Глава 7

Доверительное оценивание

7.1 Базовая часть

7.1.1 Введение

Квантили и распределение хи-квадрат

Для построения доверительных интервалов нам понадобится следующее определение:

Определение 1. Квантилью уровня α функции распределения F называют число $x_\alpha = \inf\{x : F(x) > \alpha\}$.

Для непрерывного строго монотонного распределения это, таким образом, просто $F^{-1}(\alpha)$ в обычном смысле этого слова, для остальных такую величину также считают значением обратной функции $F^{-1}(\alpha)$.

При этом для непрерывного распределения естественным образом имеем $\mathbf{P}(X \in [x_a, x_b]) = b - a$, где x_α — квантиль F_X. Для общего случая

$$\mathbf{P}(X \in [x_a, x_b]) \geq b - a.$$

Квантили стандартного нормального распределения мы будем обозначать z_α: $\Phi(z_\alpha) = \alpha$, где Φ – ф.р. $\mathcal{N}(0,1)$, $\alpha \in (0,1)$.

Кроме этого нам будет удобно использовать распределение χ^2_n — распределение суммы квадратов n независимых случайных величин с $\mathcal{N}(0,1)$ распределением. Это распределение является частным случаем гамма-распределения $\chi^2_n \stackrel{d}{=} \Gamma(n/2, 2)$.

Его квантили мы будем обозначать через $y_{\alpha;n}$.

Семейства сдвига-масштаба

Отметим, что в силу формулы

$$f_{aX+b}(t) = f_X((t-b)/a)/|a|,$$

если величина имеет плотность

$$f_{a,b}(x) = \frac{1}{a} g\left(\frac{x-b}{a}\right), \quad a > 0,$$

где $g(x)$ некоторая плотность, то $(X-b)/a$ имеет плотность $g(x)$. Если $g(x)$ – некоторая плотность, не зависящая от a, b, а $f_{a,b}$ имеет описанный вид, то семейство $f_{a,b}$ называют семейством сдвига-масштаба. Параметр b здесь называют параметром сдвига, а a — параметром масштаба.

В частности, можно удобно выделять из плотности параметр: если плотность величины имеет вид

$$f_\theta(x) = \frac{1}{a(\theta)} g\left(\frac{x - b(\theta)}{a(\theta)}\right),$$

где $g(x)$ не зависит от x, то

$$X_1 \stackrel{d}{=} a(\theta)(Y_1 - b(\theta)),$$

где Y_1 имеет плотность g.

Пример 1. Пусть $X_1 \sim \mathcal{N}(\theta, \theta)$. Тогда

$$f_X(x) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x-\theta)^2}{2\theta}\right) = \frac{1}{\sqrt{\theta}} \varphi\left(\frac{x-\theta}{\sqrt{\theta}}\right).$$

Поэтому $X_1 = \theta + \sqrt{\theta} Y_1$, где $Y_1 \sim \mathcal{N}(0, 1)$.

Многие из известных вам семейств имеют такой вид: $\{R[\theta_1, \theta_2], \theta_1 < \theta_2\}$, $\{\mathcal{N}(\theta_1, \theta_2^2),\ \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+\}$, $\{Gamma(a, \theta), \theta > 0\}$, где a – фиксировано (но не в случае изменяющегося a), $\{\exp(\theta_1, \theta_2), \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+\}$, $\{Cauchy(\theta), \theta \in \mathbb{R}\}$.

Определение доверительного интервала

От построения точечных оценок перейдем к задаче доверительного оценивания.

Определение 2. Доверительным интервалом уровня доверия $1 - \alpha$ для параметра $\theta \in \mathbb{R}$ называется такая пара оценок $\widehat{\theta}_1$, $\widehat{\theta}_2$, что

$$\mathbf{P}_\theta(\widehat{\theta}_1(X_1, ..., X_n) \le \theta \le \widehat{\theta}_2(X_1, ..., X_n)) \ge 1 - \alpha$$

при всех $\theta \in \Theta$.

По возможности мы будем искать такой интервал, чтобы неравенство обращалось в равенство.

Определение 3. Доверительной областью уровня $1-\alpha$ для параметра векторного параметра $(\theta_1, ..., \theta_m)$ называют такое множество $D(X_1, ..., X_n) \subset R^m$, что

$$\mathbf{P}_\theta((\theta_1, ..., \theta_m) \in D(X_1, ..., X_n)) \ge 1 - \alpha.$$

Как уже упоминалось на первом семинаре, событие под знаком вероятности нужно рассматривать как накрытие случайным интервалом фиксированного параметра.

Например, при построении прогноза погоды для ожидаемой температуры строится доверительный интервал с некоторым уровнем доверия α, достаточно малым, чтобы интервал часто накрывал параметр и достаточно большим, чтобы интервал не оказался слишком большим. При геологической разведке место наиболее вероятного заложения руды описывается с помощью доверительной области. Область должна быть достаточно широкой, чтобы наверняка включить место заложения, но не настолько широкой, чтобы исследование стало слишком дорогим.

Примеры

Пример 2. Пусть $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$. Тогда $X_1 - \theta \sim \mathcal{N}(0, 1)$, откуда

$$\mathbf{P}_\theta(X_1 - \theta \in [z_{\alpha/2}, z_{1-\alpha/2}]) = 1 - \alpha,$$

где $z_{\alpha/2}$, $z_{1-\alpha/2}$ — квантили $\mathcal{N}(0,1)$.

Отсюда доверительным интервалом для θ будет интервал $(X_1 - z_{1-\alpha/2}, X_1 - z_{\alpha/2})$, а также, скажем, $(X_2 - z_{1-\alpha/3}, X_2 - z_{2\alpha/3})$. С другой стороны, $\sqrt{n}(\overline{X} - \theta)$ также распределено $\mathcal{N}(0,1)$, откуда доверительным интервалом является $(\overline{X} - z_{1-\alpha/2}n^{-1/2}, \overline{X} - z_{\alpha/2}n^{-1/2})$.

Кратчайший доверительный интервал

Чтобы выбрать наиболее подходящий интервал, нам нужно понять, какие именно интервалы нас интересуют (скажем, симметричные или вытянутые в одном из направлений), а среди таких интервалов выбрать интервал с наименьшей средней длиной $\mathbf{E}_\theta(\widehat{\theta}_2 - \widehat{\theta}_1)$. Зачастую интервалы берут центральные, то есть такие, что вероятность попадания левее интервала равна вероятности попадания правее.

Пример 3. Для примера 1 найдем среди интервалов вида $(\overline{X} - z_{p_2}n^{-1/2}, \overline{X} - z_{p_1}n^{-1/2})$, где $p_2 - p_1$ равно $1-\alpha$, интервал с наименьшей средней длиной. Длина каждого интервала есть $(z_{p_2} - z_{p_1})n^{-1/2}$. При каких же p_1, p_2: $p_2 - p_1 = 1 - \alpha$ эта величина минимальна? Находим минимум $z_{p_1+1-\alpha} - z_{p_1}$, откуда имеем соотношение на экстремальную точку $z'_{p_1+1-\alpha} = z'_{p_1}$. Но z_p определяется соотношением $\Phi(z_p) = p$, откуда

$$\Phi'(z_p)z'_p = 1 = \frac{1}{\sqrt{2\pi}}e^{-z_p^2/2}z'_p,$$

значит, наше уравнение имеет вид

$$\frac{1}{\sqrt{2\pi}}e^{-z_{p_1}^2/2} = \frac{1}{\sqrt{2\pi}}e^{-z_{p_1+1-\alpha}^2/2},$$

откуда $z_{p_1} = -z_{p_1+1-\alpha}$. Убедившись, что такое p_1 в действительности будет давать минимум длины интервала, из симметричности распределения $\mathcal{N}(0,1)$ имеем $p_1 = 1 - (p_1 + 1 - \alpha)$, т.е. $p_1 = \alpha/2$. Таким образом, для нормального распределения наиболее коротким из рассматриваемых интервалов будет центральный.

7.1.2 Методы построения интервалов

Какие могут быть методы построения доверительных интервалов?

1. Метод центральной функции.

   Предположим, что удастся найти функцию $h(\vec{x}, \theta)$, такую что распределение $h(X_1, ..., X_n, \theta)$ не зависит от θ, а функция h непрерывна и строго монотонна по θ при каждом x. Тогда мы можем выбрать x_{p_1}, x_{p_2} — квантили распределения величины $h(\vec{X}, \theta)$, $p_2 - p_1 = 1 - \alpha$,

   $$\mathbf{P}_\theta(x_{p_1} \le h(X_1, ..., X_n, \theta) \le x_{p_2}) = 1 - \alpha,$$

   откуда $\theta \in (h^{-1}(\vec{X}, x_{p_1}), h^{-1}(\vec{X}, x_{p_2}))$, где h^{-1} берется по последней переменной при фиксированных остальных. Зачастую удается построить такой интервал на основе какой-либо статистики, чье распределение зависит от θ простым образом.

   Пример 4. Пусть X_i — экспоненциальные с параметром масштаба θ (т.е. $f_X(x) = \theta e^{-x\theta} I_{x>0}$). Тогда достаточной статистикой будет $T = X_1 + ... + X_n$. Т.к. X_i имеют гамма-распределение с параметрами $(1, 1/\theta)$, то T имеет гамма-распределение с параметрами $(n, 1/\theta)$. При этом здесь $1/\theta$ — параметр масштаба, т.е. $T\theta \sim \Gamma(n, 1)$. Сведем задачу к распределению χ^2. Для этого заметим, что $2T\theta \sim \Gamma(n, 2) = \chi^2_{2n}$. Берем p_2, p_1 такими, что $p_2 - p_1 = 1 - \alpha$, тогда

   $$\mathbf{P}_\theta(2T(X_1, ..., X_n)\theta \in [y_{p_1}\theta, y_{p_2}\theta]) = 1 - \alpha,$$

где y_p — квантиль распределения χ^2_{2n}. Решая полученные неравенства на θ, имеем доверительный интервал

$$\left(\frac{y_{p_1}}{2(X_1+...+X_n)}, \frac{y_{p_2}}{2(X_1+...+X_n)}\right).$$

Центральный интервал будет с $p_2 = 1-\alpha/2$, $p_1 = \alpha/2$. Кратчайший интервал здесь достаточно несимпатичен.

2. Метод монотонного преобразования.

Лемма 1. *Пусть ф.р. статистики T $F_T(t;\theta)$ непрерывна и монотонно возрастает по θ при каждом t и пусть $\mathbf{P}_{\theta_1(t)}(T \le t) = \beta$, $F_{\theta_2(t)}(T > t) = \alpha - \beta$, $\beta \in (0, \alpha)$. Тогда*

$$\mathbf{P}_\theta(\theta \in (\theta_1(T), \theta_2(T))) = 1 - \alpha,$$

то есть $(\theta_1(T), \theta_2(T))$ – доверительный интервал. Если F монотонно убывает, то границы интервала следует поменять местами.

Доказательство. Величина $F_T(T(X_1, ..., X_n); \theta) \sim R[0,1]$, если F_T — функция распределения непрерывной случайной величины T, поскольку

$$\mathbf{P}(F_T(T) \le x) = \mathbf{P}(T \le F_T^{-1}(x)) = F_T(F_T^{-1}(x)) = x,\ x \in [0,1].$$

Следовательно,

$$\{\theta : \beta < F_{T;\theta}(T) < 1 - \alpha + \beta\}$$

является доверительным интервалом, поскольку

$$\mathbf{P}_\theta(\theta \notin \{\theta : \beta \le F_{T;\theta}(T) \le 1 - \alpha + \beta\}) = \mathbf{P}(F_{T;\theta}(T) \le \beta) + \mathbf{P}(F_{T;\theta}(T) \ge 1 - \alpha + \beta) = 1 - \alpha.$$

Остается воспользовать монотонностью и превратить неравенства $F_{T;\theta}(T) > \beta$ в $\theta > \theta_1(T)$, $F_{T;\theta}(T) < 1 - \alpha + \beta$ в $\theta < \theta_2(T)$. □

Пример 5. Пусть $X \sim \exp(\theta)$. Тогда

$$F_{X;\theta}(x) = 1 - e^{-\theta x}$$

монотонно возрастает по X. Значит, $\theta_1(X) = -(\ln(1-\alpha/2))/X$, $\theta_2(X) = -(\ln(\alpha/2))/X$ – границы интервала.

Модификация метода, подходящая для дискретного случая, рассмотрен в факультативе.

3. Асимптотические доверительные интервалы на основе асимптотически нормальных оценок.

Интервал из двух статистик $\widehat{\theta}_1$, $\widehat{\theta}_2$ называется асимптотическим доверительным интервалом уровня α, если

$$\liminf_{n\to\infty} \mathbf{P}_\theta(\theta \in (\widehat{\theta}_1, \widehat{\theta}_2)) \ge 1 - \alpha.$$

Построим такой интервал с помощью асимптотически нормальной оценки θ. Если оценка $\widehat{\theta}$ — асимптотически нормальная оценка θ, то

$$\frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma(\theta)} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Будем предполагать, что $\sigma(\theta)$ положительна и непрерывно зависит от θ. Если бы $\sigma(\theta)$ не зависела бы от θ, то мы получили бы асимптотический интервал $(\widehat{\theta} + z_{\alpha/2}\sigma n^{-1/2}, \widehat{\theta} + z_{1-\alpha/2}\sigma n^{-1/2})$. В случае зависимости применяют один из двух методов:

(a) Метод подстановки оценки.

Заменим в дисперсии θ на оценку $\widehat{\theta}$. Тогда

$$\frac{n^{1/2}(\widehat{\theta}-\theta)}{\sigma(\widehat{\theta})} = \frac{n^{1/2}(\widehat{\theta}-\theta)}{\sigma(\theta)}\frac{\sigma(\theta)}{\sigma(\widehat{\theta})} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

поскольку второй сомножитель в силу состоятельности $\widehat{\theta}$ сходится по вероятности к 1. Отсюда

$$(\widehat{\theta} + z_{\alpha/2}\sigma(\widehat{\theta})n^{-1/2}, \widehat{\theta} + z_{1-\alpha/2}\sigma(\widehat{\theta})n^{-1/2})$$

будет асимптотическим доверительным интервалом на уровне доверия $1-\alpha$.

(b) Метод стабилизации асимптотической дисперсии.

Если мы рассмотрим оценку $h(\widehat{\theta})$, то она будет асимптотически нормальной оценкой $h(\theta)$ с асимптотической дисперсией $(h'(\theta)\sigma(\theta))^2$. Следовательно, если выбрать $h(\theta) = \int \sigma^{-1}(\theta)d\theta$ (любую из первообразных), то $h(\widehat{\theta})$ будет иметь асимптотическую дисперсию 1. Отсюда

$$\mathbf{P}_\theta(n^{1/2}(h(\widehat{\theta}) - h(\theta)) \in (z_{\alpha/2}, z_{1-\alpha/2})) \to 1-\alpha.$$

Но функция h гладкая и монотонно возрастает, поскольку $g'(\theta) = 1/\sigma(\theta) > 0$. Значит с вероятностью, стремящейся к $1-\alpha$,

$$\theta \in \left(h^{-1}\left(h(\widehat{\theta}) - z_{1-\alpha/2}n^{-1/2}\right), h^{-1}\left(h(\widehat{\theta}) - z_{\alpha/2}n^{-1/2}\right)\right).$$

Имеем искомый доверительный интервал.

Пример 6. Для схемы Бернулли в силу ЦПТ

$$\frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Следовательно,

$$\frac{\sqrt{n}(\overline{X}-\theta)}{\sqrt{\overline{X}(1-\overline{X})}} \xrightarrow{d} \tilde{Z} \sim \mathcal{N}(0,1).$$

Отсюда первый метод дает нам асимптотический доверительный интервал

$$(\overline{X} - z_{1-\alpha/2}\sqrt{\overline{X}(1-\overline{X})}n^{-1/2}, \overline{X} - z_{\alpha/2}\sqrt{\overline{X}(1-\overline{X})}n^{-1/2}).$$

Для применения второго метода посчитаем

$$h(\theta) = \int \frac{1}{\sqrt{\theta(1-\theta)}}d\theta = 2\int \frac{1}{1-\sqrt{\theta}}d\sqrt{\theta} = 2\arcsin\sqrt{\theta} + C.$$

Отсюда

$$2\sqrt{n}(\arcsin(\sqrt{\overline{X}}) - \arcsin(\sqrt{\theta})) \to Z \sim \mathcal{N}(0,1).$$

Имеем для $\arcsin(\sqrt{\overline{X}})$ асимптотический доверительный интервал

$$(\arcsin(\sqrt{\overline{X}}) - z_{1-\alpha/2}n^{-1/2}/2, \arcsin(\sqrt{\overline{X}}) - z_{\alpha/2}n^{-1/2}/2),$$

а значит для θ — интервал

$$(\sin(\arcsin(\sqrt{\overline{X}}) - z_{1-\alpha/2}n^{-1/2}/2)^2, \sin(\arcsin(\sqrt{\overline{X}}) - z_{\alpha/2}n^{-1/2}/2)^2).$$

Стоит отметить, что в первом случае мы можем получить отрицательную левую границу, а во втором случае — границу для интервала для $\arcsin(\sqrt{\overline{X}})$. В таких случаях естественно взять интервал с левой границей 0.

7.2 Факультатив

7.2.1 Доверительные распределения для дискретных распределений

Если F_T имеет разрывы, то не всегда удается построить точный интервал, а зачастую это и вовсе невозможно, поскольку в вероятностном пространстве при некоторых θ может не быть событий вероятности $1-\alpha$.

В этом случае строят или асимптотический интервал (это можно делать описанным ранее способом), или избыточный интервал (то есть тот, вероятность попасть в который не меньше $1-\alpha$).

Метод центральной функции здесь вряд ли пригоден, а вот метод монотонного преобразования может быть модифицирован.

Лемма 2. *Пусть $F_{T;\theta}(x)$ монотонно возрастает по θ и пусть $\mathbf{P}_{\theta_1(t)}(T \le t) = \beta$, $F_{\theta_2(t)}(T \ge t) = \alpha - \beta$, $\beta \in (0, \alpha)$. Тогда*

$$\mathbf{P}_\theta(\theta \in (\theta_1(T), \theta_2(T))) \ge 1 - \alpha,$$

то есть $(\theta_1(T), \theta_2(T))$ – доверительный интервал. Если F монотонно убывает, то границы интервала следует поменять местами.

Доказательство. Идея в том, что справедливо соотношение

$$1 - \mathbf{P}(F_T(T) > x) = 1 - \mathbf{P}(T \ge F_T^{-1}(x)) = F_T(F_T^{-1}(x) - 0),\ x \in [0, 1].$$

Величина $F_T(F_T^{-1}(x) - 0)$ уже не равна x, если x лежит в разрыве между значениями x_- и x_+ функции распределения, а равна x_-, то есть меньше x. Следовательно, $\mathbf{P}_\theta(F_{T;\theta}(T) \le x) \le x$. Аналогично

$$\mathbf{P}_\theta(1 - F_{T;\theta}(T - 0) \ge x) \le 1 - x.$$

Значит,

$$\{\theta : F_{T;\theta}(T) \ge \beta, 1 - F_{T;\theta}(T - 0) \ge \alpha - \beta\}$$

является доверительным интервалом (возможно избыточным). Остается воспользовать монотонностью и превратить интервалы в описанные в лемме. □

Пример 7. Пусть $X \sim Geom(\theta)$, $\mathbf{P}(X = k) = (1-\theta)^k\theta$. Тогда

$$F_X(t; \theta) = 1 - (1 - \theta)^{[t]}$$

монотонно возрастающая функция θ при каждом t. Значит, выражая

$$\widehat{\theta}_1(x) = 1 - (1 - \alpha/2)^{1/x},\ \widehat{\theta}_2(x) = 1 - (\alpha/2)^{1/(x-1)},$$

получаем доверительный интервал с уровнем доверия не менее $1 - \alpha$.

7.2.2 Доверительные области

Для построения доверительных областей используют аналогичные методы.

1. Найдем множество $S(\theta)$, такое что $\mathbf{P}_\theta((X_1, ..., X_n) \in S(\theta)) \ge 1-\alpha$. Затем находим множество $G(x)$, такое что $x \in S(\theta) \Leftrightarrow \theta \in G(x)$. Тогда область $G(X_1, ..., X_n)$ будет доверительной для θ.
   Как и в задаче 1) множество $S(\theta)$ обычно строят на основе статистик, например, достаточных.

Пример 8. Пусть две независимых выборки $X_i \sim \mathcal{N}(\theta_1, 1)$, $Y_i \sim \mathcal{N}(\theta_1 + \theta_2, 1)$. Для построения $1-\alpha$-доверительной области для θ_1, θ_2 можно было бы построить $1-\alpha/2$-доверительные интервалы для θ_1, θ_2, а затем рассмотреть соответствующий прямоугольник, являющийся их декартовым произведением, но это чрезмерно большое множество. Воспользуемся методом 1) и построим на основе $\overline{X}$, $\overline{Y}$ доверительное множество.

$$\mathbf{P}_\theta(\overline{X} \in (\theta_1 + z_{p_1/2} n^{-1/2}, \theta_1 + z_{1-p_1/2} n^{-1/2}), \overline{Y} \in (\theta_2 + \theta_1 + z_{p_2/2}/n^{-1/2}, \theta_2 + \theta_1 + z_{1-p_2/2}/n^{-1/2})) = (1-p_1)(1-p_2).$$

Значит для $(1-p_1)(1-p_2) = 1-\alpha$ доверительной областью будет множество

$$\{(\theta_1, \theta_2) : \theta_1 \in (\overline{X} - z_{1-p_1/2} n^{-1/2}, \overline{X} + z_{p_1/2} n^{-1/2}), \theta_1 + \theta_2 \in (\overline{Y} - z_{1-p_1/2} n^{-1/2}, \overline{Y} + z_{p_1/2} n^{-1/2})\}$$

т.е. параллелограмм.

2. Если $(T_1, ..., T_m)$ — асимптотически нормальная оценка $(\theta_1, ..., \theta_m)$ с невырожденной ковариационной матрицей $\Sigma(\theta_1, ..., \theta_m)$, то

$$g(n)((T_1, ..., T_m) - (\theta_1, ..., \theta_m)) \xrightarrow{d} Z,\ Z \sim \mathcal{N}(\mathbf{0}, \Sigma(T_1, ..., T_m)).$$

   (a) Первый подход рекомендует нам записать

$$g(n)((T_1, ..., T_m) - (\theta_1, ..., \theta_m)) \in D_\alpha,$$

   где D_α — множество, в которое вектор $\mathcal{N}(\mathbf{0}, \Sigma(T_1, ..., T_m))$ попадает с вероятностью $1-\alpha$.

   (b) Второй подход рекомендует подобрать отображение $h : \mathbb{R}^m \to \mathbb{R}^m$, так что

$$g(n)(h(T_1, ..., T_m) - h(\theta_1, ..., \theta_m)) \xrightarrow{d} Z \sim \mathcal{N}(\vec{0}, \mathbf{E}),$$

   При этом h будет подбираться из аналогичных 4б) соображений — матрица Якоби отображения h должна быть обратна матрице $\Sigma(\theta)$.

Для нормальных $\mathcal{N}(\vec{0}, \mathbf{E})$ принято использовать в качестве доверительного множества сферу

$$||x_1^2 + ... + x_n^2|| \leq x_{1-\alpha},$$

где x_α — квантиль уровня α распределения χ_n^2. Для нормальных $\mathcal{N}(\vec{0}, \Sigma)$ таким множеством оказывается эллипсоид

$$\vec{x}^t \Sigma^{-1} \vec{x} \leq x_{1-\alpha}.$$

Тем самым, мы получаем методы получения доверительных эллипсоидов на основе асимптотически нормальных оценок.