Контрольна Робота з Математичної Статистики #1

Захаров Дмитро

26 жовтня, 2024

Варіант 5

Зміст

1 Вправа 1. Візуалізація вибіркових даних

Умова Задачі 1.1. Отримані наступні вибіркові дані про час безвідмовної роботи бурових штанг (у хвилинах): 280; 188; 190; 220; 288; 190; 190; 190; 280; 280; 190; 190; 300. Побудувати вибіркову функцію розподілу, гістограму вибірки та полігон частот. Знайти вибіркове середнє, вибіркову дисперсію і незміщену оцінку дисперсії.

Розв'язання. Отже, спочатку побудуємо таблицю частот:

Значення t_i	**Частота** ν_i
188	1
190	6
220	1
280	3
288	1
300	1

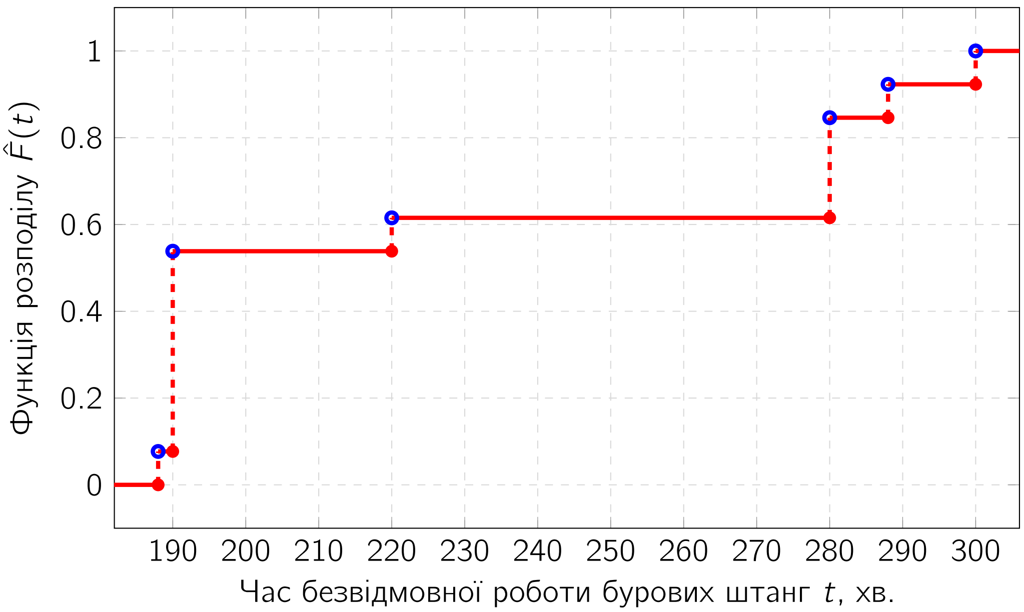
Наближено, ми вважаємо, що якщо T — наша випадкова величина, що дорівнює часу безвідмовної роботи бурових штанг, то $\Pr[T = t_i] = \nu_i / \sum_{j=1}^{n} \nu_j$ для вибірки $\{(t_i, \nu_i)\}_{1 \leq i \leq n}$. В нашому конкретно випадку, маємо наступний розподіл:

$$
\begin{aligned}
&\Pr[T = 188] = \Pr[T = 220] = \Pr[T = 288] = \Pr[T = 300] = \frac{1}{13} \\
&\Pr[T = 190] = \frac{6}{13}, \quad \Pr[T = 280] = \frac{3}{13}
\end{aligned}
$$

Отже, **вибіркова функція розподілу** виглядає наступним чином:

$$
\hat{F}_T(t) = \begin{cases} 0, & t \leq 188, \\ \frac{1}{13}, & 188 < t \leq 190, \\ \frac{7}{13}, & 190 < t \leq 220, \\ \frac{8}{13}, & 220 < t \leq 280, \\ \frac{11}{13}, & 280 < t \leq 288, \\ \frac{12}{13}, & 288 < t \leq 300, \\ 1, & t > 300. \end{cases}
$$

Отже, зобразимо графік вибіркової функції розподілу:



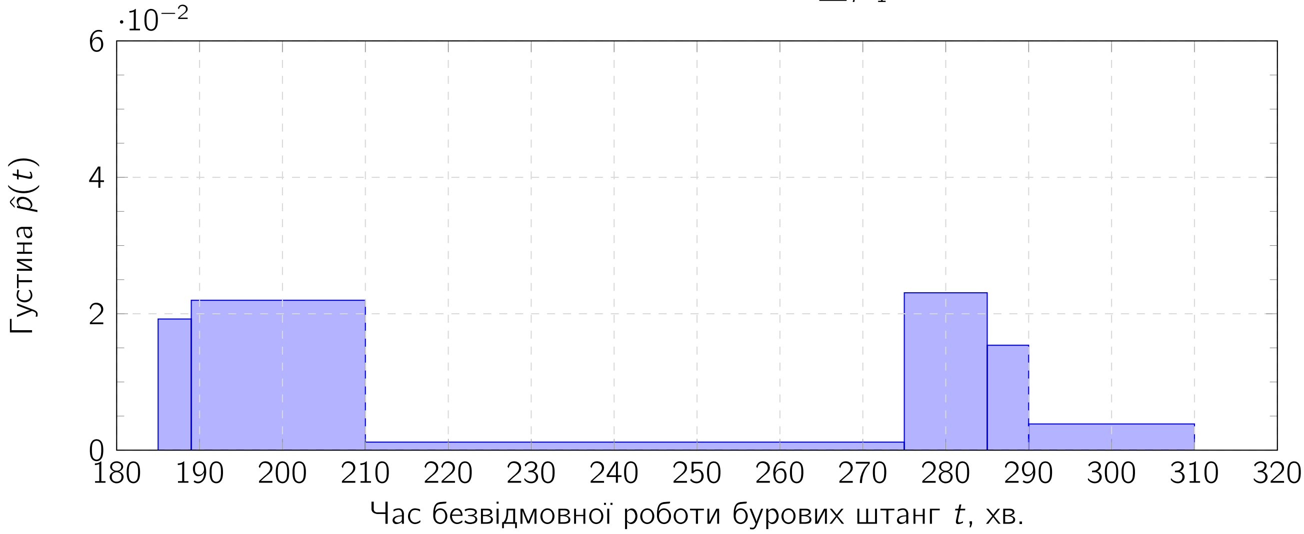
Тепер побудуємо **гістограму вибірки**. Для цього візьмемо наступні вузли:

$$a_1 = 185,\ a_2 = 189,\ a_3 = 210,\ a_4 = 275,\ a_5 = 285,\ a_6 = 290,\ a_7 = 310.$$

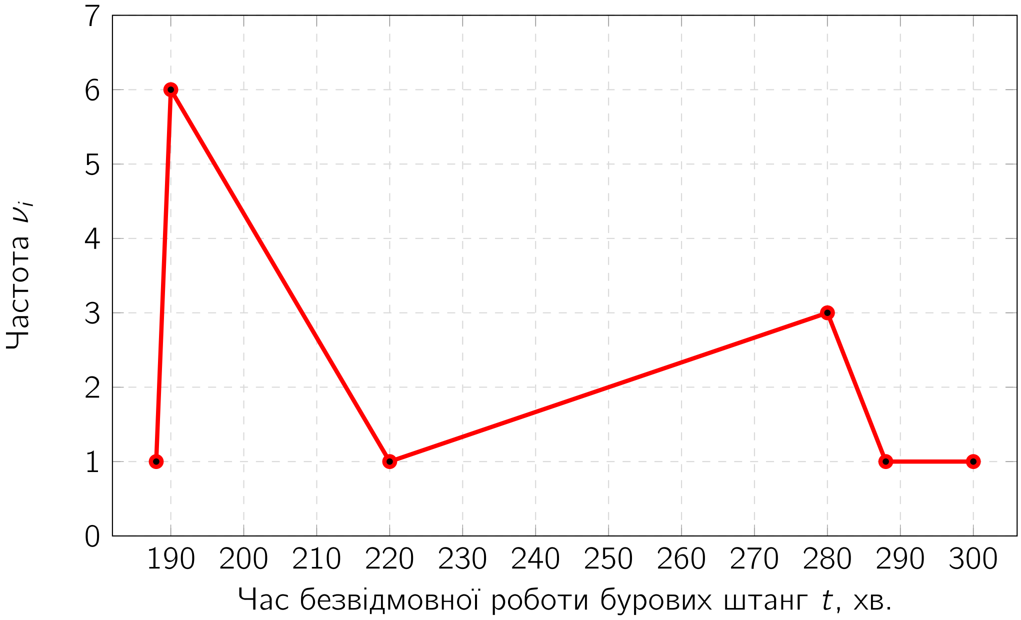
Далі будуємо гістограму наступним чином: зображуємо прямокутники, причому прямокутник i розташований між a_i та a_{i+1} вузлом, а його висота дорівнює $h_i = \nu_i/n\ell_i$ для $\ell_i = a_{i+1} - a_i$. Тут, ν_i – кількість данних на відрізку (a_i, a_{i+1}). Отже, маємо:

$$h_1 \approx 0.0192,\ h_2 \approx 0.0220,\ h_3 \approx 0.0012,$$
$$h_4 \approx 0.02308,\ h_5 \approx 0.01538,\ h_6 \approx 0.0038.$$

Можна переконатись, що при цьому площа під графіком $\sum_{i=1}^{n} h_i\ell_i = 1$.



Нарешті, **полігон частот** це просто лінія, що з'єднує $\{(t_i, \nu_i)\}_{1\leq i\leq n}$, тому маємо



Отже, тепер можемо порахувати вибіркове середнє. Воно обчислюється за формулою:

$$\overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i\nu_i, \;\; n := \sum_{j=1}^{n}\nu_j.$$

В нашому випадку,

$$\overline{t} = \frac{1}{13}\cdot(188\times 1 + 190\times 6 + 220\times 1 + 280\times 3 + 288\times 1 + 300\times 1) \approx 228.9$$

Тепер, вибіркова дисперсія обчислюється за формулою:

$$\overline{\sigma}_T^2 = \frac{1}{n}\sum_{i=1}^{n} t_i^2\nu_i - \overline{t}^2.$$

В нашому випадку,

$$\overline{\sigma}_T^2 = \frac{1}{13}\cdot(188^2\times 1 + 190^2\times 6 + 220^2\times 1 + 280^2\times 3 + 288^2\times 1 + 300^2\times 1) - 228.9^2 \approx 2100$$

У свою чергу, незміщена оцінка дисперсії обчислюється за формулою:

$$\hat{\sigma}_T^2 = \frac{n}{n-1}\overline{\sigma}_T^2 = \frac{13}{12}\cdot 2100 \approx 2280$$

Відповідь. Вибіркова функція розподілу, гістограма вибірки та полігон частот зображені у розв'язку. Вибіркове середнє $\overline{t} \approx 228.9$, вибіркова дисперсія $\overline{\sigma}_T^2 \approx 2100$, і незміщена оцінка дисперсії $\hat{\sigma}_T^2 \approx 2280$.

2 Вправа 2. Довірчий інтервал для математичного сподівання

Умова Задачі 2.1. В таблиці $\mathcal{D}$, яка приведена нижче, вказана кількість лампочок, час горіння (в тис. годин) яких потрапило у відповідний проміжок. Для довірчої ймовірності $\alpha = 0.96$ та середнього квадратичного відхилення $\sigma = 0.01$ часу горіння побудувати довірчий інтервал для середнього часу горіння лампочки.

Час горіння	$[2.1, 2.2)$	$[2.2, 2.3)$	$[2.3, 2.4)$	$[2.4, 2.5)$	$[2.5, 2.6)$	$[2.6, 2.7)$
Число лампочок	2	8	22	40	12	10

Розв'язання. Будемо вважати, що наближено кількість лампочок ξ роподілена нормально, себто $\xi \sim \mathcal{N}(\mu, \sigma^2)$. Математичне сподівання μ ми маємо оцінити, а $\sigma = 0.01$ дано. Нам треба підібрати інтервал $\mathcal{I}_\alpha = (\ell(\mathcal{D}), u(\mathcal{D}))$ так, щоб $\Pr[\mu \in \mathcal{I}_\alpha] = \alpha$.

Скористаємось наступною теоремою.

Theorem 2.2. Про довірчий інтеграл математичного сподівання нормального закону. Нехай $\xi \sim \mathcal{N}(\mu, \sigma^2)$ і ми маємо вибірку $\mathcal{D} = (x_1, \dots, x_n) \sim \xi$. Тоді, довірчий інтервал $\mathcal{I}_\alpha$ для математичного сподівання можна покласти як:

$$\mathcal{I}_\alpha = \left(\overline{\mu} - \frac{z_\alpha \sigma}{\sqrt{n}}, \overline{\mu} + \frac{z_\alpha \sigma}{\sqrt{n}}\right), \quad z_\alpha := \Phi_0^{-1}\left(\frac{\alpha}{2}\right).$$

Себто, для цього інтервала виконується умова $\Pr[\mu \in \mathcal{I}_\alpha] = \alpha$.

Отже, нам залишилось порахувати всі значення. Маємо груповані дані, тому набір $\mathcal{D}$ можна розглядати як вибірку з $n = 94$ елементів, де

$$\mathcal{D} = \{\underbrace{2.15, \dots, 2.15}_{2\text{ рази}}, \underbrace{2.25, \dots, 2.25}_{8\text{ разів}}, \underbrace{2.35, \dots, 2.35}_{22\text{ рази}}, \underbrace{2.45, \dots, 2.45}_{40\text{ разів}}, \dots\}.$$

В цьому випадку вибіркове середнє $\overline{\mu} = \frac{1}{n}\sum_{i=1}^m x_i \approx 2.4372$. Згідно таблиці, маємо величину $z_\alpha = \Phi_0^{-1}(0.48) \approx 2.055$, а тому

$$\Pr\left[2.4372 - \frac{2.055 \times 0.01}{\sqrt{94}} < \mu < 2.4372 + \frac{2.055 \times 0.01}{\sqrt{94}}\right] = \alpha.$$

Величина $2.055 \times 0.01/\sqrt{94} \approx 0.0021$. А тому наш інтервал приблизно:

$$\mathcal{I}_{0.96} = (2.4372 - 0.0021, 2.4372 + 0.0021).$$

Відповідь. $\mathcal{I}_{0.96} \approx (2.43512, 2.43935)$.

Додаток (програма). Для того, щоб переконатись, що в нас дійсно все працює, напишемо програму на мові *Python*, що реалізує цей алгоритм. Наводимо його нижче.

```python
# Importing necessary libraries
import numpy as np
from scipy.special import erf, erfinv
```

```python

def laplace_function(x: float) -> float:
    """
    Given a float x, this function returns the value of the Laplace function at x, which
    is an integral of exp(-t^2/2)/sqrt(2\pi) from 0 to x.
    """

    return erf(x/np.sqrt(2))/2.0

def inverse_laplace_function(x: float) -> float:
    """
    Given a float x, this function returns the inverse of the Laplace function at x
    """
    r = 1
    STEPS_FOR_FINDING_INVERSE = 10000

    for _ in range(STEPS_FOR_FINDING_INVERSE):
        r = r * x / laplace_function(r)

    return r

def get_mean_credible_interval(dataset: np.ndarray,
                               variance: float,
                               credibility_prob: float) -> [float, float]:
    """
    Given a dataset, consisting of an array of floats, and variance
    this function returns the credible interval for the mean of the dataset.
    """

    mu = np.mean(dataset) # Getting the mean
    n = len(dataset) # Getting the number of data points
    z_alpha = inverse_laplace_function(credibility_prob/2) # Getting the z_alpha value

    return (mu - z_alpha * np.sqrt(variance/n), mu + z_alpha * np.sqrt(variance/n))

dataset = [2.15]*2+[2.25]*8+[2.35]*22+[2.45]*40+[2.55]*12+[2.65]*10
print(get_mean_credible_interval(dataset, 0.01**2, 0.96))
```

Вихід з цієї програми: (2.4351157622922854, 2.439352322814098).

3 Вправа 3. Довірчий інтервал для дисперсії

Умова Задачі 3.1. У деяких містах України отримані наступні дані про вартість споживчого кошика (в тис. грн) 196; 208; 196; 208; 208; 222; 216; 227; 222; 216; 222; 216; 216; 222; 227; 240; 240; 240; 240; 240; 240; 240; 227; 227; 227. Для довірчої ймовірності $\alpha = 0.9$ побудувати довірчий інтервал для дисперсії вартості споживчого кошика міст України.

Розв'язання. Як і в минулому прикладі, будемо вважати, що вартість споживчого кошика (в тис. грн) ξ роподілена нормально, себто $\xi \sim \mathcal{N}(\mu, \sigma^2)$. У цій вибірці ми не знаємо ані математичного сподівання, ані дисперсії. Нам треба підібрати інтервал $\mathcal{I}_\alpha = (\ell(\mathcal{D}), u(\mathcal{D}))$ так, щоб $\Pr[\sigma^2 \in \mathcal{I}_\alpha] = \alpha$.

Скористаємось наступною теоремою.

Theorem 3.2. Про довірчий інтеграл дисперсії нормального закону. Нехай $\xi \sim \mathcal{N}(\mu, \sigma^2)$ і ми маємо вибірку $\mathcal{D} = (x_1, \dots, x_n) \sim \xi$. Тоді, якщо позначити $q := 1 - \alpha$, довірчий інтервал $\mathcal{I}_\alpha$ для дисперсії можна покласти як:

$$\mathcal{I}_\alpha = \left(\frac{n\overline{\sigma}_X^2}{\beta}, \frac{n\overline{\sigma}_X^2}{\gamma} \right), \quad \beta = \chi^2_{n-1,q/2}, \ \gamma = \chi^2_{n-1,1-q/2}$$

Себто, для цього інтервала виконується умова $\Pr[\sigma^2 \in \mathcal{I}_\alpha] = \alpha$.

Отже, залишається лише порахувати всі значення. Маємо вибірку з $n = 25$ елементів. Середнє значення $\overline{\mu}_X = \frac{1}{n}\sum_{i=1}^n x_i = 223.32$, а отже вибіркова дисперсія

$$\overline{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \overline{\mu}_X^2 \approx 177.34.$$

Знайдемо коефіцієнти α та β. Маємо $q = 1 - \alpha = 0.1$. Отже, згідно таблиці знаходимо

$$\chi^2_{24,0.05} \approx 36.42, \quad \chi^2_{24,0.95} \approx 13.85.$$

Таким чином, наш інтервал:

$$\mathcal{I}_{0.9} = \left(\frac{25 \times 177.34}{36.42}, \frac{25 \times 177.34}{13.85} \right) \approx (121.73, 320.11).$$

Відповідь. $\mathcal{I}_{0.9} = (121.73, 320.11)$.

Додаток (програма). Для того, щоб переконатись, що в нас дійсно все працює, напишемо програму на мові *Python*, що реалізує цей алгоритм. Наводимо його нижче.

```python
# Importing necessary libraries
import numpy as np
from scipy.stats import chi2

def get_variance_credible_interval(dataset: np.ndarray,
                                   alpha: float) -> [float,
    float]:
```

```
    """
    Given a dataset, consisting of an array of floats,
    this function returns the credible interval for the variance of the dataset.
    """

    mu = np.mean(dataset) # Getting the mean
    variance = np.mean(dataset**2) - mu**2 # Getting the variance
    n = len(dataset) # Getting the number of data points

    q = 1 - alpha # Getting the q value
    beta = chi2.isf(q/2, n-1)
    gamma = chi2.isf(1-q/2, n-1)

    return (n*variance/beta, n*variance/gamma)

dataset = np.array([
    196, 208, 196, 208, 208, 222, 216, 227, 222, 216, 222, 216,
    216, 222, 227, 240, 240, 240, 240, 240, 240, 240, 227, 227, 227
])

print(get_variance_credible_interval(dataset, 0.9))
```

Вихід з цієї програми: (121.74753617946827, 320.14037634618154).

4 Вправа 4. Перевірка гіпотези про математичне сподівання

Умова Задачі 4.1. Номінальний опір для резисторів, які виготовляються, складає 2000Ω. Для контролю відібрана партія з резисторів. В результаті вимірювання опору кожного зразка з точністю до 5Ω отримані наступні значення: 2130; 2090; 2030; 2080; 1920; 2020; 2015; 2000; 2045; 1940; 1980; 1970. Чи можна відхилення від номіналу (2000Ω) розглядати як випадкові (допустимі) або, навпаки, результати вказують на те, що опір резисторів відрізняється від номіналу?

Розв'язання. Введемо довірчу ймовірність $\alpha = 0.95$. Будемо вважати, що точність вимірювання — це середньоквадратичне відхилення $\sigma = 5\Omega$. Тоді, введемо дві гіпотези:

- $\mathcal{H}_0$: математичне сподівання $\mu = \mu_0$, де $\mu_0 = 2000\Omega$.
- $\mathcal{H}_1$: обернене твердження, тобто $\mu = \mu_0$, де $\mu_0 \neq 2000\Omega$.

Нехай R — випадкова величина, що описує опір резистора. Тоді, будемо вважати, що $R \sim \mathcal{N}(\mu_R, \sigma^2)$, де μ_R нам невідоме, а $\mathcal{D} = (r_1, \dots, r_n)$ — наша вибірка. Отже, знайдемо вибіркове середнє значення нашої вибірки:

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \approx 2018.33.$$

Згідно теорії, випадкова величина $y = \sqrt{n}(\overline{r} - \mu)/\sigma \sim \mathcal{N}(0, 1)$. Отже, якщо гіпотеза $\mathcal{H}_0$ вірна, то замість μ підставляємо $\mu_0 = 2000\Omega$. Тоді $y = \sqrt{12}(2018.33 - 2000)/5 \approx 12.7$. Як було сказано, довірча ймовірність $\alpha = 0.95$. Для неї добре відомий квантиль $z_\alpha := \Phi_0^{-1}(\alpha/2) = 1.96$.

Бачимо, що $12.7 > 1.96$, а отже ми відхиляємо гіпотезу $\mathcal{H}_0$ на користь $\mathcal{H}_1$. Отже, можна вважати, що опір резисторів відрізняється від номіналу. Це також достатньо добре видно і без додаткового підрахунку: як відомо, більше 95% випадкових величин з нормального розподілу лежать в межах 2σ від математичного сподівання. Отже, якщо вибіркове середнє відрізняється від номіналу більше, ніж на 2σ, то це вже досить серйозний відхил від номіналу. Також, можна це підтвердити програмою нижче:

```python
# Importing necessary libraries
import numpy as np
from scipy.special import erf, erfinv

def laplace_function(x: float) -> float:
    """
    Given a float x, this function returns the value of the Laplace function at x, which
    is an integral of exp(-t^2/2)/sqrt(2\pi) from 0 to x.
    """

    return erf(x/np.sqrt(2))/2.0

def inverse_laplace_function(x: float) -> float:
```

```
    """
    Given a float x, this function returns the inverse of the Laplace function at x
    """
    r = 1
    STEPS_FOR_FINDING_INVERSE = 10000

    for _ in range(STEPS_FOR_FINDING_INVERSE):
        r = r * x / laplace_function(r)

    return r

def test_hypothesis(dataset: np.ndarray,
                    std: float,
                    expected_mean: float,
                    alpha: float = 0.95) -> bool:
    """
    Given a dataset, consisting of an array of floats, and standard deviation,
    tests whether the expected mean is true or not.
    """

    mu = np.mean(dataset) # Getting the mean
    n = len(dataset) # Getting the number of data points
    normed_diff = (mu - expected_mean) / (std/np.sqrt(n)) # Getting the normalized difference
    z_alpha = inverse_laplace_function(alpha/2) # Getting the z_alpha value

    print(f"Got normalized difference: {normed_diff}, z_alpha is: {z_alpha}")
    return np.abs(normed_diff) < z_alpha

dataset = [2130, 2090, 2030, 2080, 1920, 2020, 2015, 2000, 2045, 1940, 1980, 1970]
result = test_hypothesis(dataset, 5, 2000, 0.95)
print("Hypothesis is true" if result else "Hypothesis is false")
```

Вихід з цієї програми: `Got normalized difference: 12.701705922171714, z_alpha is: 1.9599639845400536, Hypothesis is false`.

Відповідь. Опір резисторів відрізняється від номіналу.

5 Вправа 5

Умова Задачі 5.1. На одній з ділянок розсипного родовища золота досліджували можливість зниження витрат на розвідку. При цьому замість частини запланованих шурфів були пробурені свердловини ударно-канатного буріння (витрати на буріння свердловин менші). Результати опробування шурфів на вміст золота (в кг/м^3): 431; 397; 462; 457; 251; 221; 548; 478; 299; 541, свердловини: 322; 250; 225; 315; 399; 348; 192; 375; 381; 538; 198; 317; 293. Чи можна вважати, що в середньому результати опробування свердловин на наявність золота несуттєво відрізняються від результатів опробування шурфів?

Розв'язання. Отже, маємо вибірку вміст золота $x_1, \dots, x_n$ з генеральної сукупності шурфів X, а також вибірку вмісту $y_1, \dots, y_m$ з сукупності свердловини Y. Цілком логічно вважати X та Y незалежними та нормально розподіленими. Тобто нехай $X \sim \mathcal{N}(\mu_X, \sigma_X)$ та $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$. В нашому випадку ми не знаємо ані жодне з математичних сподівань μ_X та μ_Y, ані жодну з дисперсій σ_X^2 та σ_Y^2. Вводимо дві гіпотези:

- $\mathcal{H}_0$: $\mu_X = \mu_Y$ (основна гіпотеза).
- $\mathcal{H}_1$: $\mu_X \neq \mu_Y$ (альтернативна гіпотеза).

Перевірити гіпотезу $\mathcal{H}_0$ ми можемо, припускаючи $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Введемо випадкову величину $\xi := \overline{x} - \overline{y}$, де

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 408.5, \quad \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \approx 320.$$

Тоді, як було доведено на лекції,

$$\xi \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2\right)$$

Тому, цілком природньо ввести стандартно нормально розподілену випадкову величину:

$$\eta := \frac{\overline{x} - \overline{y} - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n}\cdot\sigma_X^2 + \frac{1}{m}\cdot\sigma_Y^2}} \sim \mathcal{N}(0,1).$$

Якщо б ми знали σ_X^2, σ_Y^2, то ми б могли одразу ввести довірчу ймовірність та дивитись, в яку частину нормального розподілу потрапляє вираз $(\overline{x} - \overline{y})/\sqrt{\frac{1}{n}\cdot\sigma_X^2 + \frac{1}{m}\cdot\sigma_Y^2}$. Тут ми так зробити не можемо, тому продовжимо далі.

Як було сказано, припускаємо, що $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Тоді, маємо випадкову величину:

$$\eta = \frac{\overline{x} - \overline{y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0,1).$$

Отже, розглянемо випадкові дисперсії:

$$\overline{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \approx 11965, \quad \overline{\sigma}_Y^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2 - \overline{y}^2 \approx 8250.$$

За теоремою про розподіл вибіркової дисперсії, $n\overline{\sigma}_X^2/\sigma^2 \sim \chi^2_{n-1}$, $m\overline{\sigma}_Y^2/\sigma^2 \sim \chi^2_{m-1}$. Враховуючи незалежність цих величин і теорему про стійкість розподілу χ^2, маємо

$$\zeta = \frac{n\overline{\sigma}_X^2}{\sigma^2} + \frac{m\overline{\sigma}_Y^2}{\sigma^2} \sim \chi^2_{m+n-2}.$$

Згідно лекції, величини η та ζ є незалежними. Тому утворимо наступну статистику:

$$\tau = \frac{\eta}{\sqrt{\zeta/(n+m-2)}} = \frac{\overline{x}-\overline{y}-(\mu_X-\mu_Y)}{\sqrt{n\overline{\sigma}_X^2 + m\overline{\sigma}_Y^2}}\sqrt{\frac{nm(m+n-2)}{m+n}} \sim \mathcal{ST}_{n+m-2}$$

Якщо ж основна гіпотеза $\mathcal{H}_0$ правильна, то

$$\tau = \frac{\overline{x}-\overline{y}}{\sqrt{n\overline{\sigma}_X^2 + m\overline{\sigma}_Y^2}}\sqrt{\frac{nm(m+n-2)}{m+n}} \sim \mathcal{ST}_{n+m-2}$$

Підставимо конкретні значення. Маємо:

$$\tau = \frac{408.5-320}{\sqrt{10\times 11965 + 13\times 8250}}\sqrt{\frac{10\times 13\times 21}{23}} \approx 2.024.$$

Оберемо довірчу ймовірність $\alpha = 0.95$, а отже відповідний рівень значущості $q := 1-\alpha = 0.05$. За таблицею розподілу Стьюдента маємо $t_{n+m-2,q} = t_{21,0.05} \approx 2.08$. Оскільки $|\tau| < t_{n+m-2,q}$, то вважаємо гіпотезу $\mathcal{H}_0$ правильною. Отже, можна вважати, що в середньому результати опробування свердловин на наявність золота несуттєво відрізняються від результатів опробування шурфів.

Відповідь. В середньому результати опробування свердловин на наявність золота несуттєво відрізняються від результатів опробування шурфів.