Метод подстановки для получения оценок функционалов от функции распределения генеральной совокупности

Идея в замене в функционале функции F на функцию F_n, или в замене распределения P на выборочное распределение P_n

Оценивание моментов генеральной совокупности

$\xi \sim P, \alpha_k = E\xi^k$ - k-ый выборочный момент, $\mu_k = E(\xi - \alpha_1)^k$ - k-ый выборочный центральный момент

Выразим их с помощью P, сделаем замену, и получим оценки

$\alpha_k = \int\limits_{\mathbb{R}^1} x^k dP$

$A_k = \int\limits_{\mathbb{R}^1} x^k dP_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$, A_k - оценка для α_k

$\overline{X} = A_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ называется выборочным средним

$\mu_k = \int\limits_{\mathbb{R}^1} (x-\alpha_1)^k dP = \int\limits_{\mathbb{R}^1} (x - \int\limits_{\mathbb{R}^1} xdP)^k dP$

$M_k = \int\limits_{\mathbb{R}^1} (x - \int\limits_{\mathbb{R}^1} xdP_n)^k dP_n$

$M_k = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^k = \overline{(X-\overline{X})^k}$

$S_n^2 = M_2 = \overline{(X-\overline{X})^2}$ называется выборочной дисперсией

$S_n^2 = \overline{X^2} - \overline{X}^2$

$M_k = \overline{\sum\limits_{i=0}^{k} C_k^i X^i \cdot (-1)^{k-i}(\overline{X})^{k-i}} = \sum\limits_{i=0}^{k} \overline{(-1)^{k-i} C_k^i X^i \cdot (\overline{X})^{k-i}} = \sum\limits_{i=0}^{k} (-1)^{k-i} C_k^i \overline{X^i} \cdot (\overline{X})^{k-i}$

Проверка для $k=2$: $(\overline{X})^2 - 2\overline{X}\cdot\overline{X} + \overline{X^2} = \overline{X^2} - (\overline{X})^2 = S_n^2$

Свойства оценок:

1) Свойство несмещённости оценок моментов

$EA_k = E(\frac{1}{n}\sum\limits_{i=1}^{n} X_i^k) = \frac{1}{n}\sum\limits_{i=1}^{n} EX_i^k = \frac{1}{n}\sum\limits_{i=1}^{n}\alpha_k = \alpha_k$

A_k - несмещённая оценка для α_k

$EM_2 = E\overline{X^2} - E(\overline{X})^2 = \alpha_2 - E(\overline{X})^2$

$E(\overline{X})^2 = E(\frac{1}{n}\sum\limits_{i=1}^{n} X_i)(\frac{1}{n}\sum\limits_{j=1}^{n} X_j) = \frac{1}{n^2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} EX_iX_j = \frac{1}{n^2}(\sum\limits_{i=j} EX_i^2 + \sum\limits_{i\neq j} EX_iX_j) = \frac{1}{n^2}(nEX_i^2 + n(n-1)EX_iX_j) = \frac{1}{n}\alpha_2 + \frac{n-1}{n}\alpha_1^2 = \frac{1}{n}(\alpha_2 - \alpha_1^2) + \alpha_1^2 = \frac{\sigma^2}{n} + \alpha_1^2$

$EM_2 = \alpha_2 - \alpha_1^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n}$

Смещение оценки M_2 для μ_2 есть $-\frac{\sigma^2}{n}$ - средняя ошибка

Также можно определить качество оценки средним разбросом, то есть дисперсией

$DA_k = D(\frac{1}{n}\sum\limits_{i=1}^{n} X_i^k) = \frac{1}{n^2}\sum\limits_{i=1}^{n} DX_i^k = \frac{1}{n}DX_1^k = \frac{1}{n}(EX_1^{2k} - (EX_1^k)^2)$

$DA_k = \frac{1}{n}(\alpha_{2k} - \alpha_k^2)$

DS_n^2 посчитать сложнее, задача на семинар

2) Свойство состоятельности оценок моментов

$A_{k,n} := A_k$

$A_{k,n} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k \to^{\mathbb{P}}_{n\to\infty} EX_1^k = \alpha_k$

Выборочный момент A_k - состоятельная оценка для α_k

Для выборочных центральных моментов понадобится лемма

Пусть имеются независимые случайные величины $\eta_1(n), \ldots, \eta_k(n)$, и $\forall i$ $\eta_i(n) \to^{\mathbb{P}}_{n\to\infty} c_i = const$ и пусть $f(x_1, \ldots, x_k)$ - функция, непрерывная в окрестности точки $(c_1, \ldots, c_n)$.

Тогда $f(\eta_1(n), \ldots, \eta_k(n)) \to^{\mathbb{P}}_{n\to\infty} f(c_1, \ldots, c_k)$

Доказательство

$C_{n,\epsilon} = \{\omega \mid |f(\eta_1(n), \dots, \eta_k(n)) - f(c_1, \dots, c_k)| > \epsilon\}$

? $\mathbb{P}(C_{n,\epsilon}) \to_{n\to\infty} 0$

$\eta_i(n) \to^{\mathbb{P}} c_i$, то есть $\forall i \forall \delta_i > 0$ $\mathbb{P}(\{|\eta_i(n) - c_i| > \delta_i\}) \to_{n\to\infty} 0$

По непрерывности $\forall \epsilon > 0\ \exists \delta_i > 0 :$ при $|x_i - c_i| < \delta_i$ $|f(x_1, \dots, x_k) - f(c_1, \dots, c_k)| < \epsilon$

Событие $C_{n,\epsilon}$ наступает, если $|\eta_i(n) - c_i| > \delta_i$ для некоторого i

$C_{n,\epsilon} \subset \bigcup\limits_{i=1}^{k} \{|\eta_i(n) - c_i| > \delta_i\}$

$\mathbb{P}(C_{n,\epsilon}) \le \sum\limits_{i=1}^{k} \mathbb{P}(\{|\eta_i(n) - c_i| > \delta_i\}) \to_{n\to\infty} 0$ - QED

$S_n^2 = A_{2,n} - (A_{1,n})^2 = f(A_{1,n}, A_{2,n})$, где $f(x_1, x_2) = x_2 - x_1^2$, f непрерывна в точке (α_1, α_2)

По лемме $S_n^2 \to^{\mathbb{P}}_{n\to\infty} \alpha_2 - \alpha_1^2 = \sigma^2$

Выборочная дисперсия является состоятельной оценкой дисперсии генеральной совокупности

$\gamma = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\alpha_2 - \alpha_1^2)^{3/2}}$ - коэффициент асимметрии (отвечает за перекос распределения)

$\kappa = \frac{\mu_4}{\sigma^4} - 3$ - коэффициент эксцесса (отвечает за ширину распределения)

$\Gamma = \frac{M_3}{S_n^3}$, $K = \frac{M_4}{S_n^4} - 3 = \frac{\overline{(X - \overline{X})^4}}{\overline{(X - \overline{X})^2}^2} - 3$ - оценки для коэффициентов

$\tilde{S_n^2} = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \overline{X})^2$ - исправленная выборочная дисперсия, несмещённая оценка для дисперсии

3) Асимптотическая нормальность оценок моментов

$A_{k,n}$ - оценка α_k, $DA_{k,n} = \frac{\alpha_{2k} - \alpha_k^2}{n}$

$A_{k,n} - \alpha_k = \frac{1}{n} \sum\limits_{i=1}^{n} (X_i^k - \alpha_k)$

$\frac{\sqrt{n}(A_{k,n} - \alpha_k)}{\sqrt{\alpha_{2k} - \alpha_k^2}} = \frac{\sum (X_i^k - \alpha_k)}{\sqrt{n}\sqrt{DX_1^k}} \to^d_{n\to\infty} \xi \sim N(0,1)$ - по центральной предельной теореме

$A_{k,n} \sim P = N(\alpha_k, \frac{\alpha_{2k} - \alpha_k^2}{n})$

Точечное оценивание параметров распределения

$F(x, \theta)$ - статистическая модель, θ - параметр

Примеры:

1) Равномерная модель $R(\theta)$, $\theta \in \Theta = (0; +\infty)$

Имеет плотность распределения $f(x, \theta) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x! \in [0; \theta] \end{cases}$

2) Нормальная модель $N(\theta_1, \theta_2^2)$, $(\theta_1, \theta_2) \in \Theta = \mathbb{R}^1 \times (0; +\infty)$

$f(x, \theta_1, \theta_2^2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp(-\frac{1}{2} \frac{(x - \theta_1)^2}{\theta_2^2})$

3) $N(a, \theta^2)$ - нормальная модель с известным средним

4) $N(\theta, \sigma^2)$ - нормальная модель с известной дисперсией

5) Пуассоновская модель $\Pi(\theta)$, $\theta > 0$

$f(x, \theta) = \frac{\theta^x}{x!} e^{-x}$, $x = 0, 1, \dots$, мера дискретная

6) $Bi(k, \theta)$ - биномиальная модель, $\theta \in (0; 1)$

$f(x, \theta) = C_k^x \theta^x (1 - \theta)^{k-x}$, $x = 0, 1, \dots, k$ (среди k испытаний ровно x успехов, θ - вероятность успеха)

6') $Bi(1, \theta)$, $f(1, \theta) = \theta^x (1 - \theta)^{1-x}$, $x = 0, 1$

7) $\overline{Bi(r, \theta)}$ - отрицательная биномиальная модель, $\theta \in (0; 1)$

$f(x, \theta) = C_{x+r-1}^x \theta^x (1 - \theta)^r$ (до наступления r успехов было ровно x неудач, θ - вероятность неудачи)

8) $\overline{Bi(1, \theta)}$ - геометрическое распределение

9) Гамма-распределение, $\lambda > 0$

$f(x, \theta) = \frac{x^{\lambda - 1} e^{-\frac{x}{\theta}}}{\Gamma(\lambda) \cdot \theta^\lambda}$, $\theta > 0$

Случайные величины, зависящие только от выборки, а не от параметра, будем называть статистиками

Пусть $\tau(\theta)$ - функция от параметра

T - статистику, оценивающую $\tau(\theta)$ будем называть оценкой для $\tau(\theta)$ (оценка не должна содержать параметр)

T^* - оптимальная оценка, если

1)$\forall\theta\ E_\theta T^* = \tau(\theta)$

2) для любой другой несмещённой оценки $T\ D_\theta T^* \leq D_\theta T$

(Значения E и D зависят от параметра, даже если оценки от него не зависят)

Оценка называется эффективной, если её дисперсия совпадает с нижней границей дисперсий всех несмещённых оценок

Свойства оптимальных оценок

Теорема 1

Оптимальная оценка, если она существует, является единственной

Доказательство

Пусть T_1 и T_2 - оптимальные оценки

$E_\theta T_1 = E_\theta T_2 = \tau(\theta)$, $D_\theta T_1 = D_\theta T_2 = v(\theta)$

$T = \frac{T_1+T_2}{2}$, $E_\theta T = \tau(\theta)$, $D_\theta T \geq v(\theta)$

$T = T_1 = T_2$ - надо доказать

$D_\theta T = \frac{1}{4}(D_\theta T_1 + D_\theta T_2 + 2cov(T_1, T_2)) \geq v(\theta)$

$2v(\theta) + 2cov(T_1, T_2) \geq 4v(\theta)$

$cov(T_1, T_2) \geq v(\theta)$

$|cov(T_1, T_2)| \leq \sqrt{D_{\theta_1} D_{\theta_2}} = v(\theta)$

Имеет место равенство в неравенстве Коши-Буняковского-Шварца: $cov(T_1, T_2) = v(\theta)$

Случайные величины $T_1 - \tau(\theta)$ и $T_2 - \tau(\theta)$ линейно зависимы, $\exists c \neq 0$, $T_1 - \tau(\theta) = c(T_2 - \tau(\theta))$

$D_\theta T_1 = c^2 D_\theta T_2 \Rightarrow c^2 = 1$

$v(\theta) = cov(T_1, T_2) = E(T_1 - \tau(\theta))(T_2 - \tau(\theta)) = cv(\theta) \Rightarrow c = 1$ - QED

Теорема 2

Пусть T_1^* - оптимальная оценка для $\tau_1(\theta)$ и T_2^* - оптимальная оценка для $\tau_2(\theta)$

Тогда $\forall c_1, c_2\ \tilde{T} = c_1 T_1^* + c_2 T_2^*$ - оптимальная оценка для $\tau = c_1\tau_1(\theta) + c_2\tau_2(\theta)$

Доказательство

$\tilde{T}$ - статистика, несмещённая для τ

Пусть T - несмещённая оценка для $\tau(\theta)$

$D_\theta T = D_\theta(T - \tilde{T} + \tilde{T}) = D_\theta(T - \tilde{T}) + D_\theta\tilde{T} + 2cov(T - \tilde{T}, \tilde{T}) \geq D_\theta\tilde{T} + 2cov(T - \tilde{T}, \tilde{T})$

Надо доказать, что $cov(T - \tilde{T}, \tilde{T}) = 0$

Заметим, что $E_\theta(T - \tilde{T}) = 0$

$\widehat{T}_1 = T_1^* + \lambda(T - \tilde{T})$ - несмещённая оценка $\tau_1(\theta)$

$D_\theta\widehat{T}_1 \geq D_\theta T_1^*\ \forall\lambda \in \mathbb{R}^1$

$D_\theta\widehat{T}_1 = D_\theta T_1^* + \lambda^2 D_\theta(T - \tilde{T}) + 2\lambda cov(T_1^*, T - \tilde{T}) \geq D_\theta T_1^*$

$\lambda^2 D_\theta(T - \tilde{T}) + 2\lambda cov(T_1^*, T - \tilde{T}) \geq 0$ - имеет 2 корня и всегда неотрицателен $\Rightarrow$ корни совпадают $\Rightarrow$ $cov(T_1^*, T - \tilde{T}) = 0$

$cov(\tilde{T}, T - \tilde{T}) = c_1 cov(T_1^*, T - \tilde{T}) + c_2 cov(T_2^*, T - \tilde{T}) = 0$ - QED