Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут
імені Игоря Сікорського»

Л. Ф. Панченко

Математичні та статистичні методи аналізу соціологічної інформації

*Рекомендовано Методичною радою КПІ ім. І. Сікорського
як навчальний посібник
для здобувачів ступеня бакалавра
для спеціальності 054 «Соціологія»*

Київ
КПІ ім. І. Сікорського
2018

УДК 303.7:316]:004] (075.8)
П16

Р е ц е н з е н т и :

Кутуєв П.В.	–	д-р соціологічних наук, професор, КПІ ім. І. Сікорського
Семеріков С.О..	–	доктор пед.наук, професор, Криворізький державний педагогічний університет
Прошкін В.В.	–	доктор пед. наук, професор, Київський університет імені Бориса Грінченка

Панченко Л. Ф.
П16 Математичні та статистичні методи аналізу соціологічної інформації : Практикум / Л. Ф. Панченко; КПІ ім. І.Сікорського, 2018. – 289 с.

Присвячено теоретичним і практичним питанням комп'ютерного аналізу даних з використанням електронних таблиць MS Excel, статистичного пакету SPSS, вільно розповсюджуваного середовища R. Розглянуто описову статистику, основи кореляційного і регресійного аналізу, теорії імовірності та перевірки гіпотез.

Для студентів спеціальностей «Соціологія», «Соціальна робота», а також всіх інших спеціальностей, які пов'язані з аналізом даних на комп'ютері

УДК 303.7:316]:004] (075.8)

Гриф надано Методичною радою КПІ ім.І.Сікорського (протокол №4 від 20.12.2018 за поданням Вченої ради Факультета соціології та права (протокол №11 від 25.06.2018)

© Панченко Л. Ф., 2018
© КПІ ім. І.Сікорського ”, ФСП, 2018

ЗМІСТ

ВСТУП..... 6
РОЗДІЛ 1. ЧАСТОТНИЙ РОЗПОДІЛ ДАНИХ. МІРИ ЦЕНТРАЛЬНОЇ ТЕНДЕНЦІЇ І МІРИ МІНЛИВОСТІ. ОПИСОВА СТАТИСТИКА В MS EXCEL, SPSS, R 7
Практична робота 1.1. Побудова частотного розподілу даних і побудова гістограми за допомогою MS Excel 12
Практична робота 1.2. Обчислення середнього і стандартного відхилення. Графічна Інтерпретація стандартного відхилення..... 19
Практична робота 1.3. Обчислення мір центральної тенденції і мір мінливості в R..... 22
Практична робота 1.4. Побудова і розрахунок параметрів розподілу для незгрупованих даних за допомогою пакета аналізу MS Excel 28
Практична робота 1.5. Побудова частотного розподілу для незгрупованих даних в R..... 35
Практична робота 1.6. Підготовка даних анкетного опитування до обробки за допомогою SPSS 39
Практична робота 1.7. Побудова частотного розподілу і обчислення статистичних показників за допомогою SPSS 45
Практична робота 1.8. Побудова і редагування графіків в SPSS 49
Практична робота 1.9. Обробка даних анкетного опитування в середовищі R .....
Практична робота 1.10. Обчислення середнього згрупованих даних в Excel та SPSS 54
Практична робота 1.11. Обчислення моди та медіани згрупованих даних у Excel та SPSS 60
71

РОЗДІЛ 2. ВИЯВЛЕННЯ ВІДМІННОСТЕЙ В РОЗПОДІЛАХ ЗА ДОПОМОГОЮ КРИТЕРІЮ ПІРСОНА χ^2 82
Практична робота 2.1. Використання критерію Пірсона χ^2 для перевірки узгодженості розподілів за допомогою MS Excel и SPSS 86
Практична робота 2. 2. Критерій Пірсона χ^2 в R 95

РОЗДІЛ 3. КОРЕЛЯЦІЯ..... 99
Практична робота 3.1. Обчислення коефіцієнта кореляції Пірсона-Браве для метричних шкал..... 102
Практична робота 3.2. Кореляція за Пірсоном за допомогою R 102
Практична робота 3.3. Обчислення коефіцієнта рангової кореляції Спірмена..... 109
Практична робота 3.4. Обчислення коефіцієнта рангової кореляції в R 115 124
Практична робота 3.5. Обчислення коефіцієнта рангової кореляції Кенделла засобами Excel 127
Практична робота 3.6. Обчислення коефіцієнтів асоціації та контингенції засобами Excel..... 131
Практична робота 3.7. Обчислення кореляційного відношення та бісеріального коефіцієнта кореляції засобами Excel..... 138

РОЗДІЛ 4. ЛІНІЙНА РЕГРЕСІЯ..... 143
Практична робота 4.1. Обчислення коефіцієнта кореляції та прогноз за допомогою лінійної регресії в MS Excel..... 146
Практична робота 4.2 Обчислення коефіцієнта кореляції та прогноз за допомогою лінійної регресії в R 154

Практична робота 4.3. Обчислення коефіцієнта кореляції та прогноз за допомогою лінійної регресії в SPSS..... …… 156
Завдання для обчислення регресії…….. 162
Практична робота 4.4. Побудова моделі множинної регресії засобами Excel та SPSS..... 170

РОЗДІЛ 5. ОСНОВИ ТЕОРІЇ ІМОВІРНОСТІ. ДИСКРЕТНІ ТА НЕПЕРЕРВНІ РОЗПОДІЛИ. ДОВІРЧІ ІНТЕРВАЛИ…… 181
Практична 5.1. Обчислення імовірностей засобами MS Excel 183
Практична 5.2. Використання теореми Баєса 189
Практична 5.3. Побудова графіку функції щільності імовірності для нормального розподілу та одиничного нормального розподілу 192
Практична 5.4. Розрахунок імовірностей нормального розподілу …….. 199
Практична 5.5. Побудова довірчих інтервалів 206
Практична 5.6. Побудова та порівняння графіків нормальної імовірності декількох розподілів 219

РОЗДІЛ 6. ПЕРЕВІРКА ГІПОТЕЗ. 236
Практична робота 6.1. Перевірка гіпотез про значущість відмінності середніх в MS Excel 248
Практична робота 6.2. Перевірка гіпотез про значущість відмінності середніх в R 252
Практична робота 6.3. Перевірка гіпотез про значущість відмінності дисперсій в MS Excel..... 257
Практична робота 6.4. Перевірка гіпотез про значущість відмінності дисперсій в R 260
Практична робота 6.5. Перевірка гіпотез про значущість відмінності середніх двох незалежних вибірок в SPSS 263
Практична робота 6.6. Перевірка гіпотез про значущість відмінності середніх двох залежних вибірок в SPSS..... 268

Таблиці критичних значень ……..... 277
Література..... 282

ВСТУП

Навчальний посібник побудовано відповідно до програми курсу «Математичні та статистичні методи аналізу соціологічної інформації» для студентів спеціальності «Соціологія» та «Соціальна робота» та рекомендаціями щодо викладання інформатики в університетах.

Видання має таку структуру: короткий виклад найбільш важливих теоретичних відомостей по темі, практичні роботи, список рекомендованої літератури. Кожна практична робота містить детальні інструкції щодо вирішення типової задачі, завдання для самостійного рішення, вимоги до звіту і контрольні питання.

Для вирішення завдань використовують електронні таблиці Microsoft Excel, статистичний пакет SPSS, середовище R. R – ефективний інструмент розроблення нових методів інтерактивного аналізу даних. Він вільно поширюється, швидко розвивається і розширюється великою колекцією пакетів, що реалізують новітні методи аналізу даних.

Виклад матеріалу супроводжується ілюстраціями – схемами, таблицями, зображенням діалогових вікон – для полегшення розуміння та засвоєння об'єктивно нелегких математичних методів аналізу даних. У виданні використано авторські матеріали з навчальних посібників «Комп’ютерний аналіз даних» та «Практикум з аналізу даних».

Навчальне видання допоможе студентам і магістрантам, які будуть використовувати комп'ютерний аналіз даних у своїй професійній діяльності, освоїти основні засоби такого аналізу з використанням Microsoft Excel, SPSS, R: описову статистику, основи кореляційного аналізу, лінійної регресії, перевірку статистичних гіпотез. Посібник може бути корисним не тільки студентам і магістрантам спеціальності «Соціологія», а й усім тим, хто прагне оволодіти математичними та статистичними методами аналізу даних.

РОЗДІЛ 1
ЧАСТОТНИЙ РОЗПОДІЛ ДАНИХ. МІРИ ЦЕНТРАЛЬНОЇ ТЕНДЕНЦІЇ І МІРИ МІНЛИВОСТІ. ОПИСОВА СТАТИСТИКА В MS EXCEL, SPSS, R

Ознаки та змінні – це вимірювані індикатори тих об'єктів, процесів і явищ, які вивчаються в ході соціологічних, маркетингових, політологічних, психолого-педагогічних досліджень, досліджень в медицині, біології, сільського господарства та ін. Значення ознаки, отримані в результаті здійснених в процесі конкретного дослідження вимірювань, називаються спостереженнями, спостережуваними значеннями, варіантами.

Абсолютна частота – це число об'єктів, що володіють даним значенням ознаки.

Відносна частота – це доля об'єктів, що володіють даним значенням ознаки, у загальному обсязі вибірки. Відносна частота отримується шляхом ділення абсолютної частоти на загальне число вимірів. Відносні частоти вимірюються у долях одиниці. Сума відносних частот для даного ряду вимірів завжди рівна одиниці.

Процентна частота – це відсоток об'єктів, що володіють даним значенням ознаки, в загальному обсязі вибірки; це відносна частота, виражена в процентах. Сума процентних частот для даного ряду вимірювань завжди рівна ста процентам.

Накопичена (кумулятивна) частота – для шкал, вище рангових, – це сума частот значень, які не перевищують дане значення.

Табулювання даних – це зведення їх у таблицю. Основні вимоги до таблиці: наявність чіткого і ясного заголовка; наявність чітких і ясних підписів колонок (стовбців) та рядків.

Частотний розподіл – упорядкований підрахунок кількості кожного значення змінної. Частотний розподіл показує, скільки раз кожне значення змінної зустрічається в сукупності спостережень.

Таблиця частотного розподілу – один з найпростіших способів представлення зібраних даних. Вона складається, принаймні, з двох стовбців: лівий містить значення, які можуть приймати змінні, а правий – абсолютна частота (число разів, які кожне значення зустрічається). Додаткові стовпчики, що відображають відносний і (або) процентний частотний розподіл.

Графік частотного розподілу – частіше всього це стовпчаста діаграма (гістограма). Дані на гістограмі – представляються у вигляді паралельних прямокутників однакової ширини. Кожний стовбець показує один клас якісний даних (один варіант). Висота стовпчику пропорціональна величині або частоті вимірюваного параметра.

Найбільш уживані **міри центральної тенденції :** мода, медіана, середнє.

Мода – значення, що найчастіше зустрічається в ряді вимірів.

Медіана – середнє значення в упорядкованому ряді вимірів.

Середнє (середнє арифметичне значення) – сума всіх значень в ряді вимірів, розділена на їх загальну кількість.

Міри мінливості характеризують розкид даних біля середніх значень. До уживаних мір мінливості належать варіаційний розмах, дисперсія, стандартне відхилення.

Розмах – різниця між найбільшим і найменшим значеннями в ряду вимірів.

Дисперсія – відношення суми квадратів відхилень від середнього арифметичного до величини (n – 1), де n – загальне число вимірів.

Стандартне відхилення – позитивне значення квадратного кореня з дисперсії; вимірює «середній» розмах значень змінної відносно її середнього арифметичного в тих самих одиницях вимірювання, що і сама змінна.

Z-оцінки дозволяють представити ряд даних в стандартизованому вигляді; середнє такого нового ряду буде дорівнювати нулю, стандартне відхилення – одиниці; при цьому кожне значення в ряду замінюється на його Z-оцінку за формулою:

$$Zi = \frac{Xi - \overline{X}}{S} \qquad (1.1)$$

$\overline{X}$ – середнє

X_i – значення ознаки

S – стандартне відхилення

Асиметрія та **ексцес** характеризують форму розподілу. Асиметрія показує чи є він симетричним, або ні; ексцес – чи є він гостровершинним, плоско вершинним або середньовершинним (нормальним).

Деякі дані представлено вже в концентрованому або згрупованому за певними інтервалами ознаки вигляді. Для обчислення **середнього згрупованих даних** використовується формула

$$\overline{X} = \frac{\sum_{i=1}^{k} \overline{Xi} \circ Ni}{N} \qquad (1.2)$$

де

$\overline{X}$ – середнє згрупованих даних

X_i – середнє в і-інтервалі

N_i – частота в і-інтервалі

N – загальна кількість об’єктів

99 процентилів (P1, ..., P99) ділять групу спостережень на на 100 рівних частин **9 децилів** (D1, ..., D9) ділять на 10 рівних частин. **4 квінтиля** (K1, ... K4) на 5 частин (квінта - п'ять).**3 квартиля** (Q1, Q2, Q3) ділять групу спостережень на чотири рівні частини (кварти).

Для обчислення медіани та децилів згрупованих даних можна скористатися формулою для обчислення **процентиля.** Рк (К-й процентиль) – це точка на числовій осі, нижче якої лежить К% спостережень.

Щоб **обчислити процентиль** необхідно наступне:

1. Побудувати криву накопичених частот.
2. По ній визначити інтервал, в який потрапляє процентиль.
3. Визначити нижню межу цього інтервалу, ширину інтервалу, накопичені частоти на нижній і верхній межі інтервалу.
4. Скористатися формулою

$$P_k = X_l + I_l \circ \frac{k - f_{l-1}}{f_l - f_{l-1}} \quad (1.3)\text{, де}$$

P_k – к-значення процентиля

X_l – нижня межа інтервалу, куди потрапляє процентиль,

I_l – ширина інтервалу,

k к значення процентилю

f_{l-1} – накопичені частоти на нижній межі інтервалу,

f_l – накопичені частоти на верхній межі інтервалу

Для медіани k=50, для першого децилю – 10, для останнього децилю – 90.

Для розрахунків зазначених вище показників описової статистики можна користуватися вбудованими функціями з категорії «статистичні» та пакетом аналізу Excel (інструменти

«Описова статистика», «Гістограма») (рис.1.1); меню ***Analyze, Descriptive statistics, Frequencies*** в SPSS (рис. 1.2).

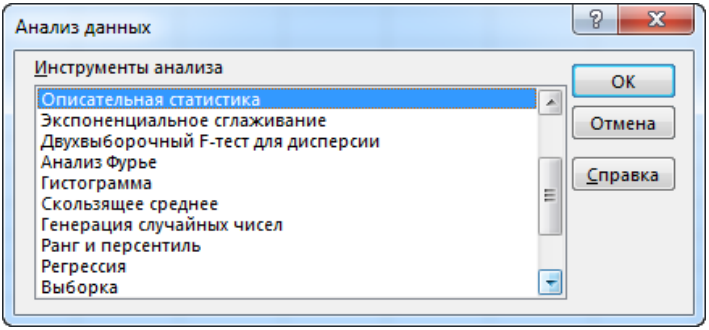


Рис. 1.1. Інструмент «Аналіз даних» електронних таблиць MS Excel

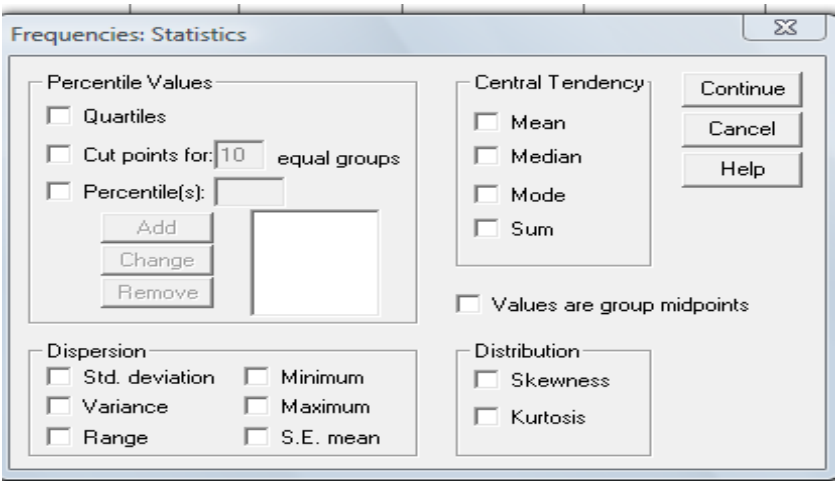


Рис.1.2. Описова статистика в SPSS. Діалогове вікно «Frequencies: Statistics».

Практична робота 1.1

Побудова частотного розподілу даних та побудова гістограми за допомогою MS Excel

Мета роботи: навчитися використовувати можливості електронних таблиць MS Excel для розрахунку відносних, процентних, накопичених частот, будувати графіки частотного розподілу.

Постановка задачі

В таблиці 1.1.1 представлено дані про розподіл користувачів мережі Facebook за континентами (2010 р.) .

Таблиця 1.1.1

Вихідні дані

Континент	**Число користувачів Facebook (осіб)**
1. Північна Америка	186126740
2. Європа	169718660
3. Азія	117151400
4. Південна Америка	52870200
5. Африка	19649500
6. Австралія та Океанія	11656460

http://statistic.su/blog/facebook_statistics/2010-11-08-99

Обчислити за допомогою MS Excel відносні, процентні і накопичені частоти і, використовуючи, графічні можливості електронних таблиць, представити графічно частотний розподіл користувачів мережі Facebook за континентами.

Хід роботи

1. Відкрити Microsoft Excel.
2. Увести дані дані відповідно до табл. 1.1.2.

Таблиця 1.1.2

Електронна таблиця з вихідними даними

	A	B	C	D	E
1	**Континент**	**Число користувачів Facebook (осіб)**	**Відносна частота**	**Відсоткова частота**	**Накопи-чена частота у %**
2	1. Північна Америка	186126740			
3	2. Європа	169718660			
4	3. Азія	117151400			
5	4. Південна Америка	52870200			
6	5. Африка	19649500			
7	6. Австралія та Океанія	11656460			
8	Разом				

3. У комірці B8 підрахувати загальну кількість користувачів.

4. У клітинах C2: C7, D2: D7, E2: E7 записати формули для розрахунку відносних, процентних та накопичених частот (табл. 1.1.3).

Таблиця 1.1.3

Електронна таблиця з формулами

	A	B	C	D	E
1	**Континент**	**Число користувачів Facebook (осіб)**	**Відносна частота**	**Відсоткова частота**	**Накопи-чена частота у %**
2	1. Північна Америка	186126740	=B2/$B$8	=C2*100	=D2
3	2. Європа	169718660	=B3/$B$8	=C3*100	=E2+D3
4	3. Азія	117151400	=B4/$B$8	=C4*100	=E3+D4
5	4. Південна Америка	52870200	=B5/$B$8	=C5*100	=E4+D5
6	5. Африка	19649500	=B6/$B$8	=C6*100	=E5+D6
7	6. Австралія та Океанія	11656460	=B7/$B$8	=C7*100	=E6+D7
8	Разом	=СУММ(B2:B7)	=СУММ(C2:C7)	=СУММ(D2:D7)	

5. В комірках C8 і D8 підрахувати за допомогою формул суми відповідних частот. У результаті отримаємо таблицю 1.1.4.

Таблиця 1.1.4

Електронна таблиця з результатами розрахунків

	A	B	C	D	E
1	**Континент**	**Число користувачів в Facebook (осіб)**	**Відносна частота**	**Відсоткова частота**	**Накопичена частота у %**
2	1. Північна Америка	186126740	0,33	33,4	33,4
3	2. Європа	169718660	0,30	30,5	63,9
4	3. Азія	117151400	0,21	21,0	84,9
5	4. Південна Америка	52870200	0,09	9,5	94,4
6	5. Африка	19649500	0,04	3,5	97,9
7	6. Австралія та Океанія	11656460	0,02	2,1	100,0
8	Разом	557172960	1	100	

6. Побудувати гістограму (стовпчикову діаграму). Для цього, утримуючи клавішу «Ctrl», виділити два несуміжних блоки даних A1:A7 і D1:D7 і обрати Вставка, Гістограма. В меню Макет додати до побудованої діаграми заголовок, підписи горизонтальної і вертикальної осей, підписи даних. Повинна вийти гістограма, представлена на рис.1.1.1.

Рис. 1.1.1. Розподіл користувачів за континентами

7. Побудувати також полігон і криву накопичених частот (рис. 1.1.2)

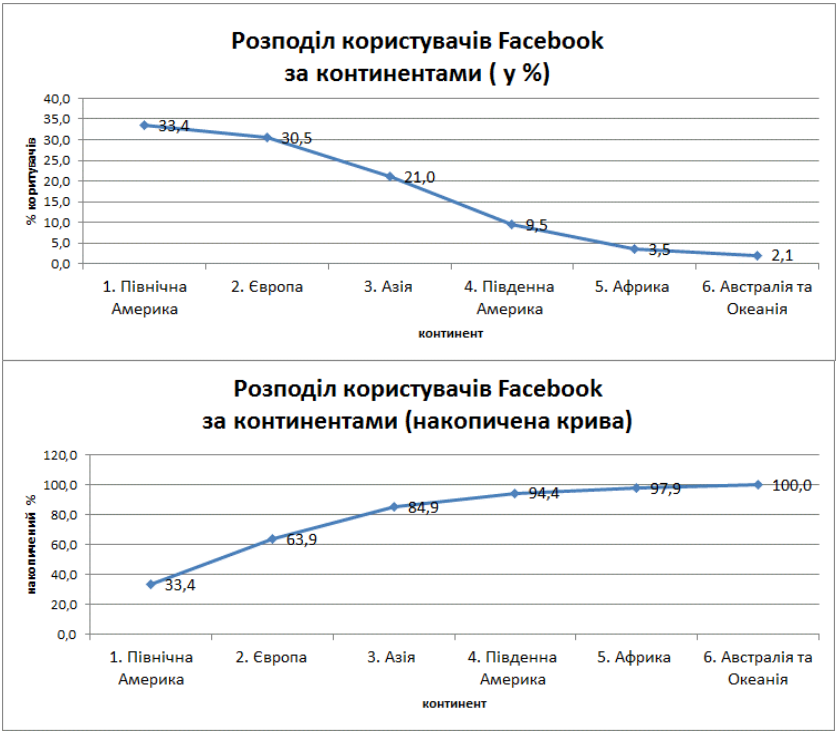


Рис. 1.1.2. Полігон та крива накопичених частот

Завдання для самостійної роботи

1. У таб. 1.1.5 представлені дані про вік й кількість українських користувачів Facebook (https://www.socialbakers.com/statistics/facebook/). Розрахуйте процентну і накопичену частоту і побудуйте гістограму, полігон і криву накопичених частот для цих даних.

Таблиця 1.1.5

Вік користувачів Facebook в Україні

Вік	Число користувачів (в тис.осіб)	% користувачів	Накопичена частота, %
13–17	159,60		
18–24	617,02		
25–34	836,92		
35–44	400,90		
45–54	188,52		
55–100	113,92		

2. Однією з перших систем з використанням природної мови була система LUNAR, яка давала відповіді на питання геологів про каміння, привезене з Місяця в ході польоту космічного апарату Апполон-11. На другій щорічній місячної наукової конференції в 1971 році системі LUNAR пред'явили 111 запитів на природній мові. Інформація про них наведена в таблиці 1.1.6. Розрахуйте відносні, відсоткові частоти і представте результати у вигляді графіку. Прокоментуйте результати.

Таблиця 1.1.6

Обробка запитів природною мовою в системі LUNAR

Запити	Абсолютна частота	Відносна частота	Відсоткова частота
Неможливо розібрати та інтерпретувати	11		
Запит, що виявився некоректним внаслідок помилок	13		
Оброблені задовільно	87		

Бондарев В.Н. Искусственный интеллект:Учеб. пособие для вузов/ В.Н. Бондарев, Ф.Г. Аде. - Севастополь: Изд-во СевНТУ, 2002.

3. П'ятнадцять професійних програмістів надали колегам для рецензування свою кращу програму. Кожен отримав по чотири програми від колег та оцінював їх за шкалою від 1 балу до 7. Результати наведені в таблиці 1.1.7 [39].

Таблиця 1.1.7

Результати peer-оцінювання

Різниця між найвищою і найнижчою з чотирьох оцінок програми, усереднена за 15 програмами	Число випадків	% випадків	Накопичений %
0 балів	2		
1 бал	7		
2 бали	17		
3 бали	15		
4 бали	9		
5 балів	7		
6 балів	3		

Розрахуйте відносні, відсоткові частоти і представте результати у вигляді графіка. Прокоментуйте результати.

4. Побудуйте частотний розподіл та представте графічно дані опитування вчителів України (таб.1.1.8), проведенного Українською асоціацією дослідників освіти у 2018 р. за методологією TALIS.

Таблиця 1.1.8.

Регіони, дн проводилося опитування

Регіон	Число респондентів-вчителівв
Схід	899
Захід	1321
м. Київ	211
Північ	389
Центр	447
Південь	333
Разом	3600

All-Ukrainian Monitoring Survey of Teaching and Learning (by TALIS Methodology), 2018.

Вимоги до звіту:

Звіт повинен містити:

– відповіді на контрольні питання;

– файли з результатами розрахунків.

Контрольні питання

1. Що таке абсолютна частота?
2. Що таке відносна частота? Як вона обчислюється?
3. Що таке процентна частота? Як вона обчислюється?
4. Що таке накопичена частота? Як вона обчислюється?
5. Що показує частотний розподіл ознаки?
6. Охарактеризуйте гістограму, полігон, криву накопичених частот. Як вони будуються та як редагуються?
7. Як виділити два несуміжних блоки в електронній таблиці MS Excel?
8. Як побудувати діаграму в MS Excel.
9. Як додати необхідні підписи до діаграми?
10. Дайте змістовну інтерпретацію отриманим результатам.

Практична робота 1.2
Обчислення середнього та стандартного відхилення. Графічна інтерпретація стандартного відхилення

Мета роботи: навчитися використовувати можливості електронних таблиць MS Excel для розрахунку мір центральної тенденції та мір мінливості, графічної інтерпретації стандартного відхилення.

Постановка задачі

Двадцяти респондентам було запропоновано дати оцінку за 10-ти бальною шкалою трьом сортам морозива. У таблиці 1.2.1 наведено оцінки, які дали респонденти**.**

Таблиця 1.2.1

Оцінка респондентами трьох сортів морозива

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Оцінка респондентами трьох сортів морозива																				
2		номер респондента																			
3	Сорт морозива	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
4	Пломбір	8	10	9	8	8	9	10	10	9	10	9	10	9	10	10	9	8	10	9	10
5	Слівочне	6	7	10	10	9	7	10	10	9	7	10	9	8	10	6	5	10	10	9	10
6	Фруктове	6	8	10	3	7	4	10	9	4	5	7	4	7	10	9	10	6	5	7	10

Визначити середню оцінку респондентами кожного виду морозива, а також стандартне відхилення для кожного виду. Побудувати графіки, що ілюструють розкид оцінок для кожного сорту морозива.

Хід роботи

1. Відкрити Microsoft Excel.
2. Увести дані відповідно до табл. 1.2.1 (щоб зробити маленькі клітинки, потрібно виділити відповідні стовпці, а потім один з стовпців зменшити).
3. У клітинці V3 написати слово «середнє», W3 – «станд. відхилення», в клітинах V4:V6 написати формули для розрахунку середнього (функція СРЗНАЧ), в клітинах W4: W6 формули для розрахунку стандартного відхилення (функція

СТАНДОТКЛОН). На рис. 1.2.1 представлено фрагмент таблиці з формулами, а на рис. 1.2.2 – з результатами.

	V	W
1		
2		
3	середнє	ст.відх.
4	=СРЗНАЧ(B4:U4)	=СТАНДОТКЛОН(B4:U4)
5	=СРЗНАЧ(B5:U5)	=СТАНДОТКЛОН(B5:U5)
6	=СРЗНАЧ(B6:U6)	=СТАНДОТКЛОН(B6:U6)

Рис. 1.2.1. Фрагмент таблиці з формулами

	V	W
1		
2		
3	середнє	ст.відх.
4	9,25	0,79
5	8,6	1,67
6	7,05	2,37

Рис. 1.2.2. Фрагмент таблиці з результатами

4. Проаналізувати отримані дані.
5. Побудувати точкову діаграму для кожної групи даних та порівняти їх.

Для цього виділити дані першої групи B4:U4, обрати *Вставка, Діаграма, Точкова*. Повторити ті ж дії для даних другої і третьої групи.

Повинні вийти три діаграми, подібні наведеним на рис. 1.2.3. Необхідно розташувати їх поруч на тому ж аркуші, що і дані.

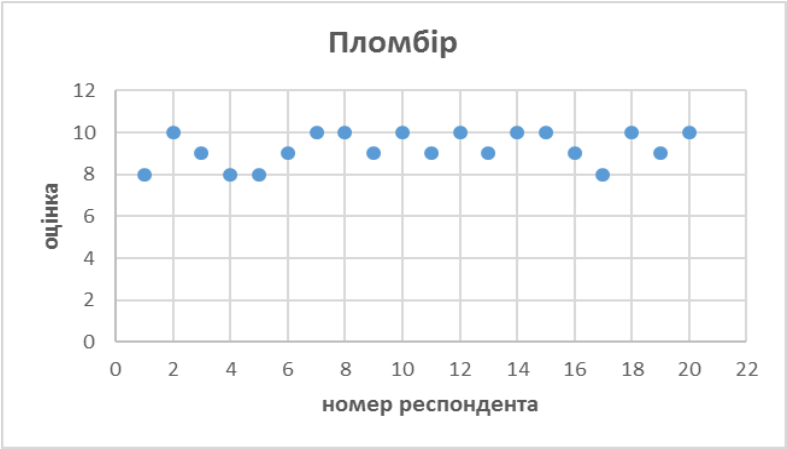


Рис. 1.2.3. Графічне представлення оцінок пломбіру

6. Збережіть результати в особистій папці.

Вимоги до звіту:

Звіт повинен містити:

– відповіді на контрольні питання;

– файли з результатами розрахунків.

Контрольні питання

1. Дайте визначення відомим вам мірам центральної тенденції.

2. Дайте визначення відомим вам мірам варіації.

3. Назвіть функції електронних таблиць для обчислення мір центральної тенденції.

4. Назвіть функції електронних таблиць для обчислення мір мінливості.

5. У чому полягає графічна інтерпретація стандартного відхилення? Дайте змістовну інтерпретацію результатами розрахунків в даній роботі.

Практична робота 1.3
Обчислення мір центральної тенденції та мір мінливості в R

Хід роботи

1. Відкрити R Studio (детальніше про R [41-46]).

2. Визначити три вектори, що представляють оцінки респондентами трьох сортів морозива:

```
> plomb <- c(8,10,9,8,8,9,10,10,9,10,9,10,9,10,10,9,8,10,9,10)
> sliv  <- c(6,7,10,10,9,7,10,10,9,7,10,9,8,10,6,5,10,10,9,10)
> fruct <- c(6,8,10,3,7,4,10,9,4,5,7,4,7,10,9,10,6,5,7,10)
```

3. Підрахувати середнє, медіану та сумарні статистики:

```
> mean(plomb); mean(sliv); mean(fruct)
[1] 9.25
[1] 8.6
[1] 7.05
```

```
> median(plomb); median(sliv); median(fruct)
[1] 9
[1] 9
[1] 7
```

```
> summary(plomb)
   Min. 1st Qu.  Median  Mean  3rd Qu.   Max.
   8.00    9.00     9.00     9.25    10.00     10.00
```

4. Підрахувати стандартне відхилення:

```
> sd(plomb);sd(sliv);sd(fruct)
[1] 0.7863975
[1] 1.667018
[1] 2.372540
```

5. Побудувати три графіки (рис. 1.3.1-1.3.2), що ілюструють розкид оцінок респондентів

```
> plot(plomb)
> title(main="Пломбір")
> plot(sliv)
> title(main="Вершкове морозиво")
> plot(fruct)
> title(main="Фруктове морозиво")
```

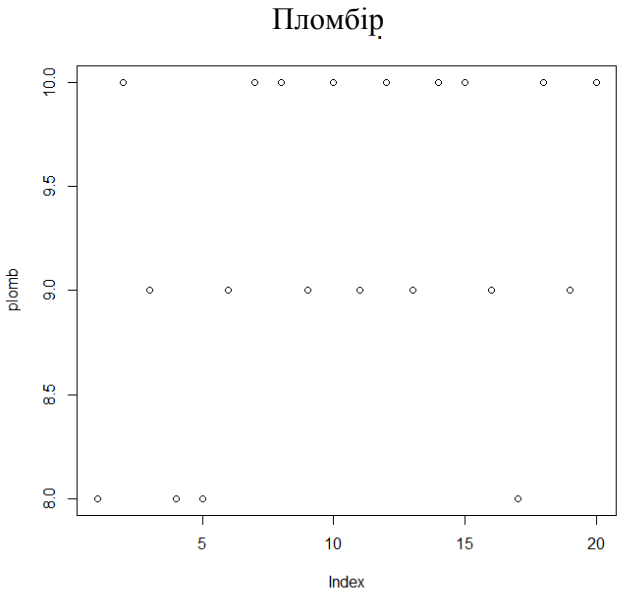
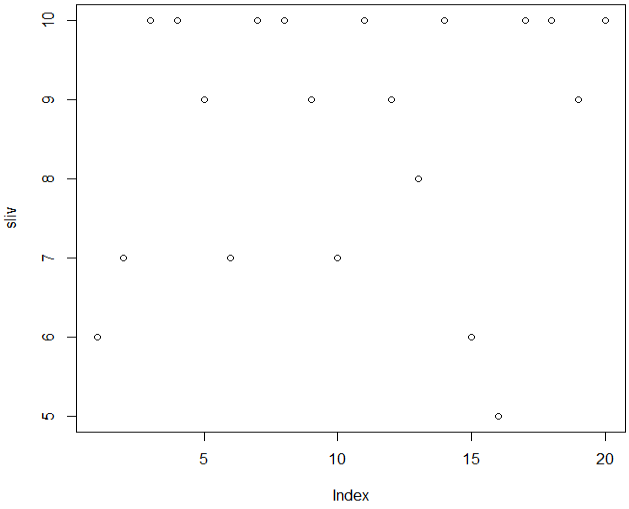


Рис. 1.3.1. Оцінка респондентами пломбіру

Вершкове морозиво



Фруктове морозиво

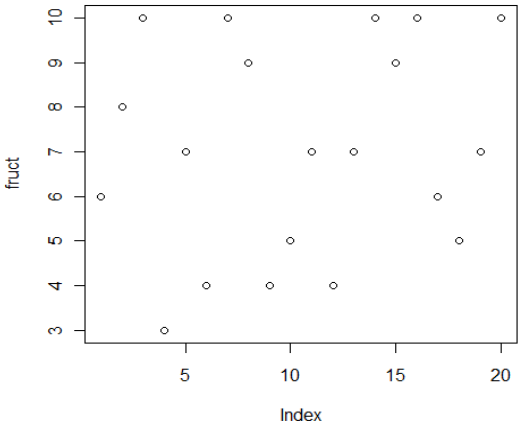


Рис. 1.3.2. Графічне представлення оцінок сортів морозива

6. Визвати справку для команд plot та title, перекласти та вивчити її.

```
>help(title)
```

Опис

This function can be used to add labels to a plot. Its first four principal arguments can also be used as arguments in most high-level plotting functions. They must be of type character or expression. In the latter case, quite a bit of mathematical notation is available such as sub- and superscripts, greek letters, fractions, etc: see plotmath

Використання

```
title(main = NULL, sub = NULL, xlab = NULL, ylab = NULL,
      line = NA, outer = FALSE, ...)
```

Аргументи

main	The main title (on top) using font and size (character expansion) par("font.main") and color par("col.main").
sub	Sub-title (at bottom) using font and size par("font.sub") and color par("col.sub").
xlab	X axis label using font and character expansion par("font.lab") and color par("col.lab").
ylab	Y axis label, same font attributes as xlab.
line	specifying a value for line overrides the default placement of labels, and places them this many lines outwards from the plot edge.
outer	a logical value. If TRUE, the titles are placed in the outer margins of the plot.

7. Об’єднати всі змінні в одну таблицю даних та переглянути її зміст

```
d <- data.frame(plombir=plomb, slivochnoe=sliv, fructovoe=fruct)

> d

   plombir slivochnoe fructovoe
1        8         6         6
2       10         7         8
3        9        10        10
4        8        10         3
5        8         9         7
6        9         7         4
7       10        10        10
8       10        10         9
9        9         9         4
10      10         7         5
11       9        10         7
12      10         9         4
13       9         8         7
14      10        10        10
15      10         6         9
16       9         5        10
17       8        10         6
18      10        10         5
19       9         9         7
20      10        10        10
```

8. Побудувати графік для пломбіру з підписами осей й ломаною червоного кольору (рис.1.3.3):

```
>plot(d$plombir, xlab="номер респондента", ylab="оцінка",
main="Пломбір", col="red", type="l")
```

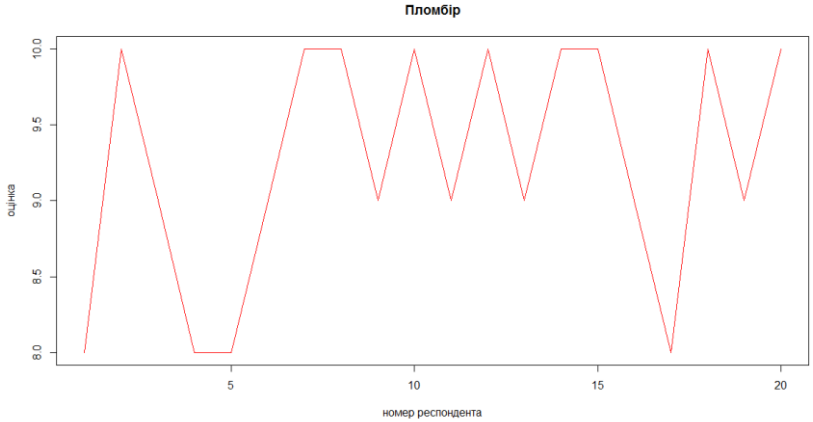


Рис. 1.3.3. Графік з підписами осей

9. Зберегти робочий простір: File, Save Workplace

Контрольні питання

1. Як визначається числовий вектор в мові R?
2. За допомогою якої команди будується графік?
3. Як дати йому назву?
4. Як об'єднати кілька векторів в одну таблицю?
5. Для чого зберігають робочий простір та яким чином це здійснюється?
6. Дайте змістовну інтерпретацію результатам розрахунків в даній роботі.
7. Які можливості R для описової статистики ми вивчили?

Практична робота 1.4
Побудова і розрахунок параметрів розподілу не згрупованих даних за допомогою пакета аналізу електронних таблиць

Мета роботи: будувати частотний розподіл і гістограму не згрупованих даних за допомогою засобу «Гістограма» з Пакета аналізу даних, розраховувати підсумкові статистики за допомогою засобу «Описова статистика».

Завдання 1. Зробити в електронних таблицях (лист 1) розрахунки за формулою Стерджерса (рис. 1.4.1)

	A	B
1	n = 1+3,322*lgN	
2		
3	N	n
4	10000	14
5	5000	13
6	1000	11
7	800	11
8	600	10
9	400	10
10	200	9
11	150	8
12	100	8
13	50	7
14	40	6
15	30	6
16	20	5

Рис. 1.4.1. Розрахунки за формулою Стерджерса

Завдання 2. У табл. 1.4.1. наведено величини продажів деякої комерційної фірми, які фіксувалися протягом 100 днів. Побудуйте гістограму і розрахуйте показники описової статистики для цих даних. Дайте змістовну інтерпретацію отриманих результатів.

Таблиця 1.4.1

Продажі за 100 днів

47	46,7	35,6	56,9	72,1
37,2	46,3	41,5	53,2	64,4
52,4	63,4	34,8	40,6	63
62,8	49,1	46,4	47,6	51,1
62	48,1	49,7	51,3	50
67,3	44,9	50,3	55,6	54,5
28,2	69,7	46,8	51,4	49,7
47,7	58,7	71,9	40,9	39,5
61	73,8	32,6	68,8	32,3
39,1	43,5	42,6	54,9	58,3
43,1	66,6	24,2	50,7	54,4
33,1	33,9	64,5	58,3	56,2
31,5	55,4	37,2	58,6	52,1
40,2	59	43,5	43,6	39,7
42,3	69,2	57,6	40,8	62,4
28,8	49,2	54,7	61,1	46,9
44,3	44,8	58,7	38	41,6
46	56,8	56	34,4	41,8
51,3	46,2	36,3	57,1	45,7
46,3	57,6	38,8	56,4	45,5

Зауваження до ходу виконання:

1. Уведіть дані в колонку А.
2. Розрахуйте за формулою Стерджерса кількість інтервалів для групування.

3. Розрахуйте максимальне, мінімальне значення та ширину інтервалу.

4. Сформуйте «карман» з нижніми межами інтервалів.

	A	B	C	D	E	F	G
1	47					карман	
2	37,2		n	8		24,2	
3	52,4		макс	73,8		30,4	
4	62,8		мін	24,2		36,6	
5	62		ширина	6,2		42,8	
6	67,3					49	
7	28,2					55,2	
8	47,7					61,4	
9	61					67,6	
10	39,1					73,8	
11	43,1						

Рис. 1.4.2. Розрахунки меж інтервалів для групування

5. Оберіть меню «Дані», «Аналіз даних», «Гістограма» (рис.1.4.3)

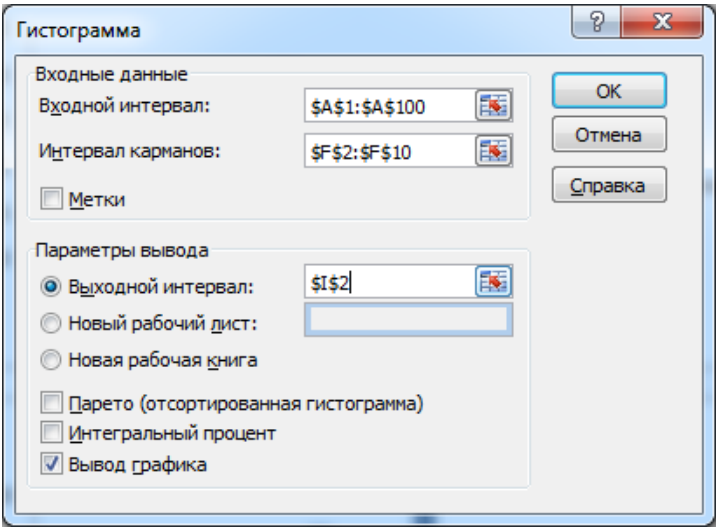
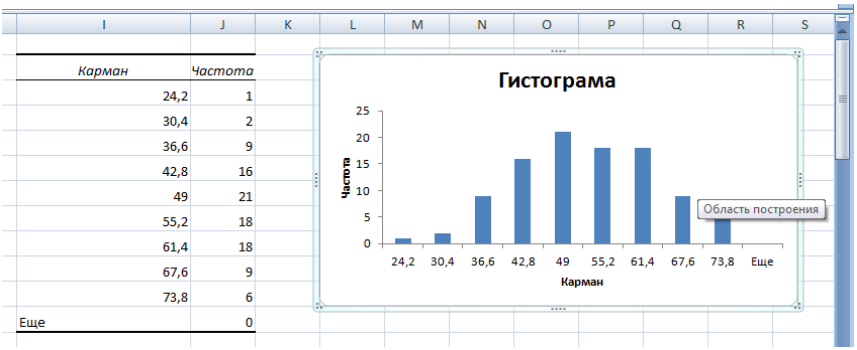


Рис. 1.4.3. Вікно «Гістограма»

Отримайте графік (рис.1.4.4):



Карман	Частота
24,2	1
30,4	2
36,6	9
42,8	16
49	21
55,2	18
61,4	18
67,6	9
73,8	6
Еще	0

Рис. 1.4.4. Побудова гістограми засобами Пакету аналізу

6. Розрахуйте показники описової статистики: меню «Дані», «Аналіз даних», «Описова статистика»:

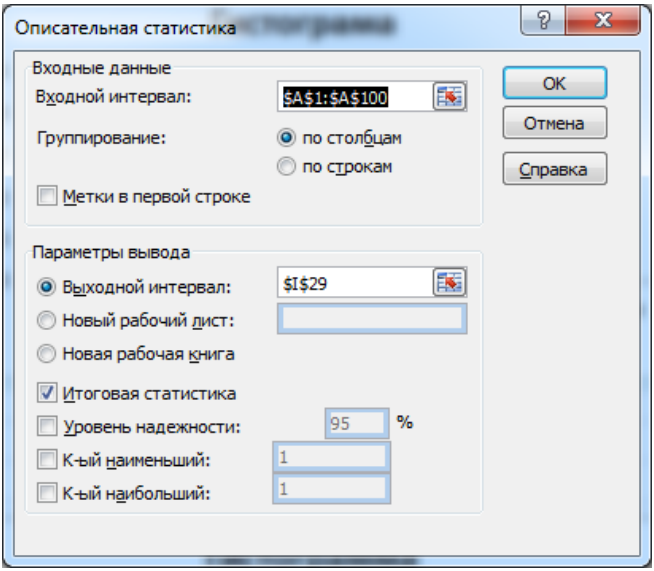


Рис. 1.4.5. Вікно «Описова статистика»

7. Надайте інтерпретацію отриманим даним (рис.1.4.5)

	Столбец1
Среднее	49,596
Стандартная ошибка	1,085417
Медиана	49,15
Мода	37,2
Стандартное отклонение	10,85417
Дисперсия выборки	117,8129
Эксцесс	-0,47441
Асимметричность	0,09077
Интервал	49,6
Минимум	24,2
Максимум	73,8
Сумма	4959,6
Счет	100

Рис. 1.4.5.Результати розрахунків показників описової статистики

Завдання 3. Узагальнити дані (табл.1.4.2), що представляють собою оцінки 75-ти дорослих людей в тесті на визначення коефіцієнта інтелектуальності Стенфорда-Біне. Побудувати гістограму та обчислити показники описової статистики.

Таблиця 1.4.2

Дані IQ 75 дорослих людей

141	104	101	130	148
92	87	115	91	96
100	133	124	92	123
132	118	98	101	107
97	124	118	146	107
110	111	138	121	129
106	135	97	108	108
107	110	101	129	105
105	110	116	113	123
83	127	112	114	105
127	114	113	106	139
95	105	95	105	106
109	102	102	102	89
108	92	131	86	134
104	94	121	107	103

Завдання 4. Знайдіть аналогічні дані та узагальніть їх шляхом групування та побудови гістограми. Надайте інтерпретацію отриманим результатам.

Вимоги до звіту:

Звіт містить назву роботи, її мету, короткий опис ходу роботи, інтерпретацію результатів всіх 4-х завдань, відповіді на контрольні запитання, файли з даними та результатами в Excel (кожне завдання на окремому аркуші книги Excel).

Контрольні питання

1. Як дається ім'я діапазону даних? Для чого потрібні такі імена?
2. Як побудувати гістограму для не згрупованих даних?
3. Як швидко розрахувати показники описової статистики?
4. Що таке «кишеня»?
5. Чи можна розрахувати накопичені частоти і побудувати криву накопичених частот, користуючись Пакетом аналізу?
6. Дайте змістовну інтерпретацію результатів розрахунків в даній роботі.

Практична робота 1.5

Побудова частотного розподілу незгрупованих даних в R

Хід роботи

1. Відкрийте R STUDIO.

2. Визначте таблицю даних з оцінками IQ 75 респондентів.

141	104	101	130	148
92	87	115	91	96
100	133	124	92	123
132	118	98	101	107
97	124	118	146	107
110	111	138	121	129
106	135	97	108	108
107	110	101	129	105
105	110	116	113	123
83	127	112	114	105
127	114	113	106	139
95	105	95	105	106
109	102	102	102	89
108	92	131	86	134
104	94	121	107	103

Уведіть ці дані у діапазон A1:A75 листа Excel, потім виділіть їх і перенесіть у буфер обміну (в R використовується десятинний розбивач "точка"). Щоб створити в R таблицю даних з буфера обміну, використовуйте таку команду:

```
iq <- data.frame(read.table("clipboard"))
```

3. Перегляньте таблицю:

```
>iq
```

4. З’ясуйте структуру об’єкта iq

```
> str(iq)
>'data.frame':   75 obs. of  1 variable:
 $ V1: num  141 92 100 132 97 110 106 107 105 83 ...
```

5. Розглянемо ще один спосіб обміну даними між Excel і R. Можна використовувати функцію scan, щоб скопіювати колонку з числами з Excel в R. Виконайте команду x <- scan (), введіть Ctrl-v щоб вставити в R, та натисніть enter щоб позначити кінець введення для даних для функції scan. Змінна x буде містити числа з Excel. Відзначимо, що scan працює з колонками, а не рядками Excel. Щоб скопіювати рядок, його треба попередньо транспонувати в колонку, а потім копіювати вже цю колонку. Функція scan () не обмежується роботою з Excel. Її можна використовувати також, щоб вставити колонку чисел з таких додатків як Notepad або Word.

6. Побудуйте гістограму оцінок 75 дорослих людей:

```
>  hist(iq$V1)
```

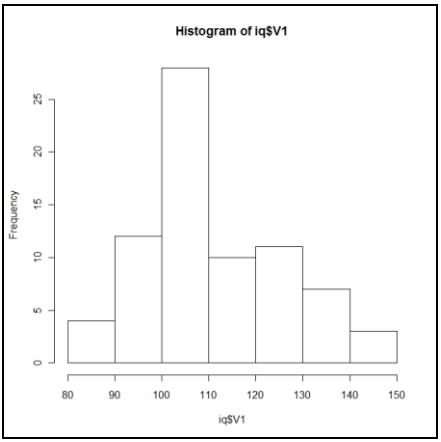


Рис. 1.5.1. Гістограма оцінок IQ 75 дорослих людей

7. Уточніть аргументи команди title і задайте всі заголовки на діаграмі:

```
> args(title)
function (main = NULL, sub = NULL, xlab = NULL, ylab =
NULL,
    line = NA, outer = FALSE, ...)
```

8. Розрахуйте показники описової статистики для даних тесту IQ:

```
> summary(iq$V1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   83.0   101.5     108.0     111.2   122.0   148.0
```

9. Розрахуйте показники мінливості: дисперсію та стандартне відхилення:

```
> var(iq$V1); sd(iq$V1)
[1] 225.2173
[1] 15.00724
```

10. Розрахуйте квартилі:

```
> quantile(iq$V1)
  0%   25%   50%   75%  100%
83.0 101.5 108.0 122.0 148.0
```

11. Найдіть 10-й и 90-й процентилі:

```
>quantile(iq$V1, probs=c(0.10, 0.90))
```

12. Збережіть робочий простір: File, Save Workplace

Контрольні питання

1. Як скопіювати дані з Excel в середовище R?
2. За допомогою якої команди будується гістограма частотного розподілу?
3. Які показники описової статистики розраховує функція summary? Чому в якості аргументу в неї використовується вираз iq$V1? Чи можна було ввести в якості аргументу тільки iq?
4. Які правила вибору імен для об'єктів в R?
5. Як розрахувати перший і дев'ятий дециль?
6. Дайте змістовну інтерпретацію результатам розрахунків в даній роботі.
7. Заповніть таблицю:

Міра	**Функція R**
Мінімальне значення	
Максимальне значення	
1-й квартиль	
3-й квартиль	
Сума	
Асиметрія, ексцес	

Практична робота 1.6
Підготовка даних анкетного опитування до аналізу за допомогою SPSS для Windows

Мета роботи: на навчальному прикладі опитування респондентів-покупців магазину вивчити процес підготовки даних в SPSS.

Постановка задачі

У таблиці 1.6.1 наведено вибірку з відповідей покупців магазину на питання анкети. До даних 11-ти респондентів входять:

- номер анкети;
- стать (1 – «чоловіча», 2 – «жіноча»);
- вік;
- задоволеність покупкою (1 – «повністю задоволений», 4 – «абсолютно незадоволений»).

Таблиця 1.6.1

Дані опитування покупців магазина

Номер анкети	Стать	Вік	Задоволеність покупками в магазині
1	1	50	3
2	2	31	2
3	1	21	3
4	1	18	4
5	2	17	1
6	2	34	1
7	2	32	1
8	2	25	1
9	2	22	2
10	2	21	2
11	2	19	4

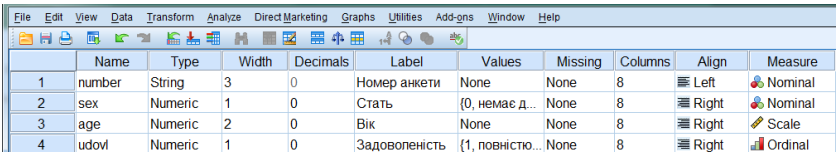
Таблиця 1.6.1 називається матрицею даних. Дані, призначені для обробки в SPSS, повинні бути представлені у ви-

гляді такої матриці. Матриця даних складається з певного числа рядків і стовпців. Рядки і стовпці утворюють прямокутну таблицю. При цьому кожен рядок відповідає одній анкеті, а кожен стовпець – однієї змінній. Так як в нашому невеликому опитуванні брали участь 11 респондентів, матриця містить 11 рядків. Кожен рядок включає чотири стовбця для змінних number (номер анкети), sex (стать), age (вік) і udovl (задоволеність покупками).

Хід роботи

1. Відкрийте SPSS.
2. Щоб описати змінні, в редакторі даних двічі клацніть на клітинці з написом Var або на ярличку Variable View в лівому нижньому кутку екрану.

Визначте змінні, як показано на рис.1.6.1. Нижче описані докладно необхідні кроки.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	number	String	3	0	Номер анкети	None	None	8	Left	Nominal
2	sex	Numeric	1	0	Стать	{0, немає д...	None	8	Right	Nominal
3	age	Numeric	2	0	Вік	None	None	8	Right	Scale
4	udovl	Numeric	1	0	Задоволеність	{1, повністю...	None	8	Right	Ordinal

Рис. 1.6.1. Визначення змінних в редакторі SPSS

Визначення змінної «номер анкети»

- Введіть в поле ***Name*** текст "number" і підтвердіть, натиснувши на клавішу <Enter> або <Tab>.
- Щоб встановити розмір змінної, клацніть в поле ***Type*** на кнопці з трьома крапками. Відкриється діалогове вікно ***Define Variable Type*** (Визначення типу змінної). Виберіть тип ***String*** (строковий) і встановіть число символів "3" (Ми припускаємо, що у нас не більше 999 анкет). Підтвердіть налаштування кнопкою ОК і перейдіть до наступного поля клавішею <Tab>.

- У полях ***Width*** і ***Decimals*** прийміть налаштування, що додаються за замовчуванням.
- Для мітки змінної задайте текст "Номер анкети".
- Для міток значень прийміть пропоновану настройку None, натиснувши <Enter>.
- Прийміть пропоновані настройки в поле ***Missing***, встановіть "8" в поле Columns, "***Right***" – в поле ***Alignment*** і виберіть настройку "***Nominal***" в поле ***Measure***.

Визначення змінної «стать»

- Уведіть в поле ***Name*** текст "sex" і підтвердіть, натиснувши на клавішу <Enter> або <Tab>.
- Щоб встановити розмір змінної, клацніть в поле ***Type*** на кнопці з трьома крапками. Відкриється діалогове вікно ***Define Variable Type*** (Визначення типу змінної). Прийміть запропоновану настройку Numeric (Чисельний) і встановіть довжину "1" і кількість десяткових розрядів "0", так як в цій змінній будуть зберігатися тільки значення 1, 2 або 0. Підтвердіть налаштування кнопкою ОК і перейдіть до наступного поля клавішею <Tab >.
- Для формату стовпця прийміть без змін значення формату "1" і кількості десяткових розрядів "0".
- Для мітки змінної задайте текст «Стать».
- Клацніть в полі ***Value Labels*** на кнопці. Відкриється діалогове вікно ***Define Value Labels*** (Визначення міток значень). Мітки значень визначаються наступним чином.

Визначення міток значень змінної «стать» (рис. 1.6.2)

- Спочатку введіть в поле ***Value*** (Значення) число "1". За допомогою клавіш <Tab>.
- Введіть в поле ***Value label*** (Метка значення) текст «жіноча».
- Клацніть на кнопці ***Add*** (Додати). Мітка значення буде додана в список.

- Повторіть ці дії для значень «2» – «чоловіча» і «0» – «немає даних».

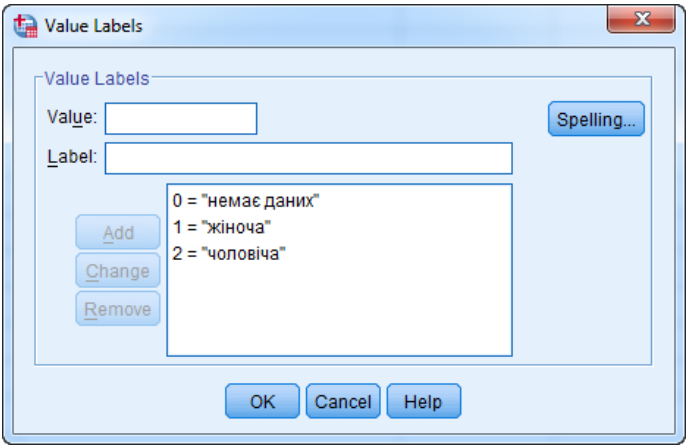


Рис. 1.6.2. Визначення міток значень зміної «стать» у вікні «Value Labels»

Визначення міток значень змінної «вік»

- Введіть в поле ***Name*** текст «age» та натисніть клавішу.
- Щоб встановити розмір змінної, клацніть в поле ***Type*** на кнопці з трьома крапками. Відкриється діалогове вікно ***Define Variable Type***. Прийміть запропоновану настройку Numeric і встановіть довжину "2" (ми припускаємо, що всі респонденти не старше 99 років) і кількість десяткових розрядів "0". Підтвердіть налаштування кнопкою ОК і перейдіть до наступного поля клавішею <Tab>.
- У полях ***Column*** format і ***Decimals*** прийміть настройки, пропоновані за умовчанням.

Для мітки змінної введіть текст «Вік», а для міток значень прийміть пропоновану настройку None, натиснувши <Enter>.

Щоб задати пропущені значення, клацніть в поле ***Missing values*** на кнопці з трьома крапками. Відкриється діалогове вікно ***Define Missing Values***. За замовчуванням обрано варіант No missing values (Немає пропущених значень), тобто всі значення розглядаються як допустимі. Введіть одиничне відсутнє значення "0" і закрийте діалогове вікно кнопкою ОК.

Прийміть пропоновані настройки "8" в поле ***Columns***, "Right" в поле ***Alignment*** і "Scale" в поле ***Measure***.

Визначення міток значень змінної «Задоволеність»
(рис. 1.6.3)

- Введіть в поле Name текст «udovl» і підтвердіть, натиснувши <Tab>.
- Щоб встановити розмір змінної, клацніть в поле ***Type*** на кнопці з трьома крапками. Відкриється діалогове вікно ***Define Variable Type***. Прийміть запропоновану настройку Numeric і встановіть довжину "1" і кількість десяткових розрядів "0", так як в цій змінній будуть зберігатися тільки значення від 1 до 4 і 0 як відсутнє значення. Підтвердіть налаштування кнопкою ОК і перейдіть до наступного поля клавішею <Tab>.
- Для формату стовпця прийміть значення "1" і кількість десяткових розрядів "0".
- Для мітки змінної задайте текст "Задоволеність".
- Клацніть в полі ***Value label*** на кнопці Відкриється діалогове вікно **Define Value Labels.**
- Спочатку введіть в поле ***Value*** (Значення) число "1". За допомогою клавіш <Tab>.
- Введіть в поле ***Value label*** (Метка значення) текст "задоволений повністю".
- Клацніть на кнопці **Add** (Додати). Мітка значення буде додана в список.

Повторіть ці дії для значень "2" – «скоріше так, ніж ні», "3" – «скоріше ні, ніж так», 4 – «повністю не задоволений» і "0" – «немає даних».

3. Перейдіть на вкладку Data View і введіть дані 11 анкет (згідно таб.1.6.1).

4. Збережіть файл з даними в особистій папці

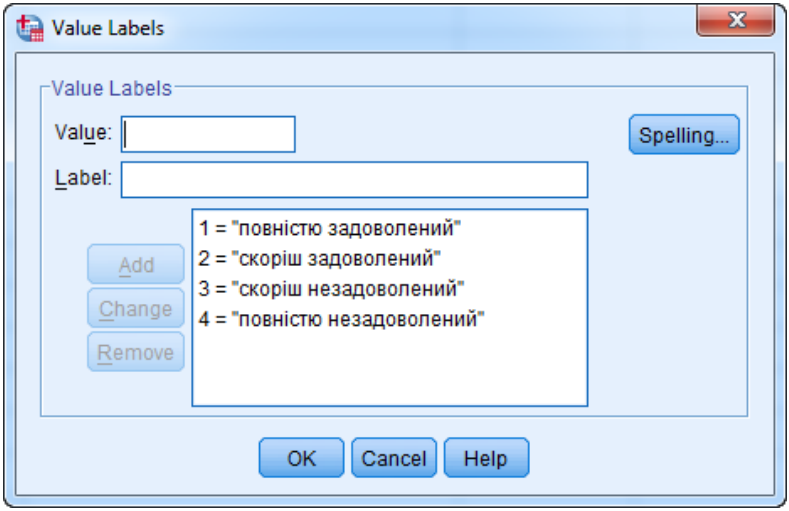


Рис. 1.6.3. Визначення міток значень змінної «задоволеність» у вікні «Value Labels»

Вимоги до звіту:

Звіт повинен містити:

– відповіді на контрольні питання;

– файл з введеними даними.

Контрольні питання

1. Як запустити SPSS?
2. Як перейти до опису змінних в SPSS?
3. Що означає тип змінної «String»?
4. Що встановлюють в колонці «Decimals» при описі змінних?
5. Що таке мітки значень?

Практична робота 1.7
Побудова частотного розподілу і обчислення статистичних показників за допомогою SPSS для Windows

Мета роботи: розглянути основні технічні прийоми при роботі з SPSS; навчитися будувати частотний розподіл, обчислювати основні описові статистики.

Хід роботи

1. Відкрити файл, який містить підсумки опитування покупців магазину з попередньої практичної роботи, вибравши меню: ***File, Open***

2. Побудувати частотний розподіл для змінної «стать». Для цього (див. рис. 1.7.1):

3. Вибрати меню ***Analyze*** (Аналіз), ***Descriptive statistics*** (Описова статистика), ***Frequencies*** ... (Частотний розподіл)

- Виділити змінну «стать», клацнувши на ній. Натиснути кнопку з трикутною стрілкою, перенісши виділену змінну зі списку вихідних змінних в список обраних змінних.
- Натиснути кнопку OK.

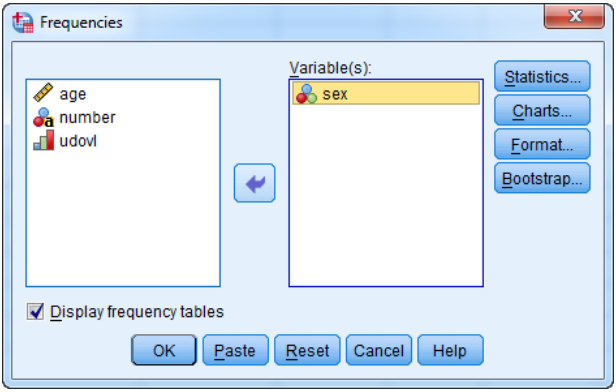


Рис. 1.7.1. Діалогове вікно «Frequencies». Перший етап побудови частотного розподілу для змінної «стать»

- Ознайомиться з отриманими результатами у вікні перегляду (рис. 1.7.2).

Стать

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	жіноча	3	27,3	27,3	27,3
	чоловіча	8	72,7	72,7	100,0
	Total	11	100,0	100,0	

Рис. 1.7.2. Таблиця частот для змінної «стать» у вікні перегляду

4. Повернутися у редактор даних
   - Або обрав меню ***Windows***
   - Або клацнути на панелі інструментів на символі редактора даних

Аналогічно розрахувати частотний розподіл для змінної «задоволеність покупками» (см. рис. 1.7.3).

Задоволеність

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	повністю задоволений	4	36,4	36,4	36,4
	скоріш задоволений	3	27,3	27,3	63,6
	скоріш незадоволений	2	18,2	18,2	81,8
	повністю незадоволений	2	18,2	18,2	100,0
	Total	11	100,0	100,0	

Рис. 1.7.3. Таблиця частот для змінної «задоволеність покупками» у вікні перегляду

5. Визначити найбільше, найменше та середнє значення змінної «вік» а також моду, медіану та стандартне відхилення (див. рис. 1.7.4).

- Обрати меню ***Analyze*** (Аналіз), ***Descriptive statistics*** (Описова статистика), ***Frequencies***… (Частотний розподіл).
- В діалоговому вікні ***Frequencies*** кликнути на кнопці ***Reset*** (Скидання), а потім перенести змінну «вік» в список обраних змінних.
- Кликнути кнопку ***Statistics…***
- У вікні ***Frequencies: Statistics*** установити флажки ***Min (***мінімальне***), Max (***максимальне***), Mean*** (середнє арифметичне), а також флажки для моди, медіани та стандартного відхилення (***Std. deviation***).
- Щелкнути на кнопці ***Continue.***
- Зняти флажок ***Display Frequencies tables*** (Показувати частотні таблиці).
- Кликнути на кнопці ОК.

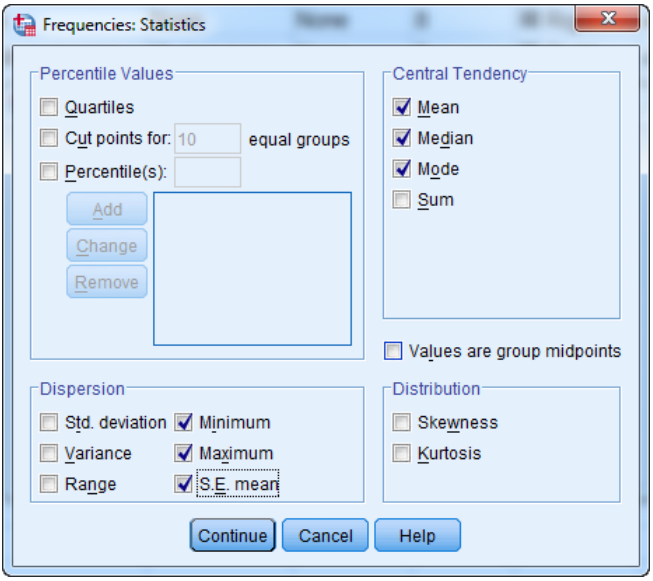


Рис. 1.7.4. Діалогове вікно «Frequencies: Statistics»

- Ознайомитися з результатами у вікні перегляду (рис. 1.7.5).

6. Зберегти результати розрахунків в особистій папці.

Statistics

Вік

N	Valid	11
	Missing	0
Mean		26,36
Median		22,00
Mode		21
Std. Deviation		9,801
Minimum		17
Maximum		50

Рис. 1.7.5. Вікно перегляду з результатами розрахунків

Вимоги до звіту:

Звіт повинен містити:

– відповіді на контрольні питання;

– файл з введеними даними.

Контрольні питання

1. Як перенести змінну зі списку вихідних змінних в список обраних змінних у вікні «Frequencies»?

2. Описати параметри вікна «Frequencies: Statistics».

3. Яке призначення п'яти наведених нижче стандартних кнопок в головному діалоговому вікні?

- ***OK***
- ***Paste***
- ***Reset***
- ***Cance***
- ***Help***

Практична робота 1.8
Побудова та редагування графіків в SPSS

Мета роботи: навчитися будувати та редагувати графіки в SPSS засобами **«Frequencies: Charts»**

Хід роботи

1. Відкрити файл, який містить дані опитування покупців магазину. Файл даних містить 4 змінних: «номер анкети», «стать», «вік» респондента та «задоволеність» покупками. Меню: ***File, Open.***

2. Побудувати графік частотного розподілу для змінної «задоволеність»**.**

Для цього:

- Обрати меню ***Analyze*** (Аналіз), ***Descriptive statistics*** (Описова статистика), ***Frequencies***… (Частотний розподіл)
- Виділити змінну «задоволеність»**,** клацнувши по ній.
- Натиснути кнопку с трикутниковою стрілкою, перенести обрану змінну з списку вихідних змінних в список обраних змінних.
- Клацнути на кнопці ***Charts***… (Діаграми).
- У діалоговому вікні ***Frequencies: Charts*** … (Частоти: Діаграмми) обрати опцію ***Bar Charts***, в області ***Chart Values*** (Значення діаграми) клацнути на опції ***Percentage*** (Відсотки) і далі на кнопці ***Continue*** (рис. 1.8.1).

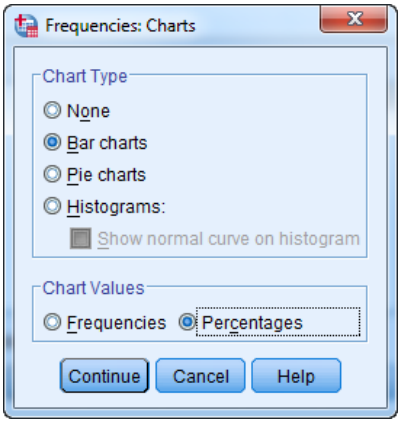


Рис. 1.8.1. Діалогове вікно «Frequencies: Charts»

- Зняти прапорець ***Display Frequency tables*** (Показывать частотные таблицы)
- Натиснути кнопку OK.
- Ознайомиться з отриманими результатами у вікні перегляду.

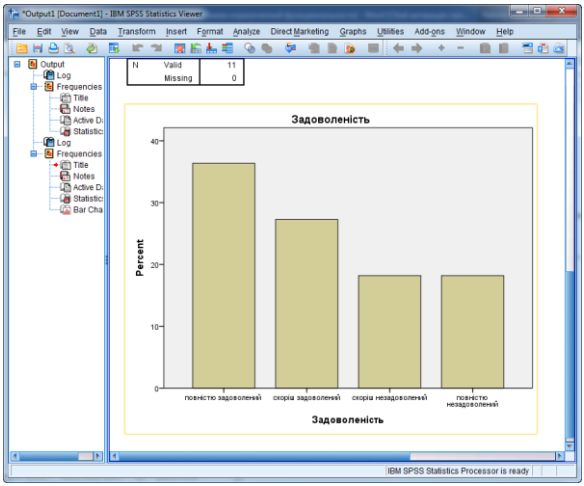


Рис. 1.8.2. Вікно перегляду

Щоб відредагувати графік, потрібно виконати наступне.

- Клацнути двічі на будь-якій точці в межах графіка, після чого він буде поміщений в редактор діаграм.
- Задати об'ємний вигляд стовпців на графіку:
- В меню редактора діаграм обрати:

Format (Формат), ***Bar Style*** (Вид столбця)

відкриється діалогове вікно ***Bar styles***

- Щелкнути на області ***3-D effects***
- В полі Depth (Глибина) ввести число “40”
- Кликнути **Apply All (**Застосувати для всіх) **і далі** на **Close.**

- Дати графіку назву
- Меню ***Chart*** (Диаграма), ***Title*** (Заголовок)
- В полі ***Title1*** (Заголовок 1) введіть текст “Опрос респондентов», а в поле ***Title2*** (Заголовок 2) – “Відношення до магазину”.
- Виділіть графік рамкою

Chart (Диаграма)

Outer Frame (Рамка ззовні)

- Помітити стовбці процентними показниками
- В меню редактора діаграм оберіть:

Format (Формат)

Bar Label Style (Мітки стовбців)

Відкриється діалогове вікно ***Bar Label Style*** (Мітки стовбців)

- Щелкнути на полі ***Framed*** (У рамкі), далі на ***Apply All***, потім на ***Close.***
- Закрити редактор діаграм.
- Зберегти відредагований графік ***File, Save As***

Завдання для самостійного виконання

У табл. 1.8. наведено дані про використання респондентами Internet для особистих цілей [6]. Ці дані 30-ти респондентів включають:

- стать (1 – «чоловіча», 2 – «жіноча»);

- ступінь знайомства з Internet (1 – «майже не знайомий», 7 – «добре знайомий»);
- використання Internet (в годинах на тиждень);
- ставлення до Internet і Internet-технологій (виміряне за шкалою: 1 – «негативне», 7 – «максимально позитивне»),
- використання Internet для придбання товарів і банківських операцій (1 – «так», 2 – «ні»)

Таблиця 1.8.1

Дані про використання Internet

Номер	Стать	Знайомство з Інтернет	Використання Internet (в часах в тиждень)	Відношення до Internet	Відношення до Internet-технологій	Використання для покупок	Використання для банківських операцій
1	1	7	14	7	6	1	1
2	2	2	2	3	3	2	2
3	2	3	3	4	3	1	2
4	2	3	3	7	5	1	2
5	1	7	13	7	7	1	1
6	2	4	6	5	4	1	2
7	2	2	2	4	5	2	2
8	2	3	6	5	4	2	2
9	2	3	6	6	4	1	2
10	1	7	15	7	6	1	2
11	2	4	3	4	3	2	2
12	2	5	4	6	4	2	1
13	1	6	9	6	5	2	1
14	1	6	8	3	2	2	2
15	1	0	5	5	4	2	2
16	2	4	3	4	3	2	2
17	1	6	9	5	3	1	1
18	1	4	4	5	4	1	2
19	1	7	14	6	6	1	1
20	2	6	6	6	4	2	2
21	1	6	9	4	2	2	2
22	1	5	5	5	4	2	1
23	2	3	2	4	2	2	2
24	1	7	15	6	6	1	1
25	2	6	6	5	3	!	2
26	1	6	13	6	6	1	1
27	2	5	4	5	5	1	1
28	2	4	2	3	2	2	2
29	1	4	4	5	3	1	2
30	1	3	3	7	5	1	2

Необхідно зробити:

1. Описати необхідні змінні та ввести дані.
2. Розрахувати частотний розподіл для змінних «стать», «використання Інтернет для покупок» і «використання Інтернет для банківських операцій».
3. Розрахувати середнє число годин в тиждень, витрачене користувачами на використання Інтернет, а також моду, медіану, стандартне відхилення і мінімальне і максимальне значення для цієї ознаки.
4. Інтерпретувати отримані дані.
5. Зберегти дані і результати розрахунків в особистій папці (повинно бути два файли).

Вимоги до звіту:

Звіт повинен містити:

– відповіді на контрольні питання;

– файли з результатами розрахунків.

Контрольні питання

1. Як відкрити файл з даними в SPSS для Windows і як зберегти результати розрахунків в файлі?
2. Що таке частотний розподіл змінної і за допомогою яких команд меню воно будується?
3. Як розрахувати для обраної змінної моду, медіану, середнє, дисперсію, стандартне відхилення?
4. Призначення кнопок в головному діалоговому вікні.
5. Як повернутися з вікна перегляду результатів SPSS «Viewer» до вікна редактора даних SPSS «Data Editor»? Указати 3 способи.
6. Як відмовитися від показу частотних таблиць у вікні перегляду результатів?
7. Як побудувати гістограму в SPSS?
8. Як відредагувати графік?
9. Як зберегти побудований графік в файлі?
10. Назвіть основні елементи вікна редактора діаграм.

Практична робота 1.9
Обробка даних анкетного опитування в середовищі R

Мета роботи: на прикладі опитування покупців магазину навчитися створювати таблицю з даними в середовищі R і будувати графіки частотного розподілу

Хід роботи

Визначте таблицю даних, що представляє дані опитування покупців магазину (див. роботу 1.6).

1. Визначте числовий вектор для змінної «вік»:

```
age <- c(50,31,21,18,17,34,32,25,22,21,19)
```

2. Визначте текстовий (категоріальний вектор) стать:

```
> sex <- c("male", "female", "male", "male", "female", "female", "female", "female", "female", "female", "female")
```

3. Перевірте, чи є він текстовим?

```
> is.character(sex)
[1] TRUE
```

4. Перевірте, чи є він вектором?

```
>  is.vector(sex)
[1] TRUE
```

5. Розгляньте його структуру:

```
> str(sex)
```

```
 chr [1:11] "male" "female" "male" "male" "female" "female" "female" "female" "female" ...
```

6. Проінформируйтеуйте R про те, що це категоріальний тип даних.

```
 > sex.f <- factor(sex)
> sex.f
 [1] male   female male   male   female female female female female female female
Levels: female male
```

7. Визначте текстовий вектор для змінної «задоволеність покупками».

```
>udovl <- c( "sknet", "skda","sknet", "neud", "ud","ud","ud","ud","skda","skda", "neud")
> udovl.f <-factor(udovl)
> udovl.f

 [1] sknet skda  sknet neud  ud    ud    ud    ud    skda  skda  neud
Levels: neud skda sknet ud
```

8. Укажіть, що це упорядкована категоріальна змінна:

```
 >udovl.o <- ordered(udovl.f, levels=c("neud", "sknet", "skda", "ud"))
> udovl.o
 [1] sknet skda  sknet neud  ud    ud    ud    ud    skda  skda  neud
Levels: neud < sknet < skda < ud
```

9. Розрахуйте частотний розподіл покупців за ознаками стать та задоволеність покупками.

```
> table(sex)
sex
female   male
     8      3
```

```
> table(udovl)
udovl
 neud  skda sknet    ud
    2     3     2     4
```

10. Побудуйте графіки для змінних стать та задоволеність покупками.

```
> plot(sex.f)
```

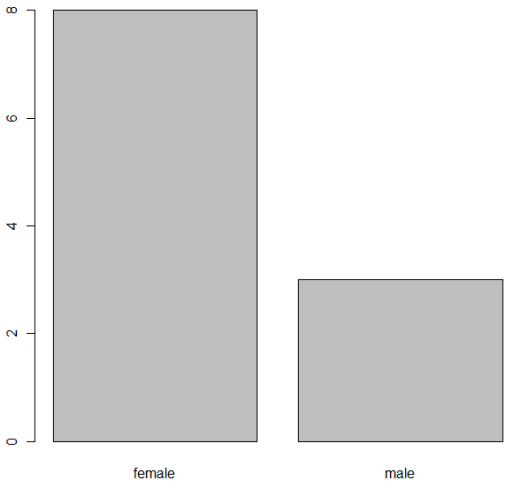


Рис. 1.9.1. Розподіл покупців за статтю

```
> plot(udovl.o)
```

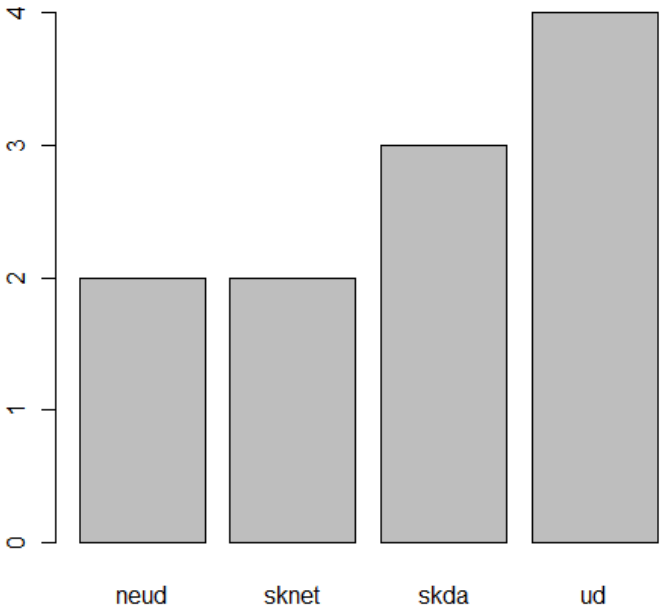


Рис. 1.9.2. Розподіл покупців за задоволеністю покупками

11. Підрахуйте самостійно середній вік покупців, а також моду, медіану, стандартне відхилення, мінімальний і максимальний вік. Заповніть таблицю за результатами розрахунків:

Міра	Значення
Середнє	
Мода	
Медіана Стандартне відхилення	
Мінімальне	
Максимальне	

12. Об’єднайте три створені вектори в одну таблицю даних:

```
>demping <- data.frame(sex1=sex.f, age1=age,
udovl1=udovl.o)
> demping
     sex1 age1 udovl1
1    male   50  sknet
2  female   31   skda
3    male   21  sknet
4    male   18   neud
5  female   17     ud
6  female   34     ud
7  female   32     ud
8  female   25     ud
9  female   22    skda
10 female   21   skda
11 female   19   neud
```

13. Перегляньте структуру створеної таблиці даних:

```
> str(demping)
'data.frame':   11 obs. of  3 variables:
 $ sex1  : Factor w/ 2 levels "female","male": 2 1 2 2 1 1 1 1 1
1 ...
 $ age1  : num  50 31 21 18 17 34 32 25 22 21 ...
 $ udovl1: Ord.factor w/ 4 levels "neud"<"sknet"<..: 2 3 2 1 4 4
4 4 3 3 ...
```

14. Виведіть інформацію по жінках-покупцях:

```
  > demping[demping$sex=="female",]
    sex1   age1 udovl1
2  female   31   skda
5  female   17     ud
6  female   34     ud
7  female   32     ud
8  female   25     ud
9  female   22   skda
10 female   21   skda
11 female   19   neud
```

15. Збережіть робочий простір: File, Save Workplace

Контрольні питання

1. Як визначається числовий вектор в середовищі R?
2. Як визначається текстовий вектор?
3. Як задаються фактори (категоріальні змінні)?
4. Яким чином задається впорядкованість в градаціях фактору?
5. Як побудувати гістограму частотного розподілу?
6. Виведіть інформацію 1) по чоловіках-покупцям: 2) по покупцям, повністю задоволеним покупками; 3) по покупцям, повністю незадоволеним покупками.
7. Дайте змістовну інтерпретацію результатам розрахунків в даній роботі.

Практична робота 1.10
Обчислення середнього значення згрупованих даних

Мета роботи: навчитися використовувати можливості електронних таблиць та SPSS для розрахунку середнього згрупованих даних, закріпити навички розрахунку процентних частот і побудови графіків.

Постановка задачі

У табл. 1.10.1 [8, С.57] наведено згруповані результати тесту лінгвістичних здібностей студентів університету (903 чоловіків і 543 жінок). Обчислити для цих даних середнє значення тесту для чоловіків і для жінок і побудувати на одному графіку два полігони для порівняння.

Таблиця 1.10.1

Результати тесту лінгвістичних здібностей студентів університету

Нижня межа інтервалу	**Верхня межа інтервалу**	**Чоловіки** (осіб)	**Жінки** (осіб)
200	250	1	0
250	300	1	1
300	350	4	1
350	400	25	9
400	450	96	32
450	500	171	86
500	550	202	128
550	600	174	117
600	650	138	85
650	700	63	56
700	750	27	28
750	800	1	4
	Разом	903	547

Для обчислення середнього згрупованих даних скористайтеся формулою:

$$\overline{X} = \frac{\sum_{i=1}^{k} \overline{Xi} \circ Ni}{N} \qquad (1.10.1)$$

де

$\overline{X}$ – середнє згрупованих даних

X_i – середнє в і-інтервалі

N_i – частота в і-інтервалі

N – загальна кількість об’єктів

Хід роботи

1. Уведіть вихідні дані в таблиці Excel (табл. 1.10.2)

Таблиця 1.10.2

A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Ніжня межа	Верхня межа	Середнє в інтервалі **Xi**	Чоловіки **Ni**	**Xi*Ni**	% чол. в інтервалі	Жін-ки Ni	Xi*Ni	% жін. в інтервалі
200	250		1			0		
250	300		1			1		
300	350		4			1		
350	400		25			9		
400	450		96			32		
450	500		171			86		
500	550		202			128		
550	600		174			117		
600	650		138			85		
650	700		63			56		
700	750		27			28		
750	800		1			4		
			Разом чол.	Середнє (чол)		Разом жін	Середнє (жін)	

2. Розрахуйте необхідні дані. Ви маєте отримати таблицю:

Ніжня межа	Верхня межа	Середнє в інтерва-лі Xi	Чоловіки Ni	Xi*Ni	% чол. в інтерва-лі	Жінки Ni	Xi*Ni	% жінок в інтервалі
200	250	225	1	225	0,11	0	0	0,00
250	300	275	1	275	0,11	1	275	0,18
300	350	325	4	1300	0,44	1	325	0,18
350	400	375	25	9375	2,77	9	3375	1,65
400	450	425	96	40800	10,63	32	13600	5,85
450	500	475	171	81225	18,94	86	40850	15,72
500	550	525	202	106050	22,37	128	67200	23,40
550	600	575	174	100050	19,27	117	67275	21,39
600	650	625	138	86250	15,28	85	53125	15,54
650	700	675	63	42525	6,98	56	37800	10,24
700	750	725	27	19575	2,99	28	20300	5,12
750	800	775	1	775	0,11	4	3100	0,73
			Разом чол.	Середнє (чол)		Разом жін	Середнє (жін)	
			903	540,89		547	561,65	

3. Побудуйте на одному графіку два полігони. Для цього виділіть дані в стовбцях C, F, I, оберіть ***Вставка, Діаграма,*** оберіть тип діаграми «точкова, поєднана відрізками». Задайте усі необхідні підписи (рис.1.10.1).

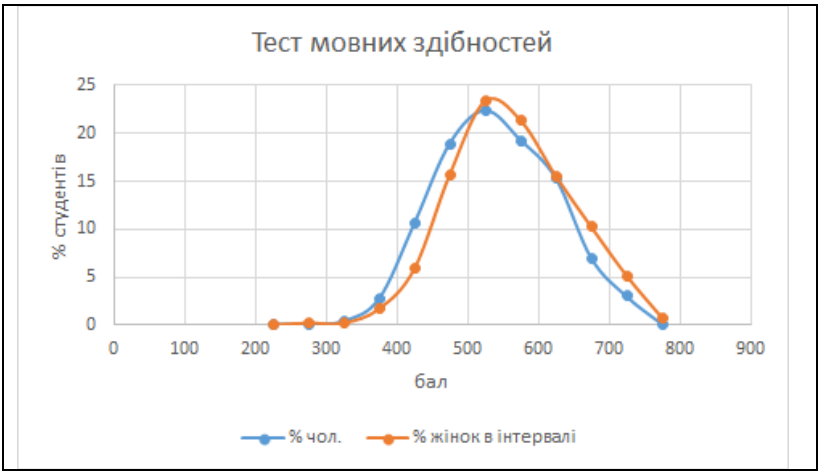


Рис. 1.10.1. Порівняння розподілів чоловіків та жінок за рівнем лінгвістичних здібностей

***Вимоги* до звіту**:

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості з теми, відповіді на контрольні питання, файл з результатами.

Контрольні питання

1.Як обчислюється середнє значення для згрупованих даних в Excel?

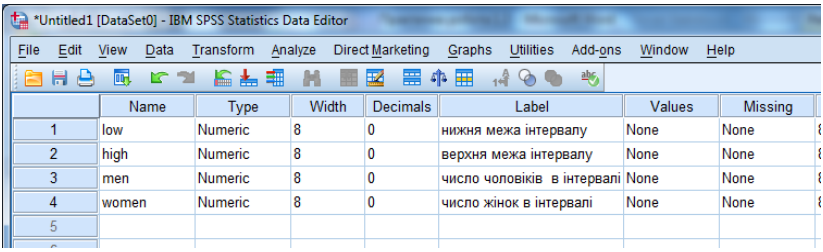
2. Як копіювати формули в Excel?

3. Як виділяти несуміжні блоки комірок в Excel?

4. Дайте змістовну інтерпретацію результатів розрахунків в даній роботі.

Рішення в SPSS

1. Описати 4 змінні (рис.1.10.2):



	Name	Type	Width	Decimals	Label	Values	Missing
1	low	Numeric	8	0	нижня межа інтервалу	None	None
2	high	Numeric	8	0	верхня межа інтервалу	None	None
3	men	Numeric	8	0	число чоловіків в інтервалі	None	None
4	women	Numeric	8	0	число жінок в інтервалі	None	None
5							

Рис. 1.10.2. Опис змінних

2. Ввести дані, як показано на рис.1.10.3:

	low	high	men	women	var
1	200	250	1	0	
2	250	300	1	1	
3	300	350	4	1	
4	350	400	25	9	
5	400	450	96	32	
6	450	500	171	86	
7	500	550	202	128	
8	550	600	174	117	
9	600	650	138	85	
10	650	700	63	56	
11	700	750	27	28	
12	750	800	1	4	
13					

Рис. 1.10.3. Уведення вихідних даних

3. Обчислити знову змінну – середнє в інтервалі **з** іменем **meantest** (рис. 1.10.4)**,** обравши ***Transform, Computer Variable***

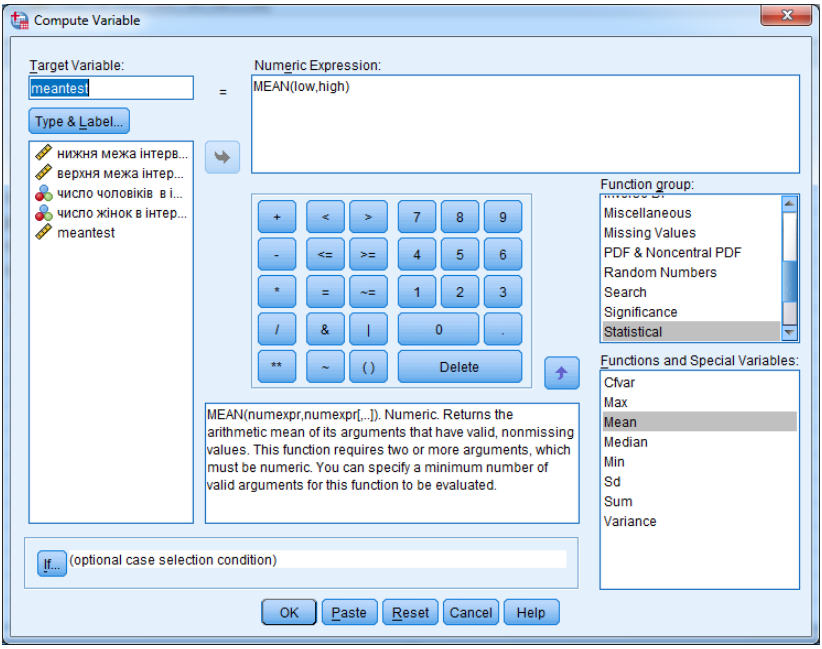
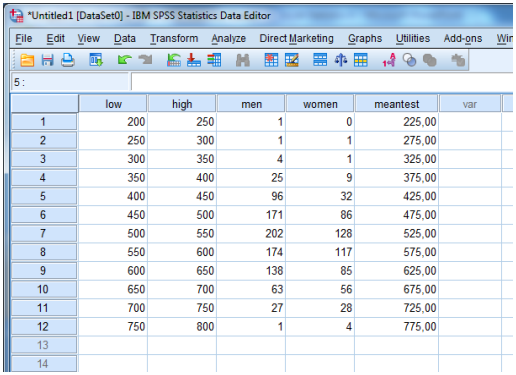


Рис. 1.10.4. Обчислення середнього в інтервалі

В файлі з’являється нова змінна:



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

	low	high	men	women	meantest	var
1	200	250	1	0	225,00	
2	250	300	1	1	275,00	
3	300	350	4	1	325,00	
4	350	400	25	9	375,00	
5	400	450	96	32	425,00	
6	450	500	171	86	475,00	
7	500	550	202	128	525,00	
8	550	600	174	117	575,00	
9	600	650	138	85	625,00	
10	650	700	63	56	675,00	
11	700	750	27	28	725,00	
12	750	800	1	4	775,00	
13						
14						

4. Зважити змінні частотою (числом чоловіків), як показано на рис. 1.10.5.

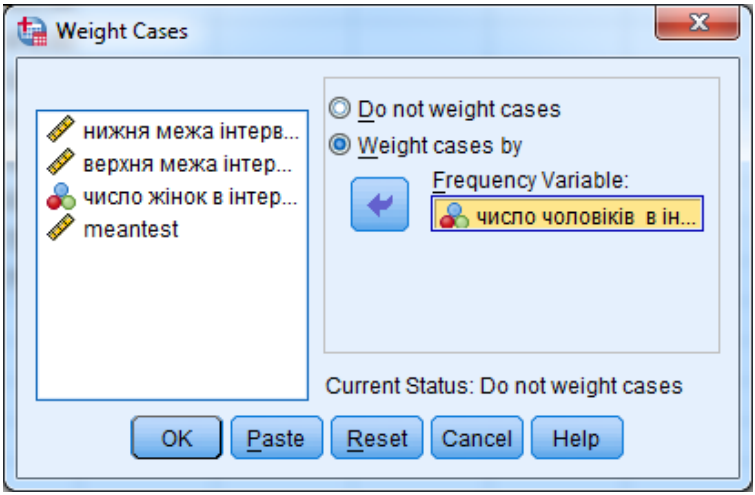


Рис. 1.10.5. Зваження даних

5. Обрати ***Analyze, Descriptive statistics, Frequencies***

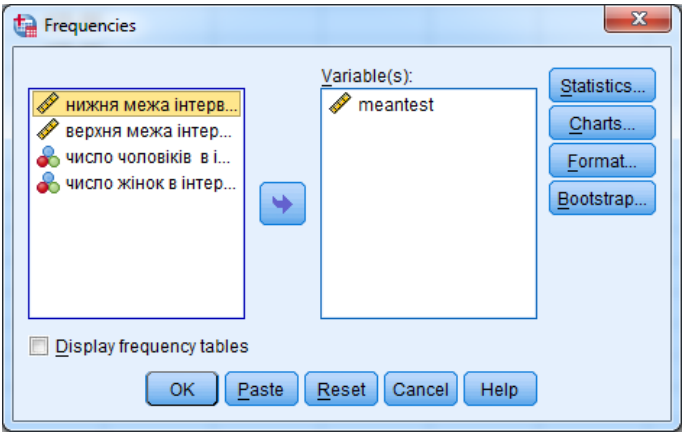


Рис. 1.10.6. Вікно «Частоти»

6. Переглянути результати у вікні виводу (рис.1.10.7):

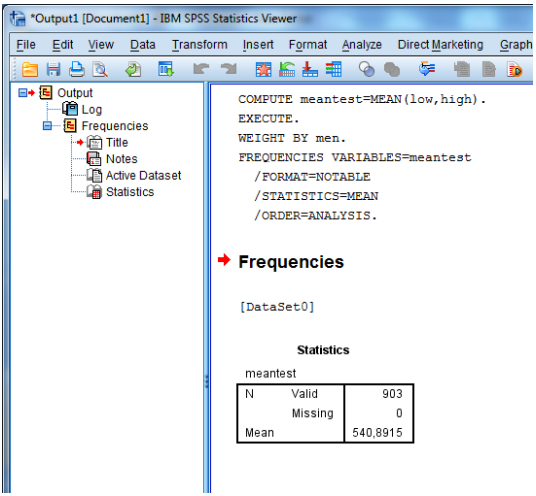


Рис.1.10.7. Результати розрахунків

7. Аналогічно, зробити підрахунки для жінок:

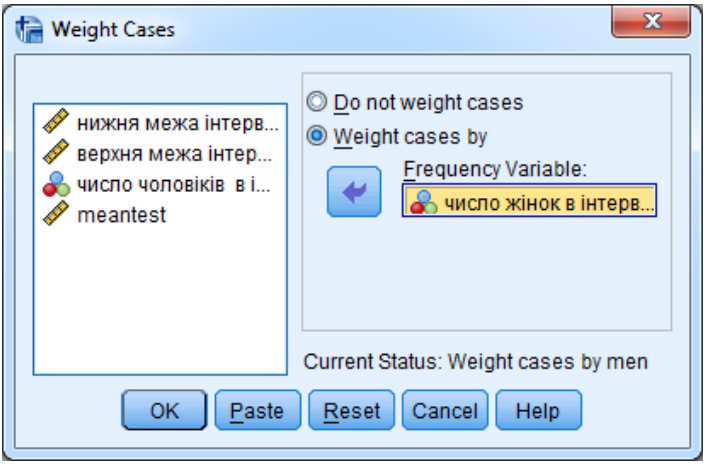


Рис.1.10.8. Розрахунки для жінок

Ви маєте отримати наступне:

`WEIGHT BY men.`
`FREQUENCIES VARIABLES=meantest`
`  /FORMAT=NOTABLE`
`  /STATISTICS=MEAN`
`  /ORDER=ANALYSIS.`
 	
Frequencies
 	
`[DataSet0]`
 	
Statistics
meantest
N Valid 903
Missing 0
Mean 540,8915
	`WEIGHT BY women.`
	`FREQUENCIES VARIABLES=mea`
	`  /FORMAT=NOTABLE`
	`  /STATISTICS=MEAN`
	`  /ORDER=ANALYSIS.`
	
	Frequencies
	
	`[DataSet0]`
	
	Statistics
	meantest
	N Valid 547
	Missing 0
	Mean 561,6545

Чоловіки	**Жінки**

Те ж саме, можна зробити через вікно синтаксису (кнопка **«Paste»),** як показано на рис. 1.10.9.

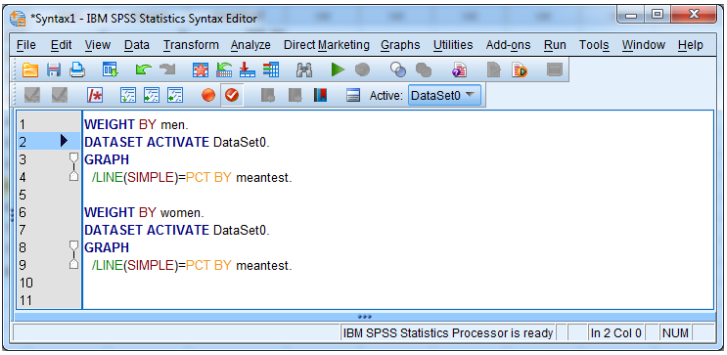


Рис. 1.10.9. Вікно синтаксису

Побудуйте графіки: ***Graphs, Legacy Dialogs, Line*** (рис.1.10.10).

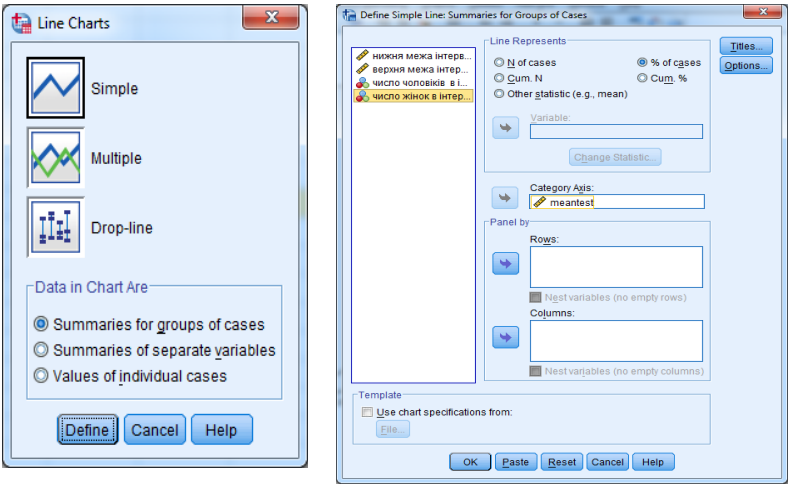


Рис. 1.10.10. Побудова графіку

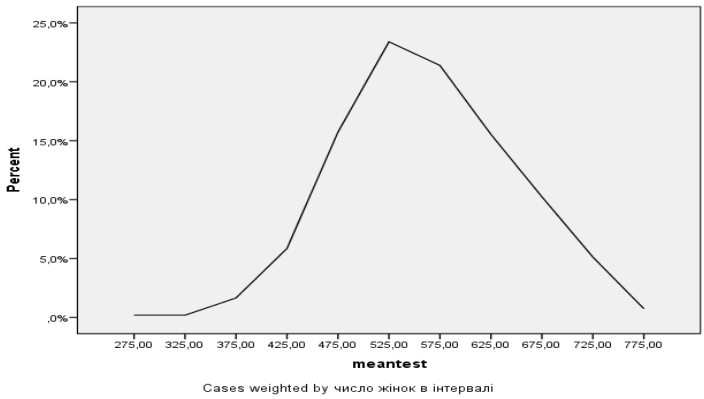


Рис. 1.10.11. Розподіл чоловіків за тестом

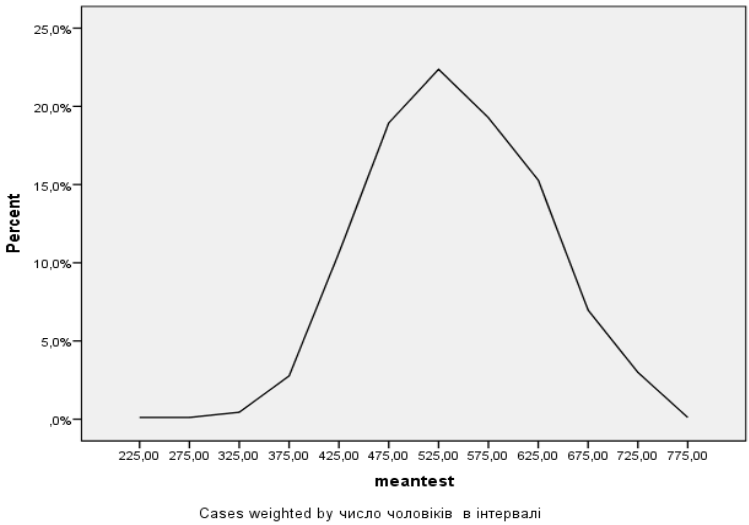


Рис. 1.10.12. Розподіл жінок за тестом

Вимоги до звіту:

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості з теми, відповіді на контрольні питання, файл з результатами.

Контрольні питання

1. Як обчислюється середнє значення для згрупованих даних в SPSS?
2. Як розрахувати значення нової змінної – середнього в інтервалі?
3. Для чого ми зважуємо дані?
4. Дайте змістовну інтерпретацію результатів розрахунків в даній роботі.

Практична робота 1.11
Обчислення моди, медіани, децилей для згрупованих даних

Мета роботи: закріпити вміння розраховувати частоти, будувати полігони, гістограму, криву накопичених частот, навчитися обчислювати моду, медіану та децилі для згрупованих даних, а також коефіцієнт децильної диференціації.

Короткі теоретичні відомості

Для обчислення медіани та децилів згрупованих даних можна скористатися формулою для обчислення процентиля. Рк (К-й процентиль) – це точка на числовій осі, нижче якої лежить К% спостережень.

Щоб **обчислити процентиль** необхідно наступне:

1. Побудувати криву накопичених частот.
2. По ній визначити інтервал, в який потрапляє процентиль.
3. Визначити нижню межу цього інтервалу, ширину інтервалу, накопичені частоти на нижній і верхній межі інтервалу.
4. Скористатися формулою

$$P_k = X_l + I_l \circ \frac{k - f_{l-1}}{f_l - f_{l-1}}$$, де

P_k –	К-Значення процентиля
X_l –	нижня межа інтервалу, куди потрапляє процентиль,
I_l –	ширина інтервалу,
k	К значення
f_{l-1}–	накопичені частоти на нижній межі інтервалу,
f –	накопичені частоти на верхній межі інтервалу

Для медіани k=50, для першого децилю – 10, для останнього децилю – 90.

Для розрахунків моди згрупованих даних скористайтеся формулою:

$$Mo = X_i + I_i \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}$$

Де:

Mo	мода
X_l	нижня межа інтервалу, куди потрапляє мода, тобто такого, що має найбільшу частку
I_l	ширина інтервалу
f_{l-1}	частоти в перед модальному інтервалі,
f_i	частоти в модальному інтервалі
f_{i+1}	частоти в після модальному інтервалі

Постановка задачі

У таблиці 1.11.1 наведено розподіл населення за рівнем середньодушових еквівалентних загальних доходів за 2014 та 2015 рік за даними Держкомстату України.

Порівняйте ці дані, обчисливши середнє, моду, медіану, 1-й та 9-й дециль (використовуйте формули для згрупованих або концентрованих даних).

Таблиця 1.11.1

Розподіл населення за рівнем середньодушових еквівалентних загальних доходів

Середньодушові загальні доходи у місяць, грн	**% домогосподарств, 2014 р.**	**% домогосподарств, 2015 р.**
до 480,0	0,1	0,1
480,1–840,0	1,4	0,9
840,1–1200,0	7,9	4,8
1200,1–1560,0	19,2	11,3
1560,1–1920,0	23,1	18,2
1920,1–2280,03	16,5	17,6
2280,1–2640,0	11,7	15,9
2640,1–3000,0	7,8	11,4
3000,1–3360,0	4,6	6,7
3360,1–3720,0	2,5	4,4
понад 3720,0	5,2	8,7

Держкомстат України

Хід роботи

1. Ввести дані відповідно до наступної таблиці у колонки А-С. Ширина інтервалу у нас 360. Середина першого інтервалу обчислюється як початок інтервалу мінус половина ширини, тобто 480-180=300.

Таблиця 1.11.2

A	B	C	D	E	F	G	H
Середньо-душові загальні доходи у місяць, грн	Домогосподарства, % ni, 2014	Домогосподарства, % ni, 2015	середина інтервалу доходу, грн, xi	Накопичена частота домогосподарств, %, 2014	Накопичена частота домогосподарств, %, 2015	xi* ni, 2014	xi* ni, 2015
до 480,0	0,1	0,1					
480,1–840,0	1,4	0,9					
840,1–1200,0	7,9	4,8					
1200,1–1560,0	19,2	11,3					
1560,1–1920,0	23,1	18,2					
1920,1–2280,0	16,5	17,6					
2280,1–2640,0	11,7	15,9					
2640,1–3000,0	7,8	11,4					
3000,1–3360,0	4,6	6,7					
3360,1–3720,0	2,5	4,4					
понад 3720,0	5,2	8,7					
					середнє		

2. Розрахувати дані в колонках D-H, а також зважене середнє (жовті комірки)

3. Перевірте отримані результати (рис. 1.11.1).

	A	B	C	D	E	F	G	H
1	Середньодушові загальні доходи у місяць, грн	Домо-господарства, % ni, 2014	Домо-господарства, % ni, 2015	середина інтервалу доходу, грн, xi	Накопичена частота домо-господарств, %, 2014	Накопичена частота домо-господарств, %, 2015	xi*ni, 2014	xi*ni, 2015
2	до 480,0	0,1	0,1	300	0,1	0,1	30	30
3	480,1–840,0	1,4	0,9	660	1,5	1	924	594
4	840,1–1200,0	7,9	4,8	1020	9,4	5,8	8058	4896
5	1200,1–1560,0	19,2	11,3	1380	28,6	17,1	26496	15594
6	1560,1–1920,0	23,1	18,2	1740	51,7	35,3	40194	31668
7	1920,1–2280,0[3]	16,5	17,6	2100	68,2	52,9	34650	36960
8	2280,1–2640,0	11,7	15,9	2460	79,9	68,8	28782	39114
9	2640,1–3000,0	7,8	11,4	2820	87,7	80,2	21996	32148
10	3000,1–3360,0	4,6	6,7	3180	92,3	86,9	14628	21306
11	3360,1–3720,0	2,5	4,4	3540	94,8	91,3	8850	15576
12	понад 3720,0	5,2	8,7	3900	100	100	20280	33930
13						середній	2048,88	2318,16
14								

Рис. 1.11.1. Результати розрахунків

4. Побудувати гістограму за накопиченими частотами.

Рис. 1.11.2. Діаграма накопичених частот

5. З'ясувати, в який інтервал потрапили мода, медіана (50-й процентиль), 1-й та 9-ий дециль.

6. Підготувати таблицю 1.11.3

Таблиця 1.11.3

Рік	Міра	X_i	I_i	K	f_{i-1}	f_i	Зна-чення
2014	Медіана	1560,1	360	50	28,6	51,7	
2014	1 дециль	1200,1	360	10	9,4	28,6	
2014	9 дециль	3000,1	360	90	87,7	92,3	

7. Розрахувати значення мір за формулою процентиля:

Таблиця 1.11.3

Обчислення процентилів

К	К-й процентиль
50	1893,61
10	1211,25
90	3180,10

8. Побудувати гістограму для накопичених частот та зробити аналогічні розрахунки для 2015 р.

9. Обчислити моду згрупованих даних:

Таблиця 1.11.4

Обчислення моди

рік	x_i	l_i	f_i	f_{i-1}	f_{i+1}	Мода
2014	1560,1	360	23,1	19,2	16,5	1772,83
2015	1560,1	360	18,2	11,3	17,6	1891,30

10. Порівняти отримані дані в таблиці, розрахувавши децильний коефіцієнт диференціації як відношення 9-го децилю до 1-го.

Перевірити розрахунки (рис. 1.11.3).

Рік	Міра	x_i	l_i	K	f_{i-1}	f_i	Значення міри
2014	медіана	1560,1	360	50	28,6	51,7	1893,61
2014	1 дециль	1200,1	360	10	9,4	28,6	1211,35
2014	9 дециль	3000,1	360	90	87,7	92,3	3180,10
2015	медіана	1920,1	360	50	35,3	52,9	2220,78
2015	1 дециль	1200,1	360	10	5,8	17,1	1333,91
2015	9 дециль	3360,1	360	90	86,9	91,3	3613,74

27			
28	**Показники**	**2014**	**2015**
29	середня	2048,88	2318,16
30	Мода	1772,83	1891,30
31	Медіана	1893,61	2220,78
32	1 деціль	1211,35	1333,91
33	9 деціль	3180,10	3613,74
34	Децільний коефіцієнт диференціації	2,63	2,71
35			

Рис. 1.11.3. Результати розрахунків

Вимоги до звіту

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості по темі, відповіді на контрольні питання, файл з результатами.

Контрольні питання

1. Як обчислюються медіана, децилі згрупованих даних?

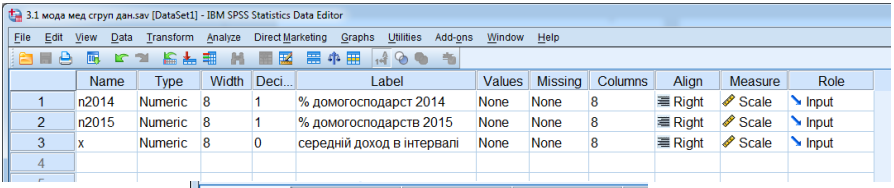
2. Чи можна в цьому завданні скористатися функціями ПЕРСЕНТІЛЬ і Квартиль? Поясніть свою відповідь.

3.Про що свідчить децильний коефіцієнт диференціації.

4. Дайте змістовну інтерпретацію результатів розрахунків в даній роботі.

Рішення задачі в SPSS:

1. Створити файл з 3-ма змінними (рис. 1.11.4):



	Name	Type	Width	Deci...	Label	Values	Missing	Columns	Align	Measure	Role
1	n2014	Numeric	8	1	% домогосподарст 2014	None	None	8	Right	Scale	Input
2	n2015	Numeric	8	1	% домогосподарств 2015	None	None	8	Right	Scale	Input
3	x	Numeric	8	0	середній доход в інтервалі	None	None	8	Right	Scale	Input
4											

	n2014	n2015	x
1	,1	,1	300
2	1,4	,9	660
3	7,9	4,8	1020
4	19,2	11,3	1380
5	23,1	18,2	1740
6	16,5	17,6	2100
7	11,7	15,9	2460
8	7,8	11,4	2820
9	4,6	6,7	3180
10	2,5	4,4	3540
11	5,2	8,7	3900
12			

Рис. 1.11.4. Вихідні дані
Провести розрахунки для 2014 року.

2. Зважити дані змінною n2014 : ***Data, Weight cases***

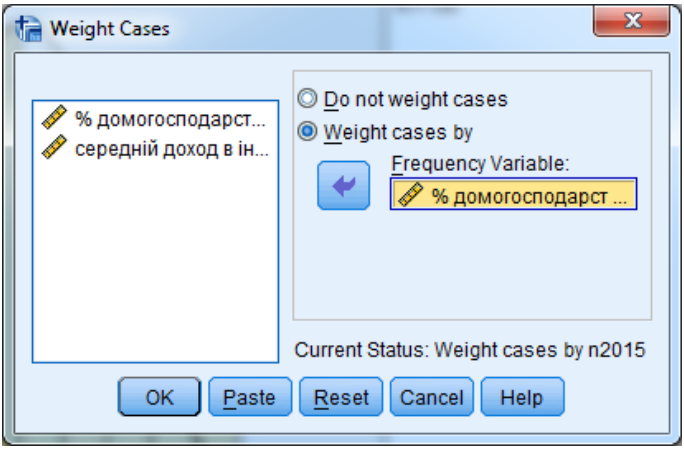


Рис. 1.11.5. Зважування

3. Отримати частотний розподіл**:** ***Analyze, Descriptive statistics, Frequencies***

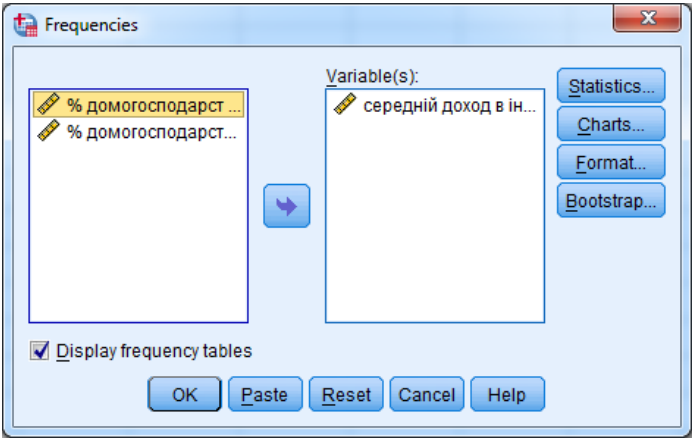


Рис. 1.11.6. Вікно Frequencies

4. Натиснути кнопку ***Statistics,*** та обрати вивід середнього, моди, медіани, та 1-го та 9-го децилю.

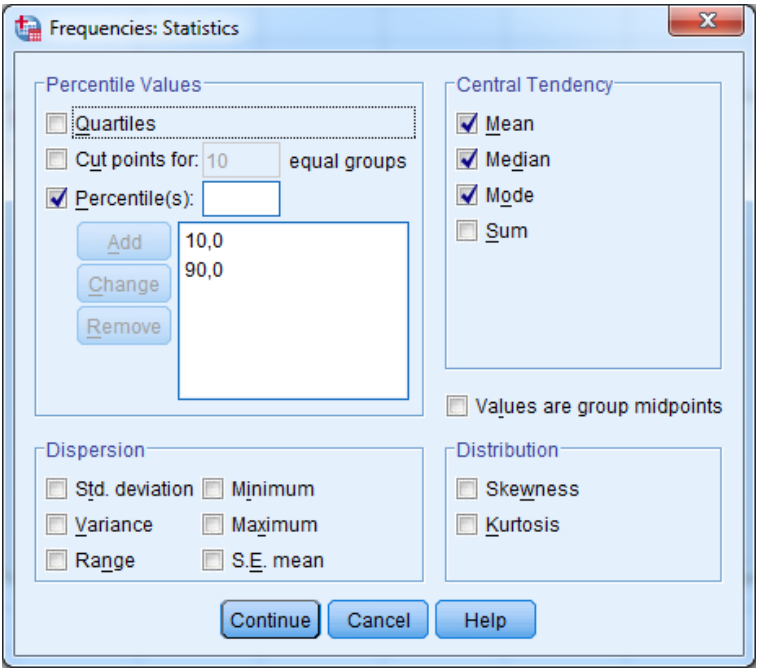


Рис. 1.11.7. Вікно Frequencies:Statistics

Результати розрахунків наступні:

Statistics

середній доход в інтервалі

N	Valid	100
	Missing	0
Mean		2048,88
Median		1740,00
Mode		1740
Percentiles	10	1272,00
	90	3180,00

середній доход в інтервалі

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	300	0	,1	,1	,1
	660	1	1,4	1,4	1,5
	1020	8	7,9	7,9	9,4
	1380	19	19,2	19,2	28,6
	1740	23	23,1	23,1	51,7
	2100	17	16,5	16,5	68,2
	2460	12	11,7	11,7	79,9
	2820	8	7,8	7,8	87,7
	3180	5	4,6	4,6	92,3
	3540	3	2,5	2,5	94,8
	3900	5	5,2	5,2	100,0
	Total	100	100,0	100,0	

Рис. 1.11.8. Середній дохід в інтервалі

5. Побудувати гістограму:

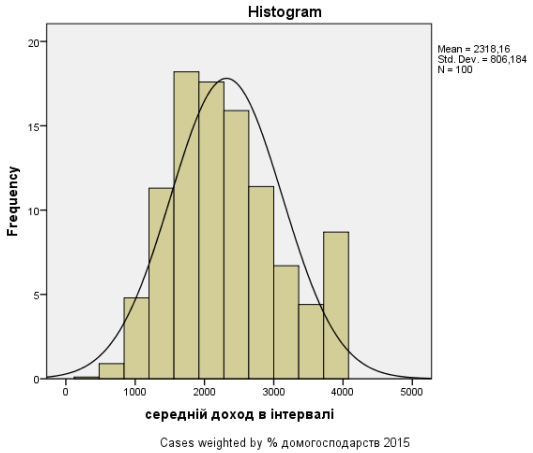


Рис. 1.11.9. Гістограма

6. Аналогічно провести розрахунки для 2015 року.

РОЗДІЛ 2
ВИЯВЛЕННЯ ВІДМІННОСТЕЙ У РОЗПОДІЛІ ОЗНАКИ ЗА ДОПОМОГОЮ КРИТЕРІЮ ПІРСОНА χ^2

Безліч завдань маркетингових, соціологічних, психологічних, політологічних, педагогічних досліджень передбачає ті чи інші зіставлення. Одна з найпоширеніших завдань в дослідженні – зіставлення емпіричного розподілу ознаки з якимнебудь теоретичним законом розподілу (рівномірним, нормальним) або двох емпіричних розподілів між собою (рис.2.1), щоб виявити схожість або відмінності у формі розподілу.

Для порівняння емпіричного розподілу ознаки з теоретичним (нормальним, рівномірним і т.і.) і для порівняння двох, трьох або більше емпіричних розподілів один з одним може використовуватися критерій Пірсона χ2 (вимовляється – "хі-квадрат»), який розраховується за наступною формулою:

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{емпj} - f_{теорj})^2}{f_{теорj}},$$

де $f_{емпj}$ – емпіричні частоти,

$f_{теорj}$ – теоретичні частоти,

k – число розрядів ознаки.

Відповіді на питання "Зрідка я відчуваю почуття раптового страху або невизначеного занепокоєння, сам (а) не знаю, чому"

Варіанти	Чоловіки	Жінки	Разом
Так	6	34	40
І так, і ні	4	3	7
Ні	23	8	71
Разом	33	45	38

Рис. 2.1. Порівняння двох емпіричних розподілів (відповідей чоловіків та жінок на питання анкети)

Отримане емпіричне значення критерію порівнюється з критичним, яке вибирається з таблиць критичних значень (табл. 2.1) в залежності від числа ступенів свободи *df* та рівня значущості *p*.

Якщо $\chi 2$емп < $\chi 2$крит, то розбіжності між розподілами статистично недостовірні (розподіли узгоджуються між собою).

Число ступенів свободи для таблиці критичних значень визначається за формулою:

$df = (k–1) * (l–1)$, де

k – число розрядів ознаки або рядків в таблиці,
l – число стовпців в таблиці.

На перетині числа ступенів свободи та рівня значущості (0,05; 0,01; 0,001) знаходиться критичне значення критерію. Так, для *p=0,05* та *df=3*, отримаємо *χ2крит=7,815* (табл. 2.1).

Таблиця 2.1

Фрагмент таблиці критичних значень критерія Пірсона χ^2

df	**P**		
	0,05	**0,01**	**0,001**
1	3,842	6,635	10,829
2	5,992	9,211	13,817
3	7,815	11,346	16,269
4	9,488	13,278	18,470
5	11,071	15,088	20,519
6	12,593	16,814	22,462
7	14,068	18,478	24,327
8	15,509	20,093	26,130
9	16,921	21,669	27,883
10	18,309	23,213	29,594

Зауваження та обмеження до застосування критерію Пірсона χ^2

- Обсяг вибірки повинен бути досить великим: n> 30.
- Теоретична частота для кожної клітини таблиці повинна бути більше 5.

- Вибрані розряди повинні охоплювати весь діапазон варіативності ознаки. А угруповання на розряди повинно бути однаковим у всіх розподілах.
- Розряди не повинні перехрещуватися, тобто якщо спостереження віднесено до одного розряду, то воно не може бути віднесено до іншого.
- Сума спостережень за розрядами дорівнює загальному числу спостережень.
- Якщо порівнюються два емпіричних розподіли, то теоретичні частоти розраховуються як відношення добутку підсумків за рядком на підсумок за стовпчиком до загальної кількості спостережень.

Практична робота 2.1

Використання критерію Пірсона χ^2 для перевірки узгодженості розподілів за допомогою MS Excel та SPSS

Мета роботи: навчитися представляти концентровані дані для обробки в Excel і SPSS, будувати для них таблиці спряженості та обчислювати критерій Пірсона χ^2.

Постановка задачі

Дослідників в області маркетингу цікавить, чи залежить обсяг придбання модного одягу від сімейного стану покупців різної статі. У таблицях 2.1.1 і 2.1.2 наведені дані для чоловіків і жінок відповідно [16, с.572].

Таблиця 2.1.1

Придбання модного одягу чоловіками

Купує	Сімейний стан	
	Одружений	Неодружений
Багато	32	12
Мало	70	19
Практично не купує	5	5
Разом	107	36

Таблиця 2.1.2

Придбання модного одягу жінками

Купує	Сімейний стан	
	Заміжня	Незаміжня
Багато	24	31
Мало	75	20
Практично не купує	7	5
Разом	106	56

За допомогою критерію хі-квадрат визначить, чи є зв'язок між сімейним станом і обсягом придбання модного одягу для чоловіків і жінок.

Рішення за допомогою електронних таблиць Excel

Хід роботи

Алгоритм розрахунку χ^2 за допомогою MS Excel

1. Занести в таблицю найменування розрядів ознаки (стовпець А)
2. Занести відповідні їм емпіричні частоти для одружених і неодружених чоловіків (стовпці В та С).
3. Підрахувати підсумки по рядках і стовпцях (стовбець D).
4. Обчислити теоретичні частоти для одружених і неодружених чоловіків як відношення добутку підсумків по рядку і підсумків по стовбчику до загальної кількості об'єктів (стовбці Е і F).
5. Для розрахунку χ2емп скористатися функціями ХИ2ТЕСТ і ХИ2ОБР. Функції ХИ2ТЕСТ передаються два папараметра: інтервал з емпіричними частотами та інтервал з теоретичними частотами. Функції ХИ2ОБР передаються два параметри: результат, отриманий в ХИ2ТЕСТ, і число ступенів свободи, що визначається за формулою: $\nu = (k–1) * (l–1)$, де k – число розрядів ознаки або рядків в таблиці (для нашого прикладу k = 3), l – число стовпців в таблиці (для нашого прикладу 2).
6. За таблицями критичних точок розподілу χ2 для даного числа ступенів свободи визначити χ2критичне.

Якщо χ2емп < χ2крит, то розбіжності між розподілами статистично недостовірні (розподіли узгоджуються між собою), а, отже, одружені і неодружені чоловіки не розрізняються за рівнем придбання модного одягу.

Формули для розрахунків наведені в таблиці 2.1.3, результати розрахунків в таблиці 2.1.4.

Таблиця 2.1.3

Таблиця з формулами

	A	B	C	D	E	F
1	**Придбання модного одягу чоловіками**					
2	Купує	Емпірична частота		Сума	Теоретична частота	
3		Сімейний стан			Сімейний стан	
4		Одружений	Неодружений		Одружений	Неодружений
5	Багато	32	12	=СУММ(B5:C5)	=$D5*B$8/$D$8	=$D5*C$8/$D$8
6	Мало	70	19	=СУММ(B6:C6)	=$D6*B$8/$D$8	=$D6*C$8/$D$8
7	Практично не купує	5	5	=СУММ(B7:C7)	=$D7*B$8/$D$8	=$D7*C$8/$D$8
8	Разом	107	36	=СУММ(D5:D7)	=СУММ(E5:E7)	=СУММ(F5:F7)
9						
10						
11			=ХИ2ТЕСТ(B5:C7;E5:F7)			
12		**χ2емп**	=ХИ2ОБР(C11;2)			

Таблиця 2.1.4

Таблиця результатів:

	A	B	C	D	E	F
1	**Придбання модного одягу чоловіками**					
2	Купує	Емпірична частота		Сума	Теоретична частота	
3		Сімейний стан			Сімейний стан	
4		Одружений	Неодружений		Одружений	Неодружений
5	Багато	32	12	44	32,92	11,08
6	Мало	70	19	89	66,59	22,41
7	Практично не купує	5	5	10	7,48	2,52
8	Разом	107	36	143	107	36
9						
10						
11			0,131			
12		**χ2емп**	4,066			

Аналогічно вирішить задачу про придбання модного одягу жінками.

Рішення задачі в SPSS

Хід роботи

1. Визначте в редакторі 3 змінні:

- Покупка с градаціями (1 – багато, 2 – мало, 3 – практично не купує);
- Сімейний стан (1 – одружений, 2 – неодружений);
- freq (частота) – указує частоту кожного поєднання.

2. Увести дані для змінних згідно наступної таблиці.

Таблиця 2.1.5

Вихідні дані для чоловіків

Покупка	Сімейний стан	freq
Багато	Одружений	32
Мало	Одружений	70
Практично не купує	Одружений	5
Багато	Не одружений	12
Мало	Не одружений	19
Практично не купує	Не одружений	5

3. Зважте дані, меню

Data (Дані)

Weight Cases (Зважити випадки).

У діалоговому вікні ***Weight Cases*** (рис. 2.1.1) обрати опцію ***Weight Cases by*** та перенести змінну freq в поле ***Frequency Variable***. Закрити діалогове вікно.

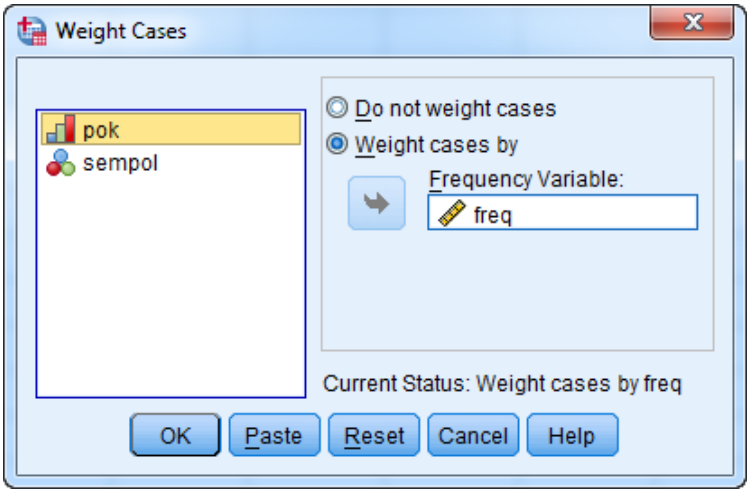


Рис. 2.1.1. Діалогове вікно «*Weight Cases*»

4. Побудувати таблиці спряженості, обравши меню ***Analyze*** (Аналіз), ***Descriptive statistics*** (Описова статистика), ***Crosstabs*** (Таблиці спряженості) (рис. 2.1.2):

- Перенести змінну «**Покупка**» в список змінних рядків.
- Змінну «**Сімейний стан**» – у список змінних стовбців.

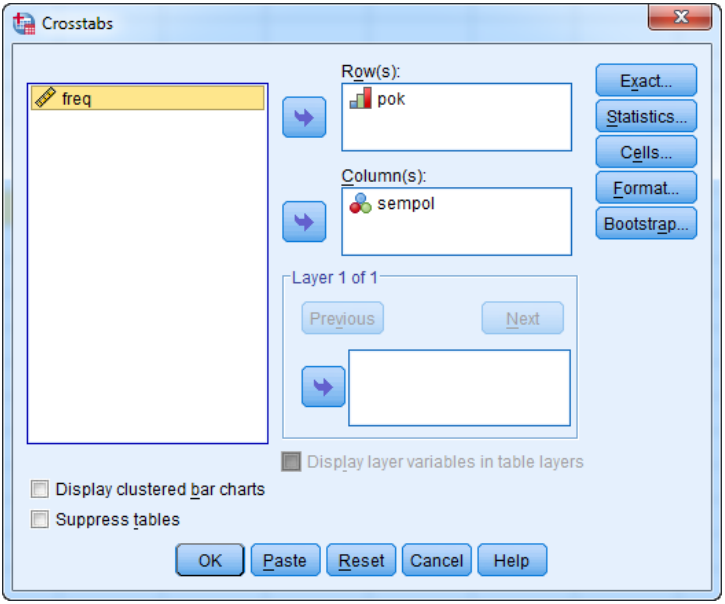


Рис. 2.1.2. Діалогове вікно «*Crosstabs*»

- Натиснуть кнопку ***Statistics***…
- У діалоговому вікні ***Crosstabs: Statistics*** … (Частоти: Статистика) обрати опцию ***Chi*-square** та натиснуть кнопку ***Continue*** (рис. 2.1.3).

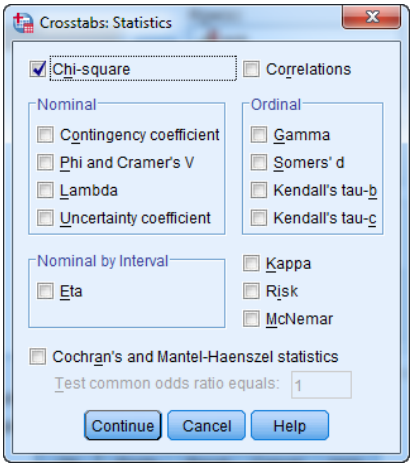


Рис. 2.1.3. Діалогове вікно «*Crosstabs: Statistics*»

- Натиснуть кнопку ***Cell*** (Ячейка). В групі ***Counts*** установити опцію для спостерігаємих та очікуваних частот. В групі ***Percentage*** поставити галочку біля опції ***Column*** (Відсоток за колонкою). Натиснути **Continue, OK** (рис. 2.1.4).

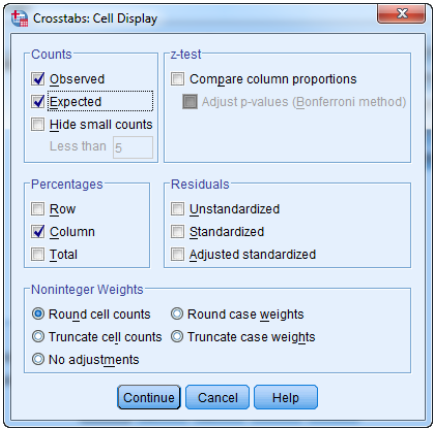


Рис. 2.1.4. Діалогове вікно «*Crosstabs: Cell Display*»

5. Розглянути та проаналізувати результати (табл. 2.1.6 – 2.1.7).

Таблиця 2.1.6

Купує * Сімейний стан Crosstabulation

			Сімейний стан		Total
			одружений	Не одружений	
Купує	Багато	Count	32	12	44
		Expected Count	32,9	11,1	44,0
		% within Сімейний стан	29,9%	33,3%	30,8%
	Мало	Count	70	19	89
		Expected Count	66,6	22,4	89,0
		% within Сімейний стан	65,4%	52,8%	62,2%
	Практично не купує	Count	5	5	10
		Expected Count	7,5	2,5	10,0
		% within Сімейний стан	4,7%	13,9%	7,0%
Total		Count	107	36	143
		Expected Count	107,0	36,0	143,0
		% within Сімейний стан	100,0%	100,0%	100,0%

Таблиця 2.1.7

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,066[a]	2	,131
Likelihood Ratio	3,649	2	,161
Linear-by-Linear Association	,279	1	,597
N of Valid Cases	143		

a. 1 cells (16.7%) have xpected count less than 5.

Аналогічно вирішити завдання про придбання модного одягу жінками.

Задачі для самостійного рішення

1. За даними таблиці 2.1.8 визначте, чи існує залежність між віком клієнта та його витратами на літній відпочинок.

Таблиця 2.1.8

Залежність між віком клієнта та його витратами на літній відпочинок

Вік	Рівень витрат на відпочинок			Разом
	Високий	Середній	Низький	
Молоді	25	40	31	
Середнього віку	52	63	47	
Старшого віку	12	23	58	
Разом				

2. За даними таблиці 2.1.9 перевірити припущення про те, що частота звернень в телефонну службу довіри нерівномірно розподіляється по днях тижня (за допомогою критерію порівняти заданий емпіричний розподіл з рівномірним).

Таблиця 2.1.9

Розподіл звернень в телефонну службу довіри за днями тижня

День тижня	Число звернень
Понеділок	18
Вівторок	12
Середа	14
Четвер	12
П’ятниця	18
Субота	30
Неділя	36
Разом	140

3. За даними наступної таблиці з’ясувати, чи є зв'язок між партійною приналежністю батька та сина (табл. 2.1.10).

Таблиця 2.1.10

Визначення партійності на основі партійності батька

Партійність батька	Партійність респондента			
	Демо-крат	Республіка-нець	Незалежний	Разом
Демократ	30	18	12	60
Республіка-нець	15	9	6	30
Незалежний	5	3	2	30
Разом	50	30	20	100

Вимоги до звіту

Звіт повинен містити:

– відповіді на контрольні питання;

– файл з результатами розрахунків.

Контрольні питання

1. Призначення критерію Пірсона $\chi 2$.
2. Обмеження критерію Пірсона $\chi 2$.
3. Що таке емпіричні частоти?
4. Що таке теоретичні частоти?
5. Як обчислюється критерій Пірсона $\chi 2$ в Excel?
6. Як подати дані для розрахунків критерію Пірсона $\chi 2$ за допомогою SPSS?
7. Дайте змістовну інтерпретацію результатів, отриманих при вирішенні задач.

Практична робота 2.2

Критерій Пірсона хі-квадрат для концентрованих даних в R

Хід роботи

Спробуємо за допомогою R вирішити задачу про порівняння двох емпіричних розподілів з попередньої практичної роботи.

1. Створюємо три вектори з вихідними даними :

```
 >a<-c(1,2,3,1,2,3)
> b<-c(1,1,1,2,2,2)
> c<-c(32,70,5,12,19,5)
```

Тут для вектору **a** цифрами закодовано: 1 багато, 2 – мало, 3 – практично не купує; для вектору **b**: 1 – одружений, 2 – не одружений. Вектор **c** представляє стовпець частот для зважування даних.

2. Об'єднуємо їх в загальну таблицю даних за допомогою команди **data.frame**:

```
 >data2<-data.frame(pokupka=a, sempol=b, freq=c)
```

3. Перевіряємо зміст таблиці даних data2

```
>data2
  pokupka sempol freq
1       1      1   32
2       2      1   70
3       3      1    5
4       1      2   12
5       2      2   19
6       3      2    5
```

Оскільки дані концентровані, відновлюємо з них неконцентрований масив шляхом дублювання:

```
>data2.x <- data2[rep(1:nrow(data2),data2$freq),]
```

4. Команда xtab дозволяє представити ці дані у вигляді кростаблиці:

```
>xtabs(freq~pokupka+sempol, data2)
       sempol
pokupka  1  2
      1 32 12
      2 70 19
      3  5  5
```

Рис. 2.2.1. Кростаблиця «покупка*сімейний стан»

5. Для отримання кростаблиці в стилі SPSS можна використувати пакет gmodels і команду CrossTable

```
> library(gmodels)
>CrossTable(data2.x$pokupka, data2.x$sempol,chisq = TRUE)
```

В результаті отримаємо таблицю (рис. 2.2.2).

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  143


                 | data2.x$sempol
data2.x$pokupka |         1 |         2 | Row Total |
----------------|-----------|-----------|-----------|
              1 |        32 |        12 |        44 |
                |     0.026 |     0.077 |           |
                |     0.727 |     0.273 |     0.308 |
                |     0.299 |     0.333 |           |
                |     0.224 |     0.084 |           |
----------------|-----------|-----------|-----------|
              2 |        70 |        19 |        89 |
                |     0.174 |     0.518 |           |
                |     0.787 |     0.213 |     0.622 |
                |     0.654 |     0.528 |           |
                |     0.490 |     0.133 |           |
----------------|-----------|-----------|-----------|
              3 |         5 |         5 |        10 |
                |     0.824 |     2.448 |           |
                |     0.500 |     0.500 |     0.070 |
                |     0.047 |     0.139 |           |
                |     0.035 |     0.035 |           |
----------------|-----------|-----------|-----------|
   Column Total |       107 |        36 |       143 |
                |     0.748 |     0.252 |           |
----------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  4.066283     d.f. =  2     p =  0.1309236
```

Рис. 2.2.2. Результати кростабуляції та обчислення критерію Пірсона хі-квадрат для концентрованих даних

Таким чином, ми отримали емпіричне значення критерію Пірсона хі-квадрат, рівне 4,06, що узгоджується з розрахунками, проведеними нами в SPSS та Excel в п.2.

Контрольні питання

1. Як самим створити таблицю даних для поточного прикладу?
2. Як відновити з цих даних неконцентрований масив?
3. Поясніть, що робить наступна команда?

xtabs (freq ~ pokupka + sempol, data2)

4. У якому пакеті знаходиться команда CrossTable?
5. За рис. 2.2.2 поясніть, які дані виводяться в результаті роботи цієї команди?
6. Вирішіть аналогічне завдання про жінок в середовищі R (див. п. 2).

Література

1. https://stat.ethz.ch/pipermail/r-help/2008-November/180925.html
2. http://statmethods.net/stats/frequencies.html

РОЗДІЛ 3. КОРЕЛЯЦІЯ

Види залежностей:		
причинно-наслідкові	прямі	функціональні
не причинно-наслідкові	Зворотні	кореляційні

Функціональна залежність має місце, якщо кожному значенню однієї змінної (X) відповідає цілком певне значення іншої змінної (Y).

Кореляційна залежність має місце, якщо певному значенню однієї змінної відповідає декілька значень іншої змінної.

Або, іншими словами, **кореляційна залежність** – це залежність, при якій певному значенню однієї величини відповідає цілий комплекс значень іншої, що представляє собою ряд розподілу, причому при зміні даної величини змінюється ряд розподілу та його середнє.

Коефіцієнти кореляції можуть набувати значень від –1 до +1.

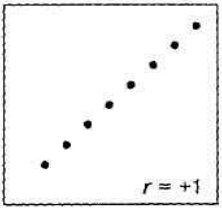
На рис. 3.1. представлено схематичне зображення сили і напряму кореляції.

Існують два підходи до оцінки сили кореляції. Перший (див. табл. 3.1) орієнтований тільки на абсолютну величину коефіцієнта кореляції, а другий – на рівень значущості даного коефіцієнта кореляції при даному обсязі вибірки (див. таб. 3.2).

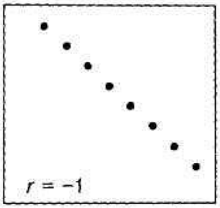
Таблиця 3.1

Оцінка сили кореляції за абсолютною величиною коефіцієнту кореляції

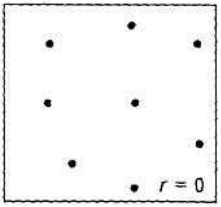
Абс. величина коефіцієнту кореляції |r|	**Оцінка сили кореляції**
$\|r\| \geq 0,70$	сильна (тісна)
$0,50 \leq \|r\| \leq 0,69$	середня
$0,30 \leq \|r\| \leq 0,49$	помірна
$0,20 \leq \|r\| \leq 0,29$	слабка
$\|r\| \leq 0,19$	дуже слабка



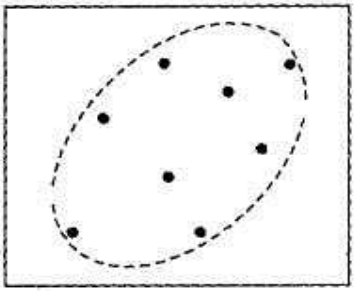
повна позитивна



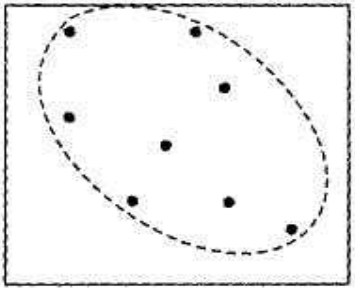
повна негативна



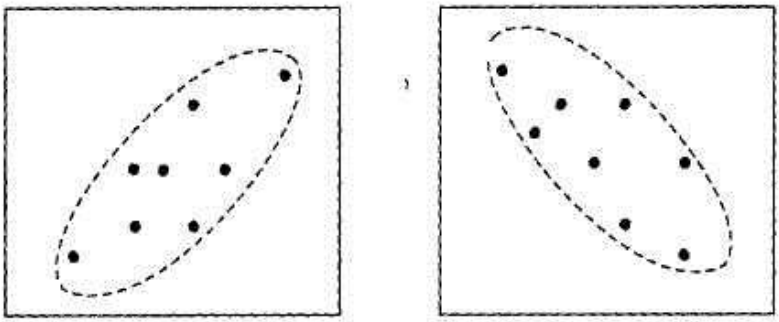
немає кореляції



слабка позитивна



слабка негативна



сильна позитивна сильна негативна

Рис. 3.1. Схематичне зображення сили та напрямку кореляції

Згідно другого підходу (див. табл. 3.2), чим більший обсяг вибірки, на якій вивчалася кореляція між змінними, тим менший за абсолютною величиною коефіцієнт кореляції визнається показником достовірності кореляційної зв'язку.

Таблиця 3.2

Оцінка сили кореляції за рівнем статистичної значущості

Рівень статистичної значущості	Оцінка сили кореляції
$p \leq 0{,}01$	високо значуща кореляція
$0{,}01 < p \leq 0{,}05$	значуща кореляція
$0{,}05 < p \leq 0{,}10$	тенденція достовірного зв’язку
$p > 0{,}10$	Незначуща кореляція

Найбільш доцільним є поєднання двох підходів, коли при формулюванні висновків про наявність чи відсутність ко-

реляції між змінними, враховується й абсолютна величина коефіцієнта кореляції, й рівень значущості цього коефіцієнта для даного обсягу вибірки.

Причинність і кореляція

Наявність кореляції між змінними аж ніяк не означає, що між ними існує причинно-наслідковий зв'язок. Можливо, або одна з змінних є частковою причиною іншої, або обидві вони є наслідком якихось загальних причин. Наявність значної кореляції між змінними дає нам підставу не відкидати гіпотезу про наявність причинно-наслідкового зв'язку між змінними, але не може служити підтвердженням такої гіпотези.

Вивчаючи кореляційну залежність між змінними, дослідник, інтерпретуючи значення коефіцієнтів кореляції, повинен враховувати сутність, природу цих змінних. Вибір коефіцієнта залежить від типу даних, які обробляються. Широко розповсюдженими в маркетингових, соціологічних, психолого-педагогічних та інших дослідженнях є коефіцієнти асоціації і контингенції, коефіцієнт Пірсона $\chi 2$, коефіцієнт рангової кореляції Спірмена, коефіцієнт рангової кореляції Кенделла, коефіцієнт Пірсона-Браве для неперервних шкал.

Практична робота 3.1
Обчислення коефіцієнта кореляції Пірсона-Браве для метричних шкал

Мета роботи: навчитися обчислювати коефіцієнт кореляції Пірсона і будувати діаграму розсіювання, визначати значимість коефіцієнта

Короткі теоретичні відомості

Для визначення кореляційної залежності між двома змінними, заданими метричними шкалами, служить коефіцієнт Пірсона-Браве. Він характеризує міру лінійного зв'язку між змінними.

Нехай X і Y дві змінних, які корелюють між собою. Коефіцієнт Пірсона-Браве обчислюється за формулою

$$r = \frac{\sum_{i=1}^{n} Z_i^x \circ Z_i^y}{n-1} \qquad (3.1.1)$$

де

$$Z_i^x = \frac{X_i - X_{cp}}{S_x} \qquad (3.1.2)$$

S_x - стандартне відхилення,
X_i - i-значення ознаки,
X_{cp} - середнє значення ознаки Y

$$Z_i^y = \frac{Y_i - Y_{cp}}{S_y}, \qquad (3.1.3)$$

S_y - стандартне відхилення,
Y_i - i-значение,
Y_{cp} - середнє значення ознаки Y

Постановка задачі

У таблиці 3.1.1 наведені дані про кількість машин, які щодня паркуються на відкритій стоянці та у гаражі біля університету. Знайдіть та інтерпретуйте кореляцію між цими двома величинами.

Таблиця 3.1.1

Кількість автомобілів, які паркувалися біля університету

День тижня	Відкрита стоянка	Гараж
Понеділок	140	180
Вівторок	120	200
Середа	130	190
Четвер	110	210
П'ятниця	160	160
Субота	135	185

Рішення за допомогою Excel

Хід роботи

1. Увести дані.

2. Обчислити коефіцієнт Пірсона для цих даних спочатку за формулами, а потім виконати те ж за допомогою функції КОРРЕЛ (рис. 3.1.1.).

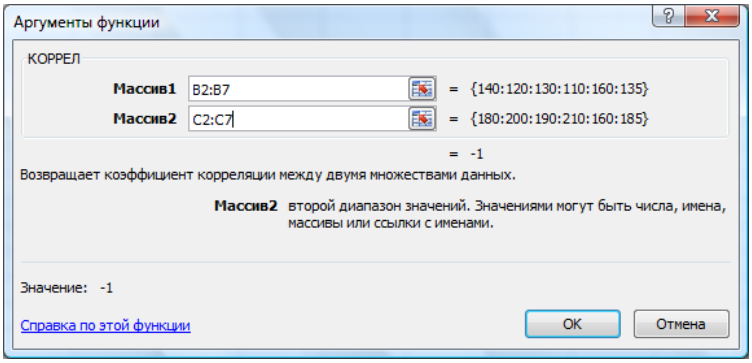


Рис. 3.1.1. Діалогове вікно «Аргументи функції КОРРЕЛ»

3. Заповнити в зошиті таблицю 3.1.2 (внести розрахункове і критичне значення коефіцієнта, визначити його значущість по таблиці критичних значень).

Таблиця 3.1.2

Обсяг вибірки	Емпіричне значення коефіцієнта	Критичне значення	Резюме (значимий/ незначимий)

4. Побудувати діаграму розсіювання (виділити дані, Вставка, Диаграмма, обрати тип диаграми – «точечна», підписати всі елементи діаграми) (см. рис. 3.1.2.).

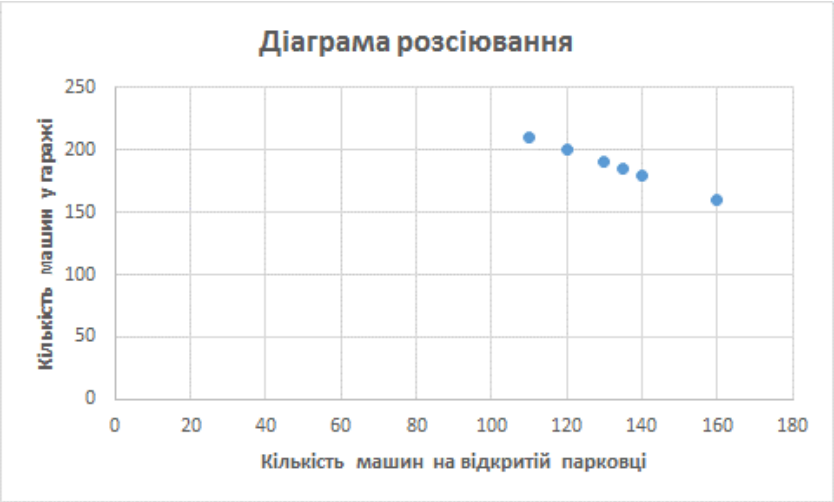


Рис. 3.1.2. Діаграма розсіювання, побудована за даними таблиці 3.1.1.

Рішення за допомогою SPSS
Хід роботи

1. Створити файл даних в SPSS: описати дві змінні (Open – число машин на відкритій стоянці, Garag – число машин в гаражі).
2. Ввести або скопіювати з файлу Excel дані.
3. Обчислити коефіцієнт Пірсона для цих даних: Меню: ***Analyze*** (Аналіз), ***Correlate*** (Кореляція), ***Bivariate*** (Двовимірна).
4. У вікні ***Bivariate Correlation*** (рис. 3.1.3) задати необхідні параметри.

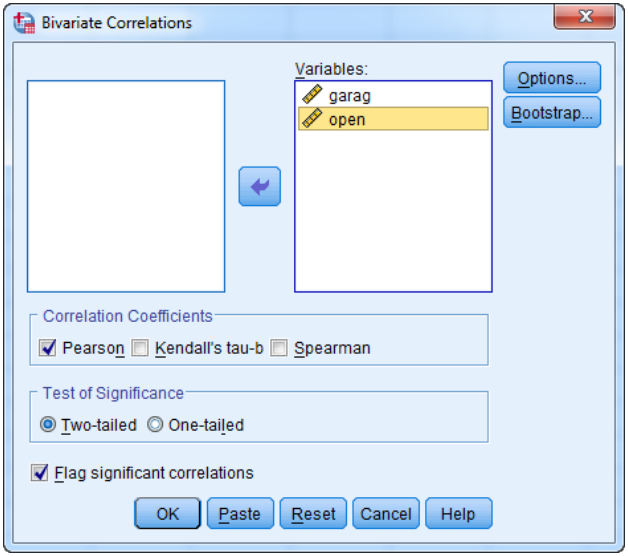


Рис. 3.1.3. Діалогове вікно «Bivariate Correlation»

5. Ознайомитися з отриманими результатами у вікні перегляду (рис. 3.1.4). Результати включають: коефіцієнт кореляції (Pearson Correlation), рівень значущості (Sig), обсяг вибірки (N). Записати їх в зошит. Визначити, чи значущий коефіцієнт кореляції.

Correlations

		garag	open
garag	Pearson Correlation	1	-1,000**
	Sig. (2-tailed)		,000
	N	6	6
open	Pearson Correlation	-1,000**	1
	Sig. (2-tailed)	,000	
	N	6	6

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 3.1.4. Вікно результатів розрахунків для прикладу з парковкою

6. Побудувати діаграму розсіювання: Меню: ***Graphs*** (Графики), ***Scatter…***(Диаграмма рассеивания), ***Simple*** (Простая) (рис. 3.1.5).

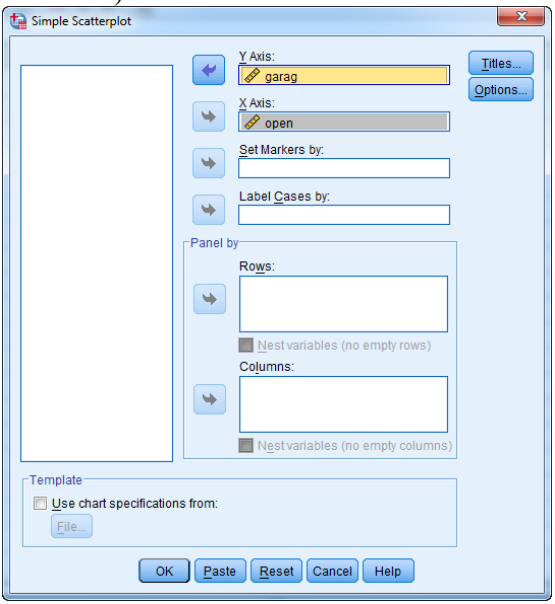


Рис. 3.1.5. Діалогове вікно «Simple Scatterplot»

Переглянути отриману діаграму розсіювання (рис. 3.1.6).

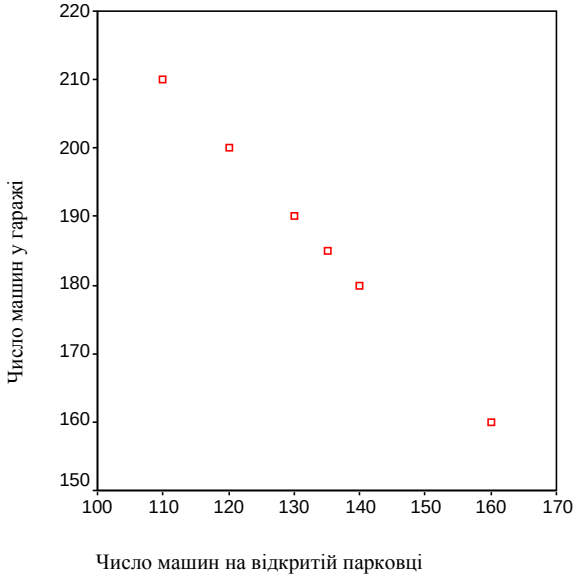


Рис. 3.1.6. Точечна діаграма для прикладу з парковкою

7. Зберегти дані розрахунків і вихідні дані в файлах.

8. Зробити висновок про наявність чи відсутність значимої кореляції між двома ознаками.

Завдання для самостійної роботи

1. У таблиці 3.1.3 [7] містяться дані про щоденний обсяг продажів в доларах лазерних принтерів, картриджів для принтерів і шкільного приладдя. Знайдіть та інтерпретуйте кореляцію між цими величинами. Вирішіть задачу в середовищі Microsoft Excel та в середовищі SPSS.

Таблиця 3.1.3

Щоденний обсяг продажів ($)

День	Лазерні принтери	Картриджі	Шкільні приналежності
1	17291	4379	3618
2	13734	2258	3514
3	18802	4206	3587
4	12171	2137	3007
5	16402	3775	3850
6	19904	4781	3675
7	14023	1991	3120
8	17513	2663	3345
9	17847	4451	2045
10	12718	1648	3292
11	12292	2342	3405
12	11846	2646	2799
13	17088	3216	2417
14	13523	2184	2405

***Вимоги* до звіту**

Звіт повинен містити:

– відповіді на контрольні питання;

– файл з результатами розрахунків.

Контрольні питання

1. Дайте визначення функціональної залежності.
2. Дайте визначення кореляційної залежності.
3. Які значення можуть приймати коефіцієнти кореляції?

4. Як обчислити коефіцієнт кореляції Пірсона-Браве за допомогою електронних таблиць?

5. Для яких шкал він використовується?

6. Як побудувати діаграму розсіювання?

7. Які параметри задаються функції КОРРЕЛ?

8. Як розрахувати коефіцієнт кореляції Пірсона-Браві і побудувати діаграму розсіювання в SPSS?

9. Як визначити значимість коефіцієнта кореляції в SPSS?

Практична робота 3.2
Кореляція за Пірсоном за допомогою R

Хід роботи

Відкрийте R STUDIO. Вирішимо в середовищі R дві задачі з попередньої практичної роботи.

Завдання 1.

1. Визначимо два вектори, що представляють число автомашин, припаркованих на відкритій стоянці та в гаражі біля університету:

```
> open <- c(140, 120,130,110,160,135)
> garag <- c(180, 200,190,210,160,185)
```

2. Підрахуємо кореляцію за Пірсоном:

```
>  cor.test(open,garag)

      Pearson's product-moment correlation

data:  open and garag
t = NaN, df = 4, p-value = NA
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 NaN NaN
sample estimates:
cor
 -1

Warning messages:
1: In sqrt(1 - r^2) : NaNs produced
2: In atanh(r) : NaNs produced
```

3. Побудуємо діаграму розсіювання:

```
>plot(open, garag,  main="Диаграма розсіювання", xlab="число машин на відкритій стоянці",  ylab="число машин у гаражі")
```

Рис. 3.2.1.Діаграма розсіювання, побудова в R

Завдання 2.

1. Уведіть дані в Excel (рис.3.2.1) та збережіть з ім'ям printer.csv в поточному каталозі.

	A	B	C	D
1	День	Принтери	Катриджі	Канцтовари
2	1	17291	4379	3618
3	2	13734	2258	3514
4	3	18802	4206	3587
5	4	12171	2137	3007
6	5	16402	3775	3850
7	6	19904	4781	3675
8	7	14023	1991	3120
9	8	17513	2663	3345
10	9	17847	4451	2045
11	10	12718	1648	3292
12	11	12292	2342	3405
13	12	11846	2646	2799
14	13	17088	3216	2417
15	14	13523	2184	2405
16				

printer

Рис. 3.2.2. Вихідні дані

2. Збережіть дані в датафрейм data4.

```
>data4<-read.table("K:\\Rexample\\printer.csv", sep=";", dec=",", header=TRUE)
```

```
> data4
   День Принтери Катриджі Канцтовари
1     1    17291     4379       3618
2     2    13734     2258       3514
3     3    18802     4206       3587
4     4    12171     2137       3007
5     5    16402     3775       3850
6     6    19904     4781       3675
7     7    14023     1991       3120
8     8    17513     2663       3345
9     9    17847     4451       2045
10   10    12718     1648       3292
11   11    12292     2342       3405
12   12    11846     2646       2799
13   13    17088     3216       2417
14   14    13523     2184       2405
```

3. Обчисліть коефіцієнт кореляції між продажами принтерів та картриджів:

```
>cor.test(data4$Принтери,data4$Катриджі)

        Pearson's product-moment correlation

data:  data4$Принтери and data4$Катриджі
t = 5.8127, df = 12, p-value = 8.306e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6035144 0.9545418
sample estimates:
      cor
0.8590239
```

У нашому випадку отримана значуща кореляція продаж принтерів та картриджів, коефіцієнт кореляції 0.8590239. Проведіть кореляційний аналіз для всій матриці даних:

```
>cor(data4)
              День        Принтери   Катриджі  Канцтовари
День        1.0000000  -0.3182177  -0.3589594 -0.6655984
Принтери   -0.3182177   1.0000000   0.8590239  0.1679915
Катриджі   -0.3589594   0.8590239   1.0000000  0.1426419
Канцтовари -0.6655984   0.1679915   0.1426419  1.0000000
```

Можна скористатися пакетом ellipse, щоб візуалізувати кореляційну матрицю. При цьому кореляційний зв'язок представляється у вигляді еліпсів. Чим ближче коефіцієнт кореляції до +1 або –1, тим вужчим стає еліпс. Скористаємося функцією plotcorr.

```
>install.packages(pkgs=c("ellipse"))
```

4. Побудуйте графік, що ілюструє результати кореляційного аналізу (см. рис.3.2.3)

```
> library(ellipse)
> plotcorr(cor(data4))
```

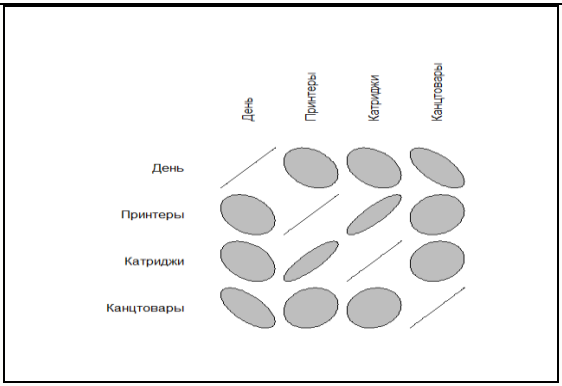


Рис. 3.2.3. Результати кореляційного аналізу

Контрольні питання

1. Як підготувати дані для кореляційного аналізу в R?
2. Як здійснюється кореляційний аналіз для двох змінних?
3. Як будується кореляційна матриця кількох змінних?
4. Як побудувати діаграми розсіювання для пари змінних?
5. Як можна візуалізувати дані кореляційного аналізу?
6. Охарактеризуйте малюнок 3.2.3.
7. Заповніть наступну таблицю за результатами розрахунків:

Кореляція між	Коефіцієнт кореляції Пірсона Пірсона	p-значення	Висновок (кореляція значима/ незначима)
Принтерами та картриджами			
Картриджами та канцтоварами			
Канцтоварами та принтерами			

Практична робота 3.3
Обчислення коефіцієнта рангової кореляції Спірмена

Мета роботи: навчитися обчислювати коефіцієнт рангової кореляції Спірмена, визначати його значущість, інтерпретувати результати.

Короткі теоретичні відомості

Припустимо *N* об'єктів можна упорядити як за ознакою *X*, так і за ознакою *Y*. Означимо

R_i^x – ранг *i*-об'єкта за ознакою *X*

R_i^y – ранг *i*-об'єкта за ознакою *Y*

$d_i = R_i^x - R_i^y$ – міра неспівпадання рангів

Для визначення кореляційної залежності між двома змінними, заданими ранговими шкалами служить коефіцієнт рангової кореляції Спірмена, який обчислюється за формулою:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} di^2}{N(N^2-1)} \qquad (3.3.1)$$

Постановка задачі

Дослідження, проведене на британському ринку, виявило наступну значимість чинників успіху нових продуктів, вироблених японськими і британськими фірмами (табл. 3.3.1, в якій n означає число опитаних фірм).

З таблиці 3.3.1 видно, що найбільш часто ключовим фактором успіху називали ступінь адаптації до вимог покупців. І тільки відносно невелика частка фірм (1/4), як японських, так і британських, вважає ефективний маркетинг важливим критерієм успіху.

Порівняння оцінок, отриманих від японських і британських фірм, виявляє, що перші з них надають більше значення наявності конкурентної переваги, а другі – більшій адаптації до запитів споживачів.

Таблиця 3.3.1

Фактори успіху нових продуктів, виявлені на британському ринку

Фактори успіху	Процент фірм	
	японських (n=116)	британських (n=86)
Хороша адаптація до потреб	69,8	75,6
Переваги над конкурентами щодо якості	79,3	59,3
Переваги над конкурентами щодо дизайну	69,9	45,3
Переваги над конкурентами щодо співвідношення якість / ціна	58,6	61,6
Переваги над конкурентами щодо конструкції	55,2	48,8
Вельми конкурентна ціна	41,4	27,9
Адаптація	39,7	34,9
Унікальність	36,2	29,1
Ефективний маркетинг	27,6	25,6
Глибокий аналіз ринку	27,6	18,6
Великий обсяг ринку	20,7	16,3
Синергія виробництво / маркетинг	16,4	18,6
Уникнення ринків з високою конкуренцією та задоволеними покупцями	7,8	10,5
Уникнення динамічних ринків з частою зміною товара	2,6	4,7

[*Edgett S.*, *Shipley D.* and *Forbes G.* Japanese and British Companies Compared: Contributing Factors to Success and Failure in NPD //Journal of Product Innovation Management, 9, 1992, pp. 3-10.]

Визначте, чи корелюють між собою оцінки японських і британських фірм.

Рішення задачі за допомогою Excel

Хід роботи

1. Увести дані в стовпці A, B, C.

2. В колонках D і E записати формули для обчислення рангів.

3. У полі F3 записати формулу для обчислення різниці рангів. Скопіювати цю формулу в комірки F4: F16.

4. У комірку G3 записати формулу для піднесення цієї різниці в квадрат. Скопіювати її в комірку G4: G16.

5. У комірці G17 підсумувати різниці квадратів.

6. У комірці B18 підрахувати число вимірювань (об'єктів) за допомогою функції СЧЕТ (в нашому випадку це число чинників успіху –14), а в комірці G18 – формулу для обчислення коефіцієнта рангової кореляції Спірмена:

=1-6*G17/(B18*(B18*B18-1)) (см. табл. 3.3.2).

Таблиця 3.3.2

Формули для розрахунків коефіцієнту рангової кореляції Спірмена

	B	C	D	E	F	G
1	Процент фірм		Ранги			
2	японських	британських	японських	британських	Різниця	Різниця в квадраті
3	69,8	75,6	=РАНГ.РВ(B3;B$3:B$16)	=РАНГ.РВ(C3;C$3:C$16)	=D3-E3	=F3*F3
4	79,3	59,3	=РАНГ.РВ(B4;B$3:B$16)	=РАНГ.РВ(C4;C$3:C$16)	=D4-E4	=F4*F4
5	69,9	45,3	=РАНГ.РВ(B5;B$3:B$16)	=РАНГ.РВ(C5;C$3:C$16)	=D5-E5	=F5*F5
6	58,6	61,6	=РАНГ.РВ(B6;B$3:B$16)	=РАНГ.РВ(C6;C$3:C$16)	=D6-E6	=F6*F6
7	55,2	48,8	=РАНГ.РВ(B7;B$3:B$16)	=РАНГ.РВ(C7;C$3:C$16)	=D7-E7	=F7*F7
8	41,4	27,9	=РАНГ.РВ(B8;B$3:B$16)	=РАНГ.РВ(C8;C$3:C$16)	=D8-E8	=F8*F8
9	39,7	34,9	=РАНГ.РВ(B9;B$3:B$16)	=РАНГ.РВ(C9;C$3:C$16)	=D9-E9	=F9*F9
10	36,2	29,1	=РАНГ.РВ(B10;B$3:B$16)	=РАНГ.РВ(C10;C$3:C$16)	=D10-E10	=F10*F10
11	27,6	25,6	=РАНГ.РВ(B11;B$3:B$16)	=РАНГ.РВ(C11;C$3:C$16)	=D11-E11	=F11*F11
12	27,6	18,6	=РАНГ.РВ(B12;B$3:B$16)	=РАНГ.РВ(C12;C$3:C$16)	=D12-E12	=F12*F12
13	20,7	16,3	=РАНГ.РВ(B13;B$3:B$16)	=РАНГ.РВ(C13;C$3:C$16)	=D13-E13	=F13*F13
14	16,4	18,6	=РАНГ.РВ(B14;B$3:B$16)	=РАНГ.РВ(C14;C$3:C$16)	=D14-E14	=F14*F14
15	7,8	10,5	=РАНГ.РВ(B15;B$3:B$16)	=РАНГ.РВ(C15;C$3:C$16)	=D15-E15	=F15*F15
16	2,6	4,7	=РАНГ.РВ(B16;B$3:B$16)	=РАНГ.РВ(C16;C$3:C$16)	=D16-E16	=F16*F16
17						=СУММ(G3:G16)
18	=СЧЁТ(B3:B17)				ρ=	=1-6*G17/(B18*(B18*B18-1))

В результаті обчислень отримаємо таблицю з результатами (табл. 3.3.3).

Таблиця 3.3.3

Результати обчислення коефіцієнту Спірмена

	A	B	C	D	E	F	G
1	Фактори успіху	Процент фірм		Ранги		Різниця	Різниця в квадраті
2		японських	британських	японських	британських		
3	Хороша адаптація до потреб	69,8	75,6	3	1	2	4
4	Переваги над конкурентами щодо якості	79,3	59,3	1	3	-2	4
5	Переваги над конкурентами щодо дизайну	69,9	45,3	2	5	-3	9
6	Переваги над конкурентами щодо співвідношення якість	58,6	61,6	4	2	2	4
7	Переваги над конкурентами щодо конструкції	55,2	48,8	5	4	1	1
8	Вельми конкурентна ціна	41,4	27,9	6	8	-2	4
9	Адаптація	39,7	34,9	7	6	1	1
10	Унікальність	36,2	29,1	8	7	1	1
11	Ефективний маркетинг	27,6	25,6	9	9	0	0
12	Глибокий аналіз ринку	27,6	18,6	9	10	-1	1
13	Великий обсяг ринку	20,7	16,3	11	12	-1	1
14	Синергія виробництво / маркетинг	16,4	18,6	12	10	2	4
15	Уникнення ринків з високою конкуренцією та задоволеними покупцями	7,8	10,5	13	13	0	0
16	Уникнення динамічних ринків з частою зміною товара	2,6	4,7	14	14	0	0
17							34
18	Об'єктів	14				ρ=	0,925

Рішення задачі за допомогою SPSS
Хід роботи

1. Описати в редакторі даних дві змінних з іменами japan, brit (японські та британські фірми).

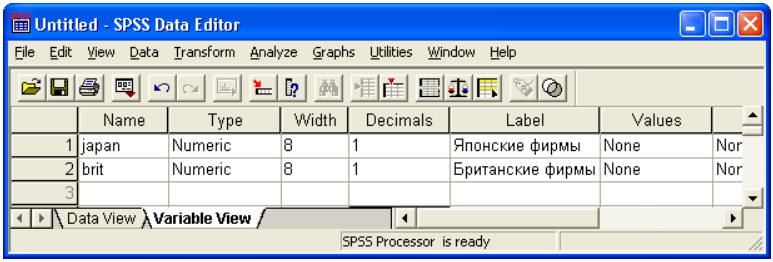


Рис. 3.3.1. Опис змінних в SPSS

Увести дані згідно таблиці, що представлена на рис. 3.2.2.

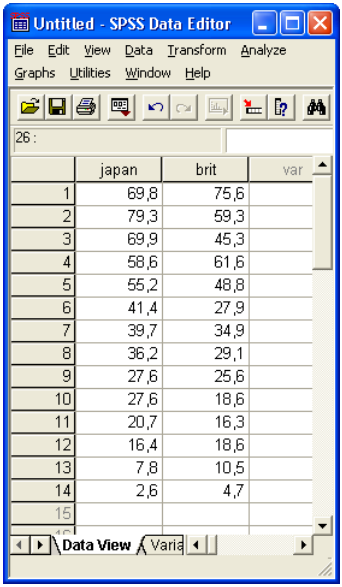


Рис. 3.3.2. Уведення даних до SPSS

2. Обчислити ранги для змінних japan, brit:

Меню: ***Transform*** (Преобразовать), ***Rank Cases***…(Ранги) (рис. 3.3.3)

У результаті з'являться дві нові змінні:

```
From      New
variable  variable  Label
--------  --------  -----
JAPAN     RJAPAN    RANK of JAPAN
BRIT      RBRIT     RANK of BRIT
```

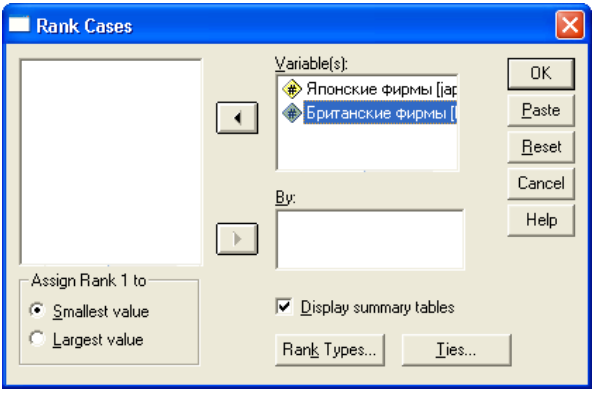


Рис. 3.3.3. Вікно «Rank Cases»

3. Обчислити коефіцієнти Спірмена та Кенделла для даних, представлених рангами:

Меню: ***Analyze*** (Аналіз), ***Correlate*** (Кореляція), ***Bivariate*** (Двумірна).

У вікні ***Bivariate Correlation*** задати необхідні параметри (рис. 3.3.4).

<u>Зауваження</u>: можна обчислювати коефіцієнти рангової кореляції без попереднього обчислення рангів, у такому випадку SPSS сам зробить необхідні розрахунки.

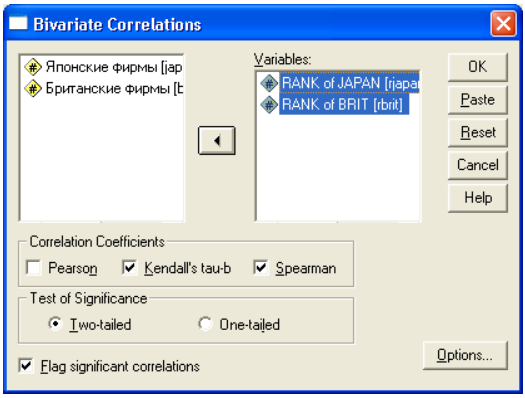


Рис. 3.3.4. Вікно «Bivariate Correlation»

4. Ознайомитися з отриманими результатами у вікні перегляду (рис. 3.3.5).

				Японські фірми	Британські фірми
Kendall's tau_b		Японські фірми	Correlation Coefficient	1,000	,811(**)
			Sig. (2-tailed)	.	,000
			N	14	14
		Британські фірми	Correlation Coefficient	,811(**)	1,000
			Sig. (2-tailed)	,000	.
			N	14	14
Spearman's rho		Японські фірми	Correlation Coefficient	1,000	,928(**)
			Sig. (2-tailed)	.	,000
			N	14	14
		Британські фірми	Correlation Coefficient	,928(**)	1,000
			Sig. (2-tailed)	,000	.
			N	14	14

Рис. 3.3.5. Результати обчислення кореляції

*<u>Зауваження</u>: коефіцієнти, позначені * - значущі на рівні 0,05;** - на рівні 0,01.*

Завдання для самостійної роботи

Керівництво компанії Marston Book Services з Великої Британії (https://www.marston.co.uk/), що займається наданням комплексу послуг з доставки споживачам друкованої продукції (книги, плакати, листівки і т.п.), розробило план посилення орієнтації діяльності компанії на запити споживачів. На одному з етапів розробки даного плану було проведено вивчення думок співробітників компанії і споживачів щодо важливості окремих показників якості послуг, що надаються. У таблиці 3.2.4 наводяться рейтинги окремих показників якості.

Таблиця 3.3.4

Рейтинг показників якості послуг

Показники якості послуг	Оцінки співробітників фірми	Оцінки споживачів
Швидкість обслуговування	1	7
Надійність	4	1
Непошкодженість упаковки замовлень	3	8
Задоволення термінових замовлень	8	3
Простота оформлення замовлень	7	6
Низький рівень рекламацій	2	4
Надання інформації за запитами	6	5
Витримування термінів виконання замовлень	5	2

Визначте, чи узгоджуються оцінки співробітників компанії і споживачів. Вирішіть задачу в середовищі Excel і в середовищі SPSS для Windows.

Вимоги до звіту

Звіт з роботи повинен містити:

– постановку задачі, файли з вихідними даними і результатами, формули для обчислення, значення обчислених коефіцієнтів;

– відповіді на контрольні питання.

Контрольні питання

1. Які коефіцієнти рангової кореляції вам знайомі?
2. Для чого призначений коефіцієнт кореляції Спірмена?
3. За якою формулою він обчислюється? Які значення приймає?
4. Як визначити його значущість? Визначте значимість коефіцієнта Спірмена в прикладі «японські-британські фірми».
5. Як обчислити коефіцієнт Спірмена за допомогою електронних таблиць Excel?
6. Призначення функції «СЧЕТ». Для чого вона використовується в даній роботі?
7. Призначення функції «РАНГ». Які параметри передаються цій функції? Для чого використовується у формулі для обчислення рангу абсолютна адресація?
8. Як перетворити дані в ранги? Які типи рангів є в SPSS?
9. Як обчислити коефіцієнти рангової кореляції в SPSS?
10. Як визначити значимість коефіцієнтів в SPSS?
11. Дайте змістовну інтерпретацію отриманих результатів.

Практична робота 3.4

Обчислення коефіцієнтів рангової кореляції за допомогою R

Хід роботи

1. Увести вихідні дані в Excel

	A	B
1	Japan	Britain
2	69,8	75,6
3	79,3	59,3
4	69,9	45,3
5	58,6	61,6
6	55,2	48,8
7	41,4	27,9
8	39,7	34,9
9	36,2	29,1
10	27,6	25,6
11	27,6	18,6
12	20,7	16,3
13	16,4	18,6
14	7,8	10,5
15	2,6	4,7
16		

Рис.3.4. Вихідні дані

2. Зберегти у форматі .CSV (кома як розбивач)
3. Відкрити RStudio. Імпортувати у змінну data.

```
>data<-read.table("K:\\Rexample\\lab321.csv", sep=";",
dec=",", header=TRUE)
> data
   Japan   Britain
1    69.8     75.6
2    79.3     59.3
3    69.9     45.3
4    58.6     61.6
5    55.2     48.8
6    41.4     27.9
7    39.7     34.9
8    36.2     29.1
9    27.6     25.6
10   27.6     18.6
11   20.7     16.3
12   16.4     18.6
13    7.8     10.5
14    2.6      4.7
```

4. Обчислити коефіцієнт кореляції Спірмена:

```
>   cor(data$Japan,data$Britain, method="spearman")
[1] 0.9284141
```

5. Вивести більш повну статистику розрахунків

```
>cor.test(data$Japan,data$ Britain, method="spearman")

        Spearman's rank correlation rho

data:  data$Japan and data$Britain
S = 32.5716, p-value = 1.663e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
```

```
      rho
0.9284141

Warning message:
In cor.test.default(data$Japan, data$Britain, method = "spearman")
:
  Cannot compute exact p-values with ties
```

6. Обчислити коефіцієнт кореляції Кенделла:

```
> cor.test(data$Japan,data$Britain, method="kendall")

        Kendall's rank correlation tau

data:  data$Japan and data$Britain
z = 4.0083, p-value = 6.115e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8111111
Warning message:
In cor.test.default(data$Japan, data$Britain, method = "kendall") :
  Cannot compute exact p-value with ties
```

Контрольні питання

1. Як обчислити коефіцієнт кореляції Спірмена в R?
2. Як обчислити коефіцієнт кореляції Кенделла в R?
3. Як визначити їх значущість?

Практична робота 3.5
Обчислення коефіцієнта Кенделла засобами Excel

Мета роботи: обчислювати коефіцієнт рангової кореляції, Кенделла, визначати його значимість по таблиці, інтерпретувати результати

Постановка задачі

Експерти аналітичного центру оцінювали шанси депутата на етапі передвиборчої компанії до міської ради наступним чином (див. табл., колонки 1 і 2). За результатами виборів депутатів ранжували за кількістю поданих за них голосів (див. колонку 3). Визначити наскільки точно результати експертної оцінки передбачили дійсні шанси депутатів бути обраними [29].

Хід роботи

1. Увести вихідні дані згідно з рис. 3.5.1.

	A	B	C
1			
2			
3	№ депутата	Ранг за оцінками експертів	Ранг за числом голосів
4	1	6	8
5	2	4	7
6	3	2	6
7	4	10	10
8	5	9	5
9	6	3	2
10	7	1	1
11	8	5	3
12	9	7	4
13	10	8	9
14			
15			
16		Об'єктів	10

Рис. 3.5.1.Вихідні дані

2. Відсортувати дані за рангами оцінок експертів (меню ***Дані, Сортування***). Має бути така таблиця (рис.3.5.2)

	A	B	C
3	№ депутата	Ранг за оцінками експертів	Ранг за числом голосів
4	7	1	1
5	3	2	6
6	6	3	2
7	2	4	7
8	8	5	3
9	1	6	8
10	9	7	4
11	10	8	9
12	5	9	5
13	4	10	10
14			
15			
16		Об'єктів	10

Рис.3.5.2. Вихідні дані, відсортовані

3. Обчислити коефіцієнт Кенделла за формулою:
де n – число об’єктів, які ранжуються, P – число пар, у яких порядок рангів по обом ознакам співпадає.

$$\tau = \frac{4P}{n(n-1)} - 1 \qquad (3.5)$$

Скористайтеся таблицею з формулами (рис. 3.5.3):

	A	B	C	D
3	№ депутата	Ранг за оцінками експертів	Ранг за числом голосів	P
4	7	1	1	=СЧЁТЕСЛИ(C5:$C$13;">1")
5	3	2	6	=СЧЁТЕСЛИ(C6:$C$13;">6")
6	6	3	2	=СЧЁТЕСЛИ(C7:$C$13;">2")
7	2	4	7	=СЧЁТЕСЛИ(C8:$C$13;">7")
8	8	5	3	=СЧЁТЕСЛИ(C9:$C$13;">3")
9	1	6	8	=СЧЁТЕСЛИ(C10:$C$13;">8")
10	9	7	4	=СЧЁТЕСЛИ(C11:$C$13;">4")
11	10	8	9	=СЧЁТЕСЛИ(C12:$C$13;">9")
12	5	9	5	=СЧЁТЕСЛИ(C13:$C$13;">5")
13	4	10	10	=СЧЁТЕСЛИ(C$13:$C14;">10")
14		P	P	=СУММ(D4:D13)
15				
16	Об'єктів	=СЧЁТ(B4:B15)	коеф.Кенделла	=4*D14/(B16*(B16-1))-1

Рис. 3.5.3. Формули для розрахунків

4. В результаті отримаєте таблицю (рис. 3.5.4):

	A	B	C	D
3	№ депутата	Ранг за оцінками експертів	Ранг за числом голосів	P
4	7	1	1	9
5	3	2	6	4
6	6	3	2	7
7	2	4	7	3
8	8	5	3	5
9	1	6	8	2
10	9	7	4	3
11	10	8	9	1
12	5	9	5	1
13	4	10	10	0
14		P	P	35
15				
16	Об'єктів	10	коеф.Кенделла	0,556

Рис. 3.5.4. Результати розрахунків

5. Збережіть таблицю. Визначте значущість коефіцієнта кореляції.

Завдання для самостійного рішення

Гаррі Потер і Герміона Грейнджер ранжували викладачів Хогвардса за рівнем компетентності. Дані ранжування наведені в таблиці. Чи узгоджуються оцінки Гаррі та Герміони?

Викладач	**Ранг Потера**	**Ранг Герміони**
Дамблдор	1	2
МакГонагал	2	1
Снег	4	3
Трилоні	5	5
Флітвик	3	4

Вирішіть ці завдання в SPSS.

Вимоги до звіту

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості по темі, емпіричні і критичні значення обчислених коефіцієнтів кореляції, висновки про їх значущість, відповіді на контрольні питання.

Контрольні питання

1. Як обчислюється коефіцієнт Кенделла. Що він показує? Для яких даних використовується?
2. Призначення функції СЧЕТЕСЛИ і функції СЧЕТ. Для чого вони використовуються в даній роботі?
3. Як визначити значимість коефіцієнта кореляції?

Практична робота 3.6
Обчислення коефіцієнтів асоціації та контингенції засобами Excel

Мета роботи: навчитися обчислювати коефіцієнти асоціації і контингенции для кореляційних таблиць 2х2

Короткі теоретичні відомості

Нижче наведено загальний вигляд кореляційної таблиці двох ознак (ознака у має дві градації у1 і у2, і ознака х має дві градації х1 і х2).

x	y		
	y1	y2	N(xi)
x1	N11	N12	N(x1)
x2	N21	N22	N(x2)
N(yi)	N(y1)	N(y2)	N

Для визначення кореляційної залежності між ознаками, заданими в таблиці 2х2 служать коефіцієнти асоціації і контингенції. Коефіцієнт асоціації характеризує односторонній зв'язок і обчислюється за формулою:

$$Q = \frac{(N11 \cdot N22 - N12 \cdot N21)}{(N11 \cdot N22 + N12 \cdot N21)}$$

(3.6.1)

Коефіцієнт контингенції характеризує двосторонній зв'язок і обчислюється за формулою:

$$\Phi = \frac{(N11 \cdot N22 - N12 \cdot N21)}{\sqrt{N(x1) \cdot N(x2) \cdot N(y1) \cdot N(y2)}} \cdot$$

(3.6.2)

Якщо абсолютне значення коефіцієнта /Ф/> 0,5, то вважається, що надійно встановлено двосторонній зв'язок між ознаками.

Постановка задачі

Визначити чи впливають дистанційні форми підвищення кваліфікації викладачів університету на рівень їх професійної майстерності за даними такої таблиці (таб.3.6.1). Вона містить результати оцінювання студентами 400 викладачів, з яких 250 підвищили свою кваліфікацію за дистанційною формою в останні два роки.

Таблиця 3.6.1

Зв'язок між підвищенням кваліфікації викладачами та оцінюванням студентами їх професійної майстерності

Групи викладачів(x)	Середній бал відносно попереднього оцінювання (y)		Разом
	не змінився або підвищився (y1)	знизився (y2)	
Підвищували кваліфікацію (x1)	150	100	250
Не підвищували кваліфікацію (x2)	80	70	150
Разом	230	170	400

Хід роботи

1. Уведіть вихідні дані, підсумуйте дані у рядках та стовбцях.

2. Обчислить коефіцієнти асоціації та контингенції.

3. Порівняйте формули для розрахунків, наведені на рис. 3.6.1.

	A	B	C	D
1				
2	Групи викладачів(x)	Середній бал відносно попередньої атестації (y)		Разом
3		не змінився або підвищився (y1)	знизився (y2)	
4	Підвищували кваліфікацію (x1)	150	100	=СУММ(B4:C4)
5	Не підвищували кваліфікацію (x2)	80	70	=СУММ(B5:C5)
6	Разом	=СУММ(B4:B5)	=СУММ(C4:C5)	=СУММ(D4:D5)
7				
8	Коефіцієнт ассоціації Q			
9	=(B4*C5-C4*B5)/(B4*C5+C4*B5)			
10	Коефіцієнт контингенції Ф			
11	=(B4*C5-C4*B5)/КОРЕНЬ(D4*D5*B6*C6)			

Рис. 3.6.1. Формули для розрахунків коефіцієнтів асоціації та контингенції

Збережіть таблицю в особистій папці. Запишіть в зошит отримані коефіцієнти. Перевірте результати (рис. 3.6.2) та надайте їм інтерпретацію.

Якщо абсолютне значення коефіцієнта /Q/ > 0,5, то вважається, що надійно встановлено односторонній зв'язок між ознаками.

Якщо абсолютне значення коефіцієнта /Ф/ > 0,5, то вважається, що надійно встановлено двосторонній зв'язок між ознаками.

	A	B	C	D
1				
2	Групи викладачів(x)	Середній бал відносно попередньої атестації (y)		Разом
3		не змінився або підвищився (y1)	знизився (y2)	
4	Підвищували кваліфікацію (x1)	150	100	250
5	Не підвищували кваліфікацію (x2)	80	70	150
6	Разом	230	170	400
7				
8	Коефіцієнт ассоціації Q			
9	0,135			
10	Коефіцієнт контингенції Ф			
11	0,065			

Рис. 3.6.2. Результати розрахунків

Завдання для самостійного рішення

1. Визначте, чи впливає надмірна суворість і дисципліна в сім'ї на впертість і негативізм дітей за даними такої таблиці. Вона містить результати дослідження 115 сімей, в 60-ти з яких було визначено сувору дисципліну.

Дисципліна в сім'ї	Дитина		разом
	уперта	неуперта	
Сувора	50	10	60
Несувора	11	44	55
Разом	61	54	115

2. За даними наступних двох таблиці з’ясувати, чи є зв'язок між расовою приналежністю виборця та його голосуванням.

Расові відмінності на президентських виборах 1964 г., %

Расова приналежність	Голосування за кандидата			
	від демо-кратів	від республікан-ців	Разом, %	Число випад-ків
Білі	59	41	100	1350
Чорні	94	6	100	150
Всі, хто голо-сував	61	39	100	1500

Расові відмінності на президентських виборах 1972 г., %

Голосування за кандидата	Расова приналежність		
	Білі	Чорні	Усі, хто го-лосував
Від демократів	34	89	40
Від республікан-ців	66	11	60
Разом, %	100	100	100
Число випадків	1350	150	1500

Мангейм Дж.Б., Рич Р.К. Политология: Методы исследования. – М. : Издательство “Весь Мир”, 1997. – 544 с. С.420-421.

3. А. Я. Шайкевич досліджував розподіл відносної частоти невизначеного артикля в мові персонажів У. Шекспіра і показав неоднорідність цього розподілу для двох, що виділилися саме завдяки цьому аналізу, груп персонажів. У слуг, городян і носіїв комічного початку (блазнів, дотепників, дурнів) відмічено підвищення частоти невизначеного артикля. І, навпаки, у героїв «серйозних» і тих, що високо стоять на соціальних сходах частота невизначеного артикля знижується. За даними такої таблиці перевірте цей факт, обчисливши коефіцієнти асоціації і контингенции.

Залежність частоти невизначеного артикля від соціального стану персонажа в п'єсах У. Шекспіра

Персонаж	Частота вживання невизначеного артикля		Разом
	Вища за середнє	Нижча за середнє	
Комічні персонажі	17	6	23
«Серйозні герої»	5	16	21
Разом	22	22	44

Шайкевич А.Я. Дистрибутивно-статистичний аналіз в семантиці // Принципи та методи семантичних досліджень. - М., 1976. - С. 353 - 378]

Вимоги до звіту

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості по темі, значення обчислених коефіцієнтів кореляції, висновки про їх значущість, відповіді на контрольні питання.

Контрольні питання

1. Для чого використовують коефіцієнти асоціації і контингенції?

2. Для яких ознак вони використовуються? В яких межах змінюються?

3. Як вони обчислюються? Як їх обчислити за допомогою електронних таблиць?

4. Дайте змістовну інтерпретацію отриманих результатів.

Практична 3.7
Обчислення кореляційного відношення та бісеріального коефіцієнта кореляції

Завдання 1. Обчислення кореляційного відношення.

У таблиці наведено дані віку респондентів та їх оцінок за тестом цифра-знак шкали інтелекта дорослих [8, с.139].

Об'єкт	X (вік)	Y (тест цифра-знак)
1	10	7
2	10	8
3	10	9
4	10	9
5	10	10
6	14	8
7	14	9
8	14	10
9	14	11
10	18	9
11	18	10
12	18	11
13	18	12
14	22	11
15	22	12
16	22	11
17	22	12
18	26	9
19	26	10
20	26	11
21	30	8
22	30	9
23	30	9
24	30	10
25	34	7
26	34	9
27	34	10
28	38	8

1. Підрахуйте коефіцієнт Пірсона-Браве. Інтерпретуйте його. Побудуйте діаграму розсіювання. Чи спостерігаємо ми залежність між змінними?

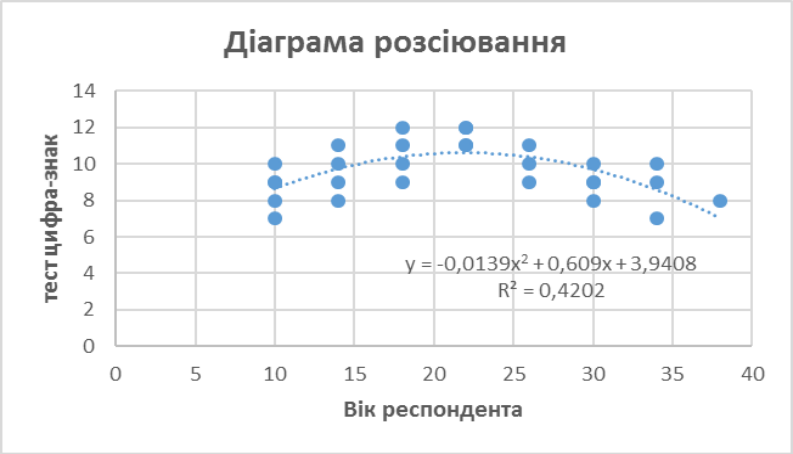


Рис. 3.7.1. Діаграма розсіювання

2. Розрахуйте кореляційне відношення за формулою:

$$\eta^2 = 1 - SSвнутр/SSзагальне$$

Спочатку розрахуємо середнє для усіх груп та загальне середнє, як показано на рис. 3.7.2.

	A	B	C	D	E	F	G	H	I
1		Вік							
2	№ респондента	10 років	14 років	18 років	22 роки	26 років	30 років	34 роки	38 років
3	1	7	8	9	11	9	8	7	8
4	2	8	9	10	11	10	9	9	
5	3	9	10	11	12	11	9	10	
6	4	9	11	12	12		10		
7	5	10							
8	середнє груп	8,60	9,50	10,50	11,50	10,00	9,00	8,67	8,00
9	загальне середнє	9,61							

Рис. 3.7.2. Розрахунок середнього груп та загального середнього

Далі розрахуємо квадрат відхилення від групового середнього (рис. 3.7.3).

	K	L	M	N	O	P	Q	R	S	T
1		Квадрат відхилення від групового середнього								
2	№ респондента	**10 років**	**14 років**	**18 років**	**22 роки**	**26 років**	**30 років**	**34 роки**	**38 років**	
3	1	2,56	2,25	2,25	0,25	1	1	2,77778	0	
4	2	0,36	0,25	0,25	0,25	0	0	0,11111		
5	3	0,16	0,25	0,25	0,25	1	0	1,77778		
6	4	0,16	2,25	2,25	0,25		1			
7	5	1,96								
8		**5,20**	**5,00**	**5,00**	**1,00**	**2,00**	**2,00**	**4,67**	**0,00**	24,87

Рис. 3.7.3. Розрахунок квадрата відхилення від групового середнього (*SSвнутр*)

На наступному кроці обчислимо квадрат відхилення від загального середнього, та кореляційне відношення (рис.3.7.4).

	A	B	C	D	E	F	G	H	I
1		**Вік**							
2	№ респондента	**10 років**	**14 років**	**18 років**	**22 роки**	**26 років**	**30 років**	**34 роки**	**38 років**
3	1	7	8	9	11	9	8	7	8
4	2	8	9	10	11	10	9	9	
5	3	9	10	11	12	11	9	10	
6	4	9	11	12	12		10		
7	5	10							
8	**середнє груп**	**8,60**	**9,50**	**10,50**	**11,50**	**10,00**	**9,00**	**8,67**	**8,00**
9	**загальне середнє**	**9,61**							
10									
11		**квадрат відхилення від загального середнього**							
12	№ респондента	**10 років**	**14 років**	**18 років**	**22 роки**	**26 років**	**30 років**	**34 роки**	**38 років**
13	1	6,80	2,58	0,37	1,94	0,37	2,58	6,80	2,58
14	2	2,58	0,37	0,15	1,94	0,15	0,37	0,37	
15	3	0,37	0,15	1,94	5,73	1,94	0,37	0,15	
16	4	0,37	1,94	5,73	5,73		0,15		
17	5	0,15							
18		**10,27**	**5,05**	**8,19**	**15,33**	**2,46**	**3,47**	**7,32**	**2,58**
19								Сумма	54,68
20									
21		eta=	0,545						

Рис. 3.7.4. Розрахунок квадрата відхилення від загального середнього (*SSзагальне*)

Завдання 2. Розрахувати бісеріальний коефіцієнт кореляції для даних, в яких для 10-ти респондентів наведено їх стать (0–хлопці, 1–дівчатка) та ранг на іспиті (Глас, с.139).

Об'єкт	X (стать студента)	Y (ранг на іспиті)
1	0	1
2	1	10
3	0	2
4	1	9
5	0	5
6	0	8
7	1	4
8	1	7
9	0	3
10	0	6

1. Введіть дані та посортуйте за ознакою X. Розрахуйте окремо середнє для хлопців та дівчат, потім – бісеріальний коефіцієнт (рис.3.7.5).

2. Для розрахунків скористайтеся наступною формулою:

$$pb=2/n * (Y_1cep-Y_0cep),$$

де

Y_1cep - середній ранг об'єктів, які мають 1 по X,

Y_0cep - середній ранг об'єктів, які мають 0 по X.

	A	B	C	D	E	F	G	H
1	Об'єкт	X	Y		Об'єкт	X	Y	
2	1	0	1		1	0	1	
3	2	1	10		3	0	2	
4	3	0	2		5	0	5	
5	4	1	9		6	0	8	
6	5	0	5		9	0	3	середнє
7	6	0	8		10	0	6	4,17
8	7	1	4		2	1	10	
9	8	1	7		4	1	9	
10	9	0	3		7	1	4	середнє
11	10	0	6		8	1	7	7,50
12								
13							pb=	0,667

Рис. 3.7.5. Результати розрахунків бісеріального коефіцієнта

Вимоги до звіту

Звіт повинен містити назву роботи, мету роботи, короткі теоретичні відомості по темі, значення обчислених коефіцієнтів кореляції, висновки про їх значущість, відповіді на контрольні питання.

Контрольні питання

1. Для чого використовують кореляційне відношення та бісеріальний коефіцієнт кореляції?

2. Для яких ознак вони використовуються? В яких межах змінюються?

3. Як вони обчислюються? Як їх обчислити за допомогою електронних таблиць?

4. Дайте змістовну інтерпретацію отриманих результатів.

РОЗДІЛ 4
ОСНОВИ РЕГРЕСІЙНОГО АНАЛІЗУ

Ще одним методом визначення зв'язку між ознаками, які варіюють, є регресійний аналіз. Вперше термін «регресія» був застосований англійським вченим-статистиком Ф.Гальтоном в 1886 р в теорії спадковості. Ф.Гальтон регресією (regression to mediocrity - повернення до середнього стану) назвав явище, яке полягало у тому, що діти тих батьків, чий зріст перевищував середнє значення на а одиниць, мали в середньому зростання, що перевищує середнє значення менш ніж на а одиниць.

Як ми вже з'ясували, завдання кореляційного аналізу – встановлення кореляційної залежності між змінними і визначення величини цієї залежності у вигляді коефіцієнтів кореляції.

Завдання регресійного аналізу – вираз кореляційної залежності у вигляді функціональних відносин. Ми розглянемо найпростішу форму цієї функціональної залежності – лінійну. Маючи функціональне відношення зв'язку між змінними X та Y, можна оцінювати Y по X, тобто встановлювати причинно-наслідкові зв'язки між змінними. Змінна, яку ми хочемо оцінити, називається залежною змінною Y, а змінна, яка використовується для її оцінки, – незалежною змінної або фактором X.

Приклади завдань на оцінювання Y по X:

- Передбачення відміток з англійської мови в університеті (Y) за оцінками з цього предмету в школі (X).
- Передбачення загальної навчальної успішності в школі (Y) за інтелектом (X).

- Передбачення обсягу продажів (Y) за витратами на рекламу (X).
- Прогнозування плинності кадрів (Y) в залежності від рівня задоволеності роботою (X) і т.п.

Щоб вивести спосіб оцінювання змінної Y на основі значень змінної X, ми повинні знати, як пов'язані між собою X і Y. Розглянемо алгоритм оцінювання на прикладі.

Нехай ми хочемо навчитися передбачати середній бал сесії у студентів (Y) в залежності від того, скільки часу вони витрачають на перегляд телесеріалів в період підготовки до іспитів (X). Для цього нам необхідно:

1) дізнатися, скільки часу (годин в тиждень) витрачають на перегляд серіалу n студентів напередодні сесії;

2) виміряти середній бал сесії у цих же студентів;

3) вивести рівняння, що зв'язує Y і X;

4) далі використовувати це рівняння для передбачення середнього балу сесії (Y) у тих студентів, для яких нам відомо, скільки часу в період підготовки до сесії вони витратили на перегляд серіалів (X).

Дані вимірювань часу на перегляд серіалів і середнього балу сесії можна представити графічно у вигляді діаграми розсіювання (див. рис. 4.1), тому що ці дві змінні пов'язані між собою кореляційної залежністю. Завдання регресійного аналізу, як уже зазначалося, представити кореляційну залежність у вигляді функціональних відношень. У разі лінійної регресії (а ми розглядаємо саме цей випадок), це завдання зводиться до знаходження такої прямої, сума квадратів відхилень від якої точок діаграми розсіювання (сума квадратів помилки оцінки) була б мінімальною. Ця пряма називається лінією передбачення або лінією тренду. Рівняння такої прямої – рівняння регресії – в загальному вигляді виглядає наступним чином:

$$y = b_1 * x + b_0 \qquad (4.1)$$

де b_1 и b_0 обрані таким чином, щоб сума квадратів помилок оцінки була б мінімальною (критерій найменших квадратів):

$$b_1 = r_{xy} * S_y / S_x \qquad (4.2)$$

$$b_0 = y_{cp} - b_1 * x_{cp} \qquad (4.3)$$

де r_{xy} – коефіцієнт кореляції Пірсона між X и Y;

S_x, S_y – стандартні відхилення для X та Y.

x_{cp}, y_{cp} – середні значення X та Y.

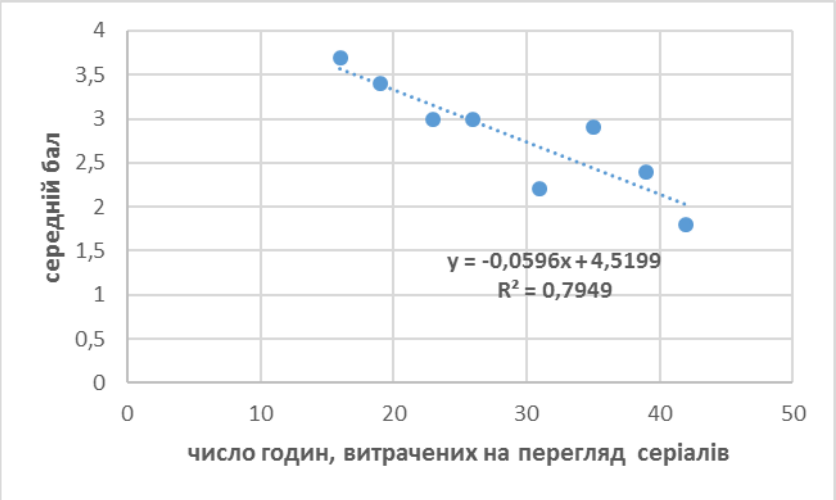


Рис. 4.1. Диаграма розсіювання, лінія передбачення, рівняння регрессії

Практична робота 4.1
Обчислення коефіцієнтів кореляції та прогноз за допомогою лінійної регресії в MS Excel

***Мета роботи*:** навчитися використовувати електронні таблиці MS Excel для обчислення коефіцієнтів кореляції і прогнозування з використанням лінійної регресії.

Постановка задачі

У таблиці 4.1.1 наведені результати тестування 20-ти абітурієнтів на вступному іспиті в університет (назвемо його «тест 1» – це ознака X) і результати тестування цих же осіб вже на випускному курсі («тест 2», ознака Y). Побудувати діаграму розсіювання для цих даних, обчислити коефіцієнт кореляції Пірсона-Браве і вивести рівняння регресії, що зв'язує ознаки X і Y.

Таблиця 4.1.1

Результати тестування в університеті

Тест 1 (X)	Тест 2 (У)
66	94
62	100
70	101
80	102
82	103
74	105
73	104
79	106
83	106
80	109
81	110
84	110
80	111
89	112
88	112
95	114
98	114
97	115
94	117
89	118

Хід роботи

1. Увести вихідні дані згідно таблиці (рис. 4.1.1).

2	№ студента	тест 1	тест 2				
3	1	66	94				
4	2	62	100				
5	3	70	101				
6	4	80	102				
7	5	82	103				
8	6	74	105				
9	7	73	104				
10	8	79	106				
11	9	83	106				
12	10	80	109				
13	11	81	110				
14	12	84	110				
15	13	80	111				
16	14	89	112				
17	15	88	112				
18	16	95	114				
19	17	98	114				
20	18	97	115				
21	19	94	117				
22	20	89	118				

Рис. 4.1.1. Уведення результатів тестування

2. Обчислити коефіцієнт кореляції Пірсона-Браве:

а) меню ***Дані, Аналіз даних, Кореляція (рис. 4.1.2);***

б) в діалоговому вікні "***Кореляція***" (рис. 4.1.3) визвати та переглянути справку;

в) задати параметри: вхідний інтервал (обидва ряди даних), групування за стовпцями. Натиснути ОК. Записати в зошит отриманий коефіцієнт кореляції. Визначити його значимість по таблиці критичних значень.

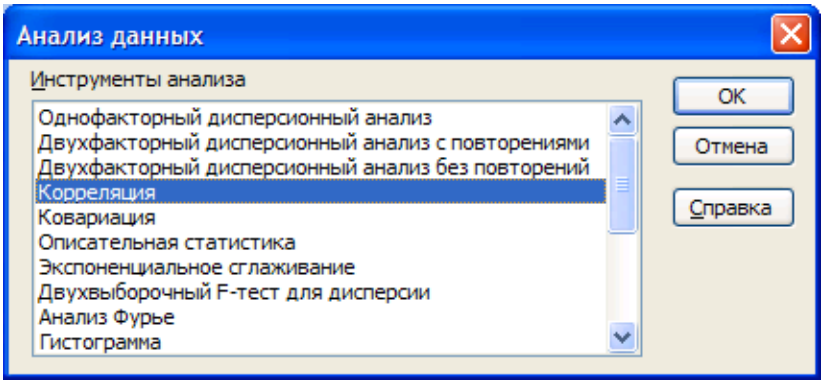


Рис. 4.1.2. Діалогове вікно «Аналіз даних»

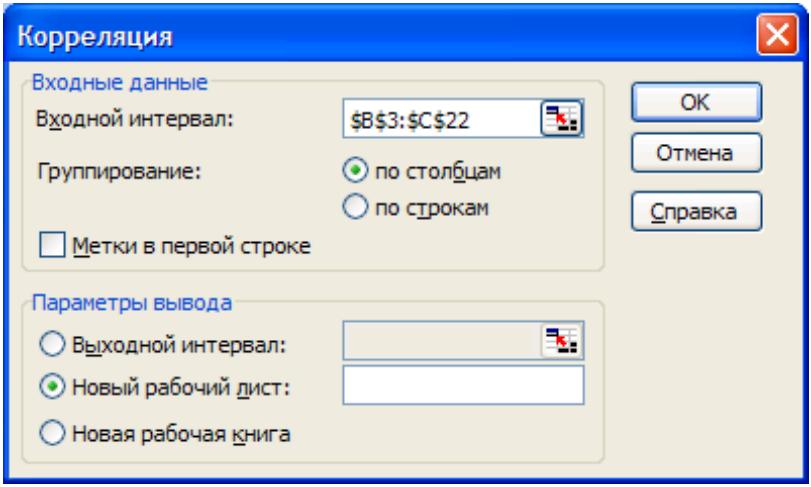


Рис. 4.1.3. Діалогове вікно «Кореляція»

3. Побудувати діаграму розсіювання (тип «точкова»). Зробити припущення про наявність чи відсутність кореляції між X і Y за зовнішнім виглядом діаграми.

4. Побудувати лінію передбачення за методом найменших квадратів (лінійна регресія):

а) натиснути правою кнопкою по точках діаграми;

б) в контекстному меню вибрати «***Додати лінію тренда»;***

в) у діалоговому вікні “***Лінія тренда***” на вкладці “***Тип***” (рис. 4.1.4) укажіть – “**лінійна**”, на вкладці “***Параметри***” – поставте прапорець у опцій “**Показувати рівняння на диаграмі**” та «**Поместити величину достовеірностиі аппроксимації**» (см. рис. 4.1.5).

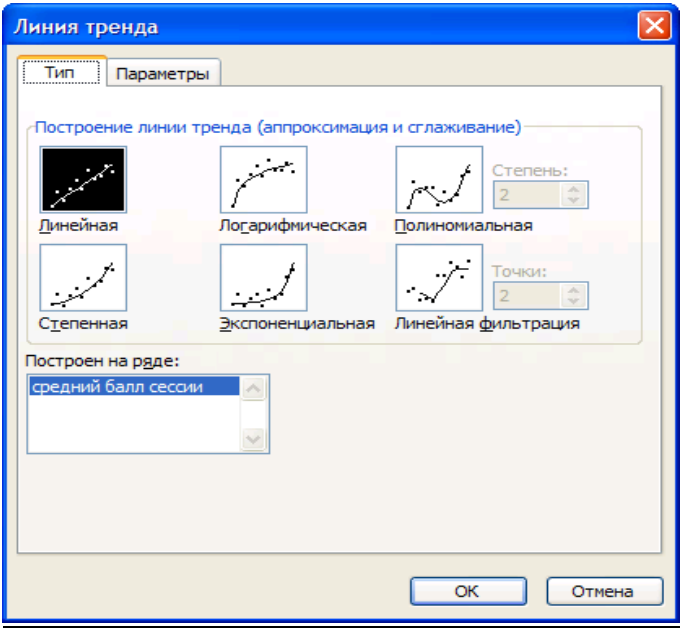


Рис. 4.1.4. Діалогове вікно «Лінія тренду», вкладка «Тип»

У результаті повинна вийти точкова діаграма з лінією тренду, рівнянням регресії та коефіцієнтом достовірності апроксимації, як це показано на рис. 4.1.6.

5. Рівняння потрібно записати в зошит. І, користуючись MS Excel, передбачити по ньому значення випускного тесту для осіб, у яких за результатами вступного тесту були такі бали:

а) 43;
б) 75;
в) 99.

6. Зберегти таблицю в особистій папці.

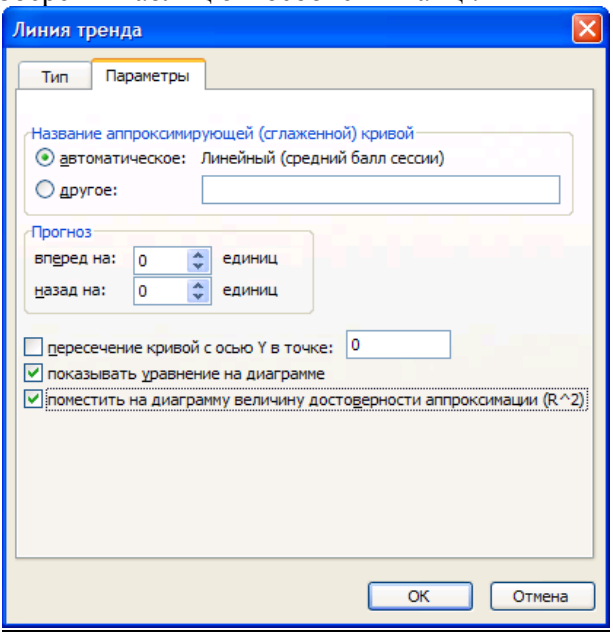


Рис. 4.1.5. Діалогове вікно «Лінія тренду», вкладка «Параметри»

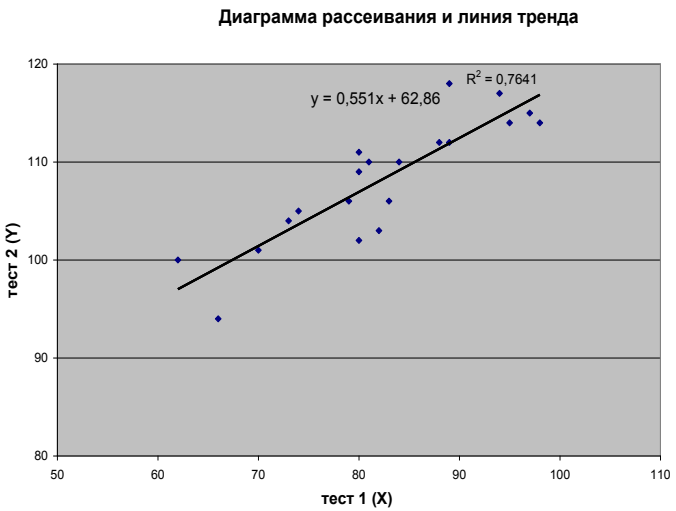


Рис. 4.1.6. Діаграма розсіювання, лінія тренду та рівняння регресії

Завдання для самостійної роботи

1. У таблиці наведено коефіцієнти інтелекту IQ батьків та дітей десяти сімей. Визначте, чи існує кореляційна залежність між цими ознаками.

IQ батьків	IQ дітей
125	110
120	105
110	95
105	125
105	120
95	105
95	75
90	95
80	90
75	80

2. Побудуйте модель лінійної регресії між роком і чисельністю науковців в Україні за даними Держкомстату України (http://www.ukrstat.gov.ua/).

Наукові кадри та кількість організацій України

Рік	Кількість організацій, які ведуть наукові розробки	Чисельність наукових співробітників, осіб.	Чисельність докторів наук в економіці України, осіб.	Чисельність кандидатів наук в економіці України, осіб.
1991	1344	295010	8133	…
1992	1350	248455	8797	…
1993	1406	222127	9224	…
1994	1463	207436	9441	…
1995	1453	179799	9759	57610
1996	1435	160103	9974	58132
1997	1450	142532	10322	59332
1998	1518	134413	10446	59703
1999	1506	126045	10233	59547
2000	1490	120773	10339	58741
2001	1479	113341	10603	60647
2002	1477	107447	11008	62673
2003	1487	104841	11259	64372
2004	1505	106603	11573	65839
2005	1510	105512	12014	68291
2006	1452	100245	12488	71893
2007	1404	96820	12845	74191
2008	1378	94138	13423	77763
2009	1340	92403	13866	81169
2010	1303	89534	14418	84000
2011	1255	84969	14895	84979
2012	1208			

Побудуйте також експоненціальну та квадратичну модель. Порівняйте їх (рис.4.1.7)

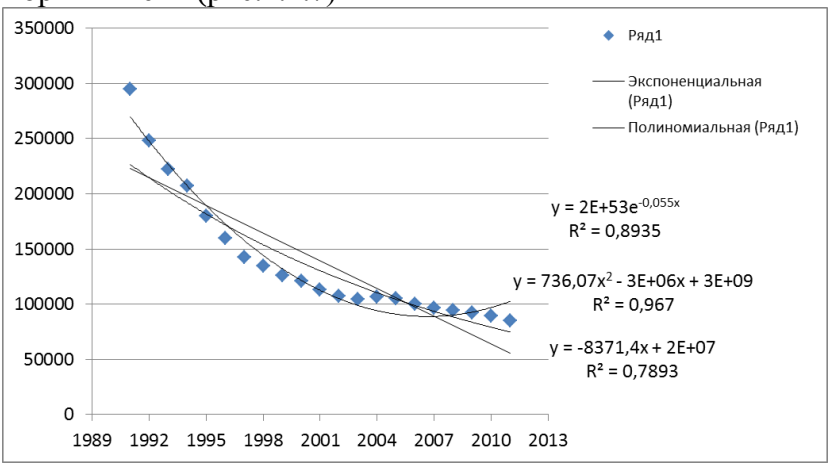


Рис. 4.1.7. Лінійна, експоненційна та квадратична модель

Вимоги до звіту

Звіт повинен містити:

– постановку задачі, вихідні дані, коефіцієнт кореляції, рівняння лінії передбачення, результати передбачення;

– файл з результатами.

Контрольні питання

1. У чому полягає завдання регресійного аналізу?
2. Що являє собою діаграма розсіювання?
3. Як отримати рівняння лінії передбачення?
4. Як здійснюються прогнози за допомогою рівняння лінії передбачення (лінії тренду)?
5. Як обчислити коефіцієнт кореляції Пірсона-Браві?
6. Що показує величина достовірності апроксимації?

Практична робота 4.2
Обчислення коефіцієнтів кореляції та прогноз за допомогою лінійної регресії в R

Хід роботи

1. Побудуємо для цих даних файл в Excel.

IQ батьків	IQ дітей
125	110
120	105
110	95
105	125
105	120
95	105
95	75
90	95
80	90
75	80

2. Прочитаємо його в R:

```
> data411<-read.table("K:\\Rexample\\lab411.csv", sep=";",
dec=",", header=TRUE)
```

3. Отримаємо рівняння регресії:

```
 fit <- lm(data411$IQ.детей ~ data411$IQ.родит)
> summary(fit)

Call:
lm(formula = data411$IQ.детей ~ data411$IQ.родит)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-22.074  -6.370  -1.888   6.370  22.074

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             41.4894    29.0066   1.430   0.1905
data411$IQ.родит         0.5851     0.2867   2.041   0.0756 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.9 on 8 degrees of freedom
Multiple R-squared: 0.3423,     Adjusted R-squared: 0.2601
F-statistic: 4.165 on 1 and 8 DF,  p-value: 0.07559
```

Коефіцієнти в рівнянні регресії представлені в стовпці Estimate. Підставивши їх в рівняння регресії, отримаємо вираз: IQ.детей = 41.4894 + 0,5851 * IQ.родит.

Крім значень коефіцієнтів R показує нам величини помилок, або стандартного відхилення, для кожного коефіцієнта. Для перевірки гіпотези про значимість коефіцієнту, використовується t-критерій Стьюдента. R обчислює як саму величину t так і ступінь значущості нашої гіпотези Pr (> | t |). Так, в нашому випадку величина 0.1905 означає, що ми на 100 * (1-0.1905) = 81% впевнені в тому, що вільний член в нашому виразі відмінний від нуля.

Далі ми можемо перевірити, наскільки точно наша модель описує дані. Для цього використовуються коефіцієнт детермінації R2. Чим ближче його величина до 1, тим краще. 1 – це ідеальний результат, що означає, що модель на 100% описує дані.

І, нарешті, останнє, що ми можемо перевірити, це те, наскільки передбачувана величина залежить від предикторів. Для цього висувається нульова гіпотеза, що передбачена величина взагалі не залежить від предикторів. Для цієї гіпотези визначається p-значення. У нашому випадку, воно вийшло рівним 0.07559. Тобто ми можемо бути впевнені приблизно на

92%, що передбачувана величина дійсно залежить від предикторів. Зазвичай, має сенс дивитися на цей параметр в першу чергу, адже він визначає, наскільки взагалі наша модель адекватна.

Контрольні питання

1. Як підготувати дані для регресійного аналізу в R?
2. За допомогою якої функції будується лінійна регресійна модель?
3. Чим відрізняється функція cor.test від функції cor?

Практична робота 4.3
Обчислення коефіцієнтів кореляції та прогноз за допомогою лінійної регресії в SPSS

***Мета роботи*:** навчитися використовувати SPSS для обчислення коефіцієнтів кореляції і прогнозування з використанням лінійної регресії.

Постановка задачі

Керівництво швейної фабрики хотіло б мати інформацію про те, скільки суконь кожного фасону кольору і розміру буде продано. З одного боку, не можна допускати перевиробництва, тому що в цьому випадку доведеться продавати товар за заниженими цінами, з іншого боку, нестача призводить до необхідності випускати більше виробів, ніж було намічено до даного терміну. Якщо мати прогноз через 5 тижнів після надходження першого найменування у продаж, можна забезпечити виготовлення необхідної кількості продукції до наміченого терміну. У таблиці 4.3.1 наведені дані про сумарний обсяг проданих виробів кожного фасону, кольору і розміру та відповідні цифри за п'ять тижнів, отримані на початок дослідження. Потрібно уявити графічно дані і отримати оцінки найменших квадратів у моделі $\boldsymbol{y = b_0 + b_1 x}$.

Таблиця 4.3.1

.

Продано в перші п’ять тижнів	Продано разом
235	392
122	190
34	74
196	307
121	200
127	185
177	291
125	191
159	314
75	140
140	235
127	188
89	142
139	253
70	96
62	96

Рішення в SPSS

Хід роботи

1. Описати в редакторі дві змінні «**Продано всього»** и «**Продано в перші п’ять тижнів»** та ввести дані таблиці 4.3.1.

2. Обрати меню ***Analyze... (Аналіз), Regression...(Регресія), Linear... (Лінійна).*** З’явиться діалогове вікно Linear Regression (рис. 4.3.1).

Перенести змінну «**Продано Всього»** в поле для залежних змінних та присвоїти змінній «**Продано в перші 5 тижнів»** статус незалежної змінної. Почати розрахунок натисканням **ОК**.

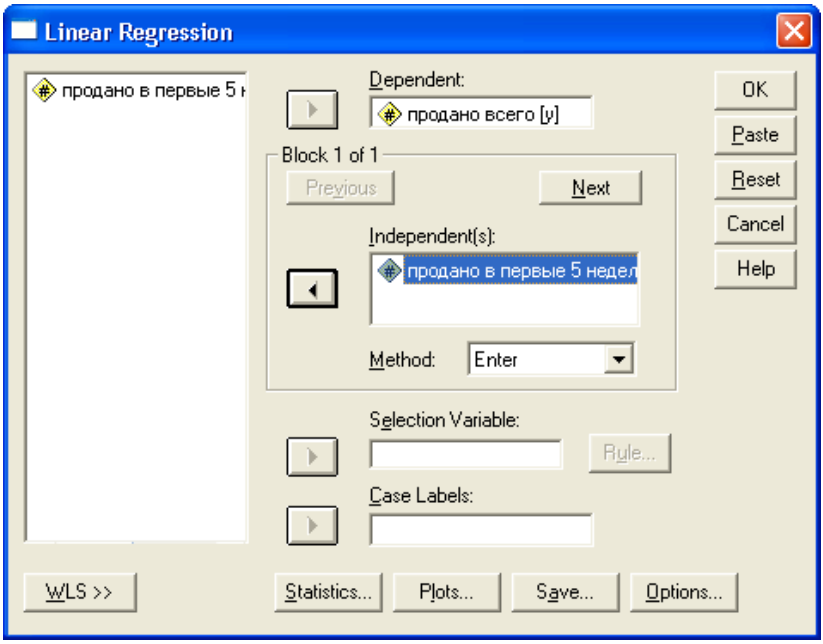


Рис. 4.3.1. Діалогове вікно «Linear Regression»

3. Вивід основних результатів представлено на рис. 4.3.2 и 4.3.3.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,974[a]	,949	,946	20,666

a. Predictors: (Constant), продано в первые 5 недель

Рис. 4.3.2. Вивід результатів. Сумарна модель

Сумарна модель регресії (рис. 4.3.2) включає: значення коефіцієнта кореляції Пірсона R = 0,974; значення коефіцієнта

детермінації R Square = 0,949; значення уточненного коефіцієнта детермінації та стандартну помилку оцінки.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,827	13,821		-,132	,897
	продано в первые 5 недель	1,663	,103	,974	16,202	,000

a. Dependent Variable: продано всего

Рис. 4.3.3. Вивід результатів.
Коефіцієнти рівняння регресії.

У таблиці «Коефіцієнти» (див. рис. 4.3.3) вікна виведення наводяться значення коефіцієнтів рівняння регресії, їх стандартної помилки, стандартизовані коефіцієнти, а також значення статистики Стьюдента t та її значимість Sig. Коефіцієнт є значущим, якщо Sig <0,05.

Отримане рівняння регресії: y = 1,663x - 1,827.

4. Побудувати лінію регресії. Для цього необхідно виконати наступні кроки:

- Побудувати діаграму розсіювання.
- Кликнути двічі на цьому графику, щоб перенести його в редактор діаграм.
- Обрати в редакторі діаграм меню ***Chart... (Діаграма) Options... (Опції)***
- ***Відкриється*** діалогове вікно Scatterplot Options (Опції для діаграмми розсіювання) (рис. 4.3.4).

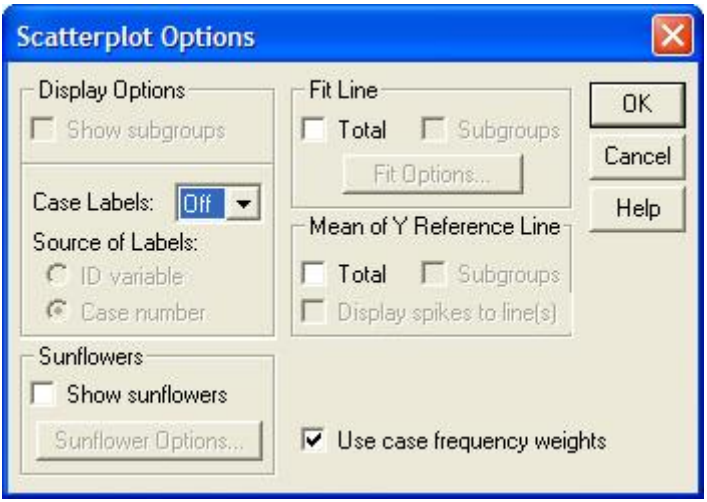


Рис. 4.3.4. Діалогове вікно «Scatterplot Options»
(Опції для диаграми розсіювання**)**

○ В рубриці ***Fit Line*** (Підбір кривої) поставити прапорець проти опції ***Total*** и щёлкнути на кнопці ***Fit Options***. Відкриється діалогове вікно ***Scatterplot Options: Fit Line*** (рис. 4.3.5).

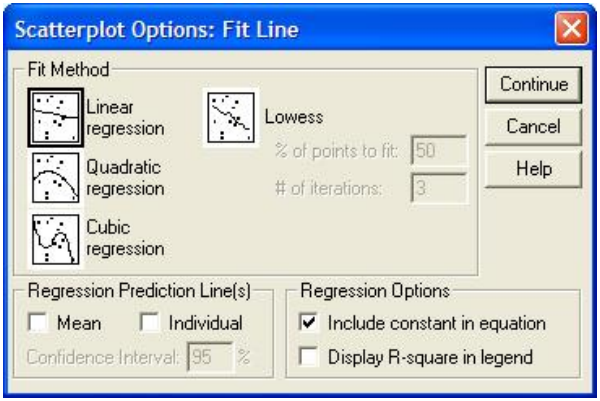


Рис. 4.3.5. Діалогове вікно «Scatterplot Options: Fit Line»

o Підтвердити попередню установку **Linear Regression** щелчком на **Continue** та **OK**.

o Закрити редактор діаграм и щёлкнути один раз де-небудь вне графіка. У вікні перегляду з'явиться діаграма розсіювання з лінією регресії (см. рис. 4.3.6).

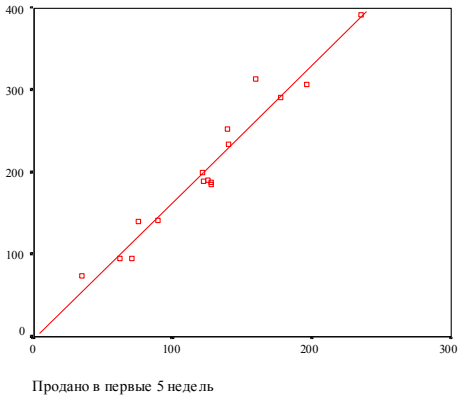


Рис. 4.3.6. Діаграма розсіювання з лінією регресії

Вимоги до звіту:

Звіт має містити:

- постановку задачі, вихідні дані, коефіцієнт кореляції, рівняння лінії передбачення, результати передбачення;
- файл з результатами

Контрольні питання

1. Яким чином ви описали змінну «Продано в перші 5 тижнів» в редакторі даних?

2. Описати елементи діалогового вікна «**Scatterplot Options**».

3. Описати елементи діалогового вікна «Scatterplot Options: Fit Line».

4. Як отримати рівняння лінії передбачення в SPSS?

5. Дайте інтерпретацію отриманих результатів.

Завдання для обчислення лінійної регресії

Завдання 1.

У таблиці задані значення змінних X та Y, які виміряні на виборці з 11 респондентів. X – кількість звернень респондента за консультацією до співробітників; Y – середній бал респондента за підсумками атестації.

№ об'єкта	**X**	**Y**
1	10	3,7
2	14	3,9
3	5	6,7
4	8	9,5
5	9	3,4
6	12	5,6
7	4	3,7
8	7	2,7
9	11	5,5
10	13	2,9
11	6	10,7

Побудуйте модель регресії, розрахуйте всі показники за допомогою пакета аналізу (регресія, рис. 4.3.9.).

Далі розрахуйте самостійно. Для цього спочатку підрахуйте середнє та стандартне відхилення для X та Y.

Потім коефіцієнт кореляції Пірсона r, коефіцієнти рівняння регресії b1 та b0.

Далі зробіть обчислення в колонках D, E, F,G (рис.4.3.7.).

- Знайдіть суми квадратів SSt (total – загальна мінливість),
- SSr (regression – мінливість, яка описується моделлю),
- SSe (error – помилкова, яка пояснюється іншими факторами, які ми не враховуємо).

- Обчисліть стандартну помилку регресії Syx

	A	B	C	D	E	F	G
1	№ обєкта	X	Y	(Yi-Ycp)^2	Ypr	(Y-Ypr)^2	(Ypr-Ycp)^2
2	1	10	3,7	2,56	5,03	1,7689	0,0729
3	2	14	3,9	1,96	3,95	0,0025	1,8225
4	3	5	6,7	1,96	6,38	0,1024	1,1664
5	4	8	9,5	17,64	5,57	15,4449	0,0729
6	5	9	3,4	3,61	5,3	3,61	0
7	6	12	5,6	0,09	4,49	1,2321	0,6561
8	7	4	3,7	2,56	6,65	8,7025	1,8225
9	8	7	2,7	6,76	5,84	9,8596	0,2916
10	9	11	5,5	0,04	4,76	0,5476	0,2916
11	10	13	2,9	5,76	4,22	1,7424	1,1664
12	11	6	10,7	29,16	6,11	21,0681	0,6561
13	сума	99	58,3	72,1		64,081	8,019
14	середнє	9,00	5,30	SSt		SSE	SSR
15	ст. відх.	3,32	2,69				
16		r	-0,3335				
17		b1	-0,27				
18		b0	7,73				

Рис. 4.3.7. Обчислення показників моделі власноруч

$$SST = \sum (Y_i - \overline{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}}$$

Рис. 4.3.8. Формули для суми квадратів та стандартної помилки

Знайдіть відповідні значення в моделі, яку розрахував Excel, та виділить їх однаковими кольорами (рис. 4.3.9).

	A	B	C	D	E	F	G	H	I	J
21		ВЫВОД ИТОГОВ								
22										
23		*Регрессионная статистика*								
24		Множественный R	0,3335							
25		R-квадрат	0,111221							
26		Нормированный R-кв	0,012467							
27		Стандартная ошибка	2,668354							
28		Наблюдения	11							
29										
30		Дисперсионный анализ								
31			*df*	*SS*	*MS*	*F*	*Значимость F*			
32		Регрессия	1	8,019	8,019	1,126246469	0,316224499			
33		Остаток	9	64,081	7,120111111					
34		Итого	10	72,1						
35										
36			*Коэффициен*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
37		Y-пересечение	7,73	2,426988615	3,185017001	0,011095002	2,239770321	13,22022968	2,239770321	13,22022968
38		Переменная X 1	-0,27	0,254417536	-1,0612476	0,316224499	-0,845532452	0,305532452	-0,845532452	0,305532452

Рис. 4.3.9. Обчислення показників моделі за допомогою пакету аналізу

	A	B	C	D	E	F	G
1	№ об'єкта	**X**	**Y**	**(Yi-Yср)^2**	**Ypr**	**(Y-Ypr)^2**	**(Ypr-Yср)^2**
2	1	10	3,7	=(C2-$C$14)^2	=$C$18+$C$17*B2	=(C2-E2)^2	=(E2-$C$14)^2
3	2	14	3,9	=(C3-$C$14)^2	=$C$18+$C$17*B3	=(C3-E3)^2	=(E3-$C$14)^2
4	3	5	6,7	=(C4-$C$14)^2	=$C$18+$C$17*B4	=(C4-E4)^2	=(E4-$C$14)^2
5	4	8	9,5	=(C5-$C$14)^2	=$C$18+$C$17*B5	=(C5-E5)^2	=(E5-$C$14)^2
6	5	9	3,4	=(C6-$C$14)^2	=$C$18+$C$17*B6	=(C6-E6)^2	=(E6-$C$14)^2
7	6	12	5,6	=(C7-$C$14)^2	=$C$18+$C$17*B7	=(C7-E7)^2	=(E7-$C$14)^2
8	7	4	3,7	=(C8-$C$14)^2	=$C$18+$C$17*B8	=(C8-E8)^2	=(E8-$C$14)^2
9	8	7	2,7	=(C9-$C$14)^2	=$C$18+$C$17*B9	=(C9-E9)^2	=(E9-$C$14)^2
10	9	11	5,5	=(C10-$C$14)^2	=$C$18+$C$17*B10	=(C10-E10)^2	=(E10-$C$14)^2
11	10	13	2,9	=(C11-$C$14)^2	=$C$18+$C$17*B11	=(C11-E11)^2	=(E11-$C$14)^2
12	11	6	10,7	=(C12-$C$14)^2	=$C$18+$C$17*B12	=(C12-E12)^2	=(E12-$C$14)^2
13	сума	**=СУММ(B2:B12)**	**=СУММ(C2:C12)**	=СУММ(D2:D12)		**=СУММ(F2:F12)**	=СУММ(G2:G12)
14	середнє	**=СРЗНАЧ(B2:B12)**	**=СРЗНАЧ(C2:C12)**	**SSt**		**SSE**	**SSR**
15	ст. відх.	=СТАНДОТКЛОН(B2:B12)	=СТАНДОТКЛОН(C2:C12)				
16		r	=КОРРЕЛ(B2:B12;C2:C12)				
17		b1	=C16*C15/B15				
18		b0	=C14-C17*B14				

Рис. 4.3.10. Формули для рохрахунку власноруч

Побудуйте діаграму розсіювання з лінією тренда та її рівнянням, знайдіть відповідність з показниками, які ми розрахували раніше (рис.4.3.11). Інтерпретійте результати.

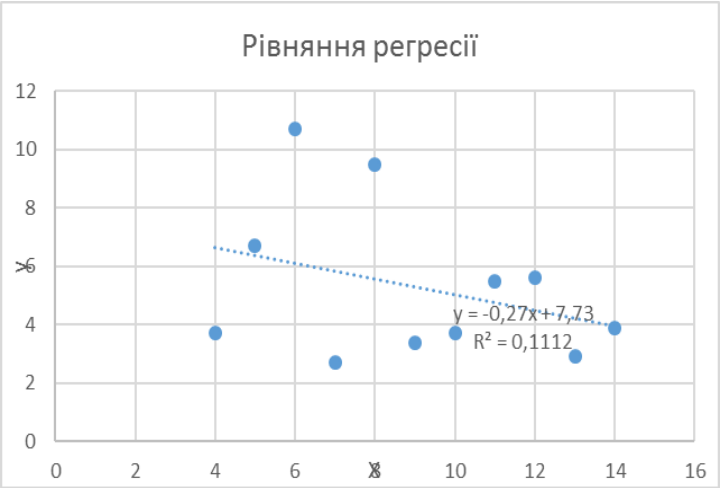


Рис. 4.3.11. Діаграма розсіювання та лінія тренду

Завдання 2. Побудуйте модель простої лінійної регресії для перевірки зв'язку між площею магазинів та обсягом їх продажу на основі даних, які зібрано в 14 магазинах [15].

Магазин	Площа (тис. кв.футів) **X**	Обсяг річних продажів (млн.$) **Y**
1	1,7	3,7
2	1,6	3,9
3	2,8	6,7
4	5,6	9,5
5	1,3	3,4
6	2,2	5,6

7	1,3	3,7
8	1,1	2,7
9	3,2	5,5
10	1,5	2,9
11	5,2	10,7
12	4,6	7,6
13	5,8	11,8
14	3	4,1

Далі показані результати, які ви маєте отримати при рішенні задачі: діаграма розсіювання, розрахунки власноруч, розрахунки за пакетом аналізу, залишки та графіки залишків (рис.4.12-4.17)

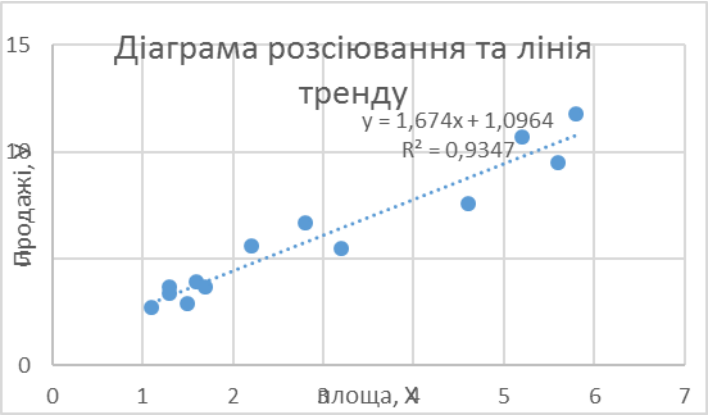


Рис. 4.12. Діаграма розсіювання для прикладу з магазинами

	A	B	C	D	E	F	G	H
2	Магазин	Площа (тис. кв.футів) X	Обсяг річних продажів (млн.$) Y	(Yi-Ycp)^2	Ypr	Y-Ypr	(Y-Ypr)^2	(Ypr-Ycp)^2
3	1	1,7	3,7	4,592	3,803	-0,103	0,011	4,160
4	2	1,6	3,9	3,775	3,636	0,264	0,070	4,869
5	3	2,8	6,7	0,735	5,640	1,060	1,123	0,041
6	4	5,6	9,5	13,375	10,316	-0,816	0,665	20,006
7	5	1,3	3,4	5,968	3,135	0,265	0,070	7,331
8	6	2,2	5,6	0,059	4,638	0,962	0,925	1,451
9	7	1,3	3,7	4,592	3,135	0,565	0,319	7,331
10	8	1,1	2,7	9,878	2,801	-0,101	0,010	9,251
11	9	3,2	5,5	0,118	6,308	-0,808	0,653	0,216
12	10	1,5	2,9	8,660	3,469	-0,569	0,324	5,634
13	11	5,2	10,7	23,592	9,648	1,052	1,107	14,477
14	12	4,6	7,6	3,088	8,646	-1,046	1,094	7,857
15	13	5,8	11,8	35,488	10,650	1,150	1,323	23,105
16	14	3	4,1	3,038	5,974	-1,874	3,512	0,017
17	середнє	2,92	5,84	116,954			11,207	105,748
18	ст.відх.	1,71	3,00	SST			SSE	SSR
19		r	0,951					
20		b1	1,670				r2	0,9042
21		b0	0,964				Syx	0,9664
22								середнє відхилення
23	ВЫВОД ИТОГОВ						$S_{yx} = \sqrt{\frac{SSE}{n-2}}$	від лінії регресії
24								

Рис. 4.13. Розрахунки власноруч

24									
25	*Регрессионная статистика*								
26	Множественный R	0,950883275							
27	R-квадрат	0,904179003							
28	Нормированный R-(	0,89619392							
29	Стандартная ошибк	0,966379679							
30	Наблюдения	14							
31									
32	Дисперсионный анализ								
33		*df*	*SS*	*MS*	*F*	*Значимость F*			
34	Регрессия	1	105,7476095	105,7476095	113,2335128	1,82269E-07			
35	Остаток	12	11,20667621	0,933889684					
36	Итого	13	116,9542857						
37									
38		*Коэффициенты*	*Стандартная ошибк*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
39	Y-пересечение	0,964473659	0,526193302	1,832926522	0,091726832	-0,182003058	2,110950377	-0,182003058	2,110950377
40	Площа (тис. кв.футів	1,669862317	0,156925375	10,64112366	1,82269E-07	1,327951298	2,011773337	1,327951298	2,011773337
41									

Рис. 4.14.Розрахунки у Пакеті аналізу, Регресія

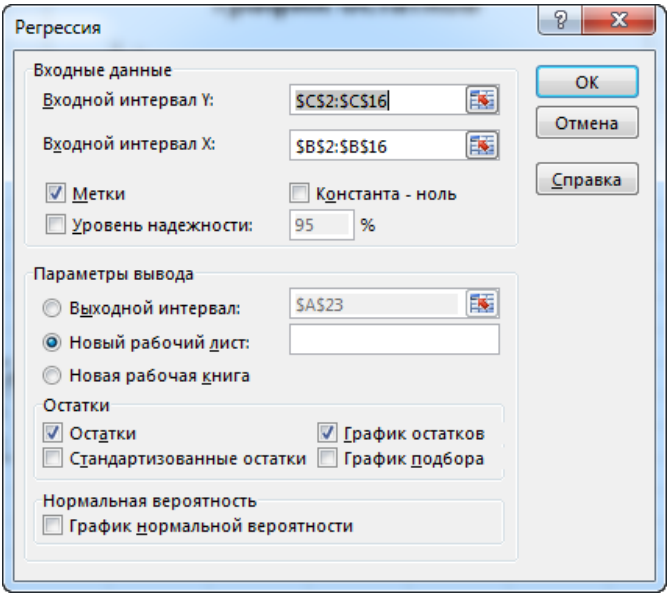


Рис. 4.15. Розрахунок залишків та побудова графіку Залишків

22	ВЫВОД ОСТАТКА		
23			
24	*Наблюдение*	*Предсказанное Обсяг річних продажів (млн.$) Y*	*Остатки*
25	1	3,803239598	-0,103239598
26	2	3,636253367	0,263746633
27	3	5,640088147	1,059911853
28	4	10,31570263	-0,815702635
29	5	3,135294672	0,264705328
30	6	4,638170757	0,961829243
31	7	3,135294672	0,564705328
32	8	2,801322208	-0,101322208
33	9	6,308033074	-0,808033074
34	10	3,469267135	-0,569267135
35	11	9,647757708	1,052242292
36	12	8,645840318	-1,045840318
37	13	10,6496751	1,150324902
38	14	5,974060611	-1,874060611

Рис. 4.16. Залишки

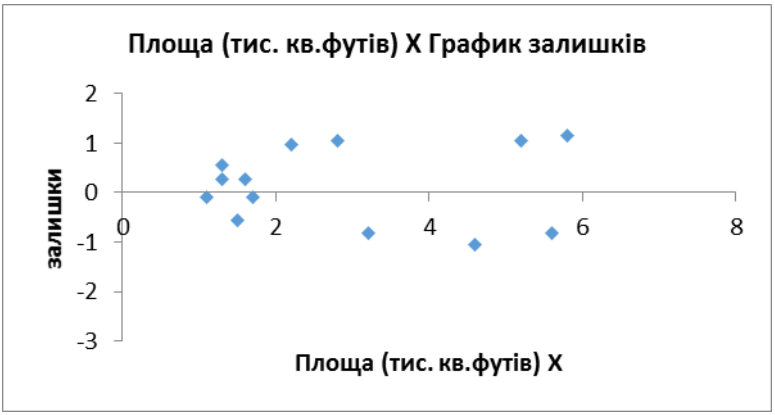


Рис. 4.17. Графік залишків

Інтерпретуйте результати.

Завдання 3. Ознайомтеся зі статтею, що присвячено побудові регресійних моделей міграції*: Р.Ф. Инглхарт, Э.Д. Понарин, М.В. Равлик. Регрессионные модели в оценке факторов международной миграции.* (http://socis.isras.ru/files/File/2014/2014_11/Inglehardt.pdf.)

У дослідженні за допомогою регресійних моделей зіставлені фактори, що мають значущий вплив на приплив мігрантів у приймаючих країнах, а також підтверджено значущий вплив неекономічних чинників (особливо освіти) на міграційні процеси в світі. Дослідження дозволило виявити, що міграція є характерною не тільки важких життєвих обставин, а й для успішності. Автори зібрали воєдино дані ООН, Всесвітньої книги фактів ЦРУ (CIA Factbook), World Values Survey, Freedom House, Світового банку (The World Bank) та отримали базу даних, що містить 179 країн і 28 незалежних змінних, здатних надати значущий вплив на приплив мігрантів.

Практична робота 4.4
Множинна регресія

***Мета роботи*:** надати досвід побудови моделі множинної регресії в Excel та SPSS

Завдання 1.

Скористуємось прикладом, вбудованим в Excel. Припустимо, що забудовник оцінює вартість групи невеликих офісних будівель в традиційному діловому районі. Забудовник може використовувати множинний регресійний аналіз для оцінки ціни офісної будівлі в заданому районі на основі наступних змінних:

X1 – площа будівлі
X2 – число офісів
X3 – число входів під офіс
X4 – термін експлуатації будівлі

Y – оцінна вартість під офіс

Користуючись даними, представленими в таблиці, побудуйте модель множинної регресії, за допомогою функції ЛИНЕЙН.

Таблиця 4.4.1.

Вихідні дані для множинної регресії

x1	x2	x3	x4	y
2310	2	2	20	142 000
2333	2	2	12	144 000
2356	3	1,5	33	151 000
2379	3	2	43	150 000
2402	2	3	53	139 000

2425	4	2	23	169 000
2448	2	1,5	99	126 000
2471	2	2	34	142 900
2494	3	3	23	163 000
2517	4	4	55	169 000
2540	2	3	22	149 000

Скористаємось формулою масиву: виділимо клітки, куди будуть виводитися результати аналізу (A14:E18), введемо наступну формулу, натиснемо CTRL+SHIFT+Enter (рис.4.4.1.)

Формула масиву

=ЛИНЕЙН(E2:E12;A2:D12;ИСТИНА;ИСТИНА)

	A	B	C	D	E
1	**Общая площадь (x1)**	**Количество офисов (x2)**	**Количество входов (x3)**	**Время эксплуатации (x4)**	**Оценочная цена (y)**
2	2310	2	2.0	20	142000
3	2333	2	2.0	12	144000
4	2356	3	1.5	33	151000
5	2379	3	2.0	43	150000
6	2402	2	3.0	53	139000
7	2425	4	2.0	23	169000
8	2448	2	1.5	99	126000
9	2471	2	2.0	34	142900
10	2494	3	3.0	23	163000
11	2517	4	4.0	55	169000
12	2540	2	3.0	22	149000
13	**Формула**				
14	-234.24	2553.21	12529.77	27.64	52317.83
15	13.27	530.67	400.07	5.43	12237.36
16	0.997	970.58	#Н/Д	#Н/Д	#Н/Д
17	459.75	6.00	#Н/Д	#Н/Д	#Н/Д
18	1732393319.23	5652135.32	#Н/Д	#Н/Д	#Н/Д

.

Рис.4.4.1. Результати виконання функції ЛИНЕЙН

Ознайомтеся з поясненням, щодо виводу, яке міститься у довідці Excel. Наводимо її.

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	$ss_{рег.}$	$ss_{ост.}$				

Рис.4.4.2. Пояснення виводу функції ЛИНЕЙН

Рівняння множинної регресії y = m1*x1 + m2*x2 + m3*x3 + m4*x4 + b тепер може бути отримано з рядку 14:

y = 27,64*x1 + 12530*x2 + 2553*x3 - 234,24*x4 + 52318

Функция ЛИНЕЙН
ЛИНЕЙН(известные_значения_у;известные_значения_х; конст; статистика)

Конст — логическое значение, которое указывает, требуется ли, чтобы константа b0 была равна 0.

- Если конст имеет значение ИСТИНА или опущено, то b0 вычисляется обычным образом.
- Если аргумент конст имеет значение ЛОЖЬ, то b0 полагается равным 0
- Статистика — логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии.

- Если аргумент статистика имеет значение ИСТИНА, то функция ЛИНЕЙН возвращает дополнительную регрессионную статистику, так что возвращаемый массив будет иметь вид: Если аргумент статистика имеет значение ЛОЖЬ или опущен, то функция ЛИНЕЙН возвращает только коэффициенты m и постоянную b.

Дополнительная регрессионная статистика:

Величина	Описание
se1,se2,...,sen	Стандартные значения ошибок для коэффициентов m1,m2,...,mn.
seb	Стандартное значение ошибки для постоянной b (seb = #Н/Д, если конст имеет значение ЛОЖЬ).
r2	Коэффициент детерминированности.. Если он равен 1, то имеет место полная корреляция с моделью, т. е. нет различия между фактическим и оценочным значениями y. В противоположном случае, если коэффициент детерминированности равен 0, то уравнение регрессии неудачно для предсказания значений y.
sey	Стандартная ошибка для оценки y.
F	F-статистика, или F-наблюдаемое значение. F-статистика используется для определения того, является ли наблюдаемая взаимосвязь между зависимой и независимой переменными случайной или нет.
df	Степени свободы. Степени свободы полезны для нахождения F-критических значений в статистической таблице. Для определения уровня надежности модели нужно сравнить значения в таблице с F-статистикой, возвращаемой функцией ЛИНЕЙН.
ssreg	Регрессионная сумма квадратов.
ssresid	Остаточная сумма квадратов.

Інтерпретація коефіцієнтів рівняння множинної регресії

Для інтерпретації напрямків зв’язку між змінними дивляться на знаки (плюс або мінус) регресійних коефіцієнтів чи B-коефіцієнтів:

- якщо B-коефіцієнт позитивний, то зв'язок цієї змінної з залежною змінною позитивний (наприклад, чим більше площа будівлі, тим вище його ціна);

- якщо B-коефіцієнт від'ємний, то і зв'язок носить негативний характер (наприклад, чим менший термін експлуатація будівлі, тим вище його ціна).
- якщо B-коефіцієнт дорівнює 0, зв'язок між змінними відсутній.

Рішення задачі в SPSS

Хід роботи

1. Створіть відповідний файл SPSS.
2. Оберіть ***Analyze, Regression, Linear*** (рис. 4.4.3)

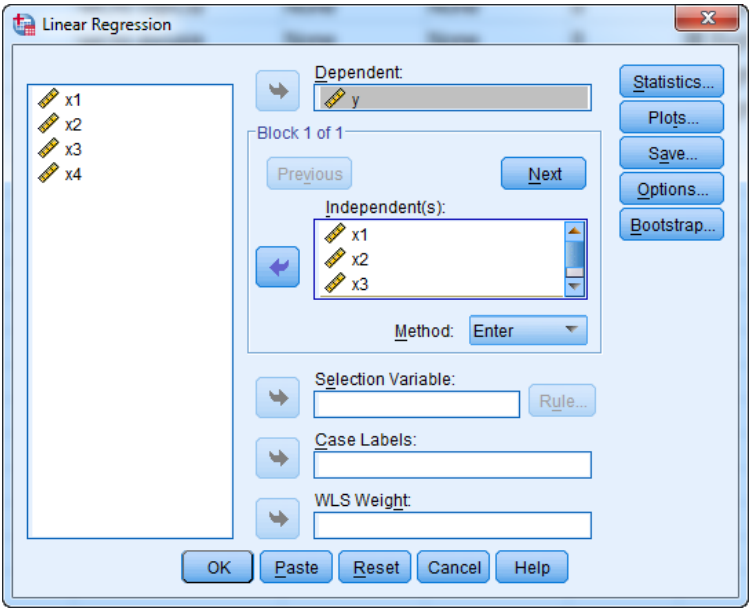


Рис. 4.4.3. Вікно «Linear Regression»

3. Передивіться результати розрахунків (рис. 4.4.4), знайдіть та інтерпретуйте R2, колонку з коефіцієнтами рівняння регресії, їх значущість. Запишіть отримане рівняння регресії.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,998[a]	,997	,996	792,474

a. Predictors: (Constant), срок эксплуатации здания, число входов под офис, число офисс площадь здания

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52317,8	7377,569		7,091	,000
	площадь здания	27,641	3,148	,179	8,780	,000
	число офисов	12529,8	299,776	,784	41,797	,000
	число входов под офис	2553,211	301,483	,172	8,469	,000
	срок эксплуатации здания	-234,237	9,827	-,441	-23,837	,000

a. Dependent Variable: оценочная цена под офис

Рис.4.4.4. Результати розрахунків

Рішення задачі в R

Хід роботи

1. Подготувати дані в Excel

	A	B	C	D	E
1	**x1**	**x2**	**x3**	**x4**	**y**
2	2310	2	2	20	142000
3	2333	2	2	12	144000
4	2356	3	1,5	33	151000
5	2379	3	2	43	150000
6	2402	2	3	53	139000
7	2425	4	2	23	169000
8	2448	2	1,5	99	126000
9	2471	2	2	34	142900
10	2494	3	3	23	163000
11	2517	4	4	55	169000
12	2540	2	3	22	149000
13					

home / Лист2 / Лист3

В RSTUDIO:

2. Прочитати дані:

```
data<-read.table("home.csv", sep=";", dec=",", header=TRUE)
> data
```

```
    x1    x2 x3 x4      y
1 2310  2 2.0 20 142000
2 2333  2 2.0 12 144000
3 2356  3 1.5 33 151000
4 2379  3 2.0 43 150000
```

5 2402 2 3.0 53 139000
6 2425 4 2.0 23 169000
7 2448 2 1.5 99 126000
8 2471 2 2.0 34 142900
9 2494 3 3.0 23 163000
10 2517 4 4.0 55 169000
11 2540 2 3.0 22 149000

3. Побудувати модель множинної регресії:

```
> fit <- lm(data$y ~ data$x1+ data$x2+ data$x3+ data$x4,
data=data)
> summary( fit)
```

```
Call:
lm(formula = data$y ~ data$x1 + data$x2 + data$x3 +
data$x4,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1092.9  -263.1  -130.2   197.3  1883.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52317.831  12237.362   4.275  0.00523
**
data$x1        27.641      5.429   5.091  0.00224
**
data$x2     12529.768    400.067  31.319 7.04e-08
***
data$x3      2553.211    530.669   4.811  0.00297
**
data$x4      -234.237     13.268 -17.654 2.12e-06
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

Residual standard error: 970.6 on 6 degrees of
freedom
Multiple R-squared:  0.9967,   Adjusted R-squared:
0.9946
```

```
F-statistic: 459.8 on 4 and 6 DF,  p-value: 1.372e-07
```

4. Порівняти результати з Excel, SPSS.

Завдання для самостійної роботи

1. Побудувати регресійну модель для збуту продукції цукеркової фабрики, яка пов'язує його з такими факторами як період збуту, витрати на рекламу, ціна, ціна конкурента та індекс споживача.

y	**x1**	**x2**	**x3**	**x4**	**x5**
Збут	**Період**	**Реклама**	**Ціна**	**Ціна конкурента**	**Індекс споживача**
126	1	4	15	17	100
137	2	4,8	14,8	17,3	98,4
148	3	3,8	15,2	16,8	101,2
191	4	8,7	15,5	16,2	103,5
274	5	8,2	15,5	16	104,1
370	6	9,7	16	18	107
432	7	14,7	18,1	20,2	107,4
445	8	18,7	13	15,8	108,5
367	9	19,8	15,8	18,2	108,3
367	10	10,6	16,9	16,8	109,2
321	11	8,6	16,3	17	110,1
307	12	6,5	16,1	18,3	110,7
331	13	12,6	15,4	16,4	110,3
345	14	6,5	15,7	16,2	111,8
364	15	5,8	16	17,7	112,3
384	16	5,7	15,1	16,2	112,9

Записати рівняння регресії. Визначити значимість коефіцієнтів рівняння регресії, при необхідності виключити з розгляду змінні, які слабо впливають на збут продукції, і побудувати нове рівняння.

Знайти залишки (різницю між фактичними значеннями збуту і передбаченими за рівнянням), побудувати графік розподілу залишків.

Завдання вирішити в SPSS, а потім в Excel та порівняти отримані рішення.

При вирішенні завдання в Excel:

1) скористатися формулою масиву і функцією ЛИНЕЙН;
2) скористатися Пакетом аналізу (регресія).

2. Квартет Енскомба – чотири набори числових даних, у яких прості статистичні властивості ідентичні, але їх графіки істотно відрізняються. Кожен набір складається з 11 пар чисел. Квартет був складений в 1973 році англійським математиком Ф. Дж. Енскомбом для ілюстрації важливості застосування графіків для статистичного аналізу і впливу викидів значень на властивості всього набору даних.

Всі набори мають приблизно такі властивості:

Характеристика	**Значення**
Середнє значення *x*	9,0
Дисперсія змінної x	10,0
Середнє значення y	7,5
Дисперсія змінної y	3,75
Кореляція між змінними x и y	0,816
Пряма лінійної регресії	Y = 3 + 0,5x

Розрахуйте перераховані статистики для усіх наборів даних.

Набор А		Набор Б		Набор В		Набор Г	
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8,04	10,0	9,1	10	7,46	8	6,58
14	9,96	14,0	8,1	14	8,84	8	5,76
5	5,68	5,0	4,7	5	5,73	8	7,71
8	6,95	8,0	8,1	8	6,77	8	8,84
9	8,81	9,0	8,8	9	7,11	8	8,47

12	10,84	12,0	9,1	12	8,15	8	7,04
4	4,26	4,0	3,1	4	5,39	8	5,25
7	4,82	7,0	7,3	7	6,42	19	12,5
11	8,33	11,0	9,3	11	7,81	8	5,56
13	7,58	13,0	8,7	13	12,74	8	7,91
6	7,24	6,0	6,1	6	6,08	8	6,89

Побудуйте 4 діаграми розсіювання, поясніть отримані результати.

Контрольні питання

1. Як здійснюється множинна регресія в Excel?
2. Як здійснюється множинна регресія в SPSS?
3. Як обчислюються залишки?
4. Як побудувати графік залишків? Для чого він будується?
5. Призначення функції ЛИНЕЙН?
6. Що таке формула масиву?
7. Які фактори залишилися в остаточному рівнянні регресії для цукеркової фабрики?
8. Що собою являє квартет Енскомба?

РОЗДІЛ 5
ОСНОВИ ТЕОРІЇ ІМОВІРНОСТІ

Імовірність – це кількісне вимірювання того факту, що подія відбудеться. Імовірність будь-якої події знаходиться між 0 та 1. **Неможливою** вважається подія, яка не може відбутися (імовірність = 0). **Достовірною** вважаємо подію, яка відбудеться обов'язково (імовірність = 1).

Існує 3 підходи до визначення імовірності. В **класичному підході і**мовірність оцінюється на основі апріорної інформації. В **емпіричному** підході імовірність оцінюється на основі апостеріорної інформації. **Суб'єктивний** підхід до імовірності – це шанс, який приписуються події конкретною особою, на основі особистого досвіду, суспільної думки або аналізу конкретної ситуації.

Кожен можливий результат випадкової змінної називаємо подією. **Елементарні** події описуються однією характеристикою (стать респондента). **Сумісна** подія описана двома або більше характеристиками (молоді жінки, чоловіки середнього віку).

Вибірковим простором називається набір усіх подій.

Події можна візуалізувати у вигляді таблиць **контингенції** або **дерев рішень**.

Подія	**Подія**		Разом
	B1	**B2**	
A1	P(A1 & B1)	P(A1 & B2)	**P(A1)**
A2	P(A2 & B1)	P(A2 & B2)	**P(A2)**
Разом	**P(B1)**	**P(B2)**	1

Рис. 5. Таблиця контингенції

Безумовна імовірність – це імовірність елементарної події. Імовірність двох або більше подій називається **сумісною** імовірністю, або імовірністю сумісних подій. **Несумісні** події або взаємно-виключні події не можуть відбутися одночасно. **Повна група** подій - обов'язково відбувається хоча б одна з подій, які складають групу. Сума імовірностей несумісних подій та повної групи подій дорівнює 1.

Умовна імовірність - імовірність однієї події при умові, що друга подія відбулася. Імовірність події А при умові що В відбулася обчислюється формулою:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \qquad (5.1)$$, де

де P(A and B) = сумісна імовірність А та В

P(A) = безумовна імовірність А

P(B) = безумовна імовірність В

Дві події **незалежні** якщо і тільки якщо, імовірність першої події не залежить від імовірності другої події.

$$P(A \mid B) = P(A) \qquad (5.2)$$

Для подій А та В

$$P(A \text{ and } B) = P(A \mid B)P(B) \qquad (5.3.)$$

Для **незалежних** подій

$$P(A \text{ and } B) = P(A)P(B) \qquad (5.4)$$

Загальне правило **додавання** імовірностей

P (A or B) = P(A) + P(B) - P(A and B) (5.5)

Для несумісних подій

P (A or B) = P(A) + P(B) (5.6)

Практична 5.1
Обчислення імовірностей

Мета роботи: надати досвід обчислювання імовірностей засобами Excel

Хід роботи

Підготуйте лист Excel, як показано на рис. 5.1.1 [15]. Введіть мітки й кількість виходів (результатів), які сприяють здійсненню події для обох подій А та В.

При необхідності мітки А1, А2, В1 і В2 в блоках В5: В6 і С4: D4 можна замінити.

В клітинках С7-С8, Е5-Е7 – формули.

	A	B	C	D	E
1	**Обчислення імовірностей**				
2					
3	**Вибірковий простір**		**Подія В**		
4			**В1**	**В2**	Разом
5	**Подія А**	**А1**	**1**	**1**	2
6		**А2**	**1**	**1**	2
7		Разом	2	2	4
8					
9	**Безумовні імовірності**				
10	Р(А1)				
11	Р(А2)				
12	Р(В1)				
13	Р(В2)				
14					
15	**Сумісні імовірності**				
16	Р(А1 та В1)				
17	Р(А1 та В2)				
18	Р(А2 та В1)				
19	Р(А2 та В2)				
20					
21	**Правило додавання**				
22	Р(А1 або В1)				
23	Р(А1 або В2)				
24	Р(А2 або В1)				
25	Р(А2 або В2)				

Рис. 5.1.1. Вихідні дані для розрахунків

В клітинці А10 запишіть формулу: **=”Р(“&B5&”)”**
Який результат її виконання?

Напишіть аналогічні формули для А11-А13, А16-А19, А22-25 (рис.5.1.2).

	A	B	C	D	E
1	**Обчислення імовірностей**				
2					
3	**Вибірковий простір**		**Подія В**		
4			**В1**	**В2**	Разом
5	**Подія А**	**А1**	**1**	**1**	=C5+D5
6		**А2**	**1**	**1**	=C6+D6
7		Разом	=C5+C6	=D5+D6	=C7+D7
8					
9	**Безумовні імовірності**				
10	="Р("&B5&")"				
11	="Р("&B6&")"				
12	="Р("&C4&")"				
13	="Р("&D4&")"				
14					
15	**Сумісні імовірності**				
16	="Р("&B5&" та "&C4&")"				
17	="Р("&B5&" та "&D4&")"				
18	="Р("&B6&" та "&C4&")"				
19	="Р("&B6&" та "&D4&")"				
20					
21	**Правило додавання**				
22	="Р("&B5&" або "&C4&")"				
23	="Р("&B5&" або "&D4&")"				
24	="Р("&B6&" або "&C4&")"				
25	="Р("&B6&" або "&D4&")"				
26					

Рис.5.1.2. Формули для розрахунків

Введіть також формули для обчислення безумовної імовірності, сумісної імовірності, та правила додавання. Маєте отримати наступні результати:

	A	B	C	D	E
4			B1	B2	Разом
5	Подія А	A1	1	1	2
6		A2	1	1	2
7		Разом	2	2	4
8					
9	Безумовні імовірності				
10	P(A1)	0,50			
11	P(A2)	0,50			
12	P(B1)	0,50			
13	P(B2)	0,50			
14					
15	Сумісні імовірності				
16	P(A1 и B1)	0,25			
17	P(A1 и B2)	0,25			
18	P(A2 и B1)	0,25			
19	P(A2 и B2)	0,25			
20					
21	Правило додавання				
22	P(A1 или B1)	0,75			
23	P(A1 или B2)	0,75			
24	P(A2 или B1)	0,75			
25	P(A2 или B2)	0,75			
26					

Рис. 5.1.3. Результати обчислення для умовної таблиці

Завдання 1. Скористайтеся цією табличкою для обчислення імовірностей наступного прикладу:

	A	B	C	D	E
1	Обчислення імовірностей				
2					
3	Вибірковий простір		Подія B: Кольор		
4			Червоний	Чорний	Разом
5	Подія A: карта	Туз	2	2	4
6		Не туз	24	24	48
7		Разом	26	26	52
8					

Рис. 5.1.4. Вихідні дані

Перевірте результати:

8		
9	**Безумовні імовірності**	
10	P(Туз)	0,08
11	P(Не туз)	0,92
12	P(Червоний)	0,50
13	P(Чорний)	0,50
14		
15	**Сумісні імовірності**	
16	P(Туз і Червоний)	0,04
17	P(Туз і Чорний)	0,04
18	P(Не туз і Червоний)	0,46
19	P(Не туз і Чорний)	0,46
20		
21	**Правило додавання**	
22	P(Туз або Червоний)	0,54
23	P(Туз або Чорний)	0,54
24	P(Не туз або Червоний)	0,96
25	P(Не туз або Чорний)	0,96
26		

Рис. 5.1 5. Результати розрахунків для приклада з картами

Завдання 2. Чи можна стверджувати, що білі робочі частіше заявляють про несправедливе звільнення? Опитування показало [15], що з 56 звільнених білих, 29 вважали це несправедливим. А з 407 звільнених чорних, заявили про несправедливість 126.

1. Створіть таблицю спряженості 2X2, та оцініть імовірність різних подій.
2. Наведіть приклад елементарної події.
3. Наведіть приклад сумісної події.
4. Яка подія є додатковою до «робочий заявив про несправедливе звільнення»?

5. Обчисліть імовірність для випадковим чином обраного робітника:
6. Він – білий.
7. Він вважає своє звільнення несправедливим.
8. Він – білий і він вважає своє звільнення несправедливим.
9. Він – чорний і він не вважає своє звільнення несправедливим.
10. Він – білий або він вважає своє звільнення несправедливим.
11. Він – чорний або він вважає своє звільнення несправедливим.
12. Поясніть різницю між 8 питанням та 10.

Підготуйте таблицю, занесіть дані, розрахуйте:

	A	B	C	D
1				
2		Подія В		
3	Подія А	**Білий**	**Чорний**	**Разом**
4	**Справедливе звільнення**			
5	**Несправедливе звільнення**			
6	**Разом**			
7				
8	**Р(білий)**			
9	**Р(несправедливе)**			
10	**Р(білий і несправедливе)**			
11	**Р(чорний і справедливе)**			
12	**Р(білий або несправедливе)**			
13	**Р(чорний або несправедливе)**			

Контрольні питання

1. Назвіть три підходи до визначення імовірності.
2. Яким чином обчислюється імовірність?
3. В яких межах вона змінюється?
4. Наведіть приклади неможливої події, достовірної події.
5. Що таке елементарна подія? Приклади.
6. Що таке сумісна подія. Приклади.
7. Що таке додаткова подія? Приклади.
8. Яким чином можна представити події в таблиці спряженості?
9. Яким чином можна представити події у дереві рішень?
10. Як визначаються несумісні події? Приклад.
11. Правила додавання імовірностей.
12. Умовна імовірність.
13. Як визначаються незалежні події?

Практична 5.2.
Обчислення імовірності за Баєсом

***Мета роботи*:** надати досвід обчислення імовірності за Баєсом

Короткі теоретичні відомості.

Теорема Баєса використовується, щоб уточнити попередньо обчислену інформацію на основі новій інформації. Вона є розширенням умовної імовірності.

Уточнена імовірність події ***Bi*** при умові, що відбулася подія ***A*** обчислюється за формулою

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)}$$

(5.2.1)

де

B – подія, яка складається з k взаємно-виключних подій

A – нова подія, яка може впливати на $P(B_i)$

Завдання 1. Бурова компанія [15] оцінила 40% шанс знайти нафту в своїй новій свердловині. Детальний тест було призначено для отримання додаткової інформації. Історично 60% успішних свердловин проходять детальні випробування, та 20% безуспішних свердловин те ж проходять детальні випробування. З огляду на деталізований тест, яка ймовірність того, що нафту буде знайдено в новій свердловині?

Хай S = успішна подія (successful), U = неуспішна (unsuccessful).

P(S) = 0.4, P(U) = 0.6 (апріорні імовірності)

Визначимо деталізований тест як подію D.

Умовні імовірності: P(D|S) = 0.6; P(D|U) = 0.2

Мета – знайти P(S|D).

За формулою Баєса отримаємо:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D|S)P(S) + P(D|U)P(U)}$$

$$= \frac{(0.6)(0.4)}{(0.6)(0.4) + (0.2)(0.6)}$$

$$= \frac{0.24}{0.24 + 0.12} = 0.667$$

Вирішить завдання в Excel, представив обчислення наступним чином (рис. 5.2.1):

	A	B	C	D	E	F	G	H
1	Баєса теорема							
2								
3			Імовірність					
4	Подія	Апріорна		Умовна		Сумісна		Уточнена
5	S успішна	**P(S)**	0,4	**P(D|S)**	0,6	**P(D|S)*P(S)**	0,24	0,667
6	U неуспіш	**P(U)**	0,6	**P(D|U)**	0,2	**P(D|U)*P(U)**	0,12	0,333
7					Разом		0,36	
8								

Рис. 5.2.1. Рішення завдання 1

Уточнена імовірність 0,667 відрізняється від начальної 0,4.

Завдання 2. Імовірність того, що людина страждає від певної хвороби, дорівнює 0,03 [15]. Медичний тест дозволяє це перевірити. Якщо людина дійсно хворіє, імовірність точного діагнозу (який стверджує, що людина дійсно хворіє) дорівнює 0,9. Якщо людина здорова, імовірність хибно позитивного діагнозу (який стверджує, що людина хворіє, якщо вона здорова) дорівнює 0,02. Припустимо, що тест дав позитивний результат. Яка імовірність того, що людина хворіє? Яка імовірність точного діагнозу?

Введемо означення:

D – людина хворіє; **D’** – здорова

T – діагноз позитивний; **T’** – діагноз негативний

Задайте дані та проведіть обчислення в Excel (рис. 5.2.2).

	A	B	C	D	E	F	G	H	I
1	Баєса теорема								
2									
3			Імовірність						
4	Подія Di	Апріорна P(Di)		Умовна P(T|Di)		Сумісна P(T|Di)P(Di)		Уточнена P(Di|T)	
5	D хворий	**P(D)**	0,03	**P(T|D)**	0,9	**P(T|D)*P(D)**	0,0270	**P(D|T)**	0,582
6	D' здоров	**P(D')**	0,97	**P(T|D')**	0,02	**P(T|D')*P(D')**	0,0194	**P(D'|T)**	0,418
7					Разом		0,046		
8									

Рис. 5.2.2. Рішення завдання 2

Надайте інтерпретацію отриманим результатам.

Завдання 3. Рекламна агенція вивчила уподобання одружених чоловіків, які дивляться телевізор в кращий ефірний час [15]. З’ясували, що чоловіки дивляться телевізор 60% ефірного часу. Далі, з’ясувалося, що коли чоловіки дивляться телепередачі, 40% часу жінки теж дивляться з ними. Якщо ж чоловік не дивиться, дружина дивиться 30% часу. Визначте імовірність подій: 1) коли дружина дивиться, чоловік також дивиться 2) жінка дивиться телевізор в кращий ефірний час.

Задайте дані та проведіть обчислення в Excel.

	A	B	C	D	E	F	G	H	I
1	Баєса теорема								
2									
3			Імовірність						
4	Подія Di	Апріорна P(Di)		Умовна P(T|Di)		Сумісна P(T|Di)P(Di)		Уточнена P(Di|T)	
5	D дивитьс	**P(D)**	0,6	**P(T|D)**	0,4	**P(T|D)*P(D)**	0,2400	**P(D|T)**	0,667
6	D'чоловік	**P(D')**	0,4	**P(T|D')**	0,3	**P(T|D')*P(D')**	0,1200	**P(D'|T)**	0,333
7					Разом		0,36		

Рис. 5.2.3. Рішення завдання 3

Контрольні питання

1. Дайте визначення умовній імовірності. Як вона розраховується.
2. Сформулюйте теорему Баєса. Для чого вона використовується?

Практична 5.3
Побудова нормальної кривої

***Мета роботи*:** надати досвід побудови графіку функції щільності імовірності для нормального розподілу засобами Excel

***Короткі* теоретичні відомості**

Математичний вираз, який описує розподіл значень випадкової неперервної величини називається функцією щільності імовірності. Для нормального розподілу маємо:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2} \qquad (5.3.1)$$

де e – 2.71828

π – 3.14159

μ – середнє генеральної сукупності

σ – стандартне відхилення генеральної сукупності

X – будь-яке значення випадкової змінної

Для **стандартизованого нормального розподілу** формула щільності імовірності наступна:

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)Z^2} \qquad (5.3.2)$$

де e – 2.71828

π – 3.14159

Z – будь-яке значення стандартизованого нормального розподілу.

Завдання 1. **Побудова одиничної нормальної кривої**

1. Занесіть дані в стовбець Z.

Z	**F(Z)**
-4	
-3,5	
-3	
-2,5	
-2	
-1,5	
-1	
-0,5	
0	
0,5	
1	
1,5	
2	
2,5	
3	
3,5	
4	

2. Обчисліть f(Z) за формулою:

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)Z^2}$$

Z	F(z)
-4	0,000134
-3,5	0,000873
-3	0,004433
-2,5	0,017533
-2	0,054005
-1,5	0,12955
-1	0,242032
-0,5	0,352155
0	0,399043
0,5	0,352155
1	0,242032
1,5	0,12955
2	0,054005
2,5	0,017533
3	0,004433
3,5	0,000873
4	0,000134

Зауваження. Формула для f(Z) має вид =1/(2*3,14)^0,5*EXP(-0,5*A4^2), якщо Z знаходиться в клітинці A4.

Побудуйте графік одиничної нормальної кривої (рис. 5.3.1)

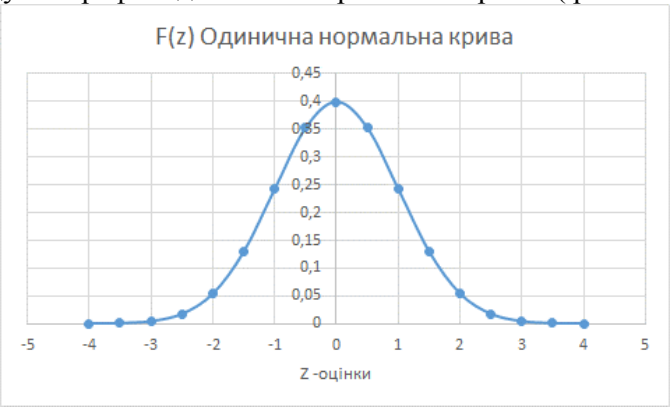


Рис. 5.3.1. Одинична нормальна крива

Завдання 2. Побудувати графік щільності імовірності за наступними даними з урахуванням того, що середнє генеральної сукупності для коефіцієнта IQ дорівнює 100, а стандартне відхилення – 15.

IQ
X
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145

Розрахуйте f(x) за формулою функції щільності імовірності наступним чином:

	A	B	C	D	E	F
1	IQ			$f(X)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$	σ	15
2	X	1/(корень(2*pi)*σ)	Z=(X-μ)/σ	F(X)=	μ	100
3	55	0,027	-3,00	0,00030		
4	60	0,027	-2,67	0,00076		
5	65	0,027	-2,33	0,00175		
6	70	0,027	-2,00	0,00360		
7	75	0,027	-1,67	0,00663		
8	80	0,027	-1,33	0,01094		
9	85	0,027	-1,00	0,01614		
10	90	0,027	-0,67	0,02130		
11	95	0,027	-0,33	0,02517		
12	100	0,027	0,00	0,02660		
13	105	0,027	0,33	0,02517		
14	110	0,027	0,67	0,02130		
15	115	0,027	1,00	0,01614		
16	120	0,027	1,33	0,01094		
17	125	0,027	1,67	0,00663		
18	130	0,027	2,00	0,00360		
19	135	0,027	2,33	0,00175		
20	140	0,027	2,67	0,00076		
21	145	0,027	3,00	0,00030		

Рис. 5.3.2. Результати розрахунків

Побудуйте графік (рис.5.3.3):

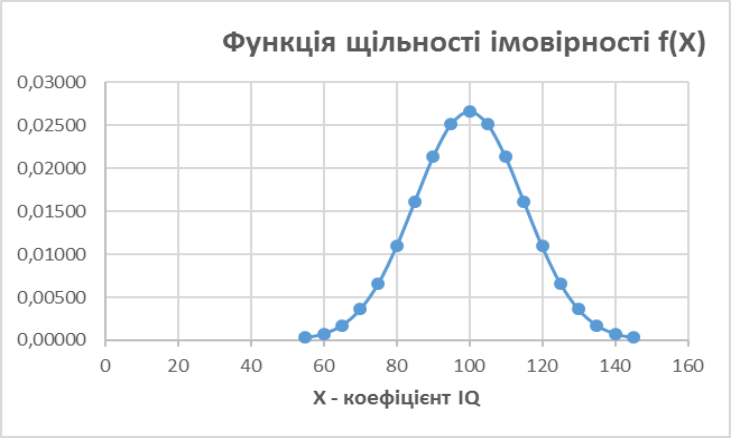


Рис. 5.3.3 . Графік нормальної кривої

Контрольні запитання

1. Як побудувати графік функції щільності імовірності в Excel.

2. Чим стандартизований нормальний розподіл відрізняється від нестандартизованого нормального розподілу?

Завдання для самостійної роботи

Побудуйте одиничну нормальну криву в SPSS

1. Скористуйтеся редактором синтаксису:

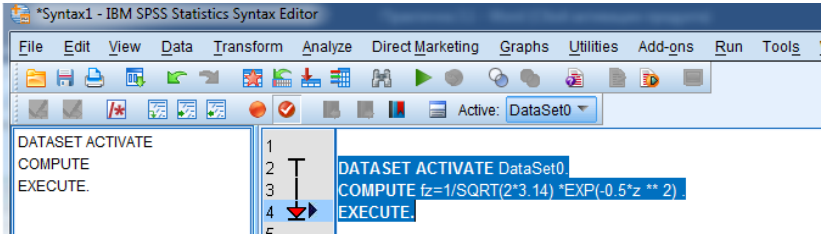


Рис. 5.3.4.Обчислення f(z) в редакторі синтаксису

	z	fz
1	-4,00	,00
2	-3,50	,00
3	-3,00	,00
4	-2,50	,02
5	-2,00	,05
6	-1,50	,13
7	-1,00	,24
8	-,50	,35
9	,00	,40
10	,50	,35
11	1,00	,24
12	1,50	,13
13	2,00	,05
14	2,50	,02
15	3,00	,00
16	3,50	,00
17	4,00	,00

Рис. 5.3.5. Результати обчислення f(z)

1. Побудуйте два графіки (рис.5.3.6-5.3.7):

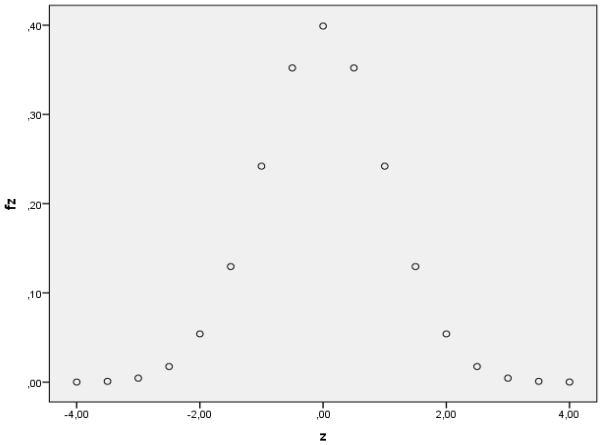


Рис.5.3.6. Точкова діаграма

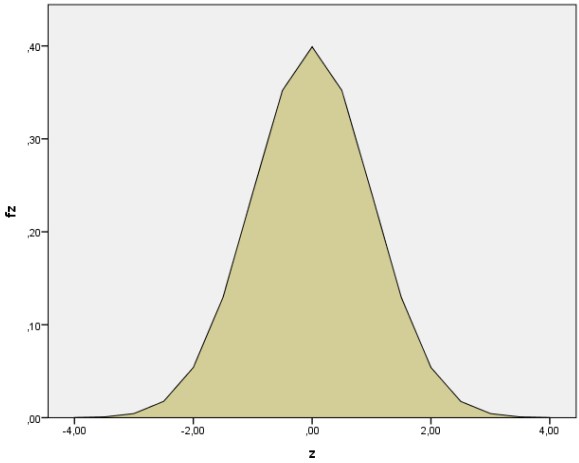


Рис. 5.3.7.Діаграма Area

Практична 5.4
Розрахунок імовірностей нормального розподілу

***Мета*:** Розраховувати імовірності для нормального розподілу засобами електронних таблиць

Постановка задачі

Припустимо ми маємо нормальний розподіл щодо коефіцієнтів IQ дорослих людей з середнім 100 та стандартним відхиленням 15. Знайдіть імовірність, що для випадкової людини його коефіцієнт IQ знаходиться: а) в межах від 85 до 115 балів; б) нижче ніж 55; в) вище 145.

Хід роботи

1. Для цього розробимо наступну таблицю таким чином, щоб вона була пригідна для розрахунків для розподілів з різними середніми та стандартними відхиленнями, та різними інтервалами (рис.5.4.1-5.4.2)[16]

	A	B	C	D	E
1	Нормальний розподіл				
2					
3	Дані розподілу				
4	Середнє	100			
5	Стандартне відхилення	15			
6				Імовірність для інтервалу	
7	Імовірність для X <=			Від X	85
8	X	55		до Y	115
9	Z			Z для 85	
10	P(X<=55)			Z для 115	
11				P(X<=85)	
12	Імовірність для X >			P(X<=115)	
13	X	145		P(85<=X<=115)	
14	Z				
15	P(X>145)			левним значенням кумулятивного від	
16				Cumulative Percentage	10,00%
17	Імовірність для X<55 or X >145			Z	
18	P(X<55 or X >145)			X	
19					

Рис. 5.4.1. Вихідні дані

	A	B	C	D	E
1	Нормальний розподіл				
2					
3	Дані розподілу				
4	Середнє	100			
5	Стандартне відхилення	15			
6				Імовірність для інтервалу	
7	Імовірність для X <=			Від X	85
8	X	55		до Y	115
9	Z			="Z для " &E7	
10	="P(X<="&B8&")"			="Z для " & E8	
11				="P(X<="&E7&")"	
12	Імовірність для X >			="P(X<="&E8&")"	
13	X	145		="P("&E7&"<=X<="&E8&")"	
14	Z				
15	="P(X>"&B13&")"			найти X та Z за значенням кумулятивного відсотку	
16				Cumulative Percentage	0,1
17	="Імовірність для X<"&B8&" or X >"&B13			Z	
18	="P(X<"&B8&" or X >"&B13 & ")"			X	
19					

Рис. 5.4.2. Формули для вихідних даних

2. Розрахуємо в клітинках B9, B14, E9-E10 всі Z-оцінки за допомогою функції нормалізація, яка для даного X, середнього та стандартного відхилення розраховує його Z-оцінку (стандартизоване або нормалізоване значення).

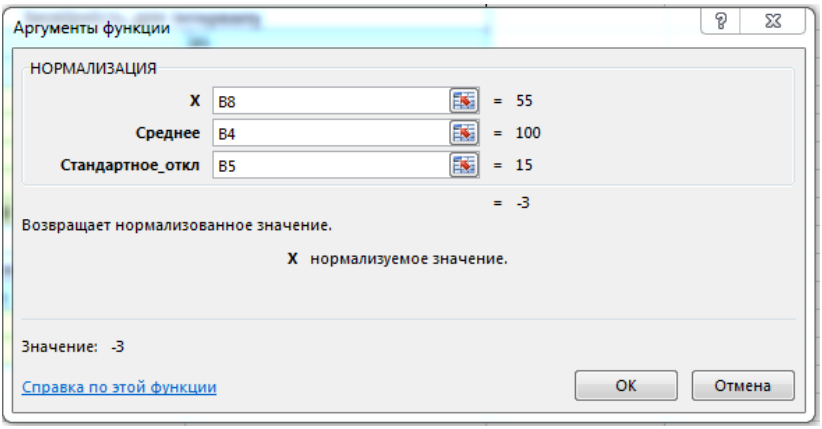


Рис. 5.4.3. Функція нормалізації

3. Розрахуємо в клітинках B10, B15, E11-E13 всі імовірності P за допомогою функції НОРМРАСП

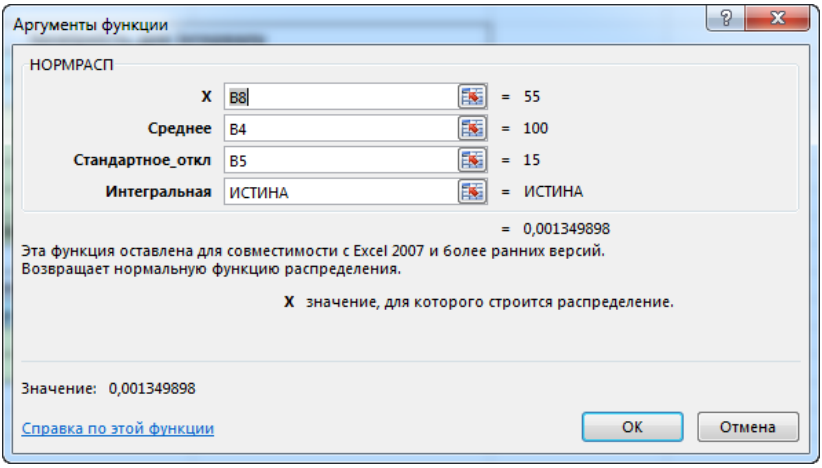


Рис. 5.4.4. Функція НОРМРАСП

Ця функція для даного Z повертає долю площі під нормальною кривою, яка лежить нижче Z (рис. 5.4.5.).

Рис. 5.4.5. Доля площі під нормальною кривою, яка лежить нижче Z=2

4. Для даного кумулятивного відсотка площі знайдемо X та Z. Використовуйте функцію НОРМСТОБР для знаходження Z та функцію НОРМОБР для знаходження X.

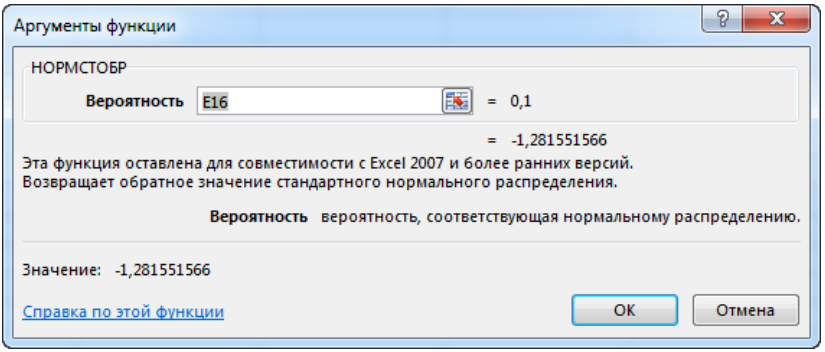


Рис. 5.4.6. Функція НОРМСТОБР

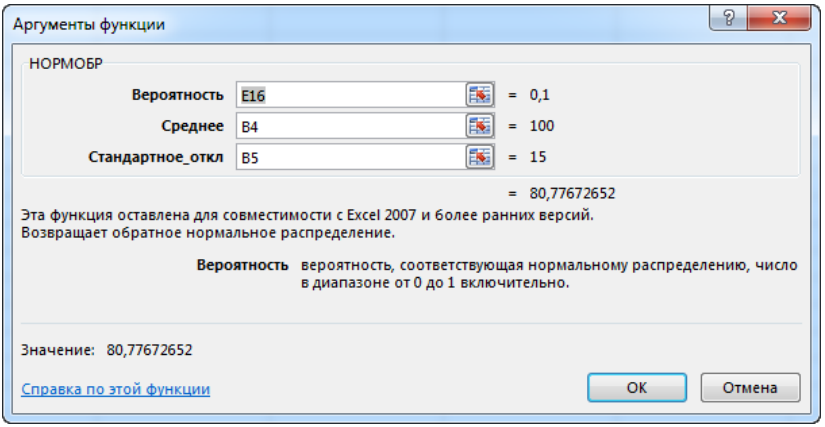


Рис. 5.4.7. Функція НОРМОБР

5. Перевірте результати всіх розрахунків та надайте їх інтерпретацію (рис.5.4.8.)

	A	B	C	D	E
1	Нормальний розподіл				
2					
3	Дані розподілу				
4	Середнє	100			
5	Стандартне відхилення	15			
6				Імовірність для інтервалу	
7	Імовірність для X <=			Від X	85
8	X	55		до Y	115
9	Z	-3		Z для 85	-1
10	P(X<=55)	0,001349898		Z для 115	1
11				P(X<=85)	0,1587
12	Імовірність для X >			P(X<=115)	0,8413
13	X	145		P(85<=X<=115)	0,6827
14	Z	3			
15	P(X>145)	0,0013		ıевним значенням кумулятивного від	
16				Cumulative Percentage	10,00%
17	Імовірність для X<55 or X >145			Z	-1,28
18	P(X<55 or X >145)	0,0027		X	80,78
19					

Рис. 5.4.8. Результати всіх розрахунків

6. Перевірте формули (рис. 5.4.9).

	A	B
1	Нормальний розподіл	
2		
3	Дані розподілу	
4	Середнє	100
5	Стандартне відхилення	15
6		
7	Імовірність для X <=	
8	X	55
9	Z	=НОРМАЛИЗАЦИЯ(B8;B4;B5)
10	="P(X<="&B8&")"	=НОРМРАСП(B8;B4;B5;ИСТИНА)
11		
12	Імовірність для X >	
13	X	145
14	Z	=НОРМАЛИЗАЦИЯ(B13;B4;B5)
15	="P(X>"&B13&")"	=1-НОРМРАСП(B13;B4;B5;ИСТИНА)
16		
17	="Імовірність для X<"&B8&" or X >"&B13	
18	="P(X<"&B8&" or X >"&B13 & ")"	=B10+B15
19		

	D	E
1		
2		
3		
4		
5		
6	Імовірність для інтервалу	
7	Від X	85
8	до Y	115
9	="Z для " &E7	=НОРМАЛИЗАЦИЯ(E7;B4;B5)
10	="Z для " & E8	=НОРМАЛИЗАЦИЯ(E8;B4;B5)
11	="P(X<="&E7&")"	=НОРМРАСП(E7;B4;B5;ИСТИНА)
12	="P(X<="&E8&")"	=НОРМРАСП(E8;B4;B5;ИСТИНА)
13	="P("&E7&"<=X<="&E8&")"	=ABS(E12-E11)
14		
15	Знайти X та Z за певним значенням кумулятивного відсотку Cum. Pctage.	
16	Cumulative Percentage	0,1
17	Z	=НОРМСТОБР(E16)
18	X	=НОРМОБР(E16;B4;B5)
19		

Рис. 5.4.9. Формули для розрахунків

Завдання для самостійної роботи

Скористуйтеся створеною таблицею (зробіть її копію на новому листі) для вирішення наступних задач [15].

Протягом 2001 р. 61,3 % сімей США покупали зерна кофе, потративши 36,16 доларів. Дослідить річні витрати на покупку кофе, якщо припустити, що вони мають нормальний розподіл з середнім 36,16 та стандартним відхиленням 10 доларів.

1. Яка імовірність того, що американська сім'я витратить на кофе менше 25$.
2. Яка імовірність того, що американська сім'я витратить на кофе більше 50$.
3. Яка імовірність того, що американська сім'я витратить на кофе більше 57$.
4.Яка доля сімей, які витрачають на покупку кофе від 30 до 40 $.
5. Менше якої суми витрачають на покупку кофе 99% американські сім'ї?
6.. Більше якої суми витрачають на покупку кофе 80% американських сімей?
7. Чи обґрунтовано припущення про те, що розподіл витрат на покупку кофе є приблизно нормальним?

Контрольні питання

1. Опишіть призначення та аргументи всіх функцій, які ми використовували в цій роботі.
2. Як побудовані таблиці нормального розподілу?
3. Які властивості нормального розподілу?
4. Яким чином розраховуються Z-оцінки?

Практична робота 5.5

Побудова довірчих інтервалів для середнього генеральної сукупності

Мета роботи: надати досвід обчислення довірчих інтервалів засобами електронних таблиць

Короткі теоретичні відомості

Довірчий інтервал для параметра генеральної сукупності, це інтервал в якому, з імовірністю, наприклад 95% знаходиться невідомий параметр цієї сукупності. На рис.5.4. наведені види довірчих інтервалів.

Загальна формула для довірчого інтервалу:

Точкова оцінка ± (критичне значення)(стандартна помилка)

- Точкова оцінка - це вибіркова статистика, яку ми використовуємо, щоб знайти параметр генеральної сукупності (наприклад, вибіркове середнє)
- Критичне значення - це табличне значення, яке базується на вибірковому розподілі точкового оцінювання та бажаному довірчому інтервалі
- Стандартна помилка - це стандартне відхилення точкової оцінки

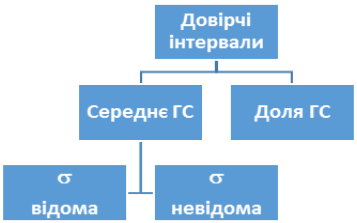


Рис. 5.5.1. Види довірчих інтервалів

Довірчий інтервал для середнього розраховується за формулою, якщо **σ відоме**:

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\overline{X}$	Середнє вибірки (точкова оцінка μ)
$Z_{\alpha/2}$	Критичне значення Z стандартного розподілу для імовірності α/2 в кожному хвості
σ	Стандартне відхилення генеральної сукупності
n	Обсяг вибірки

Рис. 5.5.2. Побудова довірчих інтервалів для середнього, якщо σ відоме [15]

Довірчий інтервал ***для середнього*** *розраховується за наступною формулою, якщо* ***σ невідоме:***

$$\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \quad (5.5.1)$$

де

$\overline{X}$	Середнє вибірки
$t_{\alpha/2}$	Критичне значення розподілу t Стьюдента для ступенів свободи n-1 та α/2
S	Стандартне відхилення вибірки
n	Обсяг вибірки

Довірчий інтервал для долі ознаки розраховується за формулою:

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \quad (5.5.2)$$

де

p	Доля ознаки у виборці
$Z_{\alpha/2}$	Критичне значення Z стандартного розподілу для імовірності α/2 в кожному хвості
n	Обсяг вибірки

Обсяг вибірки для середнього розраховується за формулою

$$n = \frac{Z_{\alpha/2}^{2}\,\sigma^{2}}{e^{2}}, \quad (5.5.3)$$ де

σ	Стандартне відхилення генеральної сукупності
$Z_{\alpha/2}$	Критичне значення Z стандартного розподілу для імовірності α/2 в кожному хвості
e	Помилка вибірки

Обсяг вибірки для долі ознаки розраховується за формулою

$$n = \frac{Z^2 \pi (1-\pi)}{e^2} \qquad (5.5.4)$$

π	Доля ознаки в Генеральній сукупності
$Z_{\alpha/2}$	Критичне значення Z стандартного розподілу для імовірності α/2 в кожному хвості
e	Помилка вибірки

Хід роботи

Завдання 1.Розрахувати найбільш уживані значення Z для відповідних довірчих інтервалів

На рис. 3 наведені вихідні дані для розрахунків. Для обчислення Z скористайтеся функцією НОРМ.СТ.ОБР.

Довірчий рівень, %	**Довірчий рівень, 1-α**	**α**	**α/2**	**$Z_{\alpha/2}$**
80				
90				
95				
98				
99				
99,8				
99,9				

Рис. 5.5.3. Вихідні дані для розрахунків

Порівняйте отримані результати з рис.5.5.4.

	A	B	C	D	E
1	**Довірчий рівень, %**	**Довірчий рівень, 1-α**	**α**	**α/2**	**$Z_{\alpha/2}$**
2	80	0,8	0,2	0,1	1,28
3	90	0,9	0,1	0,05	1,645
4	95	0,95	0,05	0,025	1,96
5	98	0,98	0,02	0,01	2,33
6	99	0,99	0,010	0,005	2,58
7	99,8	0,998	0,002	0,001	3,09
8	99,9	0,999	0,001	0,0005	3,29

Рис. 5.5.4 Результати розрахунків: рівні довіри та відповідні значення Z

Перевірте формули для розрахунків (рис.5.5.5):

	A	B	C	D	E
1	**Довірчий рівень, %**	**Довірчий рівень, 1-α**	**α**	**α/2**	**$Z_{\alpha/2}$**
2	80	=A2/100	=1-B2	=C2/2	=ABS(НОРМ.СТ.ОБР(D2))
3	**90**	=A3/100	=1-B3	=C3/2	**=ABS(НОРМ.СТ.ОБР(D3))**
4	**95**	=A4/100	=1-B4	=C4/2	**=ABS(НОРМ.СТ.ОБР(D4))**
5	98	=A5/100	=1-B5	=C5/2	=ABS(НОРМ.СТ.ОБР(D5))
6	**99**	=A6/100	=1-B6	=C6/2	**=ABS(НОРМ.СТ.ОБР(D6))**
7	99,8	=A7/100	=1-B7	=C7/2	=ABS(НОРМ.СТ.ОБР(D7))
8	99,9	=A8/100	=1-B8	=C8/2	=ABS(НОРМ.СТ.ОБР(D8))
9					

Рис. 5.5.5. Формули для розрахунків

Завдання 2. Складіть таблицю для обчислень довірчого інтервалу для середнього наступним чином:

	A	B
1	**Обчислення довірчого інтервалу**	
2		
3	**Дані**	
4	Вибіркове стандартне відхилення	
5	Вибіркове середнє	
6	Обсяг вибірки	
7	Довірчий рівень	
8		
9	**Проміжні обчислення**	
10	Стандартна помилка середнього	
11	Величина Z	
12	Половина довірчого інтервалу	
13		
14	Довірчий інтервал	
15	Нижня межа інтервалу	
16	Верхня межа інтервалу	

Рис. 5.5.6. Підготовча таблиця

Скористуйтеся таблицею для рішення задачі.

Задача1. У великій компанії працює 3000 співробітників. Адміністрація провела опитування щодо з'ясування середньорічної величини витрат на медичні послуги. Для цього з генеральної сукупності співробітників обрана вибірка обсягом 10: **110, 362, 246, 85, 510, 208, 173, 425, 316, 179.** Розрахуйте 90-% довірчий інтервал для витрат співробітників компанії на медичні послуги.

Представте дані наступним чином, розрахуйте середнє та стандартне відхилення.

№ співробітника	**Витрати**
1	110
2	362
3	246
4	85
5	510
6	208
7	173
8	425
9	316
10	179
середнє	261,4
ст.відх.	138,80

Занесіть вихідні дані в таблицю згідно рис. 5.5.6. Результати розрахунків наведено на рис. 5.5.7.

Проведіть розрахунки ще раз з обсягом вибірки 10. Порівняйте отримані довірчі інтервали для цих двох ситуацій.

	A	B	C	D	E
1	**Обчислення довірчого інтервалу**				
2					
3	**Дані**			**№ сотрудника**	**Витрати**
4	Вибіркове стандартне відхилення	138,80		1	110
5	Вибіркове середнє	261,4		2	362
6	Обсяг вибірки	3000		3	246
7	Довірчий рівень	0,9		4	85
8				5	510
9	**Проміжні обчислення**			6	208
10	Стандартна помилка середнього	2,53		7	173
11	Величина Z	1,645		8	425
12	Половина довірчого інтервалу	4,1683		9	316
13				10	179
14	Довірчий інтервал			середнє	261,40
15	Нижня межа інтервалу	257,23		ст.відх.	138,80
16	Верхня межа інтервалу	265,57			
17					

Рис. 5.5.7. Результати розрахунків

Для отримання ширини інтервалу використовується функція **ДОВЕРИТ**. Функція визивається з наступними параметрами:

ДОВЕРИТ(1-довірчий рівень; стандартне відхилення, обсяг вибірки).

Для обчислення нижньої та верхньої межі довірчого інтервалу ширина цього інтервалу ділиться пополам та додається до вибіркового середнього.

До проміжних обчислювань ми віднесли обчислювання стандартної помилки та величини Z. Формули для розрахунків наведено на рис. 5.5.8.

	A	B
1	**Обчислення довірчого інтервалу**	
2		
3		**Дані**
4	Вибіркове стандартне відхилення	138,8
5	Вибіркове середнє	261,4
6	Обсяг вибірки	3000
7	Довірчий рівень	0,9
8		
9	**Проміжні обчислення**	
10	Стандартна помилка середнього	=B4/КОРЕНЬ(B6)
11	Величина Z	=ABS(НОРМ.СТ.ОБР((1-B7)/2))
12	Половина довірчого інтервалу	=ДОВЕРИТ(1-B7;B4;B6)
13		
14	Довірчий інтервал	
15	Нижня межа інтервалу	=B5-B12
16	Верхня межа інтервалу	=B5+B12

Рис. 5.5.8. Формули для розрахунків

Скористуйтеся побудованою таблицею для рішення наступних задач [15].

Задача 2. Хай середнє ваги генеральної сукупності коробок з кукурудзяними пластівцями μ =368,3 г, стандартне відхилення σ=15. З генеральної сукупності обирається вибірка, яка складається з 25 коробок. Побудуйте 95% довірчий інтервал для μ.

Задача 3. Відоме, що 25 з 100 осіб – ліворукі. Побудуйте 95% довірчий інтервал для долі ліворуких у генеральній сукупності.

Створіть наступну таблицю.

	A	B
1	**Обчислення довірчого інтервалу для долі ліворуких**	
2		
3	**Дані**	
4	Вибіркове стандартне відхилення ***s*** ***= корень(p*(1-p))***	
5	Доля у вибірці (***p***)	0,25
6	Обсяг вибірки (**n**)	100
7	Довірчий рівень	0,95
8		
9	**Проміжні обчислення**	
10	Стандартна помилка середнього ***=s/корень(n)***	
11	Величина ***Z***	
12	Половина довірчого інтервалу	
13		
14	**Довірчий інтервал**	
15	Нижня межа інтервалу	
16	Верхня межа інтервалу	
17		

Рис. 5.5.9. Вихідні дані

Проведіть розрахунки.

	A	B
1	**Обчислення довірчого інтервалу для долі ліворуких**	
2		
3	**Дані**	
4	Вибіркове стандартне відхилення ***s*** ***= корень(p*(1-p))***	0,43
5	Доля у вибірці (***p***)	0,25
6	Обсяг вибірки (**n**)	100
7	Довірчий рівень	0,95
8		
9	**Проміжні обчислення**	
10	Стандартна помилка середнього ***=s/корень(n)***	0,0433
11	Величина ***Z***	-1,9600
12	Половина довірчого інтервалу	0,0849
13		
14	**Довірчий інтервал**	
15	Нижня межа інтервалу	0,1651
16	Верхня межа інтервалу	0,3349
17		

Рис. 5.5.10. Результати

Перевірте формули:

	A	B
1	**Обчислення довірчого інтервалу для д**	
2		
3	**Дані**	
4	Вибіркове стандартне відхилення ***s = корень(p*(1-p))***	=КОРЕНЬ(B5*(1-B5))
5	Доля у вибірці (***p***)	0,25
6	Обсяг вибірки (**n**)	100
7	Довірчий рівень	0,95
8		
9	**Проміжні обчислення**	
10	Стандартна помилка середнього ***=s/корень(n)***	=B4/КОРЕНЬ(B6)
11	Величина ***Z***	=НОРМСТОБР((1-B7)/2)
12	Половина довірчого інтервалу	=ДОВЕРИТ(1-B7;B4;B6)
13		
14	**Довірчий інтервал**	
15	Нижня межа інтервалу	=B5-B12
16	Верхня межа інтервалу	=B5+B12
17		

Рис. 5.5.11. Формули

Самостійно вирішить задачу 4.

Задача 4. Журнал США, присвячений проблемам маркетингу, опитував продавців, які неєтично поводили себе [15]. З'ясувалося, що в 58% випадків вони надавали невірні звіти, в 50% - працювали в робочий час ще в іншому місці, в 22% випадків оформляли візіти до стриптиз бару, як до ресторану,

в 19% випадків давали клієнту откат. Припустимо, що в опитуванні взяли участь 200 продавців-менеджерів.

1. Побудуйте 95-% довірчий інтервал , який містить долю менеджерів, які скривдили звітність.
2. Побудуйте 95-% довірчий інтервал , який містить долю менеджерів, які працювали в робочий час ще в іншому місці.
3. Побудуйте 95-% довірчий інтервал , який містить долю менеджерів, які оформляли візіти до стриптиз бару, як до ресторану.
4. Побудуйте 95-% довірчий інтервал , який містить долю менеджерів, які давали клієнту откат.
5. Визначте обсяг вибірки, який необхідний для оцінки долі менеджерів, які працювали в іншому місці в робочий час, якщо точність вибіркового дослідження дорівнює +-0,02, а довірчий рівень 95 %.

Контрольні питання

1. В чому різниця між точковим та інтервальним оцінюванням?
2. Які види довірчих інтервалів ви знаєте?
3. Як будується інтервал для середнього?
4. Як будується інтервал для долі ознаки?
5. Яким чином розраховується обсяг вибірки? Які для цього потрібні дані?

Практична 5.6
Побудова та порівняння графіків нормальної імовірності кількох розподілів

Мета роботи: будувати та анлізувати графіки нормального розподілу емпіричних даних в Excel SPSS

Короткі теоретичні відомості

Графіком нормальної імовірності розподілу є графік, де по осі X розташовані квантілі нормального розподілу, а по осі У – значення даних, які спостерігаються.
Якщо точки, які відповідають даним лежать на графіку близько до прямої, яка йде від лівого нижнього до правого верхнього кута, то дані розподілені приблизно нормально.

Для побудови графіку необхідно:

- Упорядочити дані за зростанням
- Обчислити квантілі стандартизованого нормального розподілу
- Представити графік, де по осі У – наші дані, по X- відповідні квантілі
- Оцініти імовірність того, що дані розподілені приблизно нормально.

Перший квантіль стандартизованого нормального розподілу – це величина Z, яка відповідає площі фигури, яка лежить під кривою щільності імовірності ***1/(n+1).***

i*-ий квантіль** стандартизованого нормального розподілу – це величина Z, яка відповідає площі фигури, яка лежить під кривою щільності імовірності ***i/(n+1).

n*-ий** квантіль стандартизованого нормального розподілу – це величина Z, яка відповідає площі фигури, яка лежить під кривою щільності імовірності ***n/(n+1).

Постановка задачі

В таблиці наведені дані 19 студентів [15], які вони отримали за 4 тести з математичних методів, в останній колонці - стандартизовані оцінки для нормального розподілу (квантилі), які треба розрахувати самостійно.

№ студ.	1 тест	2 тест	3 тест	4 тест	Q
1	48	47	47	38	-1,65
2	52	54	48	41	-1,28
3	55	58	50	44	-1,04
4	57	61	51	47	-0,84
5	58	64	52	50	-0,67
6	60	66	53	53	-0,52
7	61	68	53	56	-0,39
8	62	71	54	59	-0,25
9	64	73	55	62	-0,13
10	65	74	56	65	0
11	66	75	57	68	0,13
12	68	76	59	71	0,25
13	69	77	62	74	0,39
14	70	77	64	77	0,52
15	72	78	66	80	0,67
16	73	79	69	83	0,84
17	75	80	72	86	1,04
18	78	82	76	89	1,28
19	82	83	83	92	1,65

Якщо n=19, то перший квантиль буде 1/(19+1)=1/20=0,05 (що відповідає 5 % площі).

Рішення задачі в Excel

Хід роботи

1. Для розрахунку останньої колонки використовуйте проміжну таблицю, де Z розраховано за допомогою НОРМ.СТ.ОБР (передаємо їй імовірність, тобто відсоток площі/100).

	A	B	C
1	№	% площі	Z
2	1	5,00	-1,645
3	2	10,00	-1,282
4	3	15,00	-1,036
5	4	20,00	-0,842
6	5	25,00	-0,674
7	6	30,00	-0,524
8	7	35,00	-0,385
9	8	40,00	-0,253
10	9	45,00	-0,126
11	10	50,00	0,000
12	11	55,00	0,126
13	12	60,00	0,253
14	13	65,00	0,385
15	14	70,00	0,524
16	15	75,00	0,674
17	16	80,00	0,842
18	17	85,00	1,036
19	18	90,00	1,282
20	19	95,00	1,645

Рис. 5.6.1. Розрахунки значень для 19 квантилів розподілу

2. Побудуйте графіки усіх чотирьох розподілів, які відповідають 4-м тестам, наступним чином: по осі X розташуйте Z-оцінки, по Y – бали, які студенти отримали (рис.5.6.2–5.6.5).

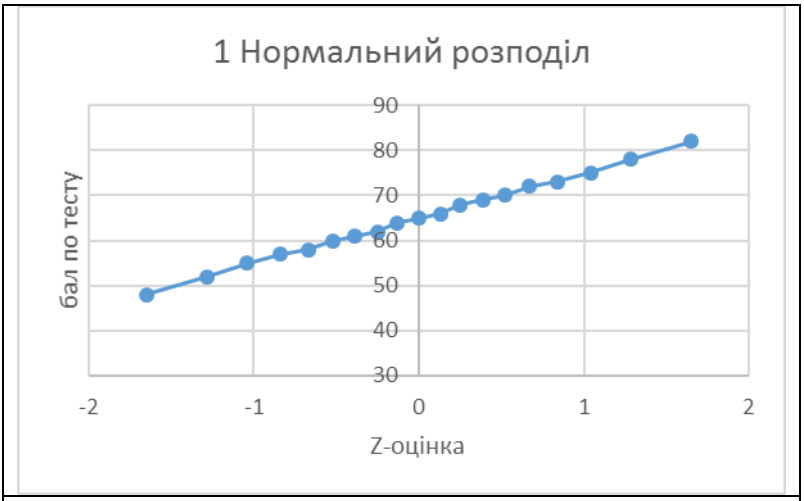


Рис. 5.6.2. Нормальний розподіл

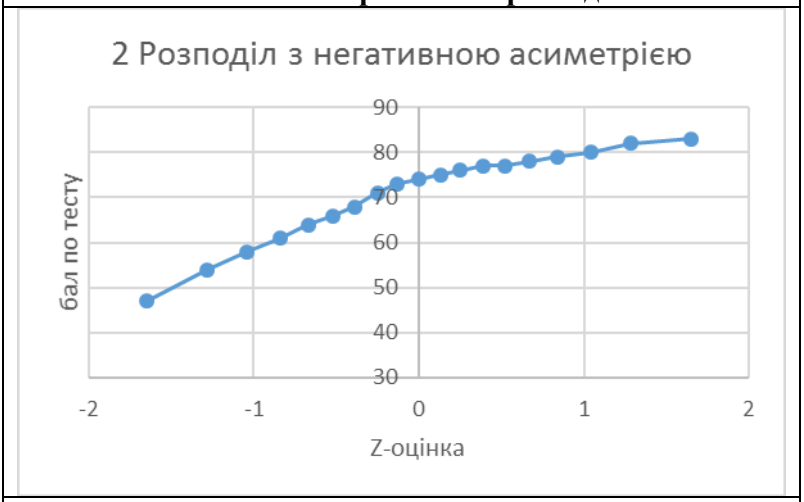


Рис. 5.6.3. Лівостороння асиметрія

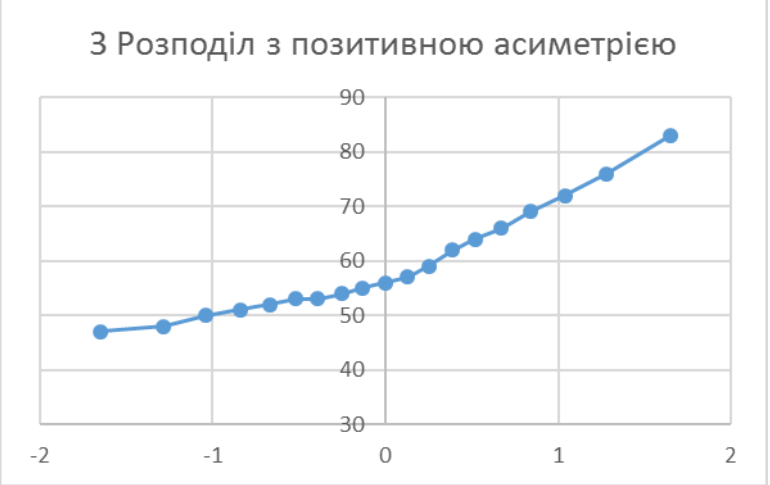


Рис. 5.6.4. Правостороння асиметрія

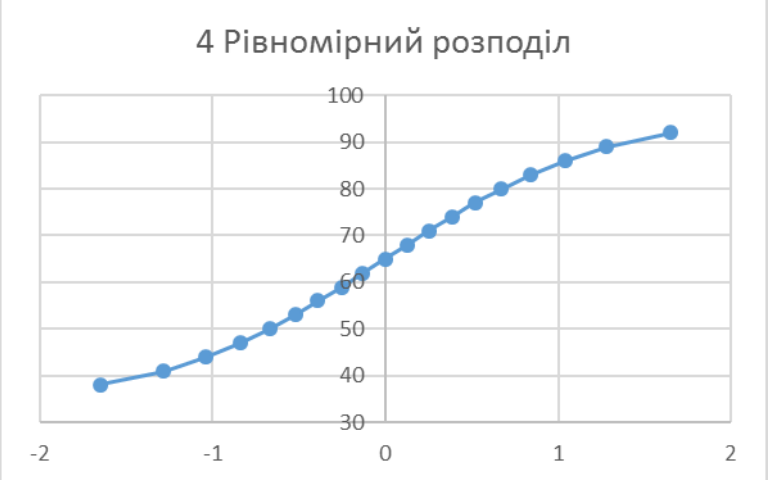


Рис. 5.6.5. Рівномірний розподіл

3.З'ясуйте, яку форму мають ці графіки. Побудуйте також відповідні діаграми частотного розподілу за допомогою пакету аналізу (рис.5.6.6.) та порівняйте їх з попередньо побудованими графіками (рис.5.6.2–5.6.5).

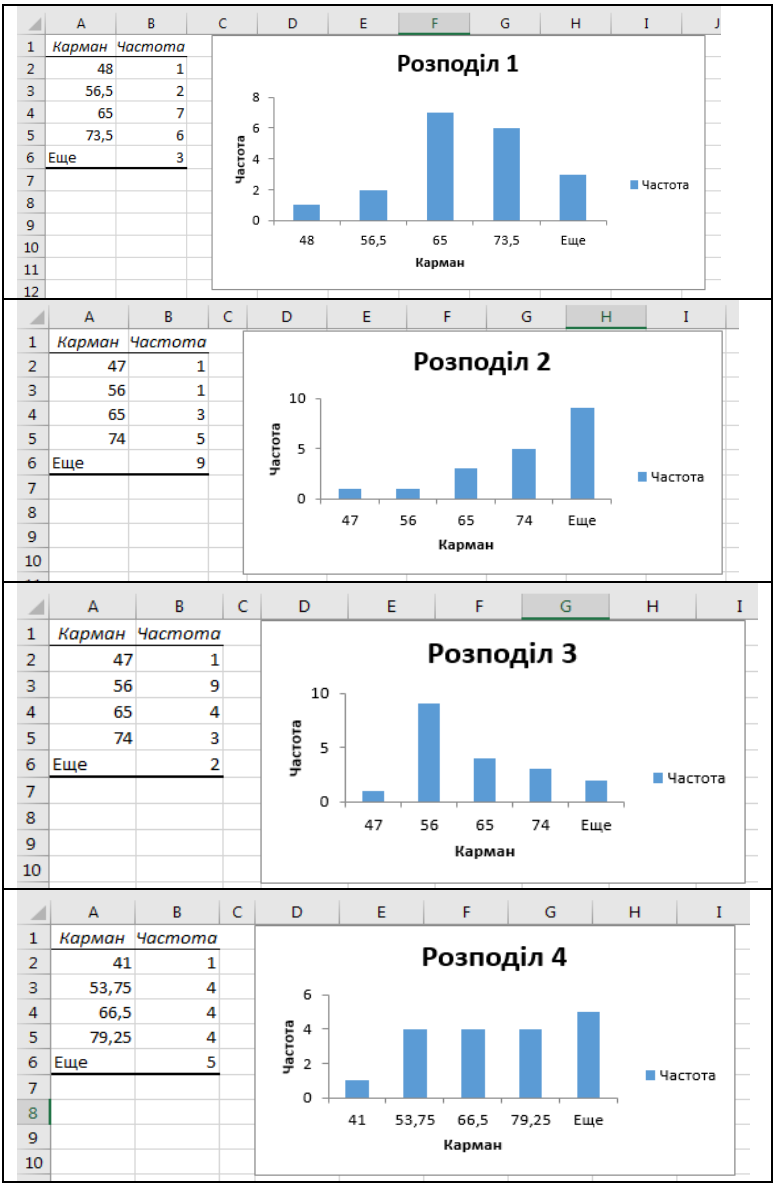


Рис. 5.6.6. Гістограми частотних розподілів 4-х тестів

Рішення задачі в SPSS

Хід роботи

1. .Розрахуйте Z за допомогою функції IDF.NORMAL(perArea/100,0,1). Їй передається імовірність (% площі, який поділений на 100, середнє=0, стандартне відхилення=1).

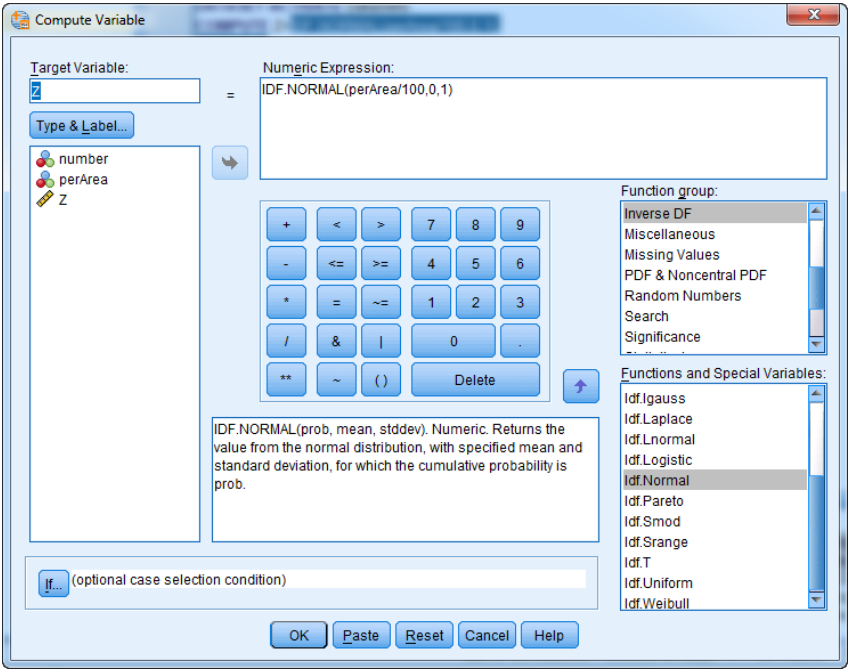


Рис. 5.6.7. Обчислення Z

В результаті отримаєте наступні значення Z.

	number	perArea	Z
1	1	5	-1,64
2	2	10	-1,28
3	3	15	-1,04
4	4	20	-,84
5	5	25	-,67
6	6	30	-,52
7	7	35	-,39
8	8	40	-,25
9	9	45	-,13
10	10	50	,00
11	11	55	,13
12	12	60	,25
13	13	65	,39
14	14	70	,52
15	15	75	,67
16	16	80	,84
17	17	85	1,04
18	18	90	1,28
19	19	95	1,64

Рис. 5.6.8. Результати обчислення Z

2. Додайте в файл 4 змінні з балами по тестам

	number	perArea	Z	test1	test2	test3	test4
1	1	5	-1,64	48	47	47	38
2	2	10	-1,28	52	54	48	41
3	3	15	-1,04	55	58	50	44
4	4	20	-,84	57	61	51	47
5	5	25	-,67	58	64	52	50

Рис. 5.6.9. Додаткові змінні у файлі

3. Побудуйте імовірнісні діаграми (***Analyze, Descriptive Statistics, P-P Plots***), для усіх тестів:

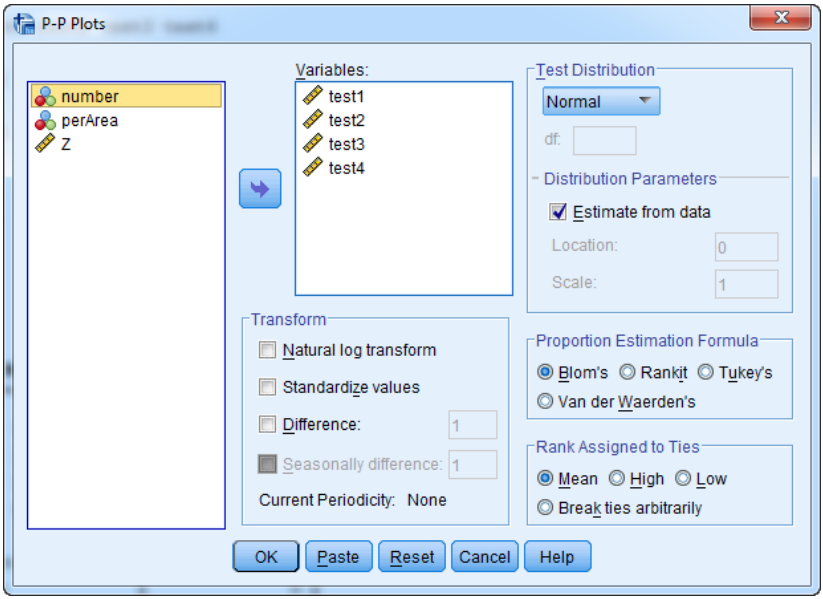


Рис. 5.6.10. Вікно «P-P Plot»

Отримаєте:

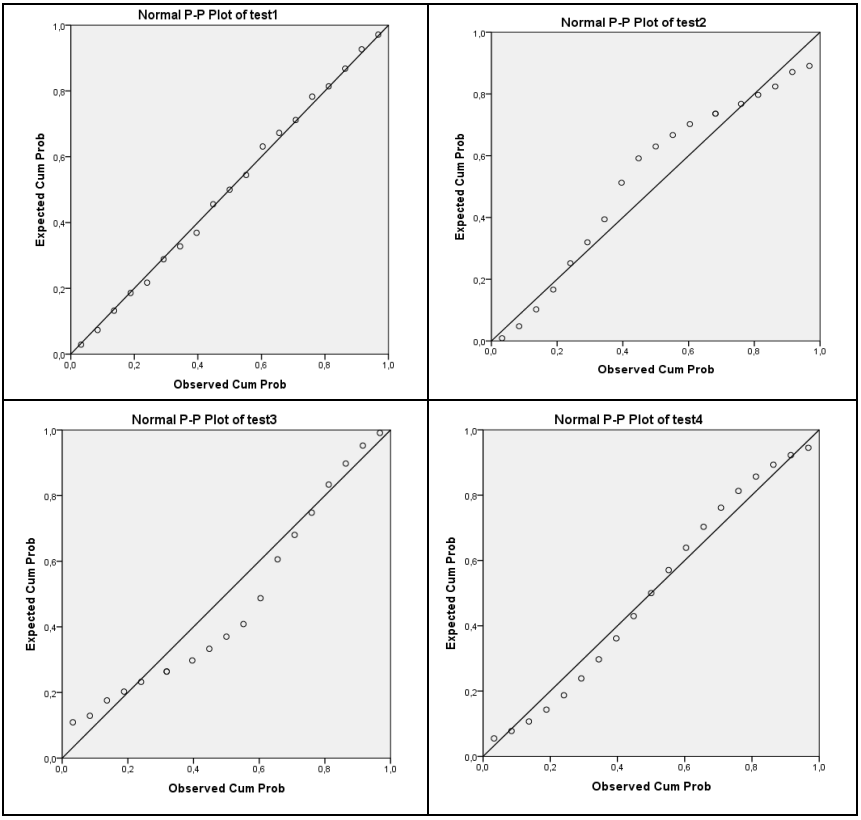
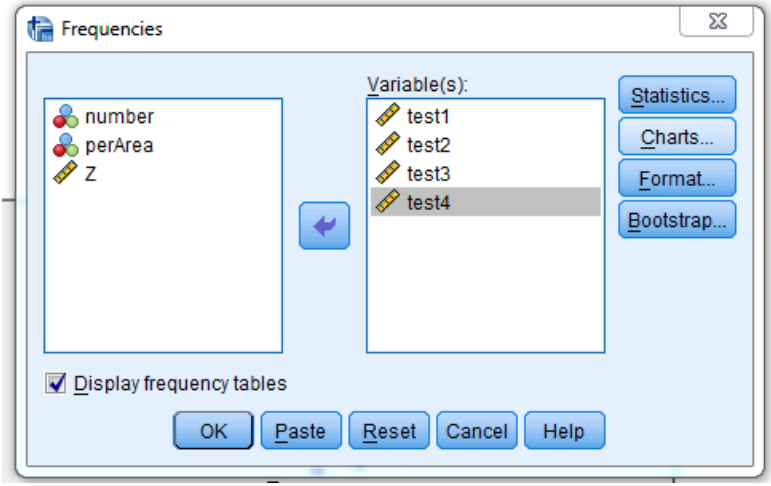


Рис. 5.6.11.Тести 1-4.

4. Побудуйте також гістограму з нормальною кривою для 4-х тестів.



Оберіть Гістограму, з нормальною кривою

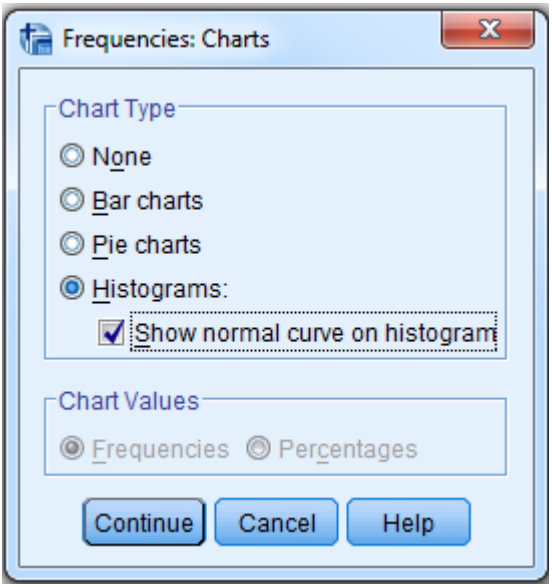
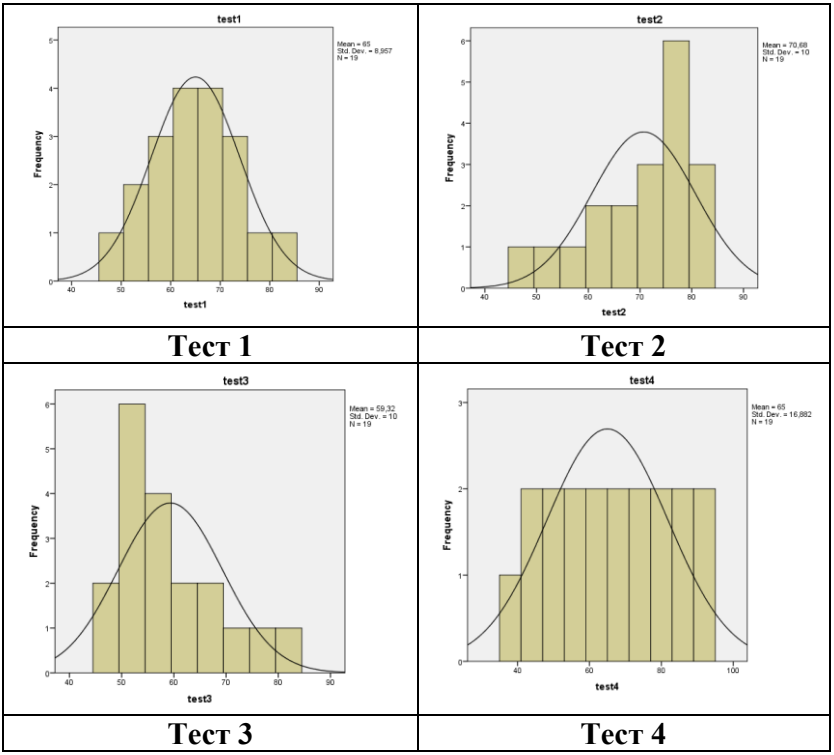


Рис.5.6.12. Вікно «Charts»

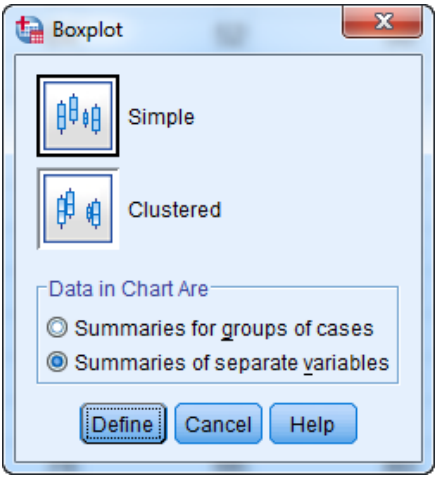
5. Перегляньте та порівняйте графіки:



Тест 1	Тест 2
Тест 3	Тест 4

Рис. 5.6.13. Гістограми з нормальною кривою

6. Побудуйте ящичкові діаграми **Graphs, Legacy Dialogs, Boxplot** (Рис.5.6.15-5.6.16)



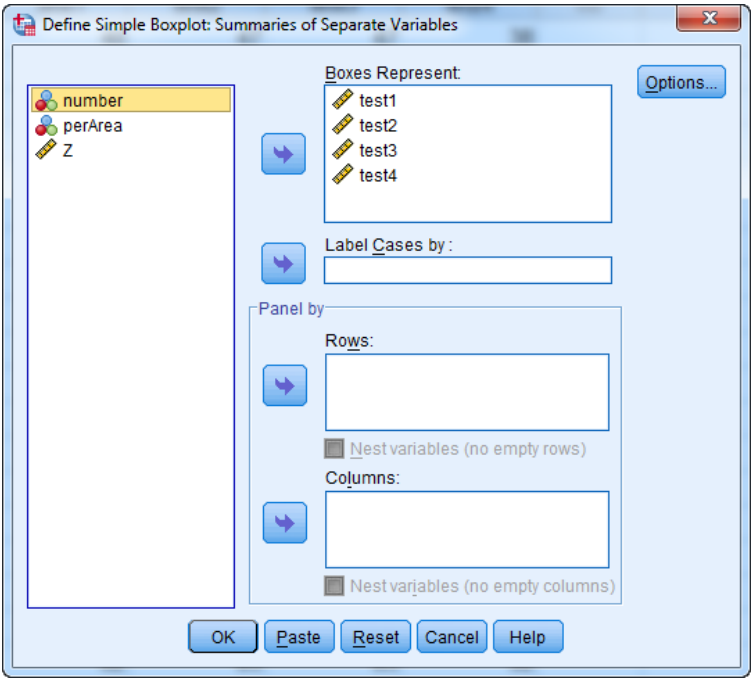


Рис. 5.6.14. Побудова ящичкових діаграм

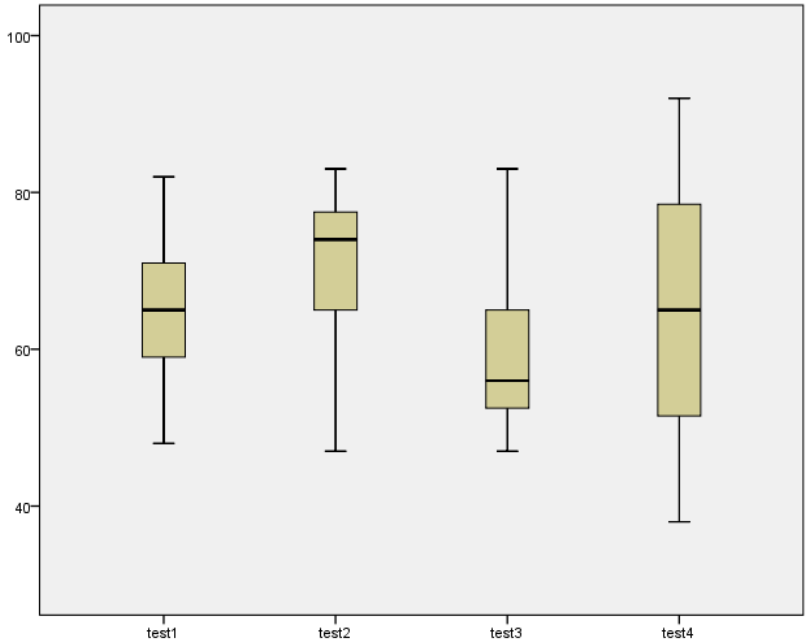


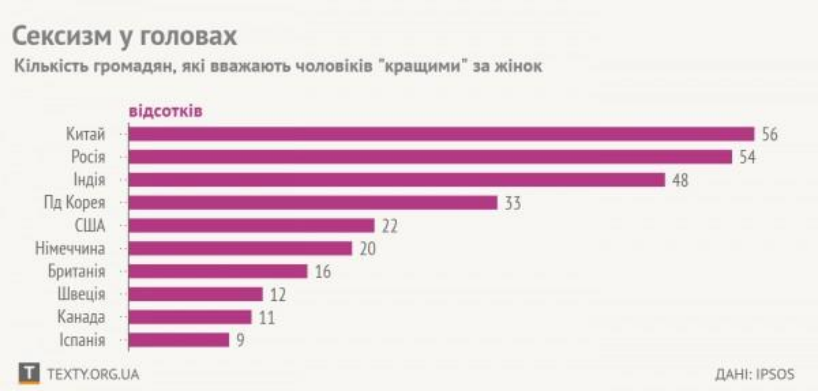
Рис. 5.6.15. Ящичкові діаграми для розподілів балів за 4-мя тестами

Контрольні питання

1. Як будується графік нормального розподілу емпіричних даних?
2. Як розрахувати квантилі розподілу?
3. Призначення функції НОРМ.СТ.ОБР.
4. Яка функція відповідає їй в SPSS?
5. Як побудувати гістограму разом з нормальною кривою в SPSS?

Завдання на повторення

Завдання 1. Проаналізуйте відповіді на питання "*Я вважаю, що чоловіки краще за жінок можуть впоратись із такими справами, як заробляння грошей, отримання освіти та навчання інших*" (Згодні + переважно згодні) за даними, наданими texty.org.ua.



Побудуйте 90%, 95% інтервал для долі в ГС, тих, хто вважає, що чоловік все робіть краще: 1) для однієї з країн (Китай, Росія, Індія, Південна Корея); для однієї з країн (Німеччина, Британія, Швеція, Канада, Іспанія). Припустимо, що в опитуванні брало участь по 600 осіб з кожної країни.

Завдання 2. Обговоріть результати соціологічного дослідження студентів Запорізького національного університету, який вперше в історії України сам провів дослідження про рівень та форми корупції у власних стінах [texty.org.ua]. Навчальний заклад оприлюднив дані та виступає за поширення цієї інформації. Побудуйте 95%, 99% довірчі інтервали для долі тих в ГС, хто стискався з хабарями у формі грошей; канцелярських товарів; купівлі літератури. Припустимо, n=400.

Завдання 3. Оцінки, які отримують студенти на іспиті з «Соціальної статистики» є нормально розподіленими випадковими величинами з середнім μ=75 та стандартним відхиленням 7.

1. Яка імовірність отримати 90 балів?
2.Яка доля студентів, які отримали отримали від 75 до 90 балів.
3.Яка доля студентів, які отримали від 90 до 100 балів.
4.Укажіть суму балів, яку перевищили тільки 10% студентів; 5% студентів.

Завдання 4. Студенти університету при виборі дисципліни за вибором проявили сваєбажання навчатися демографії. Відомо, що з 16 студентів, які обрали цей курс, 9 виявилося жіночої статі. Серед 6 студентів, які не обрали курс, було три хлопця.

Стать студента	Демографія		Разом
	обрали	Не обрали	
жіноча			
чоловіча			
Разом			

1. Створіть таблицю спряженості 2Х2, та оцініть імовірність різних подій.
2. Наведіть приклад елементарної події.
3. Наведіть приклад сумісної події.
4. Яка подія є додатковою до «обрав курс демографії»?
5. Яка подія є додатковою до «стать - жіноча»?

   Обчисліть імовірність для випадковим чином обраного студента:
6. Він – хлопець.
7. Особа обрала курс демографії.
8. Він – хлопець і він обрав курс.
9. Вона – дівчина і вона не обрала курс.
10. Він – хлопець або він обрав курс.
11. Вона – дівчина або вона обрала курс.
12. Чи вірно, що чоловіки частіше обирають курс демографії, ніж жінки?

РОЗДІЛ 6
ПЕРЕВІРКА ГІПОТЕЗ

Гіпотеза – це науково обґрунтоване припущення про структуру досліджуваних об'єктів, про характер елементів і зв’язків, що утворюють ці об'єкти, про механізми їх функціонування та розвитку.

У ході дослідження можуть бути висунуті структурні або пояснювальні гіпотези.

Структурними називаються гіпотези про структуру досліджуваних об'єктів.

Приклад структурної гіпотези: найбільш важливими факторами вибору покупця є: ціна, зовнішній вигляд, якість товару.

Пояснювальними називаються гіпотези про причинно-наслідкові зв'язки в досліджуваних об'єктах.

Приклади пояснювальних гіпотез: 1) самоосвітня діяльність майбутніх менеджерів сприяє підвищенню рівня їх професійної підготовленості; 3) збільшення витрат на рекламу не вплине на доходи фірми.

До наукових гіпотез пред'являються дві основні вимоги:

1. Гіпотеза повинна бути несуперечливою (формулювання гіпотези повинно піддаватися логічному аналізу, який встановлює її несуперечність).

2. Гіпотеза повинна бути доступна перевірці в ході дослідження (інакше немає сенсу її висувати, адже перевірити, істинна гіпотеза чи ні, буде неможливо).

Слід розрізняти наукові і статистичні гіпотези.

Наукова гіпотеза – це, як уже зазначалося, розумне, обґрунтоване і розвинене припущення про структуру досліджуваного об'єкту або причинно-наслідкові зв'язки в ньому, про можливі шляхи вирішення тієї чи іншої проблеми.

Наукова гіпотеза формулюється як теорема і перевіряється в ході експерименту – досвід відповідає на питання, чи є

гіпотеза істинною, чи ні. Грамотне формулювання наукової гіпотези – це справді творчий акт і важливіший етап дослідження.

В економіці, психології, педагогіки, соціології, і багатьох інших науках часто висувають гіпотези, які можна і потрібно перевіряти статистично, тобто спираючись на результати вимірювань в вибірці.

Під **статистичною гіпотезою** розуміють всяке висловлення про генеральну сукупність, що перевіряється за виборкою. При цьому під генеральною сукупністю розуміється той об'єкт, на який будуть поширюватися висновки за підсумками дослідження. Під вибіркової сукупністю розуміється сформована відповідно до певних правил (сформульованими в теорії вибіркового методу) частина генеральної сукупності, на якій безпосередньо проводиться дослідження.

Статистична гіпотеза являє собою твердження щодо невідомого параметра (невідомих параметрів) генеральної сукупності.

Статистичну гіпотезу прийнято позначати буквою Н.

Приклади статистичних гіпотез:

1.Твердження $H: \mu = 125$ – статистична гіпотеза, яка говорить, що невідоме середнє конкретної сукупності $\mu = 125$.

Таке твердження або справедливо, або помилково.

2. $H: \sigma^2_1 = \sigma^2_2$ - статистична гіпотеза, яка стверджує, що дисперсії 1-й і 2-й сукупностей рівні.

3. Н: $\mu 1 = \mu 2 = \mu 3$ - статистична гіпотеза, яка стверджує, що середні значення трьох сукупностей рівні.

Історія перевірки статистичних гіпотез бере початок з XYIII століття. Перший відомий приклад випробування статистичної гіпотези - в роботі Дж.Арбутнота (1667 - 1735 рр.), датованій 1710-м роком, «Доводи на користь божественних пророцтв, виведені на основі постійних та систематичних спостережень над народженням обох статей».

Відзначивши, що записи на протязі 82-х років свідчили х про більшу кількість народжених хлопчиків, Арбутнот показав, що ці дані спростовують гіпотезу про те, що народження хлопчиків і дівчаток різновірогідні. Бо, якщо ймовірність народження хлопчика точно = 1/2, то ймовірність того, що за 82 роки народилося більше хлопчиків, ніж дівчаток, була б нескінченно мала. Арбутнот дійшов висновку, що велика частка народження чоловіків - результат втручання Провидіння.

Грамотне формулювання статистичних гіпотез систематизує припущення дослідника та представляє їх у чіткому і лаконічному вигляді. Завдяки такій гіпотезі «дослідник не втрачає дороговказною нитки в процесі розрахунків та йому легко зрозуміти після їх закінчення, що, власне, він виявив».

Перевірка гіпотез є одним з напрямків в теорії статистичного висновку (див. рис. 6.1).

Статистичний висновок – це твердження про параметри генеральної сукупності на підставі вивчення вибіркової сукупності.

Числові показники, що характеризують генеральну сукупність, називають параметрами і позначаються буквами грецького алфавіту. Ті ж характеристики для вибірки називаються статистиками і позначаються буквами латинського алфавіту.

Наприклад:

μ – середнє в генеральній сукупності, x – середнє у виборці;

σ^2 –дисперсія в генеральній сукупності, S^2 – дисперсія в виборці.

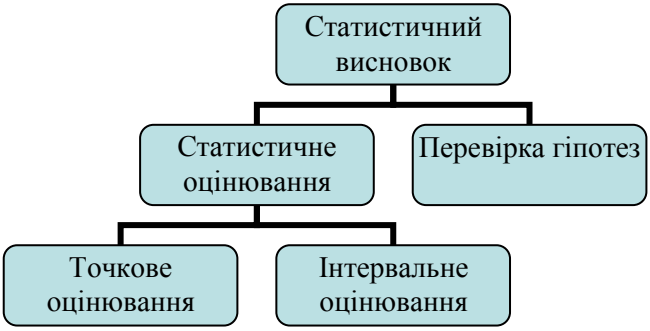


Рис. 6.1. Напрями в статистичному висновку

Статистичне оцінювання полягає в тому, що дослідник за вибіркою шукає показник, найбільш близький до оцінюваного параметру (точкове оцінювання), або інтервал, в межах якого з великою ймовірністю лежить цей параметр (інтервальне оцінювання).

Перевірка гіпотез полягає в тому, що дослідник заздалегідь формулює деяке твердження про параметрах ГС (гіпотезу), потім оцінює ступінь відповідності результатів, отриманих у вибірковому дослідженні, сформульованій гіпотезі і приймає рішення про істинність або хибність гіпотези.

Статистичні гіпотези можна розділити на: гіпотези про закони розподілу (наприклад: продуктивність праці робітників, що виконують однакову роботу в однакових умовах, має нормальний розподіл) і гіпотези про параметри розподілу, тобто про середні, дисперсії, коефіцієнти кореляції, частки ознак і т.п. (Наприклад: 1) середні розміри деталей, вироблених на типових верстатах, які паралельно працюють, не розрізняються між собою; 2) середня успішність студентів 1 курсу

спеціальності «Соціологія» не відрізняється від середньої успішності студентів 1 курсу спеціальності «Маркетинг») (рис.6.2.)

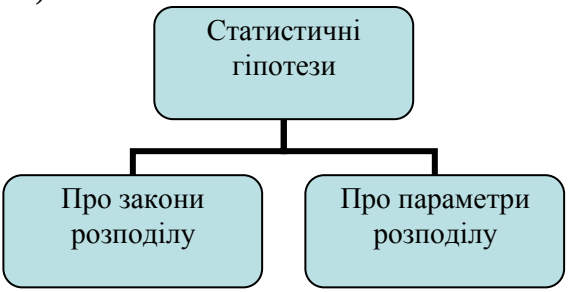


Рис. 6.2. Види статистичних гіпотез в залежності від їх змісту

Крім того, статистичні гіпотези підрозділяються на нульові й альтернативні, спрямовані й неспрямовані (див. рис. 6.3).

Нульова гіпотеза – це гіпотеза про відсутність відмінностей. Вона позначається як ***H0*** і називається нульовою тому, що містить у своїй формулюванні число нуль:

$$H0: X1 - X2 = 0$$

де ***X1, X2*** – значення ознак, які зіставляються. Нульова гіпотеза спростовується, якщо відмінності виявяться значущими.

Альтернативна гіпотеза – це гипотеза про значимість відмінностей. Вона позначається як ***H1***.

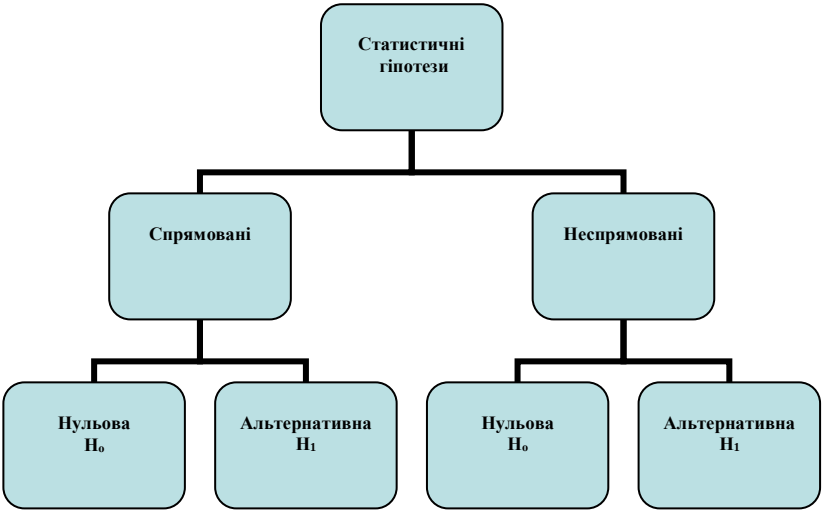


Рис. 6.3. Види статистичних гіпотез в залежності від їх спрямованості та змісту

І нульова, і альтернативна гіпотези можуть бути спрямованими і неспрямованими.

Спрямовані гіпотези:

H0: X1 не перевищує X2
H1: X1 перевищує X2

Неспрямовані гіпотези:

H0: X1 не відрізняється від X2
H1: X1 відрізняється від X2

Якщо дослідник припускає, що в групі, члени якої брали участь в спеціальному тренінгу, індивідуальні значення досліджуваних за будь-якою ознакою, наприклад, з ініціативності, вище, а в групі, члени якої не брали участь в тренінгу, - нижче, то для перевірки значущості цих відмінностей необхідно сформулювати спрямовану гіпотезу.

Тобто, якщо ми хочемо перевірити, чи дійсно в групі А під впливом певних експериментальних впливів відбулися більш виражені зміни, ніж в групі В, ми повинні сформулювати спрямовану статистичну гіпотезу.

Якщо ми хочемо перевірити, чи розрізняються форми розподілу досліджуваного ознаки в групах А і В, нам необхідно сформулювати неспрямовану статистичну гіпотезу.

Статистичні гіпотези перевіряються шляхом розрахунку статистичних критеріїв.

Статистичний критерій – це вирішальне правило, що забезпечує надійну поведінку, тобто прийняття істинної та відхилення помилкової гіпотези з високою ймовірністю.

Статистичні критерії позначають також метод розрахунку певного числа і саме це число.

За співвідношенням емпіричного і критичного значень критерію ми можемо судити про те, підтверджується або спростовується нульова гіпотеза. Наприклад, якщо $\chi 2$емп> $\chi 2$крит, Н0 відкидається.

У більшості випадків для того, щоб ми визнали відмінності значущими, необхідно, щоб емпіричне значення критерію перевищувало критичне, хоча є критерії (наприклад, критерій Манна-Уїтні, критерій знаків, Т-критерій Уілкоксона), в яких ми повинні дотримуватися протилежного правилу.

Алгоритм перевірки статистичних гіпотез

1. Формулювання нульової та альтернативної гіпотези.

2. Вибір відповідного статистичного критерію, назвемо його k.

3. Розрахунок за даними вибірки емпіричного значення kемп.

4. На підставі обсягу вибірки, рівня значущості, числа ступенів свободи визначення критичного значення критерію kкрит.

5. Порівняння емпіричного і критичного значень критерію.

6. Якщо kемп> kкрит, то нульова гіпотеза відкидається (Виключення: критерій Манна-Уїтні, критерій знаків, критерій Т-Уілкоксона).

Статистичні критерії діляться на параметричні і непараметричні.

Параметричні статистичні критерії включають в формулу розрахунку параметри розподілу, тобто середні і дисперсії (t-критерій Стьюдента, F-критерій і ін.).

Непараметричні статистичні критерії не включають в формулу розрахунку параметри розподілу і розраховуються на оперуванні частотами або рангами (критерій Розенбаума, критерій Уілкоксона і ін.).

Потужність статистичного критерію – це його спроможність виявляти відмінності, якщо вони є.

Рівні статистичної значущості

Рівень **значущості** – це ймовірність того, що ми признали випадкові відмінності істотними, достовірними.

Так, наприклад, якщо ми стверджуємо, що відмінності істотні (або достовірні) на 5%-му рівні значущості (або при $p \leq 0,05$), це означає, що ймовірність того, що вони все-таки не значимі (недостовірні, несуттєві) – не більше 5%.

Нижчим рівнем статистичної значимості прийнято вважати 5% (або 0,05); достатнім – 1% (0,01); вищим – 0,1% (0,001).

Помилки при перевірці гіпотез

При перевірці будь-статистичної гіпотези рішення дослідника ніколи не приймається зі стовідсотковою впевненістю – завжди є ризик прийняття неправильного рішення.

Сутність перевірки статистичної гіпотези укладається в тому, що ця перевірка є засобом контролю і оцінки цього ризику.

Помилка, яка полягає в тому, що ми відхилили нульову гіпотезу, в той час як вона вірна, називається помилкою I роду.

Імовірність такої помилки зазвичай позначається буквою α; тоді ймовірність правильного рішення – $(1 - \alpha)$. Таким чином, чим менша ймовірність помилки (α), тим більша ймовірність правильного рішення.

Помилка, яка полягає в тому, що ми прийняли нульову гіпотезу, в той час, як вона неправильна, називається помилкою II роду.

Імовірність такої помилки позначається буквою β.

$(1 - \beta)$ – це фактично потужність статистичного критерію, його здатність не припуститися помилки II роду.

Вибір статистичного критерію для перевірки гіпотези

Обираючи той чи інший статистичний критерій для перевірки статистичних гіпотез, слід керуватися таким:

- яка сама гіпотеза, а відповідно, і який тип статистичного завдання треба вирішувати;
- в яких шкалах здійснювалися вимірювання даних;
- які розміри вибірки (вибірок);

- якщо дослідження спирається на більш ніж одну вибірку, то чи можна застосовувати обраний критерій до нерівних за обсягом вибірок;
- яка потужність критерію;
- які можливості має дослідник для розрахунку критерію (чи можна розрахувати критерій «власноруч» або для його розрахунку необхідні певні пропрограмні продукти, чи є в розпорядженні дослідника такі продукти) і т.п.

Основні випадки перевірки гіпотез

Основні випадки перевірки гіпотез про параметри генеральної сукупності:

- гіпотези про середні (рис. 6.4);
- гіпотези про дисперсії (рис. 6.5);
- гіпотези про коефіцієнти кореляції (рис. 6.6);
- гіпотези про частки ознаки (рис. 6.7);
- гіпотези про незалежність ознак в кореляційній таблиці.

Гіпотези про середні можна розділити на гіпотези про рівність середнього певному значенню та гіпотези про значущість відмінності між середніми двох сукупностей (рис. 6.4).

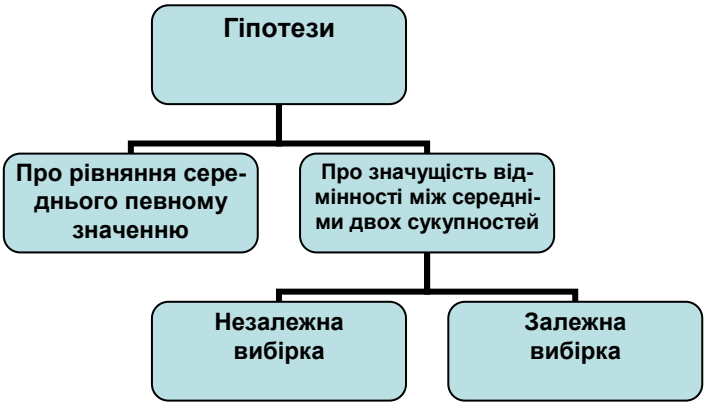


Рис. 6.4. Види гіпотез про середні

На рис. 6.5 представлено види гіпотез про дисперсії в генеральній сукупності.

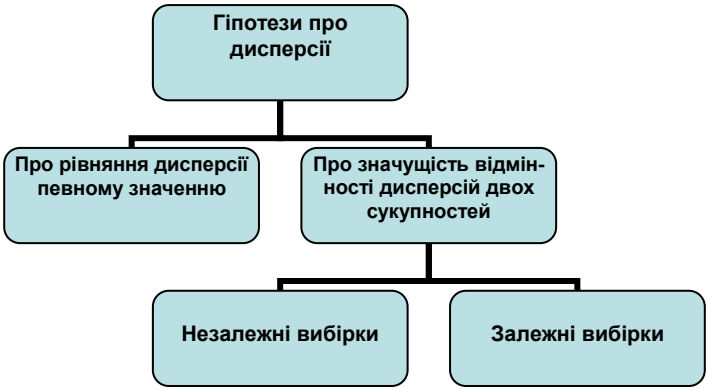


Рис. 6.5. Види гіпотез про дисперсії

На рис. 6.6 представлено види гіпотез про коефіцієнти кореляції.

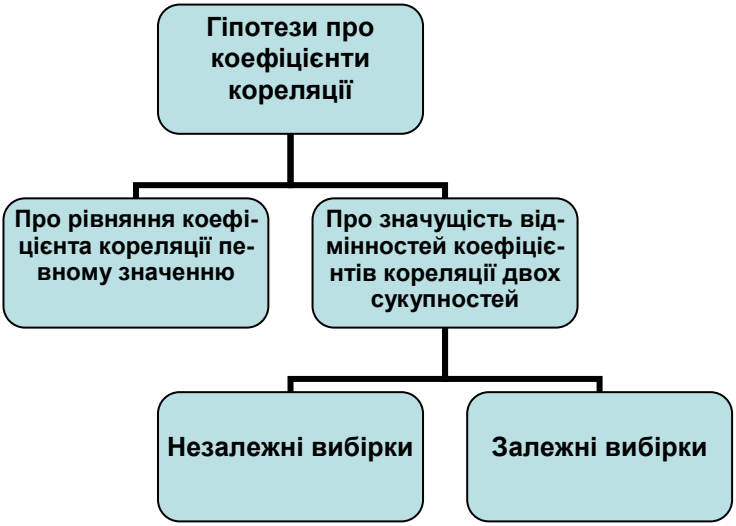


Рис. 6.6. Види гіпотез про коефіцієнти кореляції

На рис. 6.7 представлено види гіпотез про долі ознаки.

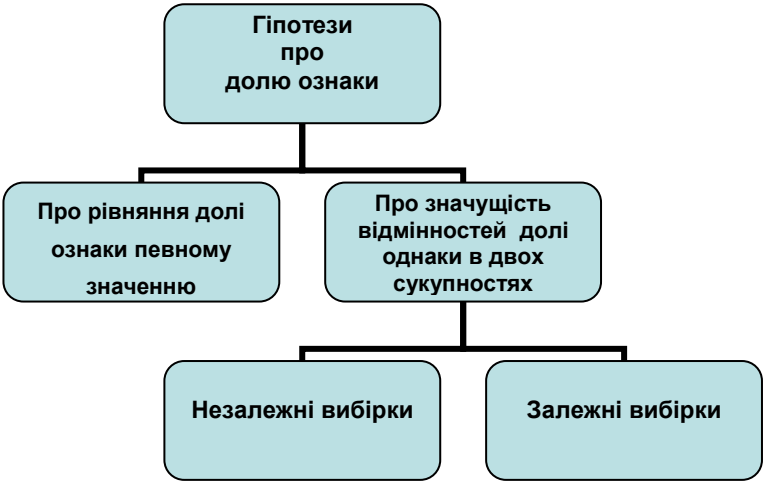


Рис. 6.7. Види гіпотез про долі ознаки

Порівняння середніх

До найбільш часто застосовуваних методів статистичного аналізу відноситься порівняння середніх значень двох вибірок.

При порівнянні середніх значень вибірок вважається, що обидві вибірки підпорядковані нормальному розподілу.

Якщо це не так, то використовуються непараметричні тести (докладніше про них – в роботах Дж.Полларда і Р.Руніона [27].

При порівнянні середніх значень вибірок виділяють 4 тестові ситуації:

Порівняння середніх двох незалежних вибірок	→	***T***-тест Стьюдента для незалежних вибірок
Порівняння середніх двох залежних вибірок	→	***T***-тест Стьюдента для залежних вибірок
Порівняння середніх двох або більше незалежних вибірок	→	Однофакторний дисперсійний аналіз
Порівняння середніх двох або більше залежних вибірок	→	Однофакторний дисперсійний аналіз з повторними вимірюваннями

Практична робота 6.1
Перевірка гіпотез про значущість відмінності середніх в MS Excel

Мета роботи: навчитися використовувати пакет статистичного аналізу MS Excel для перевірки статистичних гіпотез.

Постановка задачі

Фірма найняла 40 осіб для поширення своєї продукції і організувала їх навчання, розділивши на дві групи по 20 осіб в кожній. У програму навчання першої групи входило знайомство з особливостями продукції, яку учасникам групи належало поширювати, в програму навчання другої групи додатково було введено основи психології. У таблиці 6.1.1 наведено число продажів, які здійснили представники обох груп в перший місяць роботи. Чи свідчать представлені в таблиці 6.1.1. дані про відмінності в підготовленості до роботи співробітників, які навчалися за різними програмами?

Таблиця 6.1.1

Число продаж в залежності від виду попереднього навчання

Число продаж	
Група 1	**Група 2**
16	9
17	7
5	5
7	6
7	7
8	8
8	4
6	6
6	6
8	8
8	8
15	12
8	4
10	4
10	10
13	5
4	4
4	3
8	8
9	3

Хід роботи

1. Відкрити Excel. Увести вихідні дані згідно таблиці 5.1.1.

2. Для кожного стовбця даних обчислити середнє та дисперсію із використанням функцій **СРЗНАЧ** та **ДИСП**. Подивитися, чи рівні отримані середні та дисперсії в двох групах.

3. Відкрити пакет аналізу: меню ***«Дані», «Аналіз даних», «Двохвибірковий t-тест з вілмінними дисперсіями»*** (рис. 6.1.1). Прочитати довідку та записати її в зошит.

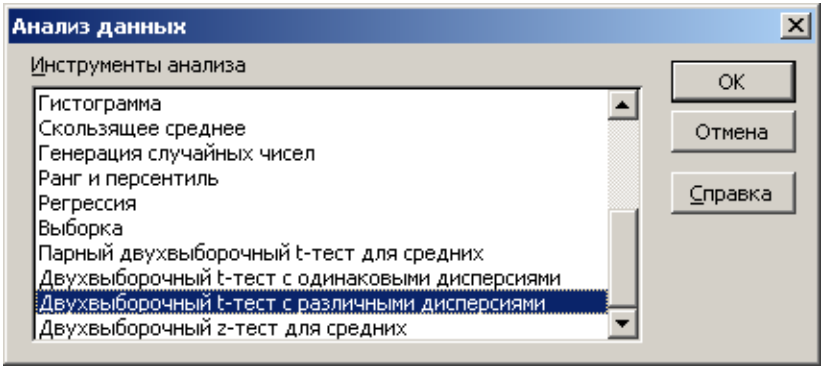


Рис. 6.1.1. Діалогове вікно «Аналіз даних»

4. У вікні «***Двохвибірковий t-тест…***» задати необхідні параметри: інтервал змінної 1, інтервал змінної 2, гіпотетичну різницю середніх – 0, рівень значущості 0,05. Натиснути кнопку *ОК* (рис. 6.1.2).

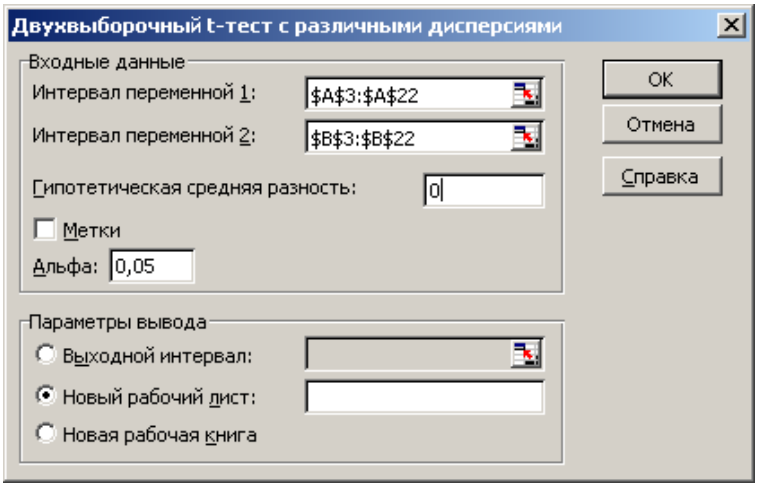


Рис. 6.1.2. Діалогове вікно «*Двохвибірковий t-тест…*»

5. Виписати таблицю результатів (рис. 6.1.3) та зробити висновок щодо значущості відмінностей.

Двохвибірковий t-тест с відмінними дисперсіями		
	Змінна 1	*Змінна 2*
Середнє	8,85	6,35
Дисперсія	13,92368421	5,923684211
Спостереження	20	20
Гіпотетична різниця середніх	0	
df	33	
t-статистика	2,509594424	
P(T<=t) однобічне	0,008586329	
t критичне однобічне	1,692360258	
P(T<=t) двобічне	0,017172658	
t критичне двобічне	2,034515287	

Рис. 6.1.3. Результати розрахунків двохвибіркового t-тесту з відмінними дисперсіями

Збережіть результати.

Вимоги до звіту:

Звіт про роботу повинен містити:

- постановку задачі, вихідні дані, опис двохвибіркового t-тесту с різними дисперсіями;
- файл з результатами, висновки про істинність або хибність гіпотези.

Контрольні питання

1. Назвіть види статистичних гіпотез. Наведіть приклади. Сформулюйте нульову і альтернативну гіпотезу для завдання.

2. Охарактеризуйте поняття «статистичний критерій». У чому полягає відмінність між параметричними і непараметричними критеріями?

3. Які види статистичних критеріїв реалізовані в Excel? Як вони здійснюються?

Практична робота 6.2
Перевірка гіпотези про рівняння середніх двох груп в R
Хід роботи

1. Відкрийте R Studio.

Визначте два вектори, що представляють дані експериментальної та контрольної групи за допомогою функції scan. Функція scan () приймає значення з файлу або стандартного пристрою введення (клавіатури) і записує їх в змінні. Це означає, що після її запуску, треба просто вводити значення змінної, при цьому кожне значення вводиться з нового рядка. Її зручно використовувати для роботи з невеликим числом даних, які не записані в файл, доступний для імпорту в R. Аргумент **what** задає тип значень, які будуть міститися у змінній, а **n** – кількість цих значень. Це не єдині доступні аргументи, про решту можете дізнатися в довідці з функції. Слід зазначити, що число значень (**n**) можна не вказувати, в цьому випадку, для припинення введення даних необхідно двічі натиснути Enter.

2. Визначте числовий вектор для першої групи:

```
> gr1<-scan(what=numeric(), n=20)
1: 16
2: 17
3: 5
4: 7
5: 7
6: 8
7: 8
8: 6
9: 6
10: 8
11: 8
12: 15
13: 8
14: 10
```

```
15: 10
16: 13
17: 4
18: 4
19: 8
20: 9
Read 20 items
```

3. Визначте числовий вектор для другої групи:

```
gr2<-scan(what=numeric(), n=20)
1: 9
2: 7
3: 5
4: 6
5: 7
6: 8
7: 4
8: 6
9: 6
10: 8
11: 8
12: 12
13: 4
14: 4
15: 10
16: 5
17: 4
18: 3
19: 8
20: 3
Read 20 items
```

4. Проведіть порівняння груп за критерієм Стьюдента:

```
>t.test(gr1,gr2)

        Welch Two Sample t-test

data:  gr1 and gr2
t = 2.5096, df = 32.689, p-value = 0.01722

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:
 0.4725301 4.5274699

sample estimates:
mean of x mean of y
     8.85      6.35
```

Пояснимо результати розрахунків. Метод Уелша не вимагає рівності дисперсій двох груп і використовує інший спосіб обчислення ступенів свободи. У середовищі R для функції t.test () за замовчуванням використовується метод Уелша [http://voliadis.ru/r-two-samples].

Після назви тесту нам показано, які групи було протестовано: група1 та група 2.

```
data:  gr1 and gr2
```

Далі показано: t-статистику, число ступенів свободи та точне значення p.

```
t = 2.5096,  df = 32.689,  p-value = 0.01722
```

Далі наведено формулювання альтернативної гіпотези:

```
alternative hypothesis: true difference in means is not equal to
0
```

Після цього наведено довірчий 95% інтервал. Для зміни довірчого інтервалу на, припустимо, 99% довірчий інтервал можна використовувати аргумент conf.level=0.99.

```
95 percent confidence interval:
 0.4725301 4.5274699
```

Наприкінці, наведено середні значення для кожної з груп:

```
sample estimates:
mean of x mean of y
    8.85      6.35
```

Отримані результати розрахунків свідчать про значущі відмінності в середніх значеннях між першою і другою групою респондентів (p <0,05).

5. Об'єднайте два створених вектори в одну таблицю даних

```
> d <- data.frame(group1=gr1, group2=gr2)
> d
  group1 group2
1     16     9
2     17     7
3      5     5
4      7     6
5      7     7
6      8     8
```

7	8	4
8	6	6
9	6	6
10	8	8
11	8	8
12	15	12
13	8	4
14	10	4
15	10	10
16	13	5
17	4	4
18	4	3
19	8	8
20	9	3

Контрольні питання

1. Як вводяться дані за допомогою функції scan?
2. Які інші способи введення даних ви знаєте?
3. Яким чином проводиться порівняння середніх значень груп за критерієм Стьюдента?
4. Як інтерпретуються його результати?
5. Заповніть наступну таблицю за результатами розрахунків:

Міра	**Група1**	**Група2**
Середнє		
Спостереження		
t-статистика		
Число ступенів свободи		
p		
Рівень значущості		
Довірчий інтервал		

Практична робота 6.3
Перевірка гіпотез про значущість відмінності дисперсій в MS Excel

Мета роботи: навчитися використовувати пакет статистичного аналізу MS Excel для перевірки статистичних гіпотез.

Постановка задачі. В одному з досліджень [8] дітям давалися звичайні арифметичні задачі, а потім однієї випадково вибраної половині учнів повідомлялося, що вони не витримали випробування, а іншим – зворотне. Потім кожну дитину запитували, скільки секунд їй знадобилося б для рішення нового завдання. Експериментатор обчислював різницю між очікуваним часом рішення задачі (в сек.), який назвав учень, і результатами раніше виконаного завдання.

Гіпотеза, що перевіряється на рівні значущості 0,05 складається з того, що дисперсія сукупності дитячих оцінок постійна незалежно від того, повідомлялося дітям про погані результати випробування чи ні. Результати експерименту наведені в таблиці 5.3.1. Групі 1 повідомлялося про позитивні результати, групі 2 – про невдачу. У кожній групі було по 12 дітей.

Таблиця 5.3.1

Результати експерименту

Різниця між очікуваним часом рішення задачі та реальним, сек.	
Група 1	Група 2
8	40
9	10
10	30
11	20
7	38
6	22
5	30
9	30
12	45
15	26
9	25
6	25

Хід роботи

1. Увести вихідні дані згідно з таблицею 5.3.1.

2. Для кожного стовпця даних обчислити дисперсію за допомогою функції ДИСП. Подивитися, рівні чи ні виявилися дисперсії в двох групах.

3. Відкрити пакет аналізу: меню ***«Дані», «Аналіз даних», «Двохвибірковий F-тест для дисперсій*** (рис. 5.3.1).

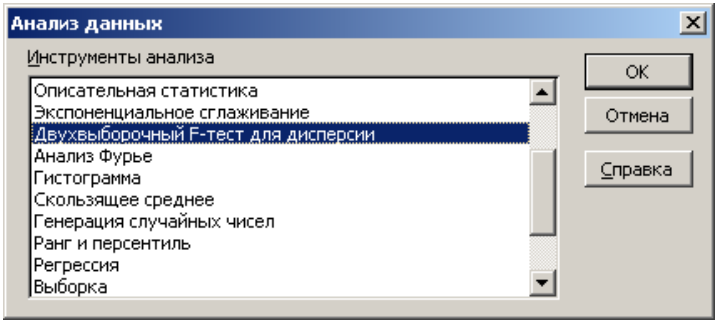


Рис. 6.3.1. Вікно «Двохвибірковий F-тест для дисперсій»

4. Прочитати довідку та записати її у зошит.

5. У вікні «***Двохвибірковий F-тест для дисперсій***» задати необхідні параметри.

6. Виписати таблицю з отриманими результатами (рис. 6.3.2).

Двохвибірковий F-тест для дисперсій		
	Група 1	*Група 2*
Середнє	8,916666667	28,41666667
Дисперсія	8,083333333	89,90151515
Спостереження	12	12
Df	11	11
F	0,089913205	
F критичне однобічне	0,354869911	

Рис. 6.3.2. Результати розрахунків двохввибіркового *F*-теста з відмінними дисперсіями

7. Зробити висновок про підтвердження або спростування гіпотези про рівність дисперсій. (Вказівка: тому що значення F-статистики 0,089 менше нижнього критичного значення 0,354, тобто потрапляє в критичну область, то гіпотеза про рівність дисперсій двох груп даних відхиляється на рівні значущості 0.05)

8. Зберегти таблицю в особистій папці.

Вимоги до звіту

Звіт про роботу повинен містити:

–постановку задачі, вихідні дані, опис двохвибіркового F-тесту, результати;

– висновки про істинність або хибність гіпотези;

– файл з результатами.

Контрольні питання

1. Наведіть приклади відомих вам параметричних і непараметричних статистичних критеріїв.

2. Сформулюйте нульову і альтернативну гіпотезу для завдання з даної лабораторної роботи.

3. Які види статистичних критеріїв реалізовані в Excel?

Практична робота 6.4

Перевірка гіпотези про тотожність дисперсій двох груп в *R*

1. Уведіть дані в електронні таблиці Excel (см. рис. 6.4.1)

	A	B	C
1	**bal**	**group**	
2	8	kg	
3	9	kg	
4	10	kg	
5	11	kg	
6	7	kg	
7	6	kg	
8	5	kg	
9	9	kg	
10	12	kg	
11	15	kg	
12	9	kg	
13	6	kg	
14	40	eg	
15	10	eg	
16	30	eg	
17	20	eg	
18	38	eg	
19	22	eg	
20	30	eg	
21	30	eg	
22	45	eg	
23	26	eg	
24	25	eg	
25	25	eg	
26			
27			

lab52 / Лист2 / Лист3

Рис. 6.4.1.Вихідні дані

2. Збережіть в форматі .CSV (розбивач – кома)
3. Імпортуйте у змінну data в R

```
data<-read.table("K:\\Rexample\\lab52.csv", sep=";", dec=",",
                 header=TRUE)
> data
   bal group
1    8    kg
2    9    kg
3   10    kg
4   11    kg
5    7    kg
6    6    kg
7    5    kg
8    9    kg
9   12    kg
10  15    kg
11   9    kg
12   6    kg
13  40    eg
14  10    eg
15  30    eg
16  20    eg
17  38    eg
18  22    eg
19  30    eg
20  30    eg
21  45    eg
22  26    eg
23  25    eg
24  25    eg
```

Функція read.table () імпортує таблицю з файлу в змінну data. Тип змінної – data.frame. Перший аргумент – це ім'я файлу, якщо файл лежить в робочій директорії, то повний шлях вказувати необов'язково. Параметр sep задає роздільник полів, dec – вказує роздільник в десятинних дробах. Цей аргумент

слід задавати кожен раз, коли десятковий роздільник в вихідному файлі не є точкою, як це прийнято в середовищі R. І останнім вказано аргумент header. Якщо задати йому значення TRUE, то перший рядок таблиці буде розцінюватися як заголовок та він буде використовуватися для визначення імен стовпців у змінній.

4. Для того, щоб провести порівняння груп за «класичним» методом Стьюдента, необхідно перевірити тотожність дисперсій

```
> var.test(bal ~ group, data)

        F test to compare two variances

data:  bal by group

F = 11.1218, num df = 11, denom df = 11, p-value = 0.0003821

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:
  3.201727 38.633914

sample estimates:
ratio of variances
          11.12184
```

Функція var.test () перевіряє гіпотезу тотожності дисперсій, а точніше того, що їх співвідношення дорівнює 1. В аргументах задано ім'я змінної, назва датафрейму data і т.і. формула моделі – це спеціальний синтаксис для завдання статистичних моделей, він використовується в багатьох функціях. Параметри var і class – це імена стовпців, що містять значення залежної і незалежної змінних відповідно. Оскільки p-value = 0.0003821, вважаємо, що дисперсії значуще різняться.

Практична робота 6.5
Перевірка гіпотез про рівняння середніх двох незалежних вибірок в SPSS

Мета роботи: навчитися проводити перевірку гіпотез про рівність середніх двох незалежних вибірок с допомогою Excel і SPSS та інтерпретувати результати

Постановка задачі

50 студентів університету були випадково розподілені за двома групами. 25 студентів вивчали курс аналізу даних за традиційною вузівської методикою, інші 25 - дистанційно. В кінці експерименту було проведено тест на засвоєння знань. Результати його наведені в таблиці 6.5.1. Чи свідчать результати тесту про відмінності в методиці викладання теми?

Таблиця 6.5.1

Тест в групі 1	Тест в групі 2
6	6
7	7
5	5
7	7
7	7
8	8
8	4
6	6
6	6
8	8
8	8
15	12
8	4
10	4
10	10
13	5
4	4
4	3
8	8
9	3
8	8
8,25	3
4	3
6	3
8	8

Нульова гіпотеза: Середні бали за результатами тесту в двох групах не відрізняються.

Рішення за допомогою Excel

Хід роботи

1. Відкрити Excel. Увести вихідні дані згідно таблиці 6.5.1.

2. Для кожного стовпця даних обчислити середнє і дисперсію допомогою функцій СРЗНАЧ і ДИСП. Подивитися рівні чи ні вийшли дисперсії.

3. Відкрити пакет аналізу: меню «*Дані», «Аналіз даних», «Двохвибірковий тест з різними дисперсіями».* Прочитати в ньому довідку і записати її в зошит.

4. У вікні «Двохбибірковий тест ...» задати необхідні параметри: інтервал змінної 1 (дані першої групи), інтервал змінної 2 (дані другої групи), гипотетична різниця середніх - 0, рівень значущості 0,05. Натиснути кнопку ОК.

5. Виписати таблицю з отриманими результатами (рис.6.5.1)

Двохвибірковий t-тест з відмінними дисперсіями		
	Змінна 1	*Змінна 2*
Середнє	7,65	6
Дисперсія	6,5	5,916666667
Спостереження	25	25
Гипотетична різниця середніх	0	
Df	48	
t-статистика	2,341269661	
t критичне двобічне	2,01063358	

Рис. 6.5.1. Таблиця з результатами

6. Порівняти емпіричне значення t-критерію Стьюдента (t-статистику) з критичним двостороннім.

7. Зробити висновок про прийняття або відкидання нульової гіпотези (якщо емпіричне значення більше критичного, то нульова гіпотеза відкидається).

8. Зберегти таблицю в особистій папці

Рішення за допомогою SPSS

Хід роботи

Перевірка гіпотези за допомогою SPSS включає наступні кроки:

1. Завдання двох змінних: залежної змінної test (бал по тесту), незалежної змінної-фактору group (група) з градаціями: «1 - перша», «2 друга»).

2. Введення вихідних даних (представлення даних в SPSS відрізняється від їх представлення в електронних таблицях) і наводиться на рис. 6.5.2.

	test	group
1	6,00	перша
2	7,00	перша
3	5,00	перша
4	7,00	перша
5	7,00	перша
6	8,00	перша
7	8,00	перша
8	6,00	перша
9	6,00	перша
10	8,00	перша

…

	test	group
26	6,00	друга
27	7,00	друга
28	5,00	друга
29	7,00	друга
30	7,00	друга
31	8,00	друга
32	4,00	друга
33	6,00	друга
34	6,00	друга
35	8,00	друга

…

Рис. 6.5.2. Представлення вихідних даних в SPSS

1. Вибір меню ***Analyze, Compare means, Independent-Samples T Test*** (для независимых выборок)

2. Перенесення змінної **бал за тестом** в список змінних для тестування, а змінної **група** – до змінної групування. (см. рис. 5.5.3)

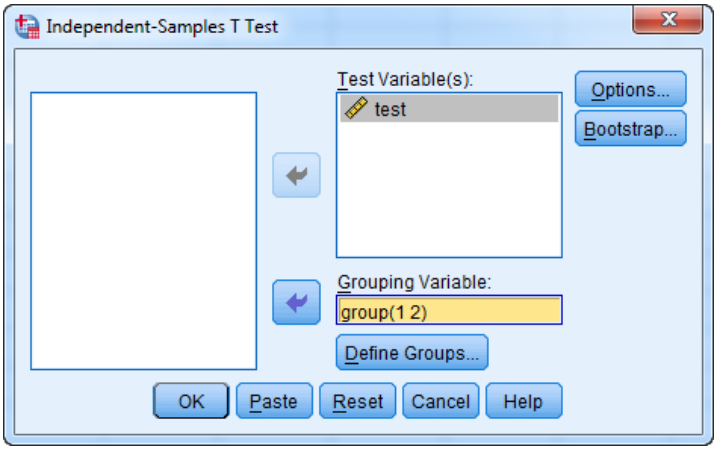


Рис. 6.5.3. Вікно Independent Samples T-Test

Клік на кнопці «**Define Groups**» задає кодировку, що показана на рис. 6.5.4.

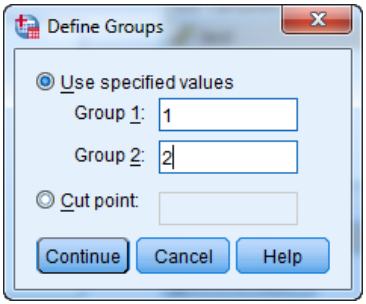


Рис. 6.5.4. Вікно «Define Groups»

3. Перегляд результатів у вікні виводу (рис.6.5.5).

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
test	перша	25	7,6500	2,54951	,50990
	друга	25	6,0000	2,43242	,48648

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	,261	,612	2,341	48	,023	1,6500	,7047	,2330	3,0670
Equal variances not assumed			2,341	47,894	,023	1,6500	,7047	,2329	3,0671

Рис. 6.5.5. Результати порівняння середніх двох вибірок

Для нашого прикладу значення критерію Стьюдента темпіричне = 2,341 і значимість p = 0,023. Нульова гіпотеза відхиляється, якщо значимість менше 0,05.

Практична робота 6.6
Перевірка гіпотез про тотожність середніх двох залежних вибірок

Мета роботи: навчитися проводити перевірку гіпотез про рівність середніх двох залежних вибірок в Excel і SPSS та інтерпретувати результати.

Постановка задачі. Визначте за даними такої таблиці, значимо чи ні розрізняються середні процентних часток працюючих жінок в 1968 і 1972 роках у 19 містах США (U.S. Bureau of Labor Statistics).

Таблиця 6.6.1.

Відсоток працюючих жінок в 1968 і 1972 роках в 19 містах США

Місто	1968	1972
N.Y.	0,42	0,45
L.A.	0,5	0,5
Chicago	0,52	0,52
Philadelphia	0,45	0,45
Detroit	0,43	0,46
San Francisco	0,55	0,55
Boston	0,45	0,6
Pitt.	0,34	0,49
St. Louis	0,45	0,35
Connecticut	0,54	0,55
Wash.D.C.	0,42	0,52
Cinn.	0,51	0,53
Baltimore	0,49	0,57
Newark	0,54	0,53
Minn/St. Paul	0,5	0,59
Buffalo	0,58	0,64
Houston	0,49	0,5
Patterson	0,56	0,57
Dallas	0,63	0,64

U.S. Bureau of Labor Statistics

Нульова гіпотеза: Середні процентних часток працюючих жінок в 1968 і 1972 роках в 19 містах США значущо не розрізняються.

Альтернативна гіпотеза: Середні процентних часток працюючих жінок в 1968 і 1972 роках в 19 містах США значимо різняться.

Рішення завдання за допомогою електронних таблиць Excel

Хід роботи

1. Завантажити Excel. Ввести вихідні дані згідно з таблицею 6.6.1.

2. Для кожного стовпця даних обчислити середнє і дисперсію допомогою функцій СРЗНАЧ і ДИСП.

3. Відкрити пакет аналізу: меню «Дані», «Аналіз даних», «Парний двохвибірковий t-тест для середніх». Прочитати в ньому довідку і записати її в зошит.

4. У вікні «Парний двухвибірковий t-тест ...» задати необхідні параметри: інтервал змінної 1, інтервал змінної 2, гіпотетична різниця середніх – 0, рівень значущості 0.05. Натиснути кнопку ОК (див.рис. 6.6.1).

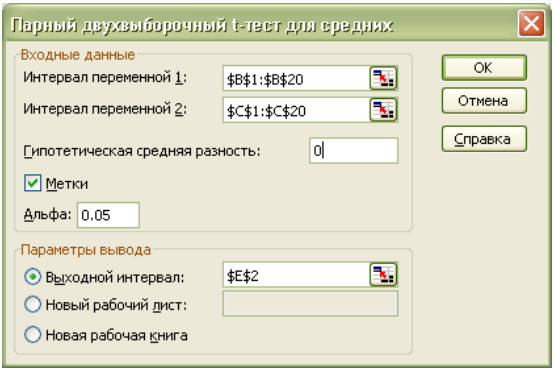


Рис. 6.6.1. Вікно «Парний двохвибірковий t- тест для середніх»

5. Виписати таблицю з отриманими результатами (таб. 6.6.2).

Таблиця 6.6.2.

Результати парного двохвибіркового t-теста для середніх

Парний двохвибірковий t-тест для середніх		
	1968	*1972*
Середнє	0.493158	0.526842105
Дисперсія	0.004623	0.005011696
Спостереження	19	19
Кореляція Пірсона	0.630073	
Гіпотетична різниця середніх	0	
df	18	
t-статистика	-2.4577	
P(T<=t) однобічне	0.012176	
t критичне однобічне	1.734064	
P(T<=t) двобічне	0.024353	
t критичне двобічне	2.100922	

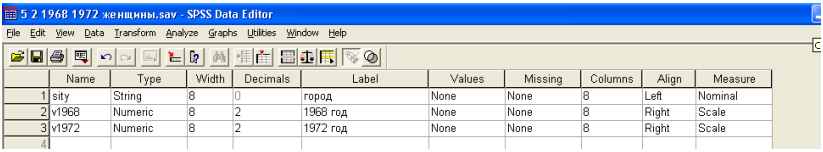
Порівняти емпіричне значення t-статистики Стьюдента з критичним (двобічним). Зробити висновок про прийняття або відкидання нульової гіпотези.

6. Зберегти таблицю в особистій папці.

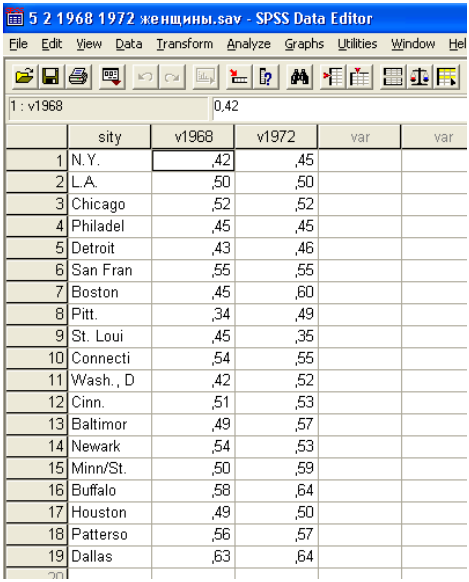
Рішення за допомогою SPSS

Хід роботи

1. Опишіть три змінні, як показано на рис. 6.6.2, заповніть даними таблиці 6.6.1 та збережіть у файлі women.sav.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	sity	String	8	0	город	None	None	8	Left	Nominal
2	v1968	Numeric	8	2	1968 год	None	None	8	Right	Scale
3	v1972	Numeric	8	2	1972 год	None	None	8	Right	Scale
4										



	sity	v1968	v1972	var	var
1	N.Y.	,42	,45		
2	L.A.	,50	,50		
3	Chicago	,52	,52		
4	Philadel	,45	,45		
5	Detroit	,43	,46		
6	San Fran	,55	,55		
7	Boston	,45	,60		
8	Pitt.	,34	,49		
9	St. Loui	,45	,35		
10	Connecti	,54	,55		
11	Wash., D	,42	,52		
12	Cinn.	,51	,53		
13	Baltimor	,49	,57		
14	Newark	,54	,53		
15	Minn/St.	,50	,59		
16	Buffalo	,58	,64		
17	Houston	,49	,50		
18	Patterso	,56	,57		
19	Dallas	,63	,64		

Рис. 6.6.2 Вкладка «Variable View» та «Data View» для задачі про відсоток працюючих жінок

2. Виберіть в меню Analyze, Compare means, Paired Samples T-Test (тест для парних вибірок).

3. Тепер в полі тестованих змінних потрібно виділити дві необхідні змінні і цю пару перенести в поле для спарених змінних. У нашому прикладі такими змінними є v1968 і v1972 (рис. 6.6.3.)

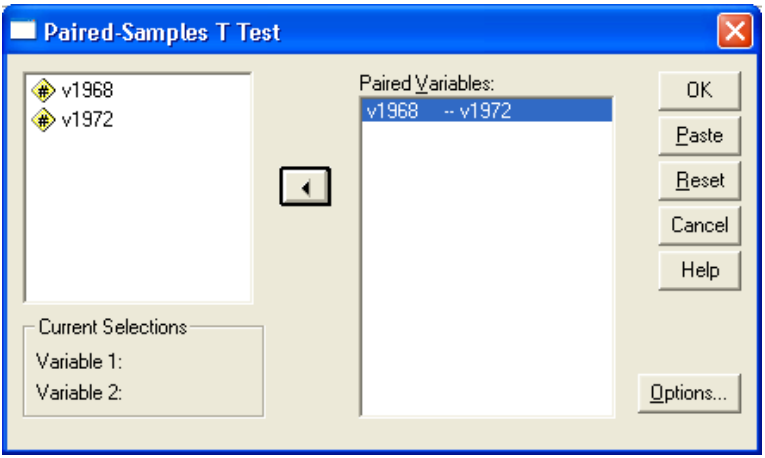


Рис. 6.6.3. Вікно «Paired-Samples T Test»

4. Запустіть тест на виконання натисканням клавіші ОК. У вікні перегляду з'являться результати розрахунку (див.рис. 6.6.3).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	V1968	,4932	19	,06799	,01560
	V1972	,5268	19	,07079	,01624

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1 V1968 - V1972	-,0337	,05974	,01371	-,0625	-,0049	-2,458	18	,024

Рис. 6.6.4. Результати розрахунків парного тесту для середніх

Зверніть увагу на емпіричне значення статистики Стьюдента – 2,458 і її значимість – 0,024. Нульова гіпотеза відкидається при значущості $p < 0,05$.

Отримана в наведеному прикладі величина $p = 0,024$ свідчить про значну різницю в середніх.

Завдання для самостійного рішення

1. За результатами оцінки розмірів дебіторської заборгованості, проведеної фінансовим директором фірми, було внесені зміни в її кредитну політику. По закінченню звітного кварталу було вирішено проаналізувати, чи дійсно зміни в кредитній політиці фірми впливають на скорочення терміну оплати дебіторської заборгованості. У таблиці 6.6.3 представлені дані про тривалість збору дебіторської заборгованості (в днях) при старій та новій кредитній політиці фірми.

Таблиця 6.6.3.

Вихідні дані

Тип кредитної політики	Тривалість збору дебіторської заборгованості (дні)
Стара	39 42 28 35 38 30 30 29 36 34
Нова	28 31 29 30 24 37 27 22 33 31

2. Майерс [37] в 1976 р провів цікавий експеримент з перевірки ефективності індивідуальної та групової форми тестування програм. Випробовувані однієї групи перевіряли програму з 63 операторів, призначену для обробки рядків, індивідуально, причому використовували набір тестів, складених після вивчення специфікації програми. Вони не мали лістингу програми. Випробовувані другої групи мали в розпорядженні ти самі засоби плюс лістинг. Випробовувані третьої групи були розбиті на бригади з трьох осіб, які мали тестувати програми власноруч методом перевірки коду. Результати експерименту наведено в таблиці 6.6.4.

Таблиця 6.6.4.

Результати експерименту з тестування програм

Характеристика	Індивідуальний перегляд + специфікація + термінал	Індивідуальний перегляд + специфікація + термінал + лістинг	Груповий перегляд + специфікації + лістинг
Середнє число знайдених помилок	4,5	5,4	5,7
Дисперсія	4,8	5,5	3,0
Мінімальне число помилок	1	2	3
Максимальне число помилок	7	9	9
Витрати на помилку, осіб.-хвилину	37	29	75

- Визначте за критерієм Стьюдента значимо чи відрізняється середнє число знайдених помилок в цих трьох групах, якщо в кожній з груп було по 25 осіб.
- Визначте, значимо чи різняться дисперсії і середні витрати на помилку трьох груп?

Вказівка. При порівнянні середніх двох незалежних вибірок емпіричне значення критерію Стьюдента обчислюється за формулою:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n1-1)\cdot S_1^2 + (n2-1)\cdot S_2^2}{n1+n2-2}\cdot\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

Де

X1	Середнє першої вибірки
X2	Середнє другої вибірки
S1	Стандартне відхилення першої вибірки
S2	Стандартне відхилення другої вибірки
n1	Обсяг першої вибірки
n2	Обсяг другої вибірки

та порівнюється з критичним

$$t_{1-(\alpha/2)\,n1+n2-2}$$

Якщо емпіричне значення перевищує критичне, то нульова гіпотеза про рівність середніх двох вибірок відкидається на рівні значущості α.

Вимоги до звіту

Звіт про роботу повинен містити: постановку задачі, вихідні дані, файли з результатами, висновки про істинність або хибність гіпотези.

Контрольні питання

1. Назвіть види статистичних гіпотез. Наведіть приклади. Сформулюйте нульову і альтернативну гіпотезу для завдання.
2. Охарактеризуйте поняття «статистичний крите-рій». У чому полягає відмінність між параметричними і не-параметричними критеріями?
3. Які види статистичних критеріїв реалізовані в Excel? Як вони здійснюються?
4. Як здійснюється перевірка гіпотез про рівність середніх для двох незалежних груп за допомогою SPSS?
5. Поясніть, що виводиться в таблицях «Статистика груп» і «Тест для незалежних вибірок» у вікні виводу?
6. Порівняйте з результатами, отриманими в Excel.
7. Які групи вважаються залежними або парними?
8. Як здійснюється перевірка гіпотез про рівність середніх для двох залежних груп?

Таблиці критичних значень. Критичні значення критерію Пірсона χ^2

df	P			df	P			df	P		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
1	3,842	6,635	10,829	**31**	44,993	52,203	61,118	**61**	80,232	89,591	100,887
2	5,992	9,211	13,817	**32**	46,202	53,498	62,508	**62**	81,381	90,802	102,165
3	7,815	11,346	16,269	**33**	47.408	54.789	63.891	**63**	82,529	92,010	103,442
4	9,488	13,278	18,470	**34**	48.610	56.074	65.269	**64**	83,675	93,217	104,717
5	11,071	15,088	20,519	**35**	49.810	57.356	66.641	**65**	84,821	94,422	105,988
6	12,593	16,814	22,462	**36**	51.007	58.634	68.008	**66**	85,965	95,626	107,257
7	14,068	18,478	24,327	**37**	52.201	59.907	69.370	**67**	87,108	96,828	108,525
8	15,509	20,093	26,130	**38**	53.393	61.177	70.728	**68**	88,250	98,028	109,793
9	16,921	21,669	27,883	**39**	54.582	62.444	72.080	**69**	89,391	99,227	111,055
10	18,309	23,213	29,594	**40**	55.768	63.707	73.428	**70**	90,531	100,425	112,317
11	19,677	24,729	31,271	**41**	56.953	64.967	74.772	**71**	91,670	101,621	113,577
12	21,028	26,221	32,917	**42**	58.135	66.224	76.111	**72**	92,808	102,816	114,834
13	22,365	27,693	34,536	**43**	59.314	67.477	77.447	**73**	93,945	104,010	116,092
14	23,688	29,146	36,132	**44**	60.492	68.728	78.779	**74**	95,081	105,202	117,347

15	24,999	30,583	37,706	**45**	61.668	69.976	80.107	**75**	96,217	106,393	118,599
16	26,299	32,006	39,262	**46**	62,841	71,221	81,431	**76**	97,351	107,582	119,850
17	27,591	33,415	40,801	**47**	64,013	72,463	82,752	**78**	99,617	109,958	122,347
18	28,873	34,812	42,323	**48**	65,183	73,703	84,069	**79**	100.749	111.144	123.595
19	30,147	36,198	43,832	**49**	66,351	74,940	85,384	**80**	101.879	112.329	124.839
20	31,415	37,574	45,327	**50**	67,518	76,175	86,694	**90**	113.145	124.116	137.208
21	32,675	38,940	46,810	**51**	68,683	77,408	88,003	**100**	124.342	135.807	149.449
22	33,929	40,298	48,281	**52**	69,846	78,638	89,308	**110**	135.480	147.414	161.582
23	35,177	41,647	49,742	**53**	71,008	79,866	90,609	**120**	146.567	158.950	173.618
24	36,420	42,989	51,194	**54**	72,168	81,092	91,909	**130**	157.610	170.423	185.573
25	37,658	44,324	52,635	**55**	73,326	82,316	93,205	**140**	138.613	181.841	197.450
26	38,891	45,652	54,068	**56**	74,484	83,538	94,499	**150**	179.581	193.207	209.265
27	40,119	46,973	55,493	**57**	75,639	84,758	95,790	**200**	233.994	249.445	267.539
28	41,343	48,289	56,910	**58**	76,794	85,976	97,078	**250**	287.882	304.939	324.831
29	42,564	49,599	58,320	**59**	77,947	87,192	98,365	**300**	341.395	359.906	381.424
30	43,780	50,904	59,722	**60**	79,099	88,406	99,649	**350**	394.626	414.474	437.487

Критичні значення коефіцієнта кореляції r Пірсона

(для перевірки неспрямованих альтернатив, n — обсяг вибірки)

n	P			n	P			n	P		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
5	0,878	0,959	0,991	**33**	0,344	0,442	0,547	**61**	0,252	0,327	0,411
6	0,811	0,917	0,974	**34**	0,339	0,436	0,539	**62**	0,250	0,325	0,408
7	0,754	0,875	0,951	**35**	0,334	0,430	0,532	**63**	0,248	0,322	0,405
8	0,707	0,834	0,925	**36**	0,329	0,424	0,525	**64**	0,246	0,320	0,402
9	0,666	0,798	0,898	**37**	0,325	0,418	0,519	**65**	0,244	0,317	0,399
10	0,632	0,765	0,872	**38**	0,320	0,413	0,513	**66**	0,242	0,315	0,396
11	0,602	0,735	0,847	**39**	0,316	0,408	0,507	**67**	0,240	0,313	0,393
12	0,576	0,708	0,823	**40**	0,312	0,403	0,501	**68**	0,239	0,310	0,390
13	0,553	0,684	0,801	**4!**	0,308	0,398	0,495	**69**	0,237	0,308	0,388
14	0,532	0,661	0,780	**42**	0,304	0,393	0,490	**70**	0,235	0,306	0,385
15	0,514	0,641	0,760	**43**	0,301	0,389	0,484	**80**	0,220	0,286	0,361
16	0,497	0,623	0,742	**44**	0,297	0,384	0,479	**90**	0,207	0,270	0,341

Критичні значення коефіцієнта кореляції r Пірсона (продовження)

17	0,482	0,606	0,725	**45**	0,294	0,380	0,474	**100**	0,197	0,256	0,324
18	0,468	0,590	0,708	**46**	0,291	0,376	0,469	**110**	0,187	0,245	0,310
19	0,456	0,575	0,693	**47**	0,288	0,372	0,465	**120**	0,179	0,234	0,297
20	0,444	0,561	0,679	**48**	0,285	0,368	0,460	**130**	0,172	0,225	0,285
21	0,433	0,549	0,665	**49**	0,282	0,365	0,456	**140**	0,166	0,217	0,275
22	0,423	0,537	0,652	**50**	0,279	0,361	0,451	**150**	0,160	0,210	0,266
23	0,413	0,526	0,640	**51**	0,276	0,358	0,447	**200**	0,139	0,182	0,231
24	0,404	0,515	0,629	**52**	0,273	0,354	0,443	**250**	0,124	0,163	0,207
25	0,396	0,505	0,618	**53**	0,271	0,351	0,439	**300**	0,113	0,149	0,189
26	0,388	0,496	0,607	**54**	0,268	0,348	0,435	**350**	0,105	0,138	0,175
27	0,381	0,487	0,597	**55**	0,266	0,345	0,432	**400**	0,098	0,129	0,164
28	0,374	0,479	0,588	**56**	0,263	0,341	0,428	**450**	0,092	0,121	0,155
29	0,367	0,471	0,579	**57**	0,261	0,339	0,424	**500**	0,088	0,115	0,147
30	0,361	0,463	0,570	**58**	0,259	0,336	0,421	**600**	0,080	0,105	0,134

Критичні значення коефіцієнта рангової кореляції r Спірмена

(по В.Ю. Урбаху, 1964)

Связь достоверна, если $r_{s\ эмп} \geq r_{s\ 0.05}$, и тем более достоверна, если $r_{s\ эмп} \geq r_{s\ 0.01}$.

n	*p*		n	*p*		n	*p*	
	0.05	**0.01**		**0.05**	**0.01**		**0.05**	**0.01**
5	0.94	-	17	0.48	0.62	29	0.37	0.48
6	0.85	-	18	0.47	0.60	30	0.36	0.47
7	0.78	0.94	19	0.46	0.58	31	0.36	0.46
8	0.72	0.88	20	0.45	0.57	32	0.36	0.45
9	0.68	0.83	21	0.44	0.56	33	0.34	0.45
10	0.64	0.79	22	0.43	0.54	34	0.34	0.44
11	0.61	0.76	23	0.42	0.53	35	0.33	0.43
12	0.58	0.73	24	0.41	0.52	36	0.33	0.43
13	0.56	0.70	25	0.39	0.51	37	0.33	0.43
14	0.54	0.68	26	0.39	0.50	38	0.32	0.41
15	0.52	0.66	27	0.38	0.49	39	0.32	0.41
16	0.50	0.64	28	0.38	0.48	40	0.31	0.40

Джерело таблиць: *Сидоренко Е. В.* Методы математической обработки в психологии - СПб.: ООО "Речь", 2000, [30]

Література

1. Адаменко Е. В. Математические методы в педагогике и психологии : учеб. пособие для студентов высш. учеб. заведений / Е. В. Адаменко. — Луганск : Альма-матер, 2008. — 96 с.
2. Адаменко Е. В. Кластерный анализ в педагогических исследованиях / Е. В. Адаменко, Л. Ф. Панченко // Вісник Луганського держ. пед. ун-ту ім. Тараса Шевченка. — 2002. — №8(52). — С. 6–10.
3. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин — М. : Финансы и статистика, 1989. — 607 с.
4. Айвазян С. А. Прикладная статистика. Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М. : Финансы и статистика, 1985. — 487 с.
5. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров / З. Брандт. — М. : Мир, 2003. — 688 с.
6. Бююль А. SPSS: искусство обработки информации / А. Бююль, П. Цефель – СПб. : ДиаСофтЮП, 2001. – 608 с.
7. Винстон У. А. Microsoft Excel: анализ данных и построение бизнес моделей / У. А. Винстон. — М. : Русская редакция, 2005. — 576 с.
8. Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стэнли. — М. : Прогресс, 1976. — 496 с.
9. Дубров А. М. Многомерные статистические методы / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. — М. : Финансы и статистика, 2003. — 352 с.
10. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл. — М. : Статистика, 1977. — 128 с.
11. Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных / М. Дэйвисон. — М. : Финансы и статистика, 1988. — 254 с.

12. Ефимова М. Р. Общая теория статистики : учебник/ М. Р. Ефимова, Е. В. Петрова, В. Н. Румянцев. — М.: ИНФРАМ, 2000. — 416 с.

13. Иберла К. Факторный анализ / К. Иберла. — М. : Статистика, 1980. — 398 с.

14. Кендэлл М. Ранговые корреляции / М. Кендэлл. — М. : Статистика, 1975. — 12 с.

15. Левин Д. М. Статистика для менеджеров с использованием Microsoft Excel, 4-е изд. / Д.М. Левин, Д.С. Кребиль і др. — М. : Вильямс, 2004. — 1312 с.

16. Малхотра Н. К. Маркетинговые исследования. Практическое руководство, 3-е издание. / Н. К. Малхотра. — М. : Вильямс, 2002. — 960 с.

17. Мидлтон М. Р. Анализ статистических данных с использованием Microsoft Excel / М. Р.Мидлтон. — М. : Бином, 2005. — 296 с.

18. Наследов А. Д. SPSS: компьютерный анализ данных в психологии и социальных науках./ А. Д. Наследов. — СПб. : Питер, 2005. — 416 с.

19. Паниотто В. И. Количественные методы в социологических исследованиях / В. И. Паниотто, В. С. Максименко. — Київ : Наукова думка, 1982. — 272 с.

20. Паніотто В. І. Статистичний аналіз соціологічних даних. / В. І. Паниотто, В. С. Максименко, Н. М. Харченко. — Київ : КМ Академія, 2004. — 270 с.

21. Панченко Л.Ф. Використання вільного програмного забезпечення для навчання студентів аналізу даних / Л. Ф. Панченко // Вісн. Луган. нац. ун-ту імені Тараса Шевченка : Педагогічні науки. — 2010. — № 17(204). — С. 67–77.

22. Панченко Л. Ф. Математические методы в психологии : метод. указания к лабораторному практикуму / Л. Ф. Панченко. — Луганск, Альма-матер, 2005. — 60 с.

23. Панченко Л. Ф. Компьютерный анализ данных : учеб. пособие для студентов высш. учеб. заведений. / Л.Ф. Панчен-

ко, Е.В. Адаменко. — Луганск : Изд-во ГУ „ДЗ ЛНУ імені Тараса Шевченка”, 2010. — 188 с.

24. Панченко Л. Ф. Практикум по анализу данных : учеб. пособие для студентов высших учебных заведений / Л. Ф. Панченко. — Луганск, Изд-во ГУ «ЛНУ имени Тараса Шевченко», 2013. — 269 с.

25. Петрунин Ю. Ю. Информационные технологии анализа данных. / Ю. Ю. Петрунин. — М. : КДУ, 2008. — 292 с.

26. Плис А. И. Практикум по прикладной статистике в среде SPSS / А. И. Плис, Н. А. Сливина. — М. : Финансы и статистика, 2004. — 284 с.

27. Поллард Дж. Справочник по вычислительным методам статистики / Дж. Поллард. — М. : Финансы и статистика, 1982. — 344 с.

28. Саймон Д. Анализ данных в Excel / Д. Саймон. — М.: Вильямс, 2004. — 528 с.

29. Сидоренко Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. — СПб. : Речь, 2000. — 350 с.

30. Статистическое обеспечение маркетинга продукта. — М. : МЭСИ, 2000. — 150 с.

31. Таганов Д. Н. SPSS: статистический анализ в маркетинговых исследованиях / Д. Н. Таганов. — СПб. : Питер, 2004. — 192 с.

32. Технологии анализа данных / Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. — СПб. : BHV-Петербург, 2007. — 384 с.

33. Тюрин Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин. — М. : Инфра-М, 2003. — 544 с.

34. Факторный, дискриминантный и кластерный анализ. — М. : Финансы и статистика, 1989. — 216 с.

35. Чекотовский Э. В. Графический анализ статистических данных в Microsoft Excel 2000 / Э.В. Чекотовский. — Киев : Диалектика, 2002. — 464 с.

36. Харман Г. Современный факторный анализ / Г. Харман. — М. : Статистика, 1972. — 489 с.

37. Шнейдерман Б. Психология программирования: Человеческие факторы в вычислительных и информационных системах / Б. Шнейдерман. — М. : Радио и связь, 1984. — 304 с.

38. Эддоус М. Методы принятия решений / М. Эддоус, Р. Стэнсфилд — М. : ЮНИТИ, 1997. — 590 с.

39. StatSoft, Inc. (2001). Электронный учебник по статистике, StatSoft [Електронний ресурс]. — Режим доступу: http://www.statsoft.ru/home/textbook/default.htm.

Додаткова література

40. Панченко Л.Ф. Підготовка майбутніх соціологів до комп'ютерного аналізу демографічних процесів та структур // Information Technologies and Learning Tools. – 2018. – Vol.65(3). – С.166–183.

41. Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R / Роберт И. Кабаков. — М. : ДМК Пресс, 2014. — 588 с.

42. Рекомендации по преподаванию программной инженерии и информатики в университетах — Software Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering; Computing Curricula 2001: Computer Science: пер.с англ. — М.: ИНТУИТ.РУ «Интернет-Университет Информационных Технологий », 2007.

43. Шипунов А. Б. и др. Наглядная статистика. Используем R! / А. Б. Шипунов. — М. : ДМК ПРЕСС, 2012. — 298 с.

44. R–project [Електронний ресурс]. — Режим доступу: http://www.r-project.org/about.html.

45. Quick-R [Електронний ресурс]. — Режим доступу: http://www.statmethods.net/

46. Using R for Multivariate Analysis [Електронний ресурс]. — Режим доступу: http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

Панченко Л. Ф. Математичні та статистичні методи аналізу соціологічної інформації : Практикум. – Навчальний посібник.

У навчальному посібнику розглядаються теоретичні та практичні питання комп'ютерного аналізу соціологічної інформації з використанням електронних таблиць MS Excel, статистичного пакету SPSS, середовища R. Посібник включає такі розділи аналізу як описова статистика, основи кореляційного й регресійного аналізу, основи теорії імовірності та перевірки гіпотез. Кожен розділ містить короткі теоретичні відомості; завдання з аналізу даних з детальними поясненнями для виконання в середовищах Microsoft Excel, SPSS, R; завдання для самостійної роботи.

Призначений для студентів та магістрантів спеціальностей „Соціологія", „Соціальна робота", а також усіх інших спеціальностей, які пов'язані з аналізом даних на комп'ютері.

Ключові слова: аналіз даних, комп'ютерний аналіз даних, описова статистика, кореляція, регресія, перевірка гіпотез, Microsoft Excel, SPSS, R.

Панченко Л. Ф. Математические и статистические методы анализа социологической информации: Практикум. – Учебное пособие.

В учебном пособии рассматриваются теоретические и практические вопросы компьютерного анализа данных с использованием электронных таблиц MS Excel, статистического пакета SPSS, свободно распростарняемой среды R. Пособие включает такие разделы анализа как описательная статистика, основы корреляционного и регрессионного анализа, теории вероятности, проверки гипотез. Каждый раздел содержит краткие теоретические сведения; задачи по анализу данных с подробными пояснениями для выполнения в среде Microsoft Excel, SPSS, R; задания для самостоятельной работы.

Предназначено для студентов и магистрантов специальностей „Социология”, „Социальная работа”, а также всех других специальностей, которые связаны с анализом данных на компьютере.

Ключевые слова: анализ данных, компьютерный анализ данных, описательная статистика, корреляция, регрессия,основы теории вероятностей, проверка гипотез, Microsoft Excel, SPSS, R.

Panchenko L. F. Mathematical and statistical methods of analyzing sociological information. – Textbook.

The textbook discusses the theoretical and practical aspects of computer data analysis in sociology using spreadsheets Microsoft Excel, the statistical package SPSS, environment R. The manual includes such sections as descriptive statistics, the basics of correlation and regression analysis, probability theory, hypothesis testing. Each section of the textbook contains short theoretical information; examples of data analysis with detailed explanations for execution in the Microsoft Excel, SPSS, R environments; tasks for self study.

The textbook is addressed for “Sociology” students as well as for students of other specialties that are associated with the computer data analysis.

Key words*:* data analysis, computer data analysis, descriptive statistics, correlation, regression, probability theory, hypothesis testing, Microsoft Excel, SPSS, R.

Навчальне видання

ПАНЧЕНКО Любов Феліксівна

Математичні та статистичні методи аналізу соціологічної інформації

Навчальний посібник
для здобувачів ступеня бакалавра
спеціальності 054 «Соціологія»

Розглядаються теоретичні та практичні питання комп'ютерного аналізу соціологічних даних з використанням електронних таблиць MS Excel, статистичного пакету SPSS, середовища R, які складаються з таких розділів аналізу як описова статистика, основи кореляційного й регресійного аналізу, теорії імовірності, перевірки гіпотез.

Для студентів та магістрантів спеціальностей „Соціологія", «Соціальна робота», а також усіх інших спеціальностей, які пов'язані з аналізом даних на комп'ютері.

За редакцією автора
Комп'ютерний макет – Л. Ф. Панченко

Видавець і виготовлювач
Видавництво КПІ імені І.Сикорського"
, м. Київ, т/ф: (044).
e-mail:
Свідоцтво суб'єкта видавничої справи.