Математическая статистика. ДЗ 2.

Прохоров Юрий, 776

Задача 1

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — простая выборка из неизвестного непрерывного распределения $F(x)$. Доказать, что распределение случайной величины

$$Z_n = \int\limits_{-\infty}^{+\infty} \left(\widehat{F}_n(x) - F(x)\right)^2 dF(x),$$

где $\widehat{F}_n(x) = \dfrac{1}{n}\sum\limits_{k=1}^{n} \mathbb{I}\big\{X_k < x\big\}$ — эмпирическая функция распределения, не зависит от функции $F(x)$.

Так как противного не сказано, будем понимать данный интеграл как интеграл Римана-Стилтьеса по траекториям подынтегральной случайной функции.

Решение:

Введем случайные величины $Y_k = F(X_k)$. Найдем их распределение:

$$\begin{aligned}
x \le 0 \qquad &\Longrightarrow \qquad F_{Y_k}(x) = \mathbb{P}\{Y_k < x\} = 0 \\
0 < x < 1 \qquad &\Longrightarrow \qquad F_{Y_k}(x) = \mathbb{P}\big\{F^{-1}(Y_k) < F^{-1}(x)\big\} = \mathbb{P}\{X_k < F^{-1}(x)\} = F(F^{-1}(x)) = x \\
x \ge 1 \qquad &\Longrightarrow \qquad F_{Y_k}(x) = \mathbb{P}\{Y_k < x\} = 1
\end{aligned}$$

Если функция $F(x)$ не строго монотонна, то под обратной функцией $F^{-1}(y)$ подразумевается

$$F^{-1}(y) = \inf\{x \mid F(x) = y\}$$

Таким образом, $Y_k \sim \mathcal{U}[0,1]$. С учетом этого перепишем случайную величину Z:

$$Z_n = \int\limits_{-\infty}^{+\infty} \left(\frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\big\{X_k < x\big\} - F(x)\right)^2 dF(x) = \int\limits_{-\infty}^{+\infty} \left(\frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\big\{Y_k < F(x)\big\} - F(x)\right)^2 dF(x) =$$

$$= \Bigg/ \begin{array}{c}\text{замена:}\\ y = F(x)\end{array} \Bigg/ = \int\limits_{0}^{1} \left(\frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\big\{Y_k < y\big\} - y\right)^2 dy$$

Видно, что это выражение уже не зависит от F, что и требовалось доказать.

Оказывается, что для случайной величины Z_n выполнено

$$nZ_n \xrightarrow[n\to\infty]{d} \int\limits_0^1 B^2(t)\,dt, \qquad B(t) = W(t) - t\cdot W(1),$$

где $B(t)$ — броуновский мост, $W(t)$ — винеровский процесс, а интеграл понимается как стохастический интеграл Римана.

Распределение квадратичного интеграла броуновского моста табулировано, про него можно прочитать в статье “On the Distribution of the Square Integral of the Brownian Bridge”, Leonid Tolmatz, 2002.

Задача 2

(a) Пусть $\xi_1, \ldots, \xi_n \sim \mathrm{Be}(p)$ — i.i.d., $S_n = \xi_1 + \ldots + \xi_n$. Доказать, что при $z \geq 0$:

$$\mathbb{P}\big\{S_n - np \geq z\big\} \leq \exp\left\{-nH\left(p + \frac{z}{n}\right)\right\} \leq \exp\left\{-\frac{2z^2}{n}\right\},$$

$$\mathbb{P}\big\{S_n - np \leq -z\big\} \leq \exp\left\{-nH\left(p - \frac{z}{n}\right)\right\} \leq \exp\left\{-\frac{2z^2}{n}\right\},$$

где функция $H(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}$.

(b) Пусть $\widehat{F}_n$ — эмпирическая функция распределения произвольной случайной величины с теоретической функцией распределения F.

Оценить вероятность
$$\mathbb{P}\Big\{\sqrt{n}\,\big|\widehat{F}_n(x) - F(x)\big| \geq \Delta\Big\}$$

Решение:

(a) Докажем первое неравенство.

Для произвольного $\lambda > 0$:

$$\mathbb{P}\{S_n - np \geq z\} = \mathbb{P}\big\{e^{S_n} \geq e^{np+z}\big\} = \mathbb{P}\big\{e^{\lambda S_n} \geq e^{\lambda(np+z)}\big\} = \Big/ \begin{array}{c}\text{неравенство}\\ \text{Маркова}\end{array} \Big/ \leq$$

$$\leq \frac{\mathbb{E}e^{\lambda S_n}}{e^{\lambda(np+z)}} = \Big/ \begin{array}{c}S_n\text{ — сумма}\\ \text{независимых с.в.}\end{array} \Big/ = \frac{\prod\limits_{k=1}^{n} \mathbb{E}e^{\lambda \xi_k}}{e^{\lambda(np+z)}} = \frac{\big[1 + p(e^{\lambda} - 1)\big]^n}{e^{\lambda(np+z)}} =$$
$$= \left[\frac{1 + p(e^{\lambda} - 1)}{e^{\lambda\left(p + \frac{z}{n}\right)}}\right]^n = f(\lambda)$$

Отсюда следует, что

$$\mathbb{P}\{S_n - np \geq z\} \leq \inf_{\lambda > 0} \left[\frac{1 + p(e^{\lambda} - 1)}{e^{\lambda\left(p + \frac{z}{n}\right)}}\right]^n = \inf_{\lambda > 0} \Big[q e^{-\lambda(p+\alpha)} + p e^{\lambda(q-\alpha)}\Big]^n = \inf_{\lambda > 0} f(\lambda),$$

где $q = 1 - p,\ \alpha = \frac{z}{n}$.
По сути мы применили сейчас неравенство Чернова.

Из условия $f'(\lambda) = 0$ находим минимум

$$\lambda^* = \ln \frac{q(p+\alpha)}{p(q-\alpha)}, \qquad f(\lambda^*) = \left(q\left[\frac{q(p+\alpha)}{p(q-\alpha)}\right]^{-(p+\alpha)} + p\left[\frac{q(p+\alpha)}{p(q-\alpha)}\right]^{q-\alpha}\right)^n = \left[\frac{p^{p+\alpha}\, q^{q-\alpha}}{(p+\alpha)^{p+\alpha}\,(q-\alpha)^{q-\alpha}}\right]^n$$

После преобразований получаем
$$\mathbb{P}\{S_n - np \geq z\} \leq e^{-nH(p+\alpha)}$$

Для дальнейшей оценки, заметим, что

$$H(p) = H'(p) = 0, \qquad H''(x) = \frac{1}{x(1-x)} \leq 4, \qquad H(p+\alpha) \geq \frac{1}{2} \cdot \max|H''| \cdot \alpha^2 = 2\alpha^2$$

Строгое доказательство:

$$\frac{d}{d\alpha}\big[H(p+\alpha) - 2\alpha^2\big] = 0 \qquad \Longrightarrow \qquad \varphi(\alpha) \equiv H'(p+\alpha) - 4\alpha = \ln\frac{p+\alpha}{p} - \ln\frac{1-p-\alpha}{1-p} - 4\alpha = 0$$

$$\left.\begin{array}{r}\varphi'(\alpha) = H''(p+\alpha) - 4 \geq 0\\ \varphi(-p) = -\infty\\ \varphi(1-p) = +\infty\end{array}\right\} \qquad \Longrightarrow \qquad \begin{array}{c}\text{по теореме о промежуточном значении:}\\ \exists!\text{ решение уравнения } \varphi(\alpha) = 0\end{array}$$

Это решение можно угадать: $\alpha^* = 0$, поэтому

$$H(p+\alpha) - 2\alpha^2 \geq H(p+\alpha^*) - 2(\alpha^*)^2 = H(p) = 0$$

Наконец, получаем требуемое неравенство

$$\mathbb{P}\{S_n - np \geq z\} \leq e^{-nH(p+\alpha)} \leq e^{-2\alpha^2 n}$$

Второе неравенство получается аналогично.

(b) Пусть $\widehat{F}_n$ строится по выборке $X_1, \ldots, X_n$. Обозначим случайные величины

$$\xi_k = \mathbb{I}\{X_k < x\} \sim \mathrm{Be}\big(F(x)\big) = \mathrm{Be}(p) \text{ — i.i.d.}$$

$$S_n = \xi_1 + \ldots + \xi_n$$

Тогда можно воспользоваться пунктом (a):

$$\mathbb{P}\Big\{\sqrt{n}\,\big|\widehat{F}_n(x) - F(x)\big| \geq \Delta\Big\} = \mathbb{P}\big\{|S_n - np| \geq \sqrt{n}\Delta\big\} \leq 2e^{-2\Delta^2}$$