Data Analysis
Індивідуальне завдання 1

Бойченко Вікторія

1 Завантажити данні в електронну таблицю. Скласти інтервальний статистичний ряд (таблицю частот). Кількість інтервалів групування – формула Стерджесса. (1)

1. Скопіювавши дані в ексель, обираємо відповідний варіант (№1)
2. Складаємо варіаційний ряд (за зростанням)
3. Скласти інтервальний статистичний ряд (таблицю частот)

   (а) Розмах вибірки:

   $$w = x_{max} - x_{min} = 0,843 - 0,081 = 0,762$$

   (б) Число інтервалів групування за формулою Стерджесса:

   $$k = 1 + \log_2 n = 7,781 \approx 8$$

   (в) Довжина інтервалу групування:

   $$\Delta = \frac{w}{k} = \frac{0,762}{8} = 0,09525 \approx 0,1$$

	Manual	Functions
n	110	110
max	0,843	0,843
min	0,081	0,081
w	0,762	0,762
k	7,781359714	8
Δ	0,09792633	0,1

(г) $x_i^* = a_{i-1} + \frac{\Delta}{2}$ — середина і-го інтервалу, $i = \overline{1,k}$

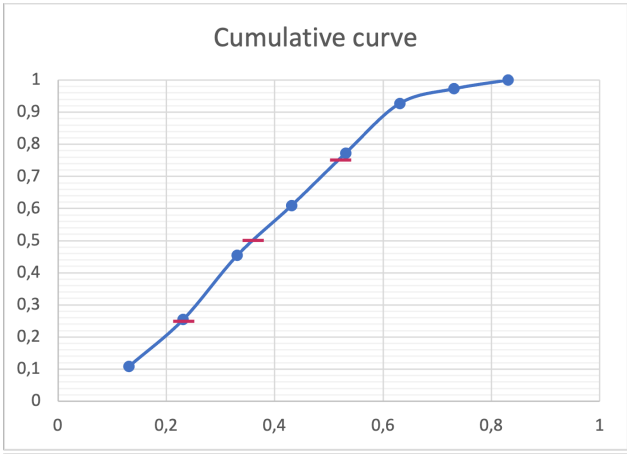
де $a_0 = x_{min},\ a_k = x_{max}$

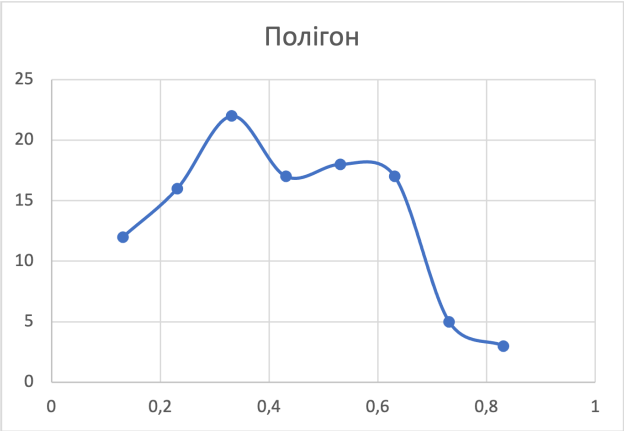
(д) n_i^* - кількість значень, що потрапляли у відповідний інтервал

(е) Формуємо згруповану вибірку:

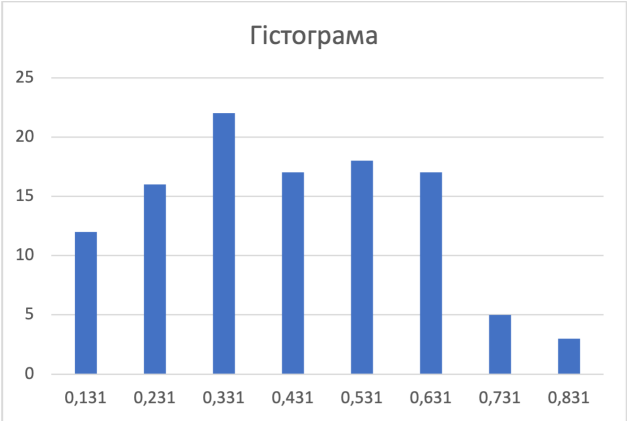
Номер інтервалу	1	2	3	4	5	6	7	8
Up.interval	0,081	0,181	0,281	0,381	0,481	0,581	0,681	0,781
down.interval	0,181	0,281	0,381	0,481	0,581	0,681	0,781	0,881
x_j^*	0,131	0,231	0,331	0,431	0,531	0,631	0,731	0,831
n_j^*	12	16	22	17	18	17	5	3
n_j / n	0,109090909	0,145454545	0,2	0,154545455	0,163636364	0,154545455	0,045454545	0,027272727
F^*	0,109090909	0,254545455	0,454545455	0,609090909	0,772727273	0,927272727	0,972727273	1

2 Візуалізувати дані. Побудувати полігон, гістограму, емпіричну функцію розподілу, кумулятивну криву, відмітити на ній медіану та квартилі. (2)






Емпірична функція розподілу
1,2
1
0,8
0,6
0,4
0,2
0
0,131
0,231
0,331
0,431
0,531
0,631
0,731
0,831


Гістограма
25
20
15
10
5
0
0,131
0,231
0,331
0,431
0,531
0,631
0,731
0,831

3 Обчислити числові характеристики центральної тенденції та розкиду: вибіркове середнє, дисперсію, середньоквадратичне відхилення, моду, медіану, коефіцієнти асиметрії та ексцесу. Для обчислення застосувати табл.1 з прикладу 1.(2)

	Manual	Functions
quartile 1/4	0,275	0,279
mediana	0,3975	0,3975
quartile 3/4	0,577	0,57575
interquart.range	0,302	0,29675
moda	0,319	0,319
aver	0,423018182	0,423018182
std.dev	0,187771772	0,187771772
var	0,035258238	0,035258238
kursity (екцесс)	-0,908389411	-0,85411353
skew	0,198382546	0,203910025

- Значення ексцесу < 3 характеризує плосковершинний розподіл.

- Коефіцієнт асиметрії є додатнім, тому це вказує на те, що розподіл має правий хвіст. Але є майже симетричним оскільки значення наближається до 0.

- За припущенням тоді вибіркове середнє має бути більше за медіану.

  $aver = 0,423 > 0,3975 = mediana$, тому це твердження виконується

4 Запропонувати своє бачення про природу даних і зробити висновок в предметній області. і скласти звіт. Завантажити на ДІСТЕДУ.(1)

Дані варіанту №1 - можна припустити, що це система нарахування новорічної премії до заробітньої платні в компанії "Київтрансгаз оскільки вони в межах від 0 до 1, невід'ємні. Для зручності можна перевести значення в відсотковий вигляд.

(або ще можна інтепретувати як знижки в супермаркеті на товари)

Висновок предметної області

В отриманому прикладі розглядалось 110 значень нарахувань премії до заробітньої платні.

Мінімальне нарахування = 8,1%

Максимальне нарахування = 84,3%

Вибіркове середнє нарахування = $\sim 42,3\%$, медіана = $\sim 39,75\%$, що вказує на правий хвіст у розподілі.

Отримані дані було переведено в інтервальний статистичний ряд з інтервалами в 10%. Найбільша кількість працівників мали нарахування в межах 28% - 38% = 22 людей.

Найменша кількість працівників мали нарахування в межах 68% - 83% = сумарно 8 людей (5 та 3 людей у відповідних інтервалах).

Інші нарахування були розподілені майже рівномірно по своїх інтервалах = від 12 до 18 людей.

Середнє квадратичне відхилення від середнього значення нарахування до 20% = 18,8%

Значення ексцесу -0,908 $<$ 3 характеризує плосковершинний розподіл.

Коефіцієнт асиметрії є додатнім, тому це вказує на те, що розподіл має правий хвіст.

За припущенням тоді вибіркове середнє має бути більше за медіану.

$aver = 0,423 > 0,3975 = mediana$, тому це твердження виконується.

Це можна інтерпретувати як більшість нарахувань помірні з невеликою кількістю дуже вигідних для працівників ($>$ 68,1% нарахувань).

Міжквартильний розмах = 57,7% - 27,5% = 30,2% показує нарахування премії для людей, що потрапили в середні 50% вибірки.

Найчастішим нарахуванням до заробітньої платні стало 31,9% (мода)-тричі