Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Фізико-технічний інститут

Статистичний аналіз результатів ЗНО 2019

предмет «Математична статистика»

2022

Зміст

Програмний етап

Для статистичного аналізу було обрано відкриті дані із загальнодоступного освітнього сайту https://osvita.ua/news/data/. З-поміж інших даних, у переліку доступні «Деперсоніфіковані дані учасників ЗНО 2019 року з кожного навчального предмета». Надалі саме ця інформація буде розглянута.

Ініціалізація даних

Зі сторінки сайту можна завантажити архів даних `opendatazno2019.zip`. У розархівованій папці наявні два файли з різними іменами: `opendatazno2019.xlsx` та `opendatazno2019_info.xls`. У першому з них містяться власне дані, а у другому – опис і роз'яснення назв стовпців таблиці даних.

Файл `opendatazno2019.xlsx` має великий обсяг – понад 230 MB, тому жоден онлайн редактор не буде в змозі його відкрити. Рівно як і програмне забезпечення `Microsoft Excel`, `Google Sheets` чи `LibreOffice Calc`. Тож для подальшого опрацювання вхідних даних буде використано засоби мови `Python`.

Для читання файлів розширення `.xlsx` можна використати бібліотеку `xlrd` версії `1.2.0` й далі працювати безпосередно із рядками таблиці, пробігаючи кожну комірку так, як це показано у рядку 13 Лістингу 1:

Лістинг 1: Використання бібліотеки xlrd

```
1  import xlrd
2
3  # open the Workbook
4  workbook = xlrd.open_workbook("opendatazno2019.xlsx")
5
6  # open the worksheet
7  worksheet = workbook.sheet_by_index(0)
8
9  # iterate the rows and columns
10 for i in range(0, 5):
11     for j in range(0, 3):
12         # print the cell values with a tab space
13         print(worksheet.cell_value(i, j), end="\t")
14     print("")
```

Проте, у такому разі обробка файлу `opendatazno2019.xlsx` триватиме близько 5 хвилин, тому такий спосіб опрацювання великого обсягу даних є неефективним. Натомість, користуючись тією ж бібліотекою `xlrd` у додачу до засобів бібліотек `pandas` та `csv`, можна зчитати й порядково перевторити файл `.xlsx` у файл `.csv`, як це наведено на Лістингу 2. Надалі це значно зменшить тривалість виконання обробки даних. Більше про різні способи зчитування й обробки файлів розширення `.xlsx` можна довідатися за цим посиланням.

Лістинг 2: Конвертація у .csv файл

```
import pandas as pd
import xlrd
import csv

# open workbook by sheet index, optional - sheet_by_index()
sheet = xlrd.open_workbook("opendatazno2019.xlsx").sheet_by_index(0)

# writer object is created
column = csv.writer(open("opendatazno2019.csv", "w", newline=""))

# write the data into csv file
for row in range(sheet.nrows):
    # row by row write operation
    column.writerow(sheet.row_values(row))

# read csv file and convert into a dataframe object
df = pd.DataFrame(pd.read_csv("opendatazno2019.csv", dtype="unicode"))
```

Як це зображено на Рис. 1, початкові дані мають рядки невідповідного формату, тобто ці дані є «брудними». На Лістингу 3 коротко вказані команди, за допомогою яких можна прибрати нульові значення чи комірки з невідповідним форматом. Як результат – матимемо готові «чисті» дані для подальшої обробки.

Whole initial dataframe:

	Birth	SEXTYPENAME	UkrBall100	mathBall100	engBall100
0	2001	жіноча	100.0	NaN	NaN
1	1985	жіноча	NaN	NaN	NaN
2	2001	жіноча	166.0	NaN	NaN
3	2000	чоловіча	127.0	0.0	NaN
4	2001	жіноча	171.0	NaN	116.0
5	2001	чоловіча	0.0	NaN	NaN
6	1999	чоловіча	0.0	NaN	NaN
...	...	...	...	...	...
353806	2001	чоловіча	107.0	NaN	NaN
353807	2002	жіноча	127.0	NaN	NaN
353808	2001	жіноча	197.5	180.0	173.0
353809	2001	жіноча	122.0	100.0	NaN
353810	2001	чоловіча	134.0	140.0	NaN
353811	2002	жіноча	131.0	NaN	NaN
353812	2000	жіноча	NaN	NaN	NaN

Рис. 1: Початкові дані

Лістинг 3: Чистка даних

```
df = pd.DataFrame(pd.read_csv("opendatazno2019.csv", dtype="unicode"))

# convert string to float
df["UkrBall100"] = df["UkrBall100"].astype("float")

# remove all rows with NULL values:
cleaned_df = df.dropna(subset=["UkrBall100"])

# resetting indexes after removing rows from dataframe
cleaned_df.reset_index(drop=True, inplace=True)

# remove all rows with "0.0" values:
cleaned_df = cleaned_df.loc[cleaned_df["UkrBall100"] != 0.0]
cleaned_df.reset_index(drop=True, inplace=True)
```

Реалізація рандомізованого формування елементів вибірок

Важливим етапом статистичного аналізу є реалізація випадкового, рандомізованого формування вибірок із усього наявного масиву даних. Програмно таку реалізацію наведено на Лістингу нижче.

Лістинг 4: Рандомізоване формування вибірок

```
import random

sample_size = 500

# create a list of all possible indexes
index = [i for i in range(0,cleaned_df.last_valid_index()+1)]

# create an empty dictionary
random_elements = {}

# add one column to a dictionary
elements = []
random_elements.update({"UkrBall100": elements})

for j in range(sample_size):
    random_index = random.choice(index)
    elements.append(cleaned_df.loc[random_index, "UkrBall100"])
    index.remove(random_index)

random_selected_df = pd.DataFrame(random_elements)
```

Із усім програмним кодом, який використано в роботі, можна ознайомитися у github репозиторії.

Результати з української мови

Після програмного етапу завантаження й чистки даних маємо змогу безпосередньо оглянути отримані 286 413 результати, при цьому зазначимо, що жіночих виявилося на 19 341 більше за чоловічих. Зобразимо гістограми результатів ЗНО з української мови 2019 року для вибірки, наприклад, 10 000 учнів:

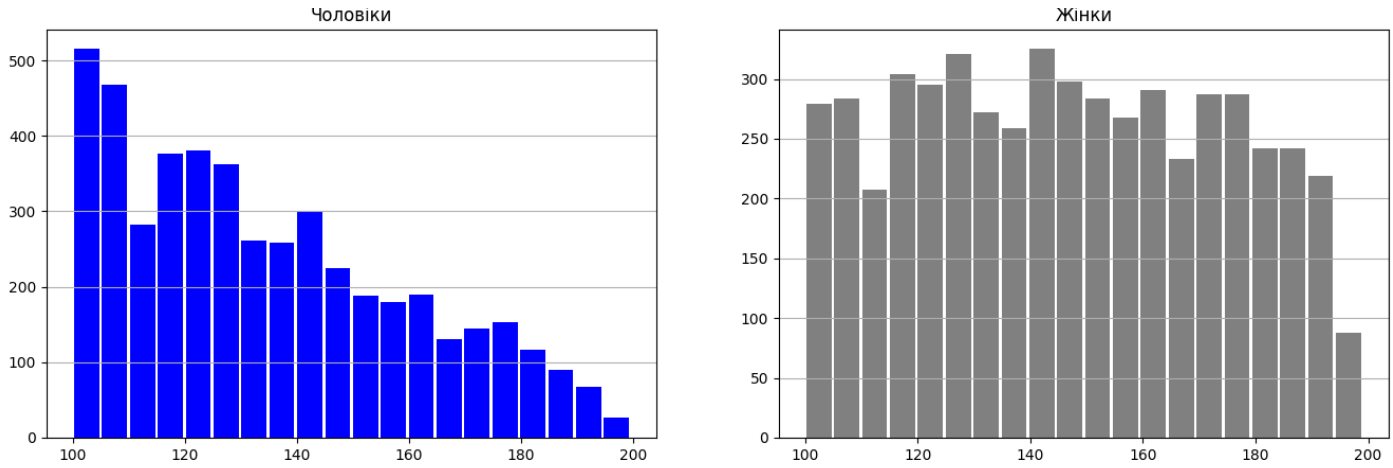

Рис. 2: Результати з української мови

На малюнку наведено розподіл балів чоловіків та жінок: неозброєним оком бачимо наявну відмінність отриманих оцінок в залежності від статі. З'ясуємо, чи ця відмінність є статистично значущою.

Перевірка гіпотези однорідності даних

Нехай маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$$\begin{aligned} X_1 \dots X_n &\sim F_x && \text{результати чоловіків} \\ Y_1 \dots Y_m &\sim F_y && \text{результати жінок} \end{aligned}$$

Перевіримо гіпотезу про однорідність статистичного матеріалу, тобто гіпотезу, що ймовірності спостереження умовно високих, помірних та низьких балів в обох вибірках є однаковими. Таким чином матимемо $k = 2$ вибірок, в яких елементи приймають $s = 3$ різних значень. Для формалізації задачі введемо позначення:

$$p_{ij} = \left\{ \begin{array}{l} \text{оцінка з чоловічої}_{(i=1)} \text{ чи жіночої}_{(i=2)} \text{ вибірки належить} \\ \text{множині високих}_{(j=1)}, \text{ помірних}_{(j=2)} \text{ чи низьких}_{(j=3)} \text{ балів} \end{array} \right\}$$

Категорізація на множини оцінок того чи іншого рівня базувалася на прохідних балах на різні факультети КПІ у 2019 році (із переліком прохідних балів можна ознайомитися за посиланням).

При цьому «високі» оцінки обиралися із міркувань, що вступнику із таким балом доступно для вступу на бюджет більше ніж 65% усіх факультетів, натомість абітурієнту із «низькими» оцінками рівень доступності значно нижчий – лише 30%. До прикладу:

$$
\begin{aligned}
A_1 &= \{\text{оцінки, вищі за 178 балів}\} \\
A_2 &= \{\text{оцінки між 144 та 178 балами}\} \\
A_3 &= \{\text{оцінки, нижчі за 144 бали}\}
\end{aligned}
$$

Остаточно гіпотеза перевірки однорідності спосережуваних даних формулюватиметься так:

$$
\begin{array}{ll}
\text{нульова гіпотеза} & H_0 : p_{11} = p_{21},\ p_{12} = p_{22},\ p_{13} = p_{23} \\
\text{проти альтернативи} & H_1 : \exists i, j \quad p_{1,j} \neq p_{2,j}
\end{array} \tag{1}
$$

А тоді критерієм перевірки гіпотези слугуватиме правило

$$
\delta(x_1 \dots x_n,\ y_1 \dots y_m) = \begin{cases} H_1, & \rho \geqslant \chi^2_{1-\alpha;\ (k-1)(s-1)} \\ H_0, & \rho < \chi^2_{1-\alpha;\ (k-1)(s-1)} \end{cases}, \tag{2}
$$

де статистика критерію ρ обчислюється так:

$$
\rho \equiv \rho(x_1 \dots x_n,\ y_1 \dots y_m) = (n+m) \left(\sum_{i=1}^{k} \sum_{j=1}^{s} \frac{\vartheta_{ij}^2}{\vartheta_{i\cdot}\vartheta_{\cdot j}} - 1 \right), \tag{3}
$$

величина критичної точки $\chi^2_{1-\alpha;\ (k-1)(s-1)}$ є квантилем рівня $(1-\alpha)$ розподілу Ст'юдента із $(k-1)(s-1)$ степенями свободи, а також

$$
\vartheta_{1j} = \sum_{i=1}^{n} \mathbb{1}(X_i \in A_j)
$$
частота потрапляння елементів вибірки результатів чоловіків в одну із j категорій

$$
\vartheta_{2j} = \sum_{i=1}^{m} \mathbb{1}(Y_i \in A_j)
$$
частота потрапляння елементів вибірки результатів жінок в одну із j категорій

$$
\vartheta_{i\cdot} = \sum_{j=1}^{s} \vartheta_{ij},\ \vartheta_{\cdot j} = \sum_{i=1}^{k} \vartheta_{ij}
$$
суми відповідних рядків чи стовпців таблиці спостережуваних даних

Таблиця спостережуваних даних

Складемо таблицю за спостережуваними даними навмання обраних $n = 500$ та $m = 500$ елементів із вибірок результатів ЗНО з української мови для чоловіків та жінок:

	Низькі бали	Помірні бали	Високі бали	Всього
Чоловіки	343	125	32	500
Жінки	231	186	83	500
Всього	574	311	115	1000

Табл. 1: Таблиця спостережуваних значень

Обчислимо значення статистики критерію:

$$\rho = 1000 \left(\frac{343^2}{574 \cdot 500} + \frac{125^2}{311 \cdot 500} + \ldots + \frac{186^2}{311 \cdot 500} + \frac{83^2}{115 \cdot 500} - 1 \right) = 56.44$$

Водночас на рівні значущості $\alpha = 0.01$ значення критичної точки

$$\chi^2_{0.99;\ (2-1)(3-1)} = \chi^2_{0.99;\ 2} = 9.21$$

Висновок

Оскільки значення статистики критерію перевищує значення критичної точки, то згідно критерію (2) гіпотеза про однорідність статистичних даних відхиляється. Отже, ймовірності спостереження оцінок різного рівня у вибірках для чоловіків та жінок, відповідно, різняться, тобто наявна ситуація неоднакового розподілу балів в залежності від статі.

Значення `p-value`

Нехай випадкова величина τ має такий самий розподіл як і статистика критерію й при цьому не залежить від неї: $\tau \sim \chi^2(2)$. Тоді величина `p-value` обчислюється як така ймовірність:

$$P(\tau > \rho \mid H_0) = P(\tau > 56.44) = 1 - P(\tau \leqslant 56.44) = 1 - F_\tau(56.44) = 0.0001$$

Тож при значенні $\alpha <$ `p-value` гіпотеза H_0 приймалася б.

Побудова довірчого інтервалу для різниці середніх

Знову маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$$X_1 \dots X_n \sim F_x \qquad \text{результати чоловіків}$$
$$Y_1 \dots Y_m \sim F_y \qquad \text{результати жінок}$$

Спробуємо оцінити, в якому інтервалі лежить значення різниці середніх балів для вибірок результатів ЗНО з української мови чоловіків та жінок. Якщо у вказаному довірчому інтервалі буде значення $\{0\}$, тоді можна стверджувати про відсутність значної відмінності між теоретичними математичними сподіваннями цих двох вибірок.

Нормалізація даних

Хоча розподіли F_x та F_y оригінальних вибірок $\{X_i\}$ й $\{Y_j\}$ невідомі, через великий обсяг наявних даних можна виокремити N незалежних вибірок виду $X^1 \dots X^N$ й $Y^1 \dots Y^N$, які в силу центральної граничної теореми (далі – ЦГТ) мають нормальний розподіл.

Виконаємо ланцюжок перетворень на прикладі вибірки результатів чоловіків. ЦГТ для великої фіксованої кількості n елементів цієї послідовності незалежних однаково розподілених випадкових величин матиме вид:

$$\frac{\sum\limits_{i=1}^{n} X_i - M\sum\limits_{i=1}^{n} X_i}{\sqrt{D\sum\limits_{i=1}^{n} X_i}} \approx N(0,1)$$

Використовуючи позначення $\mu_x = MX_i,\ \sigma_x^2 = DX_i$, спростимо вираз, скориставшись лінійними властивостями дисперсії та математичного сподівання:

$$\frac{n\overline{X} - n\mu_x}{\sqrt{n\sigma_x^2}} \approx N(0,1),\ \text{де}\ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Залишивши n виключно у знаменнику, виконаємо перетворення:

$$\frac{n\overline{X} - n\mu_x}{\sqrt{n\sigma_x^2}} \approx N(0,1) \ \Rightarrow\ \frac{\overline{X} - \mu_x}{\sqrt{\sigma_x^2/n}} \approx N(0,1) \ \Rightarrow\ \overline{X} \approx N(\mu_x, \tfrac{1}{n}\sigma_x^2)$$

При розгляді такої ж кількості спостережень n аналогічним чином отримуємо знормоване значення результатів жінок, де $\mu_y = MY_i,\ \sigma_y^2 = DY_i$:

$$\overline{Y} \approx N(\mu_y, \tfrac{1}{n}\sigma_y^2)$$

Отже, великий обсяг початкових даних дозволяє розбити оригінальні результати $\{X_i\}$ й $\{Y_j\}$ на значну кількість достатньо великих неперетинних множин, сформованих випадковим чином, для того, щоб мати змогу застосувати ЦГТ і надалі розглядати набори $\overline{X^1}\dots\overline{X^N}$ та $\overline{Y^1}\dots\overline{Y^N}$ як незалежні дослідження:

$$\overline{X^1}\dots\overline{X^N} \sim N(\mu_x, \tfrac{1}{n}\sigma_x^2) \qquad \text{середні результати чоловіків,} \tag{4}$$

$$\overline{Y^1}\dots\overline{Y^N} \sim N(\mu_y, \tfrac{1}{n}\sigma_y^2) \qquad \text{середні результати жінок,} \tag{5}$$

при цьому

$$\overline{X^i} = \frac{1}{n}\sum_{j=1}^{n} X_j,\ \overline{Y^i} = \frac{1}{n}\sum_{j=1}^{n} Y_j,\ i = \overline{1,N}$$

Тож початкові спостережувані результати вдалося перетворити у випадковим чином сформовані незалежні вибірки із відомими розподілами. Фактично, проведено процес нормалізації початкових даних. Суть процесу схематично можна зобразити на прикладі вибірки $\{X_i\}$:

$$\begin{array}{ccc} X_1^1\dots X_n^1 \sim F_x & & \sqrt{n}\cdot\dfrac{\overline{X^1}-\mu_x}{\sqrt{\sigma_x^2}} \sim N(0,1) \\ \dots & \xRightarrow{\text{ЦГТ}} & \dots \\ X_1^N\dots X_n^N \sim F_x & & \sqrt{n}\cdot\dfrac{\overline{X^N}-\mu_x}{\sqrt{\sigma_x^2}} \sim N(0,1) \end{array} \tag{6}$$

Повертаючись до отриманих наборів (4) та (5), переконаємося, що новоутворені вибірки справді мають нормальний розподіл. Для цього побудуємо гістограми, які за означенням є наближеннями істинних щільностей, а потім візуально оглянемо отримані криві:

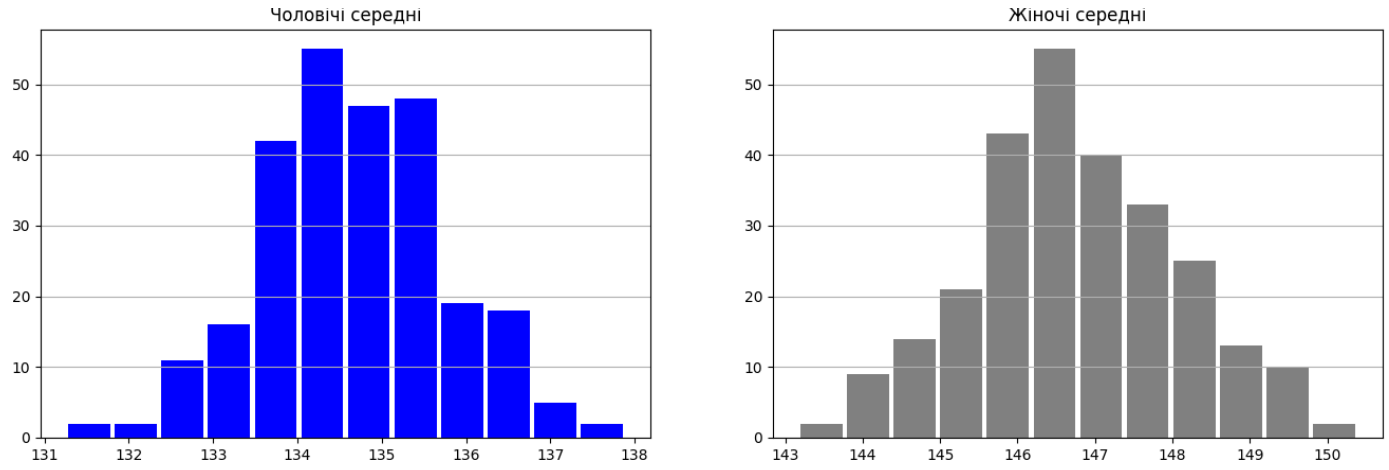

Рис. 3: Гістограми наборів усереднених результатів ЗНО з української мови

Зауважимо, що криві в обох випадках мають схожі риси (мова про «висоту» та «ширину» графіків). Тож висунемо припущення, що дисперсії цих вибірок однакові: $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Таким чином вирази (4) та (5) зведуться до такого виду:

$$\overline{X^1} \dots \overline{X^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2),\ \overline{Y^1} \dots \overline{Y^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2) \tag{7}$$

Пошук центральної статистики

Взявши за основу вибірки (7), спробуємо віднайти для параметра $\theta = \mu_y - \mu_x$ так звану центральну статистику $G(\overline{X^1} \dots \overline{X^N},\ \overline{Y^1} \dots \overline{Y^N},\ \theta) \equiv G$, яка має задовільняти двом умовам: по-перше, розподіл $f_G(x)$ не залежить від параметра θ, а по-друге, функція G неперервна і монотонна за θ.

Перш за все, виходячи з властивостей нормального розподілу

$$\overline{X} \equiv \tfrac{\overline{X^1}+\dots+\overline{X^N}}{N} \sim N(\mu_x, \tfrac{1}{Nn}\sigma^2),\ \overline{Y} \equiv \tfrac{\overline{Y^1}+\dots+\overline{Y^N}}{N} \sim N(\mu_y, \tfrac{1}{Nn}\sigma^2)$$

А отже, розподіл різниці $\overline{Y} - \overline{X}$ матиме вид

$$\overline{Y} - \overline{X} \sim N(\mu_y - \mu_x,\ \tfrac{2}{Nn}\sigma^2)$$

Тоді позначимо

$$\xi \equiv \frac{\overline{Y} - \overline{X} - (\mu_y - \mu_x)}{\sqrt{\frac{2}{Nn}\sigma^2}} \sim N(0,1) \tag{8}$$

Крім того, зазначимо, що

$$\frac{(n-1)\left(S_x^2\right)_i}{\sigma^2} \sim \chi^2(n-1),\ \frac{(n-1)\left(S_y^2\right)_i}{\sigma^2} \sim \chi^2(n-1),\ i = \overline{1, N} \tag{9}$$

де

$$\left(S_x^2\right)_i = \tfrac{1}{n-1}\sum_{j=1}^{n}(X_j^i - \overline{X^i})$$
$$\left(S_y^2\right)_i = \tfrac{1}{n-1}\sum_{j=1}^{n}(Y_j^i - \overline{Y^i})$$

вибіркові дисперсії двох наборів вибірок середніх результатів

А отже, сума випадкових величин (9) як сума незалежних випадкових величин матиме розподіл χ^2 із $N(n-1) + N(n-1)$ степенями свободи:

$$\eta \equiv \frac{(n-1)}{\sigma^2}\left(\sum_{i=1}^{N}\left(S_x^2\right)_i + \sum_{i=1}^{N}\left(S_y^2\right)_i\right) \sim \chi^2\left(2N(n-1)\right) \tag{10}$$

Тоді випадкова величина

$$\zeta \stackrel{\text{def}}{=} \frac{\xi}{\sqrt{\frac{\eta}{2N(n-1)}}} \sim t\left(2N(n-1)\right) \tag{11}$$

Підставивши вирази (8) й (10) у формулу (11), отримаємо шукану центральну статистику для параметра $\theta = \mu_y - \mu_x$:

$$G = \frac{(\overline{Y} - \overline{X}) - \theta}{\sqrt{\frac{2}{Nn}}} \cdot \sqrt{\frac{2N(n-1)}{(n-1)(S_X^2 + S_Y^2)}} \sim t\left(2N(n-1)\right), \tag{12}$$

де $S_X^2 = \sum\limits_{i=1}^{N} \left(S_x^2\right)_i$, $S_Y^2 = \sum\limits_{i=1}^{N} \left(S_y^2\right)_i$, а статистика G має розподіл Ст'юдента із відповідною кількістю степенів свободи.

Побудова довірчого інтервалу

Побудуємо довірчий інтервал рівня довіри $\gamma = 0.95$:

$$\gamma \stackrel{\text{def}}{=} P(g_1 < G < g_2) = \int\limits_{g_1}^{g_2} f_G(x)\, dx$$

В силу симетричності розподілу Ст'юдента, найкоротший центральний довірчий інтервал матиме вид:

$$\gamma = P(g_1 < G < g_2) = P(|G| < g) \tag{13}$$

На малюнку нижче схематично зображено ідею пошуку довірчого інтервалу та визначення відповідного квантиля g :

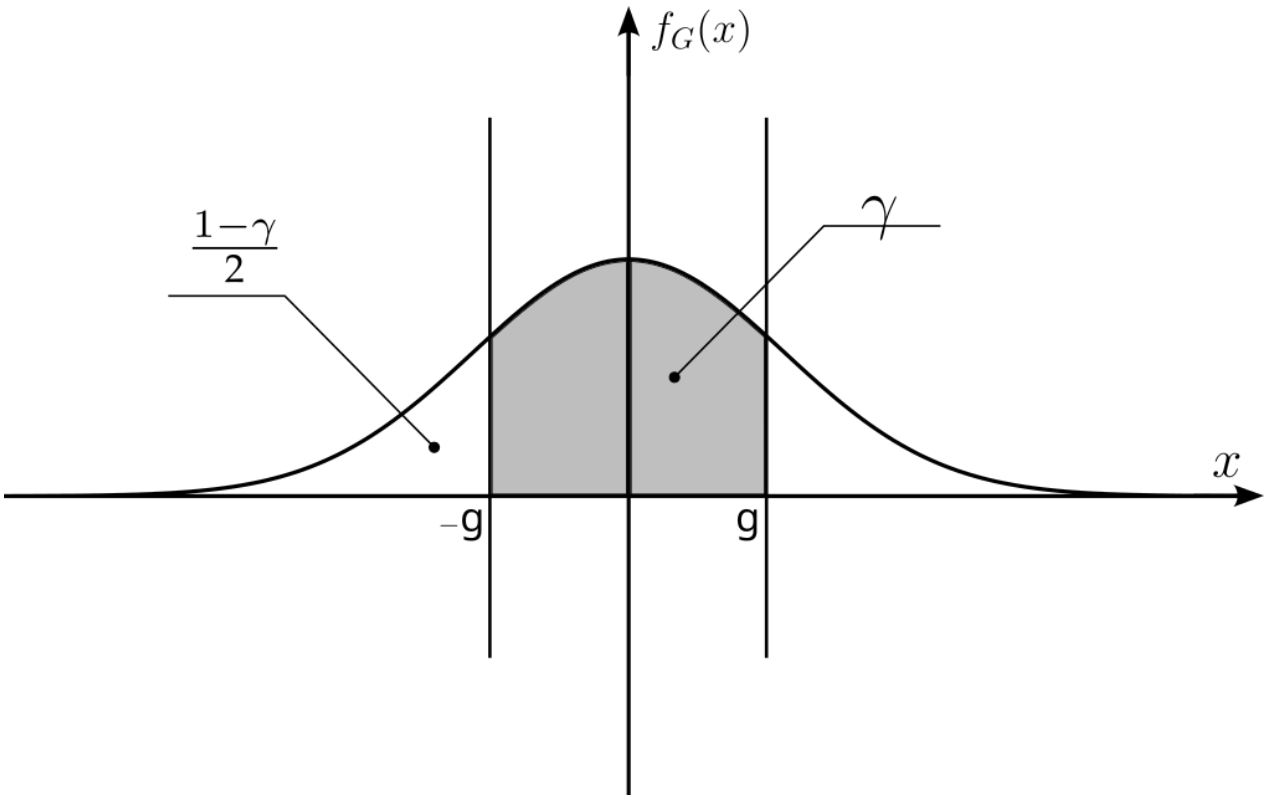

Рис. 4: Пошук квантиля розподілу Ст'юдента

Тож враховуючи, що $\int\limits_{-\infty}^{\infty} f_G(x)\,dx \overset{\text{def}}{=} 1$, маємо такий ланцюжок міркувань:

$$\int\limits_{-\infty}^{g} f_G(x)\,dx = \frac{1+\gamma}{2} \;\Rightarrow\; F(g) = \frac{1+\gamma}{2} \;\Rightarrow\; g = t_{\frac{1+\gamma}{2};\ 2N(n-1)}$$

Знайдено квантиль рівня $\frac{1+\gamma}{2}$ розподілу Ст'юдента із $2N(n-1)$ степенями свободи. Підставляючи усі отримані результати у формулу (13), отримаємо:

$$\gamma = P\left(-t_{\frac{1+\gamma}{2};\ 2N(n-1)} < \left((\overline{Y}-\overline{X})-\theta\right)\cdot\sqrt{\frac{N^2 n}{S_X^2+S_Y^2}} < t_{\frac{1+\gamma}{2};\ 2N(n-1)}\right) \tag{14}$$

Останнім кроком вкажемо конкретні значення усіх необхідних величин:

n	N	$g = t_{0.975;\ 2N(n-1)}$	S_X^2	S_Y^2	$\overline{Y}-\overline{X}$
500	267	1.96	175017.2397	200773.7347	12.0722

Табл. 2: Значення шуканих параметрів

Підставивши у вираз (14) значення, які наведені у таблиці вище, отримаємо такий довірчий інтервал для параметра $\theta = \mu_y - \mu_x$:

$$P(11.87 < \theta < 12.27) = 0.95$$

Висновок

Отримано довірчий інтервал рівня довіри $\gamma = 0.95$ для величини різниці середніх значень оцінок з української мови вибірок результатів чоловіків та жінок: $(11.87,\ 12.27) \ni \mu_y - \mu_x$. Оскільки у вказаному проміжку немає нульового значення, гіпотезу про нерозрізнюваність середніх можна відхилити.

Крім того, візуалізовані дані на Рис. 3 узгоджуються із отриманим відрізком значень, оскільки вершини кривих відрізняються за віссю абсцис приблизно на знайдені показники. Можемо стверджувати, що при порівнянні результатів чоловіка та жінки оцінка з української мови у жінки буде більшою за оцінку у чоловіка на бал у 12.07 ± 0.2 пункти, при цьому в середньому у п'яти зі ста таких порівнянь вказане наближення може бути хибним.

Результати з математики

Оглянемо отримані 122 026 результатів, при цьому зауважимо, що чоловічих на 11 984 більше за жіночих. Зобразимо гістограми результатів ЗНО з математики 2019 року для вибірки, наприклад, 10 000 учнів:

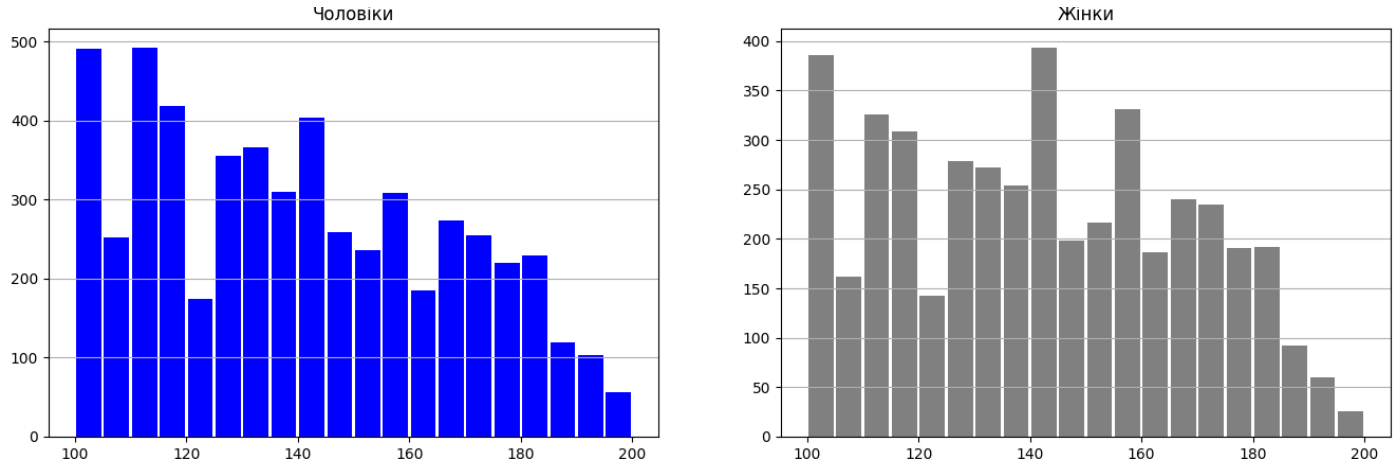

Рис. 5: Результати з математики

З'ясуємо, чи відмінність між результатами з математики для жінок та чоловіків є статистично значущою.

Перевірка гіпотези однорідності даних

Нехай маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$X_1 \dots X_n \sim F_x$	результати чоловіків
$Y_1 \dots Y_m \sim F_y$	результати жінок

Перевіримо гіпотезу про однорідність статистичного матеріалу, тобто гіпотезу, що ймовірності спостереження умовно високих, помірних та низьких балів в обох вибірках є однаковими. Таким чином матимемо $k = 2$ вибірок, в яких елементи приймають $s = 3$ різних значень. За аналогічних позначень гіпотеза перевірки однорідності спосережуваних даних формулюватиметься так само як і у виразах (1):

нульова гіпотеза	$H_0 : p_{11} = p_{21},\ p_{12} = p_{22},\ p_{13} = p_{23}$
проти альтернативи	$H_1 : \exists i, j \quad p_{1,j} \neq p_{2,j}$

А тоді критерієм перевірки гіпотези рівно як і у випадку обробки результатів з української мови (2) слугуватиме правило

$$\delta(x_1 \dots x_n,\ y_1 \dots y_m) = \begin{cases} H_1, & \rho \geqslant \chi^2_{1-\alpha;\ (k-1)(s-1)} \\ H_0, & \rho < \chi^2_{1-\alpha;\ (k-1)(s-1)} \end{cases},$$

де статистика критерію ρ обчислюється аналогічно до формули (3):

$$\rho \equiv \rho(x_1 \dots x_n,\ y_1 \dots y_m) = (n+m)\left(\sum_{i=1}^{k}\sum_{j=1}^{s}\frac{\vartheta_{ij}^2}{\vartheta_{i\cdot}\vartheta_{\cdot j}} - 1\right)$$

Таблиця спостережуваних даних

Складемо таблицю за спостережуваними даними навмання обраних $n = 500$ та $m = 500$ елементів із вибірок результатів ЗНО з математики для чоловіків та жінок:

	Низькі бали	Помірні бали	Високі бали	Всього
Чоловіки	270	165	65	500
Жінки	268	174	58	500
Всього	538	339	123	1000

Табл. 3: Таблиця спостережуваних значень

Обчислимо значення статистики критерію:

$$\rho = 1000\left(\frac{270^2}{538\cdot 500} + \frac{165^2}{339\cdot 500} + \ldots + \frac{174^2}{339\cdot 500} + \frac{58^2}{123\cdot 500} - 1\right) = 0.64$$

Водночас на рівні значущості $\alpha = 0.01$ значення критичної точки

$$\chi^2_{0.99;\ (2-1)(3-1)} = \chi^2_{0.99;\ 2} = 9.21$$

Висновок

Оскільки значення статистики критерію є меншим за значення критичної точки, то згідно критерію гіпотеза про однорідність статистичних даних приймається. Отже, ймовірності спостереження оцінок різного рівня у вибірках для чоловіків та жінок, відповідно, є однаковими. Тобто стать ніяким чином не впливає на рівень отриманого балу.

Значення `p-value`

Нехай випадкова величина τ має такий самий розподіл як і статистика критерію й при цьому не залежить від неї: $\tau \sim \chi^2(2)$. Тоді величина `p-value` обчислюється як така ймовірність:

$$P(\tau > \rho \mid H_0) = P(\tau > 0.64) = 1 - P(\tau \leqslant 0.64) = 1 - F_\tau(0.64) = 0.7261$$

Тож для довільних значень $\alpha <$ `p-value` гіпотеза H_0 приймається.

Побудова довірчого інтервалу для різниці середніх

Маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$$X_1 \dots X_n \sim F_x \qquad \text{результати чоловіків}$$
$$Y_1 \dots Y_m \sim F_y \qquad \text{результати жінок}$$

Спробуємо оцінити, в якому інтервалі лежить значення різниці середніх балів для вибірок результатів ЗНО з математики чоловіків та жінок.

Нормалізація даних та пошук центральної статистики

Проведемо нормалізацію даних крок за кроком, як це зображено на схемі (6). У результаті отримаємо такі знормовні вибірки:

$$\overline{X^1} \dots \overline{X^N} \sim N(\mu_x, \tfrac{1}{n}\sigma_x^2) \qquad \text{середні результати чоловіків,}$$
$$\overline{Y^1} \dots \overline{Y^N} \sim N(\mu_y, \tfrac{1}{n}\sigma_y^2) \qquad \text{середні результати жінок,}$$

при цьому

$$\overline{X^i} = \frac{1}{n}\sum_{j=1}^{n} X_j,\ \overline{Y^i} = \frac{1}{n}\sum_{j=1}^{n} Y_j,\ i = \overline{1, N}$$

Побудуємо гістограми отриманих вибірок та переконаємося, що вони справді мають нормальний розподіл:

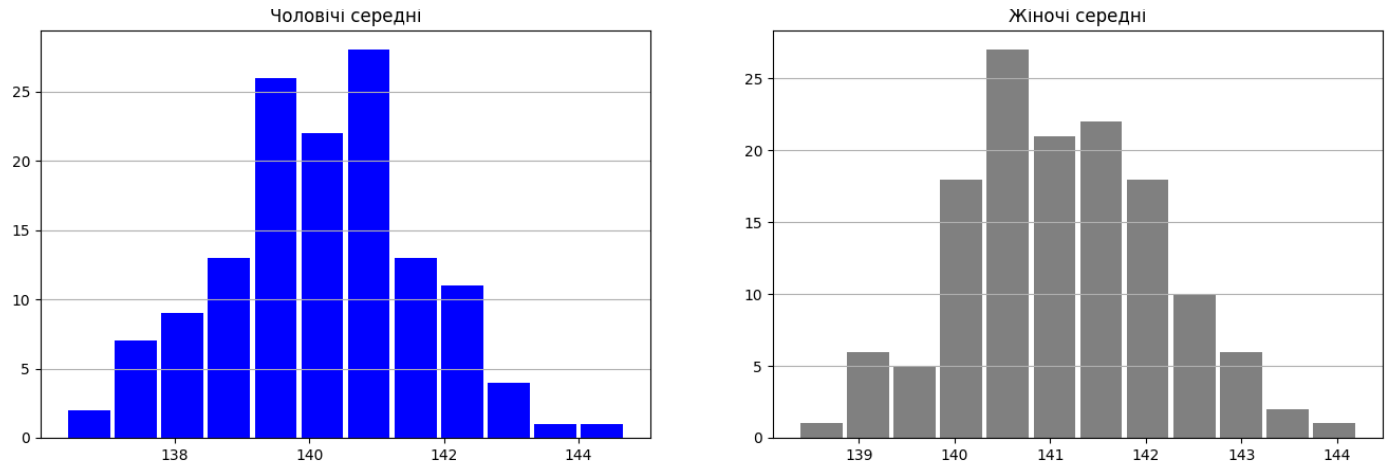

Рис. 6: Гістограми наборів усереднених результатів ЗНО з математики

Як бачимо, криві в обох випадках мають схожі риси. Тож знову висунемо припущення, що дисперсії цих вибірок однакові, тобто $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Таким чином:

$$\overline{X^1} \dots \overline{X^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2),\ \overline{Y^1} \dots \overline{Y^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2)$$

Тоді виконавши аналогічні перетворення, які наведенні на стр. 10, отримаємо шукану центральну статистику для параметра $\theta = \mu_y - \mu_x$:

$$G = \left((\overline{Y} - \overline{X}) - \theta\right) \cdot \sqrt{\frac{N^2 n}{S_X^2 + S_Y^2}} \sim t\left(2N(n-1)\right)$$

Побудова довірчого інтервалу

Побудуємо довірчий інтервал рівня довіри $\gamma = 0.95$:

$$\gamma \stackrel{\text{def}}{=} P(g_1 < G < g_2) = \int_{g_1}^{g_2} f_G(x)\, dx$$

В силу симетричності розподілу Ст'юдента, найкоротший центральний довірчий інтервал матиме вид:

$$\gamma = P(g_1 < G < g_2) = P(|G| < g),$$

де значення $g = t_{\frac{1+\gamma}{2};\ 2N(n-1)}$ – квантиль рівня $\frac{1+\gamma}{2}$ розподілу Ст'юдента із $2N(n-1)$ степенями свободи. Отже:

$$\gamma = P\left(-t_{\frac{1+\gamma}{2};\ 2N(n-1)} < \left((\overline{Y} - \overline{X}) - \theta\right) \cdot \sqrt{\frac{N^2 n}{S_X^2 + S_Y^2}} < t_{\frac{1+\gamma}{2};\ 2N(n-1)}\right) \tag{15}$$

Останнім кроком вкажемо конкретні значення усіх необхідних величин:

n	N	$g = t_{0.975;\ 2N(n-1)}$	S_X^2	S_Y^2	$\overline{Y} - \overline{X}$
400	137	1.96	95092.9605	90122.9510	0.9639

Табл. 4: Значення шуканих параметрів

Підставивши у вираз (15) значення, які наведені у таблиці вище, отримаємо такий довірчий інтервал для параметра $\theta = \mu_y - \mu_x$:

$$P(0.66 < \theta < 1.27) = 0.95$$

Висновок

Отримано довірчий інтервал рівня довіри $\gamma = 0.95$ для величини різниці середніх значень оцінок з математики вибірок результатів чоловіків та жінок: $(0.66,\ 1.27) \ni \mu_y - \mu_x$. Оскільки у вказаному проміжку немає нульового значення, гіпотезу про нерозрізнюваність середніх можна відхилити.

Крім того, візуалізовані дані на Рис. 6 узгоджуються із отриманим відрізком значень, оскільки вершини кривих відрізняються за віссю абсцис приблизно на знайдені показники. Можемо стверджувати, що при порівнянні результатів чоловіка та жінки оцінка з математики у жінки буде більшою за оцінку у чоловіка на бал у 0.96 ± 0.3 пункти, при цьому в середньому у п'яти зі ста таких порівнянь вказане наближення може бути хибним.

Результати з англійської мови

Оглянемо отримані 78 797 результатів, де жіночих є на 8 339 більше за чоловічих. Зобразимо гістограми результатів ЗНО з англійської мови 2019 року для вибірки, наприклад, 10 000 учнів:

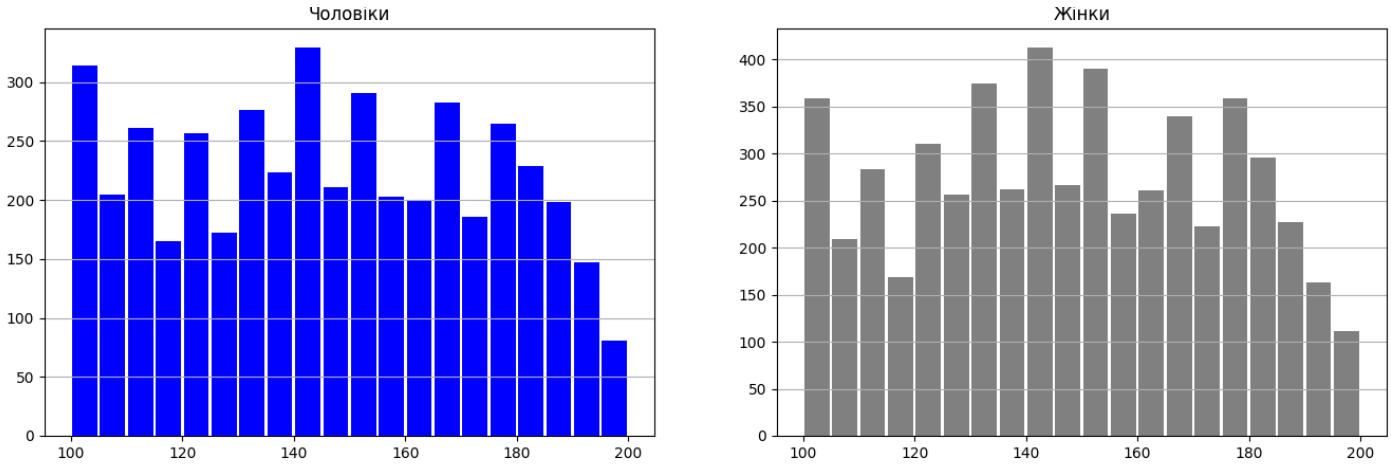

Рис. 7: Результати з англійської мови

З'ясуємо, чи відмінність між результатами для жінок та чоловіків є статистично значущою.

Перевірка гіпотези однорідності даних

Нехай маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$X_1 \dots X_n \sim F_x$	результати чоловіків
$Y_1 \dots Y_m \sim F_y$	результати жінок

Перевіримо гіпотезу про однорідність статистичного матеріалу, тобто гіпотезу, що ймовірності спостереження умовно високих, помірних та низьких балів в обох вибірках є однаковими. Таким чином матимемо $k = 2$ вибірок, в яких елементи приймають $s = 3$ різних значень. Знову за аналогічних позначень гіпотеза перевірки однорідності спосережуваних даних формулюватиметься так само як і у виразах (1):

$$\begin{array}{ll} \text{нульова гіпотеза} & H_0 : p_{11} = p_{21},\ p_{12} = p_{22},\ p_{13} = p_{23} \\ \text{проти альтернативи} & H_1 : \exists i, j \quad p_{1,j} \neq p_{2,j} \end{array}$$

А тоді критерієм перевірки гіпотези рівно як і у випадку обробки результатів з української мови (2) слугуватиме правило

$$\delta(x_1 \dots x_n,\ y_1 \dots y_m) = \begin{cases} H_1, & \rho \geqslant \chi^2_{1-\alpha;\ (k-1)(s-1)} \\ H_0, & \rho < \chi^2_{1-\alpha;\ (k-1)(s-1)} \end{cases},$$

де статистика критерію ρ обчислюється аналогічно до формули (3):

$$\rho \equiv \rho(x_1 \dots x_n,\ y_1 \dots y_m) = (n+m)\left(\sum_{i=1}^{k}\sum_{j=1}^{s}\frac{\vartheta_{ij}^2}{\vartheta_{i\cdot}\vartheta_{\cdot j}} - 1\right)$$

Таблиця спостережуваних даних

Складемо таблицю за спостережуваними даними навмання обраних $n = 500$ та $m = 500$ елементів із вибірок результатів ЗНО з англійської мови для чоловіків та жінок:

	Низькі бали	Помірні бали	Високі бали	Всього
Чоловіки	248	168	84	500
Жінки	224	196	80	500
Всього	472	364	164	1000

Табл. 5: Таблиця спостережуваних значень

Обчислимо значення статистики критерію:

$$\rho = 1000\left(\frac{228^2}{472\cdot 500} + \frac{168^2}{364\cdot 500} + \ldots + \frac{196^2}{364\cdot 500} + \frac{80^2}{164\cdot 500} - 1\right) = 3.47$$

Водночас на рівні значущості $\alpha = 0.01$ значення критичної точки

$$\chi^2_{0.99;\ (2-1)(3-1)} = \chi^2_{0.99;\ 2} = 9.21$$

Висновок

Оскільки значення статистики критерію є меншим за значення критичної точки, то згідно критерію гіпотеза про однорідність статистичних даних приймається. Отже, ймовірності спостереження оцінок різного рівня у вибірках для чоловіків та жінок, відповідно, є однаковими. Тобто так само, як і у випадку результатів з математики, стать ніяким чином не впливає на рівень отриманого балу.

Значення `p-value`

Нехай випадкова величина τ має такий самий розподіл як і статистика критерію й при цьому не залежить від неї: $\tau \sim \chi^2(2)$. Тоді величина `p-value` обчислюється як така ймовірність:

$$P(\tau > \rho \mid H_0) = P(\tau > 3.47) = 1 - P(\tau \leqslant 3.47) = 1 - F_\tau(3.47) = 0.1764$$

Тож для довільних значень $\alpha <$ `p-value` гіпотеза H_0 приймається.

Побудова довірчого інтервалу для різниці середніх

Нехай маємо дві незалежні вибірки $\{X_i\}$ та $\{Y_j\}$ однаково розподілених випадкових величин, розподіли F_x й F_y яких нам невідомі:

$X_1 \dots X_n \sim F_x$	результати чоловіків
$Y_1 \dots Y_m \sim F_y$	результати жінок

Спробуємо оцінити, в якому інтервалі лежить значення різниці середніх балів для вибірок результатів ЗНО з англійської мови для чоловіків та жінок.

Нормалізація даних та пошук центральної статистики

Аналогічним чином роведемо нормалізацію даних крок за кроком, як це зображено на схемі (6). У результаті отримаємо такі знормовні вибірки:

$\overline{X^1} \dots \overline{X^N} \sim N(\mu_x, \frac{1}{n}\sigma_x^2)$	середні результати чоловіків,
$\overline{Y^1} \dots \overline{Y^N} \sim N(\mu_y, \frac{1}{n}\sigma_y^2)$	середні результати жінок,

при цьому

$$\overline{X^i} = \frac{1}{n}\sum_{j=1}^{n} X_j,\ \overline{Y^i} = \frac{1}{n}\sum_{j=1}^{n} Y_j,\ i = \overline{1, N}$$

Побудуємо гістограми отриманих вибірок та переконаємося, що вони справді мають нормальний розподіл (Рис. 8).

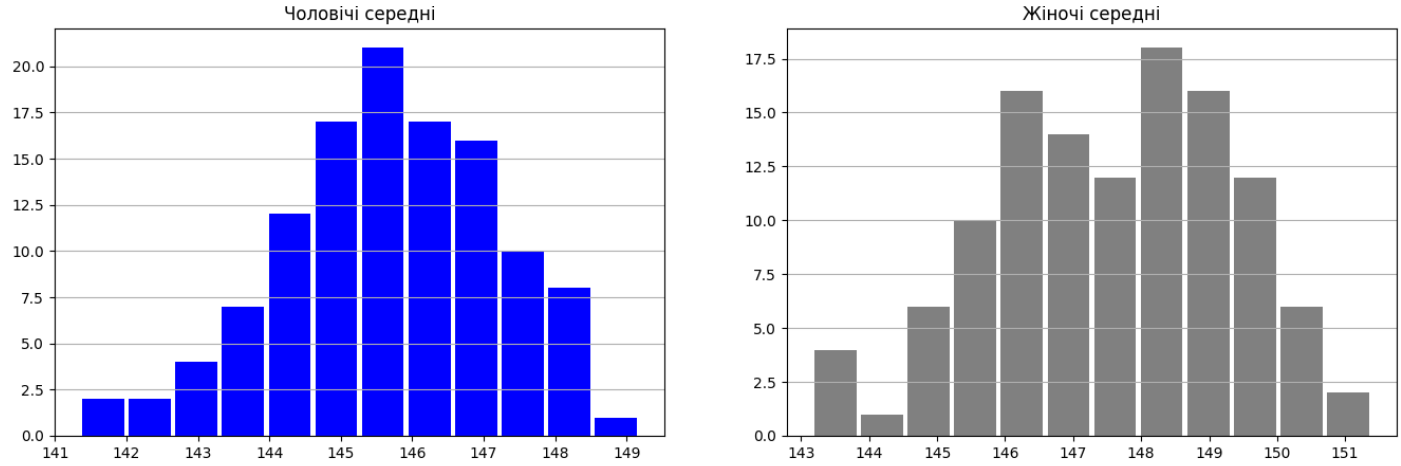

Рис. 8: Гістограми наборів усереднених результатів ЗНО з англійської мови

Знову ж таки бачимо, що криві в обох випадках мають схожі риси. Тому висунемо припущення, що дисперсії цих вибірок однакові: $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Таким чином:

$$\overline{X^1} \dots \overline{X^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2),\ \overline{Y^1} \dots \overline{Y^N} \sim N(\mu_x, \tfrac{1}{n}\sigma^2)$$

Тоді виконавши аналогічні перетворення, які наведенні на стр. 10, отримаємо шукану центральну статистику для параметра $\theta = \mu_y - \mu_x$:

$$G = \left((\overline{Y} - \overline{X}) - \theta\right) \cdot \sqrt{\frac{N^2 n}{S_X^2 + S_Y^2}} \sim t\left(2N(n-1)\right)$$

Побудова довірчого інтервалу

Побудуємо довірчий інтервал рівня довіри $\gamma = 0.95$:

$$\gamma \overset{\text{def}}{=} P(g_1 < G < g_2) = \int_{g_1}^{g_2} f_G(x)\ dx$$

В силу симетричності розподілу Ст'юдента, найкоротший центральний довірчий інтервал матиме вид:

$$\gamma = P(g_1 < G < g_2) = P(|G| < g),$$

де значення $g = t_{\frac{1+\gamma}{2};\ 2N(n-1)}$ – квантиль рівня $\frac{1+\gamma}{2}$ розподілу Ст'юдента із $2N(n-1)$ степенями свободи. Отже:

$$\gamma = P\left(-t_{\frac{1+\gamma}{2};\ 2N(n-1)} < \left((\overline{Y} - \overline{X}) - \theta\right) \cdot \sqrt{\frac{N^2 n}{S_X^2 + S_Y^2}} < t_{\frac{1+\gamma}{2};\ 2N(n-1)}\right) \qquad (16)$$

Останнім кроком вкажемо конкретні значення усіх необхідних величин:

n	N	$g = t_{0.975;\ 2N(n-1)}$	S_X^2	S_Y^2	$\overline{Y} - \overline{X}$
300	117	1.96	87866.7150	88050.4315	1.8911

Табл. 6: Значення шуканих параметрів

Підставивши у вираз (16) значення, які наведені у таблиці вище, отримаємо такий довірчий інтервал для параметра $\theta = \mu_y - \mu_x$:

$$P(1.49 < \theta < 2.29) = 0.95$$

Висновок

Отримано довірчий інтервал рівня довіри $\gamma = 0.95$ для величини різниці середніх значень оцінок з англійської мови вибірок результатів чоловіків та жінок: $(1.49,\ 2.29) \ni \mu_y - \mu_x$. Оскільки у вказаному проміжку немає нульового значення, гіпотезу про нерозрізнюваність середніх можна відхилити.

Крім того, візуалізовані дані на Рис. 8 узгоджуються із отриманим відрізком значень, оскільки вершини кривих відрізняються за віссю абсцис приблизно на знайдені показники. Можемо стверджувати, що при порівнянні результатів чоловіка та жінки оцінка з математики у жінки буде більшою за оцінку у чоловіка на бал у 1.89 ± 0.4 пункти, при цьому в середньому у п'яти зі ста таких порівнянь вказане наближення може бути хибним.

Загальний висновок

Метою статистичного аналізу було з'ясувати, чи є значущою відмінність між балами ЗНО для чоловіків та жінок. При цьому розглядалися результати трьох різних предметів: української мови, математики та англійської мови.

Перш за все вдалося встановити, що на рівні значущості $\alpha = 0.01$ лише для української мови наявна залежність між рівнем набраного балу та чинником статі. При спробі побудови інтервалу можливих значень різниці середніх балів жінок та чоловіків, було отримано проміжок у 12.07 ± 0.2 пункти, що свідчить про значне перевищення оцінок жінок у порівнянні з оцінками чоловіків. Крім того, значення вказаного довірчого інтервалу рівня довіри $\gamma = 0.95$ додатково підтверджує хибність гіпотези однорідності статистичних даних результатів з української мови для чоловіків та жінок.

У випадку аналізу балів з математики та англійської мови вже на першому кроці виявлено відсутність впливу статі на рівень отриманого балу. До того ж довірчі

інтервали різниці середніх оцінок різних статей мають невеликі значення: 0.96 ± 0.3 для математики та 1.89 ± 0.4 для англійської мови.

Отже, як результат статистичного аналізу можна стверджувати, що лише з української мови наявна статистична значущість відмінності результатів ЗНО в залежності від статі. Водночас на набрані бали з математики чинник статі не має значного впливу.