Контроль набутих знань
практика (тиждень 4)

Завдання

Для роботи використовується файл Titanic.csv.

Список обов'язкових кроків для виконання:

1. Створити код в SAS Studio з назвою у форматі - lastname_cw4.sas;
2. Створити бібліотеку(назва бібліотеки=прізвище інтерна), в якій будете зберігати лише фіналізовані датасети *(проміжні датасети зберігайте в директорії WORK).*
3. Побудувати частотну таблицю та гістограму розподілу відсотків значень для змінних Survived, Gender, Class *(див. Додаток 3: Таблиця 1-3 і рис.1-3).*
4. Побудувати двохфакторну частотну таблицю та гістограму розподілу відсотків для комбінацій значень (Gender та Survived), і (Class та Survived) *(див. Додаток 3: Таблиця 4-5 і рис.4-5).*
5. Ознайомтесь з теоретичними відомостями по обчисленню хі-квадрат. *(див. Додаток 2).*
6. Побудувати таблицю розподілу пасажирів в залежності від статі *(див. Додаток 3: Таблиця 6).*
7. Розрахувати хі-квадрат за допомогою дата кроку.
8. Створіть звіт у форматі .rtf відповідно до прикладу *(Додаток 3).*
9. Розрахувати хі-квадрат за допомогою процедурного кроку. Порівняйте отримані результати, які ви отримали (крок даних) із еталонним (процедурний крок), за допомогою процедури PROC COMPARE *(результати повинні зійтись).*
10. Реалізувати п.3 та 4 за допомогою макропрограми.
11. Реалізувати п.6 та 7 за допомогою макропрограми, яка буде розраховувати хі-квадрат на дата кроці для випадку MxM.

Загалом, можливі деякі моменти, що не досить формалізовані в постановці завдання, в цьому випадку раджу проявити ініціативу та креативність, яку опишіть у вигляді коментаря на початку вашої програми.

*Для тестування програми п.11 можете використовувати будь який датасет з бібліотеки SASHELP або датасет Тітанік (в такому разі можете в датасет вносити корективи. Наприклад для тестування випадку 3*3 (MxM) можете додати змінну AGE_GROUP (див. попередні практичні роботи) і порахувати хі-квадрат для AGE_GROUP і CLASS).*

Оцінювання:

- правильне виконання п.1-8 – 40%;
- правильне виконання п.1-10 – 50%;
- правильне виконання п.1-11 – 100%

Додаток 1. Теоретична інформація про дані для аналізу

VARIABLE DESCRIPTIONS

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Survival	Survival (0 = No; 1 = Yes)
name	Name
Sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
Home_dest	Home/Destination

SPECIAL NOTES

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)

If the Age is estimated, it is in the form xx.5

Fare is in Pre-1970 British Pounds ()

Conversion Factors: 1 = 12s = 240d and 1s = 20d

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic

Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations

Додаток 2. Теоретична інформація по аналіз даних

Нуль гіпотеза:

- Не існує зв'язку між статтю та фактом виживання при аварії.

- Ймовірність виживання при катастрофі Титаніка як для жінок, так і чоловіків була **однакова**.

Альтернативна гіпотеза:

- Існує зв'язку між статтю та фактом виживання при аварії.

- Ймовірність виживання при катастрофі Титаніка як для жінок, так і чоловіків **не була однаковою**.

Статистика хі-квадрат.

Зазвичай, для перевірки наявності зв’язку між двома категоріальними змінними використовується статистика хі-квадрат Пірсона, який обчислює різницю між частотами що спостерігаються та емпіричними. Якщо статистика хі-квадрат значима, то це вагомі докази того, що існує зв’язок між змінними.

Тест хі-квадрат і p −значення що йому відповідає

- призначений для виявлення наявності зв’язку між змінними,
- не вимірює силу зв’язку між змінними,
- залежить від розміру вибірки.

Відповідно до нульової гіпотези про відсутність зв'язку між змінною рядка та стовпця, "очікуваний" відсоток у будь-якій осередку R * C буде дорівнює відсоткам у рядку цієї комірки (R / T) у відсотках у стовпці комірки (C / T). Очікуваний показник тоді лише очікуваний відсоток від загального розміру вибірки. Очікуваний відлік = (R / T) * (C / T) * T = (R * C) / T.

Під нульовою гіпотезою щодо відсутності зв’язку між змінними Row та Column розуміють що очікувана частота в будь-якій комірці R*C, значення комірок матриці по рядку (R / T) та стовпчику (C / T) будуть рівні. Очікувана кількість – це лише очікуваний відсоток від загального розміру вибірки, що обчислюється за формулою:

$$Expected\ count = \left(\frac{R}{T}\right) \cdot \left(\frac{C}{T}\right) \cdot T = \frac{(R \cdot C)}{T}$$

p-значення для тесту хі-квадрат лише вказує, наскільки ви можете бути впевненими, щодо нульової гіпотези відносно відсутності зв’язку. Цей тест не надає інформації щодо величини сили зв’язку. Значення статистики хі-квадрат також не говорить про це. Якщо навіть ви збільшуєте розмір вибірки, наприклад дублюючи кожне спостереження, відбувається подвоєння значення статистики хі-квадрат, навіть якщо сила зв’язку не змінюється.

Приклад обчислення статистики хі-квадрат.

Нижче наведений приклад обчислення статистики хі-квадрат (таблиці 2х2).

Приклад таблиць для обчислення хі-квадрат для набору даних Titanic (комбінації змінних Gender та Survived) наведенні в Додатку 3 (Таблиця 7-9).

По-перше, обчислюється частотна матриця значень. В загальному вигляді вона може бути представлена як показано в таблиці нижче.

Частотна матриця, загальний вигляд

	Survived=0	**Survived =1**	**Total**
Female	A	B	A+B
Male	C	D	C+D
Total	A+C	B+D	A+B+C+D

По-друге, обчислюється частотна матриця частот що очікуються. В таблиці нижче в загальному вигляді наведені розрахунки, для кожного елемента матриці.

Матриця очікуваних частот, загальний вигляд

	Survived=0	**Survived =1**
Female	(A+B)*(A+C) / (A+B+C+D)	(A+B)*(B+D)/ (A+B+C+D)
Male	(C+D)*(A+C)/ (A+B+C+D)	(C+D)*(B+D)/ (A+B+C+D)

На третьому кроці, обчислюється значення хі-квадрат за формулою:

$$\chi^2 = \sum_{i=}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

де

i −й номер рядка (від 1 до r),

j −й номер стовпця (від 1 до c),

O_{ij} − фактична кількість спостережень для (i, j) елемента матриці,

E_{ij} − очікувана кількість спостережень для для (i, j) елемента матриці.

Матриця індивідуальних значень статистики хі-квадрат

	Survived=0	**Survived =1**	**Total**
Female	XXXX	XXXX	
Male	XXXX	XXXX	
Total			**XXXX**

Додаток 3. Приклад звіту

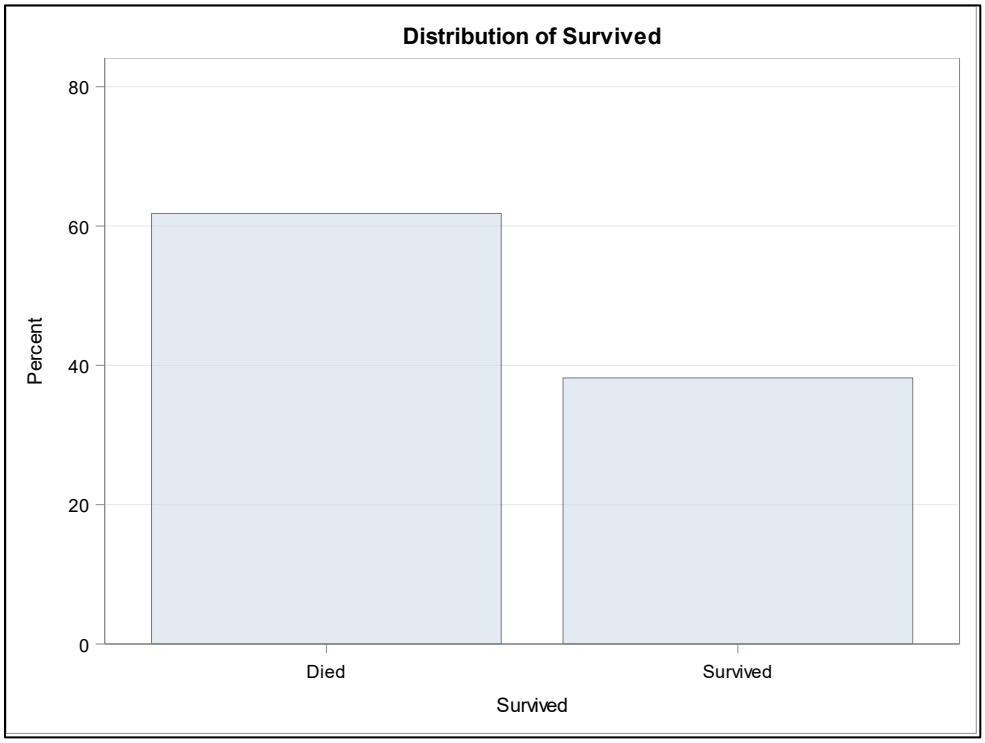
Прізвище ім’я інтерна

Заголовок звіту

Таблиця 1 – Частотна таблиця для Survived

Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Died	809	61.8	809	61.8
Survived	500	38.2	1309	100

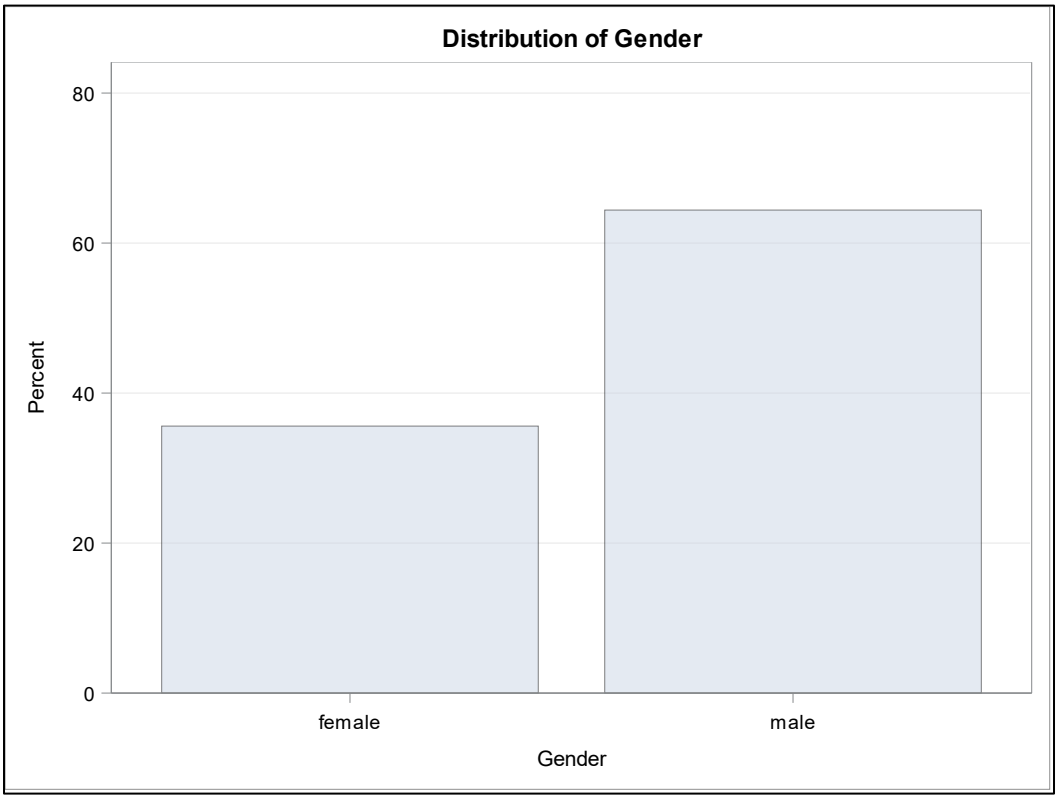
Рис. 1 – Гістограма розподілу відсотків значень змінної Survived



Таблиця 2 – Частотна таблиця для Gender

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	466	35.6	466	35.6
male	843	64.4	1309	100

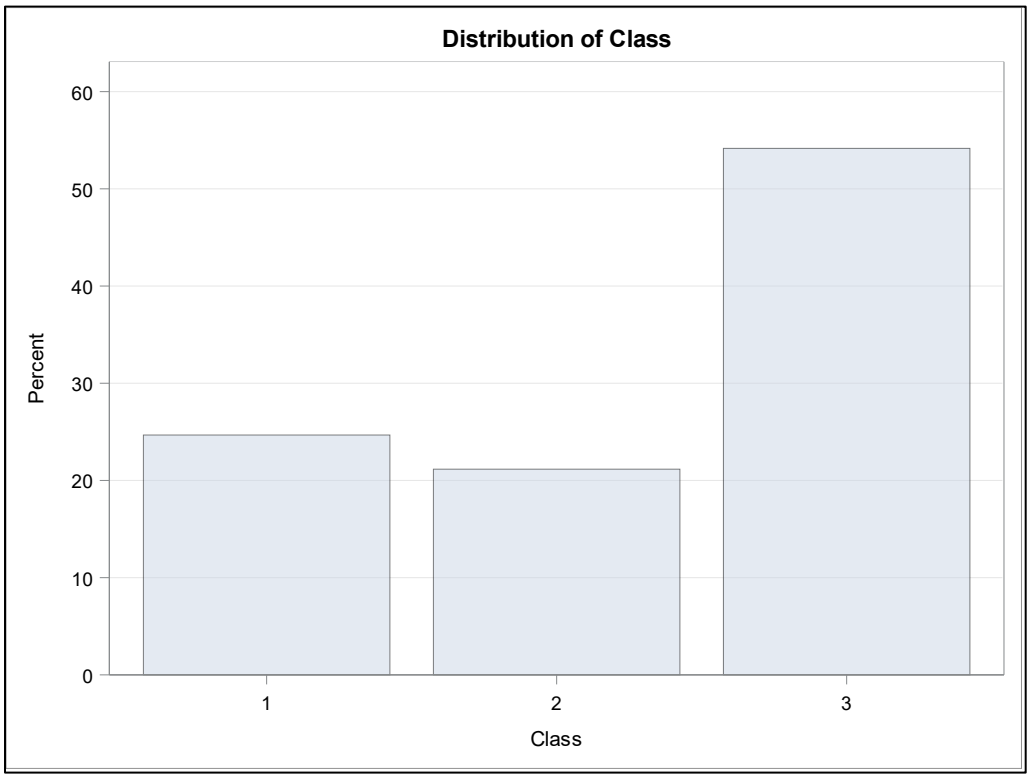
Рис. 2 – Гістограма розподілу відсотків значень змінної Gender



Таблиця 3 – Частотна таблиця для Class

Class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	323	24.68	323	24.68
2	277	21.16	600	45.84
3	709	54.16	1309	100

Рис. 3 – Гістограма розподілу відсотків значень змінної Class

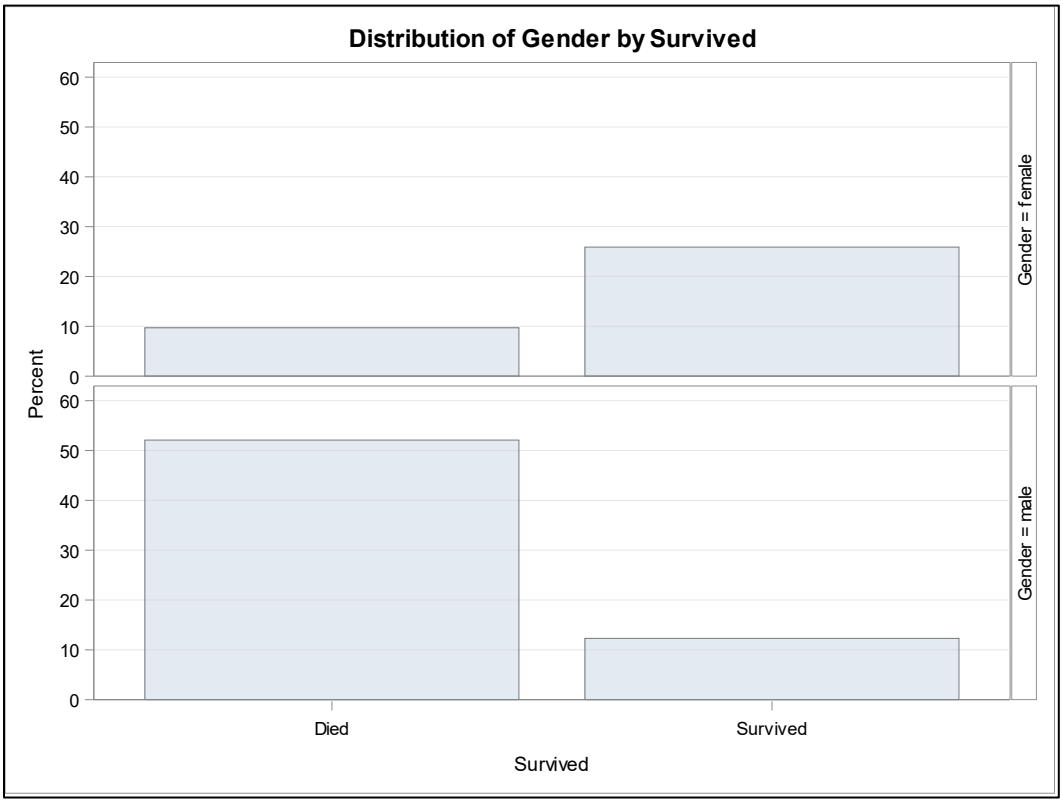


(Дивлячись на результати роботи програми, можна вирішити, що немає ніяких незвичайних значень в даних. Побудуємо двохфакторні частотні таблиці. Можна запровадити апріорну ідею щодо існування зв'язку між відгуком та регресорами. В якості відгука будемо розглядати Survived, а в якості регресорів Gender та Class).

Таблиця 4 – Двохфакторна частотна таблиця для Gender та Survived

Table of Gender by Survived			
Gender	**Survived**		
Frequency **Percent** **Row Pct** **Col Pct**	**Died**	**Survived**	**Total**
female	127 9.70 27.25 15.70	339 25.90 72.75 67.80	466 35.60
male	682 52.10 80.90 84.30	161 12.30 19.10 32.20	843 64.40
Total	809 61.80	500 38.20	1309 100.00

Рис. 4 – Гістограма розподілу відсотків всіх комбінацій значень змінних Gender та Survived

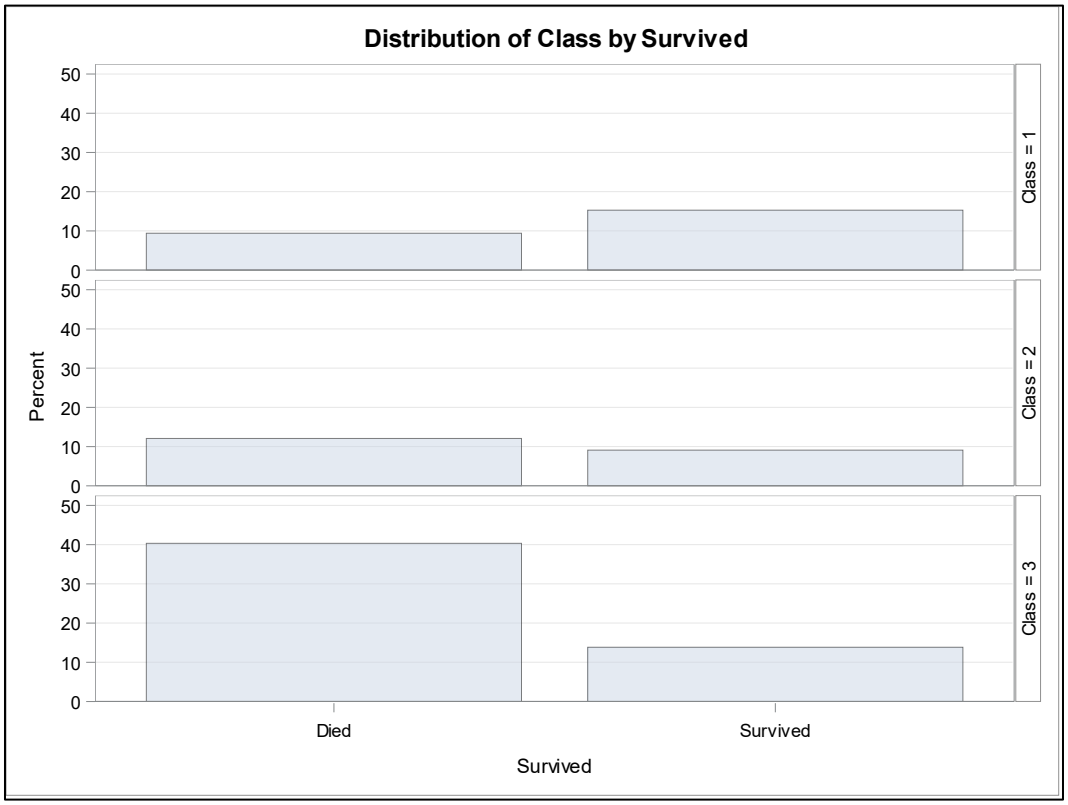


(Аналіз гістограм розподілу процентів всіх комбінацій значень змінних Gender та Survived, однозначно вказує на наявність зв'язку між змінними)

Таблиця 5 – Двохфакторна частотна таблиця для Class та Survived

Table of Class by Survived			
Class	**Survived**		
Frequency **Percent** **Row Pct** **Col Pct**	**Died**	**Survived**	**Total**
1	123 9.40 38.08 15.20	200 15.28 61.92 40.00	323 24.68
2	158 12.07 57.04 19.53	119 9.09 42.96 23.80	277 21.16
3	528 40.34 74.47 65.27	181 13.83 25.53 36.20	709 54.16
Total	809 61.80	500 38.20	1309 100.00

Рис. 5 – Гістограма розподілу відсотків всіх комбінацій значень змінних Class та Survived



(Як можна побачити з гістограм розподілу процентів всіх комбінацій значень змінних Class та Survived, також є зв'язок між змінними. Набагато більше шансів вижити у пасажирів вищих класів)

Таблиця 6 – Розподіл пасажирів в залежності від статі (Gender) та відгука (Survived)

Стать	**Відгук**		
Row Pct	**Помер**	**Живий**	**Загалом**
Жіноча	27,75%	72,25%	N=466
Чоловіча	80,90%	19,10%	N =843
Загалом	N=809	N=500	N=1309

(За результатами значень наведених в таблиці вище, можна зробити висновок про існування зв'язку між статтю та виживанням, оскільки ймовірності рядків різні в кожному стовпчику. Для перевірки наявності зв'язку виконується оцінювання різниці між ймовірністю виживання жінок (72,25%) та виживанням чоловіків (19,10%), як видно різниця значно більша аніж наявність випадковості)

Таблиця 7 - Частотна матриця, емпіричні значення

(для комбінації змінних Gender та Survived)

	Survived=0	**Survived =1**	**Total**
Female	127	339	466
Male	682	161	843
Total	809	500	1309

Таблиця 8 - Матриця очікуваних частот

(для комбінації змінних Gender та Survived)

	Survived=0	**Survived =1**
Female	288,0015279	177,9984721
Male	520,9984721	322,0015279

Таблиця 9 - Матриця індивідуальних значень статистики хі-квадрат

(для комбінації змінних Gender та Survived)

	Survived=0	**Survived =1**	**Total**
Female	90,0047	145,6276	
Male	49,75349	80,50115	
Total			**365,8869**