Міністерство освіти і науки України

Львівський національний університет імені Івана Франка

Відділення: денне
Група: ПМі-34
Факультет
прикладної математики
та інформатики
Кафедра
Інформаційних систем

Звіт

з теорії ймовірностей та математичної статистики

Виконав студент
Кізло Тарас Михайлович
Перевірив асистент
Коркуна Наталія Михайлівна

2018

План

1. Постановка задачі
2. Теоретичні відомості і програмна реалізація
3. Висновок

Постановка задачі

1. Згенерувати вибірку заданого об’єму з певного проміжку для непевної та дискретної статистичної змінної (та зчитати з файлу у вигляді частотної таблиці)
2. На підставі отриманих вибіркових даних:
    1) Побудувати варіаційний ряд
    2) Побудувати статистичний розподіл варіанти
    3) Представити графічно статистичний матеріал
    4) Побудувати емпіричну функцію розподілу
    5) Обчислити всі числові характеристики
3. Користуючись критерієм Пірсона, на підставі наведених статистичних даних при заданому рівні значущості, перевірити правильність висунутої гіпотези щодо закону розподілу генетичної сукупності.

Виконання поставленої задачі відбувається на мові програмування C# 4.7.1 з використанням графічного інтерфейсу Windows Forms.

Теоретичні відомості і програмна реалізація

Перед тим як перейти до програмної реалізації розберемо декілька важливих означень.

Часто потрібно вивчати явище, точний перебіг яких не можливо передбачити і які виступають не поодиноко, а масово. Такі явища називаються **масовими випадковими**.

Масові випадкові явища залежні від часу називаються **випадковими, або стохастичними процесами**.

Різні прояви стохастичного процесу називаються мінливими **величинами, варіантами або статистичною змінною.**

Тут і надалі використовуватимемо термін *статистична змінна*.

Випадкові явища, стохастичні процеси, мінливі величини пізнаємо спостереженнями, тобто у результаті відповідно поставлених експериментів.

Кількість спостережень називається обсягом **(розміром, об'ємом, довжиною, тривалістю)** спостережень. Використовуватиме термін *обсяг*.

Сукупність спостережень називається **статистичним матеріалом**.

Кожне окреме спостереження називається **елементом статистичного матеріалу**.

Нехай $x_1, \ldots\ldots, x_n$ – результат *n* спостережень над одновимірною кількісною мінливою величиною. Тоді така послідовність представляє собою статистичний матеріал обсягом *n* спостережень.

Нехай серед спостережень зустрічаються такі можливі значення одновимірної дискретної варіанти *x* , впорядковані за величиною

$$x_{(1)} < x_{(2)} < \cdots < x_{(k)}$$

і нехай ці значення зустрічаються відповідно часто:

$$n_1, n_2, \ldots, n_k; \sum_{i=1}^{k} n_i = n$$

Тоді статистичний матеріал зручно записати в формі таблички з двома рядками у першому рядку виписуємо в зростаючому порядку можливі значення варіанти, а в другому – відповідні їм частоти. Дістанемо **частотну таблицю** (статистичним розподілом дискретної варіанти x) .

$x_{(1)} < x_{(2)} < \cdots < x_{(k)}$	$\sum$
$n_1, n_2, \dots, n_k$	n

Саме у такому вигляді зручно зберігати статистичний матеріал і в пам'яті комп'ютера. Для цього візьмемо посортоване бінарне дерево, вузол якого міститиме значення спостереження і його частоту, а сортування відбуватиметься за значенням.

```
SortedDictionary<double, int> statisticalTable;
```

Записавши спостереження у формі

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(k)}$$

Ми отримаємо **варіаційний ряд**

```
public double[] GetVariationSeries()
{
            List<double> series = new List<double>(size);
            foreach (KeyValuePair<double, int> element in statisticalTable)
            {
                series.AddRange(
                            Enumerable.Repeat(
                                  element: element.Key,
                                  count: element.Value)
                              );
            }
            return series.ToArray();
}
```

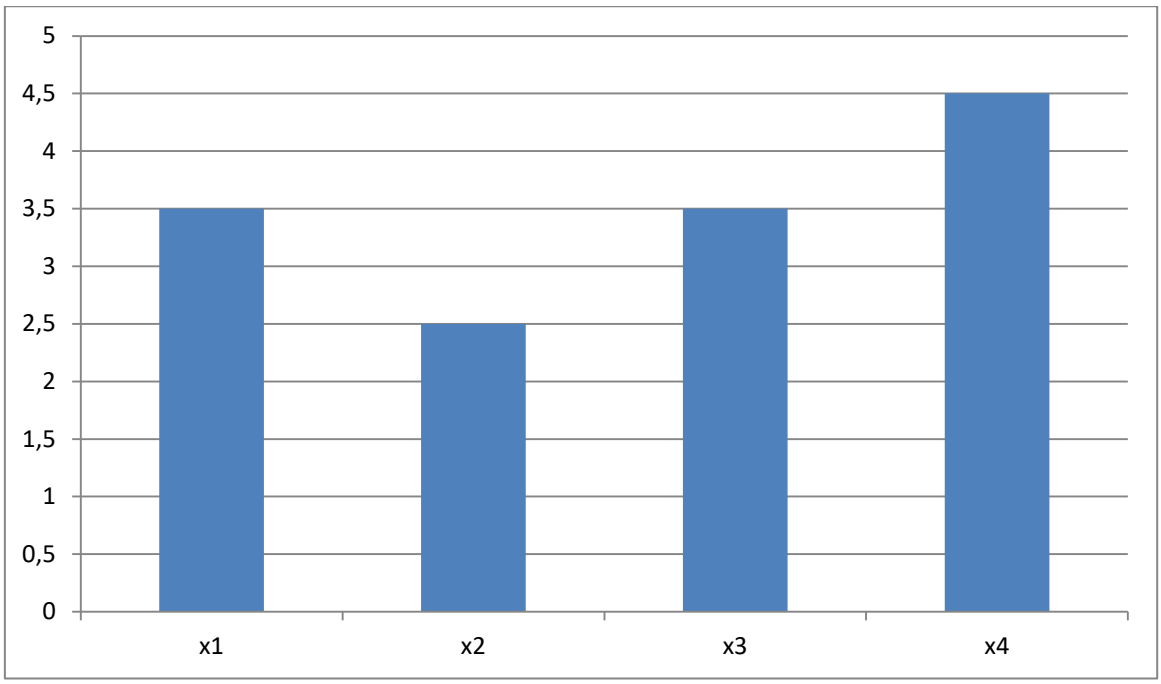
Тоді число **ступенів вільності** — кількість незалежних змінних, які однозначно описують стан фізичної системи, буде

$$d.f = n - 1$$

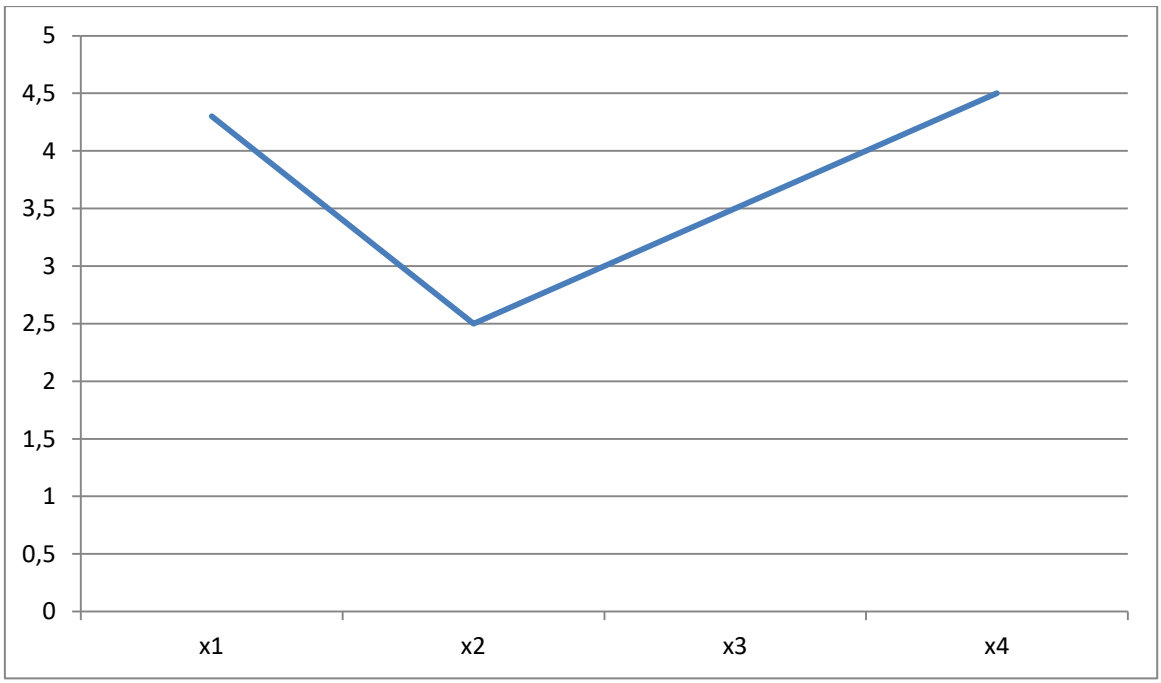
```
public int d_f => size >= 2 ? size - 1 : 1;
```

Для графічного представлення частотної таблиці на вісь абцис наносимо можливі значення дискретної мінливої величини та відкладемо в цих точках відповідні частоти *n (i =1, 2,....)* .

Отримаємо **діаграму частот**.

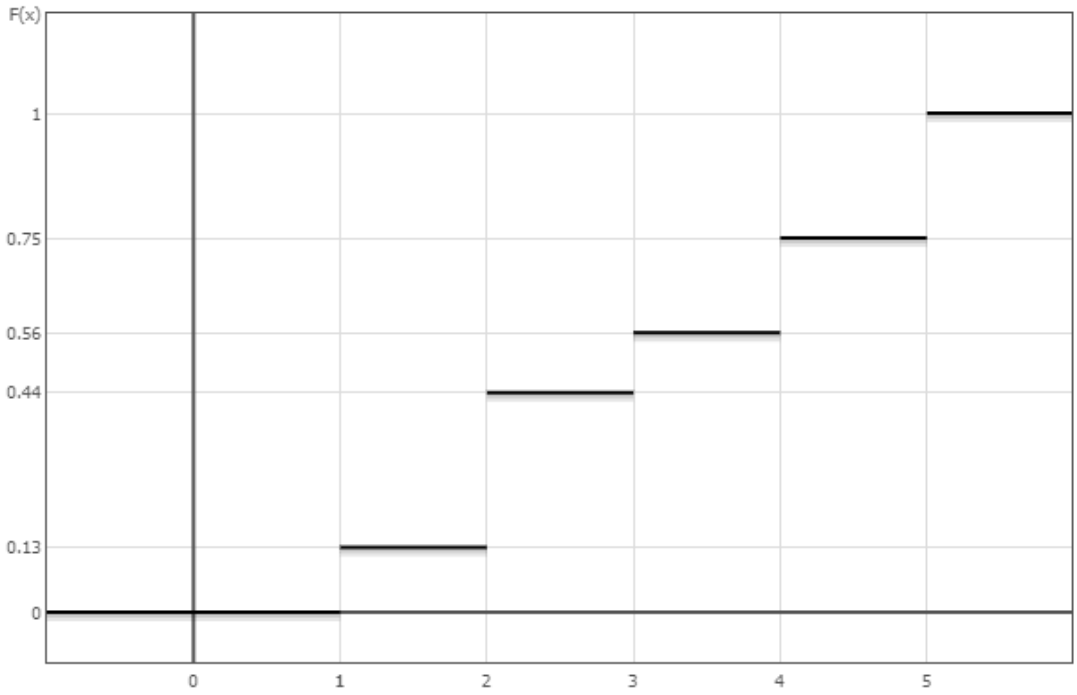


Якщо з’єднати відрізками сусідні пункти, то дістанемо **полігон частот**.



Функція розподілу ймовірностей — в теорії ймовірностей це функція, яка повністю описує розподіл ймовірностей випадкової величини.

$$F(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_k \le x < x_{k+1}\ (k = 1, \dots n-1) \\ 1, & x_{(n)} \le x \end{cases}$$



Різниці між найбільшим і найменшим елементами статистичного матеріалу називається **розмахом** статистичного матеріалу

$$p = x_{(n)} - x_{(1)}$$

```
                    public double p =>
  statisticalTable.Last().Key - statisticalTable.First().Key;
```

Інтервал розмаху ділимо досить довільним способом на (r +1) однакові або неоднакові інтервали, де r – натуральне , r = 1, 2,.....

$$2^r < n \le 2^{r+1}$$

```
public int r_1
{
            get
            {
                int r = 1;
                int count = 2;
```

```
                while (count < size)
                {
                    count *= 2;
                    ++r;
                }
                return r;
            }
}
```

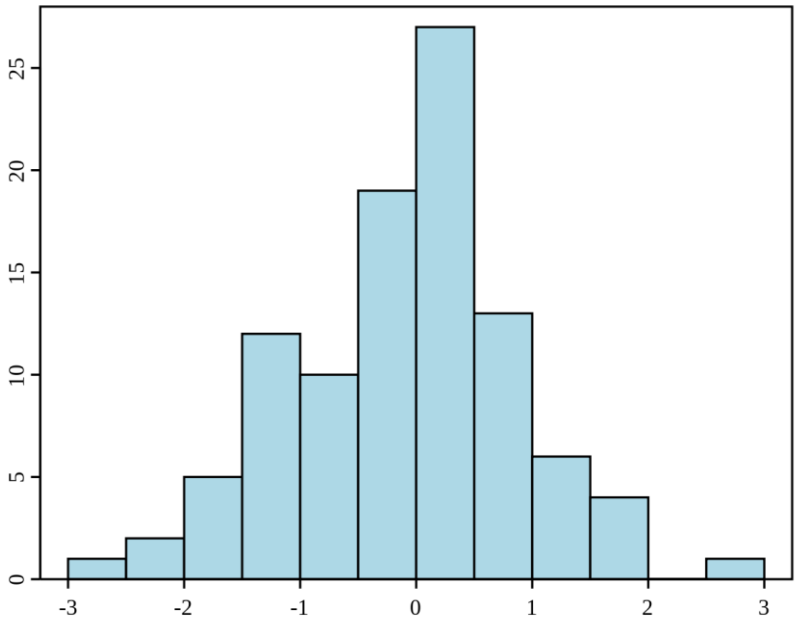
Центри одержаних інтервалів позначимо в зростаючому

$$z_1 < z_2 < \cdots < z_n$$

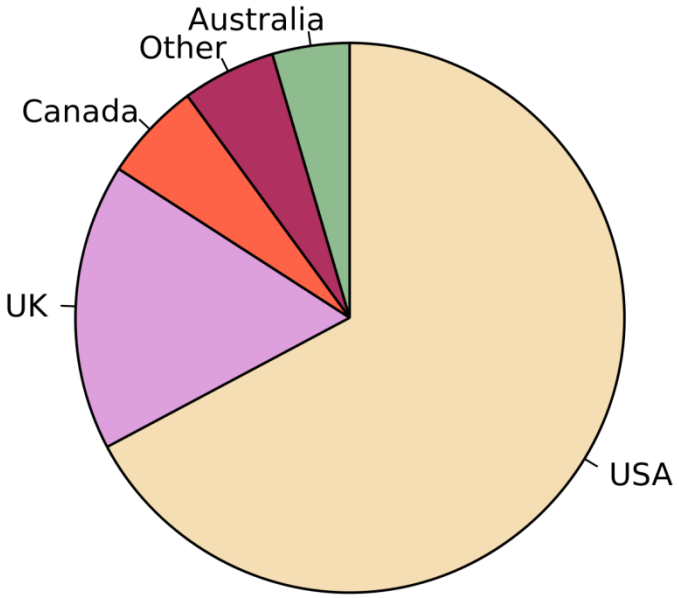
Тоді статистичний матеріал представимо у вигляді таблиці з двох рядків:

- 1-й в зростаючому порядку- центри інтервалів
- 2-й - відповідні частоти

Якщо з’єднати верхушки сусідніх вершин графіка частот відрізками, то одержимо многокутник частот або **полігон** частот . Якщо над інтервалом з центром в т. і z поставити прямокутник висотою ni , то одержимо **гістограму** частот



Існують і інші графічні представлення, наприклад **секторна діаграма**, тощо.



Групи статистик

1. Числові характеристики центральної тенденції
    a. медіана
    b. мода
    c. середнє арифметичне
2. Числові характеристики розсіяння
    a. варіанса
    b. стандарт
    c. розмах
    d. варіація
    e. інтерквантильність широт
3. Числові характеристики форми
    a. асиметрія
    b. ексцес

Центральної локації

Медіана — це елемент статистичного матеріалу, який ділить відповідний варіаційний ряд на 2 рівні за обсягом частини.

Якщо обсяг непарний, то

$$M_e = x_{(k+1)}$$

Якщо обсяг парний, то

$$M_e = \frac{x_k + x_{k+1}}{2}$$

```
public double Me
{
            get
            {
                int k = (int)Floor((float)size / 2);
                double[] series = this.GetVariationSeries();

                if (size % 2 == 0)// even
                {
                    return (series[k] + series[k + 1]) / 2;
                }
                else // odd
                {
                    return series[k];
                }
            }
}
```

Мода — елементи статистичного матеріалу, які найчастіше зустрічаються

1 2 3 3 4 4 4 5

$$M_o = 4$$

```
public double[] Mo
{
            get
            {
                int max = statisticalTable.Values.Max();
                return statisticalTable
                         .Where(elem => elem.Value == max)
                         .Select(elem => elem.Key)
                         .ToArray();
            }
}
```

Середнє арифметичне — сума всіх елементів статистичного матеріалу поділена на його обсяг

$$\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
                        public double _x =>

     statisticalTable.Sum(elem => elem.Key * elem.Value) / size;
```

Розсіяння

Девіація — сума квадратів відхилень елементів статистичного матеріалу від середнього квадратичного

$$Dev = \sum_{i=1}^{n} n_i * (x_i - \tilde{x})^2$$

```
public double Dev
{
            get
            {
                double xAvg = this._x;

                return statisticalTable.Sum(el => el.Value * Pow(el.Key - xAvg, 2));
            }
}
```

Варіанса — девіація поділена на число степенів вільності.

$$s^2 = \frac{Dev}{d.f.}$$

```
public double s2 => Dev / d_f;
```

Стандарт — арифметичний корінь з варіанси

$$s = \sqrt{s^2}$$

```
public double s => Sqrt(s2);
```

Варіація — відношення стандарту до середнього квадратичного

$$v = \frac{s}{\tilde{x}}$$

```
public double V => s / _x;
```

Інтерквантильні широти

Квантилем порядку α , *якщо він існує*, називається цей елемент статистичного матеріалу (відповідного варіаційного ряду), до якого включно маємо α% елементів статистичного матеріалу (відповідного варіаційного ряду).

Статистичний матеріал має квантилі тільки порядків кратних $\frac{100}{n}\cdot$, інші квантилі не існують.

Елемент x_i є кантилем порядку

$$x_{(i)} = i * \frac{100}{n} \ (i = 1, \dots, n)$$

При $\alpha < \beta$, різницю між квантилем порядку β і квантилем порядку α називають **інтерквантильною широтою** порядку $\beta - \alpha$.

Існують інтерквантильні широти наступних порядків:

$$q_{ij} = (j - i) * \frac{100}{n}, j > i \ (i = 1{,}2 \dots, n-1, j = 2{,}3 \dots n)$$

Квартиль (Q) — квантиль порядку 25, 50, 75
Октиль (O) — квантиль порядку 12.5, 25,…87.5

Децель (D) — квантиль порядку 10, 20, … 90

Центель (C) — квантиль порядку 1, 2, 3 … 98, 99

Міліль (M) — квантиль порядку 00.1, 00.2, …, 99.9

```
public double x(int i)
{
    if (i < 0 || i > size)
            throw new System
                        .IndexOutOfRangeException("current index is not allowed");

    return i * 100 / size;
}
public double q(float i)
{
            if (i % (100 / size) == 0)
            {
                float xI = (i * size) / 100;
                if (xI % 1 == 0)
                {
                    // -1 cuz indices start with 0 not with 1
                    return GetVariationSeries()[(int)xI - 1];
                }
                else
                {
                    throw new System
                         .IndexOutOfRangeException("something is wrong with your index");
                }
            }
            else
            {
                throw new System.ArgumentException("must be a multiple");
            }
}

public double Q(int i) => q(IL.Q(i));
public double D(int i) => q(IL.D(i));
public double O(int i) => q(IL.O(i));
public double C(int i) => q(IL.C(i));
public double M(int i) => q(IL.M(i));
```

```
private static class IL
{
            public static int Q(int i) => new int[] { 25, 50, 75 }[i - 1];
            public static int D(int i) => Enumerable.Range(1, 10)
                                              .Select(x => x * 10)
                                              .ElementAt(i - 1);
            public static float O(int i) => Enumerable.Range(125, 1000)
                                              .Where(x => x % 125 == 0)
                                              .Select(x => (float)x / 10)
                                              .ElementAt(i - 1);
            public static int C(int i) => Enumerable.Range(1, 100).ElementAt(i - 1);
            public static float M(int i) => Enumerable.Range(1, 1000)
                                              .Select(x => (float)x / 10)
                                              .ElementAt(i - 1);
}
```

Моменти

Моментом порядку H відносно сталої a називається вираз

$$\mu_H(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^H (H = 1,2 \dots)$$

При a = 0 момент називається **початковим** і позначається m_H

При a = x момент називається **центральним** і позначається M_H

1-ий початковий момент є *математичним сподіванням*

Перший центральний момент будь-якої випадкової змінної дорівнює 0.

Другий центральний момент будь-якої випадкової змінної є *дисперсією.*

```
public double Mh(int h, double a)
{
     if (h < 1 || h > size) throw new System.ArgumentException("Out range");

     return GetVariationSeries().Sum(el => Pow(el - a, h)) / size;
}
public double mh(int h)
{
     return Mh(h, 0);
}
public double Mh(int h)
{
     if (h == 1) return 0;

     return Mh(h, _x);
}
```

Форми

Асиметрія — це відношення 3-го центрального моменту до 2-го центрального моменту в степені півтора. Визначає скошеність статистичного матеріалу.

$$A_S = \gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

```
public double As => Mh(3) / Pow(Mh(2), 1.5);
```

Ексцесом (крутістю, сплющеністю) — це відношення 4-го центрального моменту до 2-го центрального моменту в квадраті мінус три

$$E_k = \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

```
public double Ek => Mh(4) / Pow(Mh(2), 2) - 3;
```

Схема статистичного доведення

Генеральна сукупність — сукупність усіх можливих значень випадкової змінної.

Вибірка з генеральної сукупності — ряд незалежних спостережень з випадкової величини.

Гіпотеза — твердження про генеральну сукупність на основі вибірки.

Статистичне доведення — міркування на основі яких приходимо до висновків про гіпотезу.

В кожному статистичному доведенні є наступні кроки:

- Формується гіпотеза H
- Вибирається рівень значущості d
- Вибирається відповідно гіпотеза статистики St
- Знаходиться розподіл цієї статистики
- На основі знайденого розподілу визначаємо критичну область для статистики.
- Знаходимо емпіричне значення статистики.
- Приймаємо вирішення про гіпотезу.
    - Якщо емпіричне значення статистичної гіпотези попадає в критичну для гіпотези область, то гіпотезу ***відкидаємо***.

- Якщо емпіричне значення статистичної гіпотези не попадає в критичну для гіпотези область, то гіпотезу ***приймаємо*** і вона не суперечить експериментальним даним.

Статистичні гіпотези відносно генеральної сукупності можуть бути дуже різноманітними. Наприклад: розподіл, його параметр, нерівність, сподівання і так далі.

Можливі 2 типи похибки:

- Відкидання правдивої гіпотези
- Прийняття хибної гіпотези

Саме рівень значущості задає імовірність допустити похибку 1 рівня.

Критерій узгодженості Пірсона

Нехай дана вибірка з генеральної сукупності і нам варто перевірити гіпотезу про вид розподілу, тобто про приналежність розподілу вибірки деякому параметричному сімейству.

$$X = (x_1, x_2, \dots, x_n)$$

$$H: F(x)$$

Зауважимо, що критерій Пірсона можна застосовувати при умовах:

- n > 4
- в кожному класі не менше 10 спостережень, в іншому разі класи варто об’єднати

Поділимо генеральну сукупність довільним чином на r+1 частин.

За міру відхилення теоретичного розподілу від вибірки Пірсон прийняв величину

$$\mathrm{X}^2(r, n, f) = \sum_{i=1}^{r+1} \frac{(m_i - np_i)^2}{np_i}$$

Також він довів, що для вибірок великого обсягу статистика має розподіл, який задається густиною

$$P_{x^2}(n)(x) = \begin{cases} 0; x < 0 \\ \frac{1}{\Gamma(\frac{r}{2})2^{\frac{r}{2}}} * x^{\frac{r}{2}-1} e^{-\frac{x}{2}}; x \geq 0 \end{cases}$$

Якщо $X^2_{\text{емпіричне}} > X^2_{\text{критичне}}$, то гіпотезу відхиляємо

Висновок

Отже, під час розв’язання індивідуального завдання з курсу «Математична статистика» я поглибив свої знання про основні поняття математичної статистики: статистичну змінну, методи формування вибірки, представлення статистичного матеріалу, числові характеристики статистичного матеріалу; та закріпив знання отримані раніше на лекційних та практичних заняттях. Окрім цього, я також перевірив, за допомогою критерію Пірсона, приналежність розподілу вибірки до нормального розподілу.

Усі набуті знання та вміння були вжиті безпосередньо на практиці при реалізації програмної аплікації.

Організував свою роботу відповідно до вимог індивідуального завдання, дотримуючись пунктів плану та структури звіту.