**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**Національний авіаційний університет**

**Лабораторна робота № 1**

З дисципліни: «Аналіз даних»

Виконав:  
студент 351 групи ІІДС  
Подлєсний Максим

Перевірив:

завідувач кафедри  
Приставка П.О.

Київ-2016

**Зміст**

**Постановка задачі………………………………………………………..1**

**Теоретична частина……………………………………………………..2**

**Практична частина……………………………………………………..9**

**Висновки…………………………………………………………………12**

**Список літератури……………………………………………………..12**

**Постановка задачі**

1.Організувати роботу з вхідними даними таким щоб уможливити

обробку однієї або кількох вибірок які характеризують одновимірні або багатовимірні об'єкти спостережень.

-Лабораторну роботу виконати на основі лабораторної роботи IV- семестру в рамках єдиної автоматизованої системи аналізу статистичних даних.

-Реалізувати обчислювальні процедури перевірки однорідності двох вибірок, що характеризують одновимірні об'єкти спостережень:

-Перевірку збігу дисперсій та середніх для вибірок, розподілених за нормальним законом;

-критерій однорідності Смирнова-Колмогорова.

-критерій Вілкоксона, Манна-Уїтні, різниці середніх рангів

-Реалізувати обчислювальні процедури перевірки однорідності множини

вибірок:

-Критерій Бартлетта та однофакторний дисперсійний аналіз для виоірок. розподілених за нормальним законом;

-критерій знаків:

-Н-критерій;

-Q-критерій;

5.Реалізувати критерій Аббе

6. Провести тестування програмного забезпечення на реальних даних

7.За результатами виконання лабораторної роботи оформити звіт.

**Теоретична частина**

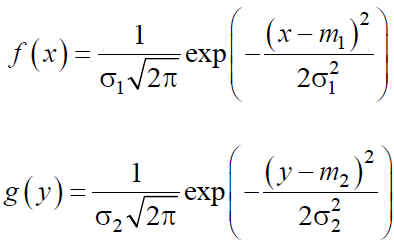
**Задача двох вибірок**

Задачу однорідності й незалежності в більшості випадків можна звести

до задачі двох вибірок.

Нехай маємо дві генеральні сукупності Ω1, Ω2, із яких вибрані вибірки Ω1,N1 = {x1, x2, … xN1}, Ω1,N2 = {y1, y2, … yN2}. Відносно Ω1 і Ω2 припускаються розподіли відповідно F (x) і G( y) . Необхідно перевірити гіпотезу Η0 : F(x) ≡ G(y) за альтернативи Η1 : F(x) ≠ G(y).

Таке подання задачі є загальне, і її розв’язок одержують за допомогою як параметричних, так і непараметричних критеріїв. Розглянемо розв’язання такої задачі за параметричним критерієм. Припустимо, що закони розподілів F (x) і G( y) є нормальні, а їх функції щільності такі:



Для того щоб F (x) і G( y) були однаковими, необхідний збіг їх відповід-

них параметрів. У цьому випадку гіпотези H0, Η1 можемо переписати у вигляді

H0: m1 = m2 σ1 = σ2

H1: m1 ≠ m2 σ1 ≠ σ2

Для перевірки гіпотез Η0, Η1 існують критерії, розглянуті нижче

**Перевірка збігу середніх**

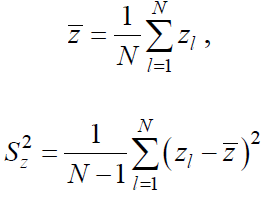
Перевірка збігу середніх двох вибірок здійснюється за t-тестом, проведення якого для аналізованих вибірок потребує певних операцій перетворення.

Будемо розрізняти випадки залежних і незалежних вибірок Ω1,N1, Ω1,N2.

*Випадок залежних вибірок*. Такий варіант дозволяє оцінювати вибірки,

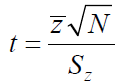
що характеризують однакові фізичні процеси або явища, які вивчаються різними методами. Для цього вимірюють один і той же параметр за різними методами, одержуючи вибірки однакового обсягу відносно xl та yl , l =1,N.

Обчисливши різницю zl = xl – yl одержують нову вибірку Ω1,N2 = {z1, z2, … zN2}, для якої визначають



Оскільки xi та yi – реалізації випадкових величин ξ та η, які мають нор-

мальні розподіли з Ν1 (x;m1,σ), Ν2(y;m2 ,σ2) , маємо, що zl – реалізація випадкової величини ζ , для якої E{ζ} = E{ξ} − E{η} . Тоді гіпотезу Η0:m1= m2 переписують у вигляді Η0 :m1 −m2 = 0 або Η0 : E ζ = 0 і для її перевірки використовують таку статистичну характеристику:



Результат порівняння t > tα/2,ν, (ν = N − 2 ) свідчить про те, що значення статистичної характеристики потрапило до критичної області, отже, головну гіпотезу слід відхилити. Подальший висновок відносно того, яке із середніх більше, робиться за знаком z .

Випадок незалежних вибірок. Даний варіант дозволяє оцінювати вибір-

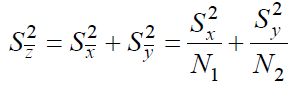
ки, які характеризують різні фізичні процеси або явища. У такому разі обсяги вибірок можуть відрізнятися. Можливі два випадки:

1) обсяг вибірок є представницький;

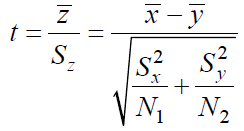
2) обсяг вибірок обмежений.

Нехай вибірки є представницькі. Враховуючи, що різниця z¯ = x¯ − y¯

розподілена нормально з дисперсією

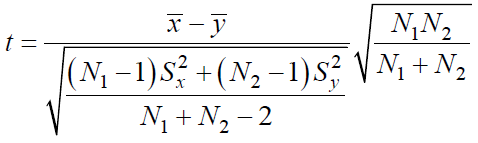


для перевірки головної гіпотези 0 Η на основі t -тесту застосовують статистику



яка має t -розподіл Стьюдента з кількістю степенів вільності ν = N1 + N2 − 2 .

За обмеженого обсягу вибірок (N1 + N2 ≤ 25) як статистичну характеристику використовують величину

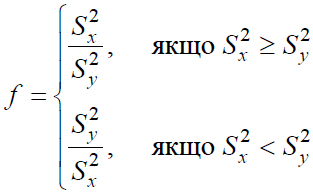


**Перевірка збігу дисперсій**

Поряд з t -тестом у статистичній теорії перевірки гіпотез особливе місце займають параметричні критерії, що базуються на F -статистиках, розподілених за законом розподілу Фішера, – так звані F-тести. За наявності S12, S22 – незалежних оцінок для дисперсій σ1, σ1 – F-тест дозволяє перевіряти гіпотезу про їх збіг

Η0 :σ1 = σ2

Під час розв’язання задачі перевірки збігу дисперсій двох вибірок за статистичну характеристику беруть значення



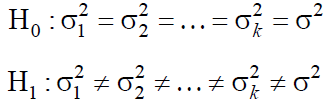
Статистика f має F-розподіл Фішера з кількістю степенів вільності

ν1 = N1 −1 та ν2 = N2 −1. Враховуючи, що f > 0, за відомого α обчислюють критичне значення і, якщо

f ≤ fα,ν1,ν2 ,

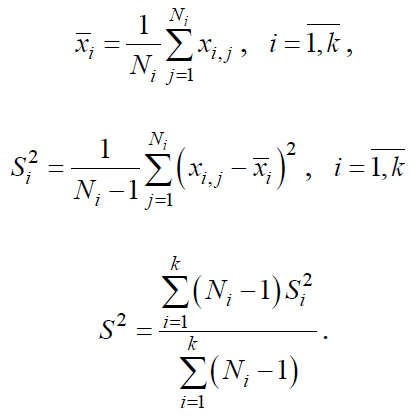
приймають головну гіпотезу.

За необхідності перевірити гіпотезу збіг дисперсій k вибірок

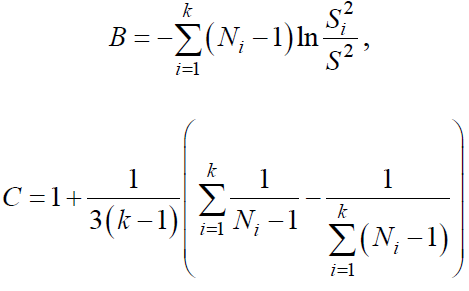


використовують критерій Бартлетта. Нехай заданий багатовимірний набір даних {xi,j i = 1…k, j = 1…Ni}, що являє собою k вибірок (можливо, різного обсягу).

Для перевірки головної гіпотези спочатку обчислюють значення



За статистичну характеристику беруть величину χ2 = B/C, де В і С визначаються як



Для заданого рівня значущості α і кількості степенів вільності ν = k −1

знаходять критичне χα,ν (табл. Б.3) і приймають головну гіпотезу, якщо

χ2 <= χ2α,ν

**Однофакторний дисперсійний аналіз**

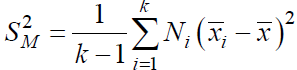
Однофакторний дисперсійний аналіз застосовують для перевірки того, чи різняться поміж себе значення середніх множини k незалежних вибірок, що є реалізаціями відповідних нормально розподілених випадкових величин. Однофакторний дисперсійний аналіз порівнює два джерела варіації даних: міжгрупову варіацію (варіацію поміж вибірками) та варіацію всередині кожної вибірки. Припускаючи, що дисперсії всіх k вибірок однакові висувають головну гіпотезу

Н0 : m1 = m2 =…= mn

за альтернативи

Н1 : mi ≠ mj;i≠j

Міжгрупова варіація SM2 дає оцінку відмінностей середніх вибірок, що аналізуються:



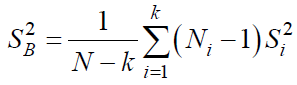
де

Ni – обсяг i -ї вибірки;

x¯I - оцінка математичного сподівання i -ї вибірки

x¯ - загальне середнє

Варіація всередині кожної вибірки S2В визначається згідно з виразом



де Si2 – оцінка дисперсії кожної вибірки

Перевірка головної гіпотези проводиться на основі статистичної характеристики F = SM2/SB2

яка має розподіл Фішера з кількістю степенів вільності ν1 = k −1, ν2 = N − k .

Головну гіпотезу Η0 приймають у разі виконання умови

F ≤ fα,ν1,ν2

роблячи висновок, що середні вибірок невеликою мірою різняться поміж собою.

Якщо остання нерівність не виконується, роблять висновок про існування істотної різниці між вибірковими середніми, а отже, про неможливість пояснити розходження в їх значеннях лише випадковістю. Подальший аналіз може полягати у визначенні того, які саме вибірки попарно різняться між собою.

Останнє з’ясовується на основі t -статистик, уведених для випадку незалежних вибірок, з урахуванням наявних обсягів аналізованих вибірок.

**Критерії порядкових статистик**

Наведені нижче критерії однорідності належать до так званих рангових. Вони ґрунтуються на вивченні послідовності реалізацій випадкової величини та можуть застосовуватися навіть у тих випадках, коли закони розподілу аналізованих вибірок відмінні від нормального. Головна гіпотеза формулюється так: дві вибірки Ω1,N2 = {x1, x2, … xN2} та Ω1,N2 = {у1, у2, … уN2} мають однакові закони розподілу.

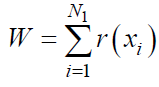
Н0 : F(x) ≡ G(y)

Критерії Вілкоксона та U-критерій Манна–Уїтні є найчастіше використовувані.

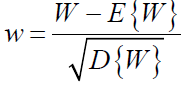
Для перевірки головної гіпотези про значущість різниці двох незалежних вибірок з останніх формують загальний варіаційний ряд (обсягом N = N1 + N2 ), приписуючи кожному значенню варіанти ранг rxi або ryj, тобто порядковий номер. Якщо в загальному варіаційному ряді виявляється декілька варіант, які збігаються, то кожній присвоюють ранг, що дорівнює середньому арифметичному їх порядкових номерів у сумісній послідовності.

Критерій суми рангів Вілкоксона базується на обчисленні статистичної характеристики W , що визначається як сума рангів вибірки Ω1,N1 (або

Ω1,N2) у загальному варіаційному ряді.

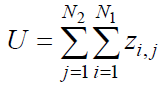


Порівнюючи значення



з критичним значенням uα нормального закону розподілу, головну гіпотезу приймають або відхиляють.

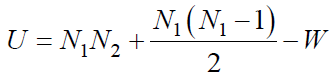
В основі U-критерію Манна–Уїтні лежить дослідження кількості способів, за допомогою яких в одній вибірці можна знайти значення, що перевищує значення в іншій вибірці. Аналізуючи загальний ряд даних, встановлюють, що має місце перерозподіл значень випадкових величин. Ступінь перерозподілу x та y визначають через інверсію. Якщо у варіаційному ряді деякому x передує y , то таке явище називають однією інверсією, якщо ж певному x передує k значень y , говорять, що значення x має k інверсій. Під час реалізації U -критерію Манна–Уїтні розраховують статистичну характеристику U , яка визначає кількість інверсій відносно x (або y ) у загальному ряду



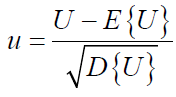
zi,j = 1; xi > yj

0; xi <= yj

Слід відзначити, що поміж статистиками U та W існує така залежність:



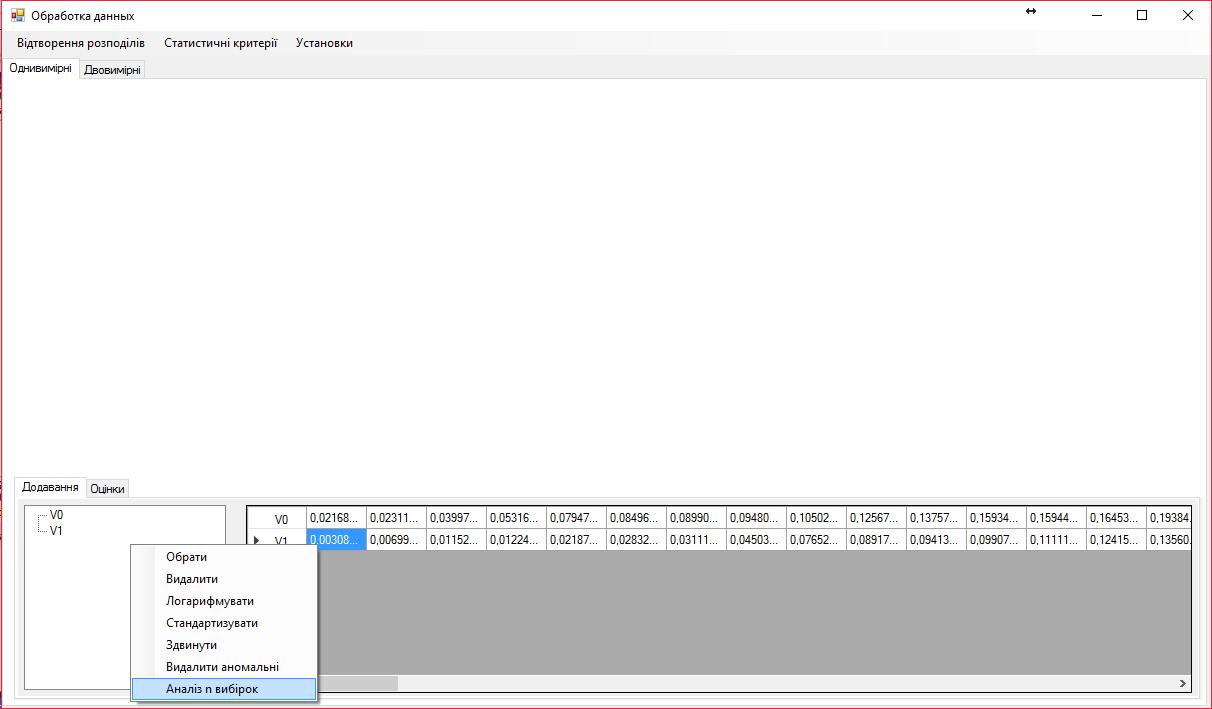
Для перевірки гіпотези Η0 обчислюють статистичну характеристику



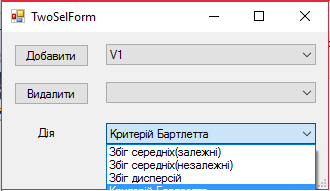
Якщо обсяг загального варіаційного ряду N < 25, слід застосовувати точні апроксимації законів розподілу статистик W та U або звертатися до їх табульованих значень.

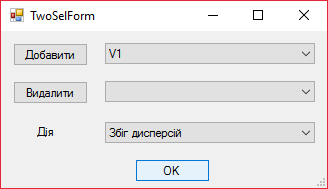
**Практична частина**

Для входу в меню перевірки критеріїв потрібно обрати вибірки та натиснути пункт меню «Аналіз N вибірок».

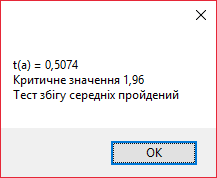


Додаємо потрібні вибірки для аналізу та обираємо потрібний критерій

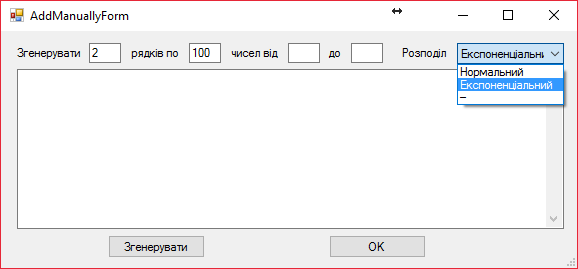


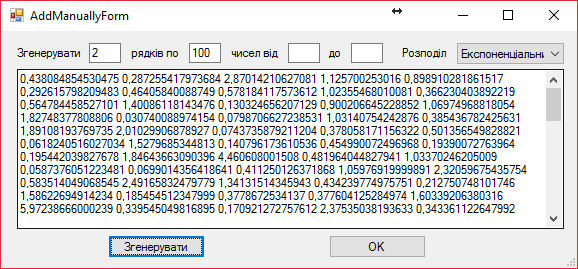


Програма покаже результат критерію



Щоб згенерувати вибірки за певним законом використаємо меню «увести вручну», та оберемо кількість вибірок та потрібний розподіл





Для задання рівня значущості та параметрів моделювання вибірок використаємо меню «установки»

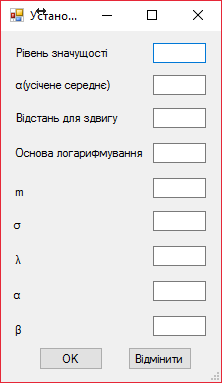


Схема програми

Зчитування з файлу

Генерація вибірки за розподілом

Обрання вибірок для аналізу

Обрання потрібного критерію

Рангові критерії(Вілкоксона, Манна-Уітні, різниці рангів)

Критерії багатьох вибірок(Бартлетта, однофакторний, Н-, Q- критерії)

Збігу дисперсій

Збігу середніх

Смирнова

Аббе

Математичні обрахунки

Порівняння емпіричного значення критерію з критичним

Робимо висновки про істинність Н0

Виведення результатів користувачу

Висновки

1.Я виконав практичну роботу з дисципліни «Аналіз даних». У ній я реалізував низку критеріїв для перевірки однорідності даних та можливість генерування вибірок за певним розподілом. Таким чином для вибірок з однаковим розподілом та вектором параметрів критерії, як правило підтверджують головну гіпотезу, що свідчить про однорідність даних.

При виконанні даної роботи я детально ознайомився з задачею двох вибірок, засвоїв знання про критерії однорідності даних та закрепив їх на практиці.

Список використаної літератури

П.О. Приставка, О.М.Мацуга: «АНАЛІЗ ДАНИХ» - Електронний посібник для студентів спеціальності «прикладна математика»