Оценка распределения и статистические функционалы

а) Эмпирическая функция распределения

Пусть задана i.i.d. выборка $X_1, \ldots, X_n \sim F$

Необходимо оценить распределение F

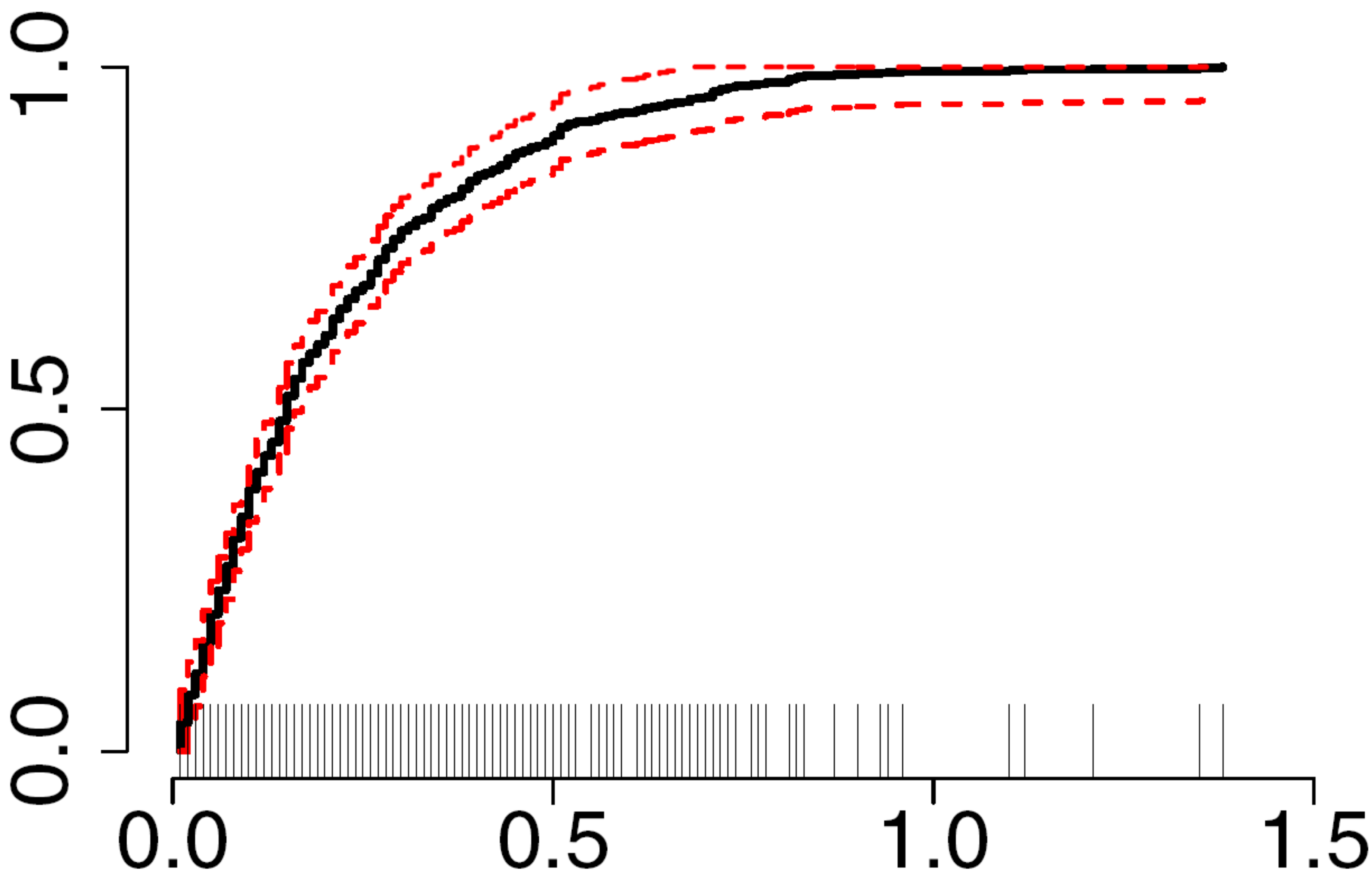
1 **Эмпирическая функция распределения** $\widehat{F}_n$ **имеет вид**

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

, где

$$I(X_i \leq x) = \begin{cases} 1, X_i \leq x \\ 0, X_i > x \end{cases}$$

2



Эмпирическая функция распределения. Моменты между последовательными импульсами вдоль нервного волокна

<u>Утверждение</u>. Для любого фиксированного x

$$\mathbb{E}\left(\widehat{F}_n(x)\right) = F(x),$$

$$\mathbb{V}\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n},$$

$$\mathrm{MSE} = \frac{F(x)(1-F(x))}{n} \to 0,$$

$$\widehat{F}_n(x) \xrightarrow{\mathrm{P}} F(x).$$

<u>Утверждение</u> (Th. Гливенко-Кантелли). Пусть $X_1, \ldots, X_n \sim F$ **- i.i.d. выборка. Тогда**

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\mathrm{P}} 0$$

<u>Утверждение</u> (Неравенство Dvoretzky, Kiefer, Wolfowitz). Пусть $X_1, \ldots, X_n \sim F$ **- i.i.d. выборка. Тогда для любого** $\epsilon > 0$

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Пусть

4

$$\begin{aligned} L(x) &= \max\{\widehat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\widehat{F}_n(x) + \epsilon_n, 1\} \\ \epsilon_n &= \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)} \end{aligned}$$

для заданного значения α **. Тогда для любого распределения** F

$$\mathbb{P}\left(L(x) \leq F(x) \leq U(x) \quad \textit{для } \forall x \right) \geq 1 - \alpha$$

b) Статистические функционалы

Пусть задана i.i.d. выборка $X_1, \ldots, X_n \sim F$

Статистическим функционалом $T(F)$ **называется любая функция от** F**, например,**

5

$$\mu = \int x \, dF(x)$$

$$\sigma^2 = \int (x - \mu)^2 \, dF(x)$$

$$m = F^{-1}(1/2)$$

Оценка $T(F)$ **получается с помощью величины** $\widehat{\theta}_n = T(\widehat{F}_n)$

Функционалы вида $T(F) = \int r(x) dF(x)$ **для некоторой функции** $r(x)$ **называются линейными статистическими функционалами, поскольку**

$$T(aF + bG) = aT(F) + bT(G)$$

Для линейных статистических функционалов

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

Пусть se **- стандартное отклонение для** $T(\widehat{F}_n)$**, а** $\widehat{\mathsf{se}}$ **- его оценка.**
Тогда во многих случаях

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2)$$

Приближенный доверительный интервал с доверительной вероятностью $1-\alpha$ **для** $T(F)$ **будет иметь вид**

$$T(\widehat{F}_n) \pm z_{\alpha/2}\,\widehat{\mathsf{se}}$$

<u>Пример</u> (среднее значение). Пусть $\mu = T(F) = \int x\, dF(x)$**. Тогда**

$$\widehat{\mu} = \int x\, d\widehat{F}_n(x) = \overline{X}_n$$

$$\mathsf{se} = \sqrt{\mathbb{V}(\overline{X}_n)} = \sigma/\sqrt{n}$$

$$\overline{X}_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}$$

<u>Пример</u> (дисперсия).

7 **Пусть** $\sigma^2 = T(F) = \mathbb{V}(X) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$**. Тогда**

$$\begin{aligned} \widehat{\sigma}^2 &= \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \end{aligned}$$

Однако полученная оценка является смещенной. Несмещенная оценка имеет вид

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

<u>Пример</u> (коэффициент асимметрии).

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}$$

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_i (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}$$

<u>Пример</u> (корреляция). Пусть $Z = (X, Y)$ **- двумерная случайная величина. Определим корреляцию по формуле**

$$\rho = T(F) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)/(\sigma_x \sigma_y),$$

где $F(x, y)$ **- двумерная функция распределения**

$$T_1(F) = \int x\, dF(z), \quad T_2(F) = \int y\, dF(z), \quad T_3(F) = \int xy\, dF(z)$$

9 $$T_4(F) = \int x^2\, dF(z), \quad T_5(F) = \int y^2\, dF(z),$$

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}$$

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n))$$

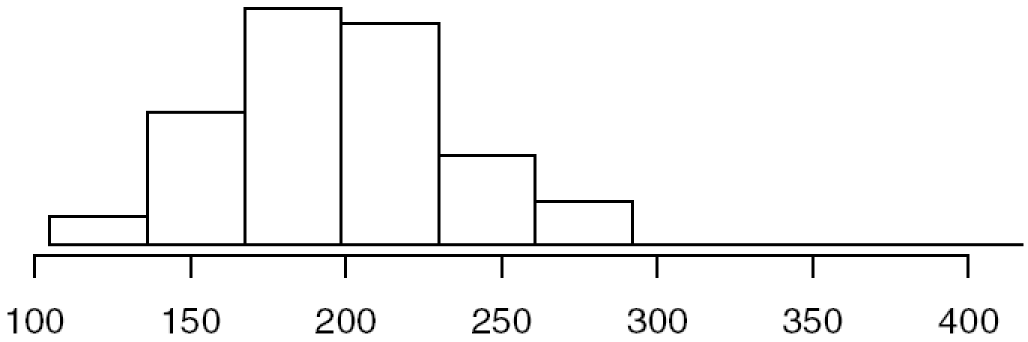
$$\widehat{\rho} = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_i (X_i - \overline{X}_n)^2}\sqrt{\sum_i (Y_i - \overline{Y}_n)^2}}$$

Пример (квантиль). Пусть распределение F строго возрастает с плотностью f. p-квантилью при $0 < p < 1$ распределения F называется величина $T(F) = F^{-1}(p)$. Поскольку $\widehat{F}_n$ - необратимая функция, то

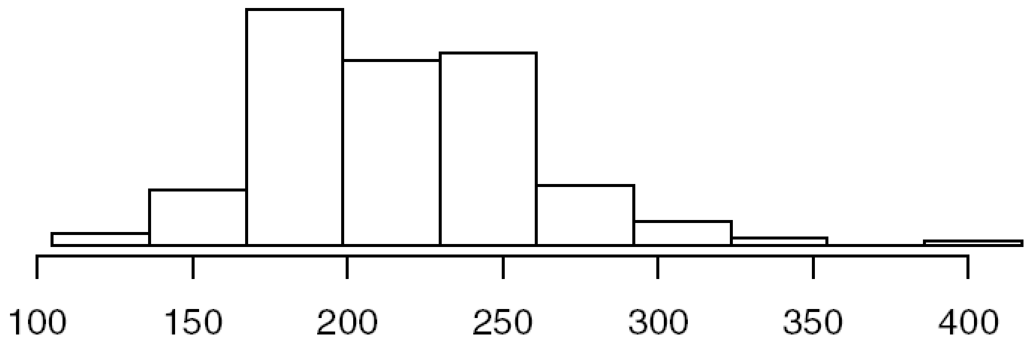
10

$$T(\widehat{F}_n) = \widehat{F}_n^{-1}(p) = \inf\{x : \ \widehat{F}_n(x) \geq p\}$$

<u>Пример</u> (проверка гипотезы равенства средних значений). 371 пациент с болью в груди. У 51 нет признаков болезни сердца, у 320 – наблюдается сужение артерий.



Распределение в % пациентов в зависимости от величины холестерина в плазме (здоровые)



Распределение в % пациентов в зависимости от величины холестерина в плазме (сердечники)

$$\widehat{\mu}_1 = \int x d\widehat{F}_{n,1}(x) = \overline{X}_{n,1} = 195.27$$

$$\widehat{\mu}_2 = \int x d\widehat{F}_{n,2}(x) = \overline{X}_{n,2} = 216.19$$

$$\mathsf{se}(\widehat{\mu}) = \sqrt{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i)} = \sqrt{\frac{n\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$$

$$\widehat{\mathsf{se}}(\widehat{\mu}) = \frac{\widehat{\sigma}}{\sqrt{n}}$$

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$\widehat{\mathsf{se}}(\widehat{\mu}_1) = 5.0$, $\widehat{\mathsf{se}}(\widehat{\mu}_2) = 2.4$

95% доверительны интервалы для μ_1 и μ_2

$$\widehat{\mu}_1 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_1) = (185, 205)$$

$$\widehat{\mu}_2 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_2) = (211, 221)$$

13 **Рассмотрим функционал**

$$\theta = T(F_2) - T(F_1)$$

$$\widehat{\theta} = \widehat{\mu}_2 - \widehat{\mu}_1 = 216.19 - 195.27 = 20.92$$

$$\mathsf{se} = \sqrt{\mathbb{V}(\widehat{\mu}_2 - \widehat{\mu}_1)} = \sqrt{\mathbb{V}(\widehat{\mu}_2) + \mathbb{V}(\widehat{\mu}_1)} = \sqrt{(\mathsf{se}(\widehat{\mu}_1))^2 + (\mathsf{se}(\widehat{\mu}_2))^2}$$

$$\widehat{\mathsf{se}} = \sqrt{(\widehat{\mathsf{se}}(\widehat{\mu}_1))^2 + (\widehat{\mathsf{se}}(\widehat{\mu}_2))^2} = 5.55$$

95% доверительны интервалы для θ

$$\widehat{\theta} \pm 2\,\widehat{\mathsf{se}}(\widehat{\theta}_n) = (9.8, 32.0)$$

Таким образом, у пациентов с суженными артериями повышенное содержание холестерина. Однако, это совершенно не означает, что холестерин – причина сужения. Может существовать какой-то третий фактор, который влияет и на сужение артерий и на повышение уровня холестерина