ВШБ Бизнес-информатика: ТВиМС 2025.
Экзаменационный вариант 2

1. (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 никто не придет составляет 0.07. Какова вероятность, что в случайный момент в течение этого времени суток ждать следующего вошедшего клиента мы будем больше 15 минут?

2. (10 баллов) Студенческий совет университета собрал случайную выборку из 525 студентов, чтобы определить, поддерживают ли они новое расписание экзаменов. Результаты обобщены в таблице ниже.

Область обучения	Размер выборки	Поддерживают новое расписание экзаменов
Гуманитарные науки	325	230
Естественные науки	200	110

(a) (2 балла) Посчитайте 92% доверительный интервал для истинной доли студентов гуманитарных наук, которые поддерживают новое расписание экзаменов.

(b) (2 балла) Посчитайте 96% доверительный интервал для истинной разности долей положительных ответов между студентами гуманитарных и естественных наук.

(c) (5 баллов) Проведите двусторонний тест на уровне значимости 4%, чтобы проверить гипотезу о том, что доля студентов гуманитарных наук, которые поддерживают новое расписание экзаменов, равна доле студентов естественных наук, которые поддерживают новое расписание экзаменов. 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:

Зеленые пометки: то, что точно должно быть, чтобы получить близкий к максимальному балл. Синие пометки: то, что желательно, как маркер хорошего понимания, но не обязательно.

(a) Данные: $n_1 = 325$ (гуманитарные науки), $\tilde{p}_1 = \frac{230}{325} \approx 0.7077$
Доверительный интервал для доли студентов гуманитарных наук p_1:

$$p_1 \in \left(\tilde{p}_1 - z_{0.04}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1}}, \tilde{p}_1 + z_{0.04}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1}} \right)$$

$z_{0.04} = 1.751$, подкоренное выражение: $\sqrt{\frac{0.7077 \cdot 0.2923}{325}} \approx 0.0252$
Интервал: $0.7077 \pm 1.751 \cdot 0.0252 = (0.6635, 0.7519)$

(b) Данные: $n_1 = 325$ (гуманитарные науки), $\tilde{p}_1 = \frac{230}{325} \approx 0.7077$; $n_2 = 200$ (естественные науки), $\tilde{p}_2 = \frac{110}{200} = 0.5500$
Доверительный интервал для разности долей $p_1 - p_2$:

$$(p_1 - p_2) \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{0.02}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \tilde{p}_1 - \tilde{p}_2 + z_{0.02}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}} \right)$$

$z_{0.02} = 2.054$, подкоренное выражение: $\sqrt{\frac{0.7077 \cdot 0.2923}{325} + \frac{0.55 \cdot 0.45}{200}} \approx 0.0433$
Интервал: $(0.7077 - 0.5500) \pm 2.054 \cdot 0.0433 = 0.1577 \pm 0.0889 = (0.0688, 0.2466)$

(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: $p_c = \frac{230+110}{325+200} = \frac{340}{525} \approx 0.6476$
- ✓ Проверка с помощью z-статистики (score):
  z-статистика: $z_{\text{score}} = \frac{\tilde{p}_1 - \tilde{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
  Подставляя значения: $z_{\text{score}} = \frac{0.7077 - 0.5500}{\sqrt{0.6476 \cdot 0.3524 \cdot \left(\frac{1}{325}+\frac{1}{200}\right)}} = \frac{0.1577}{\sqrt{0.2282 \cdot 0.00808}} \approx 3.673$
  Критическая точка для двустороннего теста: $z_{0.02} = 2.054$ (так как $\alpha/2 = 0.02$)
  Решение: $|z_{\text{score}}| = 3.673 > 2.054$, поэтому отклоняем H_0 на уровне 4%.
- ✓ p-значение: p-value $= 2 \cdot P(Z > 3.673) = 2 \cdot 0.00012 = 0.00024$
  Так как p-value $= 0.00024 < 0.04$, отклоняем H_0.

(d) ✓ Сравнение результатов:

- Двусторонний тест на уровне $\alpha = 4\%$: отклоняем H_0, доли различаются.
- 96% доверительный интервал для разности: $(0.0688, 0.2466)$ не содержит 0.
- Результаты согласованы: и тест, и доверительный интервал указывают на то, что доли студентов гуманитарных и естественных наук, которые поддерживают новое расписание экзаменов, различаются. Доля студентов гуманитарных наук выше, чем доля студентов естественных наук.

3. (21 балл) Предположим, что у нас есть реализации случайных выборок: $\mathcal{X} = \{x_1, \ldots, x_n\}$ и $\mathcal{Y} = \{y_1, \ldots, y_m\}$, 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.

   (a) (3 балла) Пусть $\bar{y}$ — реализация выборочного среднего на выборке $\mathcal{Y}$. Найдите m такое, что интервал:

   $$(\bar{y} - 0.55\,\sigma_y, \bar{y} + 0.55\,\sigma_y)$$

   является приблизительно 95% доверительным интервалом для μ_y.

   (b) (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 y случайной величины Y в попытке поймать математическое ожидание μ_y. Каким будет доверительный уровень интервала такой же ширины, как в предыдущем пункте, но построенного на основе всего лишь одной реализации y?

   (c) (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 $\sigma_x^2 = 2\sigma^2$ и $\sigma_y^2 = 2\sigma^2$ соответственно. Найдите n такое, что:

   $$P\left(\bar{X} + \bar{Y} - \frac{2\sigma}{5} < \mu_x + \mu_y < \bar{X} + \bar{Y} + \frac{2\sigma}{5}\right) = 0.86.$$

   (d) (7 баллов) Мы хотим проверить гипотезу о том, что $\mu_x = \mu_y + \Delta$ против двусторонней альтернативной гипотезы. Предположим, что известны данные: $n = 5$, $m = 9$, $\bar{x} = 9.5$, $\bar{y} = 4.5$, $\Delta = 2$, $\sigma_x = 2.2$, $\sigma_y = 3.2$. Используйте данные и проведите тест, используя уровни значимости $\alpha = 2\%,\ 5\%$.

   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:

(a) фарм баллов :)

Сравниваем условие задачи с формулой доверительного интервала для неизвестного математического ожидания:

$$(\bar{y} - 0.55\,\sigma_y,\ \bar{y} + 0.55\,\sigma_y)$$
$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma_y}{\sqrt{m}},\ \bar{y} + z_{\alpha/2}\frac{\sigma_y}{\sqrt{m}}\right)$$

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

Например, для верхней границы:

$$\bar{y} + z_{\alpha/2}\frac{\sigma_y}{\sqrt{m}} = \bar{y} + 0.55\,\sigma_y$$
$$\frac{z_{\alpha/2}}{\sqrt{m}} = 0.55$$
$$\sqrt{m} = \frac{z_{\alpha/2}}{0.55} \Rightarrow m = \left(\frac{z_{\alpha/2}}{0.55}\right)^2$$

Для 95% доверительного интервала: $\alpha = 0.05$, $\alpha/2 = 0.025$, $z_{0.025} \approx 1.960$

$$m = \left(\frac{1.960}{0.55}\right)^2 = (3.564)^2 \approx 12.70$$

Округленно: $m = 13$.

Здесь должно быть округление вверх, без вариантов.

(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 Y. И хотим узнать уровень доверия такого интервала, это вероятность, что неизвестный параметр μ_y попадет в этот интервал.

$$P\left(Y - 0.55\,\sigma_y < \mu_y < Y + 0.55\,\sigma_y\right) = ?$$
$$P\left(-Y - 0.55\,\sigma_y < -\mu_y < -Y + 0.55\,\sigma_y\right) =$$
$$P\left(-0.55\,\sigma_y < Y - \mu_y < 0.55\,\sigma_y\right) =$$
$$P\left(\frac{-0.55\,\sigma_y}{\sigma_y} < \frac{Y - \mu_y}{\sigma_y} < \frac{0.55\,\sigma_y}{\sigma_y}\right) =$$
$$P\left(-0.55 < Z < 0.55\right) = 2 \cdot \Phi(0.55) - 1$$

По таблице нормального распределения: $\Phi(0.55) \approx 0.7088$

$$P\left(-0.55 < Z < 0.55\right) = 2 \cdot 0.7088 - 1 = 0.4176$$

Доверительный уровень интервала: $\approx 41.77\%$

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:

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

где $\sigma_x^2 = 2\sigma^2$ и $\sigma_y^2 = 2\sigma^2$, поэтому $\sigma_x^2 + \sigma_y^2 = 4\sigma^2$.

А дальше непосредственно займемся вероятностью попадания в интервал:

$$P\left(\bar{X} + \bar{Y} - \frac{2\sigma}{5} < \mu_x + \mu_y < \bar{X} + \bar{Y} + \frac{2\sigma}{5}\right) = 0.86$$
$$P\left(-0.4\sigma < (\bar{X} + \bar{Y}) - (\mu_x + \mu_y) < 0.4\sigma\right) = 0.86$$
$$P\left(\frac{-0.4\sigma}{\sqrt{\frac{4\sigma^2}{n}}} < \frac{(\bar{X} + \bar{Y}) - (\mu_x + \mu_y)}{\sqrt{\frac{4\sigma^2}{n}}} < \frac{0.4\sigma}{\sqrt{\frac{4\sigma^2}{n}}}\right) = 0.86$$
$$P\left(-\frac{0.4\sigma}{2\sigma/\sqrt{n}} < Z < \frac{0.4\sigma}{2\sigma/\sqrt{n}}\right) = 0.86$$
$$P\left(-\frac{0.4\sqrt{n}}{2} < Z < \frac{0.4\sqrt{n}}{2}\right) = 0.86$$

Для 86% доверительного интервала: $\alpha = 0.14$, $\alpha/2 = 0.07$, $z_{0.07} \approx 1.476$

$$\frac{0.4\sqrt{n}}{2} = z_{0.07} = 1.476$$

$$\sqrt{n} = \frac{2 \cdot 1.476}{0.4} = \frac{2.952}{0.4} = 7.38$$

$$n = (7.38)^2 \approx 54.45$$

По идее, тут тоже нужно округление вверх, тогда интервал получается как бы чуть шире, чем 86%. Но в формулировке не было сказано ничего, например,как в пункте а), про то что это у нас должен быть "приблизительно"86% доверительный интервал, так что наверное можно любое округление зачесть.

(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: $\sigma_x = 2.2$, $\sigma_y = 3.2$. При известных дисперсиях для случайных выборок размера $n = 5$ и $m = 9$ 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-статистика: $z_{\text{score}} = \frac{\bar{x}-\bar{y}-\Delta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}$
  Подставляя значения: $z_{\text{score}} = \frac{9.5-4.5-2}{\sqrt{\frac{2.2^2}{5}+\frac{3.2^2}{9}}} = \frac{3}{\sqrt{\frac{4.84}{5}+\frac{10.24}{9}}} = \frac{3}{\sqrt{0.968+1.138}} = \frac{3}{\sqrt{2.106}} \approx 2.067$
  Критические точки для двустороннего теста:
  - При $\alpha = 2\%$: $z_{0.01} = 2.326$
  - При $\alpha = 5\%$: $z_{0.025} = 1.960$

  Сравнение с критическими точками:
  - При $\alpha = 2\%$: $|z_{\text{score}}| = 2.067 < z_{0.01} = 2.326$, не отклоняем H_0
  - При $\alpha = 5\%$: $|z_{\text{score}}| = 2.067 > z_{0.025} = 1.960$, отклоняем H_0
- ✓ p-значение: p-value $= 2 \cdot P(Z > 2.067) = 2 \cdot 0.0194 = 0.0387$
  Так как p-value $= 0.0387 > 0.02$ и p-value $= 0.0387 < 0.05$, не отклоняем H_0 на уровне 2%, но отклоняем на уровне 5%.
- Чисто теоретически еще возможно, что кто-то найдет решающую границу в оригинальных единицах. Можем на всякий случай проверить так тоже.
  Правая граница:

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

Левая граница:

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

Границы критической области:

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

где $SE=\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}=\sqrt{\frac{4.84}{5}+\frac{10.24}{9}}=\sqrt{2.106}\approx 1.451$

- При $\alpha=2\%$: $z_{0.01}=2.326$

$$K_L=\Delta-z_{0.01}\cdot SE=2-2.326\cdot 1.451\approx -1.37$$
$$K_R=\Delta+z_{0.01}\cdot SE=2+2.326\cdot 1.451\approx 5.37$$

  Проверка: $\bar{x}-\bar{y}=9.5-4.5=5.0$
  Так как $K_L=-1.37<5.0<5.37=K_R$, не отклоняем H_0
- При $\alpha=5\%$: $z_{0.025}=1.960$

$$K_L=\Delta-z_{0.025}\cdot SE=2-1.960\cdot 1.451\approx -0.84$$
$$K_R=\Delta+z_{0.025}\cdot SE=2+1.960\cdot 1.451\approx 4.84$$

  Проверка: $\bar{x}-\bar{y}=5.0$
  Так как $K_L=-0.84<5.0$, но $5.0>4.84=K_R$, отклоняем H_0

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

4. (12 баллов) Кредитные риски при выдаче кредита в компании "ВопросБанк"моделируются распределением Лапласа с параметрами $\alpha = 0.5$ и $\beta = -0.1$, 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:

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

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 0.1 менее чем 3 раз за год?

5. (14 баллов) Предположим, что у нас есть реализация случайной выборки $\mathcal{X} = (x_1, \ldots, x_n)$ неизвестной случайной величины X с плотностью

$$f_X(x;\theta) = \begin{cases} \frac{4x^3}{\theta^4}, & \text{при } x \in [0,\theta] \\ 0, & \text{иначе.} \end{cases}$$

Реализация, которая была получена: $(x_1, \ldots, x_6) = (1.5, 2.8, 5.4, 0.8, 6.1, 7.2)$.

(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:
В подобных задачах, где параметр также присутствует как граница в области определения функции плотности, возникает дополнительный аспект, который нужно учесть.
Посмотрим на часть функции плотности: $f_X(x;\theta) = \frac{4x^3}{\theta^4}, \;\; x \in [0,\theta]$, что означает, что если $x > \theta$, то функция плотности равна 0.
Запишем функцию правдоподобия, используя индикаторные функции (не обязательно, просто удобно и как возможный пример оформления).
Функция правдоподобия для выборки размера n:

$$L(\theta) = \prod_{i=1}^{n} \frac{4x_i^3}{\theta^4} \cdot I(0 \le x_i \le \theta) = \frac{4^n \prod_{i=1}^{n} x_i^3}{\theta^{4n}} \cdot I(\theta \ge \max(x_i))$$

Индикаторные функции:

$$I(0 \le x_i \le \theta) = \begin{cases} 1, & \text{если } 0 \le x_i \le \theta \\ 0, & \text{иначе} \end{cases}$$

То есть, если хотя бы один $x_i > \theta$, то индикаторная функция равна 0 и вся функция правдоподобия становится равной 0.

$$I(\theta \ge \max(x_i)) = \begin{cases} 1, & \text{если } \theta \ge \max(x_i) \\ 0, & \text{иначе} \end{cases}$$

Анализ $L(\theta)$: Для $\theta \ge \max(x_i)$ функция $L(\theta)$ пропорциональна θ^{-4n}, что является строго убывающей функцией от θ.
Нахождение максимума: Чтобы максимизировать $L(\theta)$, нужно выбрать наименьшее возможное значение θ, которое удовлетворяет условию $\theta \ge \max(x_i)$ (иначе функция правдоподобия будет равна 0).

$$\hat{\theta}_{ML} = \max(X_i)$$

Реализация на данных: $\hat{\theta}_{ML} = \max(1.5, 2.8, 5.4, 0.8, 6.1, 7.2) = 7.2$

(b) Метод моментов:
Первый момент генеральной совокупности — это математическое ожидание $E[X]$.

$$\mathbb{E}(X_i) = \int_0^\theta x \cdot \frac{4x^3}{\theta^4} dx = \frac{4}{\theta^4} \int_0^\theta x^4 dx =$$
$$\frac{4}{\theta^4} \left[\frac{x^5}{5}\right]_0^\theta = \frac{4\theta^5}{5\theta^4} = \frac{4}{5}\theta,$$
$$\frac{4}{5}\hat{\theta}_{MM} = \bar{X},$$
$$\hat{\theta}_{MM} = \frac{5}{4}\bar{X}$$

Реализация на данных: $\bar{x} = \frac{23.8}{6} \approx 3.967$, $\hat{\theta}_{MM} = \frac{5}{4} \cdot 3.967 \approx 4.958$

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.

6. (14 баллов) На дисциплине «Теория вероятностей» 25% 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 0.7. Студенты, которые решали письменный экзамен самостоятельно, на каждый вопрос по своей работе независимо отвечают с вероятностью 0.85. Студентам на устной защите задаётся 8 вопросов.

   Какой максимальный порог отсечения K нужно ввести (ответил хотя бы на K вопросов — защитился; не ответил хотя бы на K вопросов — обнуление), чтобы при количестве ответов меньше K вероятность того, что студент списал, была бы не ниже 80%?

7. (6 баллов) 55 студентов университета на определённом курсе были случайным образом распределены в две учебные группы размером 30 и 25 студентов соответственно. В конце года все студенты сдали экзамен, и их оценки обобщены в таблице ниже.

	Размер выборки	Выборочное среднее	Выборочное стандартное отклонение
Группа 1	30	75.33	7.61
Группа 2	25	71.40	6.37

(a) (2 балла) Посчитайте 95% доверительный интервал для математического ожидания экзаменационных оценок студентов группы 1.

(b) (4 балла) Используйте подходящий тест гипотез, чтобы определить, больше ли средний балл студентов группы 1, чем средний балл студентов группы 2.

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 :)

$$\bar{x} = 75.33,$$
$$s = 7.61,$$
$$t_{0.025,29} = 2.045,$$

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

Итоговый интервал:

$$\left(\bar{x} - t_{0.025,29} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,29} \cdot \frac{s}{\sqrt{n}}\right) =$$
$$\left(75.33 - 2.045 \cdot \frac{7.61}{\sqrt{30}}, 75.33 + 2.045 \cdot \frac{7.61}{\sqrt{30}}\right) =$$
$$(75.33 - 2.045 \cdot 1.389, 75.33 + 2.045 \cdot 1.389) =$$
$$(75.33 - 2.842, 75.33 + 2.842) =$$
$$(72.488, 78.172)$$

(b) **Важно:** В этом варианте не обозначили уровни значимости, на которых нужно проверять. Тогда это на усмотрение студентов. Главное, чтобы это был корректный односторонний тест, правильно посчитана t-статистика (score), и правильно найдены критические точки.

Здесь студенты могут пойти двумя путями: либо провести тест Уэлча (когда дисперсии не равны), либо предположить 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: $H_0 : \mu_1 = \mu_2$ (или $\mu_1 - \mu_2 = 0$) против $H_1 : \mu_1 > \mu_2$ (или $\mu_1 - \mu_2 > 0$) (правосторонний тест)
- ✓ Распределения при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна.

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{30} + \frac{\sigma_2^2}{25}\right), \quad \text{или} \quad \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{30} + \frac{S_2^2}{25}}} \sim t_k$$

где число степеней свободы k задаётся формулой:

$$k \approx \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}, \text{ где } V_1 = \frac{S_1^2}{n_1},\ V_2 = \frac{S_2^2}{n_2}$$

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: $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ - в числителе нет никаких дополнительных вычитаний, потому что мы верим что $(\mu_1 - \mu_2 = 0)$.

- ✓ Вычисление степеней свободы:

$$V_1 = \frac{7.61^2}{30} = \frac{57.9121}{30} \approx 1.930, \quad V_2 = \frac{6.37^2}{25} = \frac{40.5769}{25} \approx 1.623$$

$$k = \frac{(1.930 + 1.623)^2}{\frac{1.930^2}{29} + \frac{1.623^2}{24}} = \frac{12.625}{0.128 + 0.110} \approx 53.0 \approx 53$$

  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 53 степенями свободы (правосторонний тест): $t_{(53,0.01)} \approx 2.398$, $t_{(53,0.02)} \approx 2.105$, $t_{(53,0.05)} \approx 1.674$
- Критические точки t-распределения с 50 степенями свободы (правосторонний 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 (правосторонний 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-статистика: $t_{\text{score}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
  Подставляя значения: $t_{\text{score}} = \frac{75.33 - 71.40}{\sqrt{\frac{7.61^2}{30} + \frac{6.37^2}{25}}} = \frac{3.93}{\sqrt{1.930 + 1.623}} = \frac{3.93}{\sqrt{3.553}} \approx 2.085$
  Сравнение с критическими точками (для идеального случая), для 50 или 60 степеней свободы проверяйте :)
  - При $\alpha = 1\%$: $t_{\text{score}} = 2.085 < t_{(53,0.01)} = 2.398$, не отклоняем H_0
  - При $\alpha = 2\%$: $t_{\text{score}} = 2.085 < t_{(53,0.02)} = 2.105$, не отклоняем H_0
  - При $\alpha = 5\%$: $t_{\text{score}} = 2.085 > t_{(53,0.05)} = 1.674$, отклоняем H_0
- ✓ p-значение: p-value $= P(t > 2.085) \approx 0.021$

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

	$X=-1$	$X=0$	$X=1$
$Y=-1$	0.1	$0.1+\theta$	$0.3+3\theta$
$Y=0$	0	$0.2-6\theta$	$0.1+\theta$
$Y=1$	$0.1+\theta$	0.1	0

(a) (2 балла) Какой диапазон значений может принимать параметр θ, чтобы приведённая выше таблица соответствовала таблице вероятностей?

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
$$P(X=0 \mid X+Y=0).$$

(c) (2 балла) Постройте таблицу вероятностей условного распределения X при условии $Y=1$.

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

(e) (5 баллов) Предположим, что у вас есть реализация случайной выборки: $\mathcal{Y}=(y_1,\ldots,y_m)$, 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, чтобы приведённая выше таблица соответствовала таблице вероятностей?

$$\begin{cases} 0 \le 0.1+\theta \le 1, \\ 0 \le 0.3+3\theta \le 1, \\ 0 \le 0.2-6\theta \le 1, \\ 0 \le 0.1+\theta \le 1, \end{cases} \Rightarrow -0.1 \le \theta \le \frac{1}{30}$$

(b)
$$P(X=0 \mid X+Y=0) = \frac{P(X=0,Y=0)}{P\{(X=-1,Y=1),(X=0,Y=0),(X=1,Y=-1)\}} = \frac{P(X=0,Y=0)}{P(X=-1,Y=1)+P(X=0,Y=0)+P(X=1,Y=-1)} = \frac{0.2-6\theta}{(0.1+\theta)+(0.2-6\theta)+(0.3+3\theta)} = \frac{0.2-6\theta}{0.6-2\theta} = \frac{0.1-3\theta}{0.3-\theta}$$

(c) Постройте таблицу вероятностей условного распределения X при условии $Y=1$.

$$P(X=-1 \mid Y=1) = \frac{0.1+\theta}{0.2+\theta},$$
$$P(X=0 \mid Y=1) = \frac{0.1}{0.2+\theta},$$
$$P(X=1 \mid Y=1) = \frac{0}{0.2+\theta} = 0,$$

Можно проверить, что сумма условных вероятностей равна 1. Вроде все хорошо :)

(d) Вычислите $\mathrm{Corr}(X,Y)$.
Построим сначала маргинальные функции вероятности:

$$P(X=-1) = 0.1+0+(0.1+\theta) = 0.2+\theta,$$
$$P(X=0) = (0.1+\theta)+(0.2-6\theta)+0.1 = 0.4-5\theta,$$
$$P(X=1) = (0.3+3\theta)+(0.1+\theta)+0 = 0.4+4\theta,$$

$$P(Y=-1) = 0.1 + (0.1+\theta) + (0.3+3\theta) = 0.5 + 4\theta,$$
$$P(Y=0) = 0 + (0.2-6\theta) + (0.1+\theta) = 0.3 - 5\theta,$$
$$P(Y=1) = (0.1+\theta) + 0.1 + 0 = 0.2 + \theta,$$

Теперь вычислим математические ожидания:

$$\mathbb{E}(X) = (-1)\cdot P(X=-1) + 0\cdot P(X=0) + 1\cdot P(X=1) =$$
$$(-1)\cdot(0.2+\theta) + 1\cdot(0.4+4\theta) = 0.2 + 3\theta,$$
$$\mathbb{E}(Y) = (-1)\cdot P(Y=-1) + 0\cdot P(Y=0) + 1\cdot P(Y=1) =$$
$$(-1)\cdot(0.5+4\theta) + 1\cdot(0.2+\theta) = -0.3 - 3\theta,$$

Вычислим дисперсии:

$$\mathbb{E}(X^2) = (-1)^2\cdot P(X=-1) + 0^2\cdot P(X=0) + 1^2\cdot P(X=1) =$$
$$1\cdot(0.2+\theta) + 1\cdot(0.4+4\theta) = 0.6 + 5\theta,$$
$$\mathbb{E}(Y^2) = (-1)^2\cdot P(Y=-1) + 0^2\cdot P(Y=0) + 1^2\cdot P(Y=1) =$$
$$1\cdot(0.5+4\theta) + 1\cdot(0.2+\theta) = 0.7 + 5\theta,$$

$$Var[X] = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = (0.6+5\theta) - (0.2+3\theta)^2 =$$
$$0.6 + 5\theta - (0.04 + 1.2\theta + 9\theta^2) = 0.56 + 3.8\theta - 9\theta^2,$$
$$Var[Y] = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = (0.7+5\theta) - (-0.3-3\theta)^2 =$$
$$0.7 + 5\theta - (0.09 + 1.8\theta + 9\theta^2) = 0.61 + 3.2\theta - 9\theta^2,$$

Посчитаем $E[XY]$:

$$E[XY] = (-1)\cdot(-1)\cdot 0.1 + (-1)\cdot 0\cdot 0 + (-1)\cdot 1\cdot(0.1+\theta) +$$
$$0\cdot(-1)\cdot(0.1+\theta) + 0\cdot 0\cdot(0.2-6\theta) + 0\cdot 1\cdot(0.1+\theta) +$$
$$1\cdot(-1)\cdot(0.3+3\theta) + 1\cdot 0\cdot 0.1 + 1\cdot 1\cdot 0$$
$$E[XY] = 0.1 - (0.1+\theta) - (0.3+3\theta) = -0.3 - 4\theta$$

Далее собираем всё вместе:

$$Cov(X,Y) = E[XY] - E[X]E[Y] = (-0.3-4\theta) - (0.2+3\theta)(-0.3-3\theta) =$$
$$(-0.3-4\theta) - (-0.06 - 0.6\theta - 0.9\theta - 9\theta^2) =$$
$$(-0.3-4\theta) - (-0.06 - 1.5\theta - 9\theta^2) =$$
$$-0.3 - 4\theta + 0.06 + 1.5\theta + 9\theta^2 = -0.24 - 2.5\theta + 9\theta^2$$

$$\mathrm{Corr}(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}} =$$
$$\frac{-0.24 - 2.5\theta + 9\theta^2}{\sqrt{(0.56 + 3.8\theta - 9\theta^2)\cdot(0.61 + 3.2\theta - 9\theta^2)}}$$

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.

$$\bar{Y} = \mathbb{E}(Y) = -0.3 - 3\theta,$$
$$-0.3 - 3\hat{\theta}_{MM} = \bar{Y},$$
$$\hat{\theta}_{MM} = -\frac{\bar{Y} + 0.3}{3}$$

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

$$\mathbb{E}(\hat{\theta}_1) = \mathbb{E}(X) = 0.2 + 3\theta,$$
$$\mathbb{E}(\hat{\theta}_2) = \mathbb{E}\left(\frac{X+Y}{2}\right) = \frac{\mathbb{E}(X) + \mathbb{E}(Y)}{2} = \frac{(0.2+3\theta) + (-0.3-3\theta)}{2} = \frac{-0.1}{2} = -0.05,$$

Обе оценки смещенные. Посмотрим на дисперсии оценок:

$$Var[\hat{\theta}_1] = Var[X] = 0.56 + 3.8\theta - 9\theta^2,$$
$$Var[\hat{\theta}_2] = Var\left(\frac{X+Y}{2}\right) = \frac{Var[X] + Var[Y] + 2Cov(X,Y)}{4}$$
$$Var[\hat{\theta}_2] = \frac{(0.56 + 3.8\theta - 9\theta^2) + (0.61 + 3.2\theta - 9\theta^2) + 2(-0.24 - 2.5\theta + 9\theta^2)}{4} =$$
$$\frac{0.56 + 3.8\theta - 9\theta^2 + 0.61 + 3.2\theta - 9\theta^2 - 0.48 - 5\theta + 18\theta^2}{4} =$$
$$\frac{0.69 + 2\theta}{4} = 0.1725 + 0.5\theta$$

где $Cov(X,Y) = -0.24 - 2.5\theta + 9\theta^2$ (вычислено выше).
Вот дальше сравнивать кажется уже не очень целесообразно, потому что слишком много вычислений. Если кто-то дошел и просто посчитал смещения и дисперсии, то такой ответ надо сделать максимальным баллом. (или даже как-то ещё более лояльно)