Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«СТАТИСТИЧНИЙ АНАЛІЗ ТА МАТЕМАТИЧНІ ПІДХОДИ ДО ПЛАНУВАННЯ ТА ПРОВЕДЕННЯ СОЦІОЛОГІЧНОГО ДОСЛІДЖЕННЯ»**

Виконав: студент 4-го року навчання
освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Матвієнко Олег Максимович

Керівник: Щестюк Н. Ю..,
кандидат фіз.-мат. наук, ст. викладач

Рецензент_____

(прізвище та ініціали)

Кваліфікаційна робота захищена

з оцінкою _____

Секретар ЕК _____

«____» _____ 20____ р.

Київ – 2021

ЗМІСТ

Вступ

Соціологічне дослідження є фундаментальним процесом для нашого суспільства. Завдяки йому ми можемо отримати чітке розуміння про те, що для людей важливо, що їх цікавить, яке відношення вони мають до певних питань та ситуацій. Статистика ж допомагає нам розуміти основні тенденції та бажання опитуваних, що дозволяє (наприклад, керівництву країни) розуміти який вектор розвитку внутрішньої та зовнішньої політики бачить населення, як населення відноситься до пандемії і вакцинації, чи є зв'язок між певними даними, тощо.

Перші роботи зі статистики датуються аж 1663 року, публікація «Природні та політичні спостереження на списках померлих» Джона Ґраунта. Перш за все статистичне мислення використовувалось для потреб держав щоб робити висновки на тему правильності політики на основі економічних та демографічних даних. З тих пір опубліковано велику кількість наукових робіт із статистики, присвячених як теоретичним дослідженням, так і прикладним проблемам.

Минулий 2020 рік, беззаперечно, можна вважати роком небачених викликів перед людством. 31 грудня, уряд Китайської Народної Республіки повідомив Всесвітню організацію охорони здоров'я (ВООЗ) про виявлення невідомого захворювання в провінції Ухань. Згодом ця хвороба отримала назву короновірусна хвороба COVID – 19, спричинена SARS-CoV-2. Проблемі протиепідеміологічних заходів та вакцинації потребують соціологічних досліджень із статистичною обробкою результатів з метою виявлення відношення до тих чи інших заходів держави і дослідження тих чи інших залежностей між різноманітними чинниками.

Уміння застосовувати статистичний аналіз в різноманітних предметних областях є однією із невідємних компетентностей прикладного математика.

Отже **мета** мого дослідження: проведення статистичного аналізу в предметній області а саме для соціологічного опитування на тему "Відношення

жителів до можливого посилення карантинних обмежень, пов'язаних з пандемією COVID – 19”, подальшого аналізу отриманих даних, виявлення закономірностей та залежностей, перевірка гіпотез.

Поставлена мета зумовила наступне наукове завдання.

1. Провести соціологічне опитування на тему, що цікавить. Тести для опитування мають бути схвалені спеціалістами з предметних областей.
2. Провести візуалізацію даних вибірки, представити оглядову статистику. Висловити гіпотези щодо однорідності думок, залежності факторів.
3. Зробити перевірку гіпотез статистичними методами.
4. Провести кореляційний та регресійний аналіз.

Робота складається з трьох розділів, висновків і літератури.

РОЗДІЛ 1. ОПИСОВА СТАТИСТИКА ТА ВІЗУАЛІЗАЦІЯ ДАНИХ

1.1 Збір і підготовка даних для статистичного аналізу

Для проведення соціологічного дослідження було здійснено вибіркове обстеження, яке містить наступні етапи:

1. Потреба в статистичній інформації в предметній області
2. План вибіркового обстеження
3. Обробка, очищення та аналіз даних (побудова моделей, перевірка гіпотез)
4. Трактування, презентація та використання результатів

Збір даних відбувався за допомогою онлайн опитування, яке доступне за посиланням https://docs.google.com/forms/d/1aXvNh0_wwj2XDqqwAQQEMD8KpLQK9mTRMDLYjPi841c/edit.

Вибірка склала 370 осіб і була сформована по принципу цільової кластерної вибірки, оскільки це дозволяє опитати виключно жителів мікрорайону з різними соціально-демографічними характеристиками.

Генеральна сукупність - 10 000 чоловік

Довірча ймовірність – 95%

Довірчий інтервал (похибка) – 4,5 %

Вибірка – 370 чоловік.

Для репрезентативності вибірка має бути створена за одним із наступних методів:

1. Простий випадковий вибір
2. Вибір за зміною
3. Вибір без заміни
4. Стратометричний вибір
5. Кластерний вибір
6. Систематичний вибір

В нашому випадку вибірка не є репрезентативною оскільки опитування проводилось онлайн і не всі соціально-демографічні групи змогли взяти участь у опитуванні. Проте це не вплинуло на якість статистичного аналізу даних.

Отримані відповіді анкетування містять як кількісні, так і категоріальні дані. Оскільки результати мали деяку пропущену інформацію, такі рядки було видалено, або поповнено найчастішими відповідями, якщо це не суперечило загальності дослідження.

Анкета містила 16 запитань, які стосувались як загальних даних про людину, так і карантинних обмежень.

1. Ваша стать? (**Q1**)
    a. Жінка (1)
    b. Чоловік (0)
2. Ваш вік? (**Q2**)
    ---
3. Ваш щомісячний рівень доходу, грн.? (**Q3**)
    a. До 5 000 (1)
    b. 5 000 – 12 000 (2)
    c. 12 000 – 20 000 (3)
    d. Більше 20 000 (4)
4. Сімейний стан? (**Q4**)
    a. Вдівець /вдова (1)
    b. Одружений / заміжня (2)
    c. Не одружений / не одружена (3)
    d. У відносинах (4)
5. Діти? (**Q5**)
    a. Немає (0)
    b. 1 дитина (1)
    c. 2 дитини (2)
    d. 3 дитини і більше (3)

6. Освіта? (**Q6**)
    a. Середня (1)
    b. Середньо-спеціальна (2)
    c. Студент (3)
    d. Вища (4)
7. Занятість? (**Q7**)
    a. Безробітній (0)
    b. Пенсіонер (1)

c. Навчаюсь (2)
d. Працюю не повний день (3)
e. Працюю (4)
f. Підприємець/власник бізнесу (5)

8. Чи слідкуєте Ви за ситуацією з поширенням COVID – 19? (**Q8**)
   a. Ні (0)
   b. Так (1)
   c. Інколи (2)
9. Чи знаєте ви перші симптоми COVID – 19 і алгоритм дій, якщо запідозрили їх в себе? (**Q9**)
   a. Ні (0)
   b. Так (1)
   c. Мені байдуже (2)
10. Чи дотримуєтесь Ви протиепідемічних заходів (виберіть декілька варіантів)? (**Q10**)
    a. Ношу маску. (0)
    b. Ношу захисні рукавичкі (1)
    c. Регулярно дезинфікую руки (2)
    d. Дотримуюсь дистанції в 1.5 м (3)
    e. Інше (4)
11. Чи хворіли Ви на COVID – 19? (**Q11**)
    a. Ні (0)
    b. Так (1)
12. Чи є серед членів Вашої родини, близьких родичів та знайомих хворіють/перехворілі на COVID – 19? (**Q12**)
    a. Ні (0)
    b. Так (1)
13. Чи відчули Ви на собі економічні наслідки введення карантину, якщо так, то які саме? (**Q13**)
    a. Залишився без роботи (0)
    b. Затримують заробітну плату (1)
    c. Зменшилась заробітна плата/прибутки (2)
    d. Скоротили кількість робочих днів (3)
    e. У вимушеній відпустці (4)
    f. Не відчув (5)
    g. Інше (6)
14. Як ви оцінюєте діяльність уряду по протидії поширення Covid - 19? (**Q14**)
    a. дуже НЕ ефективними (1)
    b. скоріше НЕ ефективними (2)
    c. скоріше ефективними (3)
    d. ефективними (4)
    e. важко відповісти (5)
15. Середній дохід (грн/місяць) (**Q15**)

    ---

16. Чи потрібно посилювати карантин задля протидії Covid – 19 (**Q16**)

a. Відмінити всі карантинні заходи (1)
b. Достатньо буде карантину вихідного дня (2)
c. Залишити тільки особисті обмеження (маски, рукавички, соціальна дистанція), не обмежуючи роботу громадського транспорту, закладів освіти, харчування, розваг тощо (3)
d. Потрібно введення жорсткого карантину, за прикладом того, що був весною (4)

Маємо 16 змінних (Q1-Q16), дві з яких кількісні (Q2, Q15), решта - категоріальні та 357 записів.

1.2 Числові характеристики і візуалізація даних.

Дані бувають таких видів:

- Кількісні
- Категоріальні

Кількісні ж бувають:

- Дискретні
- Неперервні

А категоріальні:

- Невпорядковані
- Бінарні
- Впорядковані

Статистичний аналіз починається з характеристик центральної тенденції і розсіювання та візуалізації.

$\acute{x} = \frac{\Sigma x}{N}$ - середнє значення вибірки

Медіана ж ділить вибірку навпіл, а мода показує значення яке найчастіше трапляється в вибірці.

Квантиль відсікає у вибірці певну його частину.

Медіана – це квантиль порядку $\frac{1}{2}$

Квартилі – квантилі порядку $\frac{1}{4}, \frac{2}{4}$ та $\frac{3}{4}$.

Децилі – квантилі порядку $\frac{1}{10}, \frac{2}{10}, \ldots \frac{9}{10}$

Процентилі – кватилі порядку $\frac{1}{100}, \frac{2}{100}, \ldots \frac{99}{100}$

Якщо за характеристику центральної тенденції ми вибираємо середнє, то за характеристику розсіювання доцільно вибрати середнє квадратичне.

Розмах вибірки (range) – це різниця між найбільшим та найменшим елементом вибірки. Завдяки їй ми можемо оцінити розкид елементів у виборці.

Міжквартильний розмах (mid – spread) – різниця між нижнім та верхнім квартилем. Він дозволяє оцінити розкид 50% елементів вибірки і не враховує вплив викидів.

Почнемо описову статистику з кількісних змінних.

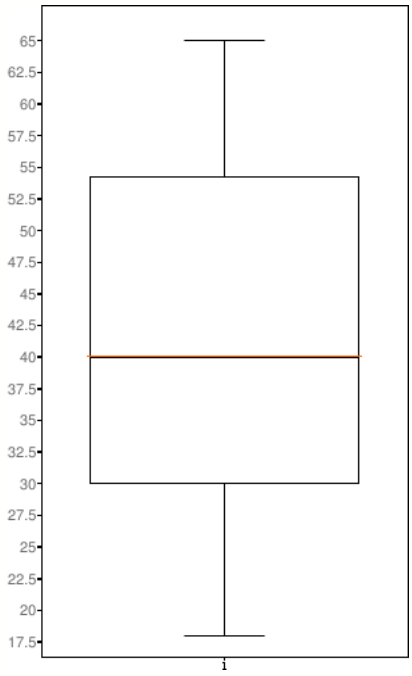
Числові характеристики кількісної змінної **Q2 було отримано за допомогою функції** `df.describe()` **бібліотеки** `pandas` **мови Python**:

```
Minmax = (18, 65)
mean = 41.389189
std = 13.577636
variance = 184.35218633
median = 40.0
ModeResult(mode=array([[55]]), count=array([[15]]))
skewness = 0.06310562
kurtosis = -1.20033886
Quantile = array([30., 40., 54.])
```

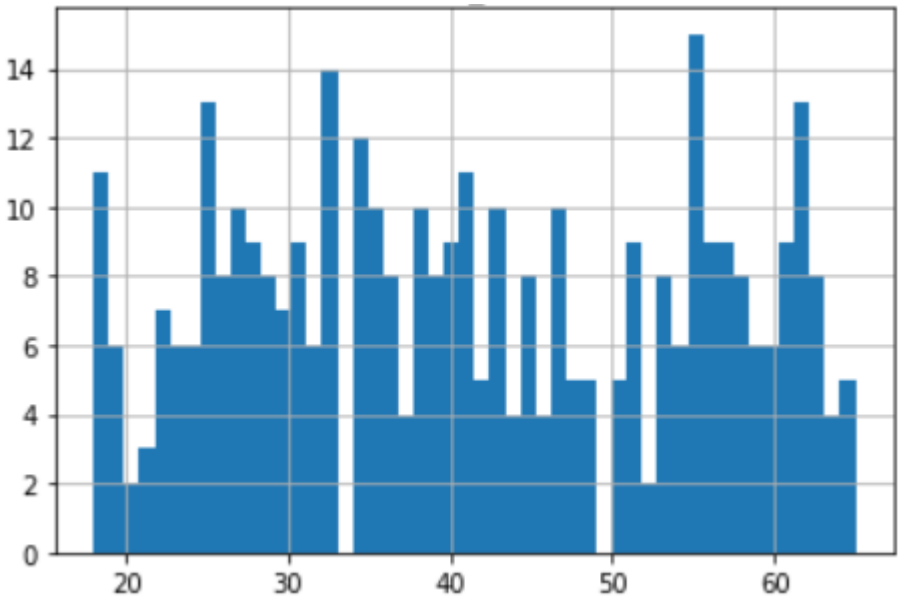
В нашому досліджені приймали участь респонденти, мінімальний вік яких є 18 років, максимальний вік 65, а середній вік 41. Мода наших респондентів є 55 років.

Середнє, мода та медіана не співпадють. Це свідчить про віддаленість до нормального розподілу. Коефіцієнт асиметрії додатній, а значить правий хвіст розподілу довший за лівий. А коефіцієнт ексцесу свідчить що він відрізняється

від нормального розподілу, бо не = 0. Оскільки є від'ємним, то пік розподілу є плоскішим за нормальний.



Боксплот з кроком 2.5 .
На ньому ми можемо оцінити середнє значення, найбільше та найменше значення, які знаходяться в межах міжквартильного розмаху.



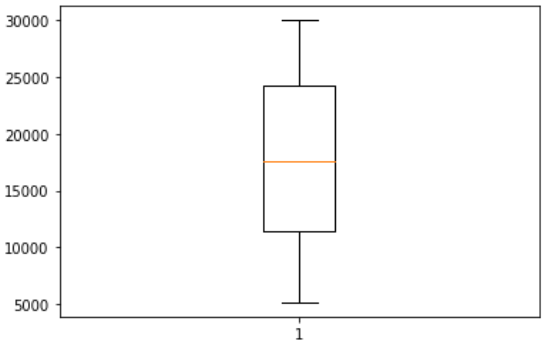
Гістограма частот з кроком 10.
На цій гістограмі ми можемо побачити що у нашій виборці зовсім не було респондентів віком 34 та 49 років. Розподіл нагадує дзіноподібну криву

Q15 було отримано за допомогою функції `df.describe()` **бібліотеки** `pandas` **мови Python**:

```
Minmax = (5149, 29994)
mean = 17588.425770308124
std = 7233.15465
variance = 52318526.194614924
median = 17644.0
ModeResult(mode=array([5701]), count=array([2]))
skewness = 0.02127201166054205
kurtosis = - 1.2188306614558682
Quantile = array([11420., 17644., 24183.])
```

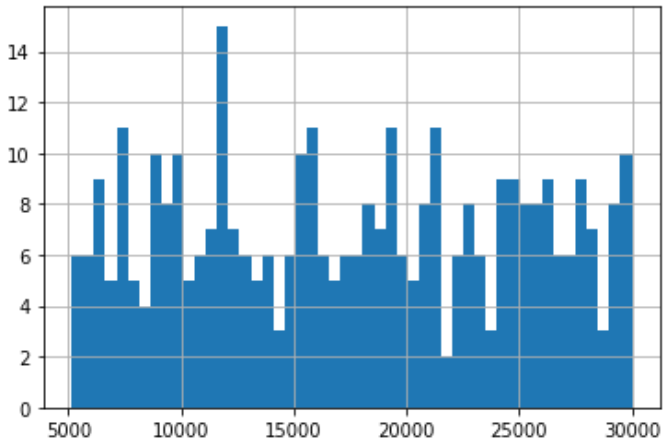
Як ми бачимо, мінімальна зарплатня наших респондентів складає 5149 грн, максимальна 29994 грн, а середня 17588 грн. Мода наших респондентів складає 5701 грн.

Ще ми бачимо, що середнє, мода медіана також не співпадють. Це також свідчить про віддаленість нормального розподілу. Коефіцієнт асиметрії знову додатній (правий хвіст розподілу довший за лівий). Коефіцієнт ексцесу від'ємний, пік розподілу плоскіший за нормальний.



Боксплот з кроком 5000.

На ньому ми бачимо що мінімальне значення 5149, а максимальне 29994. Медіана дорівнює 17644

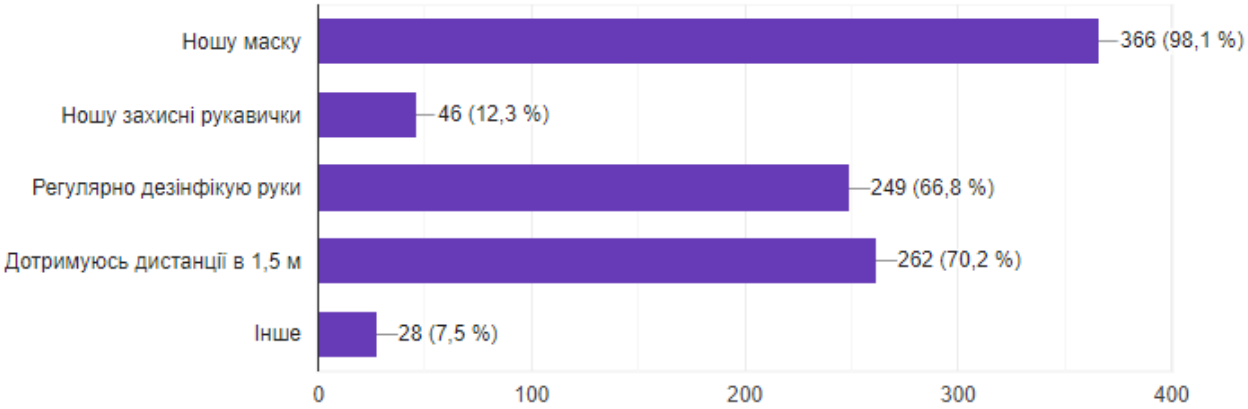


Гістограма частот з кроком 5000.

На цій гістограмі ми бачимо, що найменша кількість респондентів має заробітну плату 22000 грн. Розподіл нагадує дзвіноподібну криву.

Для аналізу категоріальних даних робимо кругові та стовпчикові діаграми.

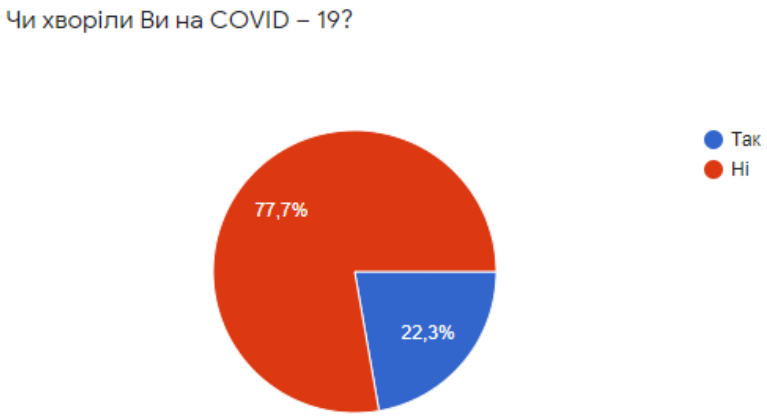
Для змінної **Q10** ми побудували стовпчикову діаграму.



На ній ми бачимо, що:

1. 98.1 % респондентів носять маски
2. 12.3 % респондентів носять захисні рукавички
3. 66.8 % респондентів регулярно дезінфікують руки
4. 70.2 % респондента дотримуються дистанції в 1.5 м

Для змінної **Q11** ми побудували кругову діаграму.



На ній ми бачимо, що:

- 290 респондентів не хворіли ковідом
- 83 респондента хворіли ковідом

Для змінної **Q14** ми також побудували кругову діаграму.

Вона показує, що:

1. 38.4 % респондентів вважають діяльність уряду скоріше НЕ ефективною.
2. 32 % респондентів вважають діяльність уряду дуже НЕ ефективною.
3. 15.9 % респондентів не можуть відповісти.
4. 12.6 % респондентів вважає діяльність уряду скоріше ефективною.

Висновки

Отже, в результаті описової статистики можемо сказати про наших респондентів:

1. У нашому опитуванні брало участь 219 жінок та 151 чоловік
2. Найбільша кількість респондентів мають вік в діапазоні 35- 49 років
3. 31.8 % респондентів мають заробітну плату більше 20000 грн
4. 63.4 % наших респондентів одружені / заміжні
5. 35.1 % респондентів мають одну дитину
6. 76.5 % респондентів мають вишу освіту
7. 58.2 % респондента працюють
8. 76.1 % респондентів слідкують за ситуацією з поширенням COVID – 19
9. 92.8 % респондентів знають перші симптоми COVID – 19 і алгоритм дії проти нього.
10. 98.1 % респондентів носять маски
11. 77.7 % респондентів не хворіли ковідом
12. 82.5 % респондентів мають родичів / знайомих які переболіли / хворіють ковідом
13. 40.9 % респондентів зменшилась заробітна плата / прибуток
14. 38.4 % респондентів вважають діяльність уряду скоріше НЕ ефективною.
15. 41.7 % респондентів вважають що потрібно залишити тільки особисті обмеження

Надалі будемо застосовувати вивідну статистику, щоб перевірити гіпотези про знайдені пропорції та дійти висновку чи однаково чоловіки і жінки реагують на посилення карантинних заходів.

РОЗДІЛ 2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

2.1 Гіпотези про нормальний розподіл

Перевірка статистичних гіпотез займає провідне місце у статистичному анілізі.

Ми маємо 5 основних стадій при перевірці гіпотез:

1. Створення нульової та альтернативної гіпотези
2. Підбір необхідних даних.
3. Обчислення статистики критерія
4. Визначення критичної області, перевірка критерія на потрапляння та критичну область.
5. Інтерпретація рівня значимості та результатів.

При формулюванні гіпотези маємо на увазі, що певний елемент генеральної сукупності знаходиться в певному інтервалі. Також дуже важливо визначати рівень статистичної значимості.

Перевірка статистичних гіпотез завжди пов’язана з ризиком прийняття помилкового рішення. Це може статися у двох випадках:

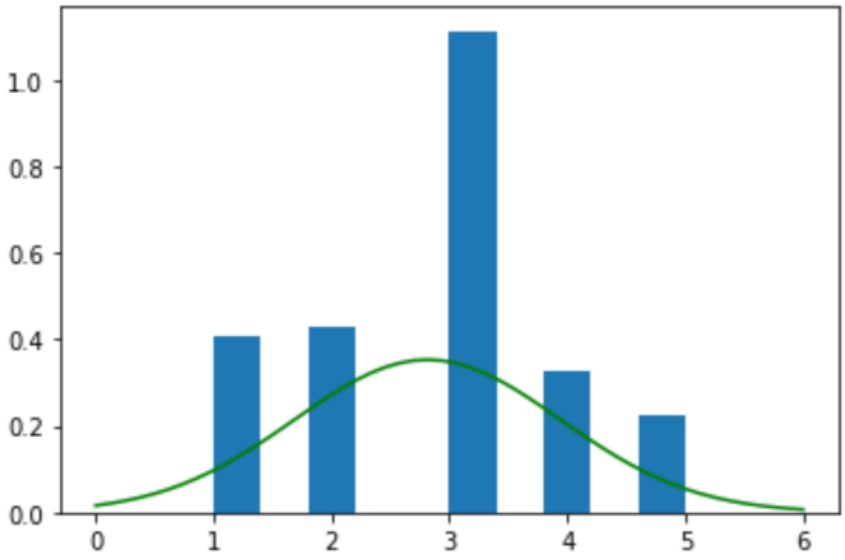
- Помилка першого роду (нульова гіпотеза відхиляється, хоча є правильною)
- Помилка другого роду (приймається нульова гіпотеза, хоча правильною є альтернативна).

Розглянемо гіпотези про нормальний розподіл на прикладі кількісної змінної.

Розглянемо змінну **Q2** (вік). Оскільки з описової статистики ми побачили, що гістограма має “дзвіноподібну” форму, то будуємо на одному графіку гістограму і щільність теоретичного нормального розподілу.

Для цього у Пайтоні застосовуємо

```
a, sigma  =scipy.stats.distributions.norm.fit(X2)
ix = np.linspace(0,6,50)
N_fitted_X2 = scipy.stats.distributions.norm.pdf(ix, a, sigma)
plt.hist(X2, density=True)
plt.plot(ix,N_fitted_X2,'g')
```



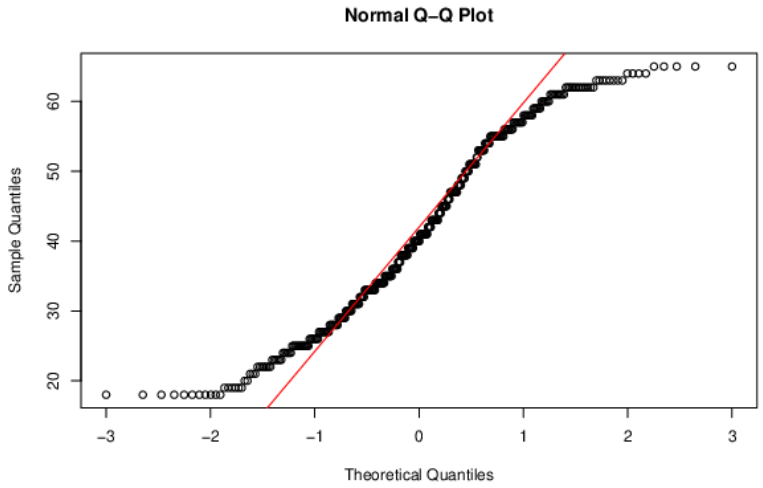
Параметри нормального розподілу це середнє та середнє квадратичне відхилення.

Щільність виглядає $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-36.67)^2}{2\sigma^2}}$

Застосуємо ще один засіб візуалізацію відповідності вибірки нормальному розподілу: Q-Q.

Для цього в Пайтон застосовуємо

```
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot
qqplot(MS, line='s')
pyplot.show()
```



З цього графіку видно що точки лежать навколо червної лінії і тільки на хвостах вони відрізняються.

З гістограми та цього графіку робимо припущення про відповідність даних вибірки нормальному розподілу.

```
H0: the sample has a Gaussian distribution.
H1: the sample does not have a Gaussian distribution.
```

Для перевірки в Пайтон використовуємо так3 бібліотеку

```
scipy.stats
```

Яка дає можливість застосувати різні крітерії.

- Критерій Шапіро – Уилка заснований на оптимальній лінійній незміщеності оцінки дисперсії до її звичайної оцінки методом максимальної правдоподібності.

Статистика:

$$W = \frac{1}{S^2}\lfloor\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\rfloor^2$$

де $s^2 = \sum_{i=1}^{n}$

- Д'Агостіно наступну статистику для перевірки розподілу

$$Y = \sqrt{n}\frac{D - 0{,}28209479}{0.02998598}$$

де $D = \frac{T}{n^2 s}, T = \sum_{i=1}^{n}\left\{i - \frac{n+1}{2}\right\}x_i, x_1 \le \cdots \le x_n; s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

- Критерій Пірсона Хі – квадрат це непараметричний метод, що дозволяє оцінити значимість відмінностей між фактичною кількістю випадків.

Статистика:

$$x_n^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$

Тест Шапиро-Уилка:

Для цього в Пайтон застосовуємо функцію

```
from scipy.stats import shapiro
stat, p = shapiro(MS)
print('stat W=%.3f, p=%.3f' % (stat, p))
alpha = 0.05
if p < alpha:
  print('The null hypothesis can be rejected. Probably not Gaussian')
else:
  print('The null hypothesis cannot be rejected. Not enough evidence to say data is not normal. Probably Gaussian')
```

Результат:

```
stat W=0.894, p=0.000
The null hypothesis can be rejected. Probably not Gaussian
```

Тест Д'Агостіно:

Для цього в Пайтон застосовуємо функцію

```
k2, p = scipy.stats.normaltest(MS)
alpha = 0.05
print('stat K_2=%.3f, p=%.3f' % (k2, p))
if p < alpha:  # null hypothesis: x comes from a normal distribution
     print("The null hypothesis can be rejected. Probably not Gaussian")
else:
     print("The null hypothesis cannot be rejected. Probably Gaussian")
```

Результат:

```
stat K_2=4.921, p=0.085
The null hypothesis cannot be rejected. Probably Gaussian
```

Тест Пірсона Хі-квадрат:

Для цього в Пайтон застосовуємо функцію

```
from scipy.stats import chisquare
Statistics_Xi_2, p=chisquare(X2)
alpha = 0.05
print('stat Xi_2=%.3f, p=%.3f' % (Statistics_Xi_2, p))
if p < alpha:  # null hypothesis: x comes from a normal distribution
     print("The null hypothesis can be rejected. Probably not Gaussian")
else:
     print("The null hypothesis cannot be rejected. Probably Gaussian")
```

Результат:

```
stat Xi_2=161.910, p=1.000
The null hypothesis cannot be rejected. Probably Gaussian
```

Як ми бачимо, два тести з трьох показали, що дані по віку розподілені нормально.

2.2 Гіпотези про однорідність

Гіпотези про однорідність можуть бути реалізовані на основі параметричних і непараметричних критеріїв для залежних та незалежних вибірок. Отже гіпотези про однорідність – це гіпотези про схожість або відмінність виборок.

Часто у соціологічних дослідженнях нас цікавить питання, чи однаково жінки та чоловіки реагують на ті, чи інші проблеми. Тобто чи однорідні вони в своєму ставленні до введння карантину, ..., тощо. Що означає, чи можемо ми за якоюсь ознакою (наприклад, відношення до посилення карантину) вважати, що жінки і чоловіки відносяться до однієї генеральної сукупності.

Для перевірки цього ми беремо, наприклад, змінну Q16 (відношення до посилення карантину) і групуємо за ознакою чоловік чи жінка. Отримаємо дві незалежні вибірки: Q16_W та Q16_M. Перша вибірка має об'єм 120 жінок , а друга 250 чоловіків.

Застосуємо критерії однорідності даних для незалежних вибірок.

- Критерій Стьюдента для незалежних вибірок полягає у перевірки рівності значень у двох вибірках.
  Статистика:
$$t = \frac{|M_1 - M_2|}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

- Критерій Мана –Уітні для незалежних вибірок використовується для оцінки різниці між двома вибірками за рівнем будь – якої ознаки.
  Статистика:
$$U = n_1 * n_2 + \frac{n_2(n_x + 1)}{2} - T_x$$

- Критерій Вілкоксона для незалежних вибірок використовується для перевірки відмінностей між двома вибірками незалежних або парних вимірювань за рівнем будь – якої кількісної ознаки.
  Статистика:
  min - $W = \frac{n(n+1)}{2}$, де $n - об'ємдругоївибірки$
  max - $W = \frac{n(n+1)}{2+mn}$, де m – об'єм першої вибірки

Тест Стьюдента

Для цього використовуємо в Пайтон функцію

```
tstat=stats.ttest_ind (Q16_W, Q16_M)
alpha = 0.05
print('t-statistics=%.3f, p-value=%.3f' % (tstat.statistic, tstat.pvalue))
if tstat.pvalue > alpha:
    print("Прийняти гіпотезу про однорідність")
else:
    print("Відхилити гіпотезу про однорідність")
```

Результат:

```
t-statistics=-31.022, p-value=0.000
Відхилити гіпотезу про однорідність
```

Тест Мана – Уітні

Для цього використовуємо в Пайтон функцію

```
mwstat=stats.mannwhitneyu(Q16_W, Q16_M)
alpha = 0.05
print('mw-statistics=%.3f, p-value=%.3f' % (mwstat.statistic, mwstat.pvalue))
if mwstat.pvalue > alpha:
    print("Прийняти гіпотезу про однорідність")
else:
    print("Відхилити гіпотезу про однорідність")
```

Результат:

```
mw-statistics=26.033, p-value=0.000

Відхилити гіпотезу про однорідність
```

Тест Вілкоксона (для перших 25 елементів вибірок).

Для цього використовуємо в Пайтон

```
wstat=stats.wilcoxon(Q16_M, Q16_W)
alpha = 0.05
print('w-statistics=%.3f, p-value=%.3f' % (wstat.statistic, wstat.pvalue))
if wstat.pvalue > alpha:
    print("Прийняти гіпотезу про однорідність")
else:
    print("Відхилити гіпотезу про однорідність")
```

```
w-statistics=36.500, p-value=0.000
Відхилити гіпотезу про однорідність
```

Дійсно, чоловіки та жінкм однаково реагують на посилення карантину. Отже ми можемо вважати що вони взяті з однієї генеральної сукупності.

2.3 Гіпотеза про пропорції

Перевіримо нульову гіпотезу «Кількість людей, що дотримуються протиепідеміологічних засобів дорівнює 50% опитаних» проти конкуруючої гіпотези «Кількість людей, що дотримуються протиепідеміологічних засобів більше 50% опитаних» на рівні значущості 0,05.

Вибірковий розподіл пропорцій має біноміальний розподіл. Однак якщо обсяг вибірки n розумно великий, тоді вибірковий розподіл пропорції приблизно нормальний з середнім.

Формула довірчого інтервалу для пропорції:

$$p \pm Z_{95\%} * se(\overline{p}), \text{де} se(\overline{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Рівень довіри α=0,05

Тестову статистику обчислюємо за формулою:

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{p(p-1)}{n}}}$$

Отже застосовуючі в Пайтон бібліотеку scipy.stats ми отримали результат.

p value = -6,5215643134.

Порівняємо її з рівнем довіри. α > p value

Отже нульову гіпотезу відхилено. Кількість людей, що дотримуються протиепідеміологічних засобів більше 50% опитаних.

РОЗДІЛ 3. КОРЕЛЯЦІЙНИЙ ТА РЕГРЕСІЙНИЙ АНАЛІЗ

3.1 Кореляційна матриця

Для вивчення залежностей між випадковими змінними в математичній статиці було введено таке поняття як кореляція. Кореляція – це залежність, коли будь – якому значенню однієї змінної величини може відповідати декілька різноманітних значень іншої змінної. Кореляція дуже корисна, бо вона указує на відношення що має передбачуваний характер і тому вона має велику практичну цінність. Кореляція може бути як додатньою, так і від'ємною. Додатня кореляція тоді, коли збільшення однієї змінної пов'язане зі збільшенням іншої і коефіцієнт кореляції додатній. Від'ємна ж кореляція тоді, коли збільшення змінної пов'язане зі зменшенням іншої і коефіцієнт кореляції від'ємний. Найвідоміша міра залежності двох величин є коефіцієнт кореляції Пірсона. Також його називають спрощено коефіцієнтом кореляції. Його розрахунок має вигляд відношення коваріації двох випадкових величин на добуток їх стандартних відхилень.

Нехай X та Y – випадкові величини за математичними сподіваннями μ_X та μ_y і стандартним відхиленням σ_X та σ_Y.

Тоді:

$$\rho(X,Y) = corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X-\mu_x)(Y-\mu_y)\right)}{\sigma_X \sigma_Y}$$

де:

- Cov(X,Y) – це коваріація
- E – оператор мат. Сподівання

Визначеною кореляція Пірсона є тільки тоді, коли обидва стандартні відхилення є скінченними і не дорівнюють 0. Це наслідок нерівності Коші – Буняковського. Вона встановлює що кореляція не може перевищувати 1 в абсолютному значенні. Кореляція Пірсона дорівнює:

- +1 (зростаюча лінійна залежність)
- -1 (спадаюча лінійна залежність)

У всіх інших випадках вона дорівнює значенню у проміжку (1; - 1) і означає ступінь лінійної залежності. Чим ближче значення до 0, тим слабкіший зв'язок між величинами. Чим ближче значення до +1 та -1, ти сильніший зв'язок між величинами.

У Пайтоні для знаходження коефіцієнта Пірсона використовують бібліотеку scipy.stats і функцію

```
corr,p=stats.pearsonr(X5,X4)
print('Pearson_corr=%.3f, p-value=%.3f' % (corr, p))
if p < alpha:
    print("Прийняти гіпотезу про наявність зв'язку")
else:
    print("Відхилити гіпотезу про наявність зв'язку")
```

Доцільним є почати вивчення залежностей з формування матриці кореляцій Пірсона між всіма змінними.

У Пайтоні для цього існує бібліотека scipy.stats і функцію

```
corr=df.corr()
corr.style.background_gradient(cmap='coolwarm')
```

Кореляційна матриця по всіх змінних має наступний вигляд:

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1	1.000000	-0.106668	-0.145843	0.107053	-0.048781	-0.001751	-0.024903	-0.022086	0.119969	0.051430	-0.053474	0.022094	-0.091891	0.154380	0.014235	-0.084435
Q2	-0.106668	1.000000	0.052393	-0.571894	0.565490	-0.008742	-0.075927	-0.076408	-0.208779	0.078332	0.062513	0.134399	0.058484	-0.000267	-0.046109	0.102616
Q3	-0.145843	0.052393	1.000000	-0.116965	0.143582	0.421618	0.602291	-0.023376	0.157391	0.121900	0.094751	0.091625	-0.182262	-0.270720	-0.025142	0.028653
Q4	0.107053	-0.571894	-0.116965	1.000000	-0.530745	-0.029881	0.036424	0.063811	0.207580	-0.117359	-0.016691	-0.146885	-0.016298	-0.034824	0.021630	-0.130125
Q5	-0.048781	0.565490	0.143582	-0.530745	1.000000	0.023236	0.057354	-0.042422	-0.059932	0.108902	0.140904	0.176307	-0.002806	0.020592	0.018472	0.019382
Q6	-0.001751	-0.008742	0.421618	-0.029881	0.023236	1.000000	0.313332	0.032413	0.151824	0.205596	0.180519	0.075980	-0.084110	-0.254631	0.019442	-0.011130
Q7	-0.024903	-0.075927	0.602291	0.036424	0.057354	0.313332	1.000000	0.022458	0.130981	0.017265	0.093922	0.008987	-0.219368	-0.161601	0.038246	-0.105544
Q8	-0.022086	-0.076408	-0.023376	0.063811	-0.042422	0.032413	0.022458	1.000000	-0.122608	-0.149500	-0.116242	0.054927	0.020089	0.058343	-0.079173	-0.064632
Q9	0.119969	-0.208779	0.157391	0.207580	-0.059932	0.151824	0.130981	-0.122608	1.000000	0.120787	0.126126	-0.059297	-0.081341	-0.070509	0.001028	-0.140686
Q10	0.051430	0.078332	0.121900	-0.117359	0.108902	0.205596	0.017265	-0.149500	0.120787	1.000000	0.075917	0.166896	0.051134	0.003289	-0.018221	0.172011
Q11	-0.053474	0.062513	0.094751	-0.016691	0.140904	0.180519	0.093922	-0.116242	0.126126	0.075917	1.000000	0.206610	-0.029438	-0.078072	0.022907	-0.006377
Q12	0.022094	0.134399	0.091625	-0.146885	0.176307	0.075980	0.008987	0.054927	-0.059297	0.166896	0.206610	1.000000	0.006518	0.026264	-0.001401	0.108624
Q13	-0.091891	0.058484	-0.182262	-0.016298	-0.002806	-0.084110	-0.219368	0.020089	-0.081341	0.051134	-0.029438	0.006518	1.000000	0.095845	-0.023731	0.117793
Q14	0.154380	-0.000267	-0.270720	-0.034824	0.020592	-0.254631	-0.161601	0.058343	-0.070509	0.003289	-0.078072	0.026264	0.095845	1.000000	-0.029497	-0.046736
Q15	0.014235	-0.046109	-0.025142	0.021630	0.018472	0.019442	0.038246	-0.079173	0.001028	-0.018221	0.022907	-0.001401	-0.023731	-0.029497	1.000000	0.074797
Q16	-0.084435	0.102616	0.028653	-0.130125	0.019382	-0.011130	-0.105544	-0.064632	-0.140686	0.172011	-0.006377	0.108624	0.117793	-0.046736	0.074797	1.000000

Як ми бачимо, залежна змінна (Q16) не має лінійного зв’язку с жодною змінною(Q1-Q15), оскільки всі коефіцієнти кореляції менші за $|R| < 0.2$. Проте спостерігається коефіцієнт кореляції 0.602291 між змінними Q3 та Q7, 0.56549 між змінними Q2 та Q5, -0.571894 між змінними Q2 та Q4, -0.530745 між змінними Q4 та Q5, що свідчить про мультиколінеарність в системі. Зрозуміло, що варто перевірити гіпотези про наявність (відсутність) лінійного зв’язку між цими парами змінних. А також варто знайти коефіцієнти рангової кореляції Спірмена та Кендела.

3.2 Перевірка на наявність лінійної та рангової кореляції

Коефіцієнт рангової кореляції Спірмена – це непараметрична міра статистичної залежності між двома змінними. Його суть в оцінки наскільки добре можна описати відношення між двома змінними за допмогою монотонної функції. Якщо дані не повторюються, то коефіцієнт Спірмена дорівнює +1 або -1. Це відбувається коли кожна змінна є монотонною функцією від іншої змінної. Він підходить для безперервних та дискретних змінних. Коефіцієнт кореляції Спірмена визначаєть:

Для вибірки обсягом n множини Ai та Bi перетворюються в ряди ai та bi, та обчислюється за формулою:

$$\rho = \frac{\sum_i (a_i - \acute{a})(b_i - \acute{b})}{\sqrt{\sum_i}}$$

Однаковим значенням присвоюється ранг. Він дорівнює середньому їхніх позицій в порядку зростання величин.

Коефіцієнт кореляції рангу Кендала зазвичай називають тау-коефіцієнт. Він використовується у статистиці для вимірювання зв'язку між двома величинами. Тау тест - це непараметричний тест статистичних гіпотез залежності на основі тау-коефіцієнта. Також він є мірою рангової кореляції. Коефіцієнт рангу Кендала зазвичай використовується для статистичної оцінки в перевірці статистичних гіпотез для визначення чи можуть дві змінні розглядатись як статистично залежні. Тест є непараметричний, так як він не залежить від будь-яких припущень про розподіл X або Y або розподіл (x, y). При нульовій гіпотезі незалежності X і Y, вибірковий розподіл τ має очікуване значення - нуль. Формула має вигляд:

$$\tau = \frac{s_1 - s_2}{\frac{1}{2}n(n-1)}$$

Де s1 – кількість узгоджених пар, а s2 – кількість неузгоджених пар.

Перевіримо чи є зв’язок між заробітною платою та працевлаштуванням.

Q3 та Q7

Нульова гіпотеза: Коефіцієнт кореляції дорівнює нулю.

Альтернативна: Наявність зв’язку.

Критерій Пірсона

Для цього використовуємо в Пайтоні функцію

```
corr,p=stats.pearsonr(Q3,Q7)
print('Pearson_corr=%.3f, p-value=%.3f' % (corr, p))
if p < alpha:
    print("Прийняти гіпотезу про наявність зв'язку")
else:
    print("Відхилити гіпотезу про наявність зв'язку")
```

Результат:

```
Pearson_corr=0.602, p-value=0.000
Прийняти гіпотезу про наявність зв'язку
```

Критерій Спірмена

Для цього використовуємо в Пайтоні функцію

```
corr,p=stats.spearmanr(Q3,Q7)
print('Spearman_corr=%.3f, p-value=%.3f' % (corr, p))
if p < alpha:
    print("Прийняти гіпотезу про наявність зв'язку")
else:
    print("Відхилити гіпотезу про наявність зв'язку")
```

Результат:

```
Spearman_corr=0.588, p-value=0.000
Прийняти гіпотезу про наявність зв'язку
```

Критерій Кендала

Для цього використовуємо в Пайтоні функцію

```
coef, p = kendalltau(Q3, Q7)
```

Результат:

```
Коефіцієнт кореляції Кендалла: 0.289
```

Отже можемо сказати що є зв’язок між заробітною платою та працевлаштуванням.

Тепер перевіримо чи є зв’язок між віком та наявність дітей у сім’ї.

Q2 та Q5

Нульова гіпотеза: Коефіцієнт кореляції дорівнює нулю.

Альтернативна: Наявність зв’язку.

Критерій Пірсона

Для цього використовуємо в Пайтоні функцію

```
corr,p=stats.pearsonr(Q2,Q5)
print('Pearson_corr=%.3f, p-value=%.3f' % (corr, p))
if p < alpha:
    print("Принйняти гіпотезу про наявність зв'язку")
else:
    print("Відхилити гіпотезу про наявність зв'язку")
```

Результат:

```
Pearson_corr=0.565, p-value=0.000
Принйняти гіпотезу про наявність зв'язку
```

Критерій Спірмена

Для цього використовуємо в Пайтоні функцію

```
corr,p=stats.spearmanr(Q2,Q5)
print('Spearman_corr=%.3f, p-value=%.3f' % (corr, p))
if p < alpha:
    print("Принйняти гіпотезу про наявність зв'язку")
else:
    print("Відхилити гіпотезу про наявність зв'язку")
```

Результат:
```
Spearman_corr=0.580, p-value=0.000
Принйняти гіпотезу про наявність зв'язку
```

Критерій Кендала

Для цього використовуємо в Пайтоні функцію

```
coef, p = kendalltau(Q3, Q7)
```

Результат:
```
Коєфіцієнт кореляції Кендалла: 0.188
```

Висновки
Отже також можемо сказати про наявність зв’язку між віком та наявність дітей у сім’ї.

3.3 Регресійний аналіз

Регресійний аналіз — розділ математичної статистики що вивчаэ методы аналізу залежності однієї величини від іншої. Відрізняється від кореляційного аналізу тим, що не з'ясовує чи істотний зв’язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії.

Регресійний аналіз використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати залежна змінна, яка обчислюється на заданому наборі значень незалежних змінних. У самому простому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія. Більшість реальних моделей не вкладаються в рамки лінійної регресії.

До регресійної моделі відносяться такі параметри і змінні:

- Невідомі параметри
- Незалежні змінні
- Залежна змінна

У статистиці **лінійна регресія** — це метод моделювання залежності між скалярною змінною y та векторною (у загальному випадку) змінною X. У разі, якщо змінна X також є скаляром, регресію називають простою.

При використанні лінійної регресії взаємозв'язок між даними моделюється за допомогою лінійних функцій, а невідомі параметри моделі оцінюються за вхідними даними. Подібно до інших методів регресійного аналізу лінійна регресія повертає розподіл умовної імовірності y в залежності від X, а не розподіл спільної імовірності y та X, що стосується області мультиваріативного аналізу.

Нелінійна регресія − це вид регресійного аналізу, в якому експериментальні дані моделюються функцією, яка є нелінійної комбінацією параметрів моделі і залежить від однієї і більше незалежних змінних.

Отже, побудуємо однофакторну регрісію між двома кількісними змінними (віком і зарплатньою).

Q2 і Q15

Для цього використовуємо в Пайтон функцію

```
x = np.array([60, 44, 40,  57, 52 ,  54,  35,  26,  28,  29,  20,  37,  32,  19,  27, 33, 29, 48]).reshape((-1, 1))
y = np.array([21026, 29557,14450, 15156, 19800, 10428, 10898, 27296, 26576, 15138, 23593, 23440, 9191, 20849, 21625, 21878, 19970,26967])
print(x)
print(y)
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
print('intercept:', model.intercept_)
print('slope:', model.coef_)
```

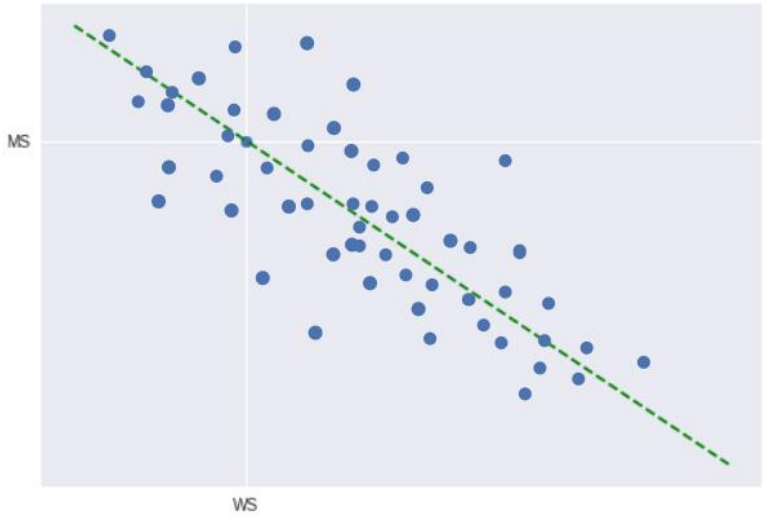
Результат:

```
[[60]
 [44]
 [40]
 [57]
 [52]
 [54]
 [35]
 [26]
 [28]
 [29]
 [20]
 [37]
 [32]
 [19]
 [27]
 [33]
 [29]
 [48]]
[21026 29557 14450 15156 19800 10428 10898 27296 26576 15138 23593 23440
  9191 20849 21625 21878 19970 26967]
coefficient of determination: 0.03327391494615062
intercept: 23174.92420638176
slope: [-88.52333689]
```

Будуємо графік в Пайтон за допомогою функції

```
y_1 = range()

plt.xlim()
plt.ylim()
plt.scatter(y,x)
plt.plot(y_1, [-88.5*y + 0.033 for y in y_1], linestyle='--', color='green')
plt.show()
```



Побудуємо ще одну однофакторну регресію між віком та відношенням до посилення карантину.

Q2 і Q16

Для цього використовуємо в Пайтон функцію

```
x = np.array([60, 44, 40,  57, 52 ,  54,  35,  26,  28,  29,  20,  37,  32,  19,  27, 33, 29, 48]).reshape((-1, 1))
y = np.array([5, 4, 2, 3, 1, 1, 2, 4, 2, 3, 4, 5, 2, 1, 3, 4, 1,5])
print(x)
print(y)
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
print('intercept:', model.intercept_)
print('slope:', model.coef_)
```

Результат:

```
[[60]
 [44]
 [40]
 [57]
 [52]
 [54]
 [35]
 [26]
 [28]
 [29]
 [20]
 [37]
 [32]
 [19]
 [27]
 [33]
 [29]
 [48]]
[5 4 2 3 1 1 2 4 2 3 4 5 2 1 3 4 1 5]
coefficient of determination: 0.0203795590984287
intercept: 2.2782380444590276

slope: [0.01640555]
```

Нам треба перевірити значущість коефіцієнтів регресії, коефіцієнтів детермінації та розподіл залишків. Перевірка показала що модель адекватна.

ВИСНОВКИ

Дана робота була присвячена проведенню соціологічного опитування на тему “ Відношення жителів до можливого посилення карантинних обмежень, пов'язаних з пандемією COVID – 19”, із подальшим статистичном аналізом для виявлення певних закономірностей.

Ререзультатами роботи стало:

1. Формулювання проблеми (робочих гіпотез) про відношення різних соціальних і вікових груп населення Лук’янівскього району до посилення карантину, дій уряду проти нього, обізнаність на тему протидій ковіду, т. д. .

2. Проведення анкетування (сумісно з блогерами мого району)

3. Описова статистика і частотний аналіз дав відповідь на те, які вікові, соціальні групи представлені в дослідження (див. Висновки до розділу 1) та яка частка респондентів вже перехворіла ковідом.

Також стало зрозуміло як відношення до дій уряду, так і до посиленню карантинних заходів. (див. Висновки до розділу 1)

4. За допомогою вивідної статистики було перевірено гіпотези про пропорцію, тобто для даного обсягу вибірки і даного рівня значущості чи відповідає вибірка частот пропорції для генеральної сукупності.

5. Було висловлено припущення, про значущу різницю між жінками та чоловіками у відношенні до посилення карантину та дій уряду. Перевірки гіпотез про однорідність показали, що за критеріями (див. Висновки до розділу 2).

6. Було досліджено залежності між (див. Висновки до розділу 3)

7. Було побудовано регресійну модель залежності між і ця модель може бути корисною для прогнозування.

Література

1. Бартіш М.Я., Дудзяний І.М. Дослідження операцій: Навч. Посібник. – Львів:ВЦ ЛНУ ім. І.Франка, 2007

2. Вентцель Е.С. Исследование операцій. Задачи, принципы, методология: Уч. Пособ. – М.Дроффа, 2004

3. Наконечний С.І., Савіна С.С. Математичне програмування: Навч. Посібник. – К.: КНЕУ, 2003

4. Моклячук М.П. Варіаційне числення. Екстремальні задачі. – К.: КНУ ім. Шевченка, 2003

5. Алексеев В.М., Галеев Э.М., Тихомиров В.М. Сборник задач по оптимизации, Наука, М., 1984