

Урок 1

Доверительные интервалы

1.1. 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.

1.1.1. Правило двух сигм

Необходимо вспомнить, как выглядит правило двух сигм. Если случайная величина имеет нормальное распределение с математическим ожиданием μ и дисперсией σ^2 ($X \sim N(\mu, \sigma^2)$), то с вероятностью примерно 95 % она принимает значение из интервала $\mu \pm 2\sigma$ (рисунок 1.1):

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

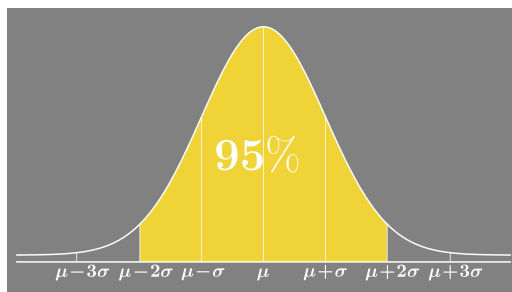


Рис. 1.1: Правило двух сигм.

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.

1.1.2. Уточнение правила двух сигм

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

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

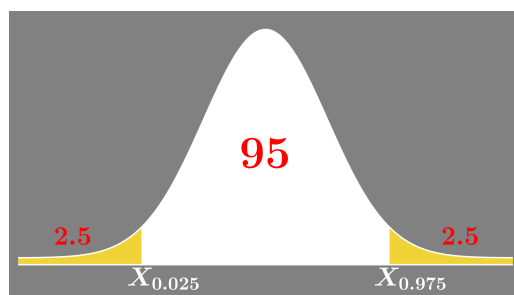


Рис. 1.2: Плотность вероятности нормально распределённой случайной величины.

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 1.2 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:

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

1.1.3. Предсказательный интервал

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

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

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 μ и σ , а также квантили z_α стандартного нормального распределения $N(0, 1)$:

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

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.

1.2. Доверительные интервалы с помощью квантилей

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

1.2.1. Точечные оценки

Пусть имеется некоторая случайная величина X , функция распределения которой зависит от неизвестного параметра θ :

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

Чтобы высказать предположение о значении параметра θ , 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 θ . Например, если θ — 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.

1.2.2. Доверительные интервалы

Помимо точечных, интерес представляют интервальные оценки, то есть доверительные интервалы. Доверительный интервал для параметра θ 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 θ 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 распределённой 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 μ и \bar{X}_n в неравенствах, которые стоят под знаком вероятности. Получается доверительный интервал для μ :

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

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 μ — это неизвестная фиксированная константа, а случайными являются границы интервала.

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

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

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

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

В свою очередь, доверительный интервал для μ , 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 распределённых 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:

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

1.3. Распределения, производные от нормального

1.3.1. 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 μ — это математическое ожидание, σ^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 1.3).

1.3.2. Распределение χ^2

Пусть есть k независимых одинаково распределённых 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 χ^2 — монотонно убывающая функция, максимум которой находится в точке $x = 0$ (рисунок 1.4). При $k > 3$ плотность перестает монотонно убывать, и с ростом k её максимум постепенно смещается вправо по числовой оси.

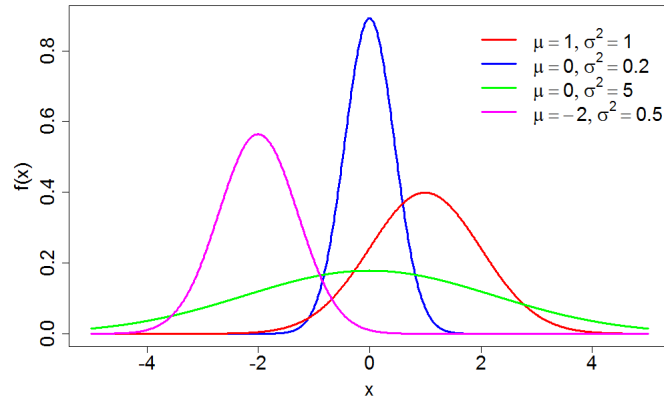


Рис. 1.3: Плотность вероятности нормального распределения с различными параметрами

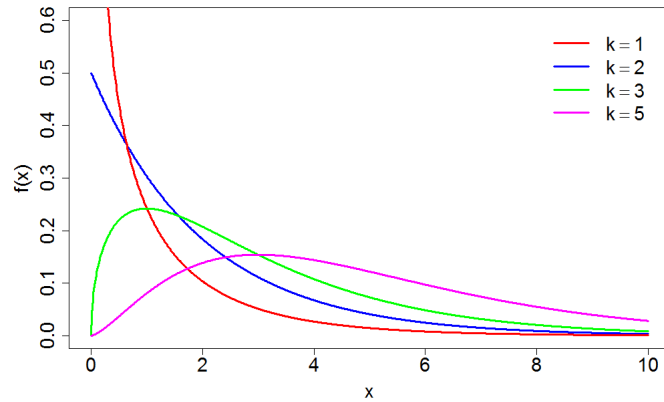


Рис. 1.4: Плотности распределений χ_k^2 с различными k

1.3.3. 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 ν .

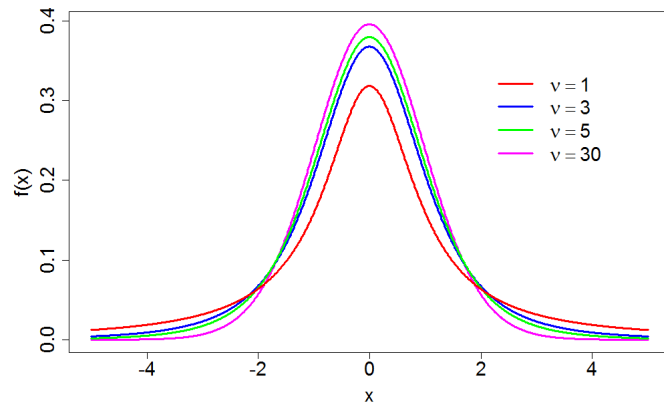


Рис. 1.5: Плотность вероятности распределения Стьюдента

На рисунке 1.5 изображены плотности вероятности распределения Стюдента при разных значениях параметра ν . 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 Стюдента более тяжелые хвосты, то есть для такой случайной величины большие по модулю значения более вероятны, чем в нормальном распределении. Однако чем больше значение параметра ν , тем меньше распределение Стюдента отличается от нормального. При $\nu > 30$ становится практически невозможно визуально различить эти распределения.

1.3.4. Распределение Фишера

Пусть теперь определены две независимые случайные величины X_1 и X_2 , принадлежащие распределению χ^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 1.6).

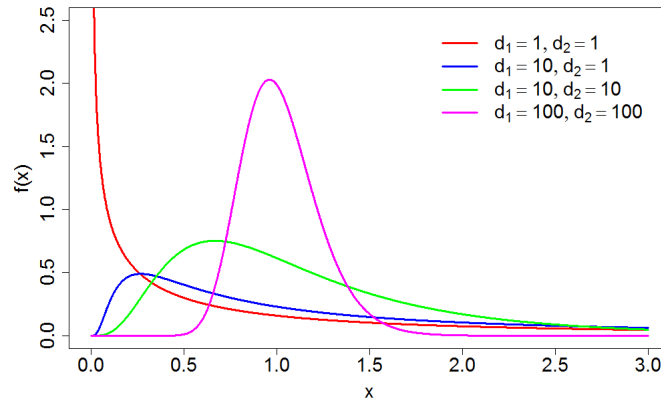


Рис. 1.6: Плотность вероятности распределения Фишера

1.3.5. 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 χ^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:

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$$

$$X_2 \sim N(\mu_2, \sigma_2^2), \quad X_2^{n_2} = (X_{21}, \dots, X_{2n_2}).$$

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

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