Задание 1 по курсу "Байесовский выбор модели"

Грабовой Андрей, группа 574

Задача 1

Пусть проводиться эксперимент по угадыванию стороны выпадения честной монеты. Известно, что оракул прав с вероятностью $p_1 = 0.9$, а обычный человек с вероятностью $p_2 = 0.5$. Известно, что человек P оказался прав во всех n=10 бросаниях. С какой вероятностью P является оракулом, если случайный человек оказывается оракулом с вероятностью $p_a = 10^{-4}$?

Пусть $A = [P - \text{оракул}]$, $B = [n \text{ из } n]$ - обозначения из лекциии.

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B|\bar{A})\mathcal{P}(\bar{A}) + \mathcal{P}(B|A)\mathcal{P}(A)}, \tag{1}$$

где $\mathcal{P}(A) = 0.0001$, $\mathcal{P}(\bar{A}) = 0.9999$.

Подставляя числа в формулу (1) получаем:

$$\mathcal{P}(A|B) = \frac{0.9^{10} \cdot 10^{-4}}{0.9^{10} \cdot 10^{-4} + 0.5^{10} \cdot 0.9999} = 0.0345$$

Пусть человек P выбран не случайно, а как лучший среди 100 человек по угадыванию $k = 100$ выпадений монеты. Вывести новую априорную вероятность того, что P оракул с учетом его неслучайного выбора.

Аналитически: Найдем вероятность того, что лучший из 100 это оракул:

$$\mathcal{P}(A|C) = \sum_{i=0}^{100} \mathcal{P}(A|C, C_O = i)\mathcal{P}(C_O = i), \tag{2}$$

где C_O это случайная величина — количества оракулов среди 100 человек, C - это событие, что P лучший из 100 человек. Из суммы останется только первых два слагаемых, так как остальные слагаемые будут много меньше:

$$\mathcal{P}(C_O = i) = C_k^i p_a^i (1 - p_a)^{k-i} = \begin{cases} 0.9900 & i = 0 \\ 0.0099 & i = 1 \\ 0.0001 & i = 2 \end{cases}, \tag{3}$$

как видно из (3) нам важны только первые два слагаемых. Получаем:

$$\mathcal{P}(A|C) = \mathcal{P}(A|C, C_O = 0)\mathcal{P}(C_O = 0) + \mathcal{P}(A|C, C_O = 1)\mathcal{P}(C_O = 1), \tag{4}$$

Найдем по формуле Байеса $\mathcal{P}(A|C, C_O = 0)$ и $\mathcal{P}(A|C, C_O = 1)$.

$$\mathcal{P}(A|C, C_O = 0) = \frac{\mathcal{P}(C|A, C_O = 0)\mathcal{P}(A|C_O = 0)}{\mathcal{P}(C|\bar{A}, C_O = 0)\mathcal{P}(\bar{A}|C_O = 0) + \mathcal{P}(C|A, C_O = 0)\mathcal{P}(A|C_O = 0)} = 0, \quad (5)$$

$$\mathcal{P}(A|C, C_O = 1) = \frac{\mathcal{P}(C|A, C_O = 1)\mathcal{P}(A|C_O = 1)}{\mathcal{P}(C|\bar{A}, C_O = 1)\mathcal{P}(\bar{A}|C_O = 1) + \mathcal{P}(C|A, C_O = 1)\mathcal{P}(A|C_O = 1)}, \quad (6)$$

Найдем вероятность того, что оракул "проиграет"простому человеку при $k >> 1$. При $k >> 1$ биномиальную случайную величину можно приблизить нормальной случайной величиной, тоесть:

$$\mathcal{P}(\text{оракул проиграл}) = \mathcal{P}(O < H) = \int_{-\infty}^{\infty} \mathcal{P}(O < t) p_{N_2}(t) dt = \int_{-\infty}^{\infty} p_{N_2}(t) \int_{-\infty}^{t} p_{N_1}(\omega) d\omega dt =$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} p_{N_2}(t) [\text{erf}\left(\frac{t - p_1 k}{\sqrt{2p_1(1-p_1)k}}\right) + 1] dt$$

где $N_2 = N(p_2 k, p_2(1-p_2)k)$, $N_1 = N(p_1 k, p_1(1-p_1)k)$

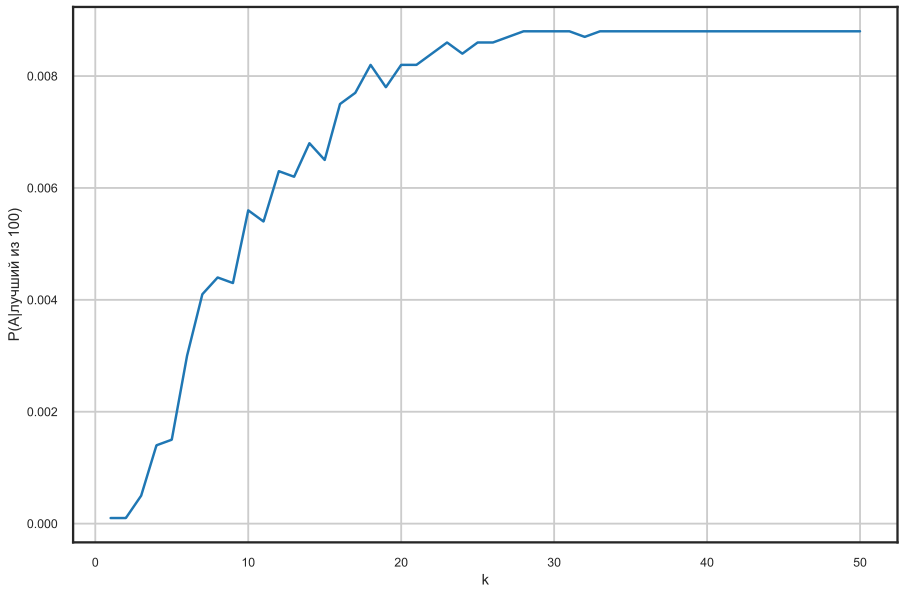

Рис. 1: График зависимости апостериорной вероятности от количества бросаний монеты

При k = 100 получаем, что $\mathcal{P}(\text{оракул проиграл}) = 10^{-14}$. Получаем, что

$$\mathcal{P}(C|A, C_O = 1) \approx 1.0, \quad (7)$$

$$\mathcal{P}(C|\bar{A}, C_O = 1) \approx 0.0, \tag{8}$$

Тогда подставляя (7-8) в (6) получаем, что $\mathcal{P}(A|C, C_O = 0) \approx 1.0$, откуда получается, что

$$\mathcal{P}(A|C) \approx 0.01, \tag{9}$$

Сэмплированием: Сгенерируем выборку учитывая наше априорное представление о модели. Пусть у нас есть 1000 групп людей по 100 человек, где каждый человек оракул с вероятностью $\mathcal{P}(A) = 10^{-4}$. После каждый человек бросает монету $k = 1...50$ раз и в каждой группе получаем ' лучшего' кандидата. После этого смотрим оракул он или нет и получаем апостериорную вероятность что лучший из 100 будет оракулом как частоту. Построим график зависимости $\mathcal{P}(A|C)$ от числа k.

Из аналитического решения и с сэмплирования Рис. 1, получаем что $\mathcal{P}(A|C) \approx 0.01$.

Задача 2

Пусть имеется i.i.d выборка $\mathbf{X} = \{x_1, \cdots x_n\}$ из неизвестного распределения с конечной плотностью. На уровне значимости $\alpha = 0.05$ проверить гипотезу о том, что деситипроцентная квантиль распределения равна $m_0 = 0$.

Начальная гипотеза эквивалентная гипотезе о том, что в выборке 10% данных меньше нуля.

Построим статистику:

$$T(\mathbf{X}) = \frac{1}{2n} \sum_{i=1}^{n} (-\text{sign}(x_i) + 1). \tag{10}$$

По ЦПТ:

$$T(\mathbf{X}) \sim N(\tau, \frac{\mathsf{S}^2}{n}), \tag{11}$$

где S^2 — это несмещенная выборочная дисперсия выборки $\frac{1}{2}(-\text{sign}(\mathbf{X}) + 1)$.

Тогда получаем, что если $T(\mathbf{X})$ попадает в область закрашенную на Рис. 2, гипотеза о том, что деситипроцентная квантиль распределения равна $m_0 = 0$ отвергается, в противном случае, если не попадает, то данные не противоречат гипотезе (о том, что $\tau = 0.1$).

Задача 3

Пусть имеется выборка пар $\mathbf{z}_i = (x_i, y_i),\ i = 1...n.\ \mathbf{z}_i \sim N([0, 0]^{\mathrm{T}}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$

Для статистики $T_1(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$, получить плотность распределения. Из ЦПТ получается, что:

$$T_1(\mathbf{Z}) \sim N(\rho, \frac{1}{n}), \tag{12}$$

Тогда для $T_1(\mathbf{Z})$ получаем следующий график плотностей распределения:

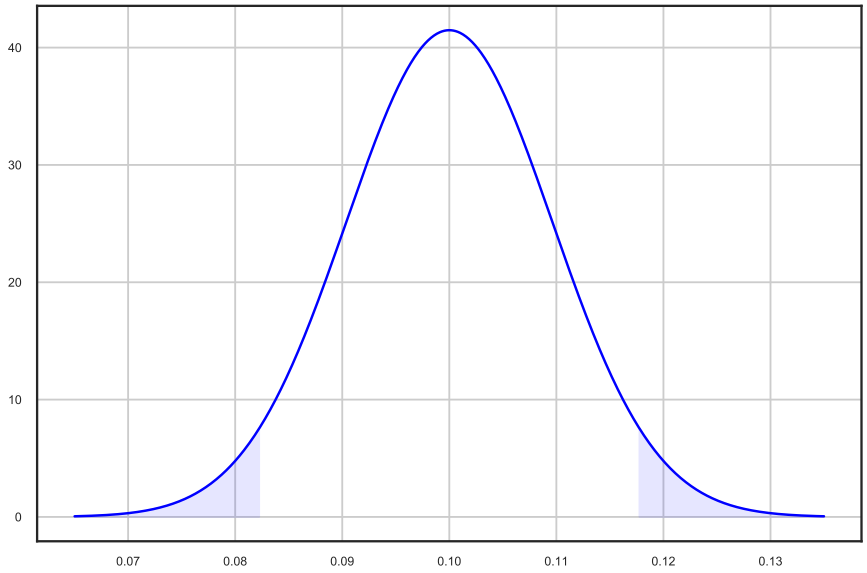

Рис. 2: График плотности распределения для n = 1000

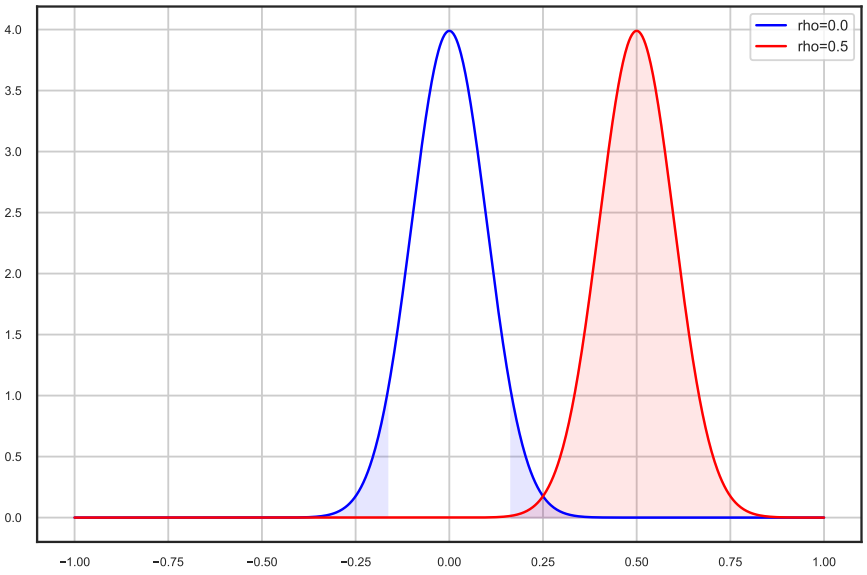

Рис. 3: График плотностей распределений для разных ρ

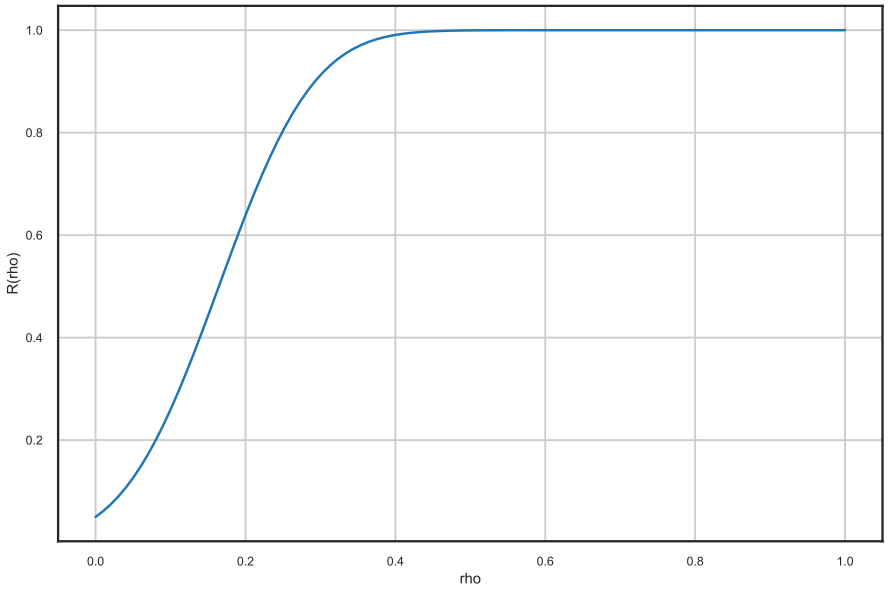

Рис. 4: График зависимости мощности от ρ аналитически

Будем проверять гипотезу H_0 о том, что $\rho = 0$ на уровне значимости $\alpha = 0.05$. На Рис. 3 показан мощность критерия для $\rho = 0.5$.

Найдем мощность критерия $R_1(\rho)$ аналитически:

$$R_1(\rho) \approx \int_{0.1645}^{\infty} p_{N(\rho,\frac{1}{n})}(x)dx = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_a^{\infty} e^{-(\frac{\sqrt{n}x}{\sqrt{2}} - \frac{\sqrt{n}\rho}{\sqrt{2}})^2} dx =$$

$$= \frac{1}{\sqrt{\pi}} \int_{a\frac{\sqrt{n}}{\sqrt{2}}}^{\infty} e^{-(x - \frac{\sqrt{n}\rho}{\sqrt{2}})^2} dx = \frac{1}{\sqrt{\pi}} \int_{a - \frac{\sqrt{n}\rho}{\sqrt{2}}}^{\infty} e^{-x^2} dx = \frac{1}{2}\text{erfc}\left((0.1645 - \rho)\sqrt{\frac{n}{2}}\right), \tag{13}$$

На Рис. 4 показан график зависимости мощности критерия $R_1(\rho)$ найденый аналитически. На Рис. 5 показан график зависимости мощности критерия $R_1(\rho)$ найденый сэмплиированием.

Сравнить мощность в зависимости от ρ со статистикой $T_2(\mathbf{Z}) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2$, рассмотренной на лекции.

Для статистики $T_2(\mathbf{Z}) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2$, из ЦПТ получается, что:

$$T_2(\mathbf{Z}) \sim N(1 - \rho, \frac{(1-\rho)^2}{n}), \tag{14}$$

Тогда для $T_2(\mathbf{Z})$ получаем график плотностей распределения показанный на Рис. 6.

Будем проверять гипотезу H_0 о том, что $\rho = 0$ на уровне значимости $\alpha = 0.05$. На Рис. 6 показан мощность критерия для $\rho = 0.5$.

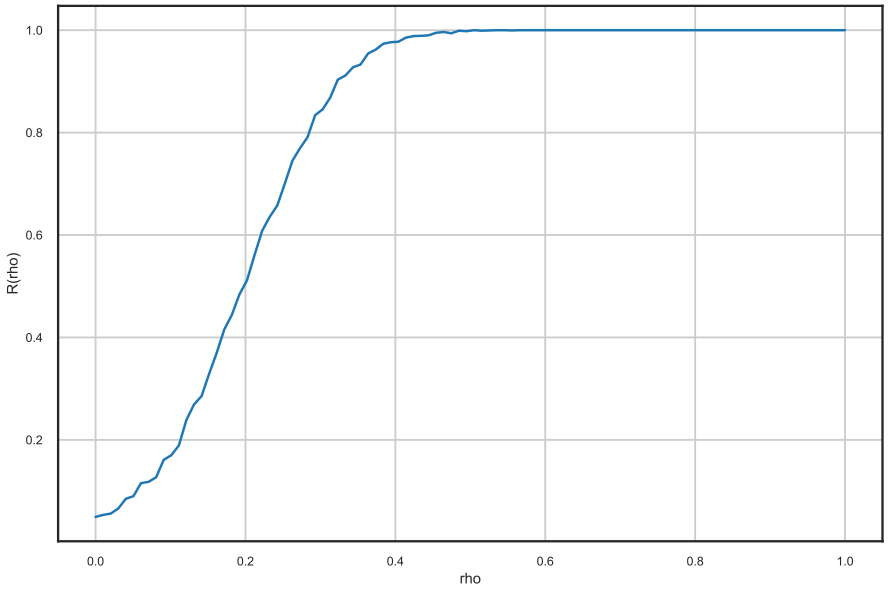

Рис. 5: График зависимости мощности от ρ сэмплированием

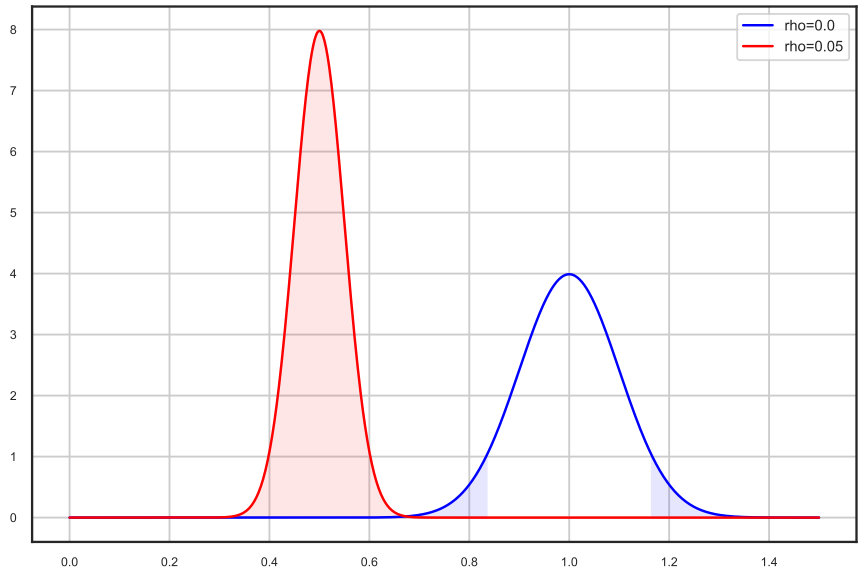

Рис. 6: График плотностей распределений для разных ρ

Найдем мощность критерия $R_2(\rho)$ аналитически:

$$R_2(\rho) \approx \int_{-\infty}^{0.8355} p_{N(1-\rho,\frac{(1-\rho)^2}{n})}(x)dx = \frac{\sqrt{n}}{\sqrt{2\pi}(1-\rho)} \int_{-\infty}^{a} e^{-(\frac{\sqrt{n}x}{\sqrt{2}(1-\rho)} - \frac{\sqrt{n}}{\sqrt{2}})^2} dx =$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{a\frac{\sqrt{n}}{\sqrt{2}(1-\rho)}} e^{-(x-\frac{\sqrt{n}}{\sqrt{2}})^2} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{a\frac{\sqrt{n}}{\sqrt{2}(1-\rho)} - \frac{\sqrt{n}}{\sqrt{2}}} e^{-x^2} dx = \frac{1}{2}\text{erf}\left((\frac{0.8355}{1-\rho} - 1)\sqrt{\frac{n}{2}}\right) + \frac{1}{2}, \quad (15)$$

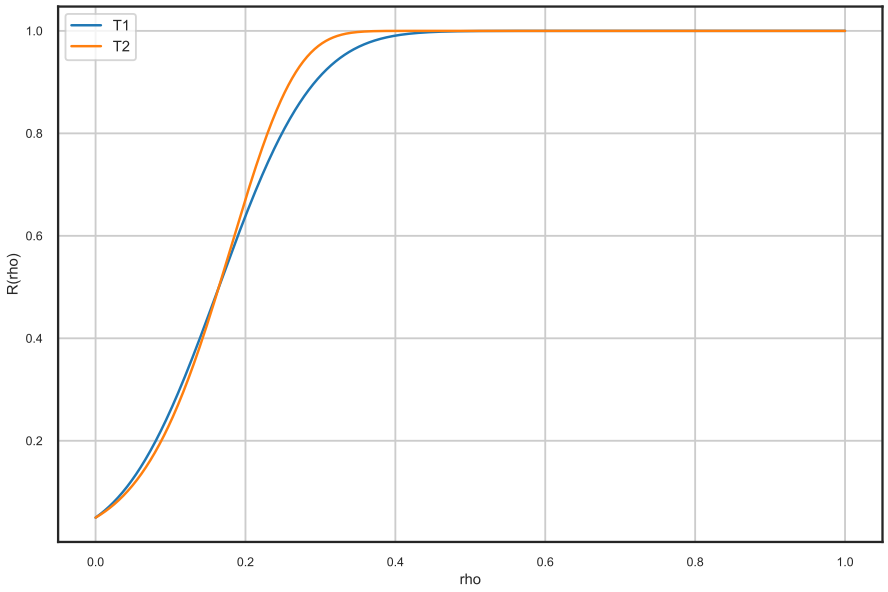

Рис. 7: График зависимости мощности от ρ для разных статистик

Как видно из Рис. 7 статистика $T_2(\mathbf{Z})$ лучше чем статистика $T_1(\mathbf{Z})$, так-как график мощности лежит выше.

Задача 4

Пусть $\mathbf{x} = \{x_1, \cdots x_n\}$, $n = 12$ есть i.i.d выборка из $N(0,1)$. Пусть $\mathbf{y} = \{y_1, \cdots y_n\}$, $n = 12$ есть i.i.d выборка из $N(0,1)$, независимая от $\mathbf{x}$. Оценить, сколько разных Kвыборок $\mathbf{y}$ нужно рассмотреть, чтобы найти ту, которая дает выборочную корреляцию с $\mathbf{x}$ не менее $\rho = 0.97$

Заметим, что распределение выборочной корреляции это $N(0, \frac{1}{12})$, в чем можно убедиться с Рис. 8 на котором изображена эмпирическая функция распределения данных и функция распределения $N(0, \frac{1}{12})$

Учитывая нормальность распределения, получаем, что $\mathcal{P}_{N(0,\frac{1}{12})}(x > 0.97) = 10^{-4}$. Математическое ожидание, времени ожидание пока выпадет $\mathbf{y}$ такой что $\rho = 0.97$, равно $\mathsf{E}t = 10^4$ (из геометрического распределения).

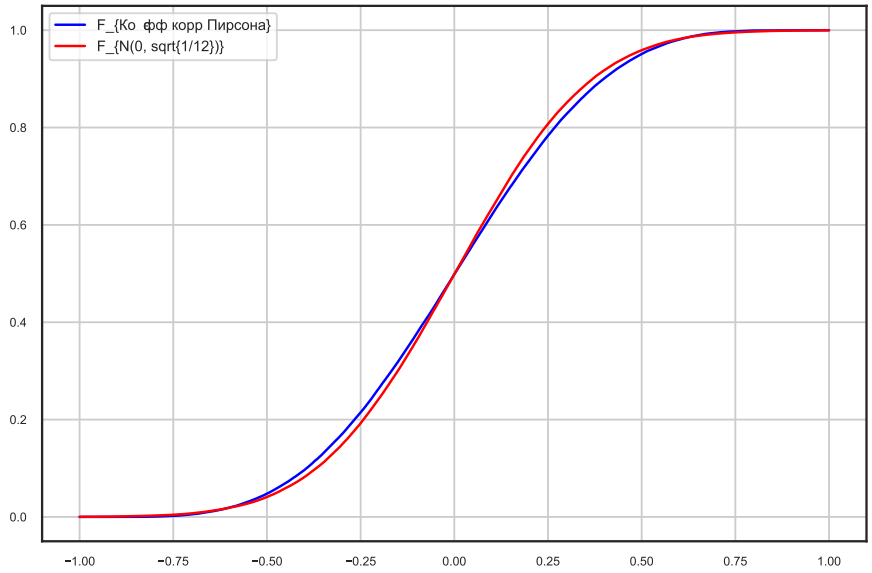

Рис. 8: Функции распределения $N(0, \frac{1}{12})$ и распределения корр. коэф. Пирсона

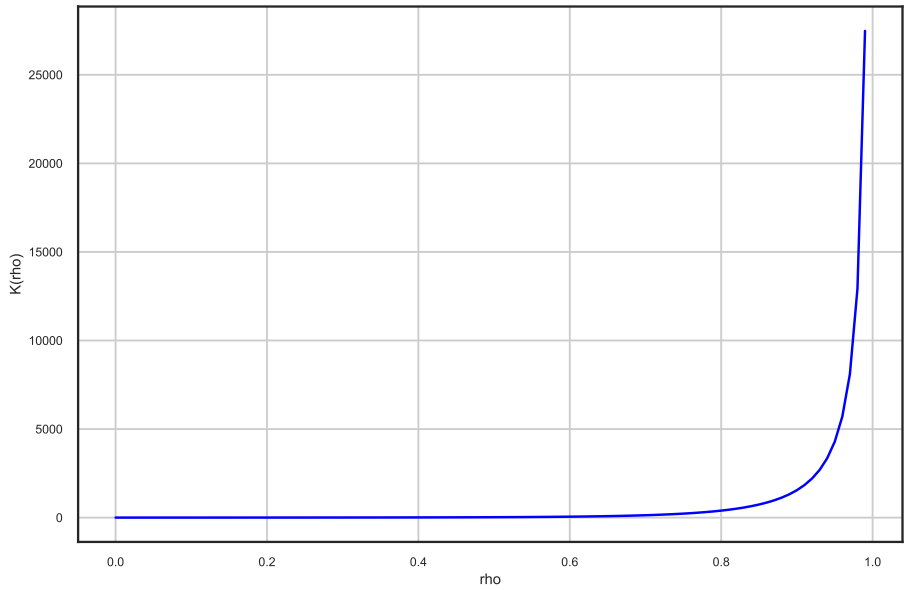

Рис. 9: График изменения времени ожидания от нужной ковариации

Построить график зависимости $K(\rho)$ в диапазоне от 0 до 0.99.

На Рис. 9 показан зависимости ожидаемого времени ожидания от заданого ρ.

Задача 5

Привести пример, когда наивный байесовский классификатор классифицирует объекты не луяше, чем наугад, хотя генеральная совокупность(все возможные объекты) идеально разделимы.

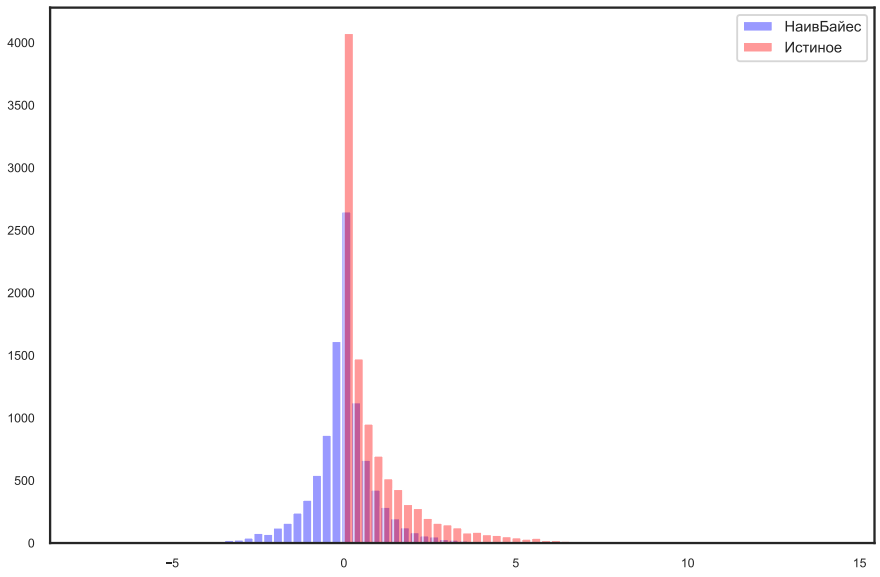

Рис. 10: Гистаграмма наивного Байеса и истинного распределения

Рассмотрим выборку распределенную по закону $\mathbf{X} = \left\{(x_i^1, x_i^2)^T | (x_i^1, x_i^2) \sim N\left((0,0)^T, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)\right\}$, то есть первый и второй признак идеально коррелируемый (то есть просто равен друг другу). Пусть ответом будет знак произведения $y_i = \text{sign}(x_i^1 x_i^2)$.

В силу того, что признаки идеально коррелируемые получим, что все ответы положительные(распределены по χ^2).

Что же мы получим с точки зрения наивного байесового класификатора. Наивный Байес будет считать что x_i^1 и x_i^2 независимые и тогда плотность распределения будет симметричная относительно оси ординат, что показано на Рис. 10. То есть наивный Байес будет классифицировать объекты не лучше чем наугад.

Задача 6

В услувиях задачи 3 для $\rho = 0.2$ и $\rho = 0.0$ при $n = 100$, сэмплировать $m = 1000$ выборок пар $z_i, i = 1...m$. С помощью статистики $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ получить достигнутые уровни значимости $p_1, \cdots p_m$.

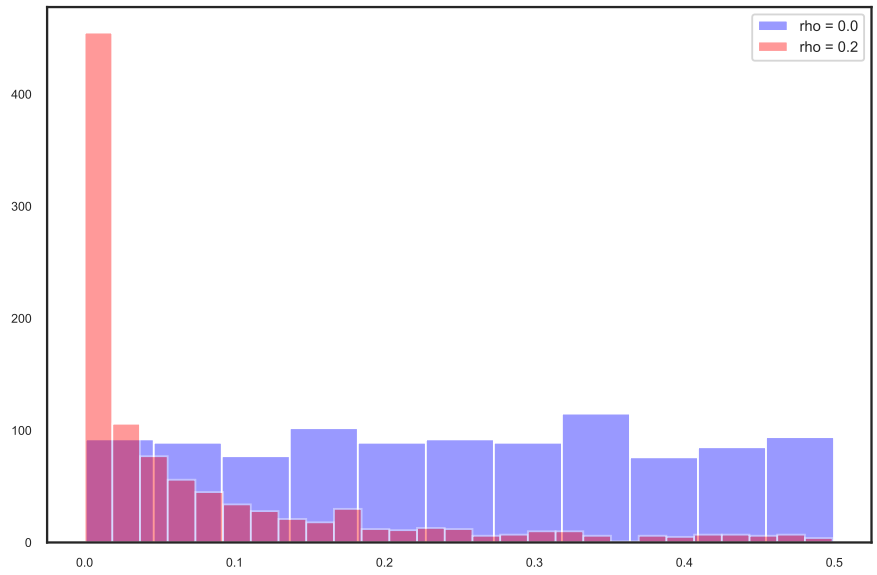

Рис. 11: Гистограма распределение p-value без поправок

Распределение достигаемых уровней показаны на Рис. [11, 12, 13]

Для уровня значимости $\alpha = 0.05$ сравним результаты применения отсутствия поправки на на множественном тестирование и с использованием поправок Бонферони и Бенджамини-Хохберга. Результаты эксперимента показаны в таб. 1.

Метод	Ложно отклоненные	Ложно принятые
Без поправки	93	400
Бонферони	0	960
Бенджамини-Хохберга	1	517

Таблица 1: Для уровня значимости $\alpha = 0.05$

Контролирует ли поправка Бенджамини-Хохберга FDR на уровне $\alpha = 0.05$? Поправка Бенджамини-Хохберга контролирует FDR на уровне $\alpha = 0.05$, так-как статистики T_i независимые.

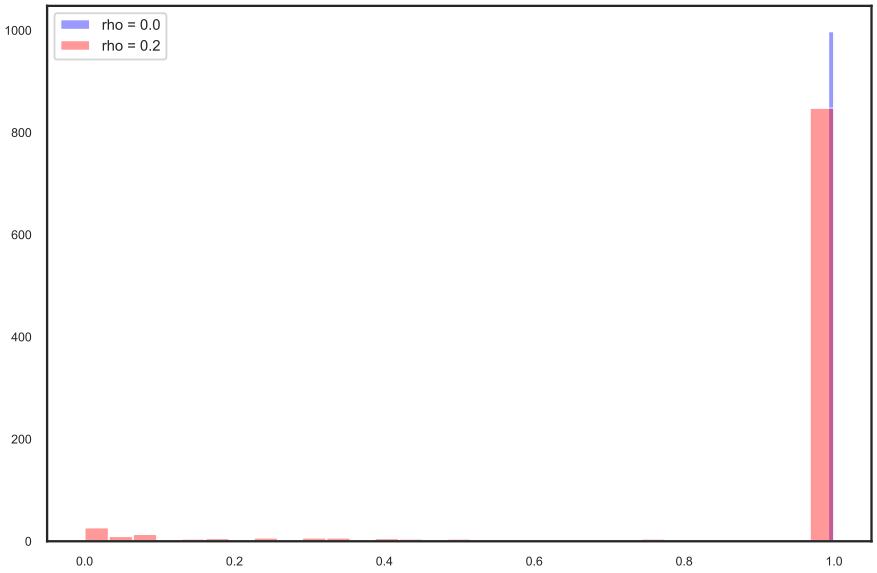

Рис. 12: Гистограма распределение p-value с Банферонии

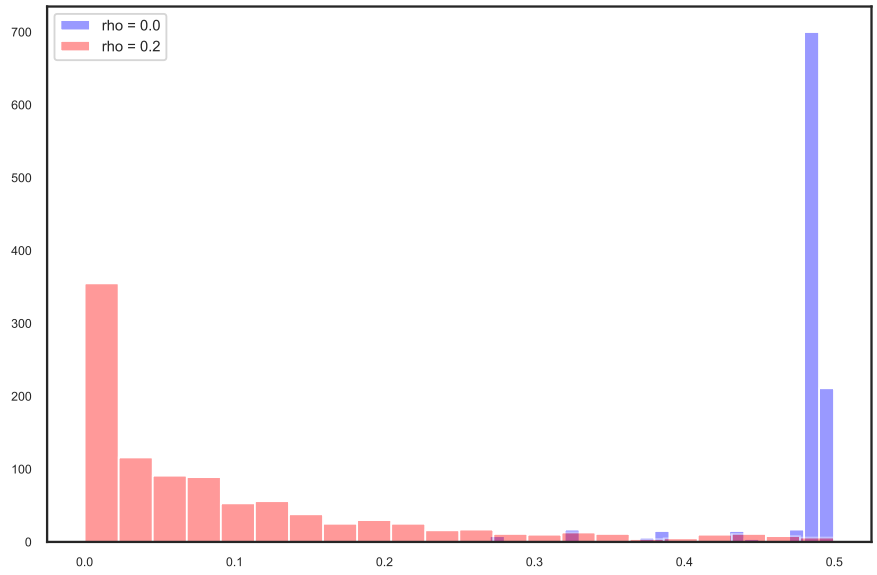

Рис. 13: Гистограма распределение p-value с Бенджамини-Хохберг

Задача 7

В условиях задачи 6 сэмплировать m = 1000 выборок пар, но с ρ_m, зависящее от номера выборки. Провести те же исследования, что и в задаче 6.

$$\rho_1 = 0, \rho_i = \begin{cases} \rho_{i-1} & \text{с вероятностью } 0.3 \\ 0.2 - \rho_{i-1} & \text{с вероятностью } 0.7 \end{cases}. \tag{16}$$

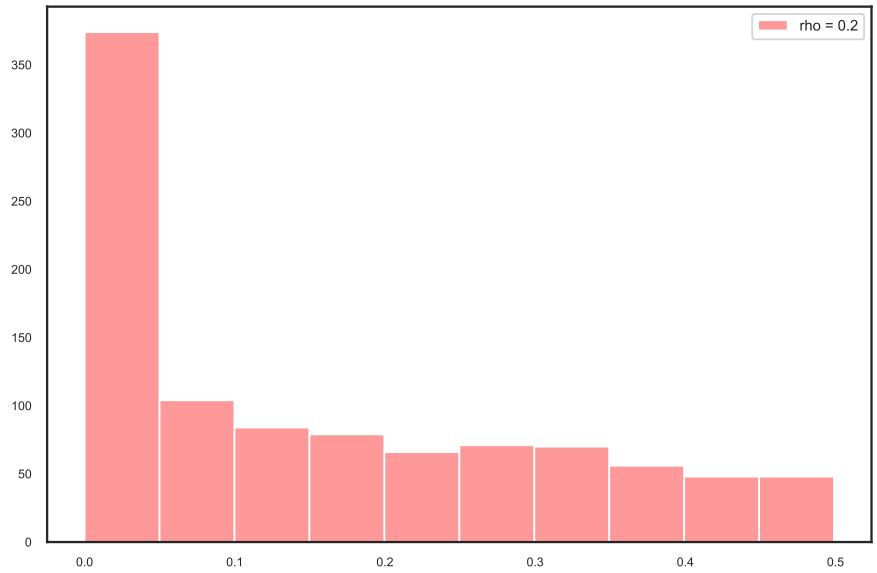

Рис. 14: Гистограма распределение p-value без поправок

Распределение достигаемых уровней показаны на Рис. [14, 15, 16]

Для уровня значимости $\alpha = 0.05$ сравним результаты применения отсутствия поправки на на множественном тестирование и с использованием поправок Бонферони и Бенджамини-Хохберга. Результаты эксперимента показаны в таб. 2.

Метод	Ложно отклоненные	Ложно принятые
Без поправки	64	193
Бонферони	0	478
Бенджамини-Хохберга	1	403

Таблица 2: Для уровня значимости $\alpha = 0.05$

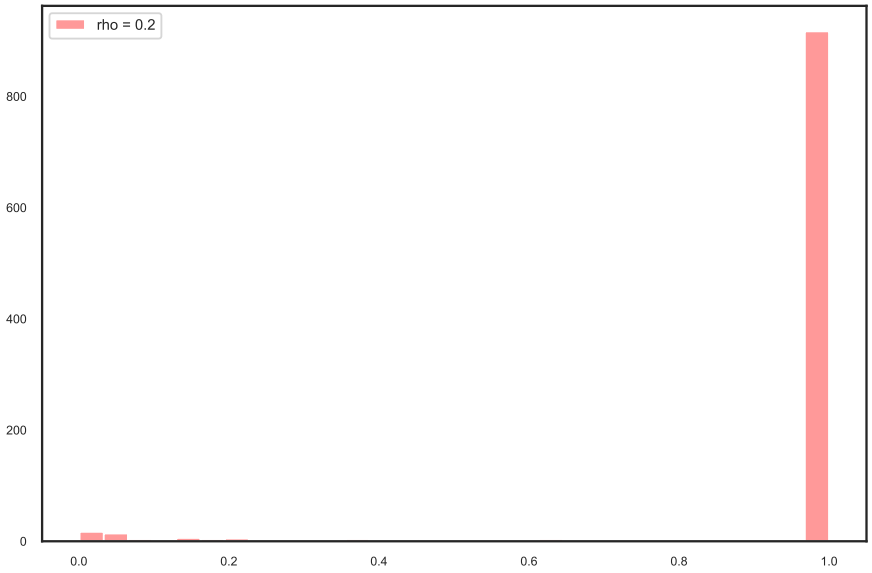

Рис. 15: Гистограма распределение p-value с Банферонии

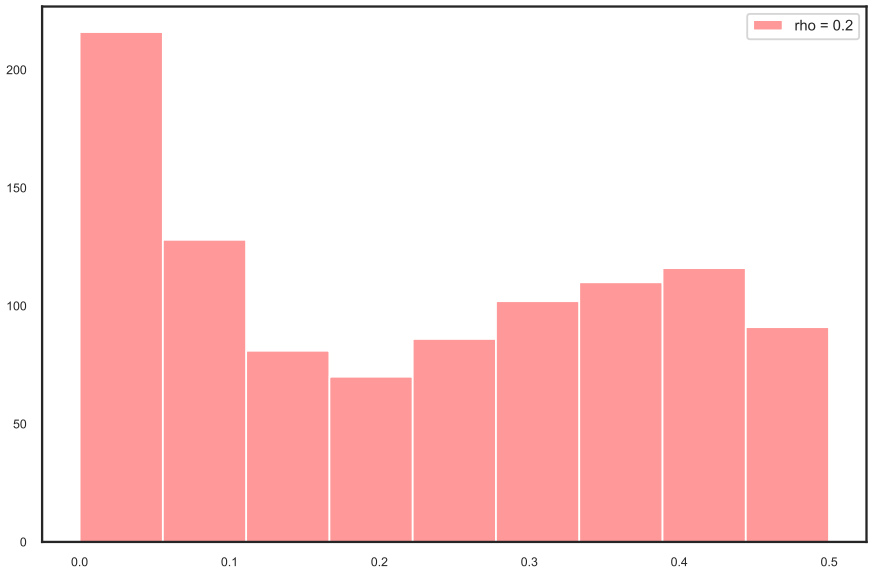

Рис. 16: Гистограма распределение p-value с Бенджамини-Хохберг