Задачи на проверку гипотез

Введение в DS на УБ и МиРА

---

1 Задачи с вариантами ответа (кандидаты в квиз)

1. При проверке гипотезы о равенстве средних p-value оказалось равно 0.04. Выберите верный вариант.

   (a) Основная гипотеза не отвергается на любом разумном уровне значимости.

   (b) Основная гипотеза не отвергается на уровне значимости 1%.

   (c) Основная гипотеза не отвергается на уровне значимости 5%.

   (d) Основная гипотеза не отвергается на уровне значимости 10%.

2. Проверяется гипотеза о равенстве средних против двусторонней альтернативы. Пусть при верной H_0 $Z \sim \mathcal{N}(0,1)$. Чему примерно равно соответствующее p-value, если

   (a) $Z_{obs} = -15$?

   (b) $Z_{obs} = 0$?

   (c) $Z_{obs} = 1.96$?

   (d) $Z_{obs} = 7$?

3. Пусть $X_1, \ldots, X_N$ – выборка независимых одинаково распределённых нормальных величин, $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Предположим, что $N = 1000$. Проверяется гипотеза $\mu = 2$ против двусторонней альтернативы. На выбор есть всего два теста: Z-тест и t-тест. Выберите все верные варианты.

   (a) Тестовая статистика может иметь распределение t_{1000}.

   (b) Тестовая статистика может иметь распределение t_{500}.

   (c) Тестовая статистика может иметь распределение, очень похожее на $\mathcal{N}(0,1)$.

   (d) Тестовая статистика может иметь распределение, очень похожее на $\mathcal{N}(1000, 2)$.

4. Определите тип ошибки (I или II рода) по описанию ситуации.

   (a) Мобильный робот врезался в стену. Нулевая гипотеза: впереди нет препятствия.

   (b) Сканер отпечатка пальца не дал согласие на разблокировку системы для зарегистрированного пользователя. Нулевая гипотеза: пользователь есть в базе.

5. Будет ли отвергнута гипотеза о независимости при использовании χ^2-критерия согласия Пирсона на уровне значимости 5%, если

(a) $\chi^2_{obs} = 1$?

(b) $\chi^2_{obs} = 50$?

(c) $\chi^2_{obs} = 100$?

В каждом случае нарисуйте картинку.

2 Задачи с открытым ответом (для тренировки техники проверки гипотез)

1. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, \sigma^2).$$

(a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

на уровне значимости 5%.

(b) Постройте 90%-ый доверительный интервал для μ.

2. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, 9).$$

(a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

на уровне значимости 10%.

(b) Постройте 95%-ый доверительный интервал для μ.

3. Рассмотрим выборку независимых одинаково распределённых случайных величин X_1, ..., X_{200}, где $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Оказалось, что

$$\sum_{i=1}^{200} X_i = 20, \ \sum_{i=1}^{200} X_i^2 = 500.$$

(a) Найдите $\bar{X}$ и $\hat{\sigma}^2$.

(b) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 2, \\ H_1 : \mu \neq 2 \end{cases}$$

на уровне значимости 1%.

(c) Постройте 95%-ый доверительный интервал для μ и при помощи него проверьте гипотезу из предыдущего пункта (уже на новом уровне значимости).

(d) **(пункт со звёздочкой)** Запишите интеграл, при помощи которого можно рассчитать p-value для полученной статистики. Посчитайте этот интеграл при помощи Wolfram Alpha. Убедитесь, что результаты проверки совпадают с предыдущими двумя пунктами.

4. **(хитрая задача)** Рассмотрим выборку независимых одинаково распределённых случайных величин X_1, …, X_{10}. Известно, что распределение X_i не является нормальным. Проверьте гипотезу $H_0 : \mathbb{E}(X_i) = 3$ против двусторонней альтернативы на уровне значимости 5%.

5. **(хитрая задача)** Рассмотрим выборку независимых одинаково распределённых случайных величин Бернулли

$$X = [0, 1, 1, 0, 1, 0, 1],$$
$$X_i \sim \mathrm{Bern}(p).$$

Проверьте гипотезу $H_0 : p = 0.3$ против двусторонней альтернативы на уровне значимости 5%.

6. Рассмотрим выборку независимых одинаково распределённых бернуллевских случайных величин X_1, …, X_{500}, где $X_i \sim \mathrm{Bern}(p)$. Оказалось, что в этой выборке ровно 300 единиц и 200 нулей.

   (a) Найдите $\hat{p}$ и оценку дисперсии X_i.

   (b) Проверьте гипотезу

$$\begin{cases} H_0 : p = 0.5, \\ H_1 : \mu \neq 0.5 \end{cases}$$

   на уровне значимости 5%.

   (c) Постройте 99%-ый доверительный интервал для p.

7. Рассмотрим две выборки случайных величин X_1, …, X_{20}, где $X_i \sim \mathcal{N}(\mu_X, 1)$ и Y_1, …, Y_{20}, где $Y_i \sim \mathcal{N}(\mu_Y, 2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково распределены, а выборки независимы между собой. Оказалось, что

$$\sum_{i=1}^{20} X_i = 10, \ \sum_{i=1}^{20} X_i^2 = 350,$$
$$\sum_{i=1}^{20} Y_i = 15, \ \sum_{i=1}^{20} Y_i^2 = 400,$$

   Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

   на уровне значимости 5%.

8. Рассмотрим две выборки случайных величин X_1, …, X_{200}, где $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ и Y_1, …, Y_{200}, где $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково

распределены, а выборки независимы между собой. Оказалось, что

$$\sum_{i=1}^{200} X_i = 10,\ \sum_{i=1}^{200} X_i^2 = 350,$$
$$\sum_{i=1}^{200} Y_i = 15,\ \sum_{i=1}^{200} Y_i^2 = 400,$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 10%.

9. Рассмотрим две выборки случайных величин $X_1, \ldots, X_{25}$, где $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ и $Y_1, \ldots, Y_{25}$, где $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково распределены, а выборки независимы между собой. Также будем предполагать, что $\sigma_X = \sigma_Y = \sigma$. Оказалось, что

$$\sum_{i=1}^{25} X_i = 10,\ \sum_{i=1}^{25} X_i^2 = 350,$$
$$\sum_{i=1}^{25} Y_i = 15,\ \sum_{i=1}^{25} Y_i^2 = 400,$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 1%.

10. Рассмотрим выборку объектов из нормального распределения до и после проведения некоторого эксперимента. Выборку ДО обозначим как X, а выборку ПОСЛЕ обозначим как Y. Известно, что

$$X = [10, 15, 20, 18, 15, 20],$$
$$Y = [13, 13, 21, 22, 14, 25],$$
$$X_i \sim \mathcal{N}(\mu_X, \sigma),$$
$$Y_i \sim \mathcal{N}(\mu_Y, \sigma),$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 10%.

11. На уровне значимости 10% проверьте гипотезу о том, существует ли зависимость между продолжением образования после окончания школы и типом местности, где выпускник окончил школу. В исследовании принимали участие по 100 школ из каждого типа местности.

	Не продолжил образование	Среднеспециальное образование	Высшее образование
Местность 1	40	40	20
Местность 2	30	50	20

12. **(сложная)** Компания «ГолденАльп» тестирует два новых вкуса шоколада: с орешками и солёной карамелью. Фокус-группа разбивают на две непересекающиеся части: N_1 человек пробуют шоколад с орешками, а N_2 — с солёной карамелью. Каждый участник пробует лишь один тип шоколада и одобряет или не одобряет опробованный вкус. Пусть X_1 — число человек, одобривших шоколад с орешками, а X_2 — одобривших шоколад с солёной карамелью. Будем предполагать, что $X_1 \sim \mathrm{Bin}(N_1, p_1)$, $X_2 \sim \mathrm{Bin}(N_2, p_2)$. Руководство компании «Голден Альп» хочет узнать, есть ли основание полагать, что один вкус шоколада предпочитается другому.

    По результатам эксперимента оказалось, что $N_1 = N_2 = 500$, $X_1 = 400$, $X_2 = 390$. Сформулируйте гипотезу, которая позволит ответить на вопрос компании, и проверьте её на уровне значимости 5%.

3 Решения

Могут содержать неточности – если какое-то решение непонятно, пишите!

3.1 Часть 1

1. c
2. a) 0, b) 1, c) 0.05, d) 0
3. a) II, b) I
4. a) Нет, b) Да, c) Да

3.2 Часть 2

1. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, \sigma^2).$$

   (a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

   на уровне значимости 5%.

   (b) Постройте 90%-ый доверительный интервал для μ.

   Решение

   а) Статистика:

$$t = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}\Bigg|_{H_0} \sim t_{n-1} = t_6.$$

$$\bar{X} = 9, \hat{\sigma}^2 = 39.33 \Rightarrow \hat{\sigma} \approx 6.27.$$

   Доверительная область: $[-t_{1-\frac{\alpha}{2};n-1}, t_{1-\frac{\alpha}{2};n-1}] = [-t_{0.975;6}, t_{0.975;6}] = [-2.4469, 2.4469]$.

   Значение статистики на выборке:

$$t_{obs} = \frac{9-3}{\sqrt{\frac{39.33}{7}}} \approx 2.5313 \notin [-2.4469, 2.4469].$$

   Значит, нулевая гипотеза отвергается на уровне значимости 5%.

   б)

$$\bar{X} - t_{1-\frac{\alpha}{2};n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t_{1-\frac{\alpha}{2};n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

$$9 - 1.9432 \cdot \frac{6.27}{\sqrt{7}} \leqslant \mu \leqslant 9 + 1.9432 \cdot \frac{6.27}{\sqrt{7}}$$

   $[4.39, 13.6]$ - 90%-ый доверительный интервал для μ

2. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, 9).$$

(a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

на уровне значимости 10%.

(b) Постройте 95%-ый доверительный интервал для μ.

Решение

а) Статистика:

$$Z = \left.\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right|_{H_0} \sim \mathcal{N}(0,1)$$

$$\bar{X} = 9, \sigma = 3$$

.

Доверительная область: $[-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}] = [-1.65, 1.65]$.

Значение статистики на выборке:

$$Z_{obs} = \frac{9-3}{\frac{3}{\sqrt{7}}} \approx 5.2915 \notin [-1.65, 1.65].$$

Значит, нулевая гипотеза отвергается на уровне значимости 10%.

б)

$$\bar{X} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$9 - 1.96 \cdot \frac{3}{\sqrt{7}} \leqslant \mu \leqslant 9 + 1.96 \cdot \frac{3}{\sqrt{7}}$$

$[6.777, 11.222]$ - 90%-ый доверительный интервал для μ.

3. Рассмотрим выборку независимых одинаково распределённых случайных величин X_1, …, X_{200}, где $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Оказалось, что

$$\sum_{i=1}^{200} X_i = 20, \ \sum_{i=1}^{200} X_i^2 = 500.$$

Решение

а)

$$\bar{X} = \frac{\sum_{i=1}^{200} X_i}{200} = \frac{20}{200} = 0.1,$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{200} X_i^2 - 2\bar{X} \cdot \sum_{i=1}^{200} X_i + n \cdot (\bar{X})^2 \right) = \frac{1}{199} \left(500 - 2 \cdot 0.1 \cdot 20 + 200 \cdot 0.01\right) \approx 2.6.$$

б) Статистика:

$$t = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}\Bigg|_{H_0} \sim t_{n-1} = t_{199} \approx \mathcal{N}(0,1).$$

Доверительная область: $[-t_{1-\frac{\alpha}{2};n-1}, t_{1-\frac{\alpha}{2};n-1}] \approx [-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}] = [-2.57, 2.57]$.

Значение статистики на выборке:

$$t_{obs} = \frac{0.1 - 2}{\sqrt{\frac{2.6}{200}}} \approx -16.66 \notin [-2.57, 2.57].$$

Значит, нулевая гипотеза отвергается на уровне значимости 1%.

в) ДИ для м. о. при неизвестной дисперсии :

$$\mathbb{P}\left(\bar{X} - t_{1-\frac{\alpha}{2};n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t_{1-\frac{\alpha}{2};n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

Его реализация на выборке:

$$0.1 - 1.96 \cdot \sqrt{\frac{2.6}{200}} \leqslant \mu \leqslant 0.1 + 1.96 \cdot \sqrt{\frac{2.6}{200}} \Rightarrow$$

$$\Rightarrow -0.123 \leqslant \mu \leqslant 0.323.$$

t_{obs} не лежит в этом доверительном интервале.

г)

$$\text{p-value} \approx 2 \int_{-\infty}^{-16.66} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \approx 2.56 \cdot 10^{-62} \approx 0.$$

p-value $< \alpha_1 = 0.01$ и p-value $< \alpha_2 = 0.05 \Rightarrow$ нулевая гипотеза отвергается на уровне значимости 1% и 5% (на любом разумном уровне значимости).

4. Так как выборка маленькая, а её распределение не нормальное, то гипотезу нельзя проверить известными нам способами.

5. Так как выборка маленькая, нельзя проверить гипотезу о доле.

6. a) p = 0,6; D(x) = 0,24 b) H0 отвергается c) (0,543475 ; 0,656525)

7. $X_1, \ldots, X_{20}, X_i \sim \mathcal{N}(\mu_X, 1)$ и $Y_1, \ldots, Y_{20}, Y_i \sim \mathcal{N}(\mu_Y, 2)$

$$\sum_{i=1}^{20} X_i = 10, \ \sum_{i=1}^{20} X_i^2 = 350,$$
$$\sum_{i=1}^{20} Y_i = 15, \ \sum_{i=1}^{20} Y_i^2 = 400,$$

Решение

$H_0 : \mu_X = \mu_Y$

$H_1 : \mu_X \neq \mu_Y$
$\alpha = 0.05$
Статистика

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

$$T\big|_{H_0} \sim \mathcal{N}(0,1)$$

Доверительная область $[-z_{1-\alpha/2}, z_{1-\alpha/2}] = [-1.96, 1.96]$
Значение статистики на выборке

$$\widetilde{T} = \frac{10/20 - 15/20}{\sqrt{\frac{1}{20} + \frac{2}{20}}} \approx -0.645$$

Наблюдаемое на выборке значение статистики попало в доверительную область, значит, у нас нет оснований отвергнуть нулевую гипотезу в пользу альтернативной гипотезы о неравенстве средних значений на уровне значимости 5%

9. $X_1, \ldots, X_{25}, X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ и $Y_1, \ldots, Y_{25}, Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$

$$\sum_{i=1}^{25} X_i = 10, \ \sum_{i=1}^{25} X_i^2 = 350,$$
$$\sum_{i=1}^{25} Y_i = 15, \ \sum_{i=1}^{25} Y_i^2 = 400,$$

Решение
$H_0 : \mu_X = \mu_Y$
$H_1 : \mu_X \neq \mu_Y$
$\alpha = 0.01$
Статистика

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{s_{XY}^2}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \ \ s_{XY}^2 = \frac{n_X s_X^2 + n_Y s_Y^2}{n_X + n_Y - 2}$$

$$T\big|_{H_0} \sim t(n_X + n_Y - 2)$$

Доверительная область $[-t_{1-\alpha/2}, t_{1-\alpha/2}] = [-2.6822, 2.6822]$
Значение статистики на выборке

$$\widetilde{T} = \frac{10/25 - 15/25}{\sqrt{15.354167}\sqrt{\frac{1}{25} + \frac{1}{25}}} \approx -0.1805$$

Наблюдаемое на выборке значение статистики попало в доверительную область, значит, у нас нет оснований отвергнуть нулевую гипотезу в пользу альтернативной гипотезы о неравенстве средних значений на уровне значимости 5%

10. Рассмотрим выборку объектов из нормального распределения до и после проведения некоторого экспе-

римента. Выборку ДО обозначим как X, а выборку ПОСЛЕ обозначим как Y. Известно, что

$$\begin{aligned} X &= [10, 15, 20, 18, 15, 20], \\ Y &= [13, 13, 21, 22, 14, 25], \\ X_i &\sim \mathcal{N}(\mu_X, \sigma), \\ Y_i &\sim \mathcal{N}(\mu_Y, \sigma), \end{aligned}$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 10%.

Решение

$$\bar{X} = 16.33, \bar{Y} = 18, \hat{\sigma}_X^2 = 14.66, \hat{\sigma}_Y^2 = 28, \hat{\sigma} = \sqrt{\frac{(n_X - 1) \cdot \hat{\sigma}_X^2 + (n_Y - 1) \cdot \hat{\sigma}_Y^2}{n_X + n_Y - 2}} \approx 4.619.$$

Статистика:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\hat{\sigma}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}\Bigg|_{H_0} \sim t_{n_X + n_Y - 2} = t_{10}$$

Доверительная область: $[-t_{0.95;10}, t_{0.95;10}] = [-1.8125, 1.8125]$.

Значение статистики на выборке:

$$t_{obs} = \frac{16.33 - 18 - 0}{4.619 \cdot \sqrt{\frac{1}{3}}} \approx -0.625 \in [-1.8125, 1.8125]$$

.

Наблюдаемое значение статистики попало в доверительную область, значит, нет оснований отвергать нулевую гипотезу в пользу альтернативной гипотезы о неравенстве средних значений на уровне значимости 10%.

11.

$$\chi^2_{obs} = 2.54, \text{p-val} = 0.28 \Rightarrow H_0 \text{ не отвергается на уровне значимости } 10\%.$$

12. Идея: используя интуицию биномиального распределения, найдите $\hat{p}_1$ и $\hat{p}_2$ и проверьте гипотезу о равенстве p_1 и p_2.