

ВШБ Бизнес-информатика: ТВиМС 2025.

Экзаменационный вариант 3

1. (21 балл) Предположим, что у нас есть реализации случайных выборок: $\mathcal{X} = \{x_1, \dots, x_n\}$ и $\mathcal{Y} = \{y_1, \dots, y_m\}$, случайных величин $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ и $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ соответственно. В следующих пунктах исследуются различные аспекты этих выборок.

- (а) (3 балла) Пусть \bar{x} — реализация выборочного среднего на выборке \mathcal{X} . Найдите n такое, что интервал:

$$(\bar{x} - 0.37 \sigma_x, \bar{x} + 0.37 \sigma_x)$$

является приблизительно 90% доверительным интервалом для μ_x .

- (б) (5 баллов) Мы привыкли строить доверительные интервалы для математического ожидания вокруг выборочного среднего. Но ничто не мешает нам забросить эту "рыболовную сеть" вокруг одной реализации x случайной величины X в попытке поймать математическое ожидание μ_x . Каким будет доверительный уровень интервала такой же ширины, как в предыдущем пункте, но построенного на основе всего лишь одной реализации x ?
- (с) (6 баллов) Пусть \bar{X} и \bar{Y} — выборочные средние двух независимых случайных выборок случайных величин X и Y , каждая размера n ($m = n$), где истинные дисперсии известны $\sigma_x^2 = 3\sigma^2$ и $\sigma_y^2 = \sigma^2$ соответственно. Найдите n такое, что:

$$P(\bar{X} + \bar{Y} - 0.27\sigma < \mu_x + \mu_y < \bar{X} + \bar{Y} + 0.27\sigma) = 0.70.$$

- (д) (7 баллов) Мы хотим проверить гипотезу о том, что $\mu_x = \mu_y + \Delta$ против двусторонней альтернативной гипотезы. Предположим, что известны данные: $n = 6$, $m = 8$, $\bar{x} = 9.5$, $\bar{y} = 4.5$, $\Delta = 2$, $\sigma_x = 2.5$, $\sigma_y = 3.5$. Используйте данные и проведите тест, используя уровни значимости $\alpha = 2\%$, 10% .

Для полного оценивания этого пункта недостаточно просто сказать, отклоняем ли мы гипотезу или нет. Вам нужно как-то обосновать ваши заключения: укажите используемую статистику и её распределение при нулевой гипотезе, процесс принятия решения (что с чем сравниваем).

Решение:

- (а) фарм баллов :)

Надо всего лишь заметить, что...

$$\begin{aligned} & (\bar{x} - 0.37 \sigma_x, \bar{x} + 0.37 \sigma_x) \\ & \left(\bar{x} - z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \right) \end{aligned}$$

Приравниваем верхнюю или нижнюю границу к теоретической границе доверительного интервала и решаем уравнение относительно n .

Например, для верхней границы:

$$\begin{aligned} \bar{x} + z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} &= \bar{x} + 0.37 \sigma_x \\ \frac{z_{\alpha/2}}{\sqrt{n}} &= 0.37 \\ \sqrt{n} &= \frac{z_{\alpha/2}}{0.37} \Rightarrow n = \left(\frac{z_{\alpha/2}}{0.37} \right)^2 \end{aligned}$$

Для 90% доверительного интервала: $\alpha = 0.10$, $\alpha/2 = 0.05$, $z_{0.05} \approx 1.645$

$$n = \left(\frac{1.645}{0.37} \right)^2 = (4.446)^2 \approx 19.76$$

Округленно: $n = 20$.

Предлагаю тут выработать не очень жесткий критерий касательно округления, потому что не факт, что на семинаре все такие задачки успели порешать. В идеале в таких задачах вроде округляют до верхнего целого, но тут, если оставят 19 я бы оставил полный балл.

- (b) Тут можно решать разными способами.

Приведу одно из возможных грамотных решений. Мы накидываем симметричный интервал такой же ширины, как в предыдущем пункте, но вокруг одной случайной величины X . И хотим узнать уровень доверия такого интервала, это вероятность, что неизвестный параметр μ_x попадет в этот интервал.

$$\begin{aligned} P(X - 0.37\sigma_x < \mu_x < X + 0.37\sigma_x) &=? \\ P(-X - 0.37\sigma_x < -\mu_x < -X + 0.37\sigma_x) &= \\ P(-0.37\sigma_x < X - \mu_x < 0.37\sigma_x) &= \\ P\left(\frac{-0.37\sigma_x}{\sigma_x} < \frac{X - \mu_x}{\sigma_x} < \frac{0.37\sigma_x}{\sigma_x}\right) &= \\ P(-0.37 < Z < 0.37) &= 2 \cdot \Phi(0.37) - 1 \end{aligned}$$

По таблице нормального распределения: $\Phi(0.37) \approx 0.6443$

$$P(-0.37 < Z < 0.37) = 2 \cdot 0.6443 - 1 = 0.2886$$

Доверительный уровень интервала: $\approx 28.86\%$

Наблюдаем красивый математический результат: действительно, из-за того, что у одной случайной величины дисперсия больше, чем дисперсия выборочного среднего, то она будет сильнее отклоняться от своего математического ожидания, и поэтому такой же интервал, накинутый вокруг одной случайной величины, будет иметь меньше вероятность "накрыть" математическое ожидание.

- (c) Тоже может быть несколько способов решения. Приведу подробный.

Тут на самом деле можно заметить, что это доверительный интервал, но не для разности, а для суммы математических ожиданий. Но давайте посчитаем подробно.

Сначала поработаем с распределением суммы двух случайных величин:

$$\bar{X} + \bar{Y} \sim \mathcal{N}\left(\mu_x + \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}\right) = \mathcal{N}\left(\mu_x + \mu_y, \frac{4\sigma^2}{n}\right)$$

А дальше непосредственно займемся вероятностью попадания в интервал:

$$\begin{aligned} P(\bar{X} + \bar{Y} - 0.27\sigma < \mu_x + \mu_y < \bar{X} + \bar{Y} + 0.27\sigma) &= 0.70 \\ P(-0.27\sigma < (\bar{X} + \bar{Y}) - (\mu_x + \mu_y) < 0.27\sigma) &= 0.70 \\ P\left(\frac{-0.27\sigma}{\sqrt{\frac{4\sigma^2}{n}}} < \frac{(\bar{X} + \bar{Y}) - (\mu_x + \mu_y)}{\sqrt{\frac{4\sigma^2}{n}}} < \frac{0.27\sigma}{\sqrt{\frac{4\sigma^2}{n}}}\right) &= 0.70 \\ P\left(-\frac{0.27\sigma}{2\sigma/\sqrt{n}} < Z < \frac{0.27\sigma}{2\sigma/\sqrt{n}}\right) &= 0.70 \\ P\left(-\frac{0.27\sqrt{n}}{2} < Z < \frac{0.27\sqrt{n}}{2}\right) &= 0.70 \end{aligned}$$

Для 70% доверительного интервала: $\alpha = 0.30$, $\alpha/2 = 0.15$, $z_{0.15} = 1.036$

$$\frac{0.27\sqrt{n}}{2} = z_{0.15} = 1.036$$

$$\sqrt{n} = \frac{2 \cdot 1.036}{0.27} = \frac{2.072}{0.27} \approx 7.67$$

$$n = (7.67)^2 \approx 58.8$$

Округленно: $n = 59$

(d) Снова выделяю зелеными пометками то, что точно должно быть, чтобы получить близкий к максимальному балл. и синим то, что желательно, как маркер хорошего понимания.

Тут не было указано, решать через score или через p -значение. Поэтому можно допускать оба варианта.

Исследуемые случайные величины: $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ и $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Истинные дисперсии известны: $\sigma_x = 2.5$, $\sigma_y = 3.5$. При известных дисперсиях для случайных выборок размера $n = 6$ и $m = 8$ можем получить распределения выборочных средних:

$$\bar{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n}\right), \quad \bar{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma_y^2}{m}\right)$$

- ✓ Гипотезы: $H_0 : \mu_x = \mu_y + \Delta$ (или $\mu_x - \mu_y = \Delta$) против $H_1 : \mu_x \neq \mu_y + \Delta$ (или $\mu_x - \mu_y \neq \Delta$) (двусторонний тест)
- ✓ Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$(\bar{X} - \bar{Y}) \sim \mathcal{N}\left(\Delta, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

или после стандартизации:

$$\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

Это хороший маркер понимания, как устроен процесс. Что у нас есть некое распределение, в параметры которого мы верим, и от него происходит расчет статистики теста. В данном случае, мы верим в то, что разница между средними равна Δ , и именно поэтому score будет считаться как: $\frac{\bar{x} - \bar{y} - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$ - в числителе есть

дополнительное вычитание Δ , потому что мы верим что $(\mu_x - \mu_y = \Delta)$.

- ✓ Проверка с помощью z -статистики (score):

$$z\text{-статистика: } z_{\text{score}} = \frac{\bar{x} - \bar{y} - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

Подставляя значения: $z_{\text{score}} = \frac{9.5 - 4.5 - 2}{\sqrt{\frac{6.25}{6} + \frac{12.25}{8}}} = \frac{3}{\sqrt{1.042 + 1.531}} = \frac{3}{\sqrt{2.573}} = \frac{3}{1.604} \approx 1.870$

Критические точки для двустороннего теста:

- При $\alpha = 2\%$: $z_{0.01} = 2.326$
- При $\alpha = 10\%$: $z_{0.05} = 1.645$

Сравнение с критическими точками:

- При $\alpha = 2\%$: $|z_{\text{score}}| = 1.870 < z_{0.01} = 2.326$, не отклоняем H_0
- При $\alpha = 10\%$: $|z_{\text{score}}| = 1.870 > z_{0.05} = 1.645$, отклоняем H_0

- ✓ p -значение: $p\text{-value} = 2 \cdot P(Z > 1.870) = 2 \cdot 0.0307 = 0.061$

Так как $p\text{-value} = 0.061 > 0.02$ и $p\text{-value} = 0.061 < 0.10$, не отклоняем H_0 на уровне 2%, но отклоняем на уровне 10%.

- Чисто теоретически еще возможно, что кто-то найдет решающую границу в оригинальных единицах. Можем на всякий случай проверить так тоже.

Правая граница:

$$\begin{aligned} P_{H_0} (\bar{X} - \bar{Y} > K_R) &= \alpha/2 \\ P_{H_0} (\bar{X} - \bar{Y} - \Delta > K_R - \Delta) &= \alpha/2 \\ P_{H_0} \left(\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} > \frac{K_R - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \right) &= \alpha/2 \\ P_{H_0} \left(Z > \frac{K_R - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \right) &= \alpha/2 \\ K_R &= \Delta + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \end{aligned}$$

Левая граница:

$$\begin{aligned}
 P_{H_0}(\bar{X} - \bar{Y} < K_L) &= \alpha/2 \\
 P_{H_0}(\bar{X} - \bar{Y} - \Delta < K_L - \Delta) &= \alpha/2 \\
 P_{H_0}\left(\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} < \frac{K_L - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}\right) &= \alpha/2 \\
 P_{H_0}\left(Z < \frac{K_L - \Delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}\right) &= \alpha/2 \\
 K_L &= \Delta - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}
 \end{aligned}$$

Границы критической области:

$$\begin{aligned}
 K_R &= \Delta + z_{\alpha/2} \cdot SE \\
 K_L &= \Delta - z_{\alpha/2} \cdot SE
 \end{aligned}$$

где $SE = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} = \sqrt{\frac{6.25}{6} + \frac{12.25}{8}} = \sqrt{2.573} \approx 1.604$

– При $\alpha = 2\%$: $z_{0.01} = 2.326$

$$\begin{aligned}
 K_L &= \Delta - z_{0.01} \cdot SE = 2 - 2.326 \cdot 1.604 \approx -1.73 \\
 K_R &= \Delta + z_{0.01} \cdot SE = 2 + 2.326 \cdot 1.604 \approx 5.73
 \end{aligned}$$

Проверка: $\bar{x} - \bar{y} = 9.5 - 4.5 = 5.0$

Так как $K_L = -1.73 < 5.0 < K_R = 5.73$, не отклоняем H_0

– При $\alpha = 10\%$: $z_{0.05} = 1.645$

$$\begin{aligned}
 K_L &= \Delta - z_{0.05} \cdot SE = 2 - 1.645 \cdot 1.604 \approx -0.64 \\
 K_R &= \Delta + z_{0.05} \cdot SE = 2 + 1.645 \cdot 1.604 \approx 4.64
 \end{aligned}$$

Проверка: $\bar{x} - \bar{y} = 5.0$

Так как $K_L = -0.64 < 5.0$, но $5.0 > 4.64 = K_R$, отклоняем H_0

Результаты проверки в двух подходах (через score и через границы в оригинальных единицах) идентичны.

2. (8 баллов) В ресторане "Вкусно и вопросительный знак" поток клиентов моделируется Пуассоновским процессом. Известно, что вероятность, что в определенное время суток за час в ресторан придет хотя бы 1 клиент составляет 0.89. Какова вероятность, что в случайный момент в течение этого времени суток ждать следующего вошедшего клиента мы будем от 10 до 20 минут?

3. (12 баллов) Кредитные риски при выдаче кредита в компании "ВопросБанк" моделируются распределением Лапласа с параметрами $\alpha = 0.01$ и $\beta = 0.3$, которое имеет следующую функцию плотности:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x-\beta|}, \quad -\infty < x < +\infty,$$

где $\alpha > 0$ - параметр масштаба, $-\infty < \beta < +\infty$ - параметр сдвига.

Начальный момент k -го порядка для распределения Лапласа может быть рассчитан по следующей формуле:

$$\mathbb{E}[X^k] = \int_{-\infty}^{+\infty} x^k f(x) dx = \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{\beta^{k-2i}}{\alpha^{2i}} \frac{k!}{(k-2i)!},$$

где $\lfloor k/2 \rfloor$ - целая часть $k/2$.

Каждый день ВопросБанк обрабатывает 1000 независимых заявок на получение кредита. Какова вероятность, что средний дневной риск по всем клиентам за день поднимется выше отметки 0.7 от 150 до 160 раз за год?

4. (6 баллов) Расход топлива двух различных моделей автомобилей (A и B) сравнивался в эксперименте. Была взята случайная выборка из 28 автомобилей модели A и 32 автомобилей модели B, и для каждого автомобиля была измерена эффективность использования топлива (в километрах на литр). Результаты обобщены в таблице ниже.

	Размер выборки	Выборочное среднее	Выборочное стандартное отклонение
Модель A	28	26.8	6.5
Модель B	32	31.2	6.3

- (a) (2 балла) Посчитайте 99% доверительный интервал для математического ожидания эффективности использования топлива автомобилей модели B.
- (b) (4 балла) Используйте соответствующий тест гипотез на любых двух уровнях значимости из (1%, 2%, 5%), чтобы определить, могут ли автомобили модели B проехать больше километров на литр, чем автомобили модели A. Сформулируйте гипотезы и ваши предположения касательно свойств и характеристик случайных величин, которые вы исследуете. Укажите используемую статистику и её распределение при нулевой гипотезе. Оформите ваши результаты и сделайте выводы.

Решение:

- (a) Фарм баллов :)

$$\begin{aligned}\bar{x} &= 31.2, \\ s &= 6.3, \\ t_{0.005,31} &= 2.744,\end{aligned}$$

Нужно удостовериться, что студенты понимают, что здесь нужно использовать t -распределение, а не нормальное.

Итоговый интервал:

$$\begin{aligned}\left(\bar{x} - t_{0.005,31} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{0.005,31} \cdot \frac{s}{\sqrt{n}} \right) &= \\ \left(31.2 - 2.744 \cdot \frac{6.3}{\sqrt{32}}, 31.2 + 2.744 \cdot \frac{6.3}{\sqrt{32}} \right) &= \\ (31.2 - 2.744 \cdot 1.114, 31.2 + 2.744 \cdot 1.114) &= \\ (31.2 - 3.056, 31.2 + 3.056) &= \\ (28.144, 34.256)\end{aligned}$$

- (b) Здесь студенты могут пойти двумя путями: либо провести тест Уэлча (когда дисперсии не равны), либо предложить равенство истинных дисперсий и провести тест Стьюдента. Это должно быть прописано в предположениях, хотя бы что минимальное "считаем истинные дисперсии разными" или "предполагаем равенство истинных дисперсий".

Приведу решение тестом Уэлча.

Выделю зелеными пометками то, что точно должно быть, чтобы получить близкий к максимальному балл. и синим то, что желательно, как маркер хорошего понимания.

- ✓ Гипотезы: $H_0 : \mu_A = \mu_B$ (или $\mu_A - \mu_B = 0$) против $H_1 : \mu_B > \mu_A$ (или $\mu_A - \mu_B < 0$) (левосторонний тест)
- ✓ Распределения при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна.

$$\bar{X} - \bar{Y} \sim \mathcal{N} \left(0, \frac{\sigma_X^2}{28} + \frac{\sigma_Y^2}{32} \right), \text{ или } \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{28} + \frac{\sigma_Y^2}{32}}} \sim t_k$$

где число степеней свободы k задаётся формулой:

$$k \approx \frac{(V_X + V_Y)^2}{\frac{V_X^2}{n-1} + \frac{V_Y^2}{m-1}}, \text{ где } V_X = \frac{S_X^2}{n}, V_Y = \frac{S_Y^2}{m}$$

Это хороший маркер понимания, как устроен процесс. Что у нас есть некое распределение, в параметры которого мы верим, и от него происходит расчет статистики теста. В данном случае, мы верим в то, что разница между средними равна 0, и именно поэтому score будет считаться как: $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$ - в числителе нет никаких дополнительных вычитаний, потому что мы верим что $(\mu_A - \mu_B = 0)$.

- ✓ Вычисление степеней свободы:

$$V_X = \frac{6.5^2}{28} = \frac{42.25}{28} \approx 1.509, \quad V_Y = \frac{6.3^2}{32} = \frac{39.69}{32} \approx 1.240$$

$$k = \frac{(1.509 + 1.240)^2}{\frac{1.509^2}{27} + \frac{1.240^2}{31}} = \frac{7.560}{0.084 + 0.050} \approx 56.4 \approx 56$$

В зависимости от используемой таблицы, у них может не быть конкретного такого значения. Значит, должно быть хоть что-то совсем маленькое написано, из разряда "у меня в таблице нет такого значения, но я использую такое-то". Безопасный выбор - брать меньшее ближайшее, и для гипотез и для интервалов. В любом случае предлагаю не сильно карать за это, пусть просто правильно найдут изначальное кол-во степеней свободы, а потом возьмут любое ближайшее, или среднее, если это между двумя соседними.

- Критические точки t -распределения с 56 степенями свободы (левосторонний тест): $t_{(56,0.01)} \approx -2.394$, $t_{(56,0.02)} \approx -2.102$, $t_{(56,0.05)} \approx -1.672$
- Критические точки t -распределения с 50 степенями свободы (левосторонний тест): $t_{(50,0.01)} \approx -2.403$, $t_{(50,0.02)} \approx -2.109$, $t_{(50,0.05)} \approx -1.676$
- Критические точки t -распределения с 60 степенями свободы (левосторонний тест): $t_{(60,0.01)} \approx -2.390$, $t_{(60,0.02)} \approx -2.099$, $t_{(60,0.05)} \approx -1.671$
- ✓ Проверка с помощью t -статистики (score):

$$t\text{-статистика: } t_{\text{score}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

$$\text{Подставляя значения: } t_{\text{score}} = \frac{26.8 - 31.2}{\sqrt{\frac{6.5^2}{28} + \frac{6.3^2}{32}}} = \frac{-4.4}{\sqrt{1.509 + 1.240}} = \frac{-4.4}{\sqrt{2.749}} \approx -2.654$$

Сравнение с критическими точками (для идеального случая), для 50 или 60 степеней свободы проверяйте :)

- При $\alpha = 1\%$: $t_{\text{score}} = -2.654 < t_{(56,0.01)} = -2.394$, отклоняем H_0
- При $\alpha = 2\%$: $t_{\text{score}} = -2.654 < t_{(56,0.02)} = -2.102$, отклоняем H_0
- При $\alpha = 5\%$: $t_{\text{score}} = -2.654 < t_{(56,0.05)} = -1.672$, отклоняем H_0

5. (14 баллов) На дисциплине «Теория вероятностей» 50% студентов списывают. Поэтому преподаватели помимо письменной части экзамена ввели ещё и обязательную устную защиту для всех студентов. Известно, что студенты, которые списывали на письменной части экзамена, на устной защите на каждый вопрос по решённой ими задаче независимо отвечают с вероятностью 0.65. Студенты, которые решали письменный экзамен самостоятельно, на каждый вопрос по своей работе независимо отвечают с вероятностью 0.95. Студентам на устной защите задаётся 9 вопросов.

Какой максимальный порог отсечения K нужно ввести (ответил хотя бы на K вопросов — защищился; не ответил хотя бы на K вопросов — обнуление), чтобы при количестве ответов меньше K вероятность того, что студент списал, была бы не ниже 75%?

6. (14 баллов) Предположим, что у нас есть реализация случайной выборки $\mathcal{X} = (x_1, \dots, x_n)$ неизвестной случайной величины X с плотностью

$$f_X(x; \theta) = \begin{cases} \theta x^{\theta-1}, & \text{при } x \in [0, 1] \\ 0, & \text{иначе.} \end{cases}$$

Реализация, которая была получена: $(x_1, \dots, x_6) = (0.57, 0.04, 0.79, 0.47, 0.86, 0.23)$.

(a) (7 баллов) Найдите оценку параметра θ - функцию от выборки $\hat{\theta}_{ML} = \hat{\theta}_{ML}(\mathcal{X})$ - методом максимального правдоподобия.

Посчитайте её реализацию на предоставленных данных.

(b) (7 баллов) Найдите оценку параметра θ - функцию от выборки $\hat{\theta}_{MM} = \hat{\theta}_{MM}(\mathcal{X})$ - методом моментов.

Посчитайте её реализацию на предоставленных данных.

Решение:

(a) Метод максимального правдоподобия:

$$\begin{aligned} L &= \prod_{i=1}^n \theta x_i^{\theta-1} = \theta^n (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{\theta-1}, \\ l &= \ln L = n \ln \theta + (\theta - 1) \left(\sum \ln x_i \right), \\ \frac{\partial l}{\partial \theta} &= \frac{n}{\theta} + \sum \ln x_i, \\ \frac{n}{\theta} + \sum \ln x_i &= 0, \\ \hat{\theta}_{ML} &= \frac{-n}{\sum \ln x_i}, \end{aligned}$$

(b) Метод моментов:

$$\begin{aligned} \mathbb{E}(X_i) &= \int_0^1 x \cdot f(x) dx = \int_0^1 x \cdot \theta x^{\theta-1} dx = \theta \int_0^1 x^\theta dx, \\ \mathbb{E}(X_i) &= \theta \left[\frac{x^{\theta+1}}{\theta+1} \right]_0^1 = \frac{\theta}{\theta+1}, \\ \frac{\hat{\theta}_{MM}}{\hat{\theta}_{MM} + 1} &= \bar{X}, \\ \hat{\theta}_{MM} &= \bar{X} (\hat{\theta}_{MM} + 1), \\ \hat{\theta}_{MM} - \bar{X} \hat{\theta}_{MM} &= \bar{X}, \\ \hat{\theta}_{MM} (1 - \bar{X}) &= \bar{X}, \\ \hat{\theta}_{MM} &= \frac{\bar{X}}{1 - \bar{X}} \end{aligned}$$

Посчитать реализации: подставить числа в общие формулы. Возможно кто-то сразу вел в числах, тогда нужно сверить финальный ответ, однако если нет формулы в общем виде, то может не ставить максимальный балл, а немного, но снять. Потому что в вопросе явно разделено: оценка как функция, и отдельно её реализация на известных числах.

7. (10 баллов) Телефонная компания собрала случайную выборку из 400 человек, чтобы определить, нравится ли им дизайн их последнего мобильного телефона. Таблица ниже обобщает ответы людей.

	Размер выборки	Положительное мнение о дизайне
Мужчины	180	91
Женщины	220	129

- (a) (2 балла) Посчитайте 90% доверительный интервал для доли мужчин, которым понравился дизайн.
- (b) (2 балла) Посчитайте 98% доверительный интервал для истинной разности долей мужчин и женщин, которым понравился дизайн.
- (c) (5 баллов) Проведите двусторонний тест на уровне значимости 2%, чтобы проверить гипотезу о том, что доля женщин, которым понравился дизайн, равна доле мужчин, которым понравился дизайн. Сформулируйте гипотезы, укажите используемую статистику и её распределение при нулевой гипотезе. Проведите тестирование через: **score** (критерий) и **p**-значение.
- (d) (1 балл) Оформите ваши результаты, сравните результаты пункта (c) с доверительным интервалом из пункта (b) и сделайте выводы.

Решение:

- (a) Данные: $n_1 = 180$ (мужчины), $\tilde{p}_1 = \frac{91}{180} \approx 0.5056$

Доверительный интервал для доли мужчин p_1 :

$$p_1 \in \left(\tilde{p}_1 - z_{0.05} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1}}, \tilde{p}_1 + z_{0.05} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1}} \right)$$

$$z_{0.05} = 1.645, \text{ подкоренное выражение: } \sqrt{\frac{0.5056 \cdot 0.4944}{180}} \approx 0.0373$$

$$\text{Интервал: } 0.5056 \pm 1.645 \cdot 0.0373 = (0.4443, 0.5669)$$

- (b) Данные: $n_1 = 180$ (мужчины), $\tilde{p}_1 = \frac{91}{180} \approx 0.5056$; $n_2 = 220$ (женщины), $\tilde{p}_2 = \frac{129}{220} \approx 0.5864$

Доверительный интервал для разности долей $p_1 - p_2$:

$$(p_1 - p_2) \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{0.01} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \tilde{p}_1 - \tilde{p}_2 + z_{0.01} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}} \right)$$

$$z_{0.01} = 2.326, \text{ подкоренное выражение: } \sqrt{\frac{0.5056 \cdot 0.4944}{180} + \frac{0.5864 \cdot 0.4136}{220}} \approx 0.0499$$

$$\text{Интервал: } (0.5056 - 0.5864) \pm 2.326 \cdot 0.0499 = -0.0808 \pm 0.1161 = (-0.1969, 0.0353)$$

- (c) • ✓ Гипотезы: $H_0 : p_1 = p_2$ против $H_1 : p_1 \neq p_2$ (двусторонний тест)

- ✓ Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$(\hat{p}_1 - \hat{p}_2) \sim \mathcal{N} \left(0, p_c(1-p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

где p_c — объединённая доля при нулевой гипотезе $p_1 = p_2 = p_c$.

• ✓ Объединённая доля: $p_c = \frac{91+129}{180+220} = \frac{220}{400} = 0.55$

- ✓ Проверка с помощью z -статистики (score):

$$z\text{-статистика: } z_{\text{score}} = \frac{\tilde{p}_1 - \tilde{p}_2}{\sqrt{p_c(1-p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\text{Подставляя значения: } z_{\text{score}} = \frac{0.5056 - 0.5864}{\sqrt{0.55 \cdot 0.45 \cdot \left(\frac{1}{180} + \frac{1}{220} \right)}} = \frac{-0.0808}{\sqrt{0.2475 \cdot 0.0101}} \approx -1.616$$

Критическая точка для двустороннего теста: $z_{0.01} = 2.326$ (так как $\alpha/2 = 0.01$)

Решение: $|z_{\text{score}}| = 1.616 < 2.326$, поэтому не отклоняем H_0 на уровне 2%.

- ✓ p -значение: $p\text{-value} = 2 \cdot P(Z > 1.616) = 2 \cdot 0.0530 = 0.106$

Так как $p\text{-value} = 0.106 > 0.02$, не отклоняем H_0 .

- (d) ✓ Сравнение результатов:

- Двусторонний тест на уровне $\alpha = 2\%$: не отклоняем H_0 , нет оснований считать, что доли различаются.

- 98% доверительный интервал для разности: $(-0.1969, 0.0353)$ содержит 0.

- Результаты согласованы: и тест, и доверительный интервал указывают на то, что нет достаточных оснований считать, что доли мужчин и женщин, которым понравился дизайн, различаются.

8. (18 баллов) Рассмотрим две случайные величины X и Y . Они обе принимают значения 0, 1 и 2. Совместные вероятности для каждой пары заданы следующей таблицей, где $\theta \in \mathbb{R}$ — параметр:

	$X = 0$	$X = 1$	$X = 2$
$Y = 0$	$1 - \frac{9\theta}{10}$	$\frac{\theta}{10}$	$\frac{\theta}{10}$
$Y = 1$	$\frac{2\theta}{10}$	0	$\frac{2\theta}{10}$
$Y = 2$	$\frac{\theta}{10}$	$\frac{\theta}{10}$	$\frac{\theta}{10}$

- (a) (2 балла) Какой диапазон значений может принимать параметр θ ?

В дальнейших пунктах предполагается, что θ находится в диапазоне, найденном в пункте 1, однако вы должны проводить все вычисления для произвольного θ .

- (b) (1 балл) Вычислите

$$P(X = 1 | X + Y = 2).$$

- (c) (2 балла) Постройте таблицу вероятностей условного распределения X при условии $Y = 0$.

- (d) (2 балла) Вычислите $\text{Corr}(X, Y)$.

- (e) (5 баллов) Предположим, что у вас есть реализация случайной выборки: $\mathcal{Y} = (y_1, \dots, y_n)$, где каждая y_i получена из закона распределения случайной величины Y из таблицы выше.

Найдите оценку параметра θ - функцию от выборки $\hat{\theta}_{MM} = \hat{\theta}_{MM}(\mathcal{Y})$ - методом моментов.

- (f) (6 баллов) Рассмотрите $\hat{\theta}_1 = X$ и $\hat{\theta}_2 = \frac{X+Y}{2}$ как оценки для неизвестного параметра θ . Какую из них вы предпочтёте и почему? (Посмотрите свойства этих точечных оценок).

Решение:

- (a) Какой диапазон значений может принимать параметр θ ?

$$\begin{cases} 0 \leq 1 - \frac{9\theta}{10} \leq 1, \\ 0 \leq \frac{\theta}{10} \leq 1, \\ 0 \leq \frac{2\theta}{10} \leq 1, \end{cases} \Rightarrow 0 \leq \theta \leq \frac{10}{9}$$

- (b)

$$P(X = 1 | X + Y = 2) = \frac{P(X = 1, Y = 1)}{P\{(X = 1, Y = 1), (X = 0, Y = 2), (X = 2, Y = 0)\}} = \frac{P(X = 1, Y = 1)}{P(X = 1, Y = 1) + P(X = 0, Y = 2) + P(X = 2, Y = 0)} = \frac{0}{0 + \frac{\theta}{10} + \frac{\theta}{10}} = 0$$

- (c) Постройте таблицу вероятностей условного распределения X при условии $Y = 0$.

$$\begin{aligned} P(X = 0 | Y = 0) &= \frac{1 - \frac{9\theta}{10}}{1 - \frac{7\theta}{10}}, \\ P(X = 1 | Y = 0) &= \frac{\frac{\theta}{10}}{1 - \frac{7\theta}{10}}, \\ P(X = 2 | Y = 0) &= \frac{\frac{\theta}{10}}{1 - \frac{7\theta}{10}}, \end{aligned}$$

- (d) Вычислите $\text{Corr}(X, Y)$.

Построим сначала маргинальные функции вероятности:

$$\begin{aligned} P(X = 0) &= 1 - \frac{9\theta}{10} + \frac{2\theta}{10} + \frac{\theta}{10} = 1 - \frac{6\theta}{10}, \\ P(X = 1) &= \frac{\theta}{10} + 0 + \frac{\theta}{10} = \frac{2\theta}{10}, \\ P(X = 2) &= \frac{\theta}{10} + \frac{2\theta}{10} + \frac{\theta}{10} = \frac{4\theta}{10}, \end{aligned}$$

$$\begin{aligned}P(Y=0) &= 1 - \frac{9\theta}{10} + \frac{\theta}{10} + \frac{\theta}{10} = 1 - \frac{7\theta}{10}, \\P(Y=1) &= \frac{2\theta}{10} + 0 + \frac{2\theta}{10} = \frac{4\theta}{10}, \\P(Y=2) &= \frac{\theta}{10} + \frac{\theta}{10} + \frac{\theta}{10} = \frac{3\theta}{10},\end{aligned}$$

Теперь вычислим математические ожидания:

$$\begin{aligned}\mathbb{E}(X) &= 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) = 1 \cdot \frac{2\theta}{10} + 2 \cdot \frac{4\theta}{10} = \theta, \\\mathbb{E}(Y) &= 0 \cdot P(Y=0) + 1 \cdot P(Y=1) + 2 \cdot P(Y=2) = 1 \cdot \frac{4\theta}{10} + 2 \cdot \frac{3\theta}{10} = \theta,\end{aligned}$$

Вычислим дисперсии:

$$\begin{aligned}\mathbb{E}(X^2) &= 0^2 \cdot P(X=0) + 1^2 \cdot P(X=1) + 2^2 \cdot P(X=2) = 1 \cdot \frac{2\theta}{10} + 4 \cdot \frac{4\theta}{10} = \frac{18\theta}{10}, \\\mathbb{E}(Y^2) &= 0^2 \cdot P(Y=0) + 1^2 \cdot P(Y=1) + 2^2 \cdot P(Y=2) = 1 \cdot \frac{4\theta}{10} + 4 \cdot \frac{3\theta}{10} = \frac{16\theta}{10},\end{aligned}$$

$$\begin{aligned}Var[X] &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{18\theta}{10} - \theta^2 = \frac{9\theta}{5} - \theta^2, \\Var[Y] &= \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \frac{16\theta}{10} - \theta^2 = \frac{8\theta}{5} - \theta^2,\end{aligned}$$

Посчитаем $E[XY]$:

$$\begin{aligned}E[XY] &= 2 \cdot 1 \cdot P(X=2, Y=1) + 1 \cdot 2 \cdot P(X=1, Y=2) + 2 \cdot 2 \cdot P(X=2, Y=2) = \\&= 2 \cdot 1 \cdot \frac{2\theta}{10} + 1 \cdot 2 \cdot \frac{\theta}{10} + 2 \cdot 2 \cdot \frac{\theta}{10} = \frac{4\theta}{10} + \frac{2\theta}{10} + \frac{4\theta}{10} = \frac{10\theta}{10} = \theta,\end{aligned}$$

Далее собираем всё вместе:

$$\text{Corr}(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var[X]Var[Y]}} = \frac{\theta - \theta^2}{\sqrt{\frac{9\theta}{5} - \theta^2} \cdot \sqrt{\frac{8\theta}{5} - \theta^2}}$$

Далее вряд ли упрощается, если пройдены все шаги корректно, то такой ответ будет достаточным.

- (e) Найдите оценку параметра θ - функцию от выборки $\hat{\theta}_{MM} = \hat{\theta}_{MM}(\mathcal{Y})$ - методом моментов.

Оказалось просто тут...

$$\begin{aligned}\bar{Y} &= E[Y] = \theta, \\\hat{\theta}_{MM} &= \bar{Y}\end{aligned}$$

- (f) Рассмотрим $\hat{\theta}_1 = X$ и $\hat{\theta}_2 = \frac{X+Y}{2}$ как оценки для неизвестного параметра θ .

Посмотрим на смещённость:

$$\begin{aligned}E[\hat{\theta}_1] &= E[X] = \theta, \\E[\hat{\theta}_2] &= E\left[\frac{X+Y}{2}\right] = \frac{E[X]+E[Y]}{2} = \frac{\theta+\theta}{2} = \theta,\end{aligned}$$

Они обе несмещенные. Посмотрим на дисперсии оценок:

$$\begin{aligned} Var[\hat{\theta}_1] &= Var[X] = \frac{9\theta}{5} - \theta^2, \\ Var[\hat{\theta}_2] &= Var\left[\frac{X+Y}{2}\right] = \frac{Var[X] + Var[Y] + 2Cov(X, Y)}{4} = \\ &\quad \frac{\frac{9\theta}{5} - \theta^2 + \frac{8\theta}{5} - \theta^2 + 2(\theta - \theta^2)}{4} = \\ &\quad \frac{\frac{17\theta}{5} + 2\theta - 4\theta^2}{4} = \frac{\frac{27\theta}{5} - 4\theta^2}{4} = \frac{27\theta}{20} - \theta^2 \end{aligned}$$

Если доводить анализ прям до конца, то нужно посмотреть, какая из дисперсий меньше. Условно, попробуем посчитать, когда дисперсия $\hat{\theta}_2$ меньше дисперсии $\hat{\theta}_1$.

$$\begin{aligned} \frac{27\theta}{20} - \theta^2 &< \frac{9\theta}{5} - \theta^2, \\ \frac{27\theta}{20} &< \frac{9\theta}{5}, \\ \frac{27}{20} &< \frac{9}{5}, \\ \frac{27}{20} &< \frac{36}{20}, \end{aligned}$$

Это неравенство всегда верно для $\theta > 0$. То есть у оценки $\hat{\theta}_2$ дисперсия меньше, чем у оценки $\hat{\theta}_1$ для всех допустимых значений θ ($0 < \theta \leq \frac{10}{9}$).

Значит, мы предпочтем оценку $\hat{\theta}_2$.