия изображено на рисунке 6.14. Значение статистики, которое реализуется в эксперименте:  $T = 3.887$ . При суммировании высоты всех столбиков, начиная от 3.887 и направо, получается достигаемый уровень значимости  $p = 0.0137$ . Нулевая гипотеза отвергается в пользу односторонней альтернативы.

#### 6.4.3. Перестановочный критерий для независимых выборок

Перестановочный критерий для независимых выборок выглядит абсолютно так же, как критерий Манна-Уитни за исключением того, что не производятся ранговые преобразования.

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$ , $X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$ ,
нулевая гипотеза:	$H_0: F_{X_1}(x) = F_{X_2}(x)$ ;
альтернатива:	$H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta < \neq 0$ ;
статистика:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$ ;
нулевое распределение:	порождается перебором $C_{n_1+n_2}^{n_1}$ размещений объединённой выборки.

Таблица 6.11: Описание двухвыборочного перестановочного критерия для несвязанных выборок

Нулевое распределение статистики этого критерия точно так же, как и для критерия Манна-Уитни, получается перебором всех  $C_{n_1+n_2}^{n_1}$  размещений объединенной выборки по выборкам  $X_1^{n_1}$  и  $X_2^{n_2}$ .

В задаче об анализе связи между кофеином и респираторным обменом (рисунок 6.8) проверялась нулевая гипотеза  $H_0$ : среднее значение показателей респираторного обмена не отличается в двух группах пациентов (в одной пациенты принимали кофеин, в другой — плацебо) — против двусторонней альтернативы  $H_1$ : что-то изменилось.

Критерий Манна-Уитни давал достигаемый уровень значимости  $p = 0.0521$ . На рисунке 6.15 показано нулевое распределение перестановочного критерия. Значение статистики, которое реализуется в эксперименте:  $T = 6.33$ , оно соответствует достигаемому уровню значимости  $p = 0.0578$ . Нулевая гипотеза все еще не отвергается.

#### 6.4.4. Особенности перестановочных критериев

У перестановочных критериев есть некоторые особенности, о которых очень важно помнить.

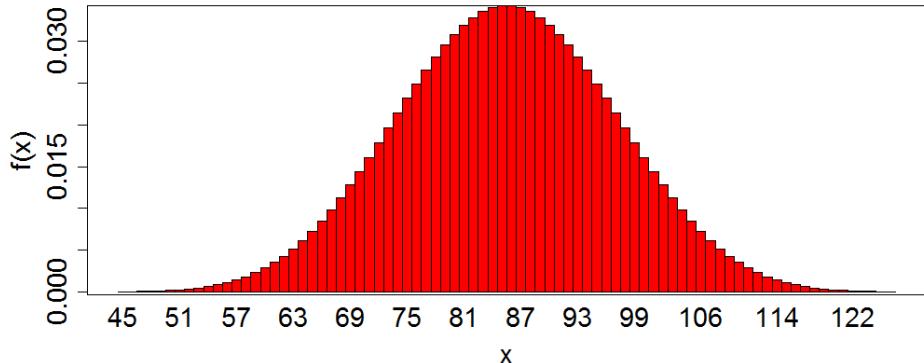


Рис. 6.15: Распределение нулевой статистики перестановочного критерия, полученной из данных эксперимента о связи между кофеином и респираторным обменом

Статистику для перестановочных критериев можно выбирать по-разному. В некоторых случаях это приводит к одному и тому же достигаемому уровню значимости, то есть ни на что не влияет. Например, в одновыборочной задаче, проверяя гипотезу о равенстве нулю математического ожидания

$$H_0: \mathbb{E}X = 0, \quad H_1: \mathbb{E}X \neq 0,$$

в качестве статистики перестановочного критерия можно использовать как сумму элементов выборки, так и выборочное среднее:

$$T_1(X^n) = \sum_{i=1}^n X_i \sim T_2(X^n) = \bar{X}.$$

Нулевые распределения этих статистик будут отличаться только сдвигом и масштабом, поэтому достигаемый уровень значимости, посчитанный по ним, будет одним и тем же.

В других случаях, по-разному выбирая статистику для перестановочного критерия, можно получать разные достижимые уровни значимости. Например, для статистик

$$T_2(X^n) = \bar{X} \approx T_3(X^n) = \frac{\bar{X}}{S/\sqrt{n}}$$

нулевые распределения отличаются не только сдвигом и масштабом, поэтому достигаемый уровень значимости у критериев с этими двумя вариантами статистик тоже будет разный. Поэтому при выборе статистики для перестановочного критерия важно думать о том, какие из свойств исходной случайной величины наиболее важны.

Перестановочные критерии придумал Рональд Фишер еще в начале XX века, однако их начали активно использовать только с появлением и широким распространением компьютеров, потому что для вычисления нулевых распределений этих критериев используются перестановки. В отличие от ранговых критериев, нормальных аппроксимаций для нулевого распределения в случае больших выборок не существует, поэтому единственный способ оценить нулевое распределение статистики — это перебрать много перестановок. Поэтому точно посчитать достигаемый уровень значимости перестановочного критерия на больших выборках достаточно сложно. Однако его можно посчитать приближенно. Для этого нужно взять какое-то случайное подмножество  $G'$  множества всех возможных перестановок  $G$ . При этом стандартное отклонение достигаемого уровня значимости будет примерно равно  $\sqrt{\frac{p(1-p)}{|G'|}}$ . На практике, чтобы получить хорошую аппроксимацию достигаемого уровня значимости, достаточно взять несколько тысяч перестановок.

## 6.5. Перестановки и бутстреп

Доверительные интервалы для параметров тесно связаны с проверкой точечных гипотез об их значениях. Например, z-критерии для средних связаны с нормальными доверительными интервалами, критерии Стьюдента соответствуют доверительным интервалам, построенным с использованием распределения Стьюдента. Для перестановочных критериев ближайшим аналогом в мире доверительных интервалов является метод бутстрепа, однако отношения между ними не такие взаимнооднозначные.

### 6.5.1. Проверка гипотез с помощью перестановочных критериев и метода бутстрепа

Перестановочные критерии принимают на вход выборку (или выборки), считают на них какую-то статистику. Далее делается дополнительное предположение о распределении, из которого эти выборки взяты. Это предположение порождает множество перестановок исходных данных, которые могли реализоваться с одинаковой вероятностью, если нулевая гипотеза справедлива. На этих перестановках вычисляется значение статистики и таким образом оценивается ее нулевое распределение.

Бутстреп-методы работают в каком-то смысле похоже. На вход они также принимают выборку или выборки и считают значение статистики, которая оценивает интересующий параметр. Далее на основании исходных данных генерируется множество бутстреп-псевдовыборок, и на этих псевдовыборках вычисляются значения интересующей статистики, то есть оценивается ее распределение.

Ключевых различий между этими методами несколько. Во-первых, перестановочный критерий использует дополнительное предположение, которое позволяет породить множество перестановок, которые используются для построения распределения. Во-вторых, перестановки, используемые в перестановочном критерии, — это выборки без возвращения, в то время как бутстреп-методы используют выборки с возвращением (бутстреп-псевдовыборки могут содержать по несколько копий элементов исходной выборки). Кроме того, распределения, которые получаются при использовании этих методов, абсолютно разные, потому что распределение статистики перестановочного критерия — это то распределение, которое статистика будет иметь при справедливости нулевой гипотезы, в то время как распределение бутстреп-статистики не подразумевает никакой нулевой гипотезы.

Чтобы лучше это понять, можно вспомнить пример с кофеином и респираторным обменом. В этой задаче проверялась гипотеза  $H_0$ : среднее значение показателя респираторного обмена не отличается в двух группах. Эта нулевая гипотеза проверялась против односторонней альтернативы  $H_1$ : под воздействием кофеина среднее значение показателя респираторного обмена снижается.

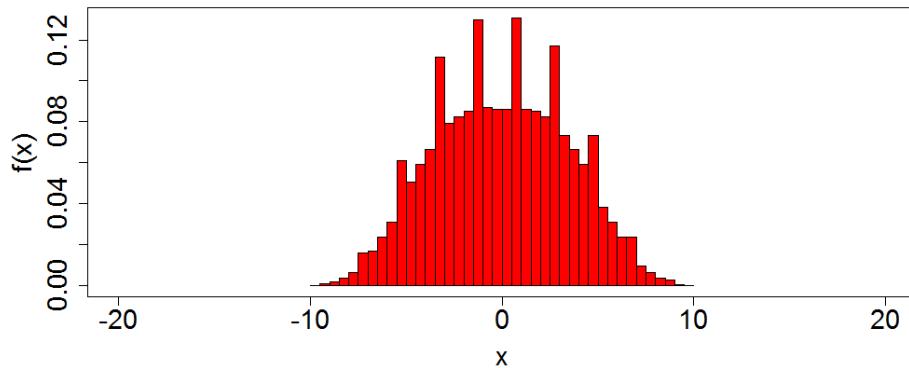


Рис. 6.16: Распределение нулевой статистики перестановочного критерия в задаче о респираторном обмене

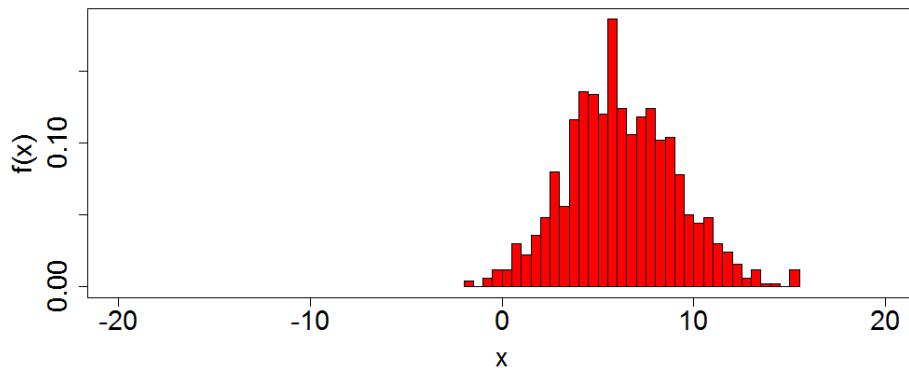


Рис. 6.17: Данные о респираторном обмене исследуемых после принятия кофеина или плацебо

На анализируемых данных разность выборочных средних значений показателей респираторного обмена у

пациентов, которые принимали плацебо и которые принимали кофеин, составляет  $\bar{X}_{1n} - \bar{X}_{2n} = 6.33$ .

На рисунке 6.16 показано нулевое распределение перестановочного критерия со статистикой  $\bar{X}_{1n} - \bar{X}_{2n}$ . На рисунке 6.17 — бутстреп-распределение той же самой статистики  $\bar{X}_{1n} - \bar{X}_{2n}$ . Ключевое различие между этими двумя распределениями в том, что они центрированы в разных местах. Перестановочное нулевое распределение центрировано в нуле — значении, соответствующем нулевой гипотезе. Бутстреп-распределение, в свою очередь, центрировано в выборочном среднем значений параметра. Параметр в данном случае — это разность средних, то есть центр бутстреп-распределения — это 6.33.

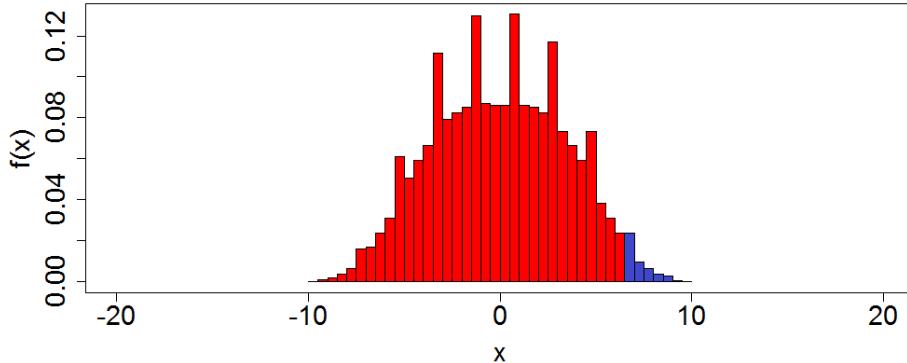


Рис. 6.18: Распределение статистики перестановочного критерия, синим показана доля перестановок, на которых среднее больше либо равно 6.33

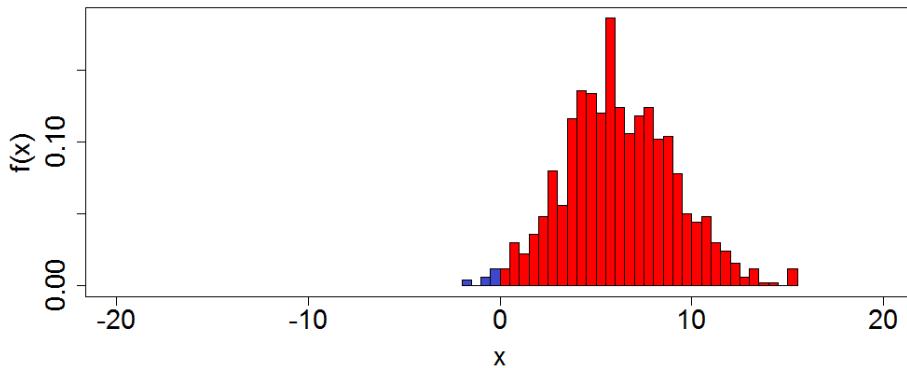


Рис. 6.19: Бутстреп-распределение статистики, синим показана доля псевдовыборок, на которых среднее меньше либо равно 0

Для перестановочного критерия доля перестановок, на которых среднее больше либо равно 6.33 (выборочное среднее, которое реализовано в данных) составляет примерно 0.03 (рисунок 6.18). Это и есть достигаемый уровень значимости перестановочного критерия, и он точный.

На бутстреп-распределении той же самой статистики доля псевдовыборок, на которых среднее меньше либо равно нулю, составляет 0.011 (рисунок 6.19). Эту величину можно считать приближенным достижимым уровнем значимости бутстреп-критерия. То есть с помощью доверительных интервалов на основе бутстрепа тоже можно проверять гипотезы, однако нужно это делать немного иначе.

### 6.5.2. Различия перестановочного критерия и бутстрепа

Перестановочный критерий с помощью нулевого распределения статистики измеряет расстояние от 0 до  $\bar{D}_n$  — значения параметра, реализовавшегося в эксперименте. Бутстреп-критерий измеряет, наоборот, расстояние от  $\bar{D}_n$  до 0.

Перестановочный критерий является более точным, потому что в нем перебирается подмножество всех возможных перестановок данных, которые равновероятны при справедливости нулевой гипотезы. Бутстреп-критерий в описанном выше виде является только приближенным, поскольку в нем всегда перебирается

только конечное подмножество всех возможных бутстреп-псевдовыборок, потому что их заведомо слишком много.

Самое важное различие между этими двумя критериями заключается в том, что они проверяют разные гипотезы, поскольку перестановочный критерий использует дополнительные предположения. Перестановочный критерий проверяет гипотезу полного равенства распределений в двух выборках

$$H_0: F_{X_1}(x) = F_{X_2}(x),$$

и проверяется она против альтернативы сдвига (в этом примере односторонней):

$$H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta > 0$$

Предполагается, что никак иначе, кроме как сдвигом, эти распределения отличаться не могут.

Бутстреп-критерий проверяет всего лишь гипотезу о равенстве математических ожиданий:

$$H_0: \mathbb{E}X_1 = \mathbb{E}X_2.$$

Гипотезу равенства он проверяет против односторонней альтернативы точно так же, как и перестановочный критерий:

$$H_1: \mathbb{E}X_1 > \mathbb{E}X_2.$$

Однако эта гипотеза заведомо более общая: не используется предположение равенства функций распределения в двух выборках.

### 6.5.3. Резюме

Проверка гипотез с помощью бутстрепа — это достаточно сложная задача. Показанный небольшой двухвыборочный пример позволяет понять плюсы и минусы бутстрепа и перестановочных критериев.

Бутстреп тоже позволяет проверять гипотезы, причём гораздо более широкого класса. Поскольку этот метод не использует дополнительных предположений, можно проверять только интересующий параметр (в примере это была разность математических ожиданий). Кроме того, с помощью бутстрепа можно проверять и другие крайне экзотические гипотезы, которые никакими другими методами проверить нельзя. Например, гипотеза о том, что распределение имеет ровно две моды.

Если предположения, лежащие в основе перестановочного критерия, выполняются, то перестановочный критерий, во-первых, точнее, поскольку его достижимый уровень значимости точный, во-вторых, всегда мощнее, чем аналогичный критерий бутстрепа. Но бутстреп при этом гораздо более гибок, потому что его можно использовать в ситуациях, когда не выполняются предположения, использующиеся в перестановочном критерии.

# Урок 7

## Поиск зависимостей в данных

### 7.1. Взаимное влияние в продажах товаров

Третья неделя курса посвящена методам выявления взаимосвязей и поиска закономерностей. Это вещи, которые статистика умеет делать лучше всего.

Одна из наиболее ярких областей, в которых постоянно требуется решать задачи с помощью статистики, — это ритейл, или розничная продажа. Такие задачи сопровождают продажи практически на всех этапах, начиная от планирования и аренды складских помещений, заканчивая разработкой гибких скидочных политик для постоянных клиентов. Например, при выборе места для склада важно, чтобы, с одной стороны, оно стоило дешево, а с другой — находилось где-то в центре и до него было удобно добираться. Помимо локации еще нужно выбрать площадь склада, а оценить ее тоже довольно сложно. Если площадь склада будет слишком маленькой, то какую-то часть времени товары может быть негде хранить, и придется арендовать дополнительную площадь, что влечёт за собой дополнительные расходы. С другой стороны, если сразу арендовать большие складские помещения, то в течение какого-то времени они будут пустовать. Это тоже дополнительные расходы, но уже из-за простоя.

Выше приведён пример небольшой задачи, которая в ритейле, фактически, решается еще до старта бизнеса. Такие задачи сопровождают бизнес на протяжении всего жизненного цикла и их количество не позволяет решать их все в голове. Получается, что для того, чтобы успешно вести бизнес в области ритейла, необходимо грамотное планирование. Заблаговременно требуется получить оценку того, какой будет спрос на различные товары, сколько места требуется на их хранение, а также сколько места будет свободно с учетом уже хранимых товаров. Это означает, что нужно уметь строить модель, способную делать прогнозы такого вида. Однако для этого недостаточно одних моделей, способных связать площадь склада и объем продаж. Помимо этого необходимо опираться на существующую статистику работы предприятия, например, на статистику по продажам. Получается, для грамотного планирования необходимо строить модели, сочетающие методы прогнозирования и методы статистики.

Также в этих задачах так много параметров, что очень легко построить модель, которая не будет интерпретируемой. Интерпретируемость — это очень важный фактор, потому что люди, принимающие решения в бизнесе, не готовы доверять моделям-черным ящикам. Поэтому необходимо, чтобы в моделях было как можно больше здравого смысла. Требуется уметь не только вносить здравый смысл в модели, но и извлекать его из данных. Именно таким методам посвящена третья неделя курса.

### 7.2. Внешние факторы, влияющие на продажи

Анализировать данные на предмет выявления различных закономерностей не только очень интересно, но и позволяет принести практическую пользу бизнесу. Так, проанализировав продажи магазина, можно сделать вывод о том, какие товары покупают наиболее часто, а эта информация предоставляет целый ряд практических преимуществ. Её можно использовать для планирования продаж, закупок, рекламных кампаний, а также для оптимизации размещения товаров в торговом зале или, в случае онлайн-магазинов, для правильного построения блока «популярные товары».

Построение промо-блоков — это очень актуальная задача. Так, проанализировав чеки оффлайн-магазина или содержимое корзин онлайн-магазина, можно сделать вывод о том, какие товары чаще всего покупают вместе. Это, в свою очередь, позволяет строить так называемые товарные рекомендации, рекомендации вида

«С товаром А наиболее часто покупают товар В». По данным различных онлайн-магазинов, такие рекомендации сильно способствуют продажам. Они позволяют принести 15–30% дополнительных продаж, а это очень много.

Имея достаточное количество информации о товарах, можно делать очень много вещей: анализировать взаимосвязи между ними (например, выявлять аксессуары, аналоги или комплементарные товары), анализировать успешность товаров или брендов у той или иной целевой аудитории, оценивать долю продукта на рынке, смотреть, как она меняется во времени, или, например, оценивать эффект каннибализации (каннибализация — это известный маркетинговый эффект, заключающийся в том, что один из товаров отбирает целевую аудиторию у похожего товара или бренда). Такие экономические явления, как, например, каннибализация, существенно сказываются на продажах. Однако важно понимать, что нужно учитывать и такие, казалось бы, незначительные вещи, как расположение товара в магазине, его близость к кассе, высота полки, на которой он находится, хорошо ли товар видно, могут ли до него дотянуться дети. Всё это тоже очень сильно сказывается на продажах. Более того, важно учитывать контекст, в котором товар существует, например, его ближайшее окружение: какие товары и бренды находятся на полке рядом с ним.

Структура продаж магазина определяется в первую очередь людьми, которые в него ходят. Часто оказывается, что в городах и в деревнях покупают совершенно разные вещи, причем иногда эти зависимости совершенно не такие, как можно было ожидать. Например, в одном из проектов были проанализированы продажи в сети доступных супермаркетов, и выяснилось, что в деревнях люди чаще покупают дорогой алкоголь.

Факторы, связанные со временем, — это, пожалуй, самые важные факторы для прогнозирования продаж. Многие товары по-разному продаются в разное время. Например, шампанское очень хорошо продаётся в Новый год, мороженое лучше продаётся летом и т.д. Кроме того, есть всякие длинные, медленно меняющиеся во времени тренды, связанные, например, с тем, что какой-то товар постепенно с рынка выходит, а какой-то товар только на нём появляется.

Также важно учитывать экономическую ситуацию регионов продаж. Так, колебания в курсах валют, изменение налогового или таможенного законодательства способны существенно повлиять на продажи. Например, в моменты нестабильной экономической ситуации товары и бренды из более низкого ценового сегмента начинают отъедать долю рынка у товаров из более высоких ценовых сегментов. Однако и обратное тоже верно: в период стабильности экономики и роста доходов населения товары из премиум и высоких классов начинают возвращать свою долю рынка обратно.

Выше перечислено большое количество факторов, которые могут быть полезны при прогнозировании продаж. На самом деле, таких факторов может быть еще больше, и методы, которые будут обсуждаться на этой неделе, способны эти факторы из данных выявлять.

Задачи ритейла очень интересные и большую часть времени речь будет идти именно о них. Однако выявление закономерностей из данных активно применяется и в других областях. Методы выявления закономерностей традиционно широко используются в социальных науках. Например для выявления гендерной дискриминации при приеме на работу или для оценки эффективности государственных социальных программ. Очень часто они применяются в биологии и медицине для выбора оптимального лечения или для поиска генов, которые в организме действуют совместно, или для поиска подгруппы пациентов, у которых побочные эффекты наблюдаются в наиболее тяжелой форме.

# Урок 8

## Корреляции

### 8.1. Корреляция Пирсона

Самый распространенный способ формализации корреляции — это коэффициент корреляции Пирсона. Корреляция Пирсона — это мера силы линейной взаимосвязи между двумя случайными величинами  $X_1$  и  $X_2$ . Определяется она следующим образом:

$$r_{X_1 X_2} = \frac{\mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2))}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}}, \quad r_{X_1 X_2} \in [-1, 1],$$

где  $r_{X_1 X_2} = 1$  соответствует идеальной линейной взаимосвязи между случайными величинами, в которой при росте  $X_1$  растет и  $X_2$ .  $r_{X_1 X_2} = -1$  — это идеальная линейная связь с отрицательным знаком, то есть, когда  $X_1$  растет,  $X_2$  падает.  $r_{X_1 X_2} = 0$  — это случай отсутствия корреляции; это значит, что две случайные величины меняются независимо друг от друга.

#### 8.1.1. Выборочный коэффициент корреляции Пирсона

Если имеется выборка пар  $(X_{1i}, X_{2i})$  объема  $n$ , по ней очень легко посчитать выборочный коэффициент корреляции Пирсона:

$$r_{X_1 X_2} = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}}.$$

#### 8.1.2. Примеры

На рисунке 8.1а показаны диаграммы рассеяния — это графики, на одной оси которых отложены значения  $X_1$ , а на другой —  $X_2$ .

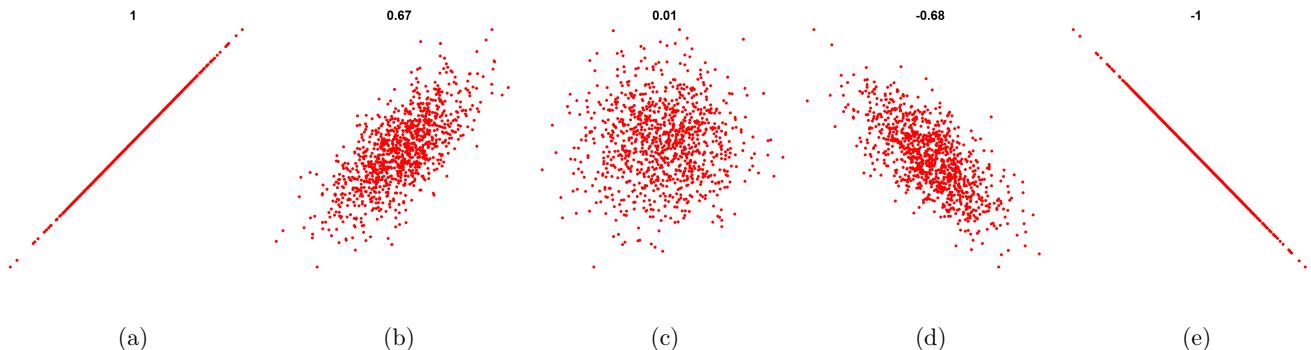


Рис. 8.1: Изменение корреляции Пирсона при размытии облака точек

Первый график (рисунок 8.1а) показывает облако точек с идеальной положительной корреляцией ( $r_{X_1 X_2} = 1$ ). Если начать это облако размывать (рисунки 8.1б, 8.1в), то коэффициент корреляции Пирсона постепенно уменьшится до 0. Если затем облако точек начать сжимать в обратном направлении, коэффициент растет по модулю и постепенно становится равным  $-1$ .

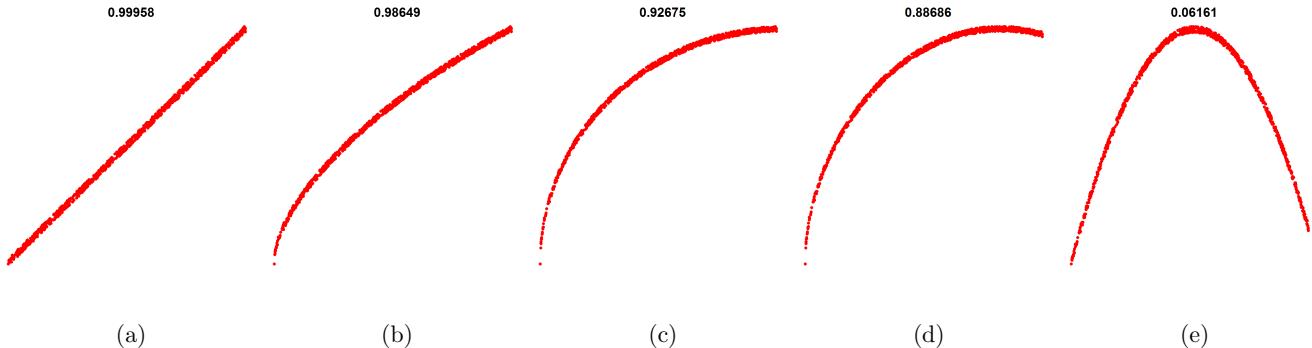


Рис. 8.2: Изменение корреляции Пирсона при увеличении изгиба облака точек

Следующий эксперимент показан на рисунке 8.2. На графике 8.2а показано облако с высокой положительной корреляцией между случайными величинами. Если начать его постепенно загибать, то коэффициент корреляции Пирсона будет уменьшаться (рисунки 8.2б, 8.2в, 8.2д). Когда форма облака становится похожей на параболу, значение выборочного коэффициента корреляции приближается к 0. Так происходит, потому что корреляция Пирсона — это мера силы линейной взаимосвязи между случайными величинами. То есть все нелинейные функциональные зависимости, даже если они очень хорошо выражены, коэффициент корреляции Пирсона не обнаруживают. Это демонстрируют примеры на рисунке 8.3. Если между случайными величинами  $X_1$  и  $X_2$  наблюдаются сложные зависимости, далекие от линейных, коэффициент корреляции Пирсона будет всё равно близким к 0 (рисунок 8.3).

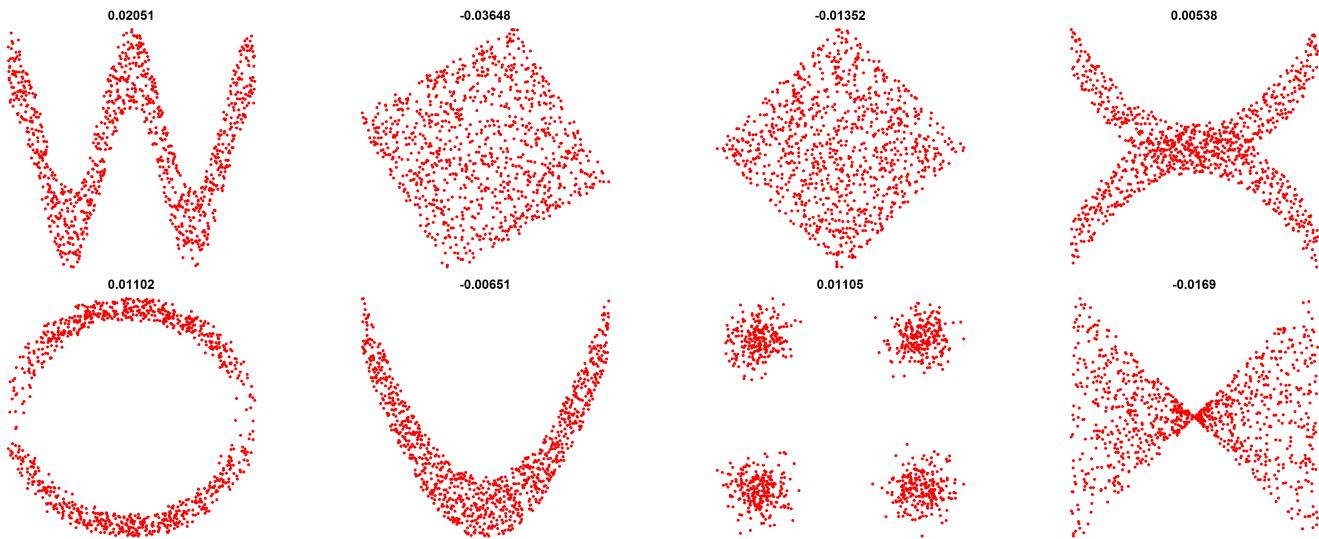


Рис. 8.3: Нелинейные зависимости между случайными величинами

Следующий важный пример изображён на рисунке 8.4. На графике 8.4а показано облако из тысячи точек с сильной отрицательной корреляцией. Если взять 5 точек из этого облака и начать постепенно отодвигать в верхний правый угол диаграммы рассеяния, то, чем дальше отодвигаются эти 5 точек, тем меньше по модулю становится значение выборочного коэффициента корреляции (рисунки 8.4б, 8.4в). С какого-то момента оно переходит через 0 и начинает расти (рисунки 8.4д, 8.4е). Достаточно сильно отодвинув всего 5 точек из тысячи, можно получить большой положительный коэффициент корреляции. Это говорит о том, что коэффициент корреляции Пирсона неустойчив к выбросам: небольшое количество точек могут оказывать на него существенное влияние, если они находятся достаточно далеко от основного облака. Это существенная

особенность корреляции Пирсона, которую нужно иметь ввиду.

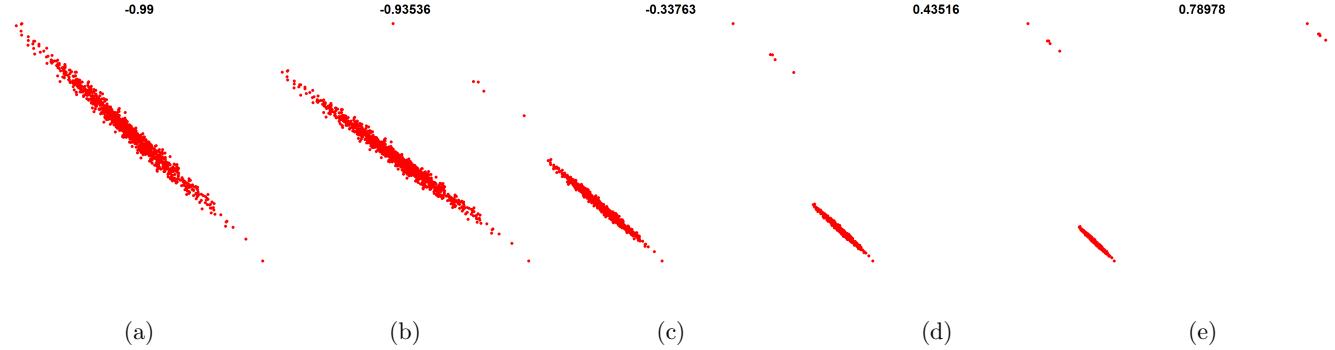


Рис. 8.4: Влияние выбросов на коэффициент корреляции Пирсона

## 8.2. Корреляция Спирмена

### 8.2.1. Определение, связь с корреляцией Пирсона

Ещё один способ формализации понятия корреляции — это корреляция Спирмена. Коэффициент корреляции Спирмена — это мера силы монотонной взаимосвязи между двумя случайными величинами, он равен коэффициенту корреляции Пирсона между рангами наблюдений. Для того, чтобы ее посчитать, нужно выборку пар  $(X_{1i}, X_{2i}), i = 1, \dots, n$  превратить наблюдение в каждой из подвыборок в ранги  $\text{rank}(X_{1i})$ , взять  $\text{rank}(X_{2i})$ , и уже на этих рангах посчитать значение коэффициента корреляции Пирсона. Именно за счет рангового преобразования получается, что корреляция Спирмена чувствительна к любой монотонной взаимосвязи между  $X_1$  и  $X_2$ , поскольку ранговое преобразование превращает любую монотонную взаимосвязь в линейную.

Корреляция Спирмена наследует часть свойств корреляции Пирсона. Она точно так же меняется от  $-1$  до  $1$ , где крайние значения отрезка соответствуют идеальной, в данной случае монотонной, взаимосвязи между случайными величинами, а  $0$  — полному отсутствию монотонной взаимосвязи между ними.

### 8.2.2. Выборочный коэффициент корреляции Спирмена

Если имеется выборка пар  $(X_{1i}, X_{2i}), i = 1, \dots, n$ , то выборочный коэффициент корреляции Спирмена вычисляется следующим образом:

$$\begin{aligned} \rho_{X_1 X_2} &= \frac{\sum_{i=1}^n (\text{rank}(X_{1i}) - \frac{n+1}{2})(\text{rank}(X_{2i}) - \frac{n+1}{2})}{\frac{1}{12}(n^3 - n)} = \\ &= 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (\text{rank}(X_{1i}) - \text{rank}(X_{2i}))^2 \end{aligned}$$

В данном случае формулу выборочной корреляции Пирсона можно немного упростить, поскольку заранее известно, чему равны средние ранги в выборках и чему равны их дисперсии.

### 8.2.3. Примеры

Чтобы посмотреть, какие из свойств корреляции Спирмена отличаются от свойств корреляции Пирсона, можно воспроизвести эксперименты с облаками точек.

Корреляция Спирмена примерно так же, как и корреляция Пирсона, реагирует на сжатие и размывание облака точек на диаграмме рассеяния (рисунок 8.5). Видно, что крайние случаи идеальной линейной взаимосвязи (графики 8.5a, 8.5e) соответствуют  $-1$  и  $1$ , а в середине получаются значения коэффициента корреляции, близкие к  $0$  (график 8.5c).

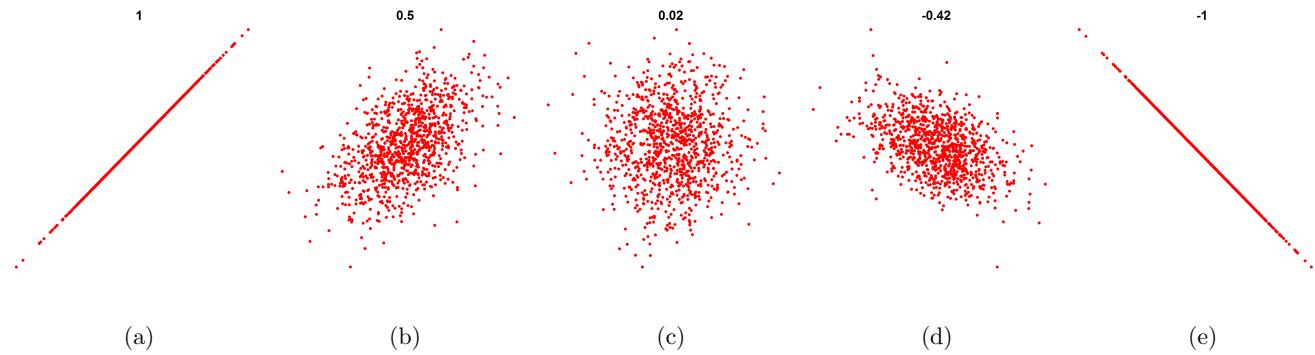


Рис. 8.5: Изменение значения коэффициента корреляции Спирмена при размытии облака точек

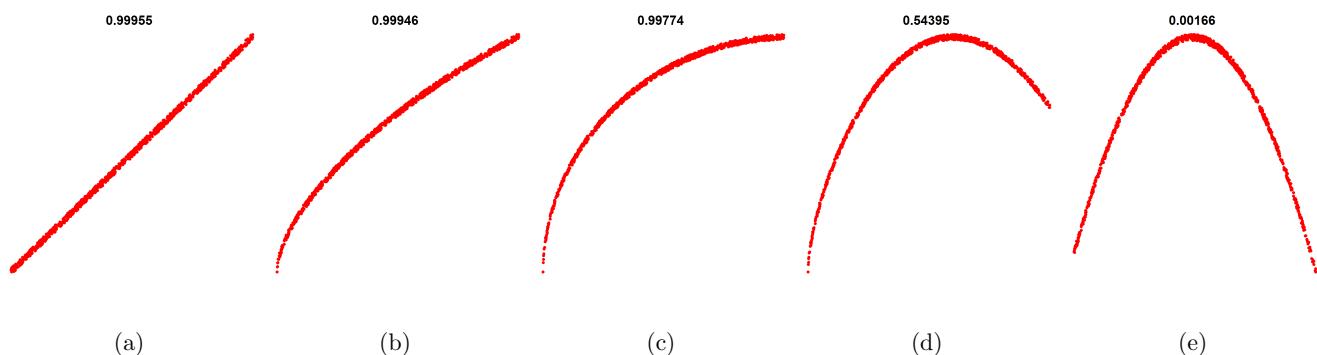


Рис. 8.6: Изменение значения коэффициента корреляции Спирмена при увеличении изгиба облака точек

Более интересные результаты получаются в эксперименте с загибанием облака точек (рисунок 8.6). Видно, что пока зависимость между  $X_1$  и  $X_2$  остается монотонной (рисунки 8.6a, 8.6b), значение коэффициента корреляции Спирмена почти не убывает<sup>1</sup>. Однако потом, когда облако точек начинает превращаться в параболу (рисунки 8.6c, 8.6d, 8.6e), значение выборочного коэффициента корреляции Спирмена постепенно превращается в 0. Корреляция Спирмена не обнаруживает взаимосвязи между  $X_1$  и  $X_2$ , отличные от монотонных. Это можно заметить и на следующих примерах. Когда между  $X_1$  и  $X_2$  есть какие-то сложные функциональные взаимосвязи (рисунок 8.7), корреляция Спирмена все равно остается близкой к 0, поскольку они далеки от монотонных.

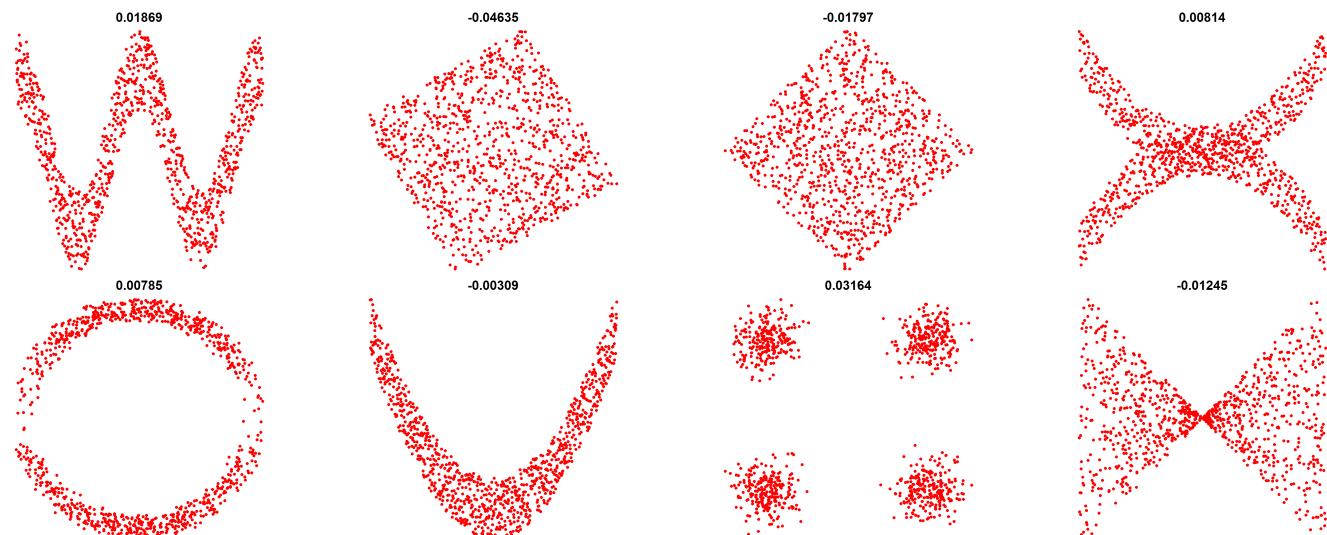


Рис. 8.7: Коэффициент корреляции Спирмена при нелинейных зависимостях между случайными величинами

<sup>1</sup>Небольшие изменения объясняются тем, что связь между признаками не в точности монотонная, а слегка зашумлена.

Гораздо более интересные результаты получаются в эксперименте с выбросами (рисунок 8.8). Когда из облака точек с сильной отрицательной корреляцией (рисунок 8.8a) пять точек начинают выдвигаться в правый верхний угол диаграммы рассеяния, значение коэффициента корреляции Спирмена сначала немного уменьшается (рисунки 8.8b, 8.8c). Однако как только эти пять точек оказываются за пределами диапазона изменений случайных величин в основном облаке, значение коэффициента корреляции Спирмена меняется перестает (рисунки 8.8d, 8.8e). Как бы далеко они ни отодвигались, не удается получить большую положительную корреляцию, как в случае с корреляцией Пирсона. Это говорит о том, что коэффициент корреляции Спирмена гораздо более устойчив к выбросам, то есть, небольшое количество точек с нетипичными значениями признаков очень слабо влияют на выборочное значение коэффициента корреляции.

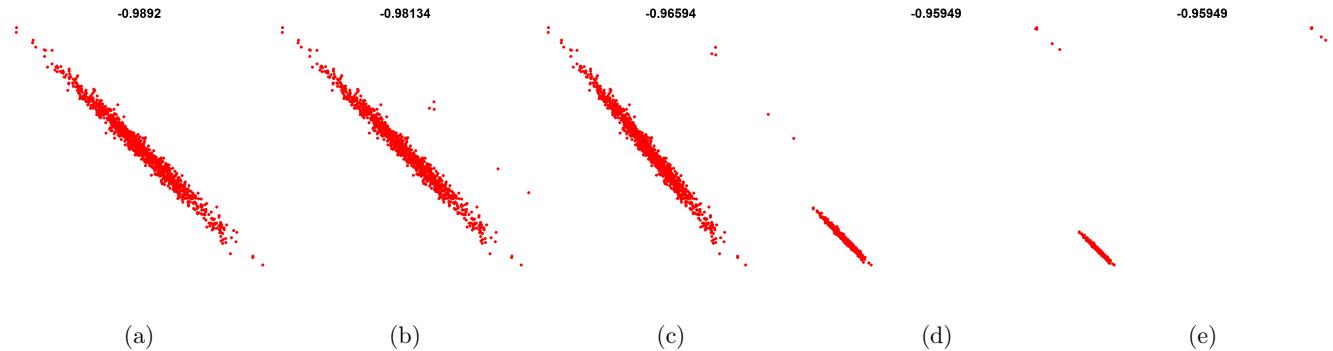


Рис. 8.8: Влияние выбросов на коэффициент корреляции Спирмена

## 8.3. Корреляция Мэтьюса и коэффициент Крамера

### 8.3.1. Корреляция Мэтьюса

Коэффициент корреляции Мэтьюса — это мера силы взаимосвязи между двумя бинарными переменными. Для того чтобы его вычислить, необходимо использовать таблицу сопряженности (таблица 8.1).

		$X_2$	0	1
		$X_1$	0	1
$X_1$	0	$a$	$b$	
	1	$c$	$d$	

Таблица 8.1: Таблица сопряжённости

В строках таблицы сопряжённости находятся значения одного признака, по столбцам — второго, в каждой ячейке — количество объектов, на которых реализовалась эта пара. Коэффициент корреляции Мэтьюса вычисляется по данным из таблицы сопряжённости следующим образом:

$$MCC_{X_1 X_2} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

Точно так же, как и коэффициенты Пирсона и Спирмена, корреляция Мэтьюса лежит в диапазоне от  $-1$  до  $1$ .  $MCC_{X_1 X_2} = 0$  точно так же соответствует случаю полного отсутствия взаимосвязи между переменными.  $MCC_{X_1 X_2} = 1$  соответствует ситуации, когда у  $X_1$  и  $X_2$  полностью совпадают, то есть  $b = c = 0$ , в выборке отсутствуют объекты, на которых значения  $X_1$  и  $X_2$  отличаются.  $MCC_{X_1 X_2} = -1$  — это противоположная ситуация: в выборке нет ни одного объекта, на которых значения двух бинарных признаков совпадают.

### 8.3.2. Коэффициент V Крамера

Этот подход можно обобщить на случай категориальных признаков. Пусть случайная величина  $X_1$  принимает  $K_1$  различных значений, а  $X_2$  —  $K_2$  разных значений. Можно составить большую таблицу сопряженности (таблица 8.2), у которой в строке  $i$  и столбце  $k$  будет стоять  $n_{ij}$  — количество объектов выборки, на которых  $X_1 = i$ , а  $X_2 = j$ .

$X_1 \backslash X_2$	1	$\dots$	$j$	$\dots$	$K_2$
1					
$\vdots$					
$i$			$n_{ij}$		
$\vdots$					
$K_1$					

Таблица 8.2: Таблица сопряжённости  $K_1 \times K_2$

На основании этой таблицы сопряженности вычисляется мера взаимосвязи между  $X_1$  и  $X_2$ . Эта мера называется коэффициентом  $V$  Крамера. Он обозначается  $\phi_c$ , и равен корню из специальным образом нормированного значения статистики хи-квадрат:

$$\phi_c(X_1^n, X_2^n) = \sqrt{\frac{\chi^2(X_1^n, X_2^n)}{n(\min(K_1, K_2) - 1)}}.$$

Далее будет описано, как статистика хи-квадрат считается для таблицы сопряженности.

Коэффициент Крамера принимает значения исключительно в интервале от 0 до 1, то есть он не может быть отрицательным. 0, как и раньше, соответствует полному отсутствию взаимосвязи, а 1 — полному совпадению переменных  $X_1$  и  $X_2$  с точностью до переименования уровней. Корреляция между двумя категориальными переменными не может быть отрицательной, поскольку уровни категориальных переменных не связаны друг с другом отношениями порядков.

### 8.3.3. Пары переменных разных видов

Итак, в этом разделе было описано, как считать корреляцию между парами бинарных переменных, парами категориальных, а до этого — как считать корреляцию между парами непрерывных переменных. Однако до сих пор не сказано, что делать, если признаки в паре разных видов.

Например, пусть  $X_1 \in \mathbb{R}$  — непрерывный признак, а  $X_2 \in \{0, 1\}$  — бинарный. Чисто теоретически на этих данных можно посчитать корреляцию Пирсона или Спирмена. Никакая из них не сломается из-за того, что одна из выборок будет не непрерывной, а бинарной. Но так делать не стоит, это очень плохо. Корреляции Пирсона и Спирмена не рассчитаны на применение к бинарным или категориальным признакам. Полученная величина будет иметь мало смысла.

На самом деле для пар признаков, один из которых непрерывный, а другой — категориальный, вообще не нужно считать никакой коэффициент корреляции.  $X_1 \in \mathbb{R}$  и  $X_2 \in \{0, 1\}$  будут положительно коррелированы, если

$$\mathbb{E}(X_1 | X_2 = 1) > \mathbb{E}(X_1 | X_2 = 0).$$

Таким образом, мерой силы взаимосвязи между  $X_1$  и  $X_2$  может служить просто разность этих математических ожиданий:

$$\mathbb{E}(X_1 | X_2 = 1) - \mathbb{E}(X_1 | X_2 = 0)$$

Эта величина не нормированная, она может меняться в любом диапазоне, от  $-\infty$  до  $+\infty$ . Однако её гораздо легче интерпретировать, чем коэффициент корреляции, который можно вычислить на такой паре выборок.

## 8.4. Значимость корреляции

В этой части пойдёт речь о том, как правильно интерпретировать значения коэффициентов корреляции. В частности, будет дан ответ на вопрос, можно ли по полученному выборочному значению коэффициента корреляции сказать, что он достаточно большой и отличается от 0.

### 8.4.1. Корреляция непрерывных величин

За 100 дней собраны данные о значениях средней дневной температуры и количестве проданных рожков мороженого. Значение коэффициента корреляции Пирсона, посчитанное по этой выборке:  $r_{X_1 X_2} = 0.45$ , Спирмена:

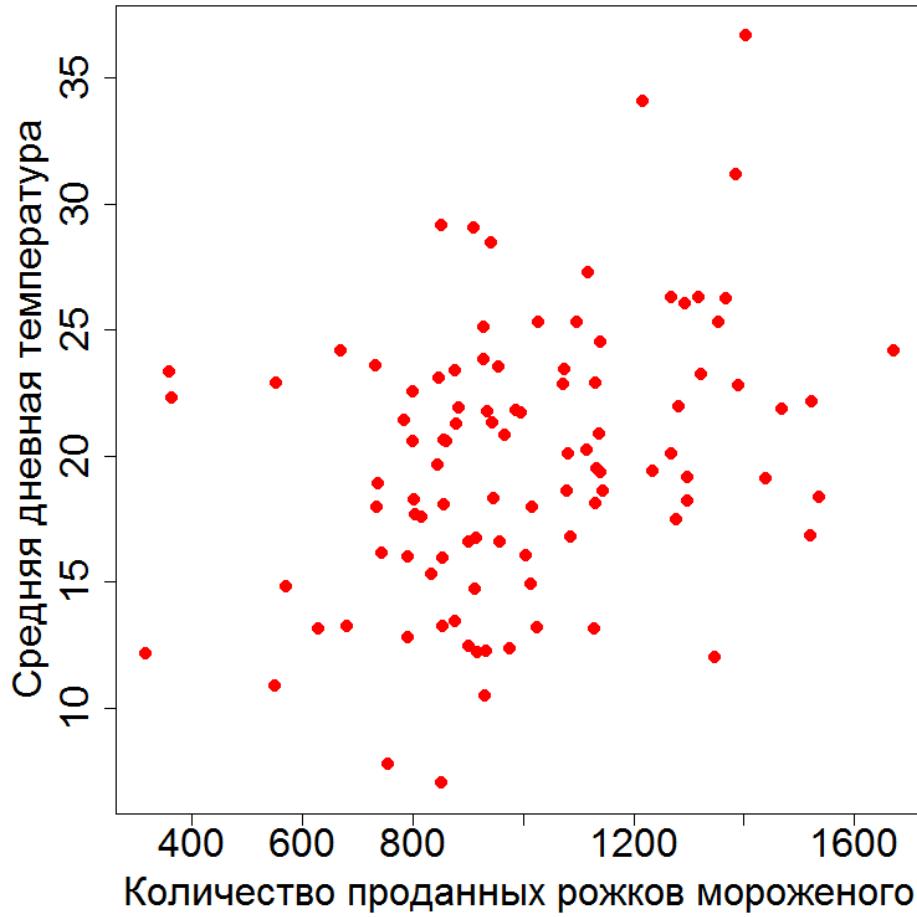


Рис. 8.9: Данные о продажах мороженого и средней дневной температуре

$\rho_{X_1 X_2} = 0.44$ . Можно ли по полученным значениям утверждать, что объем продаж мороженого и среднедневная температура статистически взаимосвязаны?

Ответить на этот вопрос позволяет статистический критерий Стьюдента (таблица 8.3).

выборки:	$(X_{1i}, X_{2i}), i = 1, \dots, n,$
нулевая гипотеза:	$H_0: r_{X_1 X_2} = 0;$
альтернатива:	$H_1: r_{X_1 X_2} < \neq > 0;$
статистика:	$T = \frac{r_{X_1 X_2} \sqrt{n-2}}{\sqrt{1-r_{X_1 X_2}^2}},$
нулевое распределение:	$T \sim St(n-2).$

Таблица 8.3: Описание статистического критерия Стьюдента

Если нулевая гипотеза справедлива, то есть, корреляции нет, эта статистика имеет распределение Стьюдента с числом степеней свободы  $n-2$  (рисунок 8.10).

Для проверки такой же точно гипотезы, но о корреляции Спирмена, а не Пирсона, можно использовать абсолютно тот же самый критерий Стьюдента.

В примере с мороженым нулевая гипотеза о том, что линейной связи нет против двусторонней альтернативы критерием Стьюдента уверенно отвергается. Признаки действительно линейно статистически взаимосвязаны. 95% доверительный интервал: [0.28, 0.59]. Такой доверительный интервал, кстати, можно построить, как на основе статистики критерия Стьюдента, так и с помощью бутстрепа.

Также можно использовать корреляцию Спирмена, чтобы проверить гипотезу об отсутствии монотонной взаимосвязи между двумя признаками:

$$H_0: \rho_{X_1 X_2} = 0,$$

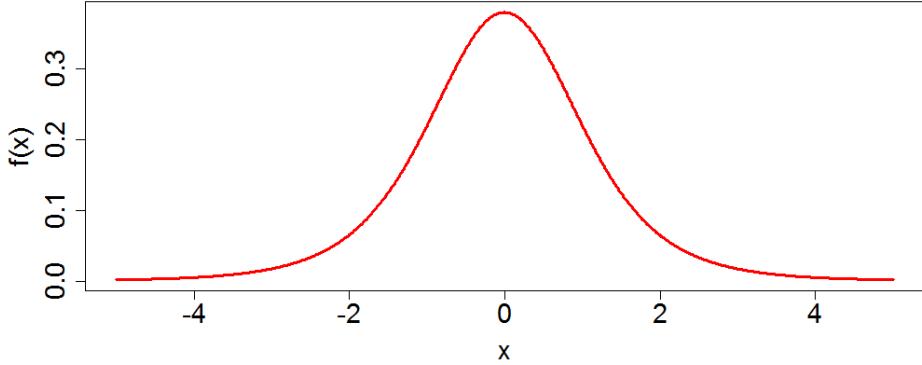


Рис. 8.10: Распределение Стьюдента

против двусторонней альтернативы

$$H_1: \rho_{X_1 X_2} \neq 0$$

Критерием Стьюдента эта гипотеза также отвергается с очень похожим достигаемым уровнем значимости  $p = 3 \times 10^{-6}$ . Признаки действительно монотонно связаны. Это не удивительно, поскольку ранее было показано, что они связаны линейно, а линейная взаимосвязь — это частный случай монотонной. 95% доверительный интервал для корреляции Спирмена: [0.26, 0.60].

#### 8.4.2. Корреляция бинарных величин

В качестве примера работы с бинарными признаками можно рассмотреть задачу оценки эффективности тромболитической терапии по данным эксперимента, который проводился в Московской городской клинической больнице №25. В эксперименте участвовало 206 пациентов. Требуется понять, влияет ли наличие сахарного диабета у этих пациентов на эффективность тромболитической терапии.

	Выздоровели	Не выздоровели
Диабет	48	30
Нет	92	36

Таблица 8.4: Данные эксперимента по оценке эффективности тромболитической терапии

Данные эксперимента представляют собой таблицу  $2 \times 2$  (таблица 8.4). Значение коэффициента корреляции Мэттьюса, подсчитанное по этой таблице:  $MCC = -0.1074$ . Возможно, наличие сахарного диабета понижает шансы на выздоровление у пациентов. Эту гипотезу можно проверить формально с помощью критерия хи-квадрат (таблица 8.5).

выборки: нулевая гипотеза: альтернатива: статистика: нулевое распределение:	$(X_{1i}, X_{2i}), i = 1, \dots, n,$ $X_1, X_2 \in \{0, 1\};$ $H_0: MCC_{X_1 X_2} = 0;$ $H_1: MCC_{X_1 X_2} \neq 0;$ $\chi^2 = n MCC_{X_1 X_2}^2;$ $\chi^2 \sim \chi_1^2.$
---	---

Таблица 8.5: Описание критерия хи-квадрат

Если нулевая гипотеза справедлива и значение коэффициента корреляции действительно равно 0, то статистика этого критерия имеет распределение хи-квадрат с одной степенью свободы (рисунок 8.11).

При рассмотрении задаче проверки нормальности уже шла речь о том, что критерий хи-квадрат достаточно капризный. Вот и в этом случае требуется, чтобы выборки были достаточно большими:  $n \geq 40$ . Кроме того, необходимо, чтобы каждая из следующих четырёх величин была больше 5:

$$\frac{(a+c)(a+b)}{n}, \frac{(a+c)(c+d)}{n}, \frac{(b+d)(a+b)}{n}, \frac{(b+d)(c+d)}{n} > 5$$

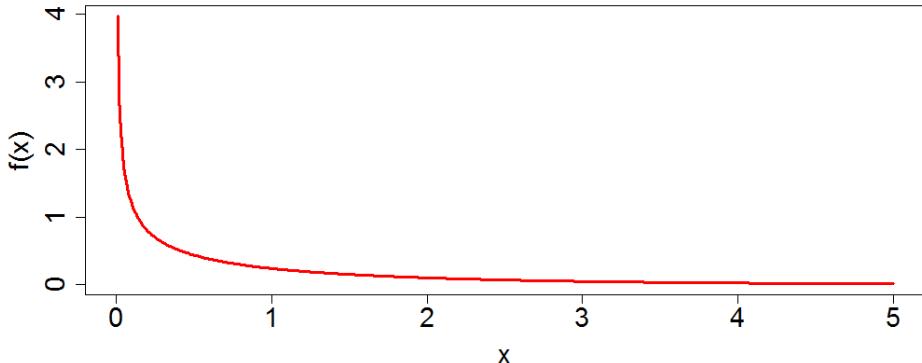


Рис. 8.11: Распределение хи-квадрат с 1 степенью свободы

Далее будет рассказано, откуда берутся эти четыре величины.

Итак, можно применить критерий хи-квадрат к данным эксперимента по оценке эффективности тромболитической терапии. Нулемая гипотеза  $H_0$ : эффективность лечения не зависит от наличия диабета, против двухсторонней альтернативы критерием хи-квадрат не отвергается. Достигаемый уровень значимости  $p = 0.1651$ , это больше, чем уровень значимости 0.05. Таким образом, нельзя утверждать, что между этими двумя признаками есть связь.

#### 8.4.3. Корреляция категориальных величин

Критерий хи-квадрат можно обобщить на случай категориальных признаков (таблица 8.6). Таблица 8.7 — это таблица сопряженности для  $X_1$  и  $X_2$ . Пусть  $X_1$  принимает  $k_1$  разных уровней,  $X_2$  —  $k_2$  разных уровней. В ячейке на пересечении строки  $i$  и столбца  $j$  стоит  $n_{ij}$  — количество объектов, на которых реализуется значение  $X_1$  под номером  $i$  и значение  $X_2$  под номером  $j$ . Дополнительно введены обозначения для сумм по строкам и столбцам:  $n_{i+}$  — это сумма по строке  $i$ , а  $n_{+j}$  — по столбцу  $j$ .

выборки: нулевая гипотеза: альтернатива: статистика:	$(X_{1i}, X_{2i}), i = 1, \dots, n,$ $X_1 \in \{1, \dots, K_1\}, X_2 \in \{1, \dots, K_2\},$ $H_0: X_1 \text{ и } X_2 \text{ независимы};$ $H_1: H_0 \text{ неверна};$ $\chi^2(X_1^n, X_2^n) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \frac{(n_{ij} - \frac{n_{i+} n_{+j}}{n})^2}{\frac{n_{i+} n_{+j}}{n}} =$ $= n \left( \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \frac{n_{ij}^2}{n_{i+} n_{+j}} - 1 \right);$ $\chi^2(X_1^n, X_2^n) \sim \chi^2_{(K_1-1)(K_2-1)}.$
---	--

Таблица 8.6: Описание критерия хи-квадрат для категориальных признаков

$X_2$	1	$\dots$	$j$	$\dots$	$K_2$	$\Sigma$
$X_1$	1					
1						
$\vdots$						
$i$			$n_{ij}$			$n_{i+}$
$\vdots$						
$K_1$						
$\Sigma$			$n_{+j}$			$n$

Таблица 8.7: Таблица сопряженности для категориальных признаков

В статистике этого критерия учитывается отклонение между  $n_{ij}$ , количеством объектов в каждой ячейке, и ожидаемым количеством объектов в этой ячейке при условии справедливости нулевой гипотезы.

При справедливости нулевой гипотезы статистика критерия имеет распределение хи-квадрат (рисунок 8.12).

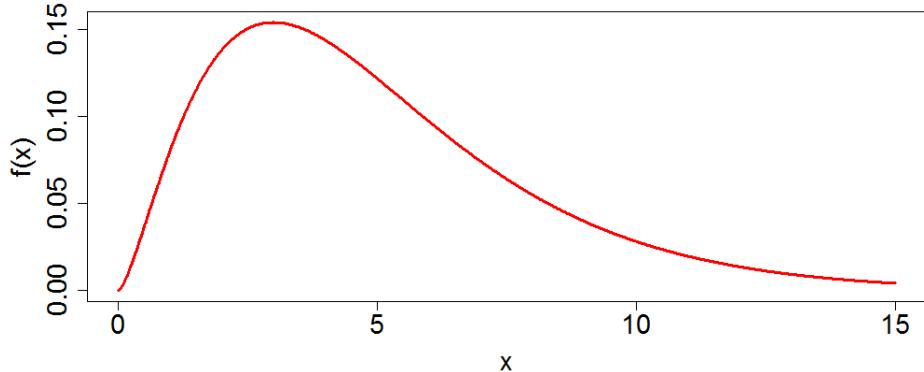


Рис. 8.12: Распределение хи-квадрат

Несложно показать, что критерий для таблиц  $2 \times 2$ , который рассматривался ранее, является частным случаем этого критерия.

Критерий хи-квадрат для таблиц сопряженности может применяться при выполнении следующих условий. Нужно, чтобы выборки были достаточно большими:  $n \geq 40$ . Кроме того, необходимо, чтобы ожидаемое количество элементов в каждой ячейке таблицы было меньше  $5 (\frac{n_{i+j}}{n} < 5)$ , не более, чем в 20% ячеек.

Можно считать, что для категориальных признаков критерий хи-квадрат проверяет гипотезу о равенстве нулю коэффициента  $V$  Крамера против альтернативы, что он нулю не равен. Вообще говоря, коэффициент  $V$  Крамера определяется как раз через статистику критерия хи-квадрат:

$$\phi_c(X_1^n, X_2^n) = \sqrt{\frac{\chi^2(X_1^n, X_2^n)}{n(\min(K_1, K_2) - 1)}}.$$

## 8.5. Буллшифт и консервативность

### 8.5.1. Исследование и поставленный эксперимент

В конце апреля 2016 года в престижном журнале PLOS one вышла статья под названием «Восприятие «буллшифта» как глубокомысленного ассоциировано с поддержкой Круза, Рубио, Трампа и консерватизмом». Круз, Рубио и Трамп — это кандидаты в президенты США от республиканской партии. По определению авторов «буллшифт» — это бессодержательное, нелогичное или явно противоречащее элементарным научным знаниям утверждение. В качестве примеров «буллшифта» они приводят такие фразы, как «скрытый смысл трансформирует беспрецедентную абстрактную красоту» или «воображение лежит в основе экспоненциальных пространственно-временных событий» (эти фразы получены специальным генератором «буллшифта»).

В статье анализируются данные эксперимента, в котором участвовали 196 граждан США, 43% из которых женщины, средний возраст испытуемых составляет 36 лет. Испытуемые (это достаточно необычно) набраны на платформе Amazon Mechanical Turk. Это платформа, на которой пользователям из Интернета можно заказать какое-то простое, достаточно механическое, не требующее высокой квалификации задание, и они его сделают за достаточно небольшие деньги. Например, типичное задание для этой платформы — это разметка картинок.

В эксперименте испытуемым нужно было ответить на ряд вопросов. Во-первых, им нужно было оценить глубокомысленность предъявляемых им утверждений по шкале от 1 (абсолютно не глубокое) до 5 (очень глубокое). Кроме того, каждый из них должен был оценить степень своей симпатии к трем кандидатам в президенты США от республиканской партии и трем кандидатам от демократической партии, также по шкале от 1 (очень не симпатичен) до 5 (очень симпатичен). Помимо этого, каждый из них должен был оценить степень консервативности собственных политических взглядов по семибалльной шкале Лайкерта, где 1 соответствует очень либеральным взглядам, а 7 — очень консервативным. Часть утверждений, предъявлен-

ных испытуемым, была «буллицитом», а часть — относительно редкими поговорками (например, «промокший человек не боится дождя»).

Данные были проанализированы следующим образом. Для каждого испытуемого была вычислена средняя склонность читать «буллицит» глубокомысленным. Для этого признака была посчитана корреляция Спирмена с консервативностью политических взглядов и степенями симпатии к шести кандидатам в президенты. Для проверки значимости отличия от нуля этой корреляции, использовался критерий Стьюдента.

Были получены следующие результаты. Обнаружена значимая положительная корреляция между тягой к «буллициту» и симпатией к Теду Крузу, Марку Рубио и Дональду Трампу (три кандидата от республиканской партии). Кроме того, тяга к «буллициту» положительно ассоциирована со степенью консервативности и эта корреляция тоже значима.

Это веселое исследование, но у него есть некоторые проблемы. Первая и самая важная проблема заключается в сомнительной репрезентативности выборки. Аудитория Amazon Mechanical Turk вовсе не является случайной выборкой из граждан США. В этой аудитории преобладают белые молодые мужчины с относительно невысоким доходом и достаточно хорошо образованные или получающие образование. Таким образом, непонятно, на какую генеральную совокупность можно обобщать результаты, полученные в исследовании: эта выборка крайне смещена для случая, когда хочется делать выводы о всех гражданах США. Связанная с этим проблема заключается в несбалансированности выборки по некоторым признакам.

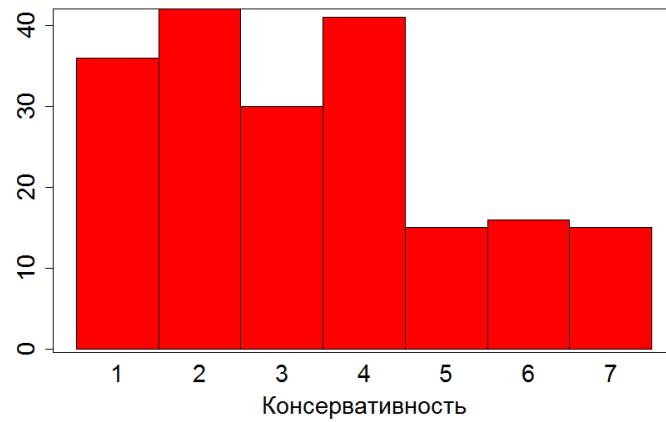


Рис. 8.13: Распределение консервативности испытуемых в выборке

На рисунке 8.13 показано распределение значения показателя консервативности в исследуемой выборке. В ней преобладают испытуемые с либеральными или умеренными взглядами, относительно небольшая часть испытуемых считает свои политические взгляды консервативными. Почему это проблема — станет понятно, если посмотреть на сырье данные.

На графике 8.14а по горизонтальной оси отложена тяга к «буллициту», а по вертикальной — степень поддержки Теда Круза. Каждая точка — это один испытуемый. Корреляция Спирмена между этими двумя признаками:  $\rho_{XY} = 0.3$ . Достигаемый уровень значимости критерия Стьюдента против двусторонней альтернативы:  $p = 2 \times 10^{-5}$ , то есть нулевая гипотеза об отсутствии корреляции отвергается. Если через это облако точек провести регрессионную прямую, то видно, что у неё, действительно, положительный наклон, но он не так уж велик. Если вместо линейной регрессии произвести локальное сглаживание методом LOESS, то получится синяя кривая, по которой видно, что картина не так однозначна. Действительно, до какой-то степени повышение тяги к «буллициту» соответствует повышению уровня поддержки Теда Круза, но это работает не на всей области определения тяги к «буллициту», признака, отложенного по горизонтальной оси.

Для сравнения можно посмотреть на графики по другим кандидатам в президенты. На рисунке 8.14б показан график для Марка Рубио. Здесь корреляция Спирмена  $\rho_{XY} = 0.2$ . Достигаемый уровень значимости соответствующего критерию Стьюдента против двусторонней альтернативы:  $p = 0.0064$ . Различия между полученным коэффициентом корреляции и нулем значимы. На графике при этом видна приблизительно та же ситуация, что и да этого, хотя угол наклона регрессионной прямой немножко уменьшился.

График для Дональда Трампа показан на рисунке 8.14с. Корреляция Спирмена здесь:  $\rho_{XY} = 0.15$ . Достигаемый уровень значимости  $p = 0.0324$ . Отличие корреляции от нуля все еще значимо на уровне значимости 0.05. Угол наклона регрессионной прямой еще немного уменьшился.

Для сравнения на рисунке 8.14д представлен график для Хиллари Клинтон. Корреляция Спирмена здесь  $\rho_{XY} = 0.09$ . Соответствующий достигаемый уровень значимости  $p = 0.212$ , то есть данные не позволяют

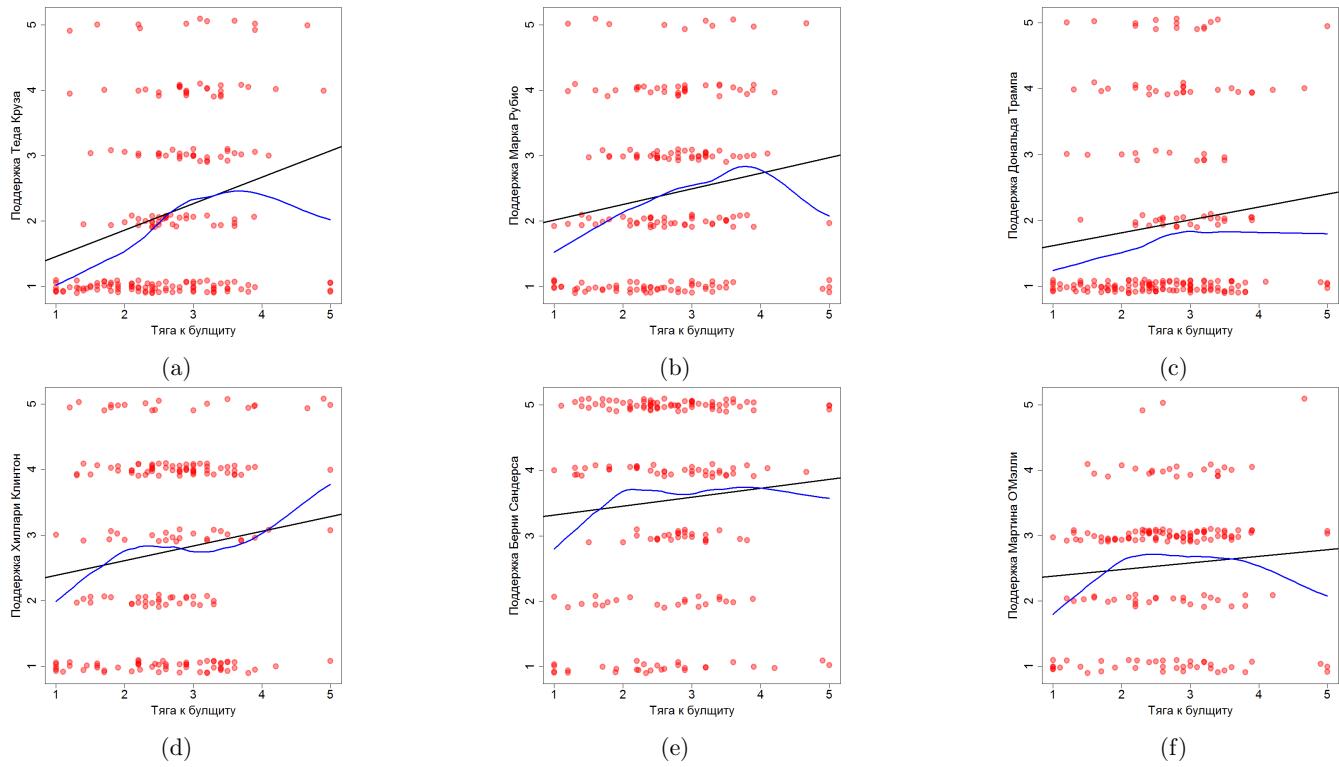


Рис. 8.14: Данные о тяге к «бульшиту» и поддержке кандидатов в президенты США. В верхнем ряду — кандидаты республиканской партии, в нижнем — демократической.

отвергнуть нулевую гипотезу о том, что признаки не коррелированы. На графике, однако, происходит примерно то же самое: положительный угол наклона регрессионной прямой сохраняется, он только еще совсем немного уменьшился. Кривая, полученная методом LOESS, здесь имеет точно такой же повышающийся вид.

На рисунке 8.14 сгруппированы графики для всех шести кандидатов в президенты. В верхнем ряду — кандидаты от республиканской партии, где корреляция между признаками везде значима. В нижнем ряду — кандидаты от демократической партии, на которых корреляция везде незначима. При этом на всех графиках происходит примерно одно и то же. Основное отличие между ними, — это то, что в верхнем ряду большая часть точек сосредотачивается в облаке, соответствующем значению признака, отложенного по вертикальной оси, равного единице. Но поскольку в выборке большая часть испытуемых не придерживается консервативных взглядов, они кандидатов от республиканской партии и не поддерживают. Именно это множество точек и оказывает наибольшее влияние на коэффициенты корреляции.

Коэффициент корреляции Спирмена предназначен для работы с непрерывными признаками, а в данном случае рассматриваемые признаки существенным образом дискретны — они измерены в шкале от 1 до 5. Кроме того, лучше всего корреляции Спирмена и Пирсона и соответствующие им критерии Стьюдента работают в ситуациях, когда признаки, между которыми вычисляется корреляция, распределены нормально. В данном случае это совсем не так.

Главная мораль этой истории заключается в том, что всегда нужно смотреть на сырье данные. Никакая статистика, никакие цифры, вычисленные на этих данных, не могут полностью описать, что происходит в данных.

Корреляционный анализ — это не сосисочная машина, которой можно подать что угодно на вход, а на выходе получить идеально правильные выводы. Нужно всегда следить за качеством исходных данных и проверять, соответствует ли распределение признаков в выборках тому распределению, которое предполагается в применяемых методах.

## 8.6. Корреляция и причинно-следственная связь

Давайте проанализируем корреляцию между суммарными продажами мороженого за день и количеством людей, которое в этот день утонуло на всех пляжах города (рисунок 8.6e). Корреляция между этими двумя

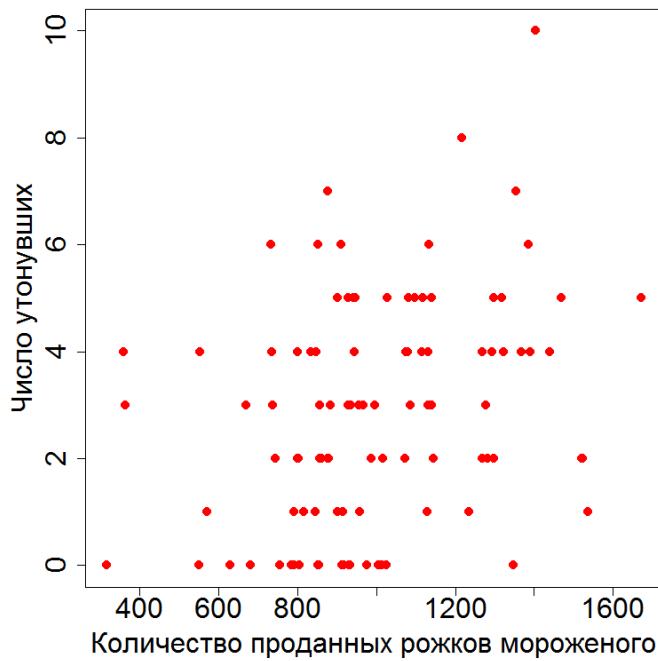
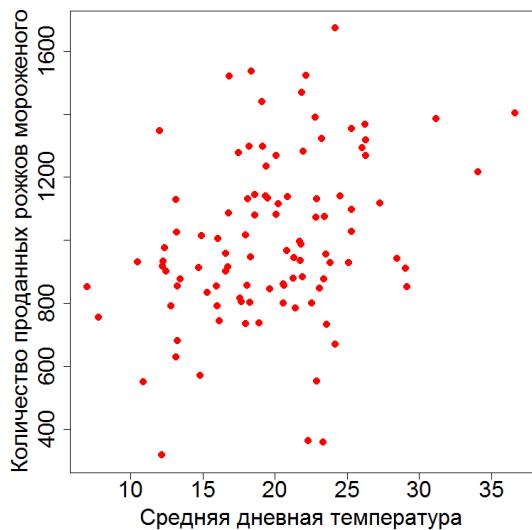
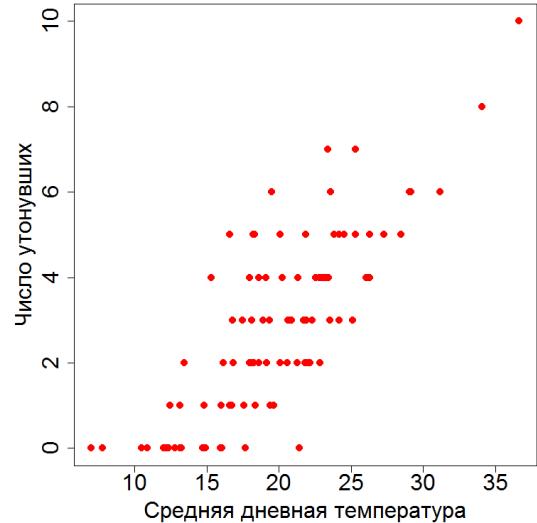


Рис. 8.15: Данные о продажах мороженого за день и количестве утонувших на пляжах города людей в этот день

признаками положительная:  $r_{X_1 X_2} = 0.33$ . Достигаемый уровень значимости критерия Стьюдента:  $p = 0.0009$ . 95% доверительный интервал для корреляции Пирсона: [0.138, 0.491].



(а) Данные о продажах мороженого за день и средней температуре за день



(б) Данные о продажах мороженого за день и средней температуре за день

Рис. 8.16

Из этих результатов можно сделать вывод, что чем больше люди едят мороженого, тем чаще они тонут. Или, например, что из-за того, что люди часто тонут, другие люди больше едят мороженого. Однако очевидно, что это не так. Ранее было показано, что продажи мороженого достаточно сильно коррелированы со среднедневной температурой (рисунок 8.16а). Если посмотреть на корреляцию между среднедневной температурой и числом утонувших людей (рисунок 8.16б), видно, что она еще больше, и это естественно. Таким образом, в данном примере значимость корреляции между продажами мороженого и числом утонувших

людей объясняется воздействием третьего признака — среднедневной температуры. Именно третий признак — единственный из трех, который оказывает причинно-следственное влияние на оставшиеся два. Никаких других причинно-следственных связей между этими тремя признаками быть просто не может.

В учебниках по статистике можно найти большое количество веселых примеров таких ложных корреляций, объясняющихся воздействием третьего скрытого признака. Например, количество самоубийств и радиоприемников на душу населения высоко положительно коррелировано, и это объясняется воздействием признака «размер города». Уровень углекислого газа в атмосфере планеты и распространенность ожирения также высоко положительно коррелированы — это объясняется ростом со временем уровня жизни. Рыночная доля браузера Internet Explorer и количество убийств в США тоже положительно коррелированы, и это объясняется в первую очередь фактором времени: во времени снижается и тот, и другой показатель.

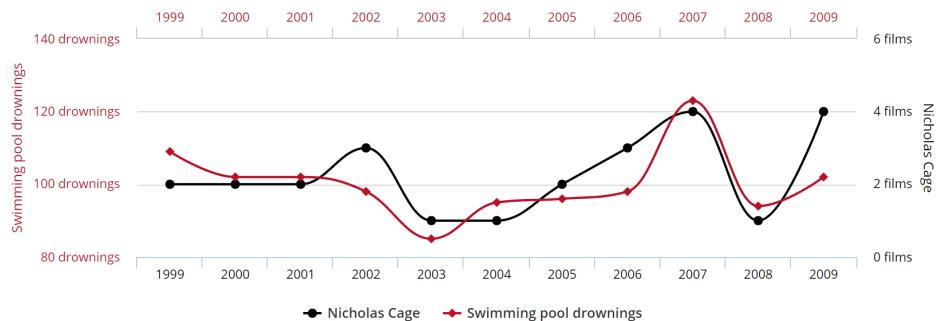


Рис. 8.17: Данные о количестве людей, утонувших в бассейне и количеством фильмов, в которых снялся Николас Кейдж

Иногда корреляцию между парой признаков нельзя объяснить даже влиянием никакого третьего другого, а эта корреляция просто случайна. Если взять достаточно большое количество величин и искать среди них все возможные попарные корреляции, найдётся очень много странного. Например, можно показать, что значима положительная корреляция между количеством людей, которые утонули при падении в бассейн, и количеством фильмов, в которых снялся за год Николас Кейдж. Корреляция Пирсона между этими двумя признаками  $r_{X_1 X_2} = 0.67$ . Достигаемый уровень значимости критерия Стьюдента:  $p = 0.0253$ . 95% доверительный интервал для корреляции Пирсона:  $[0.110, 0.905]$ . Несмотря на то, что он довольно широкий, 0 в нём не содержится. Тем не менее, абсолютно очевидно, что связать эти два признака какой бы то ни было цепочкой причинно-следственных связей не представляется возможным. Этот эффект явно случайный, и то, что его нашли, — это следствие того, что его очень хорошо искали.

Главный вывод: из корреляции никогда не следует причинно-следственная связь, но из причинно-следственной связи часто следуют корреляции. Причинно-следственная связь оставляет в данных какие-то следы, которые можно обнаружить в том числе и корреляционными методами. Однако для этого есть другие специальные методы, связанные с построением графов причинности, и лучше использовать именно их.

# Урок 9

## Множественная проверка гипотез

### 9.1. В чем проблема

Этот урок посвящен проблеме множественной проверки гипотез. Для того чтобы понять, в чем заключается эта проблема, можно рассмотреть несколько примеров.

#### 9.1.1. Поиск экстрасенсов

Первый пример связан с исследованиями Джозефа Райна. Это американский ученый 50-х годов, который занимался исследованиями возможностей экстрасенсорного восприятия. Первый этап таких исследований — это поиск экстрасенсов. Джозеф Райн придумал для этого следующий эксперимент. Испытуемому предлагалось угадать цвета десяти карт, лежащих рубашкой вверх (рисунок 9.1).

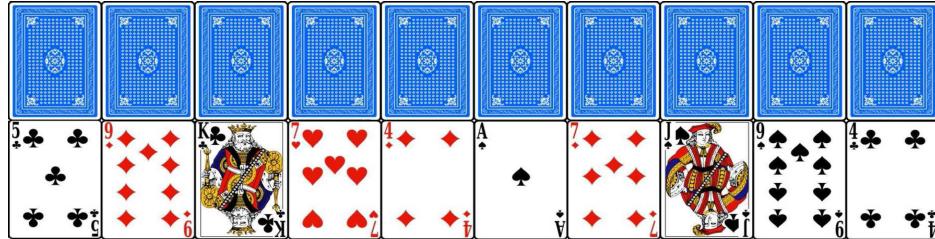


Рис. 9.1: Эксперимент по угадыванию карт

Проверялась нулевая гипотеза  $H_0$ : испытуемый выбирает ответ наугад. Альтернативная гипотеза  $H_1$ : испытуемый может предсказывать цвета карт. Статистика  $t$  — число карт, цвета которых угаданы, — при справедливости нулевой гипотезы имеет биномиальное распределение с параметрами  $n = 10, p = 0.5$ , поскольку цвета только два, и они называются наугад. Вероятность правильно назвать цвета 9 и более карт:

$$P(t \geq 9 | H_0) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875.$$

То есть, если испытуемый угадывает 9 карт, получается достигаемый уровень значимости  $p \approx 0.01$ , и нулевую гипотезу можно с чистой совестью отклонить в пользу односторонней альтернативы.

В экспериментах Джозефа Райна процедуру отбора прошли 1000 человек. Девять из них угадали цвета 9 из 10 карт, еще двое угадали все 10 карт. Ни один из этих испытуемых в последующих экспериментах не подтвердил своих способностей, из чего Джозеф Райн сделал вывод, что экстрасенсам нельзя говорить о том, что они экстрасенсы, потому что от этого их способности сразу пропадают. Однако очевидно, что проблема кроется в чем-то другом.

Если принять гипотезу о том, что экстрасенсов не существует, то вероятность того, что из тысячи человек хотя бы один случайно угадает цвета 9 или 10 из 10 карт:

$$1 - \left(1 - 11 \cdot \frac{1}{2}^{10}\right)^{1000} \approx 0.9999796.$$

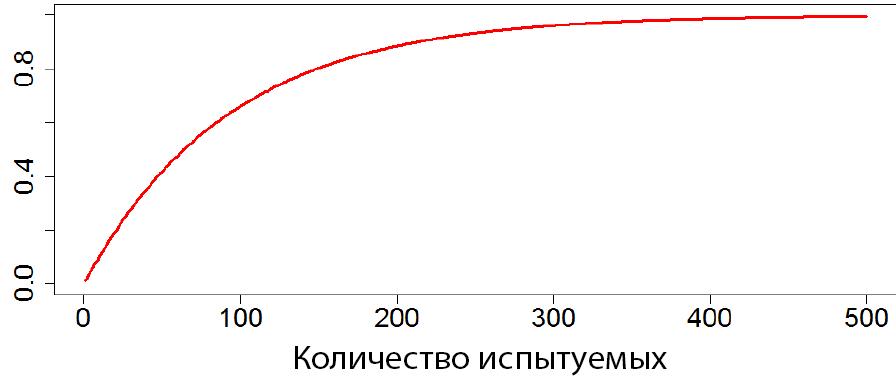


Рис. 9.2: Зависимость вероятности того, что хотя бы один испытуемый случайно угадает цвета 9 или 10 карт из 10, от количества испытуемых

На рисунке 9.2 показано, как описанная выше вероятность ведет себя в зависимости от количества испытуемых. Из графика видно, что она растет очень быстро. Уже при количестве испытуемых  $N = 100$  вероятность найти хотя бы одного экстрасенса превышает  $1/2$ . При  $N = 500$  такая вероятность уже примерно равна единице.

Тот факт, что с помощью этой статистической процедуры находятся экстрасенсы, является прекрасным примером влияния эффекта множественной проверки гипотез. При одновременной проверке большого количества гипотез вероятность совершить хотя бы одну ошибку первого рода (то есть должно отвергнуть верную нулевую гипотезу) становится очень большой.

### 9.1.2. Нейронаука

Еще один яркий пример действия эффекта множественной проверки гипотез можно найти в нейронауке.

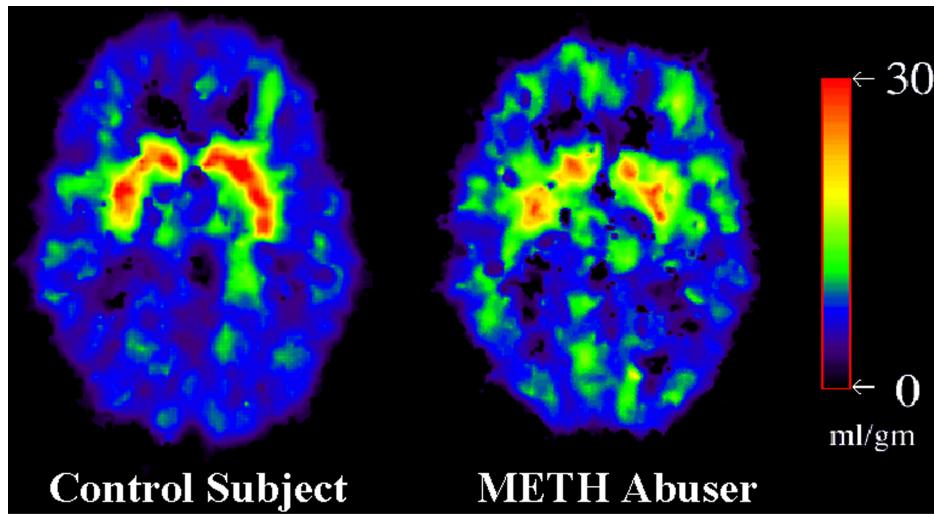


Рис. 9.3: Данные позитронно-эмиссионной томографии

Анализируются данные позитронно-эмиссионной томографии или функциональной магнитно-резонансной томографии (рисунок 9.3). Типичный дизайн такого эксперимента следующий: берётся контрольная группа испытуемых, с которыми ничего не происходит, и измеряется активность их мозга. Затем те же измерения производят с другой группой испытуемых, состояние которых каким-то образом изменили. Далее эти две выборки сравнивают, пытаясь выяснить, на какие области мозга подействовало различие между двумя экспериментальными условиями.

Решение такой задачи связано с проверкой очень большого количества гипотез. Фактически для двумерного изображения мозга гипотеза проверяется в каждой точке, для трехмерного изображения мозга, которое возникает при магнитно-резонансной томографии, гипотеза проверяется в каждом voxelе (то есть в каждом трехмерном пикселе трехмерного изображения мозга). Пикселей могут быть тысячи, voxelей могут быть миллионы. Таким образом, требуется проверить очень много гипотез. И если ничего не делать, эффект множественной проверки гипотез будет проявляться очень ярко.

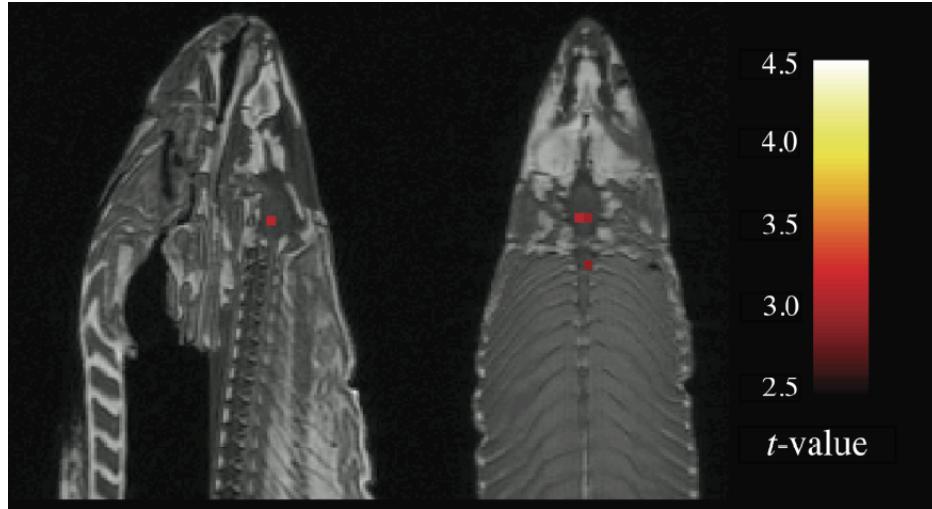


Рис. 9.4: Области мозга мёртвого лосося, в которых активность значимо изменилась при просмотре изображений

Лучше всего это демонстрирует следующий пример. Команда исследователей воспроизвела один из типичных дизайнов нейронаучных экспериментов, в котором испытуемому последовательно и много раз демонстрируются похожие стимулы, затем активность его мозга в ответ на эти стимулы сравнивается с активностью его мозга в состоянии покоя. Роль испытуемого в этом эксперименте играл мертвый лосось. В качестве стимула ему показывали картинки с изображениями людей в различных социальных ситуациях. Как видно из рисунка 9.4, задача поиска областей, которые реагируют на этот стимул, была решена успешно. В мозге лосося были выявлены области, в которых активность значимо изменилась, они на рисунке обозначены красным.

За последнее десятилетие методы анализа данных в нейронауке, в том числе методы поправки на множественную проверку гипотез, существенно улучшились. О таких методах и пойдёт речь далее.

## 9.2. Постановка

В этой части будет дана математическая постановка задачи множественной проверки гипотез. Для этого полезно вспомнить, как ставится задача однократной проверки гипотез.

### 9.2.1. Задача однократной проверки гипотез

выборка:	$X^n = (X_1, \dots, X_n)$ , $X \sim \mathbf{P}$ ;
нулевая гипотеза:	$H_0: \mathbf{P} \in \omega$ ;
альтернатива:	$H_1: \mathbf{P} \notin \omega$ ;
статистика:	$T(X^n)$ ;
нулевое распределение:	$F(x)$ ;

Таблица 9.1: Задача однократной проверки гипотез

Задача однократной проверки гипотез ставится следующим образом. Имеется некоторая выборка  $X$  объема  $n$  из неизвестного распределения  $P$ . Проверяется нулевая гипотеза  $H_0$  о распределении  $P$  против общей альтернативы  $H_1$ . Это делается с помощью статистики  $T$ , которая является функцией от выборки. Для

этой статистики известно нулевое распределение  $F(x)$ , то есть распределение при справедливости нулевой гипотезы (рисунок 9.5).

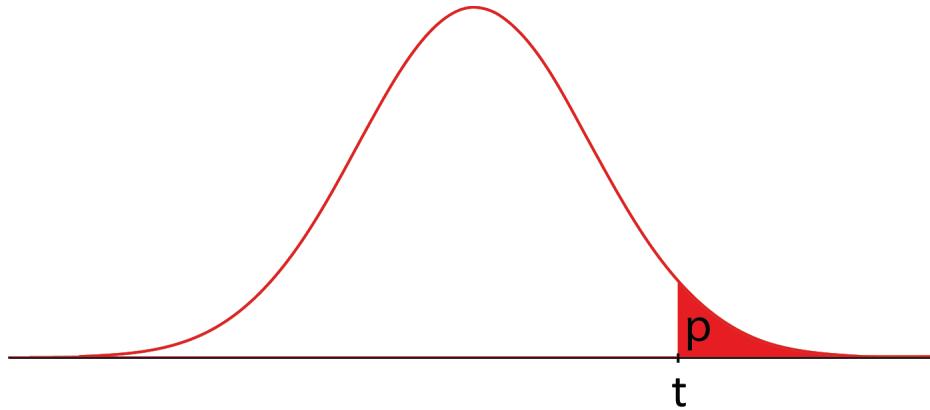


Рис. 9.5: Нулевое распределение статистики

По этому нулевому распределению, по его хвостам (разным, в зависимости от типа альтернативы) вычисляется достигаемый уровень значимости, то есть вероятность получить такое же значение статистики, какое было получено в эксперименте, или ещё более экстремальное:

$$p = \mathbf{P}(T \geq t | H_0).$$

Достигаемый уровень значимости сравнивается с порогом  $\alpha$  — уровнем значимости (типичное значение 0.05). Если достигаемый уровень значимости меньше, чем  $\alpha$ , то нулевая гипотеза отвергается в пользу альтернативы.

	$H_0$ верна	$H_0$ неверна
$H_0$ принимается	$H_0$ верно принята	Ошибка II рода
$H_0$ отвергается	Ошибка I рода	$H_0$ верно отвергнута

Таблица 9.2: Типы ошибок при проверке гипотезы

При однократной проверке гипотезы всегда есть вероятность совершить ошибку первого или второго рода (таблица 9.2). Механизм проверки гипотез построен так, что вероятность ошибки первого рода, то есть вероятность ложно отвергнуть верную нулевую гипотезу, сверху ограничена достигаемым уровнем значимости  $\alpha$ :

$$\mathbf{P}(\text{ошибка I рода}) = \mathbf{P}(T \geq t | H_0) \leq \alpha.$$

### 9.2.2. Задача множественной проверки гипотез

данные:	$\mathbf{X} = \{X_1^{n_1}, \dots, X_m^{n_m}\}, \quad X_i \sim \mathbf{P}_i;$
нулевые гипотезы:	$H_i: \mathbf{P}_i \in \omega_i;$
альтернативы:	$H'_i: \mathbf{P}_i \notin \omega_i;$
статистики:	$T_i = T(X_i^{n_i});$

Таблица 9.3: Задача множественной проверки гипотез

Теперь можно разобраться с постановкой задачи множественной проверки гипотез (таблица 9.3). Пусть имеется  $m$  выборок, каждая своего размера, и из своего распределения. Каждой выборке соответствует своя гипотеза нулевая гипотеза  $H_i$  и альтернатива  $H'_i$ . Каждая из гипотез проверяется своей статистикой  $T_i$ . Для

каждой из статистик известно свое нулевое распределение. Таким образом, можно вычислить достигаемые уровни значимости всех гипотез:

$$p_i, \quad i = 1, \dots, m.$$

Для этого вводятся следующие обозначения. Пусть  $\mathbf{M}$  — это множество индексов:

$$\mathbf{M} = \{1, 2, \dots, m\};$$

$\mathbf{M}_0$  — это множество индексов верных нулевых гипотез, пусть его мощность равна  $m_0$ :

$$\mathbf{M}_0 = \{i : H_i \text{ верна}\}, \quad |\mathbf{M}_0| = m_0.$$

Естественно, это множество неизвестно, потому что иначе не было бы смысла проверять гипотезы. Пусть  $\mathbf{R}$  — это множество индексов отвергаемых гипотез, а его мощность равна  $R$ :

$$\mathbf{R} = \{i : H_i \text{ отвергнута}\}, \quad |\mathbf{R}| = R$$

Тогда пересечение множеств  $\mathbf{R}$  и  $\mathbf{M}_0$  состоит из неверно отвергнутых гипотез. Мощность этого множества обозначается  $V$ , это есть число ошибок первого рода:

$$V = |\mathbf{M}_0 \cap \mathbf{R}|.$$

	# верных $H_i$	# неверных $H_i$	$\Sigma$
# принятых $H_i$	$U$	$T$	$m - R$
# отвергнутых $H_i$	$V$	$S$	$R$
$\Sigma$	$m_0$	$m - m_0$	$m$

Таблица 9.4: Информация о верных и неверных, принятых и отвергнутых гипотезах для случая множественной проверки гипотез

По аналогии с задачей однократной проверки гипотез можно составить таблицу  $2 \times 2$ , в которой будет стоят количество верных и неверных, принятых и отвергнутых гипотез (таблица 9.4). Из всех величин, записанных в таблице, известна только  $m$  — общее число гипотез. А единственный параметр, которым можно управлять, — это  $R$ , количество отвергаемых гипотез. При этом самая пугающая величина — это  $V$ , количество ошибок первого рода. Хочется совершать мало ошибок первого рода, но при этом единственное, что можно делать, — это перераспределять по этой таблице гипотезы из второй строки в первую. То есть, чтобы совершать мало ошибок первого рода, нужно отвергать меньше гипотез.

## 9.3. FWER. Поправка Бонферрони

Задача множественной проверки гипотез поставлена, теперь нужно её решить. Интерес представляет некоторая статистическая процедура, которая дает гарантии на значение  $V$  (таблица 9.2), оно не должно быть слишком большим.

### 9.3.1. Групповая вероятность ошибки первого рода (FWER)

Напрямую с  $V$  работать не очень удобно, поэтому, как правило, берут некоторые меры, определенные над  $V$ , и работают с ними. Одна из самых распространенных таких мер — это групповая вероятность ошибки первого рода (familywise error rate). По определению это вероятность совершить хотя бы одну ошибку первого рода:

$$\text{FWER} = P(V > 0).$$

Эту величину хочется контролировать на уровне  $\alpha$ :

$$\text{FWER} = P(V > 0) \leq \alpha.$$

То есть, хочется построить такую статистическую процедуру, что вероятность совершить хотя бы одну ошибку первого рода будет не больше, чем  $\alpha$ . Возникает вопрос, как этого добиться.

Единственный имеющийся в распоряжении инструмент — это уровни значимости  $\alpha_1, \dots, \alpha_m$ , на которых проверяются гипотезы  $H_1, \dots, H_m$ . Никаких других параметров в проверке гипотез нет. Ставится задача выбрать эти уровни так, чтобы обеспечить ограничение  $\text{FWER} \leq \alpha$ .

### 9.3.2. Поправка Бонферрони

Самый простой способ решить поставленную выше задачу — это использовать поправку Бонферрони. В методе Бонферрони достигаемые уровни значимости всех гипотез сравниваются с величиной  $\frac{\alpha}{m}$ :

$$\alpha_1 = \dots = \alpha_m = \frac{\alpha}{m}.$$

Альтернативный способ — преобразовать все достигаемые уровни значимости (p-value):

$$\tilde{p}_i = \min(1, mp_i).$$

Эти модифицированные достигаемые уровни значимости и будут сравниваться с исходным порогом  $\alpha$ :  $H_i$  отвергается при  $\tilde{p}_i \leq \alpha$ . При такой процедуре точно так же контролируется величина FWER, как и при изменении порога.

Легко показать, что метод Бонферрони контролирует групповую вероятность ошибки первого рода на уровне  $\alpha$ . Это будет единственная теорема в данном курсе.

**Теорема** Если все гипотезы  $H_i, i = 1, \dots, m$  отвергаются при  $p_i \leq \alpha/m$ , то  $\text{FWER} \leq \alpha$ .

**Доказательство** По определению Familywise error rate (FWER) — это вероятность совершил хотя бы одну ошибку первого рода:

$$\text{FWER} = \mathbf{P}(V > 0) = \mathbf{P}\left(\bigcup_{i=1}^{m_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right).$$

Вероятность объединения событий можно оценить сверху через сумму вероятностей этих событий по неравенству Буля:

$$\mathbf{P}\left(\bigcup_{i=1}^{m_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{i=1}^{m_0} \mathbf{P}\left(p_i \leq \frac{\alpha}{m}\right).$$

Далее можно воспользоваться свойством достигаемого уровня значимости:

$$\sum_{i=1}^{m_0} \mathbf{P}\left(p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m_0} \frac{\alpha}{m} = \frac{m_0}{m} \alpha.$$

$m_0 < m$ , следовательно

$$\frac{m_0}{m} \alpha \leq \alpha.$$

Исходное утверждение доказано.

### 9.3.3. Недостаток использования поправки Бонферрони

Те, кто когда-нибудь сталкивались с неравенством Буля, знают, что оценка вероятности объединения событий через сумму вероятностей этих событий очень завышенная. Действительно, чтобы получить в этом месте доказательства точное равенство, нужно вычесть вероятности всех возможных пересечений. Цепочка неравенств в доказательстве теоремы показывает, что при использовании метода Бонферрони FWER не просто меньше, чем  $\alpha$ , а намного меньше, чем  $\alpha$ . В идеале хочется, чтобы вероятность совершил хотя бы одну ошибку первого рода была в точности равна  $\alpha$ . При использовании метода Бонферрони эта вероятность ограничивается гораздо более низкой величиной, чем  $\alpha$ . Это плохо, потому что перестраховываясь в отношении ошибки первого рода, мы неизбежно совершаляем больше ошибок второго рода, то есть мощность такой статистической процедуры снижается.

### 9.3.4. Модельный эксперимент

Возьмем 50 выборок из нормального распределения  $N(1, 1)$  и еще 150 — из стандартного нормального распределения  $N(0, 1)$ . Объем всех выборок  $n = 20$ .

На каждой из этих выборок проверяется гипотеза о равенстве среднего 0:

$$H_i: \mathbb{E}X_i = 0,$$

против двусторонней альтернативы

$$H'_i: \mathbb{E}X_i \neq 0$$

с помощью критерия Стьюдента.

	# верных $H_i$	# неверных $H_i$	$\Sigma$
# принятых $H_i$	142	0	142
# отвергнутых $H_i$	8	50	58
$\Sigma$	150	50	200

Таблица 9.5: Результаты эксперимента без поправки на множественную проверку

Если не делать никакой поправки на множественную проверку, в результате получится таблица 9.5. Видно, что отвергнуты все 50 неверных гипотез, но, к сожалению, вместе с ними отвергнуты еще и 8 верных, то есть совершилось 8 ошибок первого рода.

	# верных $H_i$	# неверных $H_i$	$\Sigma$
# принятых $H_i$	150	27	177
# отвергнутых $H_i$	0	23	23
$\Sigma$	150	50	200

Таблица 9.6: Результаты эксперимента при использовании поправки Бонферрони

Если делать поправку методом Бонферрони, гипотезы из второй строчки предыдущей таблицы перераспределяются в первую, результат — таблица 9.6. В этом случае ни одна верная нулевая гипотеза не отвергается, то есть нет ни одной ошибки первого рода. Но, к сожалению, вместе с этим исчезла возможность отвергнуть больше половины неверных нулевых гипотез: из 50 удалось отвергнуть только 23. То есть за гарантии в отношении ошибки первого рода пришлось отплатить тем, что найдено меньше неверных нулевых гипотез.

## 9.4. FWER. Метод Холма

### 9.4.1. Нисходящие методы

В методе Бонферрони уровни значимости для всех гипотез выбираются одинаковыми:

$$\alpha_1 = \dots = \alpha_m = \frac{\alpha}{m}.$$

Оказывается, если значения  $a_i$  брать не одинаковыми, а разными, можно достичь лучшего результата. Для того, чтобы это сделать, необходимо использовать нисходящую процедуру множественной проверки гипотез. В общем виде она выглядит так. Из достигаемых уровней значимости составляется вариационный ряд:

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)},$$

а все гипотезы переобозначаются так, чтобы их номера соответствовали номерам достигаемых уровней значимости в этом вариационном ряду:

$$H_{(1)}, H_{(2)}, \dots, H_{(m)}.$$

Дальше нужно самый маленький достигаемый уровень значимости  $p_{(1)}$  сравнить с уровнем значимости  $\alpha_1$ . Если  $p_{(1)} \geq \alpha_1$ , то принимаются все нулевые гипотезы  $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ , и процесс останавливается.  $p_{(1)} < \alpha_1$ , то отклоняется гипотеза  $H_{(1)}$ , и процедура продолжается. На втором шаге сравниваются  $p_{(2)}$  и  $\alpha_2$ . Если  $p_{(2)} \geq \alpha_2$ , то принимаются все оставшиеся гипотезы  $H_{(2)}, H_{(3)}, \dots, H_{(m)}$ , и процедура завершается. Если нет —  $H_{(2)}$  отвергается, процедура продолжается, и т.д.

Так в общем виде выглядит нисходящая процедура множественной проверки гипотез. Процедура называется нисходящей, несмотря на то что нулевые гипотезы перебираются по возрастанию. Это немного странно и может смущать, но идея заключается в том, что нулевые гипотезы отвергаются последовательно, начиная с наиболее значимых, то есть движение происходит по убыванию значимости.

## 9.4.2. Метод Холма

Метод Холма — это нисходящая процедура множественной проверки гипотез со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha.$$

Этот метод обеспечивает безусловный контроль над FWER. Это показать немного сложнее, чем для метода Бонферрони, поэтому доказательство здесь приведено не будет.

Вместо того, чтобы сравнивать исходные достигаемые уровни значимости с модифицированными  $\alpha_i$ , можно их модифицировать и сравнивать с исходным порогом  $\alpha$ . Так выглядит формула для модифицированных достигаемых уровней значимости метода Холма:

$$\tilde{p}_{(i)} = \min (1, \max ((m-i+1)p_{(i)}, \tilde{p}_{(i-1)}))$$

Метод Холма всегда мощнее, чем метод Бонферрони, то есть, он всегда отвергает не меньше гипотез, чем метод Бонферрони, потому что его уровни значимости всегда не меньше, чем из метода Бонферрони.

## 9.4.3. Модельный эксперимент

Для демонстрации работы метода Холма можно провести такой же модельный эксперимент с 200 гипотезами, как в предыдущем разделе.

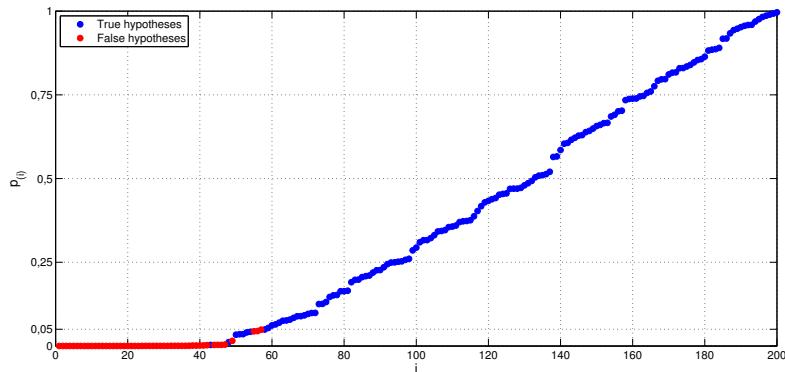
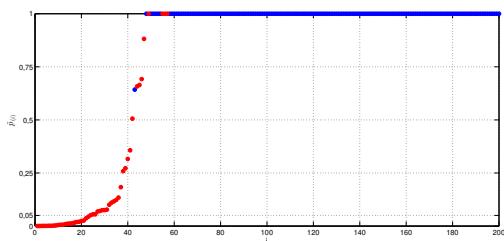
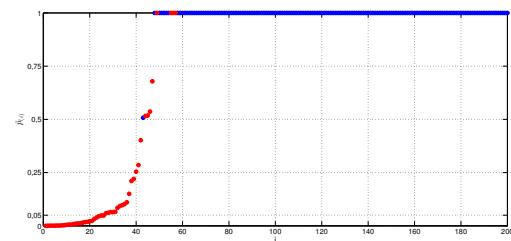


Рис. 9.6: Достигаемые уровни значимости гипотез

На графике (рисунок 9.6) показаны отсортированные достигаемые уровни значимости. По горизонтальной оси отложен номер в вариационном ряду, а по вертикальной оси — значения соответствующего достигаемого уровня значимости. Красные точки на графике — это неверные гипотезы, а синие точки — верные. Это типичный вид такого графика, на нём изображена как будто бы смесь двух треугольников: большого синего, соответствующего верным нулевым гипотезам, и маленьского красного, соответствующего неверным. В месте соединения этих двух треугольников верные и неверные гипотезы смешиваются, и задача — где-то в этом месте правильно поставить порог, чтобы обеспечить теоретические гарантии на число ошибок первого рода.



(a) Модифицированные достигаемые уровни значимости гипотез при использовании поправки Бонферрони



(b) Модифицированные достигаемые уровни значимости гипотез при использовании метода Холма

Рис. 9.7

На рисунке 9.7 слева показаны модифицированные достигаемые уровни значимости, отсортированные по неубыванию, при использовании поправки Бонферрони. Результаты были приведены в таблице 9.6. Отвергаются 23 неверные нулевые гипотезы из 50 и ни одной верной.

	# верных $H_i$	# неверных $H_i$	$\Sigma$
# принятых $H_i$	150	24	174
# отвергнутых $H_i$	0	26	26
$\Sigma$	150	50	200

Таблица 9.7: Результаты проверки гипотез при использовании метода Холма

Модифицированные достигаемые уровни значимости метода Холма показаны на рисунке 9.7 справа. Этот метод позволил отвергнуть 26 из 50 гипотез и всё ещё не совершил ни одной ошибки первого рода при этом (таблица 9.7).

С одной стороны, разница между методами Холма и Бонферрони. Метод Холма не помог случиться чуду и не отверг все неверные нулевые гипотезы. С другой стороны, этот метод позволил, не делая никаких дополнительных предположений, совершить ещё три научных открытия. Ещё три гипотезы удалось отвергнуть абсолютно бесплатно. А это уже достаточный повод, чтобы пользоваться этим методом.

## 9.5. FDR. Метод Бенджамини-Хохберга

### 9.5.1. False discovery rate

В описанных ранее поправках при множественном проверке гипотез контролировалась величина групповой вероятности ошибки, то есть ограничивалась вероятность совершить хотя бы одну ошибку первого рода:

$$\text{FWER} = P(V > 0).$$

В некоторых ситуациях, например, когда проверяются десятки тысяч или миллионы гипотез, можно допустить какое-то количество ошибок первого рода ради того, чтобы увеличить мощность процедуры и отвергнуть больше неверных гипотез, то есть совершить меньше ошибок второго рода. В таких ситуациях выгоднее использовать другую меру: не familywise error rate, а false discovery rate, ожидаемую долю ложных отклонений:

$$\text{FDR} = \mathbb{E} \left( \frac{V}{\max(R, 1)} \right).$$

Для любой фиксированной процедуры множественной проверки гипотез  $\text{FDR} \leq \text{FWER}$ . За счет этого, если контролировать FDR, а не FWER, получается более мощная процедура, поскольку она позволяет отвергать больше гипотез.

### 9.5.2. Восходящие методы

Методы, которые контролируют FDR, как правило, восходящие. В каком-то смысле это противоположность нисходящих методов (таких как метод Холма), которые рассматривались до этого.

Восходящие методы работают с тем же самым вариационным рядом достигаемых уровней значимости, что и нисходящие:

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}.$$

Отличие заключается в том, что процедура начинается с другого конца этого ряда. На первом шаге самый большой p-value  $p_m$  сравнивается с соответствующей ему константой  $\alpha_m$ . Если  $p_{(m)} \leq \alpha_m$ , то все нулевые гипотезы  $H_{(1)}, H_{(2)}, \dots, H_{(m)}$  отвергаются, и процедура останавливается. Иначе гипотеза  $H_{(m)}$  принимается, и процедура продолжается. На следующем шаге сравниваются  $p_{(m-1)}$  и  $\alpha_{m-1}$ . Если  $p_{(m-1)} \leq \alpha_{m-1}$ , то все нулевые гипотезы  $H_{(1)}, H_{(2)}, \dots, H_{(m-1)}$  отвергаются, и процедура останавливается. Иначе принимается гипотеза  $H_{(m-1)}$ , процедура продолжается. И так далее.

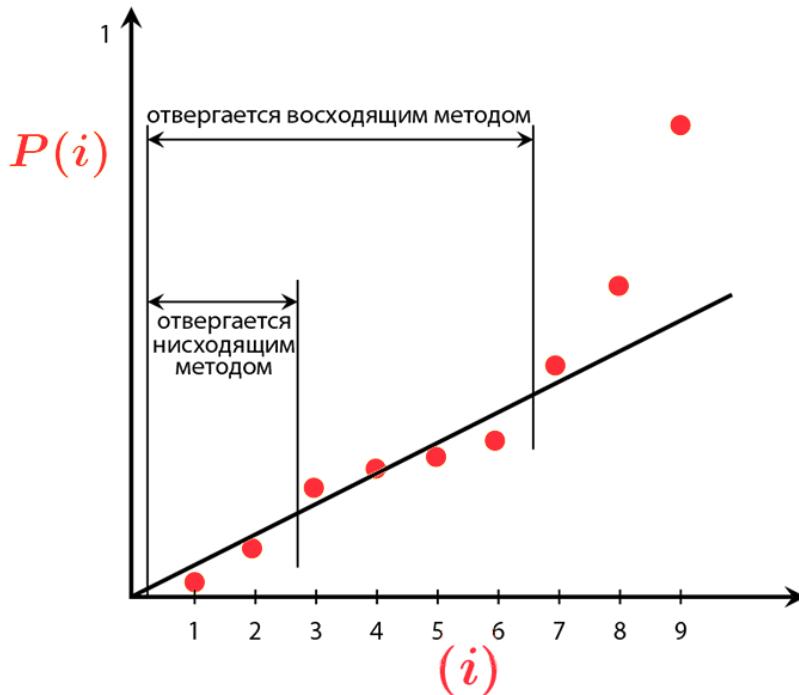


Рис. 9.8: Гипотезы, отвергаемые восходящими и нисходящими методами. Каждая точка — это гипотеза, по вертикальной оси показан соответствующий ей достигаемый уровень значимости, по горизонтальной — её номер в вариационном ряду. Линия — порог, с которым сравнивается достигаемый уровень значимости гипотез.

Если для одних и тех же  $\alpha_i$  построить восходящую и нисходящую процедуру, то восходящая процедура всегда будет отвергать не меньше гипотез, чем нисходящая. На рисунке 9.8 показано, какие гипотезы отвергаются восходящими и нисходящими методами. При использовании восходящей процедуры движение происходит от самого большого p-value к самому маленькому. В таком случае принимаются гипотезы  $H_9, H_8, H_7$ . Если используется нисходящая процедура, то наоборот, всё начинается с самого маленького p-value, и первые две гипотезы отвергаются, а оставшиеся семь — принимаются. Таким образом, в этом примере восходящая процедура отвергла в три раза больше гипотез, чем нисходящая. Это может происходить из-за того, что линия, соединяющая отсортированные достигаемые уровни значимости, может несколько раз пересекать прямую, задающую критические значения  $\alpha$ .

### 9.5.3. Метод Бенджамини-Хохберга

Для контроля над FDR чаще всего используется метод Бенджамини-Хохберга. Это восходящая процедура с уровнями значимости

$$\alpha_1 = \frac{\alpha}{m}, \dots, \alpha_i = \frac{\alpha i}{m}, \dots, \alpha_m = \alpha.$$

Крайние уровни значимости точно также же, как и в методе Холма, а вот между ними — абсолютно другие. В методе Бенджамини-Хохберга уровни значимости между  $\alpha_1$  и  $\alpha_m$  меняются линейно, в то время как в методе Холма — по гиперболе.

Модифицированные достигаемые уровни значимости для метода Бенджамини-Хохберга выглядят следующим образом:

$$\tilde{\alpha}_{(i)} = \min \left( 1, \frac{mp_{(i)}}{i}, \tilde{\alpha}_{(i+1)} \right).$$

Процедура восходящая, и каждый следующий p-value в ней не должен стать больше, чем предыдущий, поэтому берётся минимум из  $\frac{mp_{(i)}}{i}$  и  $\tilde{\alpha}_{(i+1)}$  (а также 1, поскольку это вероятность).

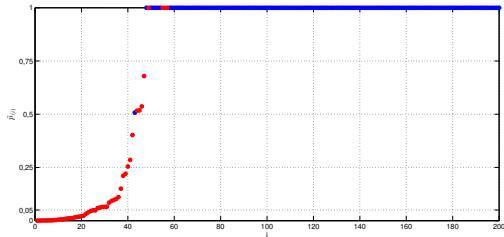
Метод Бенджамини-Хохберга обеспечивает контроль над FDR на уровне  $\alpha$  только при условии независимости статистик, которые проверяют гипотезы. Это требование достаточно сильное. Иногда его можно ослабить, и в некоторых задачах выполняется ослабленное требование. Тем не менее, важно подчеркнуть, что процедура Бенджамини-Хохберга не является универсальной и она не применима безусловно, в отличие от метода Холма.

	# верных $H_i$	# неверных $H_i$	$\Sigma$
# принятых $H_i$	148	4	152
# отвергнутых $H_i$	2	46	48
$\Sigma$	150	50	200

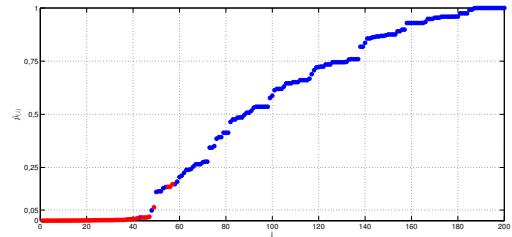
Таблица 9.8: Результаты проверки гипотез при использовании метода Бенджамини-Хохберга

#### 9.5.4. Модельный эксперимент

Давайте протестируем метод Бенджамини-Хохберга на модельных данных из предыдущего раздела.



(a) Модифицированные уровни значимости при использовании метода Холма



(b) Модифицированные уровни значимости при использовании метода Бенджамини-Хохберга

Рис. 9.9

На рисунке 9.9 показаны отсортированные модифицированные достижимые уровни значимости, полученные методами Холма и Бенджамини-Хохберга. Они отличаются довольно сильно. Метод Холма отвергает 26 неверных гипотез, при этом не совершается ни одной ошибки первого рода (таблица ??). Метод Бенджамини-Хохберга отвергает 46 неверных гипотез (таблица 9.8), при этом совершается две ошибки первого рода ( $\frac{2}{48} \approx 0.04 < 0.05$ ). Метод Бенджамини-Хохберга применим, так как в модельном эксперименте выборки генерировались независимо.

Итак, если контролировать FDR вместо FWER, допускается больше ошибок первого рода, но за счет этого можно критически увеличить количество отвергаемых неверных нулевых гипотез. Метод Бенджамини-Хохберга используется повсеместно, несмотря на то, что он работает далеко не всегда. Очень часто его применяют без проверки необходимого условия корректности, так делать не стоит.

### 9.6. Анализ подгрупп

Эффект множественной проверки гипотез ярко проявляется при анализе подгрупп. Для примера можно рассмотреть следующее исследование, в котором принимают участие 1073 пациента с ишемической болезнью сердца<sup>1</sup>. Их делят на 2 подгруппы (в зависимости от типа лечения) и исследуют взаимосвязь между выживаемостью и типом лечения. Требуется понять, какой их двух типов лечения лучше.

Важные факторы, которые влияют на выживаемость при ишемической болезни сердца, — это число пораженных артерий, (может быть 1, 2 или 3), и тип сокращений левого желудочка (нормальный и аномальный). В таких ситуациях исследователи часто хотят посмотреть на сравнительную эффективность типов лечения отдельно во всех подгруппах по уровням важных факторов. В данном случае два фактора порождают 6 подгрупп, в каждой из них сравнивается выживаемость пациентов по двум типам лечения.

<sup>1</sup>Ссылка на указанное исследование: Lee K.L., McNeer J.F., Starmer C.F., Harris P.J., Rosati R.A. (1980). Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61(3), 508–515.

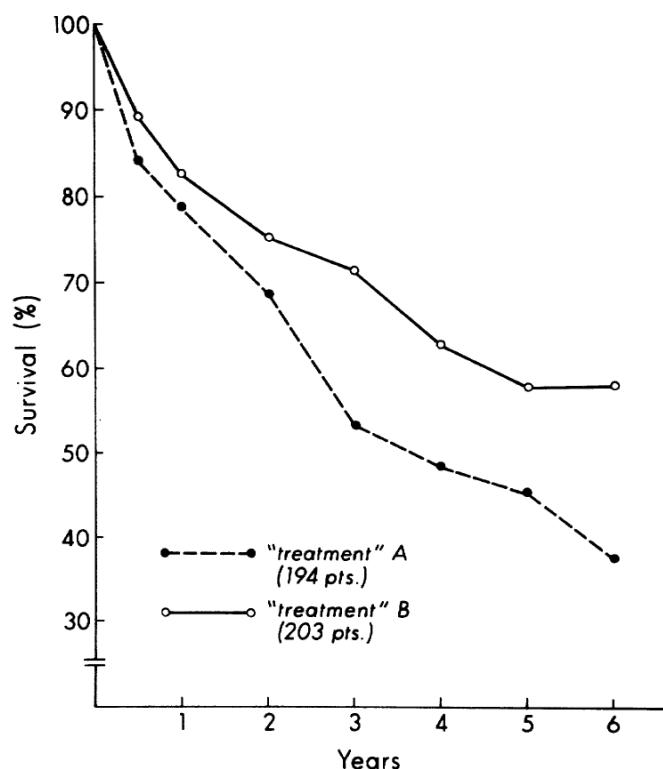


Рис. 9.10: Выживаемость людей в одной из подгрупп при разных типах лечения

В одной из 6 подгрупп были обнаружены значимые различия в выживаемости пациентов при лечении первого типа и второго. На рисунке 9.10 показаны кривые выживаемости пациентов этих подгрупп. По ним видно, что в группе лечения А после 6 лет наблюдений выжило меньше 40% пациентов, а в группе, соответствующей лечению В, — в районе 60%. Это различие статистически значимо. Кажется, что для пациентов с таким числом пораженных артерий и с таким типом сокращения левого желудочка лечение В действительно существенно эффективнее.

На самом деле эти два лечения отличаются только названием, и две группы пациентов лечились абсолютно одинаково. Эта статья была написана с целью показать необходимость поправки на множественную проверку гипотез при анализе подгрупп. Действительно, при сравнении кривых выживаемости во всех 6 подгруппах, проверяются 6 абсолютно независимых гипотез, и возникает эффект множественной проверки. Если подгрупп достаточно много, в результате будут всегда получаться какие-то значимые отклонения. Есть и исследование, в котором такая ошибка в анализе подгрупп была совершена на самом деле, — это исследование 2008 года, в котором изучалась связь между употреблением кофеина и риском возникновения рака груди<sup>2</sup>. В этой статье всего было около 50 разных подгрупп, по самым разным уровням различных факторов. В частности, было показано, что употребление более чем четырех чашек кофе в день связано с увеличением риска злокачественного рака груди (достигаемый уровень значимости  $p = 0.08$ ). Это больше, чем стандартный уровень значимости 0.05, но меньше чем либеральный уровень значимости 0.1. Кроме того, потребление кофеина связано с увеличением риска возникновения эстроген- и прогестерон- независимых опухолей, а также опухолей размером больше 2 сантиметров (достигаемый уровень значимости  $p = 0.02$ ). Еще одно открытие: потребление кофе без кофеина связано со снижением риска возникновения рака груди у женщин в постменопаузе, принимающих гормоны (достигаемый уровень значимости  $p = 0.02$ ).

Ясно, что за счет большого количества рассматриваемых подгрупп можно всегда получить какие-то значимые отклонения, если не делать поправку на множественную проверку. Какие-то из этих открытий с большой вероятностью окажутся ложными. В каком-то смысле это напоминает переобучение: оценивается эффективность лечения в разных подгруппах в зависимости от каких-то признаков пациента, и если эти признаки слишком сложные и их слишком большое количество, то происходит переобучение под анализируемую выборку. В качестве экстремального примера такого переобучения можно вспомнить цитату из Галена, II века

<sup>2</sup>Ссылка на указанное исследование: Ishitani K., Lin J. (2008). Caffeine consumption and the risk of breast cancer in a large prospective cohort of women. Archives of Internal Medicine, 168(18), 2022–2031.

до нашей эры: «*Все больные, принявшие это средство, вскоре выздоровели, за исключением тех, кому оно не помогло — они умерли. Отсюда очевидно, что средство помогает во всех случаях, кроме безнадежных*».

	Контроль (100)	Больные (100)	$p$
Мутация	1 из 100	8 из 100	0.0349
Фамилия начинается с гласной	36 из 100	40 из 100	0.6622

Таблица 9.9: Данные гипотетического эксперимента

В заключение обсуждения эффекта множественной проверки гипотез давайте рассмотрим ещё один гипотетический пример. Пусть есть 100 больных людей и 100 здоровых, и хочется понять, есть ли связь между болезнью и какой-то мутацией. В контрольной выборке из 100 человек мутация есть у одного, а в выборке больных — у 8 (таблица 9.9). По всей видимости, эта мутация достаточно редкая. Если сравнить доли людей с мутацией в выборках больных и здоровых, получится достигаемый уровень значимости  $p = 0.03$ , и гипотеза об отсутствии связи между мутацией и болезнью отвергается.

Пусть теперь выдвигается еще одна гипотеза: наличие заболевания связано с тем, с гласной или согласной буквы у пациентов начинаются фамилии. В контрольной выборке здоровых людей у 36 человек фамилия начинается с гласной буквы, а в выборке больных — у 40 из 100 (таблица 9.9). При сравнении этих долей биномиальным критерием получается достигаемый уровень значимости 0.66, гипотеза отклонена не будет. Проблема, однако, заключается в том, что теперь в исследовании проверяются две гипотезы, и необходимо делать поправку на множественность этой проверки.

Какой бы при этом ни использовался метод поправки, будь то метод Бонферрони, Холма или Бенджамина Хохберга, самый маленький достигаемый уровень значимости во всех них сравнивается с  $\frac{\alpha}{m}$ . Таким образом, если требуется обеспечить контроль над какой-то мерой числа ошибок первого рода на уровне 0.05, нужно сравнивать самое маленькое значение достигаемого уровня значимости с 0.025. Самый маленький достижимый уровень значимости в этом исследовании  $p = 0.03$ . Получается, эта нелепая гипотеза, которая была введена в исследование, замаскировала, возможно, неверную нулевую гипотезу, связанную с мутацией.

Отсюда вытекает рецепт лучшего способа борьбы с эффектом множественной проверки гипотез: проверять меньше гипотез. Необходимо до начала исследования подумать, какие из возможных гипотез на самом деле не представляют интереса, и отказаться от их рассмотрения. За счет этого появится возможность сделать более либеральную поправку на множественность и отвергнуть больше действительно неверных гипотез, совершив больше действительно интересных открытий. Важно, что такая фильтрация гипотез должна осуществляться именно до сбора данных. Если выбрасывать гипотезы уже после того, как стали известны достижимые уровни значимости, возникнет эффект переобучения.

# Урок 10

## Регрессия

### 10.1. Взаимосвязь нескольких признаков

В этом уроке будут рассмотрены методы, позволяющие проанализировать взаимосвязь между одним признаком и большим количеством других.

#### 10.1.1. Пример исследования

Пусть исследователей интересует вопрос, влияет ли употребление алкоголя на успеваемость школьников. Лучший способ это проверить — провести эксперимент. Набирается случайная выборка школьников, и каждому из них назначается случайная еженедельная доза алкоголя. По окончании учебного года требуется измерить корреляцию между назначенной дозой и успеваемостью школьников.

Этот эксперимент идеален. Поскольку доза назначается случайно, выборка автоматически балансируется по всем возможным типам школьников, которые только могут быть. Этот эксперимент мог быть лучше только, если бы школьники сами не знали, какое количество алкоголя они принимают, но это достаточно сложно обеспечить.

Существенный недостаток этого эксперимента заключается в том, что его никогда не дадут провести — это неэтично. Такие ситуации возникают достаточно часто. Не получится исследовать взаимосвязь между уровнем насилия в видеоиграх и агрессивностью детей в жизни, поскольку нельзя заставить детей играть в видеоигры с высоким уровнем насилия какое-то продолжительное количество времени, если они сами этого не хотят. Иногда проведение эксперимента не только неэтично, а попросту невозможно. Например, если хочется понять, как влияет средняя дневная температура на вероятность возникновения лесного пожара, не существует никакого способа провести эксперимент, потому что средней дневной температурой в лесу управлять нельзя.

Единственное, что остается делать в условиях, когда нельзя провести эксперимент, — это использовать обзёрвационные данные, то есть данные, которые собраны путем наблюдения за выборкой. В задаче исследования успеваемости школьников можно, например, взять данные по 633 ученикам старших классов двух португальских школ, для которых известно большое количество разных демографических показателей, в том числе, успеваемость. В частности, среди всех показателей есть уровень потребления алкоголя по выходным и финальная оценка по португальскому языку.

На рисунке 10.1 изображены эти два показателя для 633 учащихся. Видно, что эти две величины друг с другом отрицательно скоррелированы. Эта корреляция значима. Возникает вопрос: значит ли это, что потребление алкоголя влияет на успеваемость старшеклассников, или что чем больше алкоголя они потребляют, тем хуже они учатся. Чтобы точнее ответить на него, можно использовать еще 29 признаков, которые есть в наборе данных о школьниках. Эти признаки потенциально влияют на успеваемость гораздо сильнее, чем употребление алкоголя. Например, возраст учеников или доход их родителей могут определять успеваемость гораздо более явно. Теперь требуется узнать, останется ли у потребления алкоголя предсказательная сила при учёте остальных признаков, и можно ли утверждать, что потребление алкоголя вызывает снижение оценки по португальскому языку, то есть ли причинно-следственная связь между этими двумя признаками.

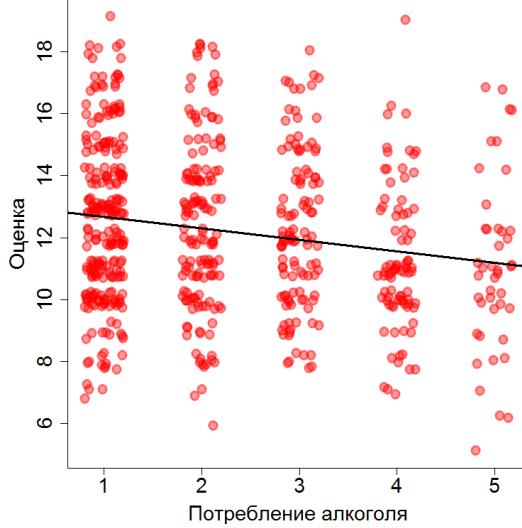


Рис. 10.1: Данные об уровне употребления алкоголя по выходным и финальной оценке по португальскому языку у учащихся двух португальских школ

### 10.1.2. Линейная регрессия

Оказывается, на такие вопросы можно отвечать с помощью линейной регрессии. Задача линейной регрессии: есть  $n$  объектов, на которых измерены значения  $k$  признаков  $x_1, \dots, x_k$ , и, кроме того, для них известно значение отклика  $y$ . Требуется найти вектор констант  $\beta$  такой, что:

$$y \approx \beta x.$$

При построении регрессии строится наилучшее линейное по  $x$  приближение условного математического ожидания  $y$  при таких  $x$ :

$$\mathbb{E}(y|x) \approx \beta_0 + \sum_{j=1}^k \beta_j x_j.$$

В линейной регрессии коэффициент  $\beta_j$  показывает, насколько в среднем увеличивается отклик  $y$ , если  $x_j$  увеличивается на 1, а все остальные  $x$  зафиксированы. Таким образом, используя регрессию, можно изолировать эффект интересующей переменной и посмотреть на него отдельно. Иногда этот эффект можно даже интерпретировать как причинно-следственную связь, при выполнении некоторых специальных условий. Строить обычную линейную регрессию очень просто. Однако если по построенной модели хочется делать какие-то выводы с использованием статистических методов, необходимо приложить дополнительные усилия. Именно этому и будет посвящен урок.

## 10.2. Свойства решения задачи

### 10.2.1. Задача линейной регрессии

Итак, решается задача линейной регрессии:

$$\mathbb{E}(y|x) \approx \beta_0 + \sum_{j=1}^k \beta_j x_j.$$

Для того, чтобы больше не думать про коэффициент  $b_0$ , можно добавить в матрицу объекты-признаки  $X$  единичный столбец:

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Теперь эта матрица размера  $n \times (k + 1)$ .

Задача регрессии будет решаться методом наименьших квадратов без использования регуляризаторов:

$$\|y - X\beta\|_2^2 \rightarrow \min_{\beta}.$$

Точное решение этой задачи известно,  $\hat{\beta}$  выражается аналитически:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Можно посчитать и  $\hat{y}$ , то есть предсказание модели на объектах, на которых она обучается:

$$\hat{y} = X (X^T X)^{-1} X^T y.$$

Чтобы найти качество решения, полученного методом наименьших квадратов, определим величину TSS (Total Sum of Squares) — разброс  $y$  относительно своего среднего:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2.$$

Оказывается, что этот разброс можно поделить на две части:

$$TSS = ESS + RSS.$$

Одна из частей, объясненная сумма квадратов, — это сумма квадратов отклонений среднего  $y$  от предсказанных  $\hat{y}$ :

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Вторая часть, остаточная сумма квадратов, — это сумма квадратов отклонений предсказанных  $\hat{y}$  от их истинных значений:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

По этим величинам, ESS и TSS, можно составить меру  $R^2$ , которая называется коэффициентом детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

По сути, это доля объясненной дисперсии отклика во всей дисперсии отклика.

### 10.2.2. Предположения МНК

Для того, чтобы решение метода наименьших квадратов обладало интересующими нас свойствами, необходимо сделать следующие предположения.

- Предполагается, что истинная модель  $y$  действительно линейна:

$$y = X\beta + \varepsilon,$$

где  $\varepsilon$  — это какая-то ошибка.

- Предполагается, что наблюдения, по которым оценивается модель, случайны, то есть объекты дают независимую выборку наблюдений  $(x_i, y_i)$ .
- Предполагается, что матрица  $X$  — матрица полного столбцового ранга:

$$\text{rank } X = k + 1,$$

то есть ни один из признаков не должен являться линейной комбинацией других. Поскольку среди столбцов есть константа, никакой из признаков в выборке не должен быть константой.

- Предполагается, что ошибка случайна:

$$\mathbb{E}(\varepsilon | x) = 0.$$

Уже из этих четырех предположений можно вывести полезное свойство оценок, получаемых методом наименьших квадратов. Если они выполняются, то оценки  $\hat{\beta}$  являются несмешенными

$$\mathbb{E}\hat{\beta}_j = \beta_j$$

и состоятельными оценками истинных  $\beta$ :

$$\forall \gamma > 0 \lim_{n \rightarrow \infty} P\left(\left|\beta_j - \hat{\beta}_j\right| < \gamma\right) = 1.$$

К четырем предположениям можно добавить еще пятое — предположение гомоскедастичности ошибок.

- Предполагается, что дисперсия ошибки не зависит от значений признака:

$$\mathbb{D}(\varepsilon | x) = \sigma^2.$$

Вместе эти пять предположений называются предположениями Гаусса-Маркова. Теорема Гаусса-Маркова утверждает, что если эти предположения выполняются, то МНК-оценки имеют наименьшую дисперсию в классе всех оценок  $\beta$ , линейных по  $y$ . То есть оценки методом наименьших квадратов при выполнении этих пяти предположений в каком-то смысле являются оптимальными.

Из сделанных предположений вытекает следующее выражение для дисперсии МНК-оценок:

$$\mathbb{D}(\hat{\beta}_j) = \frac{\sigma^2}{TSS_j(1 - R_j^2)},$$

то есть:

- чем больше  $\sigma^2$ , тем больше дисперсия  $\hat{\beta}_j$ ;
- чем больше вариация значений  $x_j$  в выборке, тем меньше дисперсия  $\hat{\beta}_j$ ;
- чем лучше признак  $x_j$  объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия  $\hat{\beta}_j$ .

По предположению о полноте столбцового ранга матрицы  $X$  коэффициент детерминации  $R_j^2 < 1$ , но, тем не менее, может быть  $R_j^2 \approx 1$ . Такая ситуация называется мультиколлинеарностью.

В матричном виде выражение для дисперсии вектора  $\hat{\beta}$  выглядит вот так:

$$\mathbb{D}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Если матрица  $X$  содержит столбцы, которые почти линейно зависимы, то матрица  $X^T X$  будет плохо обусловлена. При обращении этой матрицы будет получаться численная неустойчивость, поэтому дисперсия оценок  $\hat{\beta}_j$  будет велика.

Следует обратить внимание, что определение «мультиколлинеарности» не включает случай, когда столбцы полностью линейно зависимы. Мультиколлинеарность — это близкая к линейной зависимость признаков.

К 5 предположениям Гаусса-Маркова можно добавить еще одно, предположение о нормальности ошибки  $\varepsilon$ :

$$\varepsilon | x \sim N(0, \sigma^2).$$

Это эквивалентно следующей записи:

$$y | x \sim N(x\beta, \sigma^2).$$

Если выполняются эти 6 предположений, то оценки, даваемые методом наименьших квадратов, совпадают с оценками максимального правдоподобия. Это открывает доступ к прекрасным свойствам оценок максимального правдоподобия. Из этих 6 предположений вытекает, что оценки метода наименьших квадратов, во-первых, имеют наименьшую дисперсию среди всех несмешенных оценок  $\beta$ . Во-вторых, имеют нормальное распределение:

$$N\left(\beta, \sigma^2 (X^T X)^{-1}\right).$$

Далее, дисперсию шума  $\sigma^2$  можно оценить с помощью RSS:

$$\hat{\sigma}^2 = \frac{RSS}{n - k - 1}.$$

Кроме того, отношение RSS к истинной дисперсии будет распределено по  $\chi^2$ :

$$\frac{RSS}{\sigma^2} \sim \chi_{n-k-1}^2.$$

Наконец, следующее очень сильное свойство. Для любого вещественного вектора  $c$  длины  $k + 1$  справедливо следующее утверждение:

$$\frac{c^T (\beta - \hat{\beta})}{\hat{\sigma} \sqrt{c^T (X^T X)^{-1} c}} \sim St(n - k - 1).$$

### 10.2.3. Последствия

Если выполняются описанные предположения, то можно строить доверительные интервалы для коэффициентов  $\beta_j$ , доверительные интервалы для среднего отклика  $\mathbb{E}(y|x)$  и предсказательные интервалы для значения  $y|x$ . Далее будет описано, как всё это делать.

## 10.3. Интервалы и гипотезы

### 10.3.1. Построение доверительных и предсказательных интервалов

В предыдущей части утверждалось, что, если выполняются шесть необходимых предположений, из этого вытекают очень полезные свойства. Эти свойства можно немедленно использовать. Во-первых,  $100(1 - \alpha)\%$  доверительный интервал для дисперсии шума  $\sigma^2$  можно построить через отношение RSS к квантилям распределения  $\chi^2$ :

$$\frac{RSS}{\chi_{n-k-1, 1-\alpha/2}^2} \leq \sigma^2 \leq \frac{RSS}{\chi_{n-k-1, \alpha/2}^2}.$$

Во-вторых, чтобы построить доверительные интервалы для коэффициента  $\beta_j$ , можно использовать последнее утверждение (о распределении Стьюдента), и в качестве вектора  $c$  выбрать вектор, состоящий из всех нулей, в котором на  $j$  позиции стоит 1  $c = (0 \dots 0|1|0 \dots 0)$ . Тогда  $100(1 - \alpha)\%$  доверительный интервал для коэффициента  $\beta_j$  задаётся следующим образом:

$$\hat{\beta}_j \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{(X^T X)_{jj}^{-1}}.$$

Чтобы построить доверительный интервал для математического ожидания отклика  $y$  на новом объекте, задаваемом вектором  $x_0$ , в качестве вектора  $c$  можно использовать  $x_0$ .  $100(1 - \alpha)\%$  доверительный интервал для  $\mathbb{E}(y|x_0)$  готов:

$$x_0^T \hat{\beta} \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}.$$

Чтобы построить предсказательный интервал для значения отклика на этом же самом объекте  $y(x_0) = x_0^T \beta + \varepsilon(x_0)$ , необходимо дополнительно учесть ещё дисперсию ошибки:

$$x_0^T \hat{\beta} \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}.$$

Формула для предсказательного интервала отличается от формулы для доверительного интервала условного математического ожидания только единицей, стоящей под корнем.

### 10.3.2. Критерий Стьюдента

Для проверки гипотезы

$$H_0: \beta_j = 0$$

можно использовать Т-критерий Стьюдента (таблица 10.1). Гипотеза о равенстве нулю коэффициента  $\beta_j$  означает, что признак  $x_j$  не влияет на отклик  $y$ .

нулевая гипотеза:	$H_0: \beta_j = 0;$
альтернатива:	$H_1: \beta_j < \neq > 0;$
статистика:	$T = \frac{\hat{\beta}_j}{\sqrt{\frac{\text{RSS}}{n-k-1} (X^T X)_{jj}^{-1}}};$
нулевое распределение:	$T \sim St(n - k - 1).$

Таблица 10.1: Описание t-критерия Стьюдента

Если справедлива нулевая гипотеза, статистика данного критерия имеет распределение Стьюдента с числом степеней свободы  $n - k - 1$  (рисунок 10.2).

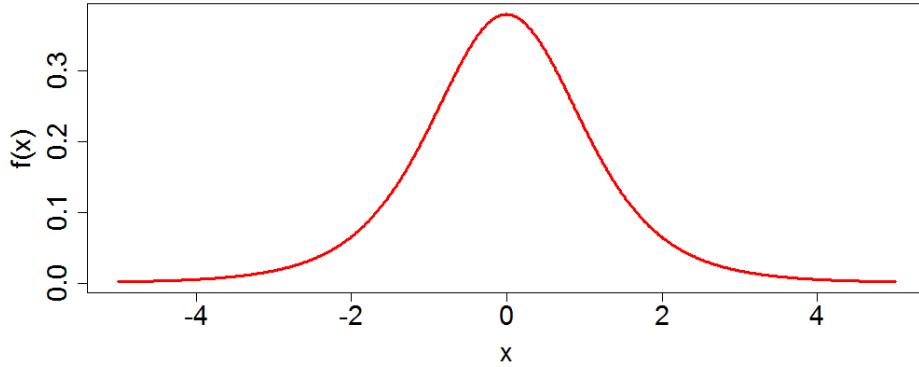


Рис. 10.2: Распределение Стьюдента

### 10.3.3. Пример

Пусть есть 12 испытуемых, и  $x$  — это результат прохождения ими составного теста на скорость реакции, а  $y$  — это их результат на симуляторе транспортного средства. Значение  $y$  получать долго и дорого, поэтому поставлена задача предсказания  $y$  по  $x$ . Необходимо понять, можно ли это делать.

Строится регрессионная модель

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Затем проверяется, что переменная  $x$  значима для предсказания  $y$ . Нулевая гипотеза

$$H_0: \beta_1 = 0$$

против двусторонней альтернативы

$$H_1: \beta_1 \neq 0$$

критерием Стьюдента отвергается. Достигаемый уровень значимости  $p = 2.2021 \times 10^{-5}$ .

### 10.3.4. Критерий Фишера

Для проверки гипотезы о том, что сразу несколько коэффициентов модели равны 0, будет использоваться не критерий Стьюдента, а критерий Фишера (таблица 10.2).

Матрицу объекты-признаки  $X$  нужно поделить на две части:

$$X_{n \times (k+1)} = \begin{pmatrix} X_1 & X_2 \\ n \times (k+1-k_1) & n \times k_1 \end{pmatrix}.$$

нулевая гипотеза:	$H_0: \beta_2 = 0;$
альтернатива:	$H_1: H_0$ неверна;
статистика:	$RSS_r = \ y - X_1\beta_1\ _2^2,$
	$RSS_{ur} = \ y - X\beta\ _2^2,$
	$F = \frac{(RSS_r - RSS_{ur})/k_1}{RSS_{ur}/(n-k-1)};$
нулевое распределение:	$F \sim F(k_1, n - k - 1).$

Таблица 10.2: Описание критерия Фишера

В первую часть  $X_1$  помещаются все признаки, которые мы хотим оставить в модели (константу нужно обязательно оставить там же). Во вторую часть  $X_2$  переносят все признаки, для которых требуется проверить гипотезу о значимости влияния на отклик. За  $\beta_1$  и  $\beta_2$  обозначаются соответствующие куски вектора параметров модели  $\beta$ :

$$\beta^T = \begin{pmatrix} \beta_1^T & \beta_2^T \end{pmatrix}^T.$$

Проверяется нулевая гипотеза о том, что все компоненты вектора  $\beta_2$  равны нулю. Это делается с помощью статистики  $F$ , которая определяется через соотношение двух RSS, где  $RSS_r$  — это RSS сокращённой модели (модель, в которой признаки из  $X_2$  вообще не используются), а  $RSS_{ur}$  — это RSS полной модели, в которой есть признаки и  $X_1$ , и  $X_2$ .

Если нулевая гипотеза справедлива, то такая статистика  $F$ , составленная из двух RSS, имеет распределение Фишера с числом степеней свободы  $k_1$  и  $n - k - 1$  (рисунок 10.3).

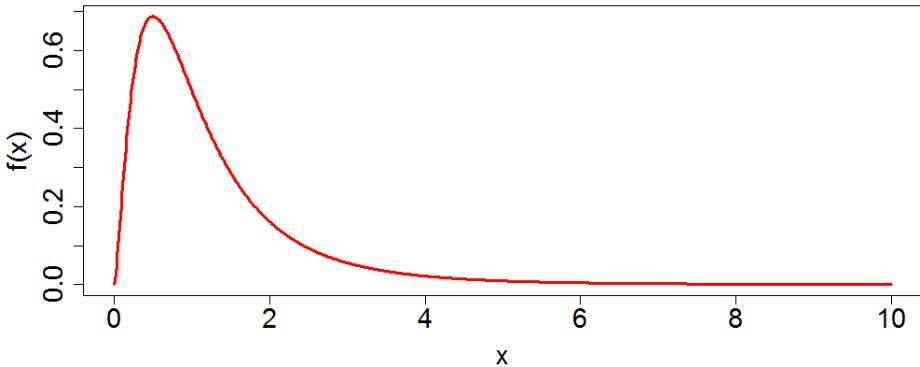


Рис. 10.3: Распределение Фишера

### 10.3.5. Пример

Пусть есть 1191 детей, для которых известны: их вес при рождении *weight*, среднее число сигарет, которые выкуривала мать за один день беременности *cigs*, номер ребёнка у матери *parity*, среднемесячный доход семьи *inc*, а также длительность получения образования в годах матерью *med* и отцом *fed*. По этим данным строится модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon.$$

Требуется проверить гипотезу о том, что образование родителей не является значимым предиктором при предсказании веса ребёнка при рождении. Для этого используется критерий Фишера. Проверяется нулевая гипотеза

$$H_0: \beta_4 = \beta_5 = 0$$

против общей альтернативы:

$$H_1: H_0 \text{ неверна.}$$

Критерий Фишера даёт достигаемый уровень значимости  $p = 0.2421$ , то есть данные не позволяют отклонить нулевую гипотезу.

### 10.3.6. Критерии Фишера и Стьюдента

Если  $k_1 = 1$ , то критерий Фишера даёт абсолютно такой же достигаемый уровень значимости, какой дал бы критерий Стьюдента для этого же самого признака при использовании двусторонней альтернативы.

Если  $k_1 > 1$ , могут возникать разные неоднозначные ситуации. Например, критерий Фишера может говорить, что гипотеза незначимости признаков  $X_2$  отвергается. При этом критерий Стьюдента не может отвергнуть никакую из гипотез о признаках, лежащих внутри  $X_2$ . Получается странная ситуация: все вместе признаки значимо определяют отклик, но отдельно ни один из них значимо на отклик не влияет. Это можно объяснить двумя способами. Во-первых, такая ситуация может возникать, если отдельные признаки из  $X_2$  недостаточно хорошо объясняют отклик, но их совокупный эффект при прогнозировании  $y$  значим. Во-вторых, признаки из  $X_2$  могут быть мультиколлинеарны. Мультиколлинеарность приводит к численной неустойчивости критериев Стьюдента и Фишера, поэтому их достигаемые уровни значимости могут быть неадекватными.

Теперь противоположная ситуация. Пусть критерий Фишера не отвергает гипотезу о незначимости признаков из  $X_2$ , а критерий Стьюдента по отдельным компонентам  $X_2$  какие-то из гипотез отвергает. То есть все вместе признаки незначимы, а какие-то из них по отдельности оказываются значимыми. Для этого тоже может быть два объяснения. Первый вариант: незначимые признаки из  $X_2$  маскируют влияние значимых. Второй вариант: значимость отдельных признаков из  $X_2$  — это результат эффекта множественной проверки гипотез. Действительно, критерии Фишера проверяют всего одну гипотезу, а критерии Стьюдента проверяют целую серию из  $k_1$  гипотез, и какие-то из них могут отклониться просто случайно.

### 10.3.7. Критерий Фишера для проверки гипотезы о незначимости всех признаков

Критерий Фишера имеет особенный вид (таблица 10.3), если требуется проверить гипотезу о том, что все признаки  $X$  для предсказания  $y$  не нужны, то есть лучшее предсказание для  $y$  — это константа.

нулевая гипотеза:	$H_0: \beta_1 = \dots = \beta_k = 0;$
альтернатива:	$H_1: H_0$ неверна;
статистика:	$F = \frac{R^2/k}{(1-R^2)/(n-k-1)};$
нулевое распределение:	$F \sim F(k, n - k - 1).$

Таблица 10.3: Описание критерия Фишера для проверки гипотезы о незначимости всех признаков

Нулевое распределение статистики точно такое же, как и раньше, — это распределение Фишера (рисунок 10.3).

**Пример.** В предыдущей задаче о весе детей при рождении можно проверить гипотезу о том, что построенная модель вообще имеет хоть какой-то смысл. Проверяем гипотезу о том, что все  $\beta$  равны нулю:

$$H_0: \beta_1 = \dots = \beta_5 = 0$$

против общей альтернативы

$$H_1: H_0 \text{ неверна.}$$

Критерием Фишера нулевая гипотеза уверено отвергается, достигаемый уровень значимости  $p = 6 \times 10^{-9}$ .

## 10.4. Проверка предположений

В этой части пойдёт речь о том, как проверять шесть предположений, лежащих в основе всей статистической машины, с помощью которой проверяется значимость коэффициентов регрессии.

### 10.4.1. Линейность отклика

Первое предположение — это предположение о линейности отклика. Утверждается, что  $y$  в действительности представляет собой линейную комбинацию  $X$  с какой-то случайной ошибкой  $\varepsilon$ :

$$y = X\beta + \varepsilon.$$

Естественно, это предположение в точности не выполняется никогда. Трудно ожидать, что отклик  $y$  в действительности — это линейная комбинация рассматриваемых признаков  $x$ . Линейная модель, как и все остальные, неверна, но очень полезна, и кроме того, устойчива к небольшим отклонениям от линейности. Поэтому единственное, что требуется проверить, — это нет ли каких-то огромных отклонений от линейности  $y$  по  $x$ . Чтобы убедиться в отсутствии больших отклонений от линейности, нужно анализировать остатки:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n,$$

где  $y$  — это истинные значения, а  $\hat{y}$  — предсказываемые.

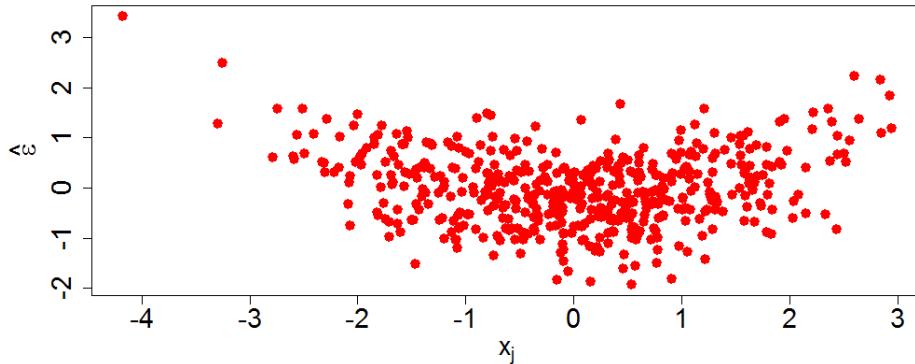


Рис. 10.4: График для проверки отклонений от линейности. По горизонтальной оси — значения признаков, по вертикальной — остатки.

Для остатков необходимо построить графики. По горизонтальной оси откладывается значение каждого из признаков  $x_j$ , по вертикальной оси — остатки, и нужно смотреть, как выглядит получающееся облако точек. Если, например, оно выглядит как на рисунке 10.4, представляет собой какую-то параболу, то, скорее всего это значит, что отклик  $y$  зависит от квадрата признака  $x_j$ . Такую зависимость можно учесть, просто добавив в матрицу  $X$  столбец, соответствующий  $x_j^2$ .

На таком графике можно обнаруживать и другие осмысленные функциональные зависимости. Если такие зависимости видны, нужно просто добавить в матрицу  $X$  соответствующий столбец.

### 10.4.2. Случайность выборки

Следующее предположение — это предположение о случайности выборки. Требуется, чтобы выборка была независимой и одинаково распределенной. Это предположение может нарушаться несколькими способами. Первый способ более тяжелый: если объекты, на которых измерены признаки и отклик, зависимы, то всё плохо: дисперсии ошибки и коэффициентов недооцениваются, и все статистические критерии, которые на этом основаны, перестают работать корректно.

Еще это предположение может нарушаться, если выборка отобрана из генеральной совокупности не случайно, а каким-то образом отфильтрована. Фильтровать генеральную совокупность по какому-то признаку  $z$  можно только в случае, если

$$\mathbb{E}(y|x, z) = \mathbb{E}(y|x),$$

то есть  $z$  не добавляет никакой новой информации об  $y$ .

Если выборка отфильтровывалась как-то иначе, например, просто по одному из признаков, содержащихся в  $x$ , то выводы, построенные по такой модели, можно распространять только на отфильтрованную генеральную совокупность. Например, если в выборке испытуемые только младше 50 лет, то нельзя ничего сказать об испытуемых в генеральной совокупности, которым больше 50 лет.

### 10.4.3. Полнота ранга $X$

Следующее предположение: матрица  $X$  должна иметь полный столбцовый ранг, то есть

$$\text{rank } X = k + 1.$$

Если в выборке есть линейно зависимые признаки, то дисперсия оценки коэффициентов при таких признаках будет бесконечной. Это не очень удобно при построении доверительных интервалов: они будут иметь бесконечную ширину. И кроме того, гипотезы тоже так не проверить.

Если возникла такая проблема, это значит, что от каких-то признаков в модели придется избавиться. Помимо всего прочего, для категориальных переменных нельзя использовать one-hot encoding, которое использовалось в предыдущих курсах. Дело в том, что при кодировании каждого уровня фактора своей бинарной переменной мы получаем, что в сумме такие переменные дают единичный столбец, а он в матрице  $X$  уже есть, поэтому столбцы получаются линейно зависимы. Вместо этого нужно использовать другой способ кодирования: dummy-кодирование. Если признак  $x_j$  принимает  $m$  различных значений, то его нужно кодировать  $m-1$  фиктивной переменной.

Тип должности	$x_1$	$x_2$
рабочий	0	0
инженер	1	0
управляющий	0	1

Таблица 10.4: Dummy-кодирование

Пусть  $y$  — это уровень заработной платы, а  $x$  — это занимаемая человеком должность: рабочий, инженер или управляющий. Эти три значения будут кодироваться двумя фиктивными переменными:  $x_1$  и  $x_2$  (таблица 10.4). В полученной регрессионной модели два признака:  $x_1$  и  $x_2$ :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Коэффициенты  $\beta_1$  и  $\beta_2$  при них кодируют среднюю разницу в уровнях зарплат инженера и рабочего и управляющего и рабочего. В регрессионных моделях с использованием dummy-кодирования интерпретация коэффициентов  $\beta$  модели всегда ведется относительно уровня фактора, который закодирован всеми 0. Можно менять кодировку dummy, используя соображения о том, какая из моделей будет удобнее интерпретироваться.

#### 10.4.4. Случайность ошибок

На очереди предположение о случайности ошибки:

$$\mathbb{E}(\varepsilon|x) = 0.$$

Гипотезу

$$H_0: \mathbb{E}(\varepsilon|x) = 0$$

можно очень легко проверить по данным. Для этого нужно построить регрессию  $y$  по  $x$ , вычислить остатки и проверить гипотезу о том, что среднее значение остатков равно 0. Это можно сделать, например, с помощью критерия Стьюдента.

#### 10.4.5. Гомоскедастичность ошибок

Пятое предположение — предположение гомоскедастичности ошибки:

$$\mathbb{D}(\varepsilon|x) = \sigma^2.$$

Это предположение можно проверять двумя способами. Первый, нестрогий, — это визуальный анализ. Нужно построить графики зависимости остатков от всех признаков  $x_j$  (рисунок 10.5) и посмотреть, выглядят ли точки на этом графике как горизонтальная полоса. Если вместо горизонтальной полосы на графике изображено что-то расширяющееся или сужающееся, значит, предположение гомоскедастичности не выполняется.

Формально это предположение можно проверять с помощью критерия Брайша-Пагана (таблица 10.5).

нулевая гипотеза:	$H_0: \mathbb{D}\varepsilon = \sigma^2;$
альтернатива:	$H_1: H_0$ неверна;
статистика:	$LM = nR_{\varepsilon^2}^2$ , $R_{\varepsilon^2}^2$ — коэффициент детерминации при регрессии $\hat{\varepsilon}^2$ на $x$ ;
нулевое распределение:	$LM \sim \chi_k^2$ .

Таблица 10.5: Описание критерия Брайша-Пагана

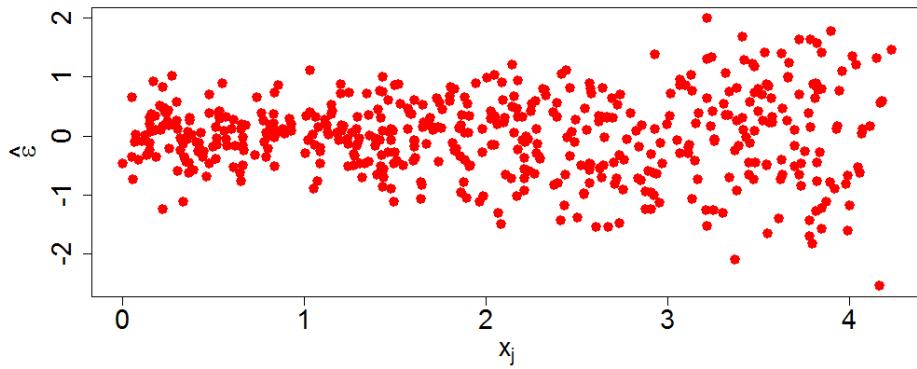


Рис. 10.5: График зависимости остатков от значения признаков  $x_j$

Если справедлива нулевая гипотеза, статистика этого критерия имеет распределение хи-квадрат с числом степеней свободы  $k$  (рисунок 10.6).

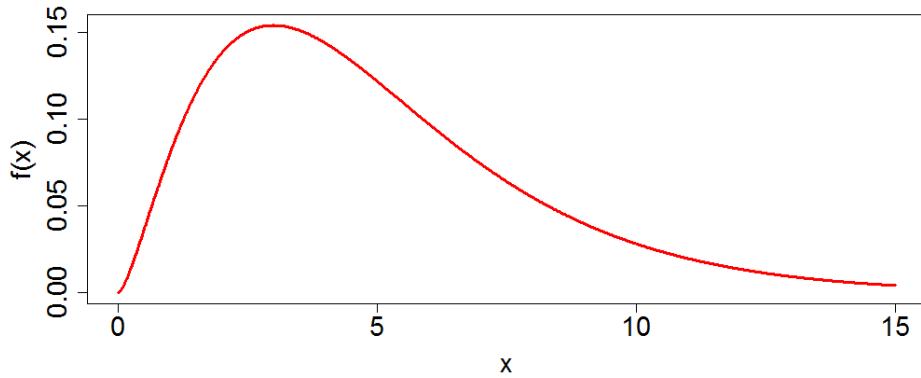


Рис. 10.6: Распределение хи-квадрат

#### 10.4.6. Нормальность ошибок

Наконец, шестое предположение — это предположение нормальности. Способы проверки нормальности уже разбирались ранее. Есть визуальный способ: нужно построить ку-ку график и посмотреть, лежат ли точки на этом графике более-менее на одной прямой. Также есть формальный способ: можно использовать статистические критерии для проверки нормальности. Среди всего разнообразия критериев рекомендуется использовать критерий Шапиро-Уилка.

## 10.5. Регрессия и причинно-следственные связи

### 10.5.1. Упражнения и холестерин

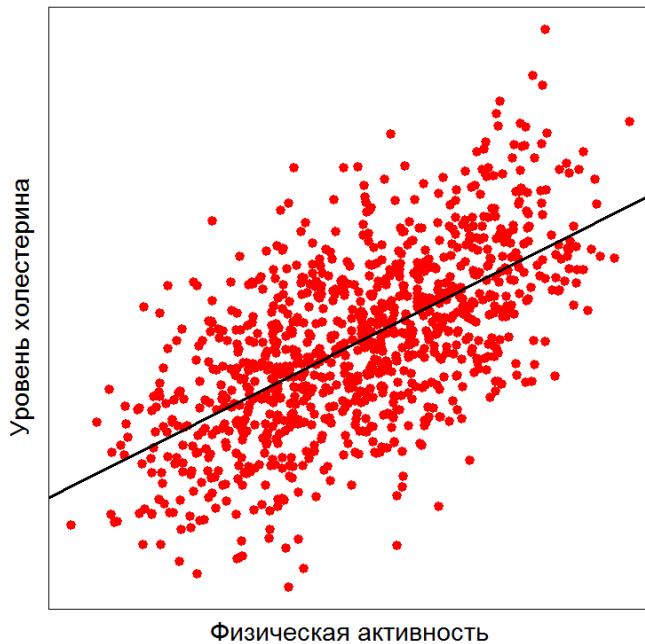


Рис. 10.7: Данные об уровне физической активности (по горизонтальной оси) и уровне холестерина в крови (по вертикальной оси)

Проводится исследование о связи уровня физической активности человека и уровня холестерина у него в крови, участвуют 10000 испытуемых. На рисунке 10.7 на графике по горизонтальной оси отложен уровень физической активности, по вертикальной — уровень холестерина. Видно, что эти два признака положительно коррелированы, поскольку облако вытянуто вдоль диагонали.

Можно проверить гипотезу о том, что по уровню физической активности  $ex$  можно каким-то образом предсказывать уровень холестерина  $chol$ . Для этого нужно построить регрессионную модель с одним-единственным признаком

$$chol = \beta_0 + \beta_1 ex$$

и проверить гипотезу

$$H_0: \beta_1 = 0$$

против альтернативы

$$H_0: \beta_1 > 0.$$

Критерий Стьюдента говорит, что нулевая гипотеза отвергается против этой альтернативы с очень маленьким достигаемым уровнем значимости  $p = 2 \times 10^{-16}$ . Даже если мы бы альтернатива была двухсторонней, получился бы достигаемый уровень значимости был бы  $p = 2 \times 10^{-16}$  — это тоже мало.

Стоит посмотреть, как эти же самые данные выглядят в разрезе возраста испытуемых. На рисунке 10.8 размечены пять возрастных групп: от левого нижнего угла к верхнему правому располагаются группы 10–20, 20–30, 30–40, 40–50 и 50–60 лет. В каждой возрастной группе уровень холестерина и количество физических упражнений друг с другом связаны отрицательно, но при этом в каждой следующей группе оба признака — и уровень холестерина, и уровень физической активности — растут с возрастом.

Теперь можно построить линейную регрессионную модель с двумя признаками: уровень холестерина  $chol$  будет предсказываться по возрасту  $age$  и количеству физических упражнений  $ex$ :

$$chol = \beta_0 + \beta_1 ex + \beta_2 age$$

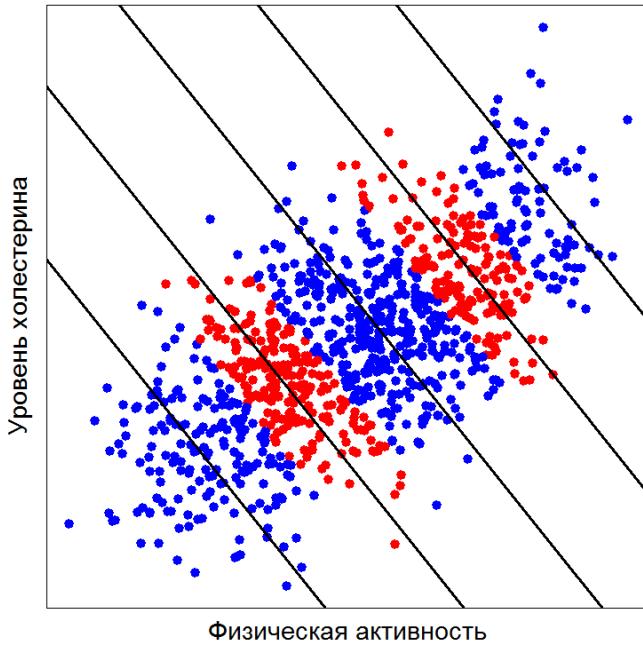


Рис. 10.8: Данные об уровне физической активности (по горизонтальной оси) и уровне холестерина в крови (по вертикальной оси) для разных возрастных групп

Проверяется утверждение, что количество физических упражнений хорошо предсказывает уровень холестерина. Гипотеза

$$H_0: \beta_1 = 0$$

будет проверяться против альтернативы

$$H_0: \beta_1 < 0.$$

Критерий Стьюдента дает достигаемый уровень значимости такой же маленький, как и раньше:  $p = 2 \times 10^{-16}$ , то есть нулевая гипотеза снова отвергается.

Итак, есть две модели. Первая модель показывает, что физические упражнения положительно влияют на уровень холестерина (то есть, чем больше вы упражняетесь, тем больше холестерина у вас в крови). Вторая модель демонстрирует ровно противоположное: если известен возраст человека, то чем больше он упражняется, тем ниже уровень холестерина у него в крови. Нужно решить, какой из этих двух выводов принять как финальный. В этой задаче можно включить здравый смысл и понять, что именно вторая модель верна. Кажется, что физические упражнения улучшают здоровье, поэтому, наверное, уровень холестерина должен снижаться. Но не во всех задачах такая опция доступна, стоит попробовать не использовать здравый смысл. Можно рассуждать так: вторая модель более подробна, она содержит больше признаков, значит, она богаче и, возможно, благодаря этому ее выводы более правильные. То есть модель, в которой больше признаков, лучше, чем модель, в которой их меньше.

### 10.5.2. Средний балл и мотивация

Пусть теперь первый признак — это средний балл выпускника из школы, а второй — это результат выпускника на мотивационном тесте во время собеседования при поступлении в вуз. Если посмотреть на облако точек, которое показано на рисунке 10.9, то кажется, что эти два признака вообще никак не связаны друг с другом. Красные точки на графике — это школьники, которые поступили в вуз, а синие — это те, которые не поступили. Видно, что правила приема в вуз устроены достаточно просто: поступают ученики, у которых или высокий средний балл, или хорошие результаты на тесте по мотивации. Требуется понять, влияет ли на результаты теста по мотивации средний балл. По конфигурации облака точек кажется, что не влияет, но это можно проверить формально: построить простую регрессионную модель, предсказывающую результат теста на мотивацию  $mot$  по среднему баллу  $SAT$ :

$$mot = \beta_0 + \beta_1 SAT.$$

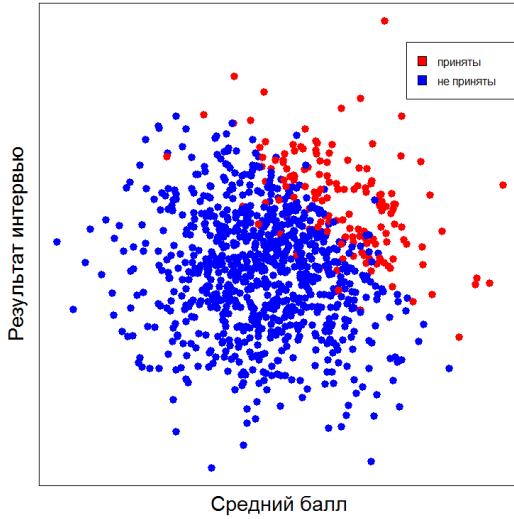


Рис. 10.9: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси). Красные точки — школьники, поступившие в вуз, синие — не поступившие.

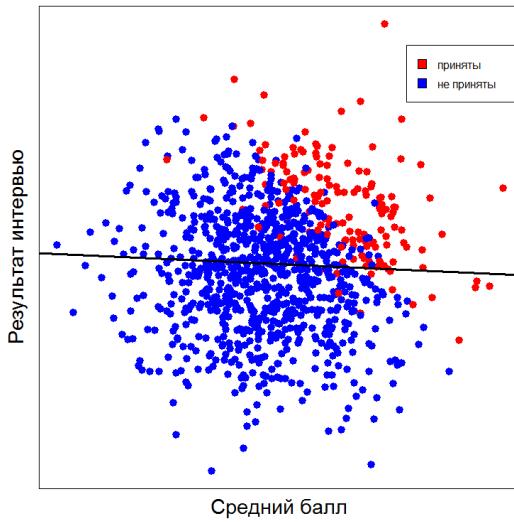


Рис. 10.10: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси) и прямая, соответствующая регрессии, которая предсказывает результат теста на мотивацию по среднему баллу

Далее нужно проверить гипотезу

$$H_0: \beta_1 = 0$$

против двухсторонней альтернативы

$$H_0: \beta_1 \neq 0$$

Критерий Стьюдента дает достигаемый уровень значимости  $p = 0.1452$ , нулевую гипотезу отвергнуть не получается, то есть нельзя утверждать, что средний балл влияет на результат теста по мотивации.

В эту регрессионную модель можно добавить еще один признак *acc*: «поступил ли человек в вуз» (10.11):

$$mot = \beta_0 + \beta_1 SAT + \beta_2 acc$$

В такой регрессионной модели снова проверяется нулевая гипотеза

$$H_0: \beta_1 = 0$$

против двухсторонней альтернативы

$$H_0: \beta_1 \neq 0.$$

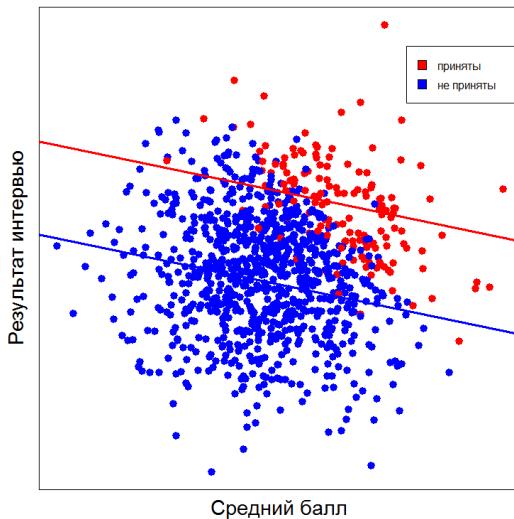


Рис. 10.11: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси) и прямая, соответствующая регрессии, которая предсказывает результат теста на мотивацию по среднему баллу и по тому, поступил ли человек в вуз

На этот раз критерий Стьюдента уверенно отвергает эту нулевую гипотезу. То есть утверждается, что в такой регрессионной модели результат теста на мотивацию значимо лучше предсказывается средним баллом студента, чем в отсутствие этого признака.

Нужно понять, как это можно интерпретировать. Снова имеются две регрессионные модели: в первой получается, что признак не влияет значимо на отклик, а во второй — влияет, причем в отрицательную сторону. Чем меньше средний балл (рисунок 10.11), тем выше результат теста на мотивацию.

### 10.5.3. Разница между двумя задачами

Задачи об уровне холестерина и о мотивации отличаются тем, что признаки, которые в этих задачах используются, связаны друг с другом совершенно разными причинно-следственными конфигурациями. В первой задаче побочный признак, возраст, влияет на оба интересующих признака: и на отклик, уровень холестерина, и на признак, количество физических упражнений. Такая конфигурация называется вилкой.

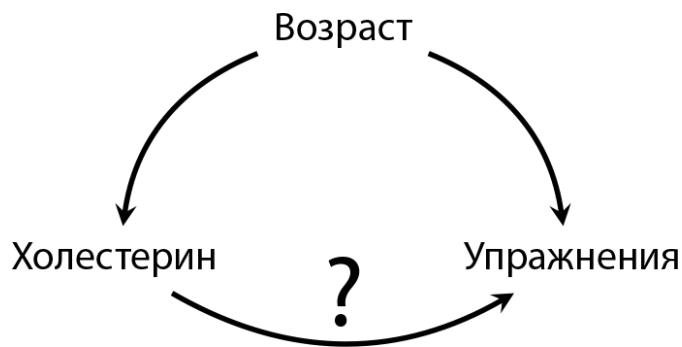


Рис. 10.12: Конфигурация "вилка"

В задаче с поступающими конфигурация противоположная. Дело в том, что и интересующий признак, средний балл, и отклик, мотивация, наоборот, влияют на третий побочный признак, факт поступления в вуз. Такая конфигурация называется коллайдером.

Оказывается, что для того чтобы коэффициент при признаке в регрессионной модели можно было интерпретировать с точки зрения причинно-следственной связи, нужно чтобы все остальные признаки в модели были предками  $x$ , то есть влияли на  $x$ , и не были ни в коем случае потомками  $x$ , которые также одновременно являются потомками  $y$ . То есть все побочные признаки в регрессионной модели не должны быть вершинами-коллайдерами.



Рис. 10.13: Конфигурация ”коллайдер”

**Причинно-следственная связь.** В линейной регрессионной модели  $\hat{\beta}_1$  — это оценка среднего эффекта, то есть среднего изменения  $y$  от увеличения  $x_1$  на 1. Этой оценке можно в некоторых случаях давать причинно-следственную интерпретацию, то есть утверждать, что если провести эксперимент, в котором зафиксированы все возможные факторы, которые могут влиять на  $y$ , и меняться будет только один из них —  $x_1$ , то именно так изменится  $y$ . Условие, при котором такую причинно-следственную интерпретацию давать можно, следующее: линейная регрессионная модель должна содержать все признаки, являющиеся причинами  $x_1$ . Кроме того, она не должна содержать признаков, которые являются следствиями одновременно  $x_1$  и  $y$ . То есть в регрессионной модели должны быть все предки  $x$  в причинно-следственном графе и не должно быть ни одной вершины коллайдера по отношению к паре  $x$  и  $y$ .

**Резюме.** Итак, линейная регрессия иногда позволяет оценивать причинно-следственные связи. Однако это можно делать только при некоторых достаточно строгих предположениях: линейная модель должна быть подобрана правильно, она должна содержать правильные признаки и не содержать неправильные. Это всё надо обязательно учитывать, если требуется провести причинно-следственную интерпретацию для модели. Плохо подобранные признаки могут привести к противоположным выводам. Интересно, что на сегодняшний день существуют методы, которые по обсервационным данным позволяют восстанавливать структуру предполагаемых причинно-следственных связей между признаками в этих данных. К сожалению, эти методы достаточно сложные, и, кроме того, реализация в Python, которая существует для этих методов, находится еще в альфа-версии, поэтому эта тема в нашем курсе не рассматривается.