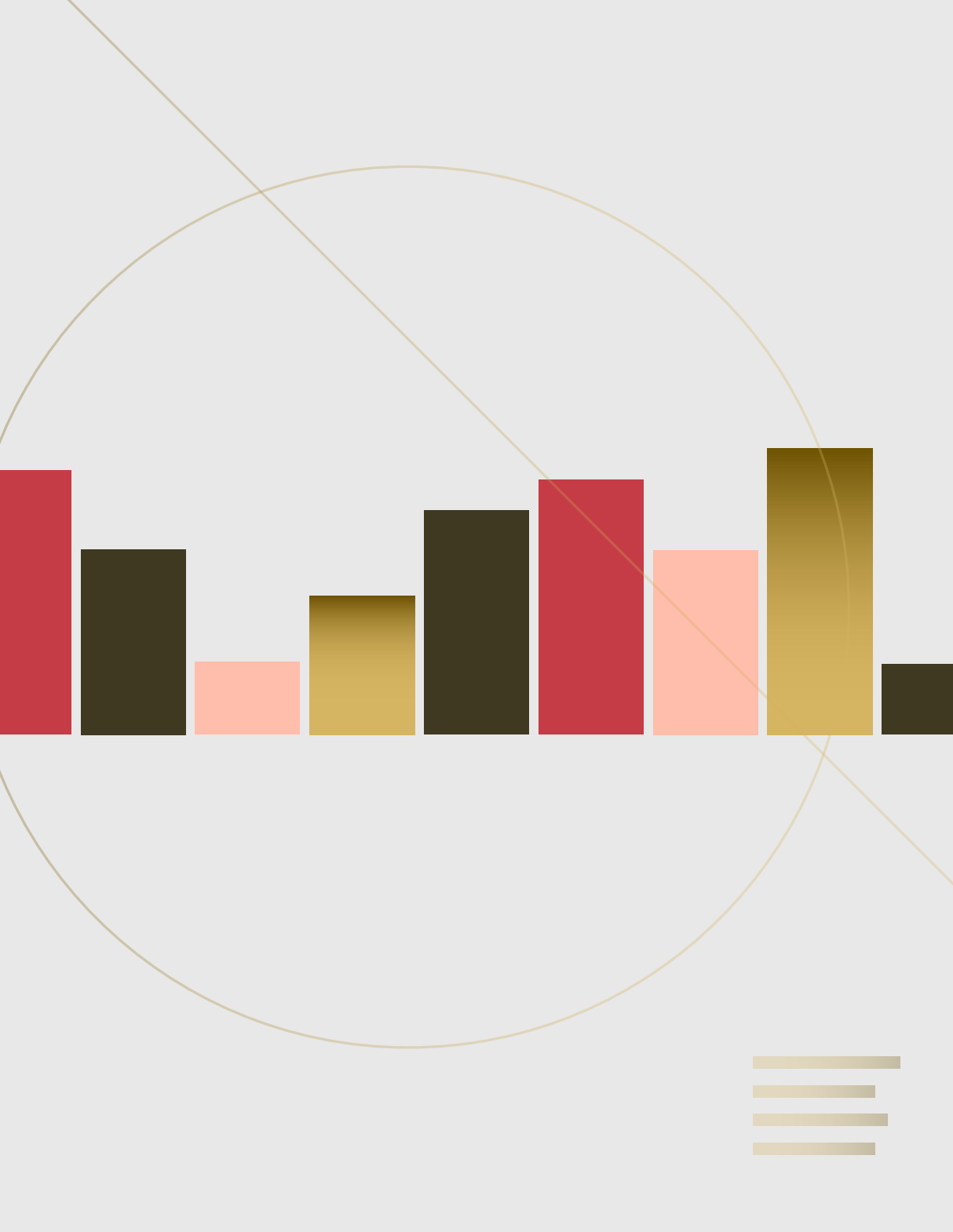


1

$p_1 = 0{,}9$ — оракул прав

$p_2 = 0{,}5$ — прав обычный человек

A_n = человек прав в n бросаниях

B = человек — оракул, $P(B) = 10^{-4}$

- $P(B \mid A_{10}) = ?$ $\quad P(B \mid A_{10}) = \dfrac{P(B \cap A_{10})}{P(A_{10})} = \dfrac{P(A_{10} \mid B) \cdot P(B)}{P(A_{10})} = \dfrac{P(A_{10} \mid B) \cdot P(B)}{P(A_{10} \mid B) \cdot P(B) + P(A_{10} \mid \bar{B}) P(\bar{B})}$

$$P(B \mid A_{10}) = \frac{0{,}9^{10} \cdot 10^{-4}}{0{,}9^{10} \cdot 10^{-4} + 0{,}5^{10} \cdot (1 - 10^{-4})} \approx 0{,}34$$

- а) 100 человек угадывают монеты на k бросаниях

$\Rightarrow$ выбираем лучшего: P $\qquad P(B) = 10^{-4}$

- Имеем: $\{\xi_i\}_{i=1}^{m}$; $\{\eta_i\}_{i=1}^{100-m}$

  ↖ рез-ты людей ↖ рез-ты оракулов

$\mathbb{E}\, Be(p_1) = p_1$

$D\, Be(p_1) = \mathbb{E}(Be(p_1) - p_1)^2 = p_1 q_1$

$\sigma^2 = p_1 q_1$

$\mu = p_1$

- $\xi_i \equiv \sum^{k} Be(p_1) \xrightarrow[k \to \infty]{d} \mathcal{N}(k p_1, k p_1^2 (1 - p_1)^2)$

- $\eta_i \xrightarrow[k \to \infty]{d} \mathcal{N}(k p_2, k p_2^2 (1 - p_2)^2)$

ЦПТ: $\dfrac{\sum^{k} \beta_i - k\mu}{\sigma \sqrt{k}} \xrightarrow{d} \mathcal{N}(0, 1)$

- При $k \to \infty$: $\Delta \mu = k (p_2 - p_1)$

  $\Delta \sigma = \sqrt{k} \sqrt{p_2^2 (1 - p_2)^2 - p_1^2 (1 - p_1)^2}$

Но тогда пики удаляются быстрее чем увеличиваются дисперсии

- Тогда $\mathbb{P}\{\eta_i \geq \xi_i\} \xrightarrow[k \to \infty]{} 1$

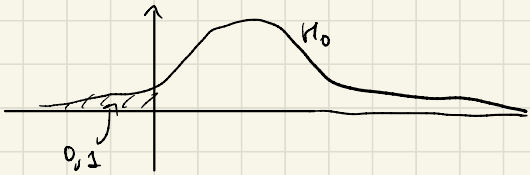
- Потому даже при сэмплировании по максимуму: $\mathbb{P}\{\eta_i \geq \max\{\xi_i\}\} \xrightarrow[k \to \infty]{} 1$

- Тогда можно найти искомое распред-е:

$\mathbb{P}$(после испытаний — оракул при k-бросаниях) $\xrightarrow[k \to \infty]{}$ $\mathbb{P}\{$выпал хотя бы один оракул в выборке из 100 спортсменов$\} = 100 \cdot P(B) = 10^{-2}$

сходится с экспериментом (task 1, ipynb)

$\boxed{2}$ $\{X\} \sim F(x)$. $\{H_0 \mid F(0) = 0{,}1\}$ $\qquad \alpha = 0{,}05$



$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{x \geq x_k\}}$$

$$\hat{F}_n(x) \xrightarrow[\text{ЗБЧ}]{\mathbb{P}} F(x)$$

$$T(X) = \hat{F}_n(0)$$

$$\delta(X) = \left[\, |T(X) - 0{,}1| > \beta \,\right]$$

$$\mathbb{P}_{H_0}\{|T(X) - 0{,}1| > \beta\} = \mathbb{P}_{H_0}\Big\{\Big|\underbrace{\frac{1}{n}\sum \mathbb{1}_{\{0 \geq x_k\}}}_{\frac{1}{n}\sum \xi_k} - 0{,}1\Big| \geq \beta\Big\} \leq \alpha$$

$$\xi_k \sim Be(F(0)) = Be(0{,}1) \qquad p = 0{,}1 \quad \overset{a = p}{\sigma = p(1-p)}$$

ЦПТ: $\dfrac{S_n - na}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \implies \dfrac{\frac{1}{n}\sum \mathbb{1}_{\{0 \geq x_k\}} - 0{,}1}{0{,}1 \cdot 0{,}9} \cdot \sqrt{n} \xrightarrow{d} \mathcal{N}(0,1)$

$$\mathbb{P}_{H_0}\Big\{\Big|\frac{1}{n}\sum \mathbb{1}_{\{0 \geq x_k\}} - 0{,}1\Big| \geq \frac{\beta \cdot 0{,}09}{\sqrt{n}}\Big\} \leq \alpha = 0{,}05$$

$$\implies \mathcal{N}(0,1)_{0{,}975} = 1{,}96 \implies \frac{\beta \cdot 0{,}09}{\sqrt{n}} = 1{,}96$$

$$\implies \beta = \frac{1{,}96}{0{,}09} \cdot \sqrt{n}$$

Т.е критерий: $\delta(X) = \left[\, |\hat{F}_n(0) - 0{,}1| > \frac{1{,}96}{0{,}09}\sqrt{n} \,\right]$

Задача 3 (25+10 баллов). Пусть имеется выборка пар $\mathbf{z}_i = (x_i,\ y_i)$, $i = \overline{1,\ n}$,

$$\mathbf{z}_i \sim N\left(\mathbf{z}_i|(0,\ 0)^{\mathsf{T}},\ \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Гипотеза $\mathbf{H}_0$: $\rho = 0$

Для статистики $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ получить

а) распределение для разных значений ρ и нарисовать плотность для $\rho = 0$ и $\rho = 0.5$ для $n = 100$ (2 балла);

б) построить критерий для проверки гипотезы $\rho = 0$ на уровне значимости $\alpha = 0.05$ (3 балла);

в) зависимость мощности данного критерия от истинного ρ сэмплированием (5 баллов) и приближенно аналитически (5 баллов), и предложить форму (3 балла) зависимости мощности критерия от n и ρ.

Сравнить мощность (7 баллов) в зависимости от ρ со статистикой $T(\mathbf{Z}) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2$, рассмотренной на лекции.

Какую статистику вы предложили бы для использования на практике? (+10 баллов тому, кто предложит свою и аргументированно обоснует, что она лучше).

$\boxed{3}$

$$z_i \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \underbrace{\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}}_{R}\right)$$

$$T(z) = \frac{1}{n}\sum x_i y_i$$

а) $$R = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}}_{S^{-1}} \underbrace{\begin{bmatrix}1+\rho & 0\\ 0 & 1-\rho\end{bmatrix}}_{D} \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}}_{S}$$

$$\frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}1+\rho & 0\\ 0 & 1-\rho\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1+\rho & \rho-1\\ 1+\rho & 1-\rho\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}2 & 2\rho\\ 2\rho & 2\end{bmatrix} = \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$$

- $$ARA^T = I \Rightarrow A = S^{-1}\frac{1}{\sqrt{D}}S \Rightarrow S^{-1}\frac{1}{\sqrt{D}}\cdot S\, S^{-1}DS\, S^{-1}\frac{1}{\sqrt{D}}S = I$$

- $$A = \frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{1+\rho}} & \\ & \frac{1}{\sqrt{1-\rho}}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\frac{1}{\sqrt{1+\rho}} & -\frac{1}{\sqrt{1-\rho}}\\ \frac{1}{\sqrt{1+\rho}} & \frac{1}{\sqrt{1-\rho}}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\frac{1}{\sqrt{1+\rho}}+\frac{1}{\sqrt{1-\rho}} & \frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}\\ \frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}} & \frac{1}{\sqrt{1+\rho}}+\frac{1}{\sqrt{1-\rho}}\end{bmatrix}$$

- $$\xi = Az \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)$$

$$2\beta = \frac{1}{4}(1+\rho-(1-\rho)) = \frac{\rho}{2}$$

$$\varphi_{\xi_1}(t) = e^{\mu_1 t}\cdot e^{-\frac{\sigma^2 t^2}{2}}$$

$$\begin{bmatrix}\alpha & \beta\\ \beta & \alpha\end{bmatrix}$$

- $$A^{-1} = \frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\sqrt{1+\rho} & 0\\ 0 & \sqrt{1-\rho}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\sqrt{1+\rho} & -\sqrt{1-\rho}\\ \sqrt{1+\rho} & \sqrt{1-\rho}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\sqrt{1+\rho}+\sqrt{1-\rho} & \sqrt{1+\rho}-\sqrt{1-\rho}\\ \sqrt{1+\rho}-\sqrt{1-\rho} & \sqrt{1+\rho}+\sqrt{1-\rho}\end{bmatrix}$$

- $$A^{-1}\xi = \begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}\alpha\xi_1+\beta\xi_2\\ \beta\xi_1+\alpha\xi_2\end{bmatrix} \Rightarrow xy = \alpha\beta(\xi_1^2+\xi_2^2) + (\alpha^2+\beta^2)\xi_1\xi_2$$

$$E xy = \alpha\beta\, \underbrace{E\chi^2(2)}_{=2} + (\alpha^2+\beta^2)\underbrace{E\xi_1}_{=0}\underbrace{E\xi_2}_{=0} = 2\alpha\beta = \rho$$

$\xi_1 \perp \xi_2$

$$\alpha^2 = \frac{1}{4}(2+2\sqrt{1-\rho^2})$$
$$\beta^2 = \frac{1}{4}(2-2\sqrt{1-\rho^2})$$

$$\alpha^2\beta^2 = \frac{1}{4}(1-(1-\rho^2)) = \frac{\rho^2}{4}$$

$$D xy = E(xy)^2 - (E xy)^2 = 1+2\rho^2-\rho^2 = 1+\rho^2 = \sigma^2$$

$$(xy)^2 = \alpha^2\beta^2(\xi_1^2+\xi_2^2)^2 + 2\alpha\beta(\alpha^2+\beta^2)\xi_1\xi_2(\xi_1^2+\xi_2^2) + (\alpha^2+\beta^2)^2\xi_1^2\xi_2^2$$

$$E(xy)^2 = \alpha^2\beta^2 E(\xi_1^2+\xi_2^2)^2 + 2\alpha\beta(\alpha^2+\beta^2)\cdot \underbrace{E\xi_1\xi_2(\xi_1^2+\xi_2^2)}_{=0} + (\alpha^2+\beta^2)^2 E\xi_1^2\xi_2^2 \quad (=)$$

$$\alpha^2\beta^2\left(\underbrace{E(\xi_1^2+\xi_2^2)}_{=2\cdot 2} + \underbrace{(E\xi_1+\xi_2)^2}_{=(2)^2}\right) = 8\alpha^2\beta^2$$

дох-во ниже

$$(\alpha^2+\beta^2)^2 \underbrace{E\xi_1^2}_{=1}\underbrace{E\xi_2^2}_{=1}$$

$$(=)\ 8\cdot\frac{\rho^2}{4} + \left(\frac{1}{4}(4)\right)^2 = 1+2\rho^2$$

$x_1, x_2 \in U[0,1]$: $\quad y_1 = \sqrt{-2\ln x_1}\,\sin(2\pi x_2)$
$x_1 \perp x_2$ $\quad y_2 = \sqrt{-2\ln x_1}\,\cos(2\pi x_2)$ $\iff \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$

$$y_1^2 + y_2^2 = -4\ln x_1$$

$$y_1 y_2 = (-\ln x_1)\sin(4\pi x_2)$$

$$\mathbb{E}(y_1^2+y_2^2)y_1y_2 = \mathbb{E}\left(4(\ln x_1)^2\right)\left(\sin 4\pi x_2\right) = \mathbb{E}(\dots)\,\underbrace{\mathbb{E}(\dots)}_{=0} = 0$$

$x_1 \perp x_2$

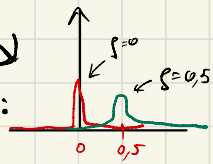
- Итак:

$$T = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$\frac{\sum x_i y_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

- $\sqrt{n}\left(\dfrac{\frac{1}{n}\sum x_i y_i - \mu}{\sigma}\right) \sim \mathcal{N}(0,1) \Rightarrow \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{d} \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$

- $T_n(z) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\rho, \dfrac{1+\rho^2}{n}\right)$



- Тогда для $n=100$ для $\rho = 0$ и $0{,}5$:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

пик для $(\rho=0) = \dfrac{1}{\sqrt{2\pi}}\sqrt{n}$

пик для $(\rho=0{,}5) = \dfrac{1}{\sqrt{2\pi}}\cdot\sqrt{\dfrac{n}{1{,}25}}$

б) $\{H_0: \rho = 0\}$ $\quad \alpha = 0{,}05$

$$\delta(x) = [\,|T(x)| > b\,]$$

$$\mathbb{P}_{H_0}\{|T(x)| > b\} \approx 2\cdot\Phi(-b) = 2\cdot\int_{-\infty}^{-b} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \le 0{,}05$$

Марков: $P(\xi > t) = P\{\xi^2 > t^2\} < \dfrac{E\xi^2}{t^2} < \dfrac{E\xi^2}{t}$ $\qquad \xi = |T_n(x)|$

$$E\xi^2 = \sigma^2 E\chi^2(1) = \sigma^2 = \frac{1+\rho^2}{n} = \frac{1}{n} \qquad \xi^2 = (T(x))^2 = (\sigma\,\eta)^2 = \sigma^2\cdot\chi^2(1)$$

тогда: $b = \dfrac{1}{0{,}05n} = \dfrac{20}{n}$ $\leftarrow$ в task3.ipynb после (сэмплирования даёт $\alpha < 0{,}05$)

в) Итак есть $T(X_n) \xrightarrow{d} \mathcal{N}(\rho, \frac{1+\rho^2}{n})$

и критерий $\left[|T(X_n)| > \frac{20}{n}\right]$

$H(\rho) = \mathcal{N}(\rho, \frac{1+\rho^2}{n})$

Пусть верна H_ρ с $\rho \neq 0$, тогда $\hookrightarrow$

$\hookrightarrow W_\rho = \mathbb{P}_{H_\rho}\left(|T(X_n)| > \frac{20}{n}\right) \approx 1 - \Phi_{H_\rho}\left(\frac{20}{n}\right) + \Phi_{H_\rho}\left(-\frac{20}{n}\right)$

Позже сравним этот ответ с результатом (эмпирирования. (спойлер: они совпадают) (прога: task3.ipynb)

Зависимость $W(n, \rho)$: 1) при $n \gg 1$: $W \approx 1 - \Phi_{H(\rho)}\left(\frac{20}{n}\right) + \Phi_{H(\rho)}\left(-\frac{20}{n}\right)$

2) при $n \sim 1$: $W(n, \rho) = 1 - \Phi_{F_T(\rho)}\left(\frac{20}{n}\right) + \Phi_{F_T(\rho)}\left(-\frac{20}{n}\right)$

, где $F_T(\rho)(x) = \mathbb{P}\left\{\frac{1}{n}\sum^n x_i y_i \leq x\right\}$ – ф распр-я статистики (приходится считать вручную – например через ЭФР)

- Дока $T(X) = \frac{1}{2n}\sum_{k=1}^n (x_k - y_k)^2$; $x_k - y_k = [1\ -1]\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}(0, ARA^T)$

$ARA^T = [1\ -1]\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = [1-\rho\ \ 1-1]\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2(1-\rho) = \sigma^2$

- $\frac{x-y}{\sqrt{2(1-\rho)}} \sim \mathcal{N}(0,1) \Rightarrow \sum \overbrace{\frac{(x-y)^2}{2(1-\rho)}}^{\chi^2(1)} \sim \chi^2(n)$

- $T(z) = \frac{1}{2n}\sum (x-y)^2 = \frac{1}{n}(1-\rho)\sum \frac{(x-y)^2}{2(1-\rho)} \sim \frac{1-\rho}{n} \cdot \chi^2(n)$

- Знаем, что $\frac{\chi^2(n) - n}{\sqrt{2n}} \to \mathcal{N}(0,1)$ ($\square$ $\chi^2(n) = \sum \underbrace{\xi_k^2}_{\sim \chi^2(1)}$; $\frac{\sum \xi_k^2 - n \cdot E\xi_k}{\sigma_k \sqrt{n}}$ $\blacksquare$)

$\Rightarrow \frac{\sum \xi_k^2 - n \cdot 1}{\sqrt{n \cdot 2}} \xrightarrow{d} \mathcal{N}(0,1)$

- тогда $\delta(X) = \left[\left|T(z) - 1\right| > \varepsilon\right]$

- хотим $\alpha < 0{,}05$: $\mathbb{P}\left\{\left|\frac{\chi^2(n)}{n} - 1\right| > \varepsilon\right\} = \mathbb{P}\left\{\left|\frac{\chi^2(n) - n}{n}\right| > \varepsilon\right\}$

$= \mathbb{P}\left\{\left|\frac{\chi^2(n) - n}{\sqrt{2n}}\right| > \frac{\varepsilon\sqrt{n}}{\sqrt{2}}\right\} \leq 0{,}05 \Rightarrow \frac{\varepsilon\sqrt{n}}{\sqrt{2}} = \mathcal{N}_{0{,}025}(0,1) \approx 1{,}96$

$\varepsilon = \frac{1{,}96\sqrt{2}}{\sqrt{n}}$ (Сравнение с 1й статистикой и проверка $\alpha < 0{,}05$ в файле task3.ipynb

- Статистика $T(z) = \frac{1}{n}\sum x_i y_i$ в применении к нашей задаче является по сути коэфф. выбор. корреляции Пирсона
- Как показывает график: мощность статистики $\frac{1}{n}\sum x_i y_i$ больше, чем у $\frac{1}{2n}\sum (x_i - y_i)^2$ при всех ρ. Поэтому оптимально использовать первую статистику.

4

Задача 4 (15 баллов). Пусть $\mathbf{x} = \{x_1, \ldots, x_n\}$, $n = 12$ есть НОР выборка из $\mathcal{N}(0, 1)$. Пусть $\mathbf{y} = \{y_1, \ldots, y_n\}$, $n = 12$ есть НОР выборка из $\mathcal{N}(0, 1)$, независимая от $\{x_1, \ldots, x_n\}$. Оценить (сэмплированием или приближенно аналитически), сколько разных (и независимых) выборок $\mathbf{y}$ нужно рассмотреть K, чтобы найти ту, которая дает выборочную корреляцию с $\mathbf{x}$ не менее $\rho = 0.97$ (5 баллов). Построить график зависимости $K(\rho)$ в диапазоне от 0 до 0.99 (5 баллов). Какой прикладной вывод можно сделать из этого эксперимента помимо известного «корреляция не означает причинность» (5 баллов)?

$\overline{x} = \{x_i\}_1^{12} \sim \mathcal{N}(0,1)$ $\quad \overline{y} = \{y_i\}^{12} \sim \mathcal{N}(0,1)$ $\quad \overline{y} \perp \overline{x}$

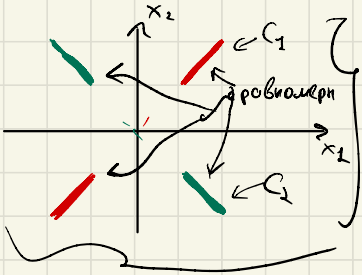
- Найти хотя бы одну $\sim \mathbb{P}\{T(x,y) \geq 0{,}97\} \geq \frac{1}{K}$
- $\mathbb{P}\left\{\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2 \cdot \sum(y_i-\overline{y})^2}} \geq 0{,}97\right\} \approx \frac{2}{10^7}$ → Нашли сэмплированием (task4.ipynb)

  т.е ОТВЕТ: $K \approx 5 \cdot 10^6$
- График для $K(\rho)$ в task4.ipynb
- Прикладной вывод:
  1) Достичь корреляции между НОР выборками очень трудно спедов-но по полученным большим значениям, к примеру, $\mathrm{corr}(x,y) > 0{,}97$ можно судить о том, что выборки не являются независимыми.
  2) Даже если выборки независимы корреляция между ними зачастую имеет значения сильно отличающиеся от нуля, т.е о некоррелиров-ти выборок можно судить только при многократном сэмплировании и проверки $|\mathrm{corr}(x,y)| < \varepsilon$

известного «корреляция не означает причинность» (5 баллов)?

Задача 5 (5 баллов). Привести пример, когда наивный байесовский классификатор классифицирует объекты не лучше, чем наугад, хотя генеральная совокупность (все возможные объекты) идеально разделима?

$$f(x) = \arg\max_k P(C_k|x) = \arg\max_k \left(P(C_k) P(x|C_k) \right)$$



Пусть $P(C_1) = P(C_2)$ наивность

тогда: $p(C_1|x) = P(C_1) \cdot \prod_{i=1}^{2} p(x_i|C_1) =$

[но $p(x_1|C_1) = p(x_2|C_1)$] $= P(C_1)\, p^2(x_1|C_1)$

$$\cdot\ P(C_2) \prod^{2} p(x_i|C_2) = P(C_2) \cdot p(x_1|C_2) \cdot p(x_2|C_2) = P(C_2) \cdot p(x_1|C_2) \cdot \underbrace{p(-x_1|C_2)}_{= p(x_1|C_2)}$$

$$= P(C_2)\, p^2(x_1|C_2) = P(C_1)\, p^2(x_1|C_1)$$

т.к $P(C_1) = P(C_2)$

а $p(x_1|C_1) = p(x_1|C_2)$, т.к мы их специально взяли одинаково распред-нными. Значит $f(x)$ - будет наугад выдавать 1 или 2.

На самом деле $P(x|C_1) \neq \prod_{i=1}^{2} p(x_i|C_1)$, а $p(x|C_1) = \begin{cases} p(x_1|C_1), & x_1 = x_2 \\ 0, & x_1 \neq x_2 \end{cases}$

и соотв $p(x|C_2) = \begin{cases} p(x_1|C_2), & x_1 = -x_2 \\ 0, & x_1 \neq -x_2 \end{cases}$

Т.е. проблема — в наивности байесовского классификатора.

$\boxed{6}$

task 6. ipynb

$FWER = \mathbb{P}\{V \geq 1\}$

$FDR = \mathbb{E}\left(\frac{V}{R} \mathbb{1}_{\{R>0\}}\right)$

можные объекты) идеально разделима?

Задача 6 (15 баллов). В условиях задачи 3 при $n = 100$ сэмплировать $m = 1000$ выборок пар $\mathbf{z}_i$, $i = \overline{1, m}$, для 500 из которых $\rho = 0$ и $\rho = 0.2$ для оставшихся. С помощью одной из рассмотренных (или своей) статистик получить достигаемые уровни значимости $p_1, \ldots, p_m$. Для уровня значимости $\alpha = 0.05$ сравнить результаты применения отсутствия поправки на множественное тестирование, поправки Бонферрони и поправки Бенджамини-Хохберга в терминах получения ложных открытий (ложно отклоненные гипотезы) или пропуска таковых (ложно принятые гипотезы). Контролирует ли поправка Бенджамини-Хохберга FDR на уровне α и почему? (8 баллов) 1

Рассмотреть отдельно 1000 выборок для $\rho = 0$ и повторить эксперимент. сравнить результаты применения отсутствия поправки на множественное тестирование, поправки Бонферрони и поправки Бенджамини-Хохберга в терминах получения ложных открытий (ложно отклоненные гипотезы) или пропуска таковых (ложно принятые гипотезы). Контролирует ли поправка Бенджамини-Хохберга FDR на уровне α и почему? (7 баллов)

2

В случаях 1 и 2 поправки Б-Х контролирует FDR на уровне α. Потому что в обоих случаях выполняются условия на поправку: незав-ть статистик для каждой из выборок.

$\boxed{7}$

Даже в этом случае поправка Б-Х. контролирует $FDR < \alpha$ так, как у нас изменилось априорное распределение выборок ложных/истинных, но это никак не повлияло на независит-ть этих выборок между собой.