Бутстреп

$X = (X_1, \dots, X_n) \sim P$ неизв.

$T(X)$ — статистика

Пример $T(X) = \dfrac{\overline{\cos X}}{\overline{X}}$

Хотим оценить дисперсию $T(X)$

Применим бутстреп

Схема:

1) Возьмём $j_1^i \ldots j_n^i \sim U\{1, \ldots, n\}$

2) Рассмотрим бутстрепную выборку $X_1^* = (X_{j_1^1} \ldots X_{j_n^1})$

$\vdots$

$X_B^* = (X_{j_1^B} \ldots B_{j_n^B})$

3) Считаем $T(X_1^*) \ldots T(X_B^*)$

4) Оцениваем искомую величину на $T(X_1^*) \ldots T(X_B^*)$

дисперсия: $\hat{\sigma}(T(X)) = \frac{1}{n} \sum T_i^2 - \left(\frac{1}{n} \sum T_i\right)^2$

Пример

Пусть $X = (5, 3, 2, 8, 0, 2, 7)$ $\qquad T(X) = \dfrac{\overline{\cos X}}{\overline{X}}$

$X_1^* = (2, 2, 7, 5, 3, 5, 0)$ $\qquad T_1 = 0.02$

$X_2^* = (0, 3, 3, 5, 7, 8, 0)$ $\qquad T_2 = 0.19$

$X_3^* = (2, 2, 8, 2, 2, 8, 0)$ $\qquad T_3 = -0.04$

$X_4^* = (3, 2, 2, 8, 3, 2, 2)$ $\qquad T_4 = 0.318$

Почему размер бутстрепной выборки $= n$?

$T(X) = \overline{X}$

Хотим дисперсию.

$D\overline{X} = D \frac{1}{n} \sum X_i = \frac{1}{n^2} D \sum X_i = \frac{1}{n} D X_1$

Пусть $T_1 - T_B$ — статистики по X_i^* размера k

$D T^* = D \overline{X}^* = \frac{1}{k} D X_1$

Если $T(X)$ „не зависит от n", то можно брать любого размера бутстрепную выборку.

Задача $X_1 \ldots X_n$ — выборка

$X_1^* \ldots X_n^*$ — бутстренная выборка

Найти $E\overline{X^*}$

Бутстреп: $X_1^* = X_1 I\{j_1 = 1\} + \ldots + X_n I\{j_1 = n\}$

$$E(X_1^* \mid X_1 \ldots X_n) = E(X_1 I\{j_1 = 1\} + \ldots + X_n I\{j_1 = n\} \mid X_1 \ldots X_n) =$$

$$= \sum_{i=1}^{n} E(X_i I\{j_1 = i\} \mid X_i) \underset{j_1 \perp\!\!\!\perp (X_1 \ldots X_n)}{=} \sum_{i=1}^{n} X_i E\, I\{j_1 = i\} =$$

$$= \frac{1}{n} \sum X_i = \overline{X}$$

$$E(\overline{X^*} \mid X_1 \ldots X_n) = E \frac{1}{n} \sum_{i=1}^{n} (X_i^* \mid X_1 \ldots X_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i^* \mid X_1 \ldots X_n) = \frac{1}{n} \sum_{i=1}^{n} \overline{X} = \overline{X}$$

$$E(\overline{X^*}) = E(E(\overline{X^*} \mid X_1 \dots X_n)) = \frac{1}{n}\sum_{i=1}^{n} EX_i = EX_1$$

$$E(\overline{X}) = EX_1$$

Каким брать B? [Кол-во бутстрепных выборок]

$\lambda = (X_1 \ldots X_n) \sim P$

1) Хотим сгенерировать $X' \sim P$

Для этого используем метод подстановки и будем генерировать

из $\hat{P}$ – эмпирическое распределение P по $X_1 \ldots X_n$

2) Когда оцениваем дисперсию $\Rightarrow$ используем Монте-Карло

оцениваемая величина

$\Theta \approx \hat{\Theta} \approx \tilde{\Theta}$

метод подстановки

Монте-Карло ск. сх. $\frac{1}{\sqrt{B}}$

ск. сх. $1/\sqrt{n}$ по т. Колмогорова – Смирнова: $\sup_A \sqrt{n}\,|P(A) - P_n(A)| \to \xi$

$\Rightarrow \frac{1}{\sqrt{B}} = o\left(\frac{1}{\sqrt{n}}\right)$

т.е $B \geq n^{1+eps}$

Метод подстановки и метод моментов

Вспомним метод моментов

$X = (X_1 \ldots X_n) \sim P \in \{P_\theta \mid \theta \in \Theta\}$

Пусть целевые функции $g_1 \ldots g_d$

Составляем систему $m(\theta) = \begin{pmatrix} E_\theta g_1(X) \\ \vdots \\ E_\theta g_d(X) \end{pmatrix} = \begin{pmatrix} \overline{g_1(X)} \\ \vdots \\ \overline{g_d(X)} \end{pmatrix}$

Пусть P — непарам.

Введём $G(P) = m^{-1}(E_P g_1(X) \ldots E_P g_d(X))$

$\forall \theta \in \Theta \quad G(P_\theta) = m^{-1}(m(\theta)) = \theta$

Тогда $\hat{\theta}$ – оценка по методу подстановки

$$\hat{\theta} = G(\hat{P}_\theta) = m^{-1}\left(\overline{g_1(X)}, \ldots, \overline{g_d(X)}\right)$$

Критерий хи-квадрат

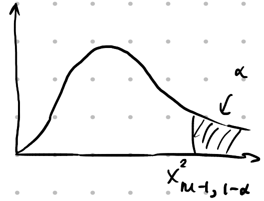
Пусть $X = (X_1 \ldots X_n) \sim P$

$H_0: P = P_0$ vs. $H_1: P \neq P_0$

Пусть $\mathcal{X} = \bigsqcup_{j=1}^{m} B_j$, тогда $\mu_j = \#\{i \mid X_i \in B_j\}$

$p_j^0 = P_0(X_1 \in B_j)$

Рассмотрим $\chi(X) = \sum_{j=1}^{m} \frac{(\mu_j - np_j^0)^2}{np_j^0} \xrightarrow{d} \chi^2_{m-1}$



Критерий $\{\chi(X) > \chi^2_{m-1,1-\alpha}\}$

Пример $X = (X_1 \ldots X_n)$ – выборка, сгенерированная псевдослучайным генератором чисел

$H_0: P = U\{1, 2, 3, 4\}$ vs. $H_1: P \neq U\{1, 2, 3, 4\}$

Проверить генератор.

Эксперимент:

#1	=	249
#2	=	254
#3	=	246
#4	=	251
Σ		1000

$B_j = \{j\} \quad j \in \overline{1,4}$

$p_j^0 = 1/4$

$$\Rightarrow \chi(X) = \frac{1 + 4^2 + 4^2 + 1}{250} = \frac{34}{250} \approx 0{,}136$$

Где отвергать? Если генератор даёт 1 2 3 4 по кругу, то

мы его не отвергнем

$\Rightarrow$ отвергаем справа и около 0, т.к. "генератор слишком идеальный"

Критерий: $S = \{ \chi(X) < \chi^2_{3, \alpha/2} \} \cup \{ \chi(X) > \chi^2_{3, 1-\alpha/2} \}$

Численные значения:

$\chi(X) = 0.136$ $\quad \chi^2_{3, 1-\alpha} = 7.81$ $\Rightarrow$ не отвергаем

$\chi^2_{3, 1-\alpha/2} = 9.35$ $\quad \chi^2_{3, \alpha/2} \approx 0.216$ $\Rightarrow$ отвергается

Вывод: генератор подозрительный.

Обобщённый критерий хи-квадрат

$X = (X_1 \dots X_n) \sim P \in \{P_\theta \mid \theta \in \Theta \in \mathbb{R}^d\}$

$H_0: \; P \in \mathcal{P}_\theta^0 \quad$ vs. $\quad H_1: \; P \in \mathcal{P}_\theta^1$

$\mathcal{X} = \bigcup_{j=1}^{m} B_j \qquad \mu_j = \#\{i \mid x_i \in B_j\} \qquad p_j^0(\theta) = P_\theta^0(X_1 \in B_j)$

$$\chi(X) = \sum_{j=1}^{m} \frac{(\mu_j - n p_j^0(\hat{\theta}))^2}{n p_j^0(\hat{\theta})} \xrightarrow{d_\theta} \chi^2_{m-d-1}$$

$\hat{\theta}$ — ОМП для $\;L_X(\theta) = \log \Big[\underbrace{\prod_{i=1}^{n} \prod_{j=1}^{m} p_j^0(\theta)^{I\{x_i \in B_j\}}}_{\text{категориальное распр.}}\Big] =$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_j \log p_j^0(\theta) \to \max_\theta$$

Критерий : $\{X(x) > \chi^2_{m-d-1,\, 1-\alpha}\}$

Задача Данные по бомбёжкам Лондона.

Поделили Лондон на сетку 24×24 $(= 576)$

Посчитали во сколько кварталов с какой частотой попадали бомбы

0	1	2	3	4	5	6	7	≥ 4
229	211	93	35	7	0	0	1	8

Используя обобщённый хи-квадрат, проанализировать случайно ли прилетают бомбы.

Если H_0 верна (бомбы случайны), то число событий (прилётов) имеет пуассоновское распределение.

Зачем-то хотим $np_j^0(\hat{\theta}) \geq 5$

$\mu_j \geq 5$

$$p_0^0(\theta) = e^{-\theta}$$

Pois $p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$

$$p_1^0(\theta) = \theta e^{-\theta}$$

$$p_2^0(\theta) = \frac{\theta^2}{2} e^{-\theta}$$

$$p_3^0(\theta) = \frac{\theta^3}{6} e^{-\theta}$$

$$p_4^0(\theta) = 1 - e^{-\theta}\left(1 + \theta + \frac{\theta^2}{2} + \frac{\theta^3}{6}\right)$$

Считаем ОМП :

$$L_x(\theta) = \sum_{j=0}^{4} \mu_j \log p_j^{\circ}(\theta) =$$

$$= 229(-\theta) + 211(-\theta + \log\theta) + 93(-\theta + 2\log\theta - \log 2) + 35(-\theta + 3\log\theta - \log 6) + 8\log p_4^{\circ}$$

$$\frac{\partial L_x(\theta)}{\partial\theta} = (-229 - 211 - 93 - 35) + \frac{211 + 186 + 105}{\theta} + \frac{e^{-\theta}\left(1+\theta+\frac{\theta^2}{2}+\frac{\theta^3}{6}\right) - e^{-\theta}\left(1+\theta+\frac{\theta^2}{2}\right)}{1 - e^{-\theta}\left(1+\theta+\frac{\theta^2}{2}+\frac{\theta^3}{6}\right)} =$$

$$= -568 + \frac{502}{\theta} + \frac{8\, e^{-\theta}\, \theta^3/6}{1 - e^{-\theta}\left(1+\theta+\frac{\theta^2}{2}+\frac{\theta^3}{6}\right)} = \text{-----}$$

$\hat{\theta} = 0.93$

$p_0^0(\hat{\theta}) = 0.384$

$p_1^0(\hat{\theta}) = 0.367$

$p_2^0(\hat{\theta}) = 0.171$

$p_3^0(\hat{\theta}) = 0.053$

$p_4^0(\hat{\theta}) = 0.015$

$\chi^2_{3, 1-\alpha} = 7.81$

$X(x) = 1.17$

$\Rightarrow$ не отвергается

p-value $= 0.759$