$\boxed{\text{Д/З №9}}$

$\boxed{№1}$

$$X_1 \ldots X_n \sim P$$

Рассм $\hat{P}$ – эмп. распр. с $\hat{F}_n(x) = \frac{1}{n} \sum I\{X_i \le x\}$

Найдём $\operatorname{cov}(\hat{F}_n(x), \hat{F}_n(y)) =$, положим $x \le y$

$$= \frac{1}{n^2} \operatorname{cov}\left(\sum I\{X_i \le x\}, \sum I\{X_j \le y\}\right) =$$

$$= \frac{1}{n^2}\left[\underbrace{\sum_{i \ne j} \operatorname{cov}(I\{X_i \le x\}, I\{X_j \le y\})}_{\substack{= P(X_i \le x,\ X_j \le y) - P(X_i \le x) P(X_j \le y) \\ = 0,\ \text{т.к. } X_i \text{ и } X_j \text{ независ.}}} + \sum_k \operatorname{cov}(I\{X_k \le x\}, I\{X_k \le y\})\right] =$$

$$= \frac{1}{n^2} \sum_{k=1}^{n} P(X_k \le x, X_k \le y) - P(X_k \le x) P(X_k \le y) =$$

$$= \frac{1}{n^2} \sum P(X_k \le x) - P(X_k \le x) P(X_k \le y) =$$

$$= \frac{1}{n^2} \sum P(X_k \le x) P(X_k > y) = \frac{1}{n} P(X_1 \le x) P(X_1 > y)$$

$\boxed{№ 2}$ $X_1 \dots X_n$ – выборка из неизв. распр. P.

$$\boxed{\gamma = \frac{1}{\sigma^4} E(X-a)^4 - 3}, \text{ где } \sigma^2 = DX = E(X-a)^2$$

коэф. эксц.

$$\hat{\gamma} = \frac{1}{\hat{\sigma}^4} \int_{\mathbb{R}} (x-\hat{a})^4 \, dF_n(x) - 3 = \frac{1}{n\hat{\sigma}^4} \sum_{i=1}^{n} (X_i - \hat{a})^4$$

где $\hat{\sigma}^2 = \overline{X^2} - \overline{X}^2$

$\hat{a} = \overline{X}$

Схема бутстрепа для оценки дисп. получ. оценки

т.е. для подсчёта $D_{\hat{P}_n} \hat{\gamma}$

1) B раз генерируем

$i_1 \dots i_n \sim U\{1 \dots n\} \Rightarrow X^* = (X_{i_1} \dots X_{i_n})$ – бутстреп. выборка

Назовём выборки $X_1^* \dots X_B^*$.

2) Посчитаем на выборках величину $\hat{\gamma}$. Получим $\hat{\gamma}_1 \dots \hat{\gamma}_B$

3) найдём выб. дисперс.

$$D_{\hat{P}_n} \hat{\gamma} \approx \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}_b^2 - \left(\frac{1}{B} \sum \hat{\gamma}_b\right)^2$$

$\boxed{№3}$ $X_1 \ldots X_n$ - реализ выборки

$X_1^* \ldots X_n^*$ - нек. бутстр выб по ней

Обозначим за k_i - кол-во вхождений X_i в бутстр. выборку

$$P(k_i > 0) = 1 - P(k_i = 0) = \boxed{1 - \left(\frac{n-1}{n}\right)^n}$$

Рассм. $Y_1 \ldots Y_n$ - случ. величина

$$Y_i \sim Bern\left(1 - \left(\frac{n-1}{n}\right)^n\right)$$

$\sum_{i=1}^{n} Y_i$ - число уник. элементов в бутстр. выб.

$$E \sum_{i=1}^{n} Y_i = \boxed{n\left(1 - \left(\frac{n-1}{n}\right)^n\right)}$$

Посмотрим куда сход. величина $1 - \left(\frac{n-1}{n}\right)^n$

$$\left(\frac{n-1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n \to e^{-1} \approx 0{,}367$$

Откуда $1 - \left(\frac{n-1}{n}\right)^n \to 1 - e^{-1} \approx 0{,}633$

Значит в среднем при больших n, доля уник. элементов: 63,3%

$\boxed{№4}$ $X_1 \dots X_n \sim U[0, \theta]$

$\hat{\theta} = X_{(n)}$

Почему метод бутстр. для оценки распр. $\hat{\theta}$ раб. плохо?

① Поймём что значит "плохо" работает.

Пусть $X_1^* \dots X_B^*$ – бутстр. выборки

$X_{(n)}^1 \dots X_{(n)}^B$ – соотв. максимумы

Заметим $X_{(n)}^b \leq X_{(n)}$

Притом более чем в 30% случаев $X_{(n)}$ не попадает в бутстр. выборку.

В таких случаях $X_{(n)}^b \leq X_{(n-1)}$

Иными словами, если изнач. выборка чуть-чуть неудачная (иными словами $X_{(n)} \leq \frac{n}{n+1}$), получим:

$E_{boots} X_{(n)} = \left(\sum_{b}^{B} X_{(n)}^b\right) \cdot \frac{1}{B} \leq E X_{(n)}$

Ан-но с другими параметрами распр.

$P_{boot} X_{(n)}$, K , γ

② Бутстреп лучше работает с оценками вроде средних, т.к. вся выборка непоср. влияет на оценку, которые бутстрепные оценки в обе стороны от настоящей.

$\Rightarrow \frac{2}{3}\left(-\frac{1}{n-1}\right) + \frac{2}{9}\left(-\frac{1}{n}\right) + \frac{2}{27}\left(-\frac{1}{n-1}\right)$

№5 $X_1 \dots X_n$ – выборка | $X_1^* \dots X_n^*$ – бутстреп.

$$\overline{X^*} = \frac{1}{n}\sum_{i=1}^{n} X_i^*$$

① $D\left(\frac{1}{n}\sum_{i=1}^{n} X_i^* \mid X_1 \dots X_n\right) = \frac{1}{n^2} D\left(\sum X_i^* \mid X_1 \dots X_n\right) \circled{=}$

$$E\left(\sum X_i^* \mid X_1 \dots X_n\right) = \sum E\left(X_i^* \mid X_1 \dots X_n\right) = n\overline{X}$$

$$E\left(\left(\sum X_i^*\right)^2 \mid X_1 \dots X_n\right) = \sum_{i \neq j} E\left(X_i^* X_j^* \mid X_1 \dots X_n\right) + n\overline{X^2}$$

$$= n(n-1)\sum_{a,b}\frac{X_a X_b}{n^2} = \frac{(n-1)}{n}\sum_{a,b} X_a X_b + n\overline{X^2}$$

$$+ n\overline{X^2}$$

$$= \frac{n-1}{n}\left(\sum X_i\right)^2 + n\overline{X^2} = n(n-1)\overline{X}^2 + n\overline{X^2}$$

$\circled{=} \frac{1}{n^2}\left(n(n-1)\overline{X}^2 + n\overline{X^2} - n^2\overline{X}^2\right) = \frac{1}{n}\overline{X^2} - \frac{1}{n}\overline{X}^2 =$

$$= \frac{\overline{X^2} - \overline{X}^2}{n}$$

② $D\left(\frac{1}{n}\sum_{i=1}^{n} X_i^*\right) = \cancel{\frac{1}{n} D\left(\sum_{i=1}^{n} X_i^*\right)}\ \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i^*\right) \circled{=}$

$$E\left(\sum X_i^*\right) = E\left(E\left(\sum X_i^* \mid X_1 \dots X_n\right)\right) = nEX_1$$

$$E\left(\left(\sum X_i^*\right)^2\right) = E\left(E\left(\left(\sum X_i^*\right)^2 \mid X_1 \dots X_n\right)\right) = n(n-1)(EX)^2 + nEX^2$$

$\circled{=} \frac{1}{n^2}\left(n(n-1)(EX)^2 + nEX^2 - n^2(EX)^2\right)$

$$= \frac{DX}{n}$$

$\boxed{№6}$ $\mathcal{X} = \{1, 2, 3, 4\}$

P_θ: $P(1) = \theta$, $P(2) = \theta$, $P(3) = 2\theta$, $P(4) = 1 - 4\theta$

H_0: $p \in \{P_\theta \mid \theta \in (0, \frac{1}{4})\}$

Воспольз. крит. хи-квадрат (обобщённый):

Найдём $\hat{\theta}$ из усл $\sum \mu_j \ln p_j(\theta) \to max$

$$f(\theta) = 1010 \ln\theta + 2200 \cdot \ln(\theta) + 950 \cdot \ln 2\theta + 840 \cdot \ln(1-4\theta)$$

$$f' = \frac{1010}{\theta} + \frac{2200}{\theta} + \frac{950}{2\theta} \cdot 2 + \frac{840}{1-4\theta} \cdot (-4) = 0$$

$$4160 \cdot (1-4\theta) - 3360 \cdot \theta = 0$$

$$\theta = \frac{26}{125} < \frac{1}{4}$$

$$\hat{\chi} = \frac{\left(1010 - 5000 \cdot \frac{26}{125}\right)^2}{5000 \cdot \frac{26}{125}} + \frac{\left(2200 - 5000 \cdot \frac{26}{125}\right)^2}{5000 \cdot \frac{26}{125}} + \frac{\left(950 - 5000 \cdot \frac{52}{125}\right)^2}{5000 \cdot \frac{52}{125}} + \frac{\left(840 - 5000 \cdot \frac{21}{125}\right)^2}{5000 \cdot \frac{21}{125}}$$

$$= \frac{(1010 - 1040)^2}{1040} + \frac{(2200 - 1040)^2}{1040} + \frac{(950 - 2080)^2}{2080} + \frac{(840 - 840)^2}{840} =$$

$$= \frac{1800 + 1300 + 1276900}{2080} = 615,625$$

Критерий отв-я: $\{\chi \geq \chi^2_{4-1-1,\ 0,95}\}$

$\{\chi \geq 6\} \Rightarrow$ ОТВЕРГАЕМ (p-value ≈ 0)

№7 Пусть имеем P_1 и P_2 – два разл. распр.

Им соотв. F_1 и F_2 – две функции распр.

$\beta(P_2) = P_2(X \in S) \to 1$ – цель док-ва.

а) Критерий: $S = \{ \sqrt{n} \sup\limits_{x \in R} |\hat{F}_n(x) - F_2(x)| \geq C_\alpha \}$

$F_1 \neq F_2 \Rightarrow \exists x = [redacted]$, где б.о.о. $\underline{\underline{F_2 > F_1 + \varepsilon}}$, $\varepsilon > 0$.

Заметим по теор. Гливенко – Кантели для почти всех X выборках

$\exists N \; \forall n \geq N: |\hat{F}_n(x) - F_1(x)| < \frac{\varepsilon}{2}$

А значит на $x = [redacted]$ $\underline{\underline{F_2 > \hat{F}_n + \frac{\varepsilon}{2}}}$ при $n \geq N$

Тогда $\sup |\hat{F}_n - F_2| \geq \frac{\varepsilon}{2}$

Тогда $\sqrt{n} \sup |\hat{F}_n - F_2| \geq C_\alpha$

для дост. больш. n.

Им $P(\sqrt{n} \sup |\hat{F}_n - F_2| \geq C_\alpha) = 1$ при $n \geq N$

ч.т.д.

Очев-но $\beta(P_2) = P_2(x \in S) \to 1$

Б) Критерий хи-квадрат:

$$\left\{ \sum_{j=1}^{k} \frac{(\mu_j - np_j^{(0)})^2}{np_j^0} \geq \chi^2_{k-1,\,1-\alpha} \right\}$$

Пусть $P_2(x_1 \in B_j) = p_j^{(2)}$

$P_1(x_1 \in B_j) = p_j^{(1)}$

Пусть б.о.о. $p_1^{(1)} \neq p_1^{(2)}$ (и также б.о.о. $p_1^{(2)} > p_1^{(1)}$)

Утв-ие $P\left(\frac{(\mu_1 - np_1^{(1)})^2}{np_1^{(1)}} \geq c\right) \to 1$, $c := \chi_{k-1,\,1-\alpha}$

где $\mu_1 = \#\{i \mid x_i \in B_1\}$ и $P(B_1) = p_1^{(2)}$

Выполн. из ЗБЧ
$\mu_1 \sim Bin(n, p^{(2)})$
$E\mu_1 = np^{(2)} > n\left(\frac{p^{(2)}+p^{(1)}}{2}\right)$

~~$P(\mu_1 \geq np^{(1)} + \sqrt{cnp^{(1)}})$~~

$P\left((\mu_1 - np_1^{(1)})^2 \geq cnp^{(1)}\right)$,

Зафикс. $\delta < 1$. Заметим, что для дост. больших n:

$P\left(\mu_1 > n\left(\frac{p_1^{(1)}+p_1^{(2)}}{2}\right)\right) \geq \delta$ (т.е. почти всегда)

$P\left(\mu_1 - np_1^{(1)} > n\,\frac{p_1^{(2)} - p_1^{(1)}}{2}\right) \geq \delta$

$P\left((\mu_1 - np_1^{(1)})^2 > n^2 \cdot z\right) \geq \delta$, $z := \frac{p_1^{(2)} - p_1^{(1)}}{2}$

А значит для дост. большого n (т.к. правая часть квадратич.)

$P\left((\mu_1 - np_1^{(1)})^2 \geq cnp^{(1)}\right) \geq \delta$.

Формулировки:

Критерии согласия – те, что проверяют принадл-ть известному классу распр (напр. $\{P_\theta\}$ или нормальн или гамма-р.)

• Теорема Колмогорова

$X = (X_1 \dots X_n)$ – выборка из неизв. распр

$H_0: F = F_0$

$D_n = \sup\limits_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$ – статистика критерия

$\sqrt{n} D_n \xrightarrow{d} \xi \sim K$ – распр Колм., если F_0 – непр

• Опр: $\hat{P}_n$ наз-ся эмпир. распр. по выборке X, если $\forall B \in \mathcal{B}$: $\hat{P}_n(B) = \frac{1}{n} \sum\limits_{i=1}^{n} I\{X_i \in B\}$

• Теор. Гливенко-Кантелли

Если F – ф.р. для P и $X = (X_1 \dots X_n)$ – выб. из распр P

то $D_n = \sup\limits_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{п.н.} 0$

• Критерий Колмогорова $\{\sqrt{n} D_n \geq C_\alpha\}$, где C_α – $(1-\alpha)$ квантиль K-распр. Колм

• Критерий хи-квадрат.

$X = (X_1 \dots X_n)$ – выб. из неизв. распр P, $X_i \in \mathcal{X}$

$H_0: P = P_0$

Разбиение $\mathcal{X} = \bigsqcup\limits_{j=1}^{k} B_j$, статистика: $\hat{\chi} = \sum\limits_{j=1}^{k} \frac{(\mu_j - np_j^0)^2}{np_j^0}$

Критерий: $\{\hat{\chi} \geq \chi^2_{k-1, 1-\alpha}\}$

Теорема Пирсона

$\hat{\chi} \xrightarrow{d} \chi^2_{k-1}$