

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

Факультет комп'ютерних наук і технологій
Кафедра системного аналізу та обчислювальної математики

Пояснювальна записка
до кваліфікаційної роботи
бакалавр
(ступінь вищої освіти)

на тему: Аналіз гендерних відмінностей результатів ЗНО

Виконав: студент 4 курсу, групи КНТ-819сп
Маньков Владислав

"підпис"

Спеціальності 124 Системний аналіз

Освітня програма (спеціалізація):

Інтелектуальні технології та прийняття рішень в
складних системах

Керівник: Бахрушин В.Є.
(прізвище та ініціали)

"підпис"

Рецензент Ігнахіна М.О.
(прізвище та ініціали)

"підпис"

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»
(повне найменування закладу вищої освіти)

Інститут, факультет Інформатики та радіоелектроніки, комп'ютерних наук і технологій

Кафедра Системного аналізу та обчислювальної математики

Ступінь вищої освіти бакалавр

Спеціальність 124 – Системний аналіз
(код і найменування)

Освітня програма

(спеціалізація) «Інтелектуальні технології та прийняття рішень в складних системах»

ЗАТВЕРДЖУЮ

Завідувач кафедри _____

_____ Корніч Г.В "підпис"
« _____ » _____ 20 ____ року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТА(КИ)

_____ Манькова Владислава Володимировича

(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи Аналіз гендерних відмінностей результатів ЗНО
керівник кваліфікаційної роботи Бахрушин Володимир Євгенович, д.ф.-м.н, професор,

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом закладу вищої освіти від «11» травня 2022 року № 121

2. Строк подання студентом проєкту (роботи) 15 червня 2022 року.

3. Вихідні дані до проєкту (роботи) Матеріали українського центру оцінювання якості освіти

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Проблематика гендерної нерівності в сфері освіти, методика виконання роботи, аналіз результатів ЗНО.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

6. Консультанти розділів проєкту (роботи) Бахрушин В.Є д.ф.-м.н., професор

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	прийняв виконане завдання
1	Бахрушин В.Є, д.ф.-м.н., професор	"підпис"	"підпис"
2	Бахрушин В.Є, д.ф.-м.н., професор	"підпис"	"підпис"
3	Бахрушин В.Є, д.ф.-м.н., професор	"підпис"	"підпис"

7. Дата видачі завдання «02» травня 2022 року.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проєкту (роботи)	Строк виконання етапів проєкту (роботи)	Примітка
1	Аналіз літератури по темі дослідження	02.05-15.05	"підпис"
2	Підготовка даних до дослідження	16.05-22.05	"підпис"
3	Відпрацювання методики дослідження	23.05-29.05	"підпис"
4	Виконання розрахунків і аналіз результатів	30.05-08.06	"підпис"
5	Написання тексту пояснювальної записки	09.06-15.06	"підпис"

Студент(ка)

"підпис"

Маньков В.В.
(прізвище та ініціали)

Керівник проєкту (роботи)

"підпис"

Бахрушин В.Є.
(прізвище та ініціали)

РЕФЕРАТ

Кваліфікаційна робота: 79с., 30 табл, 54 рисунки, 1 додаток

Об'єктом дослідження є результати Зовнішнього незалежного оцінювання (ЗНО) за 2021 рік.

Мета дослідження: визначення наявності або відсутності істотних відмінностей результатів ЗНО між чоловіками та жінками.

Предмет дослідження: статистичний аналіз результатів ЗНО.

Методи дослідження: статистичні критерії: критерій Ліллієфорса, критерій Андерсона-Дарлінга, критерій Колмогорова-Смірнова, Т-критерій Уїлкоксона.

Ключові слова: ЗНО, ГЕНДЕРНІ ВІДМІННОСТІ, ОПИСОВА СТАТИСТИКА, КРИТЕРІЙ ЛІЛЛІЄФОРСА, КРИТЕРІЙ АНДЕРСОНА-ДАРЛІНГА, ОДНОРІДНІСТЬ ВИБІРОК, КРИТЕРІЙ КОЛМОГОРОВА-СМІРНОВА, Т-КРИТЕРІЙ УІЛКОКСОНА.

ЗМІСТ

ЗАВДАННЯ	2
РЕФЕРАТ	4
ВСТУП.....	6
1. ТЕОРЕТИЧНА ЧАСТИНА	7
1.1 Світова проблематика гендерної нерівності в сфері освіти	7
1.2 Методи статистичних досліджень	9
1.3 Застосування статистичних критеріїв для аналізу даних	11
1.3.1 Критерій Уелча.....	12
1.3.2 Критерій Фішера	12
1.3.3 Критерій Ліллієфорса	13
1.3.4 Критерій Андерсона-Дарлінга	14
1.3.5 Критерій Колмогорова-Смірнова.....	15
1.3.6 Критерій Уїлкоксона.....	16
2. МЕТОДИКА ВИКОНАННЯ РОБОТИ	17
3. ПРАКТИЧНА РОБОТА	21
3.1 Аналіз даних результатів ЗНО з української мови	21
3.2 Історія України	29
3.3 Математика	33
3.4 Фізика	37
3.5 Географія.....	41
3.6 Англійська мова	45
3.7 Німецька мова.....	49
3.8 Французька мова	53
3.9 Біологія	56
3.10 Іспанська мова	60
ВИСНОВКИ.....	63
ПЕРЕЛІК ПОСИЛАНЬ	64
ДОДАТОК А. Код програми.....	66

ВСТУП

Зовнішнє незалежне оцінювання є тестуванням спрямованим на визначення рівня навчальних досягнень випускників закладів при їхньому вступі до закладів вищої освіти в Україні.

Мета зовнішнього незалежного оцінювання: забезпечення реалізації конституційних прав громадян на рівний доступ к якісній освіті, здійснення контролю за дотриманням Державного стандарту базової та повної середньої освіти [1].

Результати зовнішнього незалежного оцінювання зараховуються як результати державної підсумкової та як результати вступних іспитів до закладів вищої освіти [1].

Для досягнення мети дослідження були опрацьовані методики дослідження, розраховано основні показники описової статистики, перевірена однорідність вибірок результатів чоловіків та жінок за предметами, проаналізовано отримані результати про наявність або відсутність відмінностей між чоловіками та жінками.

1 ТЕОРЕТИЧНА ЧАСТИНА

1.1 Світова проблематика гендерної нерівності в сфері освіти

Гендерна нерівність в світі стає предметом палких дискусій та зіткнень у всіх сферах життя, у тому числі й в освіті.

Згідно даних ООН, досягнення гендерної рівності входить до однієї з цілей у сфері сталого розвитку країн до 2030 року. Ключову роль у вирішенні цього завдання грає етап здобуття освіти, на якому необхідно надати однаковий доступ до якісного навчання представникам різної статі [2].

Успіхи останніх 50 років у скороченні гендерної нерівності в доступі до освіти важко переоцінити – у всіх країнах досягнені значні покращення в цій сфері.

В 1970 році різниця в роках отриманого навчання між чоловіками та жінками складала майже 2 роки у всьому світі, а к 2018 року ця цифра скоротилась до 0,3 років. В ряді країн дана тенденція призвела до того, що жінки перевершили чоловіків за роками навчання. Довше чоловіків жінки до 2018 року стали навчатися в країнах з найбільшим рівнем доходу – в Європі та Центральній Азії [2].

Примітно, що до 2030 року прогнозується ще більш значне посилення наявного тренду на переверот гендерної нерівності в сфері освіти [2].

Не дивлячись на загальні позитивні тенденції, гендерна нерівність в школі продовжує існувати й особливо проявляється в окремих країнах.

Забезпечення рівного доступу до здобуття освіти ще не гарантує відсутність дискримінуючих практик та несправедливих відмінностей безпосередньо в процесі навчання.

Міжнародне дослідження TIMSS-2019, спрямоване на оцінку знань з математики та природознавства, свідчить про наявність розриву шкільних досягнень серед хлопців та дівчат. З математики у 2019 році в 27 країнах хлопчики-четверокласники випереджали дівчаток, а дівчатка – лише в 4 країнах [3]. До 8-го класу, тим не менш, ці відмінності повністю вирівнюються: дівчатка випереджають хлопчиків у 7 країнах, а хлопчики в 6 країнах. Відомо, що саме за математичними

дисциплінами (поряд з технічними) представники чоловічої статі демонструють більш високі досягнення.

За природничими предметами загальним трендом в світі є перевага дівчаток як в четвертому, так й у восьмому класі [3].

Абсолютно протилежну ситуацію ми можемо спостерігати, якщо звернемося до іншої предметної області – читання. Ще одне міжнародне дослідження – PIRLS – дозволяє оцінити навички четверокласників у читацькій грамотності. У переважній більшості країн-учасниць дослідження дівчатка випередили хлопців в навичках читання [4].

У PISA, яка вимірює навички застосування знань на практиці у всіх трьох предметних областях серед 15-річних підлітків, моделі гендерної нерівності приблизно такі самі [5].

Ще одне дослідження на основі даних іспитів GSCE (загальний сертифікат про середню освіту в Англії, Уельсі та Північній Ірландії) в 2017 році показало, що 70,9% жінок отримали принаймні оцінку C або 4 у порівнянні їх однолітків-чоловіків. В 2016 році математика була єдиним предметом, за яким хлопці отримали більш високі оцінки, ніж дівчатка, а в 2017 хлопці показали кращі результати в фізиці, економіці та статистиці, крім математики [6].

Результати GSCE 2019 року показують гендерний розрив в 9,8%: 71,7% жінок отримали оцінку 4/C або вище порівняно тільки з 62,9% чоловіків. З англійської мови дівчатка вчать набагато краще, ніж хлопці – вони перевершують хлопчиків приблизно на 16% по гарним оцінкам. З математики гендерний розрив складає 0,5% на користь хлопців [6].

English	Male	397993 (381731)	14.4 (14.0)	9.6 (9.8)	54.2 (54.3)	98.1 (98.3)	100.0 (100.0)
	Female	367339 (351354)	13.2 (12.8)	18.7 (18.7)	70.5 (69.9)	99.4 (99.5)	100.0 (100.0)
	Male & Female	765332 (733085)	13.8 (13.4)	14.0 (14.1)	62.0 (61.8)	98.7 (98.9)	100.0 (100.0)
Mathematics	Male	387012 (370068)	14.0 (13.6)	16.7 (16.8)	59.9 (59.7)	97.3 (97.2)	100.0 (100.0)
	Female	391846 (377101)	14.1 (13.7)	15.5 (14.9)	59.2 (59.2)	97.8 (97.6)	100.0 (100.0)
	Male & Female	778858 (747169)	14.0 (13.7)	16.1 (15.8)	59.6 (59.4)	97.6 (97.4)	100.0 (100.0)

Рисунок 1.1 – Результати GSCE з математики та англійської мови

1.2 Методи статистичних досліджень

Описова статистика охоплює методи опису статистичних даних, представлення їх у формі таблиць, розподілів. Певною мірою протиставляється статистичним висновкам, оскільки не робить висновків про генеральну сукупність на основі результатів дослідження частинних випадків, в той час як статистичний висновок припускає, що властивості та закономірності, виявлені під час дослідження об'єктів вибірки, характеризують генеральну сукупність [7].

Основні статистичні показники, які використовуються для опису набору даних – це міри центральної тенденції та міри розкиду даних стосовно центру розподілу[7].

Основними мірами центральної тенденції є:

- Середнє значення.
- Медіана.
- Мода.

Основними показниками розкиду стосовно центру розподілу є:

- стандартне відхилення (чи дисперсія),
- мінімальне та максимальне значення змінної,
- розмах, коефіцієнт варіації.

До інших показників найчастіше зараховують ексцес та коефіцієнт асиметрії.

Описова статистика використовує три основних методи подання даних [7]:

- табличне,
- графічне,
- за допомогою статистичних показників.

Найбільш простішим для розрахунків показником описової статистики є середнє арифметичне (формула 1.1)

$$\bar{X} = n^{-1} \sum_{i=1} x_i, \quad (1.1)$$

де n – кількість випадків;

x_i – значення i -го випадку.

Дисперсія є найпростішим розрахунковим показником для опису розсіяння у вибірці. Вона відображає ступінь відхилення спостережень від середнього арифметичного у вибірці (формула 1.2)[7]:

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}, \quad (1.2)$$

де x_i – значення i -го показника;

$\bar{X}$ – середнє арифметичне;

n – кількість випадків.

Квадратний корінь з дисперсії називають середньоквадратичним відхиленням або стандартним відхиленням середнього. Цей показник більш зручний для практичного використання, оскільки має однакові із середнім арифметичним одиниці виміру [7].

$$s = \sqrt{\frac{\sum (x_i - \bar{X})^2}{n - 1}}. \quad (1.3)$$

У графічному вигляді міри центральної тенденції та міри розсіяння в одній або кількох групах зручно представляти у вигляді графіків «ящик з вусами» (box-

whisker plot) (рисунок 2.1). Окремі елементи графіку («центр», «ящик», «вуса») відрізняються для різних видів розподілу (таблиця 1.1) [7].

Таблиця 1.1 – Значення графічних елементів у графіку «ящик з вусами»

№	Графічний елемент	Значення графічного елементу в залежності від виду розподілу	
		Нормальне	Відмінне від нормального
1	Вуса	Стандартна помилка	Розмах
2	Ящик	Стандартне відхилення	Інтерквартильний розмах
3	Центр	Середнє арифметичне	Медіана

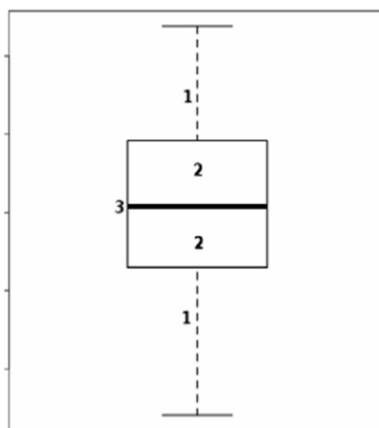


Рисунок 1.1 – Форма представлення у вигляді графіку «ящик з вусами»

1.3 Застосування статистичних критеріїв для аналізу даних

Для перевірки статистичних гіпотез можуть бути використані різні статистичні критерії. Їх поділяють на параметричні та непараметричні. Головною умовою застосування параметричних методів є нормальність розподілу змінних, що аналізуються [8].

1.3.1 Критерій Уелча

Критерій Уелча – тест оснований на розподілі Стьюдента та призначений для перевірки статистичної гіпотези про рівність математичних сподівань випадкових величин, маючих необов’язково рівні дисперсії [5].

Застосування критерія Уелча оснований на статистиці виду:

$$T = \frac{\sqrt{mn}(\bar{x} - \bar{y})}{\sqrt{ns_x^2 + ms_y^2}} \quad (1.4)$$

Правило прийняття рішень для критерію Уелча виглядає так:

- Якщо $|T| \leq \Phi\left(1 - \frac{\alpha}{2}\right)$, то гіпотеза однорідності математичних сподівань приймається на рівні значимості α .
- Якщо $|T| > \Phi\left(1 - \frac{\alpha}{2}\right)$, то гіпотеза однорідності математичних сподівань відхиляється на рівні значимості α .

1.3.2 Критерій Фішера

Критерій Фішера – статистичний критерій, тестова статистика якого при виконанні нульової гіпотези має розподіл Фішера (F-розподіл). Статистика тесту однак зводиться до відношення вибірових дисперсій (сум квадратів, поділених на ступеня свободи). Щоб статистика мала розподіл Фішера, необхідно щоб чисельник та знаменник були незалежними випадковими величинами та відповідні суми квадратів мали розподіл Хі-квадрат. Для цього потрібно, щоб дані мали

нормальний розподіл. Крім того, передбачається, що дисперсія випадкових величин, квадрати яких сумуються, однакова [8].

Тест проводиться шляхом порівняння значення статистики з критичним значенням відповідного розподілу Фішера, при заданому рівні значущості. Відомо, що якщо $F \sim F(m, n)$, то $1/F \sim F(n, m)$.

Крім того, квантілі розподілу Фішера мають властивість $F_{1-\alpha} = \frac{1}{F_{\alpha}}$. Тому зазвичай в чисельнику бере участь потенційно велика величина, у знаменнику менша і порівняння здійснюється з правою квантиллю розподілу. Тим не менш тест може бути двостороннім, і одностороннім. У першому випадку при рівні значимості α використовується квантиль $F_{\frac{\alpha}{2}}$, а при односторонньому тесті F_{α} .

Більш зручний спосіб перевірки гіпотез – за допомогою р-значення $p(F)$ – ймовірністю того, що випадкова величина з даним розподілом Фішера перевищить це значення статистики. Якщо $p(F)$ (для двозначного тесту – $2p(F)$) менше рівня значимості α , то нульова гіпотеза відхиляється, в протилежному випадку приймається [8].

В MS Excel для розрахунку рівня ймовірності виконання гіпотези рівності дисперсій застосовується функція ФТЕСТ(массив1;массив2) та процедура Пакету аналізу Двухвибірковий F-тест для дисперсій.

1.3.3 Критерій Ліллієфорса

Критерій Ліллієфорса є модифікацією критерію Колмогорова-Смірнова. Використовується для перевірки нульової гіпотези про те, що вибірка розподілена по нормальному закону для випадку, коли параметри нормального розподілу (математичне сподівання та дисперсія) невідомі. Критерій порівнює емпіричну функцію розподілу змінної з заданим теоретичним нормальним розподілом. На основі цього приймається рішення, чи є статистично значимим спостережуване

відхилення вибіркової функції розподілу від теоретичної. Якщо так, то нульова гіпотеза відхиляється [9].

В R для розрахунку критерію застосовується функція виду: `lillie.test(x)`

1.3.4 Критерій Андерсона-Дарлінга

Критерій Андерсона-Дарлінга призначений для перевірки простих гіпотез про належність аналізованої вибірки повністю відомому закону про узгодженість емпіричного розподілення $F_n(x)$ і теоретичного закону $F(x, \theta)$, тобто для перевірки гіпотези вигляду $H_0 : F_n(x) = F(x, \theta)$ з відомим вектором параметрів теоретичного закону [7].

В критерії Андерсона-Дарлінга використовується статистика виду:

$$S_{\Omega} = -n - 2 \sum_{i=1}^n \left\{ \frac{2i-1}{2n} \ln(F(x_i, \theta)) + (-F(x_i, \theta)) \right\}, \quad (1.5)$$

де n – об'єм вибірки

$x_1, x_2, \dots, x_n$, – впорядковані за зростанням елементи вибірки

В R для розрахунку критерію застосовується функція виду: `ad.test(x)`

1.3.5 Критерій Колмогорова-Смірнова

Критерій Колмогорова-Смірнова призначений для перевірки нормальності розподілу сукупностей кількісних даних. Для більшої достовірності отриманих даних обсяги аналізованих вибірок мають бути досить великими: $n > 50$ [10].

Критерій Колмогорова-Смірнова розраховується з допомогою спеціальних статистичних програм.

В основі лежить статистика виду:

$$D_n = \sup |F_n(x) - F(x)|, \quad (1.6)$$

де $\sup S$ – точна верхня грань множини S

F_n – функція розподілу досліджуваної сукупності

$F(x)$ – функція нормального розподілу

Значення ймовірності, що виводяться, засновані на припущенні, що середнє і стандартне відхилення нормального розподілу відомі і не оцінюються з даних.

Проте на практиці зазвичай параметри обчислюються безпосередньо з даних. У цьому випадку критерій нормальності включає складну гіпотезу ("наскільки ймовірно отримати D статистику даної або більшої значущості, яка залежить від середнього та стандартного відхилення, обчислених з даних"), і наводяться ймовірності Ліллієфорсу.

Якщо D статистика Колмогорова-Смірнова значуща ($p < 0,05$), то гіпотеза у тому, що відповідний розподіл нормальний, має бути відхилена.

Критерій однорідності Смірнова використовується для перевірки гіпотези про приналежність двох незалежних вибірок одному закону розподілу, тобто про те, що два емпіричних розподіли належать одному й тому закону.

Критерій Смірнова про перевірку гіпотези про однорідність двох емпіричних законів розподілу використовується найчастіше серед непараметричних критеріїв.

В R для розрахунку критерію застосовується функція виду: `ks.test(x,y)`

1.3.6 Критерій Уїлкоксона

Критерій Уїлкоксона використовується для оцінки відмінностей між двома рядами вимірювань, виконаних для однієї і тієї ж сукупності досліджуваних, але в різних умовах або в різний час. Критерій Уїлкоксона є непараметричним критерієм, тому не вимагає наявності нормального розподілу порівнюваних сукупностей. Число досліджуваних при використанні критерію має бути не менше ніж 5 [11].

Суть методу у тому, що зіставляються абсолютні величини вираженості зрушень у тому чи іншому напрямі. Для цього спочатку всі абсолютні величини зрушень ранжуються, а потім підсумовуються ранги. Якщо зрушення у той чи інший бік відбуваються випадково, то суми їх рангів виявляються приблизно рівні.

Якщо ж інтенсивність зрушень в один бік більша, то сума рангів абсолютних значень зрушень у протилежний бік буде значно нижчою, ніж це могло бути при випадкових змінах.

В R для розрахунку критерію застосовується функція виду: `wilcox.test(x,y)`

2 МЕТОДИКА ВИКОНАННЯ РОБОТИ

Для виконання дослідження були обрані програмні засоби для статистичної обробки даних: Microsoft Excel для роботи з електронними таблицями та мова програмування R.

Для аналізу даних була завантажена таблиця з сайту Українського центру оцінювання освіти результатів Зовнішнього незалежного оцінювання за 2021 рік [1].

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
OUTID	Birth	SexTypeN	RegName	AREANAM	TERNAME	RegTypeN	TerTypeN	ClassProfil	ClassLang	EONAME	EOTYPE	EORegNar	EOAreaNa	EOTerNan	EOParent	UMLTest	UMLTestS	UMLBall	
8a2abef7-3e975d01	2003	чоловіча	Дніпропетр	м.Кам'янськ	Дніпропетр	Студент з місто		Молодши українські	Кам'янський заклад фа	Дніпропетр	м.Кам'янськ	Дніпропетр	Департам	Українські	Зараховані			42	
	2003	чоловіча	Кіровоград	м.Кропив	Форттечні	Випускни місто		Технологі українські	Комуналь ліцей			Кіровоград	м.Кропив	Форттечні	Управлін	Українські	Зараховані	58	
0a1bc8b4-fe877949-f94cc47c-5c24db4a-8e2b9dcc-9398cdca-0d15e151-c1adaab1-b3e45578-6293eec3-7a9b5b1e-845b4055-bf4e0e1d-cec717c8-0f13ff3d-e	1977	чоловіча	м.Київ	м.Київ	Подільськ	Випускни місто		null	null	null	null	null	null	null	null	Українські	Зараховані	32	
	2003	чоловіча	Дніпропетр	Дніпропетр	м.Павлогр	Випускни місто		Технологі українські	Павлогр	середня з Дніпропетр	м.Павлогр	м.Павлогр	Відділ осв	Українські	Не подолі			25	
	2003	чоловіча	м.Київ	м.Київ	Подільськ	Випускни місто		Українські українські	Загальнос середня з	м.Київ	м.Київ	Пс	Подільськ	Управлін	Українські	Зараховані		45	
	2003	чоловіча	Дніпропетр	м.Дніпро	Чечелівськ	Випускни місто		Українські українські	Комуналь середня з	Дніпропетр	м.Дніпро	Чечелівськ	Управлін	Українські	Зараховані			30	
	2003	чоловіча	Луганська	Старобіль	м.Старобіль	Учень (сл місто)		Кваліфіко українські	Відокрем. заклад пр	Луганська	Старобіль	м.Старобіль	Міністерс	Українські	Не подолі			20	
	2004	чоловіча	Дніпропетр	Дніпропетр	м.Новомос	Студент з місто		Молодши українські	Новомос	заклад фа	Дніпропетр	м.Новомос	м.Новомос	Укоопспіл	Українські	Зараховані		65	
	2004	чоловіча	Закарпатс	Великобє	с.Ставне	Випускни селище	с.Інший(баг)	українські	Ставне	середня з	Закарпатс	Великобє	с.Ставне	Відділ осв	Українські	Зараховані		55	
	2004	чоловіча	Хмельниць	Хмельниць	м.Хмельн	Студент з місто		Молодши українські	Відокрем. заклад фа	Хмельниць	м.Хмельн	м.Хмельн	Міністерс	Українські	Зараховані			33	
	2003	чоловіча	Тернопіль	Заліщиць	м.Заліщи	Студент з місто		Молодши українські	Відокрем. заклад фа	Тернопіль	Заліщиць	м.Заліщи	Міністерс	Українські	Не подолі			18	
	2002	чоловіча	Одеська	с.м.Одеса	Суворова	Випускни місто		null	null	null	null	null	null	null	null	Українські	Зараховані	41	
	2002	чоловіча	Донецька	м.Маріуп	Централь	Випускни місто		null	null	null	null	null	null	null	null	Українські	Зараховані	50	
	2004	чоловіча	Волинські	Луцький	р-с мт	Торчи	Випускни селище	м	Універсал	українські	Опорний ліцей	Волинські	Луцький	р-с мт	Торчи	Гуманітар	Українські	Зараховані	73
	2004	чоловіча	Волинські	Волинські	м.Новово	Випускни місто		Історични українські	Нововоли ліцей	Волинські	м.Новово	м.Новово	Управлін	Українські	Зараховані			114	
	1997	чоловіча	Запорізьк	м.Запоріж	Олександр	Випускни місто		null	null	null	null	null	null	null	null	Українські	Зараховані	32	
	1992	чоловіча	Запорізьк	Кам'янськ	м.Кам'ян	Випускни місто		null	null	null	null	null	null	null	null	Українські	Зараховані	29	
Odata2021File Укр.мова(ж) Укр.мова(ж) Іср.Укр(ж) Іср.Укр(ж) Матем(ж) Матем(ж) Фізика(ж) Фізика(ж) ...																			

Рисунок 2.1 – Таблиця даних Українського центру оцінювання якості освіти

На основі даних Українського центру оцінювання освіти було створено таблиці з первинними (до переведення у шкалу 100-200) результатами ЗНО з усіх предметів.

За допомогою спеціальних функцій MS Excel розраховано основні показники елементів описової статистики результатів ЗНО з кожного предмету (окремо для чоловіків та жінок):

- Середнє значення.
- Медіана.
- Стандартне відхилення.
- Мода вибірки.
- Мінімум.
- Максимум.

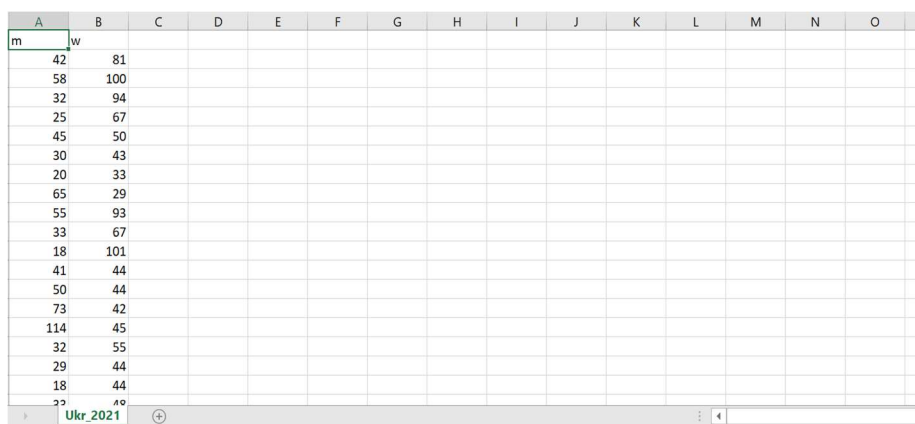
Для знаходження показника середнього балу застосовувалась функція `СРЗНАЧ(число1:число2)`, для розрахунку медіани функція `МЕДІАНА(число1:число2)`.

Наступною по важливості характеристикою вибірки є міра розсіювання елементів вибірки від середнього значення. Такою мірою є стандартне відхилення. Для визначення стандартного відхилення була застосована функція `СТАНДОТКЛОН`.

Також елементи описової статистики можна розрахувати за допомогою вбудованого в MS Excel пакету Аналіз даних.

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Перш ніж порівнювати результати ЗНО з української мови чоловіків та жінок створимо нову таблицю даних з балами чоловіків та жінок та збережемо її у форматі csv:



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
m	w													
42	81													
58	100													
32	94													
25	67													
45	50													
30	43													
20	33													
65	29													
55	93													
33	67													
18	101													
41	44													
50	44													
73	42													
114	45													
32	55													
29	44													
18	44													
22	40													

Рисунок 2.2 –Таблиця з балами чоловіків та жінок у форматі csv

Сформуємо таблицю даних за допомогою команд:

```
dat=read.table("Ukr_2021.csv", dec = ",", sep = ";", header = TRUE, nrow = 130580)
```

```
dat=as.matrix(dat)
```

```
datm = dat[1:91668,1] # Чоловіки
```

```
datw = dat[(1:130580),2] #Жінки
```

Перш, ніж перевіряти однорідність вибірок за статистичними критеріями, порівняємо їх за допомогою графічних засобів. Спочатку побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок. Для цього використовуємо такий набір команд:

```
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки", "Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

Перші дві команди використовуються для розрахунку значень емпіричної функції розподілу, відповідно, чоловіків та жінок. Наступна група команд впорядковує елементи аналізованих вибірок за зростанням. І заключна група команд використовується для побудови графіків ЕФР

Побудуємо діаграми типу “ящик з вусами”:

```
boxplot(datm, datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів", main="Українська мова", names=c("Чоловіки", "Жінки"))
old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)
```

Для більш докладного кількісного порівняння характеристик розподілів визначимо їх квантілі за допомогою команд:

```
quantile(datm); quantile(datw)
```

Для вибору адекватних критеріїв перевірки однорідності вибірок важливою є попередня перевірка їх підпорядкованості нормальному розподілу. Для цього можна використовувати різноманітні статистичні критерії. Але спочатку доцільно побудувати Q-Q або P-P діаграми. На цих діаграмах за віссю абсцис відкладають

значення квантилів (Q-Q) або функції розподілу (P-P), що відповідають теоретичній моделі розподілу, а за віссю ординат – відповідні значення для аналізованої вибірки.

Для побудови Q-Q діаграм використаємо такий набір команд:

```
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Українська мова, чоловіки", col = 3, col.main =
"brown", cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Українська мова, жінки", col = 4, col.main = "brown",
cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

За результатом побудови Q-Q діаграми можна буде казати про наявність відхилення даних від нормального розподілу.

Після цього додатково перевіряємо нульову гіпотезу за критеріями Андерсона-Дарлінга і Ліллієфорса:

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

Якщо б порівнювані вибірки відповідали нормальному розподілу, ми могли би перевіряти рівність їх середніх арифметичних за допомогою критерію Уелча – функція `t.test()`, та рівність дисперсій за допомогою критерію Фішера – функція `var.test()`. Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу;

```
ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності
між центрами розподілів)
```

3 ПРАКТИЧНА ЧАСТИНА

3.1 Аналіз даних результатів ЗНО з української мови

Розрахуємо основні показники описової статистики для чоловіків.

Для знаходження показника середнього балу застосуємо функцію **СРЗНАЧ** для діапазону значень S2:S91669 та отримаємо значення 49,7744033.

Також розрахуємо медіану за допомогою функції **МЕДИАНА**(S2:S91669) та отримаємо значення 46.

Моду за допомогою функції **МОДА**(S2:91669), отримаємо значення 34.

Наступною по важливістю характеристикою вибірки є міра розсіювання елементів вибірки від середнього значення. Такою мірою є стандартне відхилення. Для визначення стандартного відхилення необхідно застосувати функцію **СТАНДОТКЛОН**, отримаємо значення 21,639334.

Показники	Значення
Середній балл (до переведення в шкалу 100-200)	49,7744033
Медіана	46
Стандартне відхилення	21,639334
Мода вибірки	34
Vmin=	34
Vmax=	46
Відн. відхилення	26,0869565

Рисунок 3.1 – Результат розрахунків показників з української мови (чоловіки)

Також елементи описової статистики можна розрахувати за допомогою вбудованого в MS Excel пакету Аналіз даних. Виберемо діапазон даних (S2:91669) та запусимо процедуру Описова статистика.

Середнє	49,7744033
Стандартна помилка	0,07147185
Медіана	46
Мода	34
Стандартне відхилення	21,639334
Дисперсія вибірки	468,260777
Екссес	-0,2634231
Асимметричність	0,65673509
Інтервал	116
Мінімум	0
Максимум	116
Сумма	4562720
Загальна кількість	91668

Рисунок 3.2 - Результат розрахунків показників з української мови
(чоловіки) (2)

Теж саме розрахуємо і для жінок:

Показники	Значення
Середній бал (до переведення в шкалу 100-200	62,8994486
Медіана	61
Стандартне відхилення	23,5942141
Мода виборки	50
Vmin=	50
Vmax=	61
Відносне відх.	18,0327869

Рисунок 3.3 – Результат розрахунків показників з української мови (жінки)

Середнє	62,8994486
Стандартна помилка	0,06529308
Медіана	61
Мода	50
Стандартне відхилення	23,5942141
Дисперсія вибірки	556,686939
Екссес	-0,9433543
Асимметричність	0,16012506
Інтервал	116
Мінімум	0
Максимум	116
Сумма	8213410
Загальна кількість	130580

Рисунок 3.4 – Результат розрахунків показників з української мови (жінки)
(2)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Перш ніж порівнювати результати ЗНО з української мови чоловіків та жінок створимо нову таблицю даних з балами чоловіків та жінок та збережемо її у форматі csv :

m	w
42	81
58	100
32	94
25	67
45	50
30	43
20	33
65	29
55	93
33	67
18	101
41	44
50	44
73	42
114	45
32	55
29	44
18	44
22	40

Рисунок 3.5 – Таблиця даних з балами чоловіків та жінок

Сформуємо таблицю даних за допомогою команд:

```
dat=read.table("Ukr_2021.csv", dec = ",", sep = ";", header = TRUE, nrow
= 130580)
dat=as.matrix(dat)
datm = dat[1:91668,1] # Чоловіки
datw = dat[(1:130580),2] #Жінки
```

Перш , ніж перевіряти однорідність вибірок за статистичними критеріями, порівняємо їх за допомогою графічних засобів. Спочатку побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок. Для цього використовуємо такий набір команд:

```
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
```

```
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки", "Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

Перші дві команди використовуються для розрахунку значень емпіричної функції розподілу, відповідно, чоловіків та жінок. Наступна група команд впорядковує елементи аналізованих вибірок за зростанням. І заключна група команд використовується для побудови графіків ЕФР. Результати побудови наведено на рис. 3.6

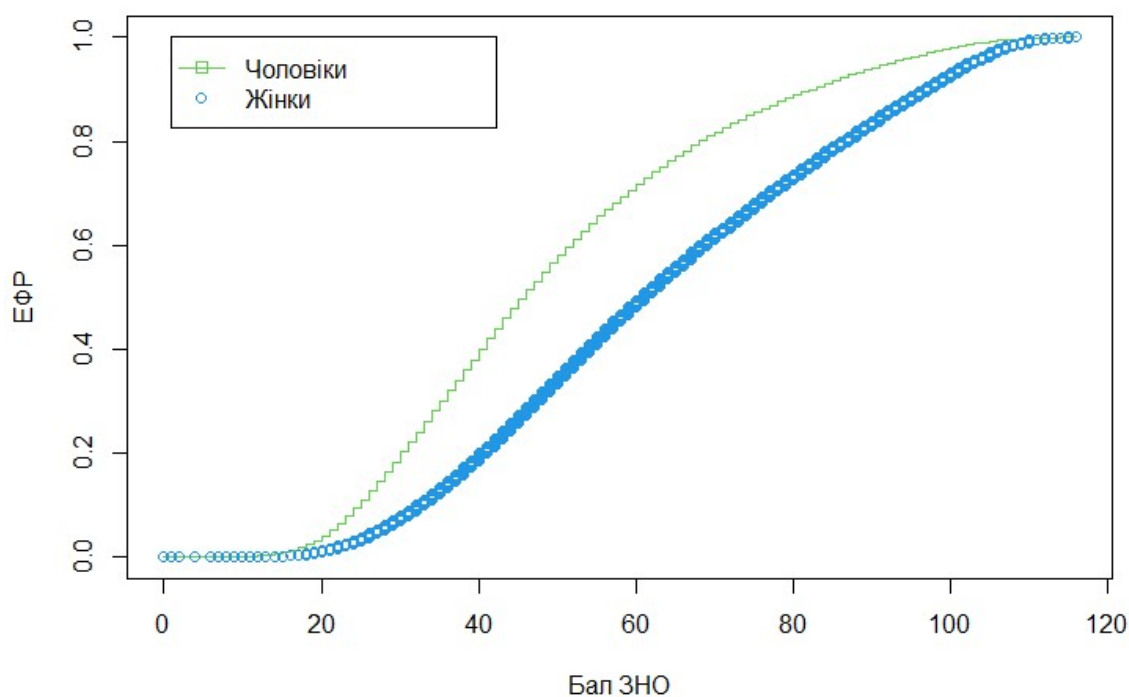


Рисунок 3.6 – Емпіричні функції розподілу показників чоловіків та жінок з української мови

З цього графіка можна казати про наявність істотної різниці між вибірками чоловіків та жінок, але є істотні області перетину між ними.

Побудуємо діаграми типу “ящик з вусами”:

```
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість балів",main="Українська мова", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
```

```
hist(datm, col = 3, xlab = "Кількість балів")
```

```
hist(datw, col = 4, xlab = "Кількість балів")
```

```
par(old.par)
```

Результати наведено на рис. 3.7

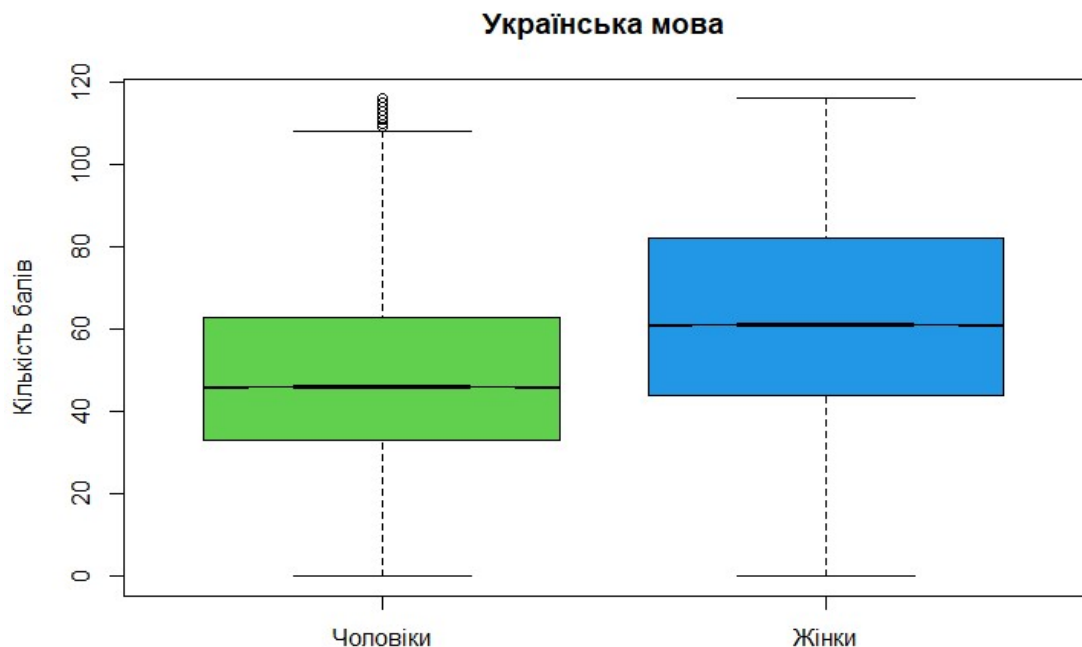


Рисунок 3.7 – Boxplot діаграми розподілу результатів чоловіків та жінок з української мови

Для більш докладного кількісного порівняння характеристик розподілів визначимо їх кuartилі за допомогою команд:

```
quantile(datm); quantile(datw)
```

При цьому отримуємо результати наведені у табл.1

Таблиця 3.1 – Квартилі розподілу даних результатів з української мови

	Мін	25%	50%	75%	Макс
Українська мова, чоловіки	0	33	46	63	116
Українська мова, жінки	0	44	61	82	116

З наведених даних бачимо, що практично всі квартилі для вибірки чоловіків є істотно нижчими, ніж для жінок. Втім частка чоловіків також має досить високі показники, максимальні результати з української мови для чоловіків та жінок однакові, 116 балів.

Для вибору адекватних критеріїв перевірки однорідності вибірок важливою є попередня перевірка їх підпорядкованості нормальному розподілу. Для цього можна використовувати різноманітні статистичні критерії. Але спочатку доцільно побудувати Q-Q або P-P діаграми. На цих діаграмах за віссю абсцис відкладають значення квантилів (Q-Q) або функції розподілу (P-P), що відповідають теоретичній моделі розподілу, а за віссю ординат – відповідні значення для аналізованої вибірки.

Для побудови Q-Q діаграм використаємо такий набір команд:

```
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Українська мова, чоловіки", col = 3, col.main =
"brown",
sex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Українська мова, жінки", col = 4, col.main =
"brown",
sex.main = 1.2)
qqline(datw, col = 2)
```

```
par(old.par)
```

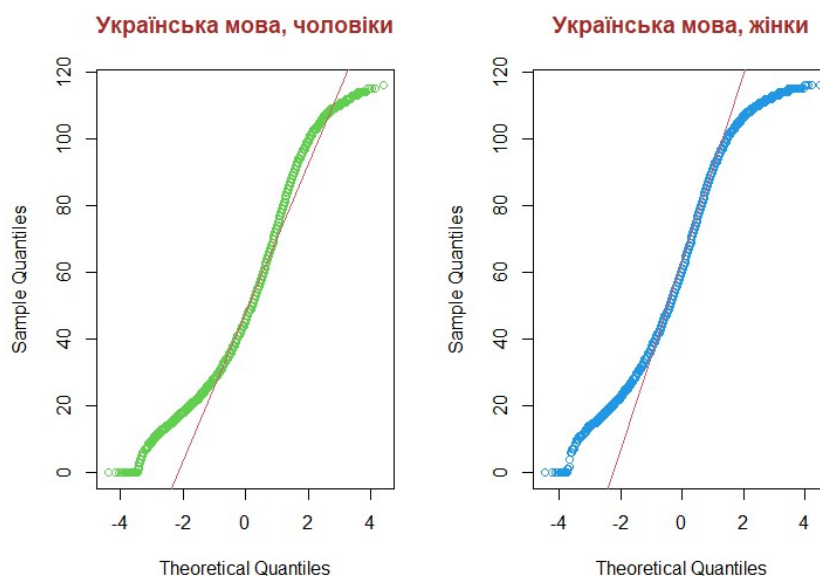


Рисунок 3.8 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Тепер перевіримо відповідність даних нормальному розподілу за допомогою більш строгих статистичних критеріїв: Шапіро-Уїлка, Андерсона-Дарлінга, Ліллієфорса.

Критерій Шапіро-Уїлка в даному випадку застосовувати не можна, оскільки обсяг вибірки перевищує 5000 елементів.

Для перевірки гіпотези про нормальний розподіл за іншими непараметричними критеріями підключаємо бібліотеку `nortest`.

Після цього додатково перевіряємо нульову гіпотезу за критеріями Андерсона-Дарлінга і Ліллієфорса:

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
```

```
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

Таблиця 3.2 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Ліллієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Укр.мова, чоловіки	0.083712	$< 2.2 \cdot 10^{-16}$	1197.5	$< 2.2 \cdot 10^{-16}$
Укр.мова, жінки	0.056998	$< 2.2 \cdot 10^{-16}$	859.99	$< 2.2 \cdot 10^{-16}$

Всі наведені у табл 2. результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Якщо б порівнювані вибірки відповідали нормальному розподілу, ми могли би перевіряти рівність їх середніх арифметичних за допомогою критерію Уелча – функція `t.test()`, та рівність дисперсій за допомогою критерію Фішера – функція `var.test()`. Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу; критерій Ансарі – Бредлі, що оцінює значущість різниці дисперсій:

```
ks.test(datm,datw) #Основний тест для подібних даних
```

```
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між центрами розподілів)
```

```
ansari.test(datm,datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності дисперсій розподілів)
```

Таблиця 3.3 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смірнова		Уїлконсона	
	D	p-value	W	p-value
Укр.мова	0.23318	$< 2.2 \cdot 10^{-16}$	4056221866	$< 2.2 \cdot 10^{-16}$

3.2 Історія України

Розрахуємо основні показники елементів описової статистики результатів ЗНО з історії України (окремо для чоловіків та жінок):

Показники	Значення
Середній бал (до переведення в шкалу 100-200)	37,2259684
Медіана	34
Стандартне відхилення	14,50526429
Мода вибірки	30
Vmin=	30
Vmax=	34
Відносне відхилення	11,76470588

Рисунок 3.9 – Результат розрахунків показників з історії України (чоловіки)

Середнє	41,01298605
Стандартна помилка	0,04815239
Медіана	37
Мода	32
Стандартне відхилення	15,84989161
Дисперсія вибірки	251,2190642
Ексцес	0,611157726
Асиметричність	1,07311629
Інтервал	94
Мінімум	0
Максимум	94
Сумма	4443634
Загальна кількість	108347

Рисунок 3.10 – Результат розрахунків показників з історії України (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

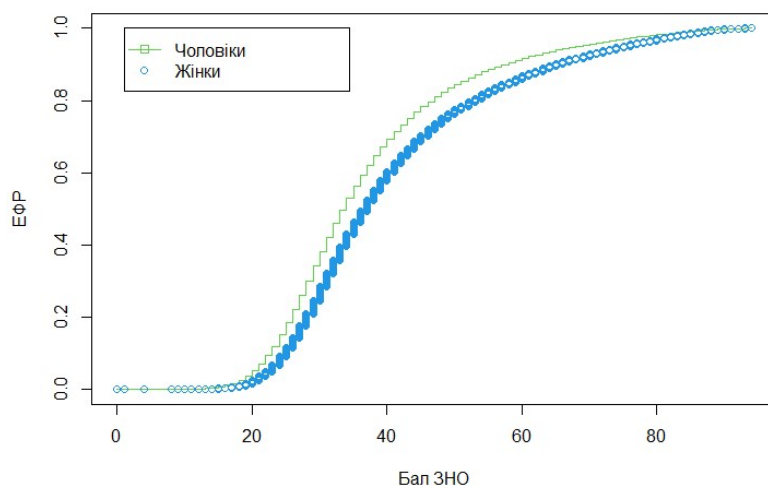


Рисунок 3.11 – Емпіричні функції розподілу показників з історії України

Побудуємо діаграми типу “ящик з вусами”:

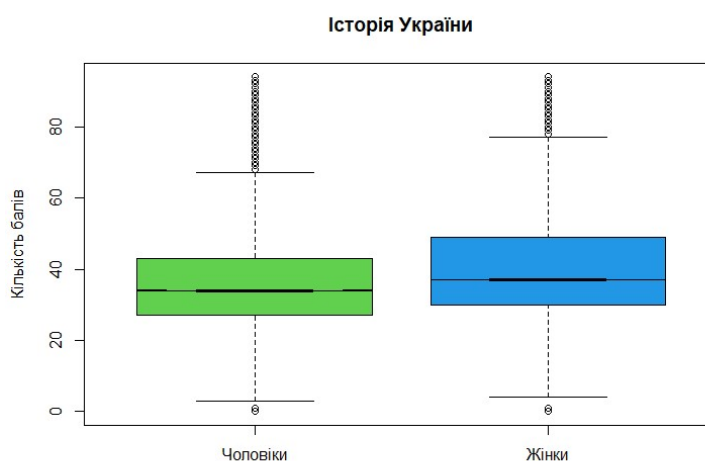


Рисунок 3.12 – Boxplot діаграми розподілу результатів з історії України

Отримаємо квартилі розподілу даних результатів з історії України:

Таблиця 3.4 – Квартилі розподілу даних результатів з історії України

	Мін	25%	50%	75%	Макс
Історія України, чоловіки	0	27	34	43	94
Історія України, жінки	0	30	37	49	94

Побудуємо Q-Q діаграми:

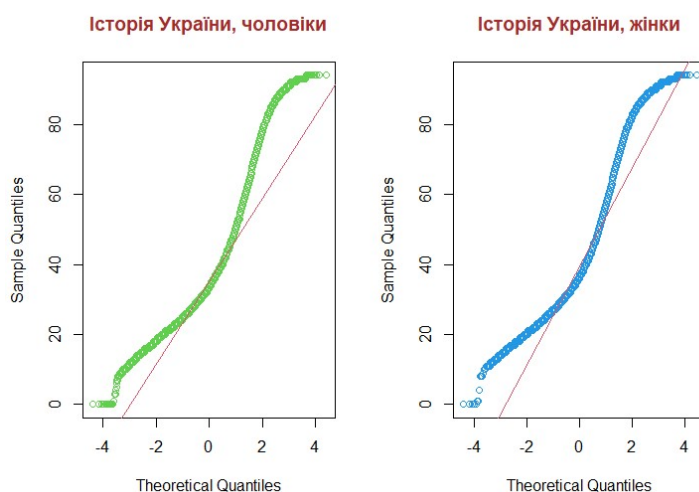


Рисунок 3.13 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.5 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Історія України, чоловіки	0.12761	$< 2.2 \cdot 10^{-16}$	2648.5	$< 2.2 \cdot 10^{-16}$
Історія України, жінки	0.012851	$< 2.2 \cdot 10^{-16}$	3145.1	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.6 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Іст. Укр	0.10207	$< 2.2 \cdot 10^{-16}$	4197671985	$< 2.2 \cdot 10^{-16}$

3.3 Математика

Розрахуємо основні показники елементів описової статистики результатів ЗНО з математики (окремо для чоловіків та жінок):

Середнє	19,641459
Стандартна помилка	0,04169991
Медіана	14
Мода	8
Стандартне відхилення	14,7619465
Дисперсія вибірки	217,915064
Ексцес	0,43978375
Асиметричність	1,15122504
Інтервал	67
Мінімум	0
Максимум	67
Сума	2461448
Загальна кількість	125319

Рисунок 3.14 – Результат розрахунків показників з математики (чоловіки)

Середнє	19,566809
Стандартна помилка	0,0411661
Медіана	14
Мода	8
Стандартне відхилення	14,193822
Дисперсія вибірки	201,46458
Ексцес	0,2277122
Асиметричність	1,0544567
Інтервал	67
Мінімум	0
Максимум	67
Сума	2326161
Загальна кількість	118883

Рисунок 3.15 – Результат розрахунків показників з математики (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

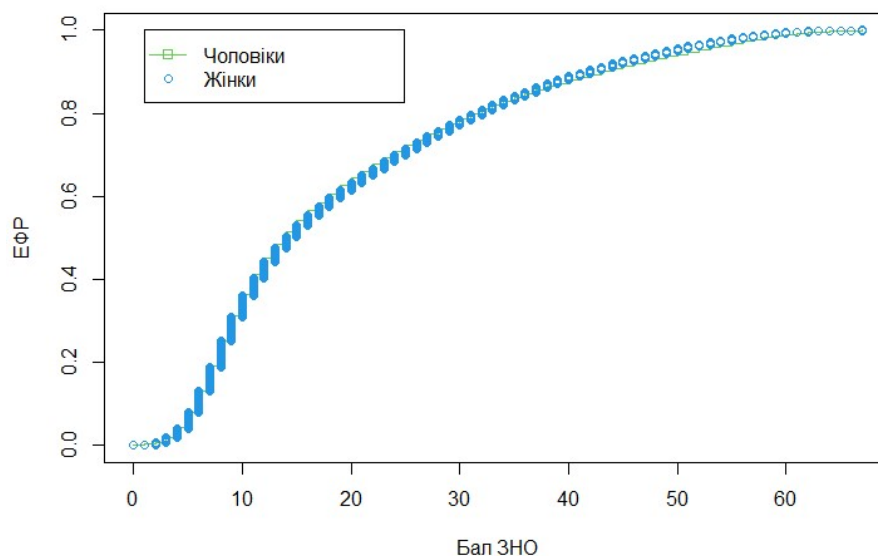


Рисунок 3.16 – Емпіричні функції розподілу показників з математики

Побудуємо діаграми типу “ящик з вусами”:

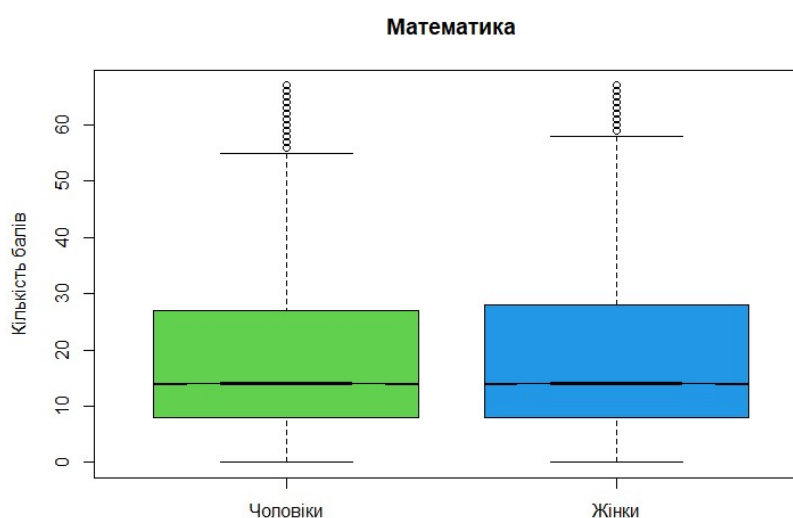


Рисунок 3.17 – Boxplot діаграми розподілу результатів з математики

Отримаємо квартилі розподілу даних результатів з математики:

Таблиця 3.7 – Квартилі розподілу даних результатів з математики

	Мін	25%	50%	75%	Макс
Математика, чоловіки	0	8	14	27	67
Математика, жінки	0	8	14	28	67

Побудуємо Q-Q діаграми:

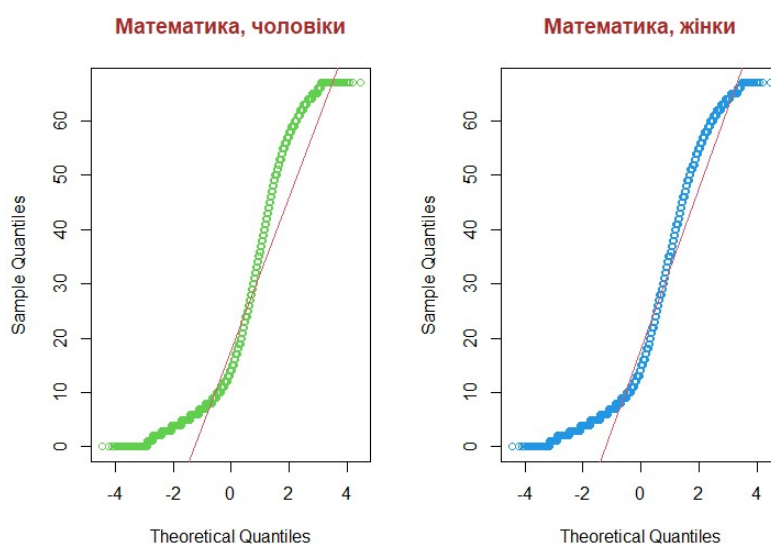


Рисунок 3.18 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.8 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Математика, чоловіки	0.16412	$< 2.2 \cdot 10^{-16}$	5862.5	$< 2.2 \cdot 10^{-16}$
Математика, жінки	0.15673	$< 2.2 \cdot 10^{-16}$	4890.2	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.9. Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Математика	0.014279	$3.14 \cdot 10^{-11}$	7406633701	0.01455

3.4 Фізика

Розрахуємо основні показники елементів описової статистики результатів ЗНО з фізики (окремо для чоловіків та жінок):

Середнє	24,25003786
Стандартна помилка	0,086716758
Медіана	21
Мода	17
Стандартне відхилення	12,20490232
Дисперсія вибірки	148,9596407
Екссес	0,812656956
Асимметричність	1,151297958
Інтервал	64
Мінімум	0
Максимум	64
Сумма	480369
Загальна кількість	19809

Рисунок 3.19 – Результат розрахунків показників з фізики (чоловіки)

Середнє	24,96021
Стандартна помилка	0,225436
Медіана	21
Мода	15
Стандартне відхилення	13,13152
Дисперсія вибірки	172,4368
Екссес	-0,0143
Асимметричність	0,905659
Інтервал	64
Мінімум	0
Максимум	64
Сумма	84690
Загальна кількість	3393

Рисунок 3.20 – Результат розрахунків показників з фізики (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

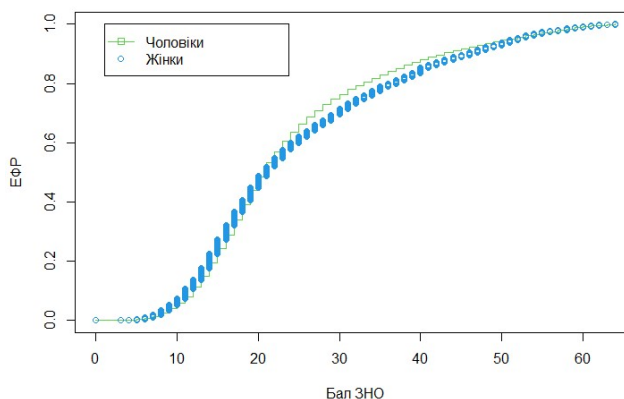


Рисунок 3.21 – Емпіричні функції розподілу показників з фізики

Побудуємо діаграми типу “ящик з вусами”:

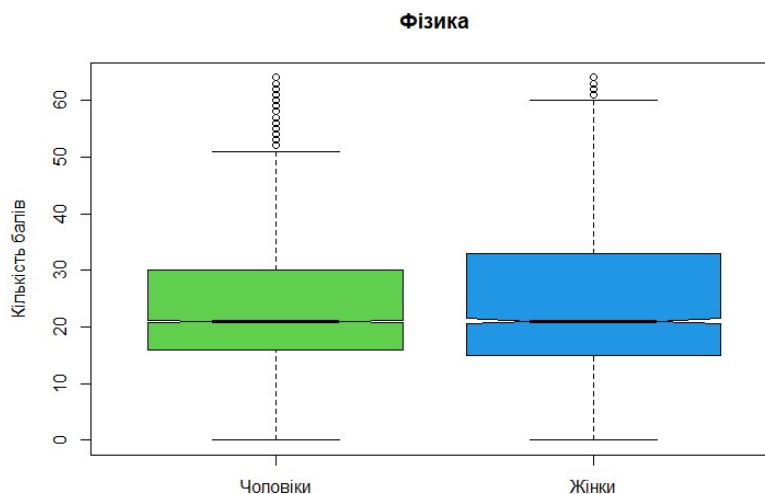


Рисунок 3.22 Boxplot діаграми розподілу результатів з фізики

Отримаємо кватилі розподілу даних результатів з фізики:

Таблиця 3.10 – Кватилі розподілу даних результатів з фізики

	Мін	25%	50%	75%	Макс
Фізика, чоловіки	0	16	21	30	64
Фізика, жінки	0	15	21	33	64

Побудуємо Q-Q діаграми:

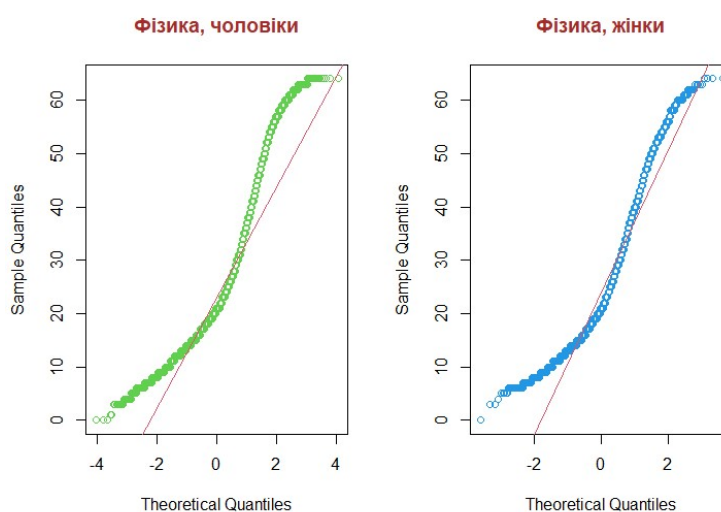


Рисунок 3.23 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.11 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Ліллєфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Фізика, чоловіки	0.14607	$< 2.2 \cdot 10^{-16}$	655.73	$< 2.2 \cdot 10^{-16}$
Фізика, жінки	0.13782	$< 2.2 \cdot 10^{-16}$	97.745	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.12 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Фізика	0.054279	$7.727 \cdot 10^{-8}$	33293766	0.3862

3.5 Географія

Розрахуємо основні показники елементів описової статистики результатів ЗНО з географії (окремо для чоловіків та жінок):

Середнє	44,14532
Стандартна помилка	0,061495
Медіана	43
Мода	41
Стандартне відхилення	14,87967
Дисперсія вибірки	221,4046
Екссес	-0,32473
Асимметричність	0,397836
Інтервал	89
Мінімум	0
Максимум	89
Сумма	2584620
Загальна кількість	58548

Рисунок 3.24 – Результат розрахунків показників з географії (чоловіки)

Середнє	45,33578
Стандартна помилка	0,063587
Медіана	43
Мода	40
Стандартне відхилення	14,8538
Дисперсія вибірки	220,6353
Екссес	-0,37448
Асимметричність	0,458485
Інтервал	89
Мінімум	0
Максимум	89
Сумма	2473883
Загальна кількість	54568

Рисунок 3.25 – Результат розрахунків показників з географії (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

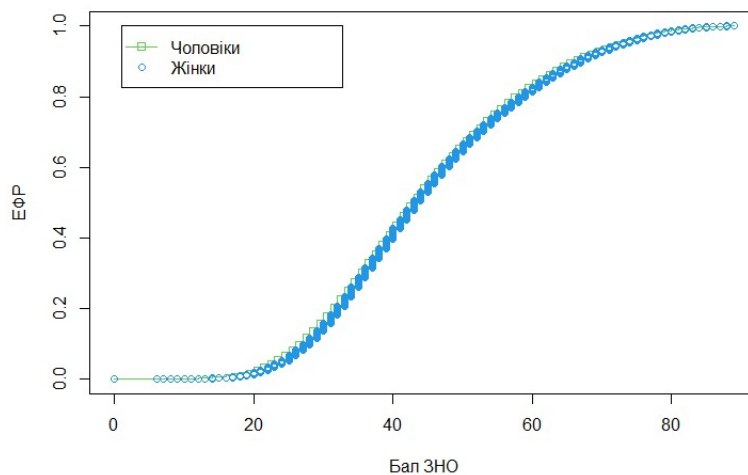


Рисунок 3.26 – Емпіричні функції розподілу показників з географії

Побудуємо діаграми типу “ящик з вусами”:

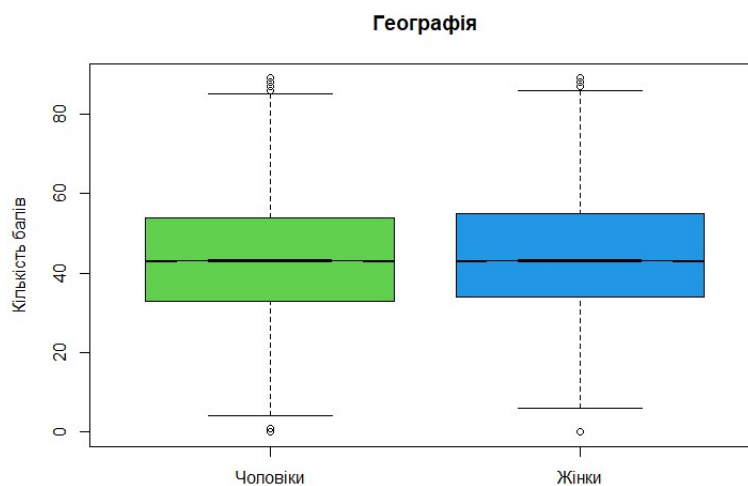


Рисунок 3.27 – Boxplot діаграми розподілу результатів з географії

Отримаємо квартилі розподілу даних результатів з географії:

Таблиця 3.13 – Квартилі розподілу даних результатів з географії

	Мін	25%	50%	75%	Макс
Географія, чоловіки	0	33	43	54	89
Географія, жінки	0	34	43	55	89

Побудуємо Q-Q діаграми:

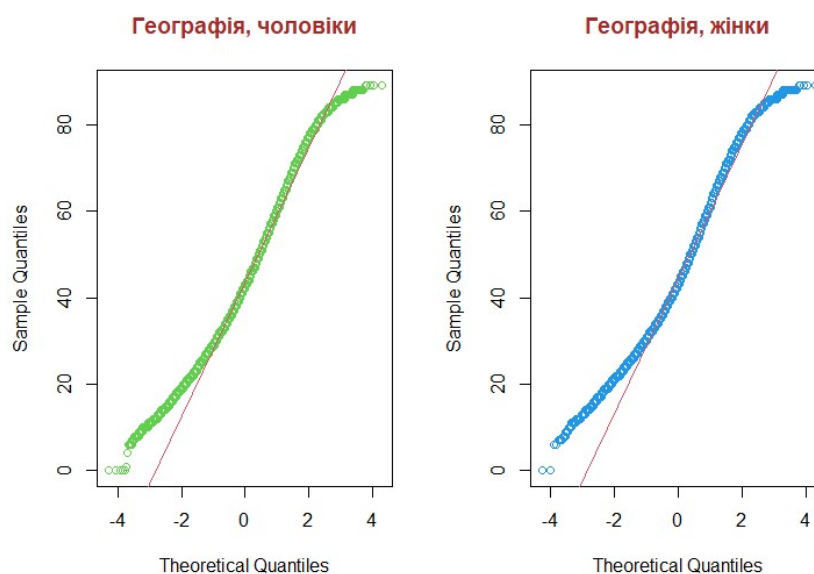


Рисунок 3.28 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.14 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Географія, чоловіки	0.057232	$< 2.2 \cdot 10^{-16}$	258.22	$< 2.2 \cdot 10^{-16}$
Географія, жінки	0.068176	$< 2.2 \cdot 10^{-16}$	369.71	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.15 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Географія	0.030197	$< 2.2 \cdot 10^{-16}$	$1.532 \cdot 10^9$	$< 2.2 \cdot 10^{-16}$

3.6 Англійська мова

Розрахуємо основні показники елементів описової статистики результатів ЗНО з англійської мови (окремо для чоловіків та жінок):

Середнє	38,42655
Стандартна помилка	0,065917
Медіана	38
Мода	18
Стандартне відхилення	16,5234
Дисперсія вибірки	273,0226
Екссес	-1,16637
Асиметричність	0,109084
Інтервал	72
Мінімум	0
Максимум	72
Сумма	2414571
Загальна кількість	62836

Рисунок 3.29 – Результат розрахунків показників з англійської мови (чоловіки)

Середнє	40,88252
Стандартна помилка	0,06633
Медіана	41
Мода	59
Стандартне відхилення	16,89514
Дисперсія вибірки	285,4457
Екссес	-1,18411
Асиметричність	-0,00861
Інтервал	72
Мінімум	0
Максимум	72
Сумма	2652376
Загальна кількість	64878

Рисунок 3.30 – Результат розрахунків показників з англійської мови (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

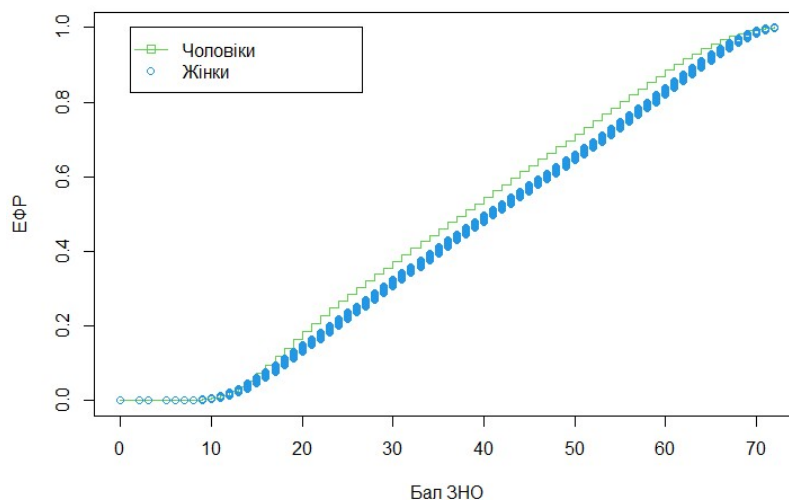


Рисунок 3.31 – Емпіричні функції розподілу показників з англійської мови

Побудуємо діаграми типу “ящик з вусами”:

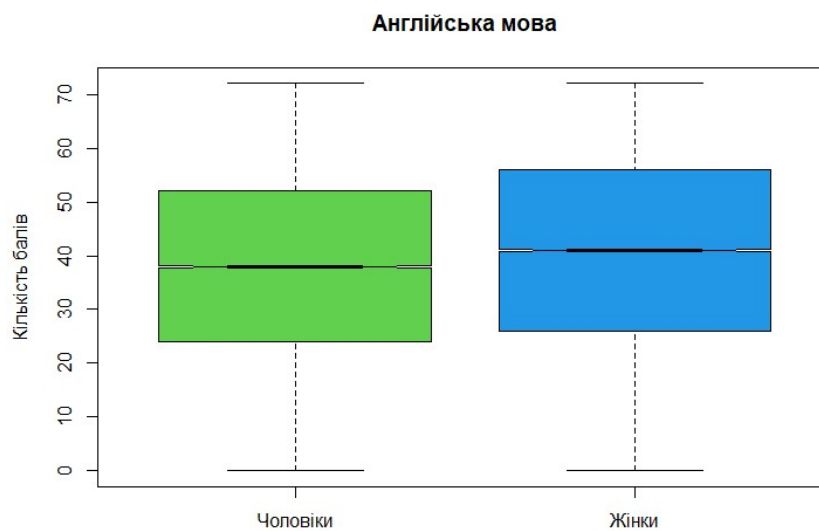


Рисунок 3.32 Boxplot діаграми розподілу результатів з англійської мови

Отримаємо квартилі розподілу даних результатів з англійської мови:

Таблиця 3.16 – Квартилі розподілу даних результатів з англійської мови

	Мін	25%	50%	75%	Макс
Англ.мова, чоловіки	0	24	38	52	72
Англ.мова, жінки	0	26	41	56	72

Побудуємо Q-Q діаграми:

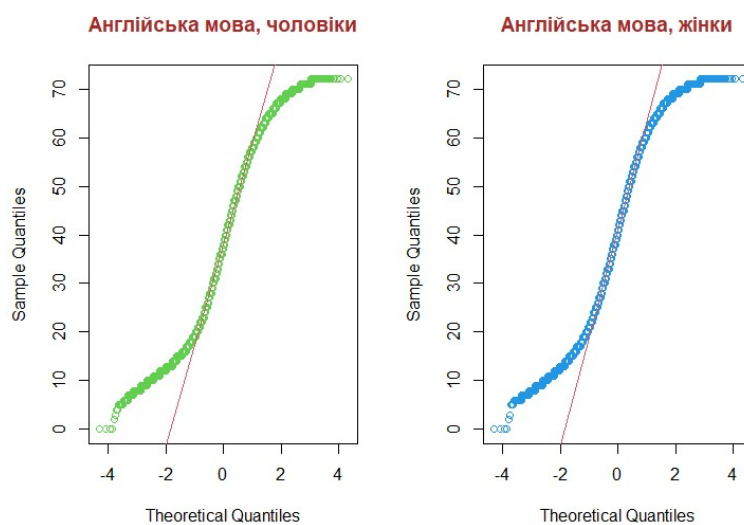


Рисунок 3.33 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.17 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Англ.мова, чоловіки	0.076669	$< 2.2 \cdot 10^{-16}$	725.69	$< 2.2 \cdot 10^{-16}$
Англ.мова, жінки	0.070049	$< 2.2 \cdot 10^{-16}$	717.4	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.18 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Англ.мова	0.056653	$< 2.2 \cdot 10^{-16}$	1866408422	$< 2.2 \cdot 10^{-16}$

3.7 Німецька мова

Розрахуємо основні показники елементів описової статистики результатів ЗНО з німецької мови (окремо для чоловіків та жінок):

Середнє	39,74641
Стандартна помилка	0,685968
Медіана	41
Мода	60
Стандартне відхилення	17,17662
Дисперсія вибірки	295,0362
Екссес	-1,2207
Асимметричність	-0,1143
Інтервал	71
Мінімум	0
Максимум	71
Сумма	24921
Загальна кількість	627

Рисунок 3.34 – Результат розрахунків показників з німецької мови (чоловіки)

Середнє	41,9808
Стандартна помилка	0,489194
Медіана	43
Мода	60
Стандартне відхилення	16,93204
Дисперсія вибірки	286,6939
Екссес	-1,19666
Асимметричність	-0,18947
Інтервал	63
Мінімум	8
Максимум	71
Сумма	50293
Загальна кількість	1198

Рисунок 3.35 – Результат розрахунків показників з німецької мови (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

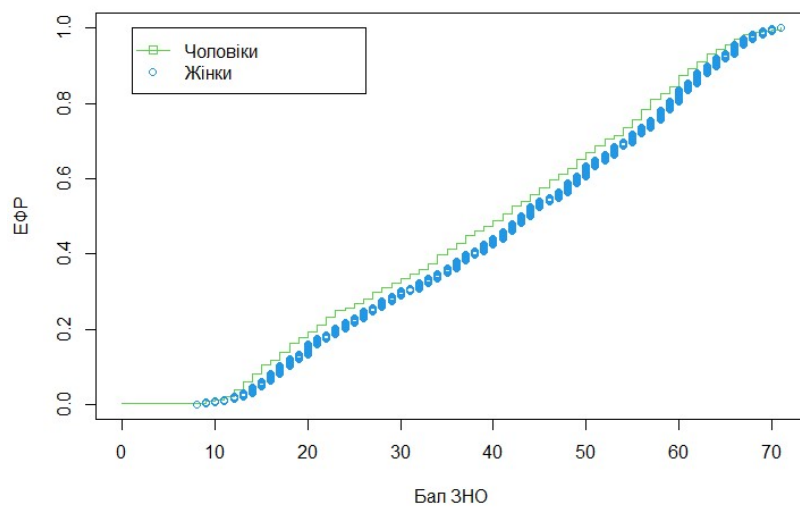


Рисунок 3.36 – Емпіричні функції розподілу показників з німецької мови

Побудуємо діаграми типу “ящик з вусами”:

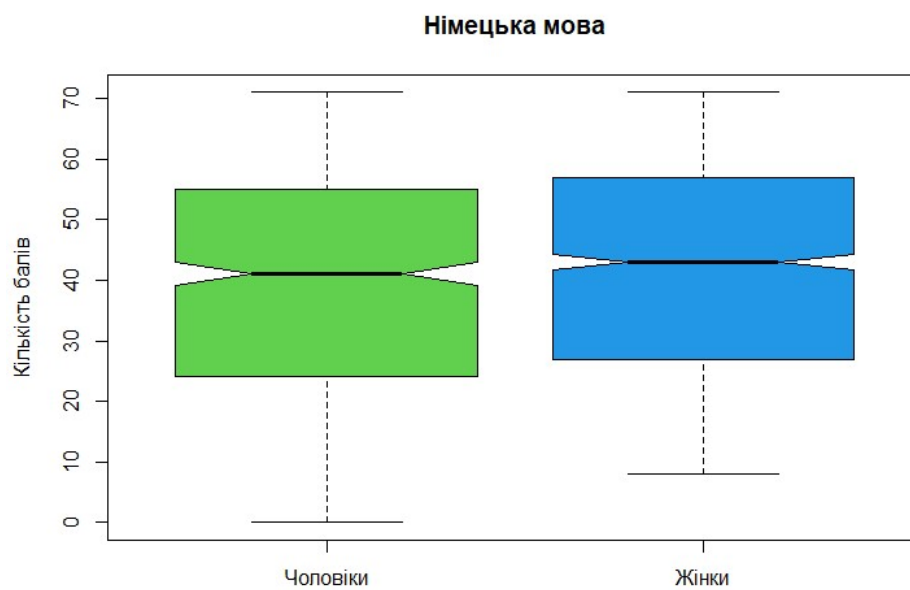


Рисунок 3.37 – Boxplot діаграми розподілу результатів з німецької мови

Отримаємо квартилі розподілу даних результатів з німецької мови:

Таблиця 3.19 – Квартилі розподілу даних результатів з німецької мови

	Мін	25%	50%	75%	Макс
Нім.мова, чоловіки	0	24	41	55	71
Нім.мова, жінки	8	27	43	57	71

Побудуємо Q-Q діаграми:

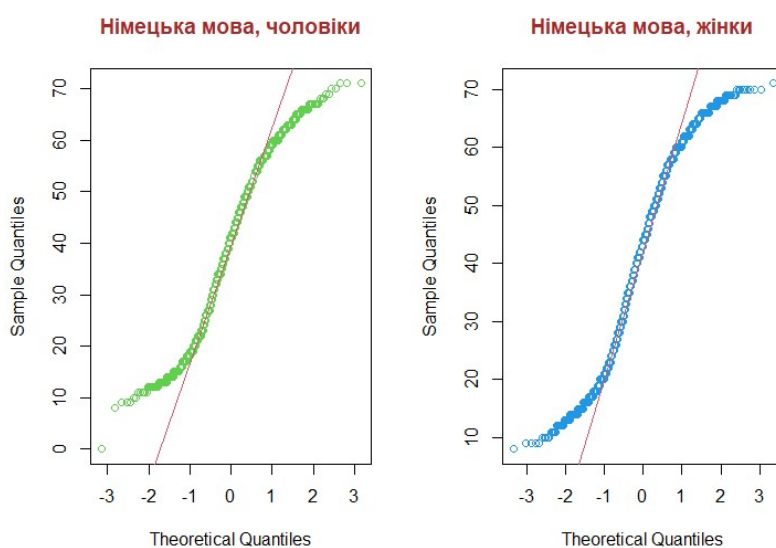


Рисунок 3.38 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.20 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Нім.мова, чоловіки	0.084012	$2.333 \cdot 10^{-11}$	8.946	$< 2.2 \cdot 10^{-16}$
Нім.мова, жінки	0.082031	$< 2.2 \cdot 10^{-16}$	17.363	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.21 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смірнова		Уїлконсона	
	D	p-value	W	p-value
Нім.мова	0.056377	0.1461	346899	0.007308

3.8 Французька мова

Розрахуємо основні показники елементів описової статистики результатів ЗНО з французької мови (окремо для чоловіків та жінок):

Середнє	42,84848
Стандартна помилка	1,65776
Медіана	43
Мода	62
Стандартне відхилення	16,4945
Дисперсія вибірки	272,0686
Екссес	-1,03322
Асимметричність	-0,28738
Інтервал	58
Мінімум	11
Максимум	69
Сумма	4242
Загальна кількість	99

Рисунок 3.39 – Результат розрахунків показників з французької мови (чоловіки)

Среднее	43,94140625
Стандартная ошибка	0,989421074
Медиана	47
Мода	58
Стандартное откл.	15,83073718
Дисперсия выбор	250,6122396
Экссес	-0,99006448
Асимметричность	-0,40414091
Интервал	58
Минимум	11
Максимум	69
Сумма	11249
Счет	256

Рисунок 3.40 – Результат розрахунків показників з французької мови (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

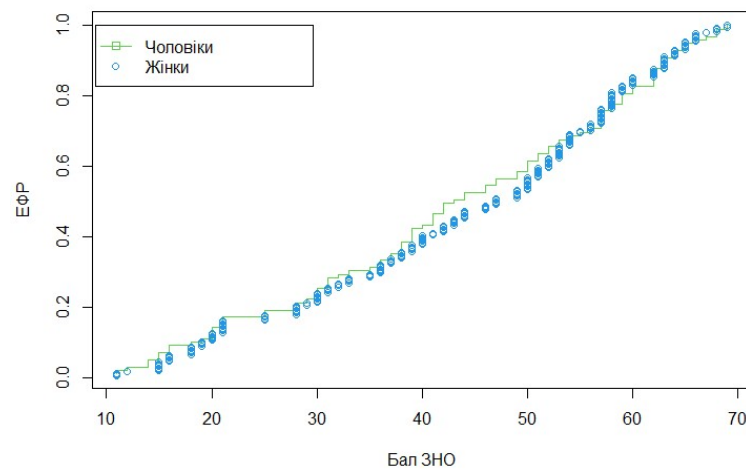


Рисунок 3.41 – Емпіричні функції розподілу показників з французької мови

Побудуємо діаграми типу “ящик з вусами”:

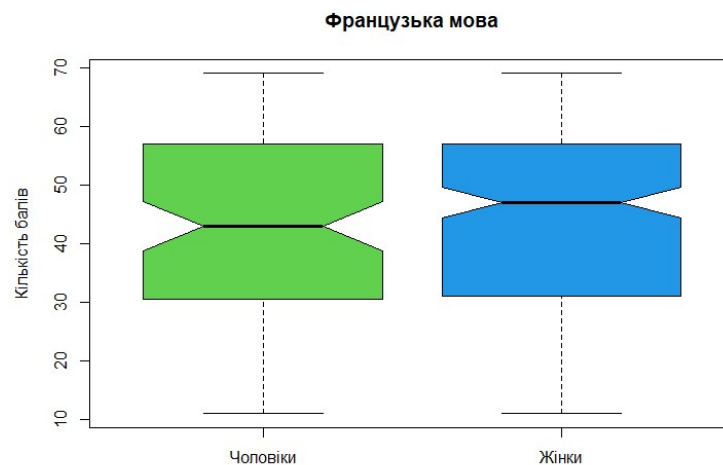


Рисунок 3.42 Voxplot діаграми розподілу результатів з німецької мови

Отримаємо квартилі розподілу даних результатів з французької мови:

Таблиця 3.22 – Квартилі розподілу даних результатів з французької мови

	Мін	25%	50%	75%	Макс
Фр.мова, чоловіки	11	30.5	43	57	69
Фр.мова, жінки	11	31	47	57	69

Побудуємо Q-Q діаграми:

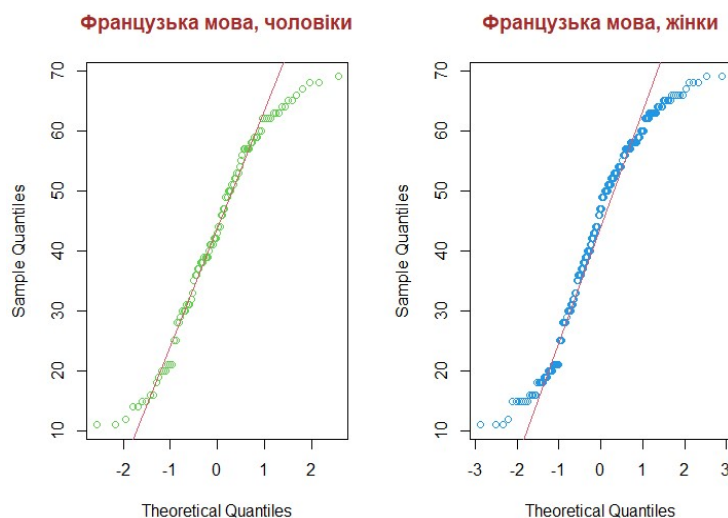


Рисунок 3.43 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.23 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Ліллєфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Фр.мова, чоловіки	0.09747	0.02145	1.2593	0.00268
Фр.мова, жінки	0.11778	$< 3.54 \cdot 10^{-9}$	4.4066	$< 5.77 \cdot 10^{-11}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.24 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Фр.мова	0.065262	0.9214	12215	0.5984

3.9 Біологія

Розрахуємо основні показники елементів описової статистики результатів ЗНО з біології (окремо для чоловіків та жінок):

Середнє	40,60762
Стандартна помилка	0,075816
Медіана	39
Мода	35
Стандартне відхилення	13,00203
Дисперсія вибірки	169,0527
Екссес	0,443302
Асимметричність	0,730883
Інтервал	82
Мінімум	0
Максимум	82
Сума	1194270
Загальна кількість	29410

Рисунок 3.44 – Результат розрахунків показників з біології (чоловіки)

Середнє	43,5858
Стандартна помилка	0,057264
Медіана	41
Мода	36
Стандартне відхилення	13,05826
Дисперсія вибірки	170,5181
Ексцес	0,034189
Асиметричність	0,680425
Інтервал	82
Мінімум	0
Максимум	82
Сумма	2266505
Загальна кількість	52001

Рисунок 3.45 – Результат розрахунків показників з біології (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

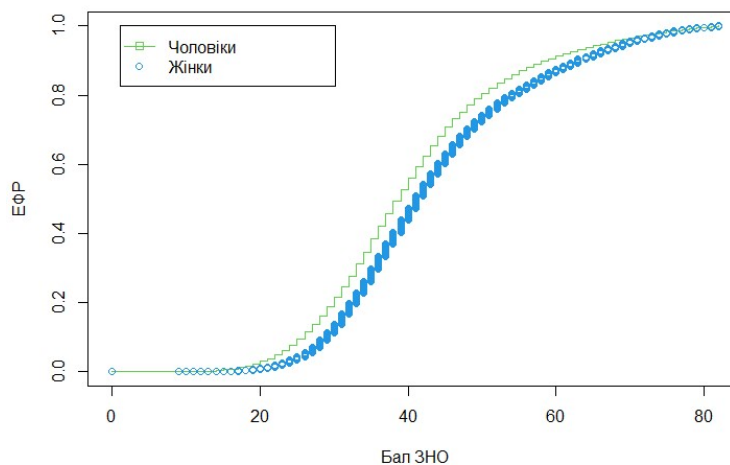


Рисунок 3.46 – Емпіричні функції розподілу показників з біології

Побудуємо діаграми типу “ящик з вусами”:

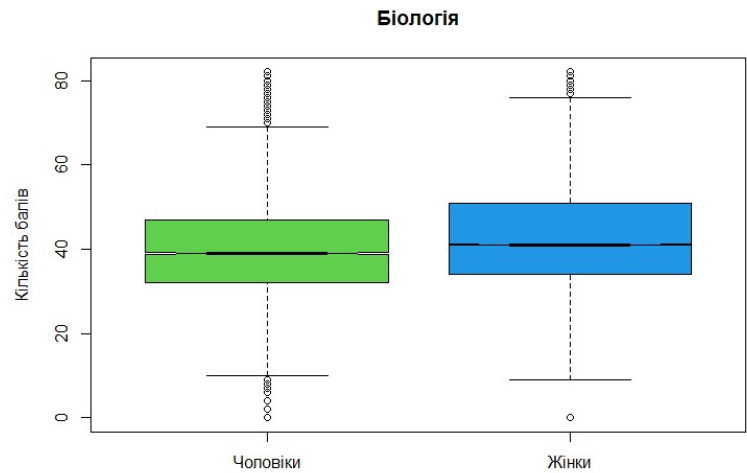


Рисунок 3.47 – Boxplot діаграми розподілу результатів з біології

Отримаємо квантілі розподілу даних результатів з біології:

Таблиця 3.25 – Квантілі розподілу даних результатів з біології

	Мін	25%	50%	75%	Макс
Біологія, чоловіки	0	32	39	47	82
Біологія, жінки	0	34	41	51	82

Побудуємо Q-Q діаграми:

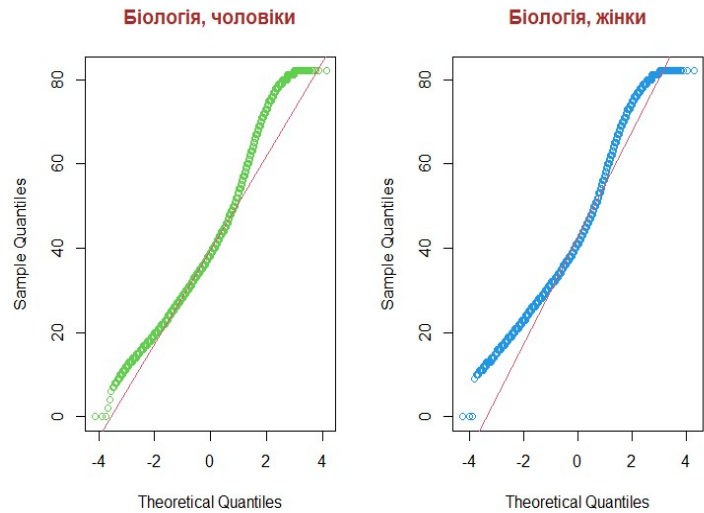


Рисунок 3.48 – Q-Q діаграми розподілу показників

За результатом побудови Q-Q діаграми можна казати про істотні відхилення даних від нормального розподілу.

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.26 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Лілієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Біологія, чоловіки	0.0814	$< 2.2 \cdot 10^{-16}$	299.06	$< 2.2 \cdot 10^{-16}$
Біологія, жінки	0.90245	$< 2.2 \cdot 10^{-16}$	640.43	$< 2.2 \cdot 10^{-16}$

Всі наведені результати свідчать про необхідність відхилення нульової гіпотези про відповідність даних нормальному розподілу, оскільки за будь-якого вибору рівня значущості (0,1, 0,05, 0,01) p-value у будь-якому випадку для всіх вибірок буде значно меншим.

Досліджувані вибірки не підпорядковуються нормальному розподілу, тому для перевірки їх однорідності потрібно буде використовувати непараметричні критерії перевірки гіпотези про однорідність. Для цього використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.27 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Біологія	0.0895	$< 2.2 \cdot 10^{-16}$	662896263	$< 2.2 \cdot 10^{-16}$

3.10 Іспанська мова

Розрахуємо основні показники елементів описової статистики результатів ЗНО з іспанської мови (окремо для чоловіків та жінок):

Среднее	44,3333333
Стандартная оши	2,92144875
Медиана	48
Мода	56
Стандартное откл	18,2444416
Дисперсия выбор	332,859649
Эксцесс	-0,93652212
Асимметричност	-0,37351556
Интервал	62
Минимум	9
Максимум	71
Сумма	1729
Счет	39

Рисунок 3.49 – Результат розрахунків показників з іспанської мови (чоловіки)

Среднее	44,16216216
Стандартная ош	1,960316399
Медиана	51
Мода	65
Стандартное отк	20,65321482
Дисперсия выбо	426,5552826
Эксцесс	-1,43908377
Асимметричнос	-0,323328683
Интервал	62
Минимум	10
Максимум	72
Сумма	4902
Счет	111

Рисунок 3.50 – Результат розрахунків показників з іспанської мови (жінки)

Далі для продовження аналізу вибірок та порівняння результатів будемо застосовувати мову програмування R для статистичної обробки даних та роботою з графікою.

Побудуємо емпіричні функції розподілу (ЕФР) сформованих вибірок:

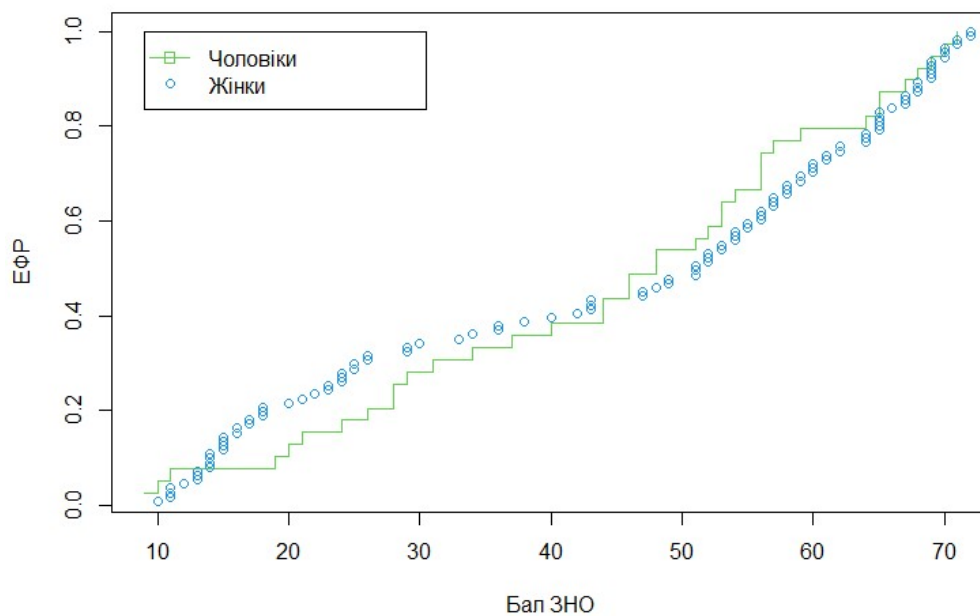


Рисунок 3.51 – Емпіричні функції розподілу показників з іспанської мови

Побудуємо діаграми типу “ящик з вусами”:

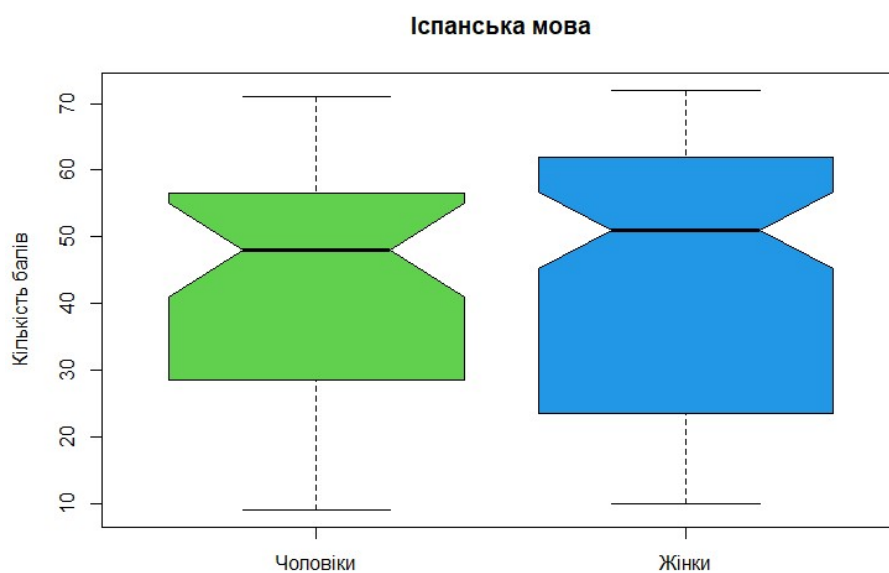


Рисунок 3.52 – Boxplot діаграми розподілу результатів з іспанської мови

Отримаємо квартилі розподілу даних результатів з біології:

Таблиця 3.28 – Квартилі розподілу даних результатів з іспанської мови

	Мін	25%	50%	75%	Макс
Іспанська мова, чоловіки	9	28.5	48	56.5	71
Іспанська мова, жінки	10	23.5	51	62	72

Перевіримо гіпотезу про нормальний розподіл за допомогою непараметричних критеріїв:

Таблиця 3.29 – Перевірка нульової гіпотези

Вибірка	Критерії			
	Ліллієфорса		Андерсона-Дарлінга	
	D	p-value	A	p-value
Іспанська мова, чоловіки	0.1081	0.2993	0.61931	0.09949
Іспанська мова, жінки	0.15223	$1.188 \cdot 10^{-6}$	4.3074	$9.12 \cdot 10^{-11}$

Використаємо; критерій Колмогорова – Смирнова, що перевіряє близькість емпіричних функцій розподілу вибірок; критерій Уїлкоксона, чутливий до зсуву центрів розподілу

Таблиця 3.30 – Результат перевірки однорідності показників

Вибірка	Критерії			
	Колмогорова-Смирнова		Уїлконсона	
	D	p-value	W	p-value
Іспанська мова	0.13208	0.5868	2114.5	0.832

ВИСНОВКИ

На основі результатів дослідження найбільш великий розрив спостерігається між чоловіками та жінками з української мови, на рівні 13,1 балів на користь жінок.

Тестування з української мови є обов'язковою умовою для вступу до вищих навчальних закладів, при вступі він має найбільший ваговий коефіцієнт, тому можна очікувати, що при вступі на бюджет ця відмінність буде відігравати істотну роль.

З історії України жінки перевершують чоловіків на 3.79 бали, з біології на 2.98, з англійської мови на 2.46, з німецької мови на 2.24, з географії на 1,19, з французької мови на 1,1, з іспанської мови на 0,83.

Абсолютна протилежна ситуація в математиці та хімії. З хімії чоловіки випереджають жінок на 2,05 бали. З математики результати середнього балу чоловіків на 0,08 більше ніж у жінок, проте при переведенні цих балів у шкалу 100-200 отримаємо однаковий бал на рівні 129 балів, тому можна казати, що на практиці різниця середніх балів чоловіків і жінок є не суттєвою.

Також за цими результатами можна казати, що відмінності між результатами приблизно такі ж як і у наведених прикладах міжнародних досліджень.

ПЕРЕЛІКИ ПОСИЛАНЬ

1. ЗНО. Загальна інформація [Електронний ресурс]. – Режим доступу: <https://testportal.gov.ua/zagalna-informatsiya-zno/>.
2. ООН. Цели устойчивого развития [Електронний ресурс]. – Режим доступу: <https://sdgs.un.org/ru/goals>.
3. Міжнародне дослідження TIMSS-2019 [Електронний ресурс]. – Режим доступу: <https://timss2019.org/reports/>.
4. Міжнародне дослідження PIRLS [Електронний ресурс]. – Режим доступу: <https://www.iea.nl/studies/iea/pirls>.
5. Міжнародне дослідження PISA [Електронний ресурс]. – Режим доступу: <https://www.oecd.org/pisa/>.
6. Результаты GSCE 2019 [Електронний ресурс]. – Режим доступу: <https://schoolsweek.co.uk/gcse-results-2019-subject-tables/>
7. Методи статистичних досліджень [Електронний ресурс]. – Режим доступу: <https://sibac.info/blog/metody-opisatelnoy-statistiki>.
8. **Лемешко, Б. Ю.** О применении и мощности критериев проверки однородности дисперсий. Ч. I. Параметрические критерии [Текст] / Б. Ю. Лемешко, С. Б. Лемешко, А. А. Горбунова // Измерительная техника. – 2010. – № 3. – С. 10-16.
9. **Lilliefors, H.W.** On the Kolmogorov-Smirnov test for normality with mean and variance unknown [Текст] / H. W. Lilliefors // J. Am. Statist. Assoc. – 1967. – V.62. – P.399-402.
10. Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия [Електронний ресурс]. – Режим доступу: <https://www.researchgate.net/publication/315330227>.
11. **Wilcoxon, F.** Individual comparisons by ranking methods [Текст] / F. Wilcoxon // Biometrics Bulletin. 1945. – Vol 1. – No 6. – P. 80-83.

ДОДАТОК А.

Код програми

```
Ukr_2021.R
```

```
#Порівняння результатів ЗНО-2021 з фізики для чоловіків і жінок
```

```
# Формування таблиці даних
```

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Ukr_2021.csv', dec = ",", sep = ";", header = TRUE, nrow = 130580)
```

```
dat=as.matrix(dat)
```

```
datm = dat[1:91668,1] # Чоловіки
```

```
datw = dat[(1:130580),2] #Жінки
```

```
#Візуалізація даних
```

```
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість балів",main="Українська мова", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
```

```
hist(datm, col = 3, xlab = "Кількість балів")
```

```
hist(datw, col = 4, xlab = "Кількість балів")
```

```
par(old.par)
```

```
#Емпірична функція розподілу
```

```
erfm = seq(1/length(datm), 1, len = length(datm))
```

```
erfw = seq(1/length(datw), 1, len = length(datw))
```

```
datms = sort(datm); datws = sort(datw)
```

```
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type = "s", col = 3, col.main = "brown", cex.main = 1.2)
```

```
points(datws, erfw, col = 4)
```

```
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
```

```
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
```

```
quantile(datm); quantile(datw)
```

```
#Проверка нормальності
```

```
#Q-Q-діаграма
```

```
old.par <- par(mfrow=c(1,2))
```

```
qqnorm(datm, main = "Українська мова, чоловіки", col = 3, col.main = "brown",
```

```

сех.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Українська мова, жінки", col = 4, col.main = "brown",
сех.main = 1.2)
qqline(datw, col = 2)
par(old.par)

# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)

#Тести Андерсона – Дарлінга і Ліллієфорса

#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)

ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса

ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансари – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)

Phys_2021.R

#Порівняння результатів ЗНО-2021 з фізики для чоловіків і жінок

# Формування таблиці даних

dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Phys_2021.csv', dec
= ",", sep = ";", header = TRUE, nrow = 19810)

dat=as.matrix(dat)
datm = dat[,1] # Чоловіки
datw = dat[(1:3393),2] #Жінки

#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Фізика", names=c("Чоловіки","Жінки"))

old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")

```

```
par(old.par)
```

```
#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки", "Жінки"), col=c(3,4), pch=c(0, 1),
btu = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
quantile(datm); quantile(datw)
```

```
#Перевірка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Фізика, чоловіки", col = 3, col.main = "brown",
cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Фізика, жінки", col = 4, col.main = "brown",
cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

```
# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)
```

```
#Тести Андерсона – Дарлінга і Ліллієфорса
```

```
#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)
```

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

```
ks.test(datm, datw) #Основний тест для подібних даних
wilcox.test(datm, datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm, datw) # тест Ансари – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)
```

Math_2021.R

#Порівняння результатів ЗНО-2021 з фізики для чоловіків і жінок

Формування таблиці даних

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Math_2021.csv',
dec = ",", sep = ";", header = TRUE, nrow = 125319)
```

```
dat=as.matrix(dat)
datm = dat[,1] # Чоловіки
datw = dat[(1:118883),2] #Жінки
```

```
#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Математика", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)
```

```
#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
quantile(datm); quantile(datw)
```

```
#Проверка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Математика, чоловіки", col = 3, col.main = "brown",
cex.main = 1.2)
qqline(datm, col = 2)
```

```

qqnorm(datw, main = "Математика, жінки", col = 4, col.main = "brown",
sex.main = 1.2)
qqline(datw, col = 2)
par(old.par)

# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)

#Тести Андерсона – Дарлінга і Ліллієфорса

#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)

ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса

ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)

HistUkr_2021.R

#Порівняння результатів ЗНО-2021 з фізики для чоловіків і жінок

# Формування таблиці даних

dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\HistUkr_2021.csv',
dec = ",", sep = ";", header = TRUE, nrow = 108347)

dat=as.matrix(dat)
datm = dat[1:91026,1] # Чоловіки
datw = dat[(1:108347),2] #Жінки

#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Історія України", names=c("Чоловіки","Жінки"))

old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)

```

```

#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки", "Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))

#Квантилі розподілів
quantile(datm); quantile(datw)

#Проверка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Історія України, чоловіки", col = 3, col.main = "brown",
      cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Історія України, жінки", col = 4, col.main = "brown",
      cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)

# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)

#Тести Андерсона – Дарлінга і Ліллієфорса

#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)

ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса

ks.test(datm, datw) #Основний тест для подібних даних
wilcox.test(datm, datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm, datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)

Geography_2021.R

```

#Порівняння результатів ЗНО-2021 з географії для чоловіків і жінок

Формування таблиці даних

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка
(2)\\Geography_2021.csv', dec = ",", sep = ";", header = TRUE, nrow = 58548)
```

```
dat=as.matrix(dat)
datm = dat[,1] # Чоловіки
datw = dat[(1:54568),2] #Жінки
```

#Візуалізація даних

```
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Географія", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)
```

#Емпірична функція розподілу

```
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
btu = "y", lty=c(1, 0))
```

#Квантилі розподілів

```
quantile(datm); quantile(datw)
```

#Проверка нормальності

#Q-Q-діаграма

```
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Географія, чоловіки", col = 3, col.main = "brown",
cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Географія, жінки", col = 4, col.main = "brown",
cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

```

# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)

#Тести Андерсона – Дарлінга і Ліллієфорса

#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)

ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса

ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)

French_2021.R

#Порівняння результатів ЗНО-2021 з французької мови для чоловіків і жінок

# Формування таблиці даних

dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\French_2021.csv',
dec = ",", sep = ";", header = TRUE, nrow = 256)

dat=as.matrix(dat)
datm = dat[1:99,1] # Чоловіки
datw = dat[(1:256),2] #Жінки

#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Французька мова", names=c("Чоловіки","Жінки"))

old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)

#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)

```

```
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(9, 1, c("Чоловіки", "Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
quantile(datm); quantile(datw)
```

```
#Проверка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Французька мова, чоловіки", col = 3, col.main = "brown",
cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Французька мова, жінки", col = 4, col.main = "brown",
cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

```
# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)
```

```
#Тести Андерсона – Дарлінга і Ліллієфорса
```

```
#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)
```

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

```
ks.test(datm, datw) #Основний тест для подібних даних
wilcox.test(datm, datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm, datw) # тест Ансари – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)
```

```
Esp_2021.R
```

```
#Порівняння результатів ЗНО-2021 з іспанської мови для чоловіків і жінок
```

```
# Формування таблиці даних
```

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Esp_2021.csv', dec
= ",", sep = ";", header = TRUE, nrow = 111)
```

```
dat=as.matrix(dat)
datm = dat[1:39,1] # Чоловіки
datw = dat[(1:111),2] #Жінки
```

```
#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Іспанська мова", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)
```

```
#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(9, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
quantile(datm); quantile(datw)
```

```
#Перевірка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Іспанська мова, чоловіки", col = 3, col.main = "brown",
cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Іспанська мова, жінки", col = 4, col.main = "brown",
cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

```
# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)
```

```
#Тести Андерсона – Дарлінга і Ліллієфорса
```

```
#install.packages("nortest") #При першому використанні треба встановити пакет  
library(nortest)
```

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
```

```
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

```
ks.test(datm,datw) #Основний тест для подібних даних
```

```
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між  
центрами розподілів)
```

```
ansari.test(datm,datw) # тест Ансари – Бредлі (перевіряє значущість відмінності  
дисперсій розподілів)
```

```
English_2021.R
```

```
#Порівняння результатів ЗНО-2021 з англійської мови для чоловіків і жінок
```

```
# Формування таблиці даних
```

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\English_2021.csv',  
dec = ",", sep = ";", header = TRUE, nrow = 64878)
```

```
dat=as.matrix(dat)
```

```
datm = dat[1:62836,1] # Чоловіки
```

```
datw = dat[(1:64878),2] #Жінки
```

```
#Візуалізація даних
```

```
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість  
балів",main="Англійська мова", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
```

```
hist(datm, col = 3, xlab = "Кількість балів")
```

```
hist(datw, col = 4, xlab = "Кількість балів")
```

```
par(old.par)
```

```
#Емпірична функція розподілу
```

```
erfm = seq(1/length(datm), 1, len = length(datm))
```

```
erfw = seq(1/length(datw), 1, len = length(datw))
```

```
datms = sort(datm); datws = sort(datw)
```

```
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =  
"s", col = 3, col.main = "brown", cex.main = 1.2)
```

```
points(datws, erfw, col = 4)
```

```
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),  
bty = "y", lty=c(1, 0))
```

```
#Квантили розподілів
quantile(datm); quantile(datw)
```

```
#Проверка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Англійська мова, чоловіки", col = 3, col.main = "brown",
  cex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Англійська мова, жінки", col = 4, col.main = "brown",
  cex.main = 1.2)
qqline(datw, col = 2)
par(old.par)
```

```
# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)
```

```
#Тести Андерсона – Дарлінга і Ліллієфорса
```

```
#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)
```

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

```
ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансари – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)
```

```
Deutsch_2021.R
```

```
#Порівняння результатів ЗНО-2021 з німецької мови для чоловіків і жінок
```

```
# Формування таблиці даних
```

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Deutsch_2021.csv',
  dec = ",", sep = ";", header = TRUE, nrow = 1198)
```

```
dat=as.matrix(dat)
datm = dat[1:627,1] # Чоловіки
```

```
datw = dat[(1:1198),2] #Жінки
```

```
#Візуалізація даних
```

```
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість  
балів",main="Німецька мова", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
```

```
hist(datm, col = 3, xlab = "Кількість балів")
```

```
hist(datw, col = 4, xlab = "Кількість балів")
```

```
par(old.par)
```

```
#Емпірична функція розподілу
```

```
erfm = seq(1/length(datm), 1, len = length(datm))
```

```
erfw = seq(1/length(datw), 1, len = length(datw))
```

```
datms = sort(datm); datws = sort(datw)
```

```
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =  
"s", col = 3, col.main = "brown", cex.main = 1.2)
```

```
points(datws, erfw, col = 4)
```

```
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
```

```
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
```

```
quantile(datm); quantile(datw)
```

```
#Проверка нормальності
```

```
#Q-Q-діаграма
```

```
old.par <- par(mfrow=c(1,2))
```

```
qqnorm(datm, main = "Німецька мова, чоловіки", col = 3, col.main = "brown",  
cex.main = 1.2)
```

```
qqline(datm, col = 2)
```

```
qqnorm(datw, main = "Німецька мова, жінки", col = 4, col.main = "brown",  
cex.main = 1.2)
```

```
qqline(datw, col = 2)
```

```
par(old.par)
```

```
# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
```

```
shapiro.test(datm)
```

```
shapiro.test(datw)
```

```
#Тести Андерсона – Дарлінга і Ліллієфорса
```

```
#install.packages("nortest") #При першому використанні треба встановити пакет  
library(nortest)
```

```
ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса
```

```
ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)
```

```
Biology_2021.R
```

```
#Порівняння результатів ЗНО-2021 з біології для чоловіків і жінок
```

```
# Формування таблиці даних
```

```
dat=read.table(file='C:\\Users\\User\\Downloads\\Новая папка (2)\\Biology_2021.csv',
dec = ",", sep = ";", header = TRUE, nrow = 52001)
```

```
dat=as.matrix(dat)
datm = dat[1:29410,1] # Чоловіки
datw = dat[(1:52001),2] #Жінки
```

```
#Візуалізація даних
boxplot(datm,datw, notch = TRUE, col = c(3,4), ylab = "Кількість
балів",main="Біологія", names=c("Чоловіки","Жінки"))
```

```
old.par <- par(mfrow=c(1,2))
hist(datm, col = 3, xlab = "Кількість балів")
hist(datw, col = 4, xlab = "Кількість балів")
par(old.par)
```

```
#Емпірична функція розподілу
erfm = seq(1/length(datm), 1, len = length(datm))
erfw = seq(1/length(datw), 1, len = length(datw))
datms = sort(datm); datws = sort(datw)
plot(datms, erfm, xlab = "Бал ЗНО", ylab = "ЕФР", type =
"s", col = 3, col.main = "brown", cex.main = 1.2)
points(datws, erfw, col = 4)
legend(1, 1, c("Чоловіки","Жінки"), col=c(3,4), pch=c(0, 1),
bty = "y", lty=c(1, 0))
```

```
#Квантилі розподілів
quantile(datm); quantile(datw)
```

```

#Перевірка нормальності
#Q-Q-діаграма
old.par <- par(mfrow=c(1,2))
qqnorm(datm, main = "Біологія, чоловіки", col = 3, col.main = "brown",
sex.main = 1.2)
qqline(datm, col = 2)
qqnorm(datw, main = "Біологія, жінки", col = 4, col.main = "brown",
sex.main = 1.2)
qqline(datw, col = 2)
par(old.par)

# Тест Шапіро-Уїлка (для виборок обсягом до 5000)
shapiro.test(datm)
shapiro.test(datw)

#Тести Андерсона – Дарлінга і Ліллієфорса

#install.packages("nortest") #При першому використанні треба встановити пакет
library(nortest)

ad.test(datm); ad.test(datw) # тест Андерсона-Дарлінга
lillie.test(datm); lillie.test(datw) # тест Ліллієфорса

ks.test(datm,datw) #Основний тест для подібних даних
wilcox.test(datm,datw) # тест Уїлкоксона (перевіряє значущість відмінності між
центрами розподілів)
ansari.test(datm,datw) # тест Ансарі – Бредлі (перевіряє значущість відмінності
дисперсій розподілів)

```