

Лекция 4. Доверительное оценивание

4.1. Интервальное оценивание

Мы рассмотрели методы оценивания, позволяющие находить точечные оценки параметров. Однако для каждой конкретной выборки значения оценки отличаются от истинного значения параметра, поэтому полезно знать возможную погрешность, возникающую при использовании оценки. Например, можно указать такой интервал, внутри которого с вероятностью γ находится истинное значение параметра θ . В этом случае говорят об интервальном оценивании. Иногда интервальное оценивание называют доверительным.

4.2 Понятие доверительного интервала

Определение

γ -доверительным интервалом для параметра θ называют интервал $(T_1(X_n), T_2(X_n))$, такой что $P\{T_1(X_n) < \theta < T_2(X_n)\} = \gamma, \forall \theta \in \Theta$.

Статистику $T_1(X_n)$ называют нижней границей доверительного интервала, а $T_2(X_n)$ - верхней.

В случае многомерного параметра определение доверительной границы для каждой компоненты векторного параметра. В общем случае доверительных интервалов может быть несколько, поэтому из них выбирают интервал минимальной длины.

4.3 Построение доверительного интервала с использованием центральных статистик

Определение

$G(\bar{X}_n, \theta)$ - центральная статистика, если

1) распределение $G(\bar{X}_n, \theta)$ не зависит от θ .

2) При любом фиксированном θ статистика $G(\bar{X}_n, \theta)$ непрерывна и строго монотонна по θ .

С помощью центральной статистики можно построить доверительный интервал.

Пусть $f_G(g)$ плотность распределения статистики $G(X_n, \theta)$, так как функция плотности не зависит от θ , то для любого $\gamma \in (0, 1)$, можно найти такие $g_1, g_2, g_1 < g_2$, что

$$P\{g_1 < G(X_n, \theta) < g_2\} = \int_{g_1}^{g_2} f_G(g) dg = \gamma$$

Составим два уравнения

$$G(X_n, T_1) = g_1$$

$$G(X_n, T_2) = g_2$$

Из решений выберем минимальное и максимальное

$$T_1 = \min\{T_1, T_2\}, T_2 = \max\{T_1, T_2\}, \text{ тогда неравенство } g_1 < G(X_n, T_1) < g_2, T_1(X_n) < \theta < T_2(X_n)$$

В силу строгой монотонности функции G следует

$$P\{g_1 < G(X_n, \theta) < g_2\} = P\{T_1(X_n) < \theta < T_2(X_n)\} = \gamma \text{ является } \gamma \text{ доверительным интервалом.}$$

Пример

Дана $\mathbb{X}_n = \{X_1, \dots, X_n\} \in N(\theta, \sigma^2)$.

Построить доверительный интервал для θ .

$$M_{X_i} = \theta; D_{X_i} = \sigma^2; \frac{\sqrt{n}(\bar{x} - \theta)}{\sigma} \in N(0, 1)$$

Таким образом, при больших n можно взять $G(\mathbb{X}_n, \theta) = \frac{\sqrt{n}(\bar{x} - \theta)}{\sigma}$, так как данная статистика имеет распределение, не зависящее от θ и G монотонна убывает по θ .

Составим два неравенства

$$\text{Пусть } g_1 < g_2, P\{g_1 < G(\mathbb{X}_n, \theta) < g_2\} = \Phi(g_2) - \Phi(g_1) = \gamma$$

$$\frac{\sqrt{n}(\bar{x} - T_1)}{\sigma} = g_1; \frac{\sqrt{n}(\bar{x} - T_2)}{\sigma} = g_2$$

$$\text{Отсюда } T_1 = \bar{x} - \frac{\sigma}{\sqrt{n}} g_1; T_2 = \bar{x} - \frac{\sigma}{\sqrt{n}} g_2$$

В силу $g_1 < g_2 \Rightarrow T_2 < T_1$, тогда $\left(\bar{x} - \frac{\sigma}{\sqrt{n}} g_2; \bar{x} - \frac{\sigma}{\sqrt{n}} g_1 \right) \gamma$ доверительный интервал для θ .

Пусть $g_1 = -g_2$, тогда $\Phi(g_2) - \Phi(g_1) = \Phi(g_2) - \Phi(-g_2) = \Phi(g_2) - (1 - \Phi(g_2)) = 2\Phi(g_2) - 1 = \gamma$

$$\Phi(g_2) = \frac{\gamma + 1}{2}; g_2 = \Phi^{-1}\left(\frac{\gamma + 1}{2}\right)$$

Для построения доверительного интервала с помощью центральной статистики основная проблема заключается в нахождении этой центральной статистики. Можно выделить класс моделей, для которых центральная статистика существует и имеет простой вид.

Пусть $F(x, \theta)$ - функция распределения наблюдаемой случайной величины и монотонна по параметру θ . (то есть $F(x, \theta_1) \geq F(x, \theta_2); \forall \theta_1 < \theta_2$ или $F(x, \theta_1) \leq F(x, \theta_2); \forall \theta_1 > \theta_2$).

Можно положить в качестве центральной статистики функцию

$$G(\mathbb{X}_n, \theta) = -\sum_{i=1}^n \ln F(X_i, \theta).$$

Действительно:

1. Непрерывность и монотонность $G(\mathbb{X}_n, \theta)$ следует из непрерывности и монотонности $F(x, \theta)$.
2. Найдем распределение случайной величины $G(\mathbb{X}_n, \theta)$. Случайная величина $\eta = F(x_i, \theta)$ подчинена $Rav(0,1)$:

$$P\{\eta < x\} = P\{F(x_i, \theta) < x\} = P\{x_i < F^{-1}(x, \theta)\} = F(F^{-1}(x, \theta)) = x$$

$x \in (0,1)$

Случайная величина $-\ln \eta$ подчинена гамма-распределению с параметром формы 1 и параметром масштаба 1:

$$P\{-\ln \eta < x\} = P\{\ln \eta > -x\} = P\{\eta > e^{-x}\} = 1 - P\{\eta > e^{-x}\} = 1 - e^{-x}.$$

Из свойств гамма-распределения известно, что сумма случайных величин, подчиненных гамма-распределению, также подчинена гамма-распределению, при этом параметры формы суммируются, а параметр масштаба остается тем же. Поэтому

$-\sum_{i=1}^n \ln F(x_i, \theta)$ подчинена гамма-распределению с параметром формы n , то есть

$$F_G(g) = P\left\{-\sum_{i=1}^n \ln F(x_i, \theta) < g\right\} = \frac{1}{\Gamma(n)} g^{n-1} e^{-g}, g > 0.$$

Таким образом, распределение случайной величины $G(\mathbb{X}_n, \theta)$ не зависит от θ , и, следовательно, $G(\mathbb{X}_n, \theta)$ - центральная статистика.

В результате получаем следующий метод построения γ - доверительного интервала:

1. Выбираем $g_1 < g_2$ так, чтобы $\frac{1}{\Gamma(n)} \int_{g_1}^{g_2} g^{n-1} e^{-g} dg = \gamma$.

2. Решая уравнения

$$-\sum_{i=1}^n \ln F(x_i, \theta) = g_1$$

$$-\sum_{i=1}^n \ln F(x_i, \theta) = g_2$$

относительно θ находим корни $T_1(X_n) < T_2(X_n)$ (в общем случае численными методами).

Тогда $(T_1(X_n), T_2(X_n))$ искомый γ - доверительный интервал.

Пример

Для выборки $X_n = (X_1, \dots, X_n)$ из $Rav(0, \theta)$ построить γ - доверительный интервал для параметра θ . В качестве статистики возьмем функцию

$$G(X_n, \theta) = -\sum_{i=1}^n \ln F(x_i, \theta) = -\sum_{i=1}^n \ln F(x_i, \theta) = g_1$$

$$-\sum_{i=1}^n \ln \frac{x_i}{\theta} = -\sum_{i=1}^n \ln x_i + n \ln \theta$$

$$\begin{cases} G(X_n, T_1) = g_1 \\ G(X_n, T_2) = g_2 \end{cases}$$

$$-\sum_{i=1}^n \ln x_i + n \ln T_1 = g_1$$

$$n \ln T_1 = g_1 - \sum_{i=1}^n \ln x_i$$

$$T_1 = e^{\frac{\left(g_1 - \sum_{i=1}^n \ln x_i\right)}{n}} = e^{\frac{g_1}{n} \left(\prod_{i=1}^n x_i\right)^{1/n}}$$

$$T_2 = e^{\frac{g_2}{n} \left(\prod_{i=1}^n x_i\right)^{1/n}}.$$

При $g_1 < g_2$ $T_1(X_n) < T_2(X_n)$, следовательно γ - доверительный интервал имеет вид

$$\left(e^{\frac{g_1}{n} \left(\prod_{i=1}^n x_i\right)^{1/n}}, e^{\frac{g_2}{n} \left(\prod_{i=1}^n x_i\right)^{1/n}} \right) \frac{1}{\Gamma(n)} \int_{g_1}^{g_2} g^{n-1} e^{-g} dg = \gamma.$$

4.4 Построение доверительного интервала с использованием распределения точечной оценки параметров

Если имеется некоторая точечная оценка $T = T(\mathbb{X}_n)$ для параметра θ и известна ее функция распределения $F_T(t, \theta)$, то доверительный интервал можно построить основываясь на этой функции.

Пусть функция распределения $F_T(t, \theta)$ непрерывна и монотонна по θ и задан доверительный уровень γ .

Найдем $t_1 = t_1(\theta)$ и $t_2 = t_2(\theta)$ такие, что $P\{t_1 < T(\mathbb{X}_n) < t_2\} =$
 $= F_T(t_2, \theta) - F_T(t_1, \theta) = \gamma$.

Чтобы процедура была однозначной, будем выбирать t_1 и t_2 из условия $F_T(t_1, \theta) = \frac{1-\gamma}{2}; 1 - F_T(t_2, \theta) = \frac{1-\gamma}{2}$, а соответствующий доверительный интервал будем называть *центральным*.

Обозначим через $D\gamma = \{(\theta, \theta') : t_1(\theta) < \theta' < t_2(\theta)\}$, тогда $P\{(\theta, T(\mathbb{X}_n)) \in D\gamma\} = P\{t_1 < T(\mathbb{X}_n) < t_2\} = \gamma; \forall \theta \in \Theta$

Рассмотрим множество $D\gamma(\theta') = \{\theta : (\theta, \theta') \in D\gamma\}$. Событие $\theta \in D\gamma(T(\mathbb{X}_n))$ произойдет тогда и только тогда, когда $T(\mathbb{X}_n) \in (t_1(\theta), t_2(\theta))$ и, следовательно, имеет вероятность γ . Так как $t_i(\theta)$ являются монотонными одного типа в силу монотонности функции распределения, то $D\gamma(\theta')$ является интервалом $(\theta_1(\theta'), \theta_2(\theta'))$ а γ - доверительный интервал есть $(T_1(\mathbb{X}_n), T_2(\mathbb{X}_n)); T_i(\mathbb{X}_n) = \theta_i(T(\mathbb{X}_n))$.

Алгоритм построения γ -доверительных интервалов

1. Находим точечную оценку - наблюдаемое значение оценки.
2. Решаем отношение θ уравнений

$$F_T(t, \theta_1) = \frac{1-\gamma}{2}$$

$$F_T(t, \theta_2) = \frac{1+\gamma}{2}$$

3. $\theta_1 = \min(\theta_1, \theta_2), \theta_2 = \max(\theta_1, \theta_2)$.

В случае дискретных моделей рассуждения аналогичны. Единственное отличие заключается в том, что в силу дискретности функции распределения можно обеспечить выполнение неравенства

$$P(t_1 < T(X_n) < t_2) = F(t_2 - 0; \theta) - F(t_1; \theta) \geq \gamma.$$

Выбирая различные точечные оценки $T(X_n)$ будут получаться различные доверительные интервалы. Доверительные интервалы с минимальной длиной будут получаться, если брать точечные оценки с минимальной дисперсией, то есть эффективные или асимптотически эффективные оценки.

Оценки максимального правдоподобия при достаточно общих условиях являются асимптотически эффективными и асимптотически нормальными, следовательно

$$P\left\{\left|\theta_n - \theta\right| \sqrt{ni(\theta_n)} \leq c_\gamma\right\} \Rightarrow \Phi(c_\gamma) - \Phi(-c_\gamma) = 2\Phi(c_\gamma) - 1 = \gamma \quad \text{если} \quad c_\gamma = \Phi^{-1}\left(\frac{\gamma + 1}{2}\right)$$

$\left(\theta_n - \frac{c_\gamma}{\sqrt{ni(\theta)}}; \theta_n + \frac{c_\gamma}{\sqrt{ni(\theta)}}\right)$, следовательно является асимптотическим кратчайшим γ -доверительным интервалом для θ .