Прикладная статистика: задание 2 **5 июля 2021**

Домашнее задание.
Кирилл Сетдеков

Задачи:

1. С помощью неравенства Чебышёва покажите,с какой вероятностью величина лежит $[\mu - 2\sigma, \mu + 2\sigma]$ и $[\mu - 3\sigma, \mu + 3\sigma]$. Сравните полученные вероятности с соответствующими вероятностями для стандартного нормального распределения N (0, 1). Какой вывод можно сделать?

   Решение:
   Формулировка неравенства Чебышева:

   $$P(|X - \mu| \geqslant a) \leqslant \frac{\sigma^2}{a^2}$$

   Нас интересует обратная вероятность:

   $$P(|X - \mu| < a) > 1 - \frac{\sigma^2}{a^2}$$

   Подставим значения для 2 сигм, заменив $a = 2\sigma$:

   $$P(|X - \mu| < 2\sigma) > 1 - \frac{\sigma^2}{4\sigma^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0.75$$

   Из нормального распределения: $P(|X - \mu| < 2\sigma) = \Phi(2) - \Phi(-2) = 0.9545$

   Аналогично для 3 сигм, заменив $a = 3\sigma$:

   $$P(|X - \mu| < 3\sigma) > 1 - \frac{\sigma^2}{9\sigma^2} = 1 - \frac{1}{9} = \frac{8}{9} \approx 0.889$$

   Из нормального распределения: $P(|X - \mu| < 3\sigma) = \Phi(3) - \Phi(-3) = 0.9973$

   Ответ: с помощью неравенства Чебышёва оценки для этого интервала имеют меньшую вероятность, чем вероятности, используя функцию распределения стандартного нормального распределения. $\Rightarrow$ если мы знаем форму распределения, лучше считать через его квантили так как оценка будет точнее, а неравенства Чебышёва использовать в случаях, когда мы совсем ничего не знаем про распределение

2. Пусть есть реализация выборки $x_1, ..., x_n$ из $Unif([0, \theta])$ с неизвестным θ.

2.1.

$x_i \sim Unif(0, \theta) \Rightarrow y_i = \frac{x_i}{\theta} \sim Unif(0, 1)$

$Y_{(n)} = \max\{y_1, \dots y_n\} = \frac{\hat{\theta}_1}{\theta} \sim F(u)$ – из условия

$P_\theta(g_1 < F(u) < g_2) = P_\theta\left(g_1 < \frac{\hat{\theta}_1}{\theta} < g_2\right) =$

$= P_\theta\left(\frac{\hat{\theta}_1}{g_2} < \theta < \frac{\hat{\theta}_1}{g_1}\right)$

$f(u)$ на $[0, 1]$ $= n u^{n-1}$ и возрастает $\Rightarrow$

ближе всего g_1 и g_2 при $\to$ к 1 $g_2 = 1$

$P_\theta(g_1 < Y_n < 1) = F(1) - F(g_1) = 1 - g_1^n = 1 - \alpha$

$g_1 = \sqrt[n]{\alpha}$

$\Downarrow$

$\cancel{1-\alpha}$ $P_\theta(\sqrt[n]{\alpha} < F(u) < 1) = P_\theta\left(\hat{\theta}_1 < \theta < \frac{\hat{\theta}_1}{\sqrt[n]{\alpha}}\right) = 1 - \alpha$

интервал от $\hat{\theta}_1$ до $\frac{\hat{\theta}_1}{\sqrt[n]{\alpha}}$

2.2) $n \to \infty$, то $\dfrac{n\overline{X} - nEx_i}{\sqrt{nDx_i}} = \dfrac{n\overline{X} - n\frac{\theta}{2}}{\sqrt{n\frac{\theta^2}{12}}} =$

$$= \frac{\sqrt{3n}\,(\overbrace{2\overline{X}}^{\hat{\theta}_2} - \theta)}{\theta} \sim N(0,1)$$

$$P_\theta\left(-c_{1-\alpha/2} < \frac{\sqrt{3n}\,(\hat{\theta}_2 - \theta)}{\theta} < c_{1-\frac{\alpha}{2}}\right) \to 1-\alpha \;\Big|\; n \to \infty$$

$$P_\theta\left(\frac{-c_{1-\alpha/2} + \sqrt{3n}}{\sqrt{3n}\,\hat{\theta}_2} < \frac{1}{\theta} < \frac{c_{1-\frac{\alpha}{2}} + \sqrt{3n}}{\sqrt{3n}\,\hat{\theta}_2}\right) \to 1-\alpha$$

$$P_\theta\left(\frac{\sqrt{3n}\,\hat{\theta}_2}{c_{1-\alpha/2} + \sqrt{3n}} < \theta < \frac{\sqrt{3n}\,\hat{\theta}_2}{-c_{1-\alpha/2} + \sqrt{3n}}\right) \to 1-\alpha$$

наш доверит. интервал

3. Про выборки, про сравнение дисперсий.

Решение:
Исходное равномерное распределение $Unif[0, \theta] \sim X$, для него $EX = \theta/2, DX = \frac{\theta^2}{12}$.
Мы оценили $\hat{\theta} = 2\overline{X}$ и сделали параметрический бутстрэп:

$$Y \sim Unif[0, \hat{\theta}]; Y = 2\overline{X}R_1$$

, где $R_1 \sim Unif[0, 1]$.

Мы знаем, что $D\overline{X} = \frac{\theta^2}{12\sqrt{n}}$, $DX = \frac{\theta^2}{12}$, $DR_1 = \frac{1}{12}$, $ER_1 = \frac{1}{2}$, $E\overline{X} = \frac{\theta}{2}$. Воспользуемся формулой для произведения дисперсии и запишем дисперсию Y:

$$D[2\overline{X}R_1] = 4D[\overline{X}R_1] = 4\left[D\overline{X}DR_1 + \left(E\overline{X}\right)^2 DR_1 + (ER_1)^2 D\overline{X}\right] =$$

$$= 4\left[\frac{\theta^2}{12\sqrt{n}}\frac{1}{12} + \frac{\theta^2}{4}\frac{1}{12} + \frac{1}{4}\frac{\theta^2}{12\sqrt{n}}\right] = \frac{\theta^2}{12} + \frac{\theta^2}{9\sqrt{n}}$$

Мы получили, что $DY = \frac{\theta^2}{12} + \frac{\theta^2}{9\sqrt{n}}$, а $DX = \frac{\theta^2}{12}$. Получается, что $DY > DX, \forall n \geqslant 1$, при этом, при $n \to \infty, DY \to DX$, то есть в бесконечно большой выборке дисперсия из параметрического бутстрэпа будет приближаться к дисперсии из оригинального равномерного распределения.

4. Про оценки для среднего в нормальном распределении, реализация на питоне.

Я решил и описал это задание в приложенном ноутбуке.