МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М.Є. Жуковського
«Харківський авіаційний інститут»

В.Л. Петрик, М.А. Голованова, С.М. Мельніков

СТАТИСТИКА В SPSS

Навчальний посібник

Харків «ХАІ» 2010

УДК 681. 3. 06 + 6196

Петрик В.Л. Статистика в SPSS : навч. посіб. / В.Л. Петрик, М.А. Голованова, С.М. Мельніков. – Х. : Нац. аерокосм. ун-т «Харк. авіац. ін-т», 2010. – 172 с.

Розглянуто комплекс статистичних методів збору, оброблення й аналізування статистичної інформації, теоретичні й методичні основи будування системи статистичних показників, які використовують для вивчення закономірностей суспільних явищ. Особливу увагу приділено статистичній методології, можливостям її використання в умовах суттєвих змін в економіці. Наведено приклади розв'язання задач за допомогою програми для оброблення статистичної інформації SPSS.

Для студентів та аспірантів економічних спеціальностей «Фінанси», «Економіка», «Менеджмент організацій», «Маркетинг» при виконанні домашніх завдань, курсових і дипломних робіт.

Іл. 117. Табл. 21. Бібліогр.: 6 назв

Рецензенти: д-р техн. наук, проф. Ю.Ю. Шабанов-Кушнаренко,
д-р техн. наук, проф. В.О. Філатов

Вступ

В усьому світі зростає інтерес до статистики. У нашій країні цю увагу тим більше загострено у зв'язку зі здійсненням економічних реформ, що стосуються інтересів усіх людей. Перехід до ринкової економіки надає нового змісту роботі маркетологів, економістів, менеджерів. Це ставить підвищені вимоги до рівня їхньої статистичної підготовки. Оволодіння статистичною методологією – це одна з неодмінних умов пізнання кон'юнктури ринку, вивчення тенденцій і прогнозування попиту й пропозиції, прийняття оптимальних рішень на всіх рівнях комерційної діяльності на ринку товарів і послуг.

Найпоширенішою програмою для оброблення статистичної інформації є SPSS. Пакет SPSS містить усі основні розділи щодо оброблення й аналізування даних, а також засобів табличного й графічного подання отриманих результатів.

1. СТАТИСТИКА

Оволодіння прийомами роботи з програмою SPSS потребує попередніх знань в області статистики. Коротко зупинимося на деяких основних поняттях, з якими неодмінно має бути ознайомленний користувач SPSS. До них належать середні, варіація, попередні оцінки, які виконуються перед проведенням будь-якого статистичного тесту, класифікація змінних за статистичними шкалами, перевірка наявності нормального розподілу й виділення незалежних і залежних вибірок.

1.1. Предмет і метод статистичної науки

Статистика (від лат. status – стан) – наука, у якій вивчаються розміри й кількісні співвідношення масових суспільно-економічних явищ і процесів у нерозривному зв'язку з їхнім змістом.

Предмет статистики – кількісна сторона масових суспільних явищ у нерозривному зв'язку з їхньою якісною стороною або їхнім змістом.

Предмет статистики вивчається за допомогою різноманітних методів, сукупність яких утворює ***статистичну методологію***.

Уся різноманітність статистичних методів вивчення комерційної діяльності систематизується згідно з їх цільовим вживанням у послідовно виконуваних трьох ***основних стадіях економіко-статистичного дослідження***:

– статистичне спостереження;

– статистичне зведення й групування первинної інформації;

– аналізування статистичної інформації.

Статистичне спостереження – це спланований, систематичний і науково організований збір масових даних про різноманітні соціально-економічні явища й процеси.

Для здійснення спостереження застосовують ***методи масового спостереження***, які дають можливість отримати певні значення досліджуваних ознак від кожної одиниці статистичної сукупності шляхом реєстрації їх на підставі розробленої програми.

Статистичне зведення й групування первинних даних необхідні для систематизації матеріалу статистичного спостереження, що полягає у перевірці даних, їх групуванні за певними ознаками; підведенні групових і загальних підсумків, розрахунку різних показників, проектуванні таблиць і внесенні до них даних. На цій стадії застосовують ***метод групувань***, який дає можливість виділити в досліджуваній сукупності соціально-економічні типи явищ, охарактеризувати їхню структуру, виявити взаємозв'язки й взаємозалежності між показниками.

При ***аналізуванні статистичної інформації*** передбачається проведення аналізу даних на основі обчислення узагальнюючих статистичних показників: абсолютних, відносних і середніх величин, статистичних коефіцієнтів, показників варіації ознак і динаміки явищ, індексів і показників, що характеризують силу зв'язку між явищами й т.ін.

Предмет статистики вивчають за допомогою певних категорій – ***понять, які відображають загальні й істотні*** властивості, ознаки, зв'язки й відносини предметів і явищ об'єктивного світу.

Основними категоріями теорії статистики є такі.

Статистична сукупність – це велика кількість одиниць явища, які об'єднані відповідно до задачі дослідження єдиною якісною основою, загальним зв'язком, але відрізняються один від одного окремими ознаками (наприклад, сукупність промислових підприємств України, сукупність сімей, сукупність підприємств, фірм, об'єднань тощо).

Сукупність називають однорідною, якщо один або декілька істотних ознак її об'єктів, що вивчаються, є загальними для всіх одиниць. Сукупність, до якої входять явища різного типу, вважається різнорідною.

Одиниця статистичної сукупності – первинний елемент статистичної сукупності, який є носієм ознак, що підлягають реєстрації. Наприклад, для перепису населення одиницею сукупності є кожна людина; для обстеження проданих на біржі квартир – кожна продана квартира.

Ознака – якісна особливість одиниці сукупності – відмінна риса, властивість, якість, що є характерною для окремих одиниць, об'єктів (явищ). Наприклад, ознаками промислового підприємства є обсяги виробництва, розмір основних виробничих фондів, кількість персоналу та ін.; демографічні й соціально-економічні ознаки людини: вік, рівень освіти, професія, стать і т.ін.

Статистичний показник – узагальнена характеристика соціально-економічного явища або процесу, в якій поєднуються якісна й кількісна характеристики останнього (наприклад, кількість населення, продукція промислового підприємства, рівень продуктивності праці, рівень рентабельності й т.ін.).

Система статистичних показників – це сукупність показників, що відображає взаємозв'язки, які об'єктивно існують між явищами.

1.2. Статистичне спостереження

Статистичне спостереження – перша стадія статистичного дослідження, науково організований за єдиною програмою облік фактів, що характеризують явища й процеси суспільного життя, а також збір отриманих на основі цього обліку масових даних.

Статистичні дані – це масові системні кількісні характеристики соціально-економічних явищ і процесів. Статистичні дані мають бути достовірними, повними (за обсягом і суттю), своєчасними, порівнянними за часом і простором, доступними.

Об'єкт спостереження – сукупність соціально-економічних явищ і процесів, які підлягають дослідженню, або точні межі, в яких реєструватимуться статистичні дані.

Статистичне спостереження може бути:

– первинним – реєстрація даних, що надходять безпосередньо від об'єкта, який їх продукує (наприклад, поточний облік);

– вторинним – збір раніше зареєстрованих і оброблених даних (наприклад, банківських звітів).

Одиниця статистичного спостереження – це та первинна ланка, з якої мають бути отримані необхідні статистичні дані.

Програма спостереження – це перелік чітко сформульованих питань, на які необхідно отримати відповіді, або перелік ознак і показників, що підлягають реєстрації.

У більшості випадків перед застосуванням статистичного тесту необхідно з'ясувати такі моменти:

– до якої статистичної шкали належить змінна;

– якщо йдеться про змінні з інтервальною шкалою, то чи підкоряються вони закону нормального розподілу;

– порівнювані вибірки є залежними чи незалежними.

Типи статистичних шкал

Шкала – засіб впорядкування й кількісного вираження ознак. Використовують такі види шкал:

1) номінальна – шкала найменувань, з допомогою якої встановлю-

ється відношення подібності елементів, для якої порядок розташування ознак значення не має, наприклад: класифікатор сфер економічної діяльності, перелік форм власності, видів підприємницької діяльності й т.ін.;

2) порядкова (рангова) – шкала, яка встановлює послідовність інтенсивності прояву ознаки; застосовується під час визначення ступеня економічного ризику підприємництва, рівня кваліфікації робітників або ставлення респондентів до якогось явища або процесу. Ознаки порядкової шкали оцифровують шляхом присвоєння рангів (балів) окремим значенням зі збільшенням або зменшенням їхньої інтенсивності;

3) метрична – кількісна шкала, в основу якої покладено результати безпосереднього вимірювання. Метричною шкалою визначаються обсяги виробництва й реалізації продукції, розміри капіталу, кількість зайнятих у виробництві осіб, кількість і вартість приватизованих об'єктів тощо.

Розглянемо різницю між цими видами шкал на прикладі кодувальної табл. 1.1. В емпіричному дослідженні можуть траплятися, наприклад, змінні, кодування яких наведено в табл. 1.1.

Таблиця 1.1

Змінна	Кодування змінних
Стать	1 = чоловічий
	2 = жіночий
Сімейний стан	1 = неодружений/незаміжня
	2 = одружений/заміжня
	3 = удівець/удова
	4 = розведений/розведена
Паління	1 = який не палить
	2 = який зрідка палить
	3 = який інтенсивно палить
	4 = який дуже інтенсивно палить
Місячний дохід	1 = до 3000 грн
	2 = 3001 – 5000 грн
	3 = понад 5000 грн

Розглянемо графу «Стать», з якої видно, що призначення відповідності цифр 1 і 2 жіночої й чоловічої статі є абсолютно довільним, їх можна поміняти місцями або позначити іншими цифрами. У такому випадку говорять про змінні, що належать до номінальної шкали. У цьому прикладі розглядається змінна з номінальною шкалою, що має дві категорії. Така змінна має ще одне назву – дихотомічна. Така сама ситуація й зі змінною «Сімейний стан». Тут також відповідність між числами й категоріями сімейного стану не має ніякого емпіричного значення. Однак на відміну від змінної «Стать», ця змінна не є дихотомічною – вона має чотири категорії замість двох. Можливості оброблен-

ня змінних, таких, що належать до номінальної шкали, є дуже обмеженими. Власне кажучи, можна провести тільки частотний аналіз таких змінних. Наприклад, розраховувати середнє значення для змінної «Сімейний стан» немає сенсу. Змінні, що належать до номінальної шкали, часто використовуються для угрупувань, за допомогою яких сукупна вибірка розбивається за категоріями цих змінних. У часткових вибірках проводяться однакові статистичні тести, результати яких потім порівнюються один з одним.

Як наступний приклад розглянемо змінну «Паління». Тут кодовим цифрам присвоюється емпіричне значення в тому порядку, в якому вони розташовані в списку. Змінну «Паління», як наслідок, сортовано в порядку значущості від низу до верху: помірний курець палить більше, ніж той, що не палить, а той, що сильно палить – більше, ніж помірний курець і т. д. Такі змінні, для яких використовуються числові значення, що відповідають поступовому зміненню емпіричної значущості, належать до порядкової шкали.

До класичних прикладів змінних з порядковою шкалою належать також змінні, які отримано внаслідок об'єднання величин в класи, як, наприклад, клас «Місячний дохід».

Окрім частотного аналізу, змінні з порядковою шкалою дають можливість також обчислити певні статистичні характеристики, такі, як медіани. У деяких випадках можна обчислити середнє значення. Якщо треба встановити зв'язок (кореляцію) з іншими змінними такого роду, для цієї мети можна використати коефіцієнт рангової кореляції.

Для порівняння різних вибірок змінних, що належать до порядкової шкали, можна застосовувати непараметричні тести, формули яких оперують рангами.

Такі змінні, у яких різниця (інтервал) між двома значеннями має емпіричну значущість, належать до інтервальної шкали. Вони можуть оброблятися будь-яким статистичним методами без обмежень. Так, наприклад, середнє значення є повноцінним статистичним показником для характеристики таких змінних.

Нарешті, досягнуто найвищої статистичної шкали, на якій емпіричної значущості надбуває й відношення двох значень. Прикладом змінної, що належить до такої шкали, є вік: якщо Максу 30 років, а Міколі 60, можна сказати, що Мікола удвічі старший за Макса. Шкалу, до якої належать дані, називають шкалою стосунків. До цієї шкали належать усі інтервальні змінні, які мають абсолютну нульову точку. Тому змінні, що належать до інтервальної шкали, як правило, мають і шкалу відношень.

Підводячи підсумки, можна сказати, що існує чотири види статистичних шкал, на яких можуть порівнюватися числові значення (табл. 1.2).

Таблиця 1.2

Статистична шкала	Емпірична значущість
Номінальна	–
Порядкова	Порядок чисел
Інтервальна	Різниця чисел
Шкала відношень	Відношення чисел

На практиці, у тому числі в SPSS, різниця між змінними, що належать до інтервальної шкали і шкали стосунків, зазвичай є несуттєвою, тобто надалі майже завжди йтиметься про змінні, що належать до інтервальної шкали.

Користувач SPSS має добре розбиратися у видах статистичних шкал і під час вибору методу звертати увагу на те, щоб були визначеними відповідні види шкал.

1.3. Зведення і групування статистичних даних

Статистичне зведення – упорядкування, систематизація й наукове оброблення даних з метою отримання узагальненої характеристики досліджуваного об'єкта або процесу.

Статистичне зведення здійснюється за науково розробленою програмою, яка містить визначення груп і підгруп, систем показників, видів таблиць.

За складністю побудови зведення поділяють на прості й групові.

Просте підсумкове зведення – не передбачається попереднього розподілу на групи отриманих відомостей, а визначається загальний підсумок усіх одиниць сукупності або загальний обсяг досліджуваного показника. Наприклад, щоб знайти загальну кількість студентів в Україні, достатньо скласти дані про кількість студентів у всіх вищих навчальних закладах.

Групове зведення – передбачається попередній розподіл одиниць на групи (наприклад, рентабельні й збиткові підприємства), що дає можливість підрахувати кількість одиниць сукупності й обсяг досліджуваної ознаки в кожній групі.

Групування – це розбиття сукупності на однорідні групи на підставі розподілу всієї сукупності досліджуваного явища на окремі групи за найбільш істотними ознаками.

Основні задачі, які вирішуються за допомогою статистичних групувань: виділення соціально-економічних типів явищ; вивчення структури досліджуваних явищ і структурних змін; дослідження взаємозв'язку і залежності між ознаками суспільних явищ.

Згідно з цими задачам розрізняють типологічні, структурні й аналітичні групування.

Типологічне групування призначене для виділення соціально-економічних типів явищ у якісно неоднорідній сукупності, визначення істотних відмінностей між ними і загальних ознак для всіх груп.

Типологічне групування використовують при вивченні поділу підприємств за формами власності й суспільного виробництва, за економічним призначенням продукції, при групуванні населення за соціальними групами тощо.

Структурне групування характеризує структуру якісно однорідної сукупності за якою-небудь однією варіювальною ознакою, яка дає можливість описати складові частини сукупності й проаналізувати структурні зміщення (наприклад, вивчення підприємств за галузями виробництва, розмірами основних виробничих фондів, рівнем механізації виробництва, кількістю працівників, обсягом продукції, для дослідження складу населення – за статтю, віком, національністю, походженням тощо).

Структурне групування, як і типологічне, можна здійснювати з урахуванням атрибутивних і кількісних ознак.

Типологічне групування відрізняється від структурного лише метою дослідження, за формою ж вони цілком збігаються.

Аналітичне групування дає можливість оцінювати взаємозв'язки двох і більше взаємодійних ознак, з яких одна розглядається як наслідок, а інші – як фактор. У кожній групі факторної ознаки визначається середня величина результативної ознаки.

За наявності зв'язку між ознаками середні групові систематично збільшуються (прямий зв'язок) або зменшуються (зворотний зв'язок).

Комбінаційне групування містить групи, що виділені за однією ознакою, які підрозділяються на підгрупи за іншою ознакою.

При проведенні групування доводиться вирішувати такі задачі:

1) виділення групувальної ознаки;

2) визначення кількості груп і величини інтервалів;

3) за наявності декількох групувальних ознак – опис того, як вони комбінуються між собою;

4) установлення показників, за якими характеризують групи.

Групувальна ознака – це ознака, за якою відбувається об'єднання окремих одиниць сукупності в однорідні групи.

Ознаки розрізняються за способами їх вимірювання й іншими особливостями, що впливають на прийоми статистичного вивчення. Це дає підставу для класифікації ознак (табл. 1.3).

Таблиця 1.3

За формою виразу	За способом обчислення	Відносно об'єкта	За характером варіації	За ознакою часу	За роллю у взаємозв'язку явищ
Атрибутивні: - номінальні, - порядкові Варіаційні	Первинні Похідні (вторинні)	Прямі Непрямі	Альтер-нативні Варіаційні	Моментні Інтервальні	Факторні Результативні

Дамо пояснення деяких ознак.

Атрибутивні ознаки характеризують властивості, якості явищ і не мають кількісного виразу (стать, національність, освіта, професія).

Номінальні – описові ознаки, за якими не можна ранжувати дані.

Порядкові – ознаки, за якими можна ранжувати, упорядковувати дані.

Кількісні (варіаційні) ознаки набувають різного кількісного виразу в певних одиницях (вік людини, заробітна плата, дохід фірми, кількість працівників).

Альтернативні ознаки – це протилежні за значенням варіанти ознаки, тобто ті, які можуть набувати тільки двох значень.

Моментні ознаки характеризують об'єкт у якийсь момент часу, встановлений планом статистичного дослідження.

Інтервальні ознаки характеризують результати процесів, явищ за певний час (рік, місяць, доба), але не на момент часу.

Факторні ознаки впливають на інші ознаки.

Результативні ознаки – це ознаки, розмір і динаміка яких формуються під впливом інших ознак.

Після визначення групувальної ознаки важливим кроком є поділ одиниць сукупності на групи. Для цього потрібно визначити кількість утворюваних груп і розмір інтервалу.

Інтервал обкреслює кількісні межі груп. Як правило, він є проміжком між максимальними й мінімальними значеннями ознаки в групі.

Кількість груп при рівних інтервалах визначають за формулою Стерджеса

$$N = 1 + 3{,}322 \cdot \lg n,$$

де n – кількість елементів сукупності.

Для однакових інтервалів їхню величину можна визначити за формулою

$$I = \frac{x_{\max} - x_{\min}}{n},$$

де x_{max}, x_{min} – відповідно максимальне й мінімальне значення ознаки;

n – кількість елементів сукупності.

Нижня межа першого інтервалу a_0 може збігатися з мінімальним значенням ознаки або визначатися за формулою

$$a_0 = x_{\min} - \frac{l}{2}.$$

Верхню межу першого інтервалу й нижню межу другого розраховують за формулою

$$a_1 = a_0 + l.$$

Межі подальших інтервалів знаходять за формулою

$$a_{j+1} = a_j + l.$$

Основою будь-якого групування є ряд розподілу.

Ряд розподілу – групування, в якому для характеристики груп, упорядкованих за значенням групувальної ознаки, застосовується один показник – чисельність групи.

Ряд розподілу складається з двох елементів – варіантів і частот.

Варіанти – окремі значення групувальної ознаки.

Частоти – числа, що показують, скільки разів повторюються окремі значення варіантів або кожної групи варіаційного ряду.

Замість частот може бути ***частість*** – частка частоти від загальної чисельності сукупності, яку виражено коефіцієнтом або відсотком.

Накопичені частоти або частості називають ***кумулятивними***.

Обсяг сукупності – сума всіх частот, що визначає кількість елементів усієї сукупності.

Ряди розподілу, побудовані за атрибутивною ознакою, називають ***атрибутивними***, наприклад: поділ населення за статтю, зайнятістю, національністю, професією тощо.

Ряди розподілу, побудовані за кількісною ознакою, називають ***варіаційними рядами***, наприклад: поділ населення за віком, робітників – за стажем роботи, за заробітною платою тощо.

Варіаційні ряди залежно від способу будування підрозділяються на дискретні й інтервальні.

За дискретною ознакою, кількість значень якої є обмеженою, утворюється ***дискретний ряд*** розподілу***.***

Якщо ознака може набувати дискретних значеннь, кількість яких є великою, або дробових значень в області свого існування, будується ***інтервальний варіаційний ряд*** – таблиця з двох рядків або граф, що заповнюються інтервалами ознаки, варіація якої вивчається, і частотами.

Засобом наочного виразу результатів статистичного дослідження є статистичні таблиці, які призначено для подання результатів зведення й групування матеріалів спостереження.

1.4. Статистичні графіки

Статистичний графік – це спосіб наочного подання й викладення статистичних даних за допомогою геометричних знаків та інших графічних засобів з метою їх узагальнення й аналізування.

Основні елементи графіка:

– ***поле графіка*** – простір, на якому розміщують графічне зображення;

– ***графічний образ*** – сукупність геометричних або графічних знаків, за допомогою яких відображаються статистичні дані;

– ***масштабні орієнтири*** – масштаб, масштабні шкали, масштабні знаки для визначення розмірів;

– ***просторові орієнтири*** – елементи, які використовують для визначення порядку розміщення графічних знаків у полі графіка;

– ***експлікація графіка*** – пояснення, що розкривають його зміст і основні елементи (заголовок, одиниці вимірювання, умовні позначення).

Класифікація статистичних графіків

За *загальним призначенням* графіки поділяють на аналітичні, ілюстративні й інформаційні.

За *функціонально-цільовим* призначенням виділяють графіки групувань, рядів розподілу, порівняння, динаміки, взаємозв'язаних показників. Можливими є комбінації цих графіків, наприклад, графічне зображення варіації в динаміці або динаміки взаємозв'язаних показників і т.ін.

За *способом побудови (за* виглядом поля) графіки поділяють на діаграми й статистичні карти (картодіаграми і картограми).

За формою графічного образу розрізняють графіки точкові, лінійні, площинні (стовпчасті, почасові, квадратні, кругові, секторні, фігурні), просторові (об'ємні й фігурні).

За *типом системи координат* виділяють графіки в прямокутній і полярній системах.

За *типом масштабних шкал* – графіки з рівномірними, нерівномірними (функціональними) і змішаними шкалами.

За функціонально-цільовим призначенням розрізняють графіки:

– порівняння статистичних величин (діаграми – стовпчасті, стрічкові, квадратні, кругові, прямокутні, фігурні);

– структури й структурних зміщень (структурні діаграми – стовпасті, стрічкові, секторні);

– зображення динаміки статистичних показників (динамічні графіки, стовпчасті, стрічкові, квадратні, кругові, фігурні, діаграми темпів, лінійні діаграми (їхній різновид – радіальні));

– контролю виконання плану (лінійні графіки виконання плану, обліково-планові графіки);

– розповсюдження в просторі (картограма, картодіаграма, центрограма);

– варіаційних рядів (полігон, гістограма, кумулята, огіва, крива концентрації (Лоренца) і т.ін.);

– взаємозв'язку й взаємозалежності (графіки кореляційної залежності).

Діаграма – креслення, на якому статистична інформація зображується з допомогою геометричних фігур, ліній або символічних знаків.

Стовпчаста діаграма – графік, на якому статистичні дані зображені у вигляді стовпчиків-прямокутників однакової ширини, розташованих вертикально на осі абсцис і будь-якої висоти. Кожний стовпець характеризує окремий об'єкт.

Різновидом стовпчастої діаграми є ***стрічкова діаграма***, для якої є характерними горизонтальна орієнтація стовпчиків (стрічок) і вертикальне розташування базової лінії. Стрічкова діаграма є особливо зручною в тих випадках, коли окремі об'єкти порівняння характеризуються протилежними за знаком показниками***.***

Квадратні й кругові діаграми використовують для порівняння декількох абсолютних значень, при цьому сторона квадрата (радіус круга) – це корінь квадратний з абсолютного значення, що характеризує явище. У середині квадратів і кругів слід проставляти величини показників, що зображаються.

В ***об'ємних діаграмах*** (наприклад, у вигляді кубів) лімітні розміри графічного образу є пропорційними кореням кубічним з порівнюваних величин.

Прямокутні діаграми використовують у випадках, коли необхідно порівняти три взаємозв'язані показники, один з яких дорівнює додатку двох інших, і показати значення кожного з них у формуванні першої величини. У разі прямокутних діаграм установлюють два масштаби: один – для співмножника, який беруть за основу, другий – для того, який беруть за висоту.

У ***фігурних діаграмах*** геометричні фігури замінюють рисунками.

Структурні діаграми – діаграми співвідношення питомої ваги, які характеризують співвідношення окремих частин сукупності в їхньому загальному обсязі (стовпчасті, стрічкові, секторні). Основна форма структурних діаграм – ***секторні діаграми*** – графічне зображення на площі круга, працюючим геометричним параметром в якому є величина кута між радіусами: 1 % на діаграмі дорівнює 3,6°, а 100 % – сума всіх кутів, яка становить 360°.

Динамічні графіки призначені для зображення економічних явищ, що відбуваються в часі. У динамічних діаграмах об'єктом відображення є процеси.

Лінійні графіки характеризують змінення явищ у часі, залежність між двома показниками. У статистиці комерційної діяльності на ринку товарів і послуг вони мають найбільше поширення.

Різновидом лінійних діаграм є ***радіальні діаграми***, які відображають процеси і явища, що періодично повторюються в часі.

Для зображення варіаційних рядів застосовують лінійні й площинні діаграми, побудовані в прямокутній системі координат.

При дискретній варіації ознаки графіком варіаційного ряду є ***полігон розподілу*** – графічне зображення варіаційного ряду в прямокутній системі координат, де значення, які варіюються, відкладаються на осі абсцис, а відповідні їм частоти – на осі ординат.

Гістограма – графічне зображення інтервального варіаційного ряду, в якому на осі абсцис відкладаються варіанти, а прямокутники є пропорційними за висотою частотам значень ознаки для кожного інтервалу.

Кумуляти (кумулятивні діаграми) використовують для графічного порівняння двох або більше варіаційних розподілів з рівними або нерівними інтервалами. На осі абсцис відкладають відрізки інтервалів групувань, на осі ординат – накопичені частоти або частості.

1.5. Середні величини

Середня величина – це узагальнююча кількісна характеристика варіювальної ознаки в статистичній сукупності за конкретних умов місця й часу, яка характеризує її рівень з розрахунку на одиницю сукупності.

Середня величина відображає те загальне, що є властивим усім одиницям досліджуваної сукупності. Однак вона ігнорує індивідуальні відмінності окремих одиниць сукупності, які обумовлені випадковими обставинами через дію закону великих чисел.

Принципи розрахунку середніх величин:

1. Середню величину визначають для сукупностей, що складаються з якісно однорідних одиниць.

2. Середню величину визначають на основі масових даних.

3. Середню величину слід обчислювати з урахуванням економічного змісту досліджуваного показника.

4. Отриману середню величину слід обчислювати так, щоб під час замінення кожного індивідуального значення осереднюваного показника його середньою величиною залишався без змінень деякий підсумковий зведений показник, що має назву визначального, який зв'язаний тим чи іншим способом з осереднюваним показником.

Види середніх величин:

1) ***степеневі середні***, до яких належать середня геометрична, середня арифметична, гармонійна й середня квадратична. Залежно

від форми подання початкових даних середні величини можуть бути простими або зваженими;

2) ***структурні середні***, до яких належатья мода, медіана, квартилі, децилі.

Степеневі середні

Проста середня обчислюється за первинними, незгрупованими даними і має загальний вигляд

$$\overline{x}_k = \sqrt[k]{\frac{\sum_{i=1}^{n} x_i^k}{n}} = \left(\frac{\sum_{i=1}^{n} x_i^k}{n} \right)^{\frac{1}{k}},$$

де x_i – рівень (варіанта) усереднюваної ознаки;

$к$ – показник степеня середньої;

n – кількість варіант.

Зважена середня обчислюється за згрупованими даними і має такий загальний вигляд:

$$\overline{x}_k = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k f_i}{\sum_{i=1}^{m} f_i}} = \left(\frac{\sum_{i=1}^{m} x_i^k f_i}{\sum_{i=1}^{m} f_i} \right)^{\frac{1}{k}},$$

де x_i – рівень (варіанта) усереднюваної ознаки або серединне значення інтервалу, в якому вимірюється варіанта;

k – показник степеня середньої;

n – кількість варіант;

f_i – відповідні частоти (ваги) – кількість одиниць сукупності в різних групах, інтервалах, що показують, скільки разів трапляється i-те значення осереднюваної ознаки, причому $\sum_{i=1}^{m} f_i = n$.

Структурні середні

Структурні середні застосовують для характеристики структури сукупності. До них відносять показники моди, медіани, квартилі й децилі.

Мода (Мо) – величина, яка найчастіше трапляється в даній сукупності, або варіанта, що найчастіше повторюється в ряду.

Трапляються ряди, які мають дві моди (бімодальний ряд) або декілька мод (полімодальний ряд).

Медіана (Ме) – варіанта, що поділяє ранжований ряд на дві рівні за чисельністю частини, внаслідок чого у однієї половини одиниць сукупності значення ознаки не перевищує медіанного рівня, а у іншої – не менша за нього.

Особливості обчислень моди й медіани

1. Для *дискретного* варіаційного ряду:

– мода – це варіанта з максимальною частотою;

– медіана при парній кількості одиниць сукупності – арифметична середня величина з двох центральних варіант, при непарному числі – це центральна варіанта, розташована в центрі ряду.

2. Для *інтервального ряду розподілу* мода розраховується тільки для рівних інтервалів, оскільки від цього залежить показник повторюваності значень ознаки X.

Для обчислення моди необхідно визначити модальний інтервал, тобто інтервал зі значенням ознаки, що найчастіше повторюється:

$$Mo = x_{mo} + h\frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})},$$

де X_{mo} і h – відповідно нижня межа й ширина модального інтервалу;

f_{mo} – частоти модального інтервалу;

f_{mo-1}, f_{mo+1} – частоти попереднього й наступного інтервалів відносно модального.

Для обчислення медіани визначають медіанний інтервал – це інтервал, кумулятивна частота (сума накопичених частот, що передують медіанному інтервалу) якого дорівнює половині суми частот $0{,}5\sum_{1}^{m} f_j$ або перевищує її. З допомогою інтерполяції в цьому медіанному інтервалі знаходять значення медіани:

$$Me = x_{Me} + h\frac{0{,}5\sum_{1}^{m} f_j \quad S_{Me\ 1}}{f_{Me}},$$

де X_{Me} і h – відповідно нижня межа й ширина медіанного інтервалу;

f_{Me} – частота медіанного інтервалу;

S_{Me-1} – кумулятивна частота передмедіанного інтервалу.

Квартилі – це варіанти, які поділяють обсяг сукупності на чотири рівні частини, ***децилі*** – на десять рівних частин, **процентилі** – на 100. Ці характеристики визначаються на основі кумулятивних частот аналогічно медіані, яка є другим квартилем або п'ятим децилем.

1.6. Показники варіації

Середня величина – це узагальнювальна характеристика ознаки статистичної сукупності. Проте вона не пояснює, як групуються навколо неї

окремі значення, лежать вони поблизу або значно відхиляються від середньої. Коливання окремих значень ознаки характеризують показники варіації.

Варіація – це кількісні зміни ознаки в межах однорідної сукупності, обумовлені впливом різних факторів.

Ступінь близькості даних окремих одиниць до середньої вимірюється системою показників варіації, до якої належать абсолютні, середні й відносні показники. До абсолютних належать: варіаційний розмах, середнє лінійне й середнє квадратичне відхилення, дисперсії. Відносні характеристики подані рядом коефіцієнтів варіації, нерівномірності, локалізації, концентрації.

Абсолютні та середні показники варіації

Розмах варіації (варіаційний розмах) R – різниця між максимальним (X_{max}) і мінімальним (X_{min}) значеннями варіант:

$$R = X_{max} - X_{min}.$$

Розмах варіації фіксує лише крайні відхилення ознаки. Повторюваність проміжних значень тут не враховується.

Для узагальнюючої характеристики розподілу відхилень обчислюють ***середнє лінійне відхилення*** варіаційного ряду – середню арифметичну з абсолютних величин відхилень варіант від їхньої середньої:

– коли дані незгруповані,

$$\overline{d} = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n};$$

– коли дані згруповані,

$$\overline{d} = \frac{\sum_{i=1}^{m} \left| x_i - \overline{x} \right| f_i}{\sum_{i=1}^{m} f_i}.$$

З допомогою середнього лінійного відхилення аналізують, наприклад, склад робітників, ритмічність виробництва, рівномірність постачання матеріалів; розробляють системи матеріального стимулювання.

На практиці ступінь варіації об'єктивніше відображає ***дисперсія*** σ^2 – середня арифметична з суми квадратів відхилень окремих варіант від їхньої середньої:

– коли дані незгруповані,

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n};$$

– коли дані згруповані,

$$\sigma^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}.$$

Обчислити дисперсію можна за формулою

$$\sigma^2 = \overline{x^2} - \bar{x}^2,$$

де $\overline{x^2} = \frac{\sum_{i=1}^{m} x^2{}_i f_i}{\sum_{i=1}^{m} f_i}$.

Середнє квадратичне відхилення σ – корінь квадратний з дисперсії:

$$\sigma = \sqrt{\sigma^2}.$$

Середнє квадратичне відхилення є мірилом надійності середньої: чим меншим є значення σ, тим краще середня арифметична відображає собою всю сукупність.

Для забезпечення порівняння абсолютних показників варіації у варіаційних рядах різних явищ обчислюють відносні показники – ***коефіцієнти варіації***. Вони дають можливість порівнювати характер розсіювання в різних розподілах (різні одиниці спостереження однієї й тієї самої ознаки в двох сукупностях, при різних значеннях середніх, при порівнянні різнойменних сукупностей).

Квадратичний коефіцієнт варіації є найпоширенішим показником, що використовується для оцінювання однорідності сукупності, тобто надійності й типовості середньої величини:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\ \%\ \ (\bar{x} \neq 0).$$

Сукупності, які мають коефіцієнт варіації понад 30 %, прийнято вважати неоднорідними.

Коефіцієнт осциляції відображає відносну коливальність крайніх значень ознаки навколо середньої:

$$K_o = \frac{R}{\bar{x}} 100\ \%\ \ (\bar{x} \neq 0).$$

Відносне лінійне відхилення (лінійний коефіцієнт варіації) характеризує частку усередненого значення ознаки абсолютних відхилень від середньої величини:

$$K_d = \frac{\bar{d}}{\bar{x}} 100\ \%\ \ (\bar{x} \neq 0).$$

1.7. Вибіркове спостереження

Вибіркове спостереження – це вид несуцільного спостереження, за характеристикою відібраної частини одиниць якого оцінюють усю сукупність.

Не завжди можна використовувати суцільне спостереження, і тоді використовують вибіркове спостереження. Крім того, його використовують для уточнення результату суцільного спостереження (наприклад, під час перепису населення нарівні із суцільним спостереженням певну групу людей досліджували спеціально за більш розширеними анкетами). Крім того, вибіркове спостереження використовують при експериментах в природничих науках, а також і в таких економічних галузях дослідження, як митне обстеження якості продукції.

Основні завдання вибіркового спостереження такі:

1) вивчення середнього розміру досліджуваної ознаки;

2) вивчення питомої ваги (частки) досліджуваної ознаки в сукупності.

Основні поняття вибіркового методу

Сукупність, з якої відбираються елементи для обстеження, називають ***генеральною,*** а сукупність, яку безпосередньо обстежують, – ***вибірковою.*** Статистичні характеристики вибіркової сукупності розглядають як ***оцінки*** відповідних характеристик генеральної сукупності. Оскільки вибіркова сукупність неточно відтворює структуру генеральної, то вибіркові оцінки також не збігаються з характеристиками генеральної сукупності. Розбіжності між ними називають ***помилками репрезентативності (граничними похибками).*** Залежно від причин виникнення ці похибки поділять на систематичні й випадкові. ***Систематичні похибки*** виникають унаслідок порушення принципів випадковості відбору. ***Випадкові похибки*** – це наслідок випадковості відбору елементів сукупності для обстеження.

При організації вибіркового обстеження важливо запобігти виникненню систематичних похибок. Що стосується випадкових похибок, то уникнути їх неможливо, проте на основі теорії вибіркового методу можна визначити їхній розмір і наскільки можливо – регулювати.

Точність результатів вибіркового методу залежить від способу відбору одиниць, ступеня варіації ознаки у сукупності, кількості одиниць, що спостерігаються.

У генеральної сукупності частку одиниць, якій властива ознака, що вивчається, називають ***генеральною часткою*** p, а середня величина ознаки – ***генеральною середньою***. У вибіркової сукупності частку одиниць, якій властива ознака, що вивчається, називають ***вибірковою часткою*** (частістю) w, а середню величину ознаки у виборці – ***вибірковою середньою*** $\bar{x}$.

Розрізняють такі види вибіркового спостереження:

1) власне ***випадкова вибірка*** передбачає випадковий відбір одиниць з генеральної сукупності; це класичний спосіб формування вибіркової сукупності, і саме на ньому ґрунтується теорія вибіркового методу;

2) ***механічна вибірка*** – це послідовний відбір одиниць через рівні проміжки за їх розташуванням у генеральній сукупності або в будь-якій іншій послідовності. Відбір елементів здійснюється через однакові інтервали, крок інтервалу залежить від частості вибірки. Так, при $n / N = 0{,}05$ (де n, N – обсяг вибіркової й генеральної сукупностей) крок інтервалу дорівнює $1 / 0{,}05 = 20$;

3) ***типова (районована) вибірка*** передбачає попередню структуризацію генеральної сукупності на однорідні групи за певною ознакою і незалежний відбір елементів у кожній складовій частині випадковим або механічним способом. Обсяг розшарованої вибірки n – це сума часткових вибірок n_i, тобто $n = \sum_{1}^{m} n_i$, де m – кількість складових частин (груп, типів, районів тощо);

4) ***серійна (гніздова) вибірка*** складається з серій елементів сукупності, зв'язаних територіально (райони, селища), організаційно (фірми, акціонерні суспільства) тощо. Серії відбираються за схемою механічної або простої випадкової вибірки, обстеженню підлягають усі елементи серії. Перевагою серійної вибірки є те, що інколи відібрати окремі одиниці складніше, ніж серії. Прикладом є 10 %-ві відбори певної серії випуску продукції;

5) в статистичній практиці часто застосовується не один, а декілька видів вибірки, таке спостереження називають ***комбінаційним***.

Можливі розходження між середніми величинами або частками ознаки вибіркової й генеральної сукупності вимірюються ***середніми похибками репрезентативності (стандартом) μ*** (табл. 1.4). У цій таблиці наведено такі значення:

σ^2 – вибіркова дисперсія;

n – чисельність вибірки;

N – чисельність генеральної сукупності;

w – частка одиниць, які мають певну ознаку.

Таблиця 1.4

Спосіб відбору	Повторна вибірка	Безповторна вибірка
Визначення середньої	$\mu = \sqrt{\frac{\sigma^2}{n}}$	$\mu = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$
Визначення частки	$\mu = \sqrt{\frac{w(1-w)}{n}}$	$\mu = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$

Фактори, що впливають на помилки репрезентативності:

– показники варіації певної ознаки (наприклад, дисперсії): чим більший показник варіації, тим більший розмір помилки;

– від чисельності вибірки: чим більша вибірка, тим менша ймовірність помилки (розмір помилки);

– від способу відбору (повторний або безповторний).

Гранична похибка вибірки Δ – це максимально можлива похибка для прийнятої ймовірності *p,* якій відповідає коефіцієнт довіри – *t*-разове значення μ. Значення коефіцієнта довіри відображає, скільки середніх похибок міститься в граничній похибці: *t* = 1 – для ймовірності 0,683; *t* = 2 – для ймовірності 0,954; *t* = 3 – для ймовірності 0,997; *t* = 4 – для ймовірності 0,999.

Формула граничної похибки має вигляд

$$\Delta = t\mu .$$

Для розрахунку граничної похибки вибірки застосовують формули, наведені в табл. 1.5.

Таблиця 1.5

Спосіб відбору	Повторна вибірка	Безповторна вибірка
Визначення середньої	$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}}$	$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$
Визначення частки	$\Delta_w = t\sqrt{\frac{w(1-w)}{n}}$	$\Delta_w = t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$

При обчисленні похибки районованої вибірки застосовують середню із групових дисперсій:

$$\overline{\sigma}^2 = \frac{\sum_1^m \sigma_i^2 n_i}{\sum_1^m n_i},$$

тоді

$$\Delta = t\sqrt{\frac{\overline{\sigma}^2}{n}(1-\frac{n}{N})}.$$

Як правило, похибка розшарованої вибірки менша, ніж похибка механічної або випадкової вибірки.

При обчисленні похибки серійної вибірки враховують міжсерійну варіацію:

$$\Delta = t\sqrt{\frac{\delta^2}{n}(1-\frac{s}{S})},$$

де δ^2 – міжсерійна дисперсія;

$$\delta^2 = \frac{\sum_1^m (\bar{x}_k - \bar{x})^2 n_k}{\sum_1^m n_k}.$$

Тут n_k і $\bar{x}_k$ – відповідно обсяг і середня *к*-ї серії s_k.

З допомогою формул граничної похибки вибірки визначають:

1) довірчі межі генеральної середньої і частки з певною ймовірністю:

– для середньої генеральної сукупності

$$\bar{x} - \Delta \le \bar{x} \le \bar{x} + \Delta;$$

– для частки одиниць w, яким притаманна ця ознака в генеральній сукупності,

$$w - \Delta \le w \le w + \Delta;$$

2) імовірність того, що відхилення між вибірковими й генеральними характеристиками не перевищує визначену величину;

3) необхідну чисельність вибірки *n*, яка із заданою ймовірністю забезпечує очікувану точність вибіркових показників і при якій вибіркові оцінки мали б основні властивості генеральної сукупності (табл. 1.6).

Таблиця 1.6

Спосіб відбору	Повторна вибірка	Безповторна вибірка
Визначення середньої	$n = \frac{t^2\sigma^2}{\Delta_{\bar{x}}^2}$	$n = \frac{t^2\sigma^2 N}{\Delta_{\bar{x}}^2 N + t^2\sigma^2}$
Визначення частки	$n = \frac{t^2 w(1-w)}{\Delta_w^2}$	$n = \frac{t^2 w(1-w)N}{\Delta_w^2 N + t^2 w(1-w)}$

Для визначення обсягу вибірки *n* використовують оцінки дисперсій σ^2 аналогічних пробних обстежень. Якщо таких обстеженнь немає, можна скористатися співвідношенням $\sigma = \frac{1}{6}(x_{\max} - x_{\min})$, а для частості взяти найбільше значення дисперсії $\sigma^2 = 0{,}25$.

Якщо в основу розрахунку *n* покласти відносну похибку вибірки, формули відповідно модифікуються:

– для середньої

$$n = \frac{t^2 V_x^2}{V_\Delta^2};$$

– для частки

$$n = \frac{t^2 q}{V_\Delta^2 p}.$$

При порівнянні точності вибіркових оцінок використовують ***відносну похибку вибірки*** V_μ, яка показує, на скільки відсотків вибіркова оцінка відхиляється від параметра генеральної сукупності:

$$V_\mu = \frac{\mu}{\bar{x}} \cdot 100.$$

Відносну похибку вибірки можна розрахувати на основі коефіцієнта варіації ознаки V_x :

– для повторної вибірки

$$V_\mu = 100 \cdot \frac{V_x}{\sqrt{n-1}};$$

– для безповторної вибірки

$$V_\mu = 100 \cdot \frac{V_x}{\sqrt{n-1}} \sqrt{1 - \frac{n}{N}}.$$

Аналогічно розраховують відносну похибку вибірки для частості:

$$V_\mu = \frac{\mu_p}{p} = \frac{\sqrt{pq/n}}{p} = \sqrt{\frac{q}{np}}.$$

Залежність і незалежність вибірок

Дві вибірки залежать одна від одної, якщо кожному значенню однієї вибірки можна закономірним і однозначним чином поставити у відповідність тільке одне значення іншої вибірки. Аналогічно визначається залежність декількох вибірок.

Найчастіше залежні вибірки виникають, коли вимір проводиться для декількох моментів часу. Залежні вибірки утворюють значення параметрів процесу, що вивчається, які відповідають різним моментам часу.

У SPSS залежні (а також зв'язані, спарені) вибірки будуть подаватися різними змінними, які зіставляються один з одним у відповідному тесті на одній і тій самій сукупності спостережень.

Якщо закономірна й однозначна відповідність між вибірками є неможливою, ці вибірки є незалежними. У SPSS незалежні вибірки містять різні спостереження (наприклад, такі, що належать до різних респондентів), які зазвичай розрізняються з допомогою групової змінної, що належить до номінальної шкали.

1.8. Статистичні методи вивчення взаємозв'язків. Кореляційний і регресійний методи аналізу зв'язку

Усі соціально-економічні явища є взаємозв'язаними. Зв'язок між ними має причиново-наслідковий характер. Ознаки, які характеризують причини й умови зв'язку, називають ***факторними*** ознаками ***x,*** а ті, що характеризують наслідки зв'язку, – ***результативними*** ознаками ***y.*** Між ознаками *x* і *y* виникають різні за природою й характером зв'язки, зокрема функціональні й стохастичні. При ***функціональному зв'язку*** певному значенню факторної ознаки *x* відповідає чітко визначене значення результативної ознаки *y*. Цей зв'язок виявляється однозначно у кожному конкретному випадку. При ***стохастичному зв'язку*** кожному значенню ознаки *x* відповідає певна множина значень *y*, які утворюють так званий умовний розподіл*.* Як закон цей зв'язок виявляється тільки у великій кількості випадків і характеризується зміненням умовних розподілів *y*. Якщо замінити умовний розподіл середньою величиною *y*, то утворюється різновид стохастичного зв'язку – ***кореляційний.*** У разі кореляційного зв'язку кожному значенню ознаки *x* відповідає середнє значення результативної ознаки *y*.

Умовні розподіли можна замінити середніми значеннями результативної ознаки, які обчислюються як середня арифметична зважена.

Поступове змінення середніх $\bar{y}_j$ від однієї групи до іншої свідчить про наявність кореляційного зв'язку між ознаками.

Характеристикою кореляційного зв'язку є ***лінія регресії*** *y* на *x* – це функція, яка зв'язує середні значення ознаки *y* зі значеннями ознаки x. Лінія регресії розглядається в двох моделях: аналітичного групування й регресійного аналізу. У моделі ***аналітичного групування*** – це емпірична лінія регресії, яка складається з групових середніх значень результативної ознаки $\bar{y}_j$ для кожного значення (інтервалу) x_j.

Оцінка щільності зв'язку ґрунтується на правилі складання дисперсій. У моделі аналітичного групування мірою щільности зв'язку є відношення міжгрупової дисперсії δ^2 до загальної σ^2, яке називають ***кореляційним відношенням:***

$$\eta^2 = \frac{\delta^2}{\sigma^2},$$

де σ^2 – загальна дисперсія, яка вимірює варіацію результативної ознаки *y*, обумовлену дією всіх можливих чинників:

$$\sigma^2 = \frac{\sum (y_i - \bar{y})^2 f_i}{\sum f_i}$$

або

$$\sigma^2 = \overline{y^2} - \bar{y}^2 = \frac{\sum (y_i)^2 f_i}{\sum f_i} - \left(\frac{\sum y_i f_i}{\sum f_i} \right)^2,$$

де δ^2 – міжгрупова дисперсія, яка вимірює варіацію результативної ознаки *y* за рахунок дії тільки групувальної ознаки *x*:

$$\delta^2 = \frac{\sum (\bar{y}_j - \bar{y})^2 f_j}{\sum f_j}.$$

Кореляційне відношення коливається від 0 до 1 або від 0 до 100 %. За умови відсутності зв'язку маємо η^2 = 0, а за умови наявності функціонального зв'язку – η^2=1. Чим більше це відношення наближається до одиниці, тим більш щільним є зв'язок.

Щільний зв'язок може виникнути випадково, тому необхідно перевірити наявність його щільности, тобто довести невипадковість зв'язку. ***Перевірка щільности зв'язку*** – це порівняння фактичного значення η^2, з його критичним значенням $\eta^2_{1-\alpha}(k_1, k_2)$ для певного рівня щільности α і числа ступенів свободи $k_1 = m - 1$ і $k_2 = n - m$, де *m* – кількість груп; *n* – обсяг сукупності.

Критичне значення є тим максимально можливим значенням кореляційного відношення, яке може виникнути випадково за умови відсутності кореляційного зв'язку.

Якщо $\eta^2 > \eta^2_{1-\alpha}(k_1, k_2)$, то зв'язок між результативною й факторною ознаками вважається істотним. Якщо $\eta^2 < \eta^2_{1-\alpha}(k_1, k_2)$, то наявність кореляційного зв'язку між результативною і факторною ознаками не доказана й зв'язок уважається неістотним.

Критичне значення вибирають таким чином, щоб імовірність отримання значення η^2, більшого за критичне (за умови відсутності зв'язку між ознаками), була достатньо малою. Таку ймовірність називають рівнем істотності α. Найчастіше зостосовують такі рівні істотності, як α = 0,05 і α = 0,01.

Для перевірки істотності зв'язку використовують також функціонально зв’язану з η^2 характеристику F-критерію (критерію Фішера), який обчислюють за формулами

$$F = \frac{\eta^2}{1-\eta^2} \cdot \frac{k_2}{k_1} \quad \text{або} \quad F = \frac{\delta^2}{\bar{\sigma}^2} \cdot \frac{k_2}{k_1}.$$

Існують таблиці критичних значень F-критерію. Перевірку істотності зв’язку з його допомогою здіснюють аналогічно описаній для кореляційного відношення η^2.

У ***моделі регресійного аналізу*** характеристикою кореляційного зв'язку є теоретична лінія регресії, що описується функцією $Y = f(x)$, яка має назву ***рівняння регресії.***

На основі рівняння регресії визначають теоретичні значення Y, тобто значення результативної ознаки за умови дії тільки чинника x при незмінному рівні інших чинників.

Залежно від характеру зв'язку використовують:

– ***лінійні рівняння*** $Y = a + bx$, коли зі зміненням x ознака y змінюється більш-менш рівномірно;

– ***нелінійні рівняння***, коли змінення взаємозв'язаних ознак відбувається нерівномірно (з прискоренням, уповільненням або зі змінним напрямом зв'язку), зокрема: степеневе $Y = ax^b$, гіперболічне $Y = a + b/x$, параболічне $Y = a + bx + cx^2$ та ін.

Частіше застосовують лінійні рівняння або рівняння, зведені до лінійного вигляду. У лінійному рівнянні параметр b – ***коефіцієнт регресії*** – означає, на скільки одиниць у середньому зміниться y зі зміненням x на одиницю. Він має одиницю вимірювання результативної ознаки. У разі прямого зв'язку b – величина додатна, а при зворотному – від'ємна. Параметр a – вільний член рівняння регресії, тобто це є значенням y при $x = 0$. Якщо x не набуває нульового значення, то цей параметр буде тільки розрахунковим. Параметри визначаються ***методом найменших квадратів (МНК),*** згідно з яким сума квадратів відхилень емпіричних значень y від теоретичних Y є мінімальною: $\sum (y - Y)^2 \to \min$. Відповідно до умови мінімізації параметри лінійного рівняння регресії обчислюються на основі системи нормальних рівнянь:

$$\begin{cases} na_0 + a_1 \sum x = \sum y, \\ a_0 \sum x + a_1 \sum x^2 = \sum xy. \end{cases}$$

Звідси

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum y};$$

$$a_0 = \bar{y} - a_1 \bar{x}.$$

Значення коефіцієнта регресії в невеликих за обсягом сукупностях ($n < 30$) може випадково змінюватися, тому здійснюється перевірка його істотності за допомогою t-критерію Стьюдента. При цьому фактичні значення t-критерію розраховують за формулою

$$t_{a_1} = \frac{a_1}{\mu_{a_1}},$$

де a_1 – коефіцієнт регресії;

μ_{a_1} – власне стандартна погрішність, яка розраховується за формулою

$$\mu_{a_1} = \frac{\sigma_e}{\sigma_x}\sqrt{(n-2)},$$

де n – обсяг сукупності;

σ_x – середнє квадратичне відхилення факторної ознаки x_i від загальної середньої $\bar{x}$,

$$\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}};$$

σ_e – середнє квадратичне відхилення результативної ознаки y_i від теоретичних значень Y,

$$\sigma_e = \sqrt{\frac{\sum(y_i - Y)^2}{n}}.$$

Фактичне значення t-критерію t_{a_1} порівнюють з критичним значенням t_k, яке отримують за таблицею Стьюдента з урахуванням рівня істотності α і числа ступенів свободи $k = n - 2$. Якщо $t_k < t_{a_1}$, то коефіцієнт регресії вважається істотним.

Характеристикою відносного змінення y за рахунок x є ***коефіцієнт еластичності***

$$K_{ел} = a_1\frac{\bar{x}}{\bar{y}},$$

який показує, на скільки відсотків у середньому змінюється результативна ознака зі зміненням чинника на 1 %.

Для статистичного оцінювання щільності зв'язку використовують такі показники варіації:

1) загальна дисперсія результативної ознаки, яка відображає дію всіх факторів,

$$\sigma_Y^2 = \frac{1}{n}\sum_1^n\left(y - \bar{Y}\right)^2;$$

2) факторна дисперсія результативної ознаки, яка відображає варіацію y, обумовленую дією тільки чинника x,

$$\sigma_Y^2 = \frac{1}{n}\sum_1^n(Y - \bar{y})^2;$$

3) залишкова дисперсія, яка відображає дію на результативну ознаку всіх інших чинників, окрім x,

$$\sigma^2{}_e = \frac{\sum (y_i - Y)^2}{n}.$$

Частка факторної дисперсії в загальній характеризує щільність зв'язку і має назву ***коефіцієнта детермінації***:

$$R^2 = \frac{\sigma_Y^2}{\sigma_y^2}.$$

Він має такі самі зміст, інтерпретацію й цифрові межі, як η^2.

Для лінійного зв'язку використовують лінійний ***коефіцієнт кореляції*** (коефіцієнт Пірсона). Критерії кількісного оцінювання залежності між змінними називають коефіцієнтами кореляції, або заходами зв'язаності:

$$r = \frac{\sum_1^n xy - n\bar{x} \cdot \bar{y}}{n\sqrt{\sigma_x^2 \cdot \sigma_y^2}}, \text{ або } r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Дві змінні корелюють між собою додатно, якщо між ними існує пряме однонапрямлене співвідношення, при якому малі значення однієї змінної відповідають малим значенням іншої змінної, великі значення – великим. Дві змінні корелюють між собою від'ємно, якщо між ними існує обернене, різнонапрямлене співвідношення, при якому малі значення однієї змінної відповідають великим значенням іншої змінної і навпаки. Значення коефіцієнтів кореляції завжди знаходяться в діапазоні від –1 до +1.

Для опису величин коефіцієнта кореляції застосовують табл. 1.7.

Таблиця 1.7

Значення коефіцієнта кореляції *r*	Інтерпретація
$0 < r \leq 0{,}2$	Дуже слабка кореляція
$0{,}2 < r \leq 0{,}5$	Слабка кореляція
$0{,}5 < r \leq 0{,}7$	Середня кореляція
$0{,}7 < r \leq 0{,}9$	Сильна кореляція
$0{,}9 < r \leq 1$	Дуже сильна кореляція

Як коефіцієнт кореляції між змінними, що належать до порядкової шкали, застосовують коефіцієнт Спірмена, а для змінних, що належать до інтервальної шкали, – коефіцієнт кореляції Пірсона. При цьому слід врахувати, що кожну дихотомічну змінну, тобто змінну, що належить до номінальної шкали і має дві категорії, можна розглядати як порядкову.

Перевірка істотності зв'язку здійснюється так само, як і в моделі аналітичного угрупування, тобто шляхом порівняння фактичного й критичного значень. Відмінності стосуються тільки визначення $k_1 = m - 1$

і $k_2 = n - m$, де m – кількість параметрів рівняння регресії. Для лінійної моделі $k_2 = 2 - 1 = 1$.

Істотність зв'язку в обох моделях можна здійснювати також за критерієм Фішера, який функціонально зв'язаний з R^2 і η^2. Тоді фактичне значення визначають за формулою

$$F = \frac{R^2}{1-R^2} \cdot \frac{k_2}{k_1} \quad \text{або} \quad F = \frac{\eta^2}{1-\eta^2} \cdot \frac{k_2}{k_1},$$

тому процедури перевірки й висновки є ідентичними.

Огляд тестів для перевірки гіпотез про рівність середніх

У найбільш поширеній ситуації, коли необхідно порівняти різні вибірки за їхніми середніми значеннями або медіанами, зазвичай застосовується один із восьми тестів, два з яких наведено нижче (табл. 1.8, 1.9).

Таблиця 1.8

Змінні, що належать до інтервальної шкали і підпорядковуються нормальному розподілу

Кількість порівнюваних вибірок	Залежність	Тест
1	Незалежні	t -тест Стьюдента
1	Залежні	t -тест для залежних вибірок
більше 2	Незалежні	Простий дисперсійний аналіз
більше 2	Залежні	Простий дисперсійний аналіз з повторними вимірами

Таблиця 1.9

Змінні, що належать до порядкової або інтервальної шкали, але не підпорядковуються нормальному розподілу

Кількість порівнюваних вибірок	Залежність	Тест
1	Незалежні	U -тест Манна й Уітни
2	Залежні	Тест Уілкоксона
більше 2	Незалежні	H-тест Крускала й Уолліса
більше 2	Залежні	Тест Фрідмана

Для кожної з цих двох груп тестів у SPSS є окремі пункти меню, а саме Analyze (Аналіз) Compare Means (Порівняння середніх) або Analyze (Аналіз) Nonparametric Tests (Непараметричні тести)

1.9. Імовірність помилки *p*

Якщо поділити статистику на описову й аналітичну, то завдання аналітичної статистики – надати методи, з допомогою яких можна було б об'єктивно з'ясувати, наприклад, чи є випадковою спостережувана різниця в середніх значеннях (або взаємозв'язок (кореляція) вибірок) чи ні.

Наприклад, якщо порівнюються два середні значення вибірок, то можна сформулювати дві попередні гіпотези:

– гіпотеза 0 (нульова) – спостережувані відмінності між середніми значеннями вибірок знаходяться в межах випадкових відхилень;

– гіпотеза 1 (альтернативна) – спостережувані відмінності між середніми значеннями не можна пояснити випадковими відхиленнями.

В аналітичній статистиці розроблено методи обчислення так званих тестових (контрольних) величин, які розраховуються за певними формулами на основі даних, що містяться у вибірках або в отриманих з них характеристиках. Ці тестові величини відповідають певним теоретичним розподілам (*t*-розподілу, F-розподілу, розподілу X^2 і т. д.), які дають можливість обчислити так звану вірогідність помилки, що дорівнює відсотку помилки, яку можна допустити, відкинувши нульову гіпотезу й прийнявши альтернативну.

Імовірність визначається в математиці як величина, що знаходиться в діапазоні від 0 до 1 і позначається буквою p: $0 \le p \le 1$. На практиці її часто виражають у відсотках.

Імовірність помилки, при якій допустимо відкинути нульову гіпотезу й прийняти альтернативну гіпотезу, залежить від кожного конкретного випадку. Значною мірою ця ймовірність визначається характером досліджуваної ситуації. Чим більш необхідною є ймовірність, з якою потрібно уникнути помилкового рішення, тим більш вузькими вибираються межі ймовірності помилки, при якій відкидається нульова гіпотеза, так званий довірчий інтервал імовірності.

Існує загальноприйнята термінологія, яка належить до довірчих інтервалів ймовірності.

Вирази, що мають імовірність помилки $p > 0{,}05$, називають незначущими; вирази, що мають імовірність помилки $p \le 0{,}05$, називають значущими; вирази з імовірністю помилки $p \le 0{,}01$ – дуже значущими, а вирази з імовірністю помилки $p \le 0{,}001$ – максимально значущими. У літературі такі ситуації позначають однією, двома або трьома зірочками.

Часи, коли не було комп'ютерів, придатних для статистичного аналізу, давали практикам принаймні одну перевагу: оскільки всі обчислення потрібно було виконувати вручну, статистик мав спочатку ретельно обдумати, які питання можна вирішити з допомогою того або іншого тесту. Крім того, особливе значення надавалося точному формулюванню нульової гіпотези.

З допомогою комп'ютера й такої потужної програми, як SPSS, дуже легко можна провести безліч тестів за дуже короткий час. Наприклад, якщо в таблицю спряженості звести 50 змінних з іншими 20 змінними і виконати тестування, то вийде 1000 результатів перевірки значущості для яких буде отримано 1000 значень *p*. Некритичний підбір значущих

величин може дати беззмістовий результат, оскільки вже при граничному рівні значущості $p = 0,05$ в п'яти відсотках спостережень, тобто в 50 можливих спостереженнях, можна чекати значущі результати.

Цим помилкам першого роду (коли нульова гіпотеза відкидається, хоча вона є правильною) слід приділяти достатньо уваги. Помилкою другого роду називають ситуацію, коли нульова гіпотеза приймається, хоча вона є помилковою. Імовірність допустити помилку першого роду дорівнює ймовірності помилки p. Імовірність помилки другого роду тим менша, чим більше ймовірність помилки p.

2. СТАТИСТИКА В SPSS

2.1. Введення в SPSS

Два студенти Норман Най (Norman Nie) і Дейл Вент (Dale Bent), що спеціалізувалися в області політології, 1965 року, намагаючись відшукати в Стенфордському університеті Сан-Франциско комп'ютерну програму для аналізу статистичної інформації і зневірившись у своїх спробах (тому що наявні програми виявлялися непридатними, невдало побудованими або не забезпечували наочність подання обробленої інформації), вирішили розробити власну програму зі своєю концепцією і єдиним синтаксисом. У їх розпорядженні тоді була мова програмування FORTRAN і обчислювальна машина типу IBM 7090. Через рік було розроблено першу версію програми, що з 1967 року могла працювати на IBM 360.

Як відомо з історії розвитку інформатики, програми тоді являли собою пакети перфокарт. Саме про це свідчить і вихідна назва програми, що автори дали своєму продукту: SPSS – це абревіатура від Statistical Package for the Social Science.

Командна мова (синтаксис) SPSS у той час була ще не так добре розвинена, як зараз, і орієнтована на перфокарти. 1983 року командну мову SPSS було повністю перероблено, синтаксис став значно зручнішим. Щоб відзначити цей факт, програму було перейменовано в SPSSX, де буква X мала служити як номером версії в римських числах, так і скороченням для extended (для розширення).

З появою персональних комп'ютерів було розроблено також і PC-версію SPSS. З 1983 року виникла PC-версія SPSS\PC+. Пізніше (з 1984 року) програма SPSS стала широко застосовуватися й у Європі.

Зараз це є найпоширенішим програмним забезпеченням для статистичного аналізу в усьому світі.

Для того щоб мати можливість використовувати програми в усіх областях, що належать до статистичного аналізу, вихідній абревіатурі присвоєно нове значення: Superior Performance Software System (система програмного забезпечення вищої продуктивності).

Версія SPSS для операційної системи Windows (SPSS for Windows) стала великим кроком уперед. Програмою можна користуватися без особливих знань в області прикладного програмування. Виклик необхідних процедур статистичного аналізу відбувається з допомогою стандартної техніки, що застосовується в Windows, тобто з допомогою миші й відповідних діалогових вікон.

Модулі SPSS

Основу програми SPSS становить базовий модуль SPSS Base, що надає різноманітні можливості доступу до даних і керування ними. Він містить методи аналізу, які застосовуються найчастіше.

Традиційно разом з SPSS Base (базовим модулем) поставляються ще два модулі: Advanced Models (просунуті моделі) і Regression Models (регресійні моделі). Нарівні із трьома згаданими існує ще декілька спеціальних додаткових модулів і самостійних програм, кількість яких постійно зростає, так що користувачам варто постійно ознайомлюватися з інформацією про нововведення в SPSS.

Базовий модуль ***SPSS Base*** входить у базову поставку. Він містить усі процедури введення, відбору й коригування даних, а також більшість пропонованих в SPSS статистичних методів. Нарівні із простими методиками статистичного аналізу, такими, як частотний аналіз, розрахунок статистичних характеристик, таблиці спряженості, кореляцій, будування графіків, цей модуль містить t-тести й велику кількість інших непараметричних тестів, а також ускладнені методи, такі, як багатовимірний лінійний регресійний аналіз, дискримінантний аналіз, факторний аналіз, кластерний аналіз, дисперсійний аналіз, аналіз придатності (аналіз надійності) і багатовимірне шкалування.

Модуль ***Regression Models*** містить різні методи регресійного аналізу, такі, як бінарна й мультіноміальна логістична регресія й нелінійна регресія.

До модуля ***Advanced Models*** належать різні методи дисперсійного аналізу (багатовимірний, з урахуванням повторних вимірів), загальна лінійна модель, аналіз виживання, включаючи метод Каплана–Майера й регресію Кокса, а також логлінійні моделі.

Модуль ***Tables*** призначено для створення презентаційних таблиць. Тут надаються більш широкі можливості порівняно зі спроще-

ними частотними таблицями й таблицями спряженості, які будуються в SPSS Base (базовому модулі).

Нижче за абеткою наведено список інших модулів і програм, пропонованих для розширення SPSS.

Amos (Analysis of moment structures – аналіз моментних структур) містить методи аналізу з допомогою лінійних структурних рівнянь. Метою програми є перевірка складних теоретичних зв'язків між різними ознаками випадкового процесу і їх опис з допомогою відповідних коефіцієнтів. Перевірка проводиться у формі причинного аналізу й аналізу траєкторії. При цьому користувач у графічному вигляді повинен задати теоретичну модель, у яку разом з даними безпосередніх спостережень можуть бути включені й так звані приховані елементи. Програму Amos включено до складу модулів розширення SPSS.

AnswerTree (дерево рішень) містить чотири різні методи автоматизованого поділу даних на окремі групи (сегменти). Поділ проводиться таким чином, що частотні розподіли цільової (залежної) змінної в різних сегментах значно розрізняються. Типовим прикладом застосування цього методу є створення характерних профілів покупців при дослідженні споживчого ринку.

Модуль ***Categories*** містить різні методи для аналізу категоріальних даних, а саме: аналіз відповідностей і три різні методи оптимального шкалування (аналіз однорідності, нелінійний аналіз головних компонентів, нелінійний канонічний кореляційний аналіз).

Clementine – це програма для data mining (одержання знань), у якій користувачеві пропонуються численні підходи до будування моделей, наприклад, нейронні мережі, дерева рішень, різні види регресійного аналізу. Clementine являє собою "верстат" аналітика, з допомогою якого можна візуалізувати процес моделювання, перевіряти ще раз моделі, порівнювати їх між собою. Для зручності користування програмою існує допоміжне середовище впровадження результатів.

Спільний аналіз ***Conjoint*** застосовується під час дослідження ринку для вивчення споживчих властивостей продуктів стосовно їхньої привабливості. При цьому опитувані респонденти на свій розсуд мають розташувати пропоновані набори споживчих властивостей продуктів за перевагами, на основі якого можна потім вивести так звані деталізовані показники корисності окремих категорій кожної споживчої властивості.

Програму ***Data Entry*** (уведення даних) призначено для швидкого складання запитальників, а також уведення й чищення даних. Задані на етапі створення запитальника запитання й категорії відповідей потім використовуються як мітки змінних і значень.

Модуль ***Exact Tests*** (точні тести) призначено для обчислення точного значення ймовірності помилки (величини *p*) в умовах обмежено-

сті даних під час перевірки за критерієм X^2 (Chi-Quadrat-Test) і при непараметричних тестах. Якщо буде необхідно, для цього також можна застосувати метод Монте-Карло (Monte-Carlo).

Програма ***GOLDMine*** містить спеціальну регресійну модель для регресійного аналізу впорядкованих залежних і незалежних змінних.

З допомогою ***SamplePower*** можна визначити оптимальний розмір вибірки для більшості методів статистичного аналізу, реалізованих у SPSS.

Модуль ***SPSS Missing Value Analysis*** призначено для аналізу й відновлення закономірностей, яким підпорядковуються пропущені значення. Він надає різні варіанти замінення пропущених значень.

Модуль ***Trends*** містить різні методи для аналізу тимчасових рядів, таких, як моделі ARIMA, експонентне згладжування, сезонна декомпозиція й спектральний аналіз.

Вікна програми SPSS

SPSS містить такі вікна:
– редактор даних (Data Editor),
– вікно перегляду (Viewer),
– вікно перегляду тексту (Text Viewer),
– редактор мобільних таблиць (Pivot Table Editor),
– редактор діаграм (Diagram Editor),
– редактор текстового висновку (Text Output Editor),
– редактор синтаксису (Syntax Editor),
– редактор скриптів (Script Editor).

Кожне вікно, крім редактора мобільних таблиць, має одну або дві панелі символів для виклику часто використовуваних команд. Короткі відомості про кожний символ можна одержати, якщо помістити на нього покажчик миші. Нижче наведено ті символи, які трапляються майже в усіх вікнах.

Символи *відкриття*, *збереження файла й друку* мають стандартний вигляд і значення, як при роботі з Windows.

Історія виклику діалогових вікон  . Цей символ виводить список 12 останніх викликаних діалогових вікон, що дає можливість швидко перейти до одного з нещодавно викликаних діалогових вікон. Вікно, викликане в останню чергу, завжди знаходиться на початку списку (рис. 2.1).

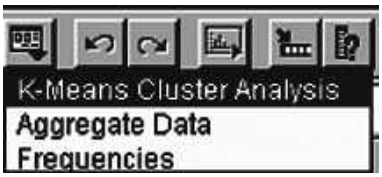


Рис. 2.1. Історія виклику діалогових вікон

Щоб наново викликати діалогове вікно, необхідно просто клацнути на відповідному пункті списку.

Перейти в редактор даних  . Цей символ забезпечує перехід в редактор даних з будь-якого вікна.

Перейти до спостереження  . Цей символ відкриває діалогове вікно Go to case (Перейти до спостереження). Його можна використовувати для переходу до певного спостереження (так в SPSS називають набір значень змінних, набраних в рядку редактора даних).

Вибрати спостереження  . Цей символ відкриває діалогове вікно Select cases (Вибрати спостереження). Його можна використовувати для відбору спостережень, для яких виконується певна умова (що фактично означає будування фільтра).

Інформація про змінні  . Цей символ відкриває діалогове вікно Variables, в якому відображаються описи виділених змінних. Про призначення символів, які виникають тільки в одному певному вікні, легко можна дізнатися з коротких відомостей (Quick Info) за цим символом.

Три наступні символи можна задіяти в редакторі даних.

Вставити спостереження  . В редакторі даних клацання на цьому символі викликає вставку спостереження над активною коміркою.

Вставити змінну  . В редакторі даних клацання на цьому символі викликає вставку нової змінної ліворуч від активної змінної.

Мітки значень  . Цей символ дає можливість перемикатися між відображеннями значень їх мітками.

2.2. Огляд статистичних методів, які застосовуються при статистичному аналізі за допомогою SPSS

Розглянемо послідовність дій, які виконуються при статистичному аналізі.

1. *Структуризація, введення і перевірка даних.* Перш ніж зможемо застосовувати статистичні методи або будувати графіки, слід подати зібрані дані у формі, придатній для оброблення. При цьому рекомендується дотримуватися такого плану дій:

– проведення структуризації набору даних; передусім з'ясування, до яких категорій належать спостереження і до яких – змінні; в більшості випадків це ясно відразу; якщо структуризацію провести не вдається, SPSS застосовувати не можна, та й усі інші статистичні програми також потребують, щоб дані були структурованими;

– визначити шкалу, до якої належать змінні;

– скласти кодувальну таблицю;

– увести дані в редакторі даних, враховуючи кодувальну таблицю; якщо для введення даних використовуються інші програми (наприклад, Excel, dBase), це є цілком допустимим, SPSS може працювати з файлами даних цих програм; не треба вводити дані, які можна обчислити на основі інших даних, ці обчислення слід дати комп'ютеру; якщо дані вже було введено в інших програмах статистики (наприклад, Statistica), їх можна перетворити у файли SPSS з допомогою таких утиліт, як, наприклад, DBMS/COPY;

– перевірити введені дані на відсутність помилок;

– установити, чи підпорядковуються нормальному розподілу змінні, що належать до інтервальної шкали.

Тепер можна починати статистичне оброблення введених даних. Слід ураховувати, що аналіз можна виконати тільки для спостережень, згрупованих певним чином.

2. *Описовий (дескриптивний) аналіз*. Цей вид аналізу містить описове подання окремих змінних. До нього належать створення частотної таблиці, обчислення статистичних характеристик або графічне подання. Частотні таблиці будуються для змінних, що належать до номінальної шкали, і для порядкових змінних, що мають не дуже багато категорій.

Для змінних, що належать до номінальної шкали, не можна обчислити ніяких значущих статистичних характеристик. Найчастіше для порядкових змінних і змінних, що належать до інтервальної шкали, але не підпорядковуються нормальному розподілу, обчислюються медіани й обидва квартилі; при невеликій кількості категорій можна використовувати варіант для концентрованих даних.

Для змінних, що належать до інтервальної шкали і підпорядковуються нормальному розподілу, найчастіше обчислюється середнє значення й стандартне відхилення або стандартна помилка. Проте слід вибрати тільки одну з цих двох характеристик розкиду. Для змінних, що належать до усіх статистичних шкал, можна побудувати велику кількість різноманітних графіків, на яких показано частоти, середні значення або інші характеристики.

3. *Аналітична статистика*. Майже будь-який статистичний аналіз разом з чисто описовими операціями містіть ті або інші аналітичні методи (тести значущості), при застосуванні яких наприкінці визначаються вірогідності помилки *p*.

Велика група тестів призначена для з'ясування того, чи розрізняються дві або більше різних вибірок за своїми середніми значеннями або медіанами. При цьому враховується різниця між незалежними вибірками (різні спостереження) і залежними вибірками (різні змінні). Залежно від кількості вибірок (дві або більше), від того, чи є вибірки залежними,

чи належать змінні до інтервальної або порядкової шкали, чи підпорядковуються нормальному розподілу, застосовують спеціалізовані тести.

Дуже часто трапляється ситуація, коли порівнюються різні групи спостережень або значень змінних, що належать до номінальної шкали. У цьому випадку будуються таблиці спряженості. Інша група тестів стосується дослідження зв'язків між двома змінними, тобто виявлення кореляцій і відновлення регресій.

Окрім цих досить простих статистичних методів існують також складніші методи багатовимірного аналізу, в яких зазвичай одночасно використовується дуже багато змінних. Наприклад, якщо потребується звести велику кількість змінних до меншої кількості «в'язок змінних», які мають назву чинників, то проводиться факторний аналіз, якщо ж мета є протилежною (об'єднати задані спостереження, утворивши з них кластери), то застосовують кластерний аналіз.

2.3. Підготовка даних

2.3.1. Кодування і кодувальна таблиця

Для того щоб отримані дані можна було обробити, передусім слід створити ***кодувальну таблицю***, з допомогою якої встановлюється відповідність між окремими запитаннями анкети й змінними, що використовують при комп'ютерному обробленні даних. Наприклад, пункту анкети «Стать» може бути поставлена у відповідність змінна «sex».

Змінні – це елементи пам'яті, в які можна записувати значення, уведені з клавіатури. Імена змінних в SPSS для Windows можуть містити до восьми символів. Інше, детальніше ім'я було б занадто довгим. Імена змінних можуть складатися з букв латинського алфавіту, цифр і спеціальних символів, причому першим символом імені має бути буква.

Змінні можуть набувати різних значень. Змінна «sex» може мати два можливі значення: «жіночий» і «чоловічий». Кодувальна таблиця визначає кодові числа, що відповідають окремим значенням змінних, наприклад, значенню «жіночий» може відповідати цифра «1», а значенню «чоловічій» – «2».

Підсумуємо завдання, які вирішуються під час складання кодувальної таблиці:

– кодувальна таблиця встановлює відповідність між окремими запитаннями анкети й змінними;

– кодувальна таблиця встановлює відповідність між можливими значеннями змінних і кодовими числами.

2.3.2. Матриця даних

Припустимо, що було заповнено 30 анкет. Наведена табл. 2.1 має назву ***матриці даних***. Дані, призначені для оброблення в SPSS для Windows, мають бути подані у вигляді такої матриці.

Таблиця 2.1

№ п/п	Sex	Age	Employment
1	Жіноча	45	Викладач
2	Чоловіча	22	Студент
3	Чоловіча	19	Студент
4	Жіноча	42	Інженер
5	Чоловіча	34	Приватний підприємець
6	Жіноча	72	Пенсіонерка

Матриця даних складається з певної кількості рядків і стовпців, які утворюють прямокутну таблицю. При цьому кожен рядок відповідає одній анкеті, а кожен стовпець – одній змінній.

2.3.3. Запуск SPSS

Почнемо з уведення даних для невеликого прикладу аналізу.

Запустіть SPSS для Windows, двічі клацнувши лівою кнопкою миші на значку SPSS, якщо він знаходиться на робочому столі, або на кнопці Start (Пуск) і вибравши в пункті (Programs) Програми пункт SPSS for Windows.

Відкриється редактор даних SPSS (рис. 2.2), з якого починається робота з програмою.

Редактор даних – це одне з багатьох вікон SPSS. Тут можна вводити нові дані або завантажувати існуючі з файлів даних з допомогою команд меню File (Файл) Open (Відкрити).

Оскільки під час запуску SPSS жоден із файлів даних ще не завантажено, в заголовку редактора даних стоїть «Untitled» (Без імені). Над зображенням таблиці в редакторі даних є рядок меню і панель символів.

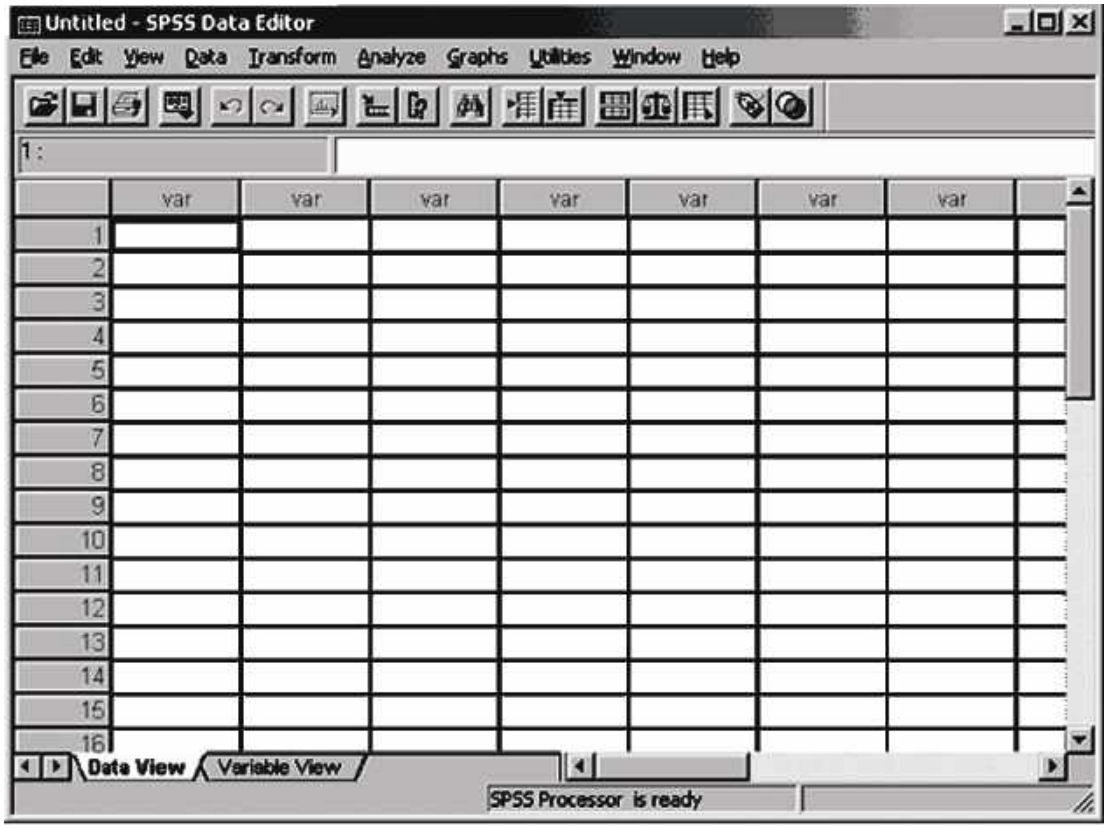


Рис. 2.2. Початковий вигляд вікна редактора даних

2.3.4. Редактор даних

Вікно редактора даних відкривається автоматично при завантаженні SPSS.

Зовнішній вигляд і властивості редактора даних SPSS схожі на зовнішній вигляд і властивості електронних таблиць (наприклад, Excel). Під електронною таблицею мається на увазі робочий аркуш, розділений на рядки й стовпці, що дає можливість просто й ефективно вводити дані. Окремі рядки таблиці відповідають окремим спостереженням. Наприклад, при оброблені даних опитування один рядок містить дані одного респондента. Окремі стовпці відповідають окремим змінним. При оброблені даних спостережень анкети в одній змінній зберігаються відповіді на окреме запитання. Окремі комірки таблиці містять значення змінних для кожного окремого спостереження (на відміну від електронних таблиць комірки в редакторі даних не можуть містити формул); у кожній комірці зберігається одне значення змінної.

Файл даних є прямокутною матрицею. У межах границь файла даних не існує «порожніх» комірок. Для числових змінних порожні комірки конвертуються в системні пропущені значення. Для строкових змінних порожня комірка є значущим значенням.

Визначення зміних

Визначити змінну – це зв'язати інформацію, вихідні дані з конкретною змінною. SPSS дає можливість не тільки визначати створювані змінні, але й перевизначати вже існуючі.

Щоб визначити змінну, слід перейти на вкладку Variable View (Перегляд даних), клацнувши на її ярличку мишею (рис. 2.3).

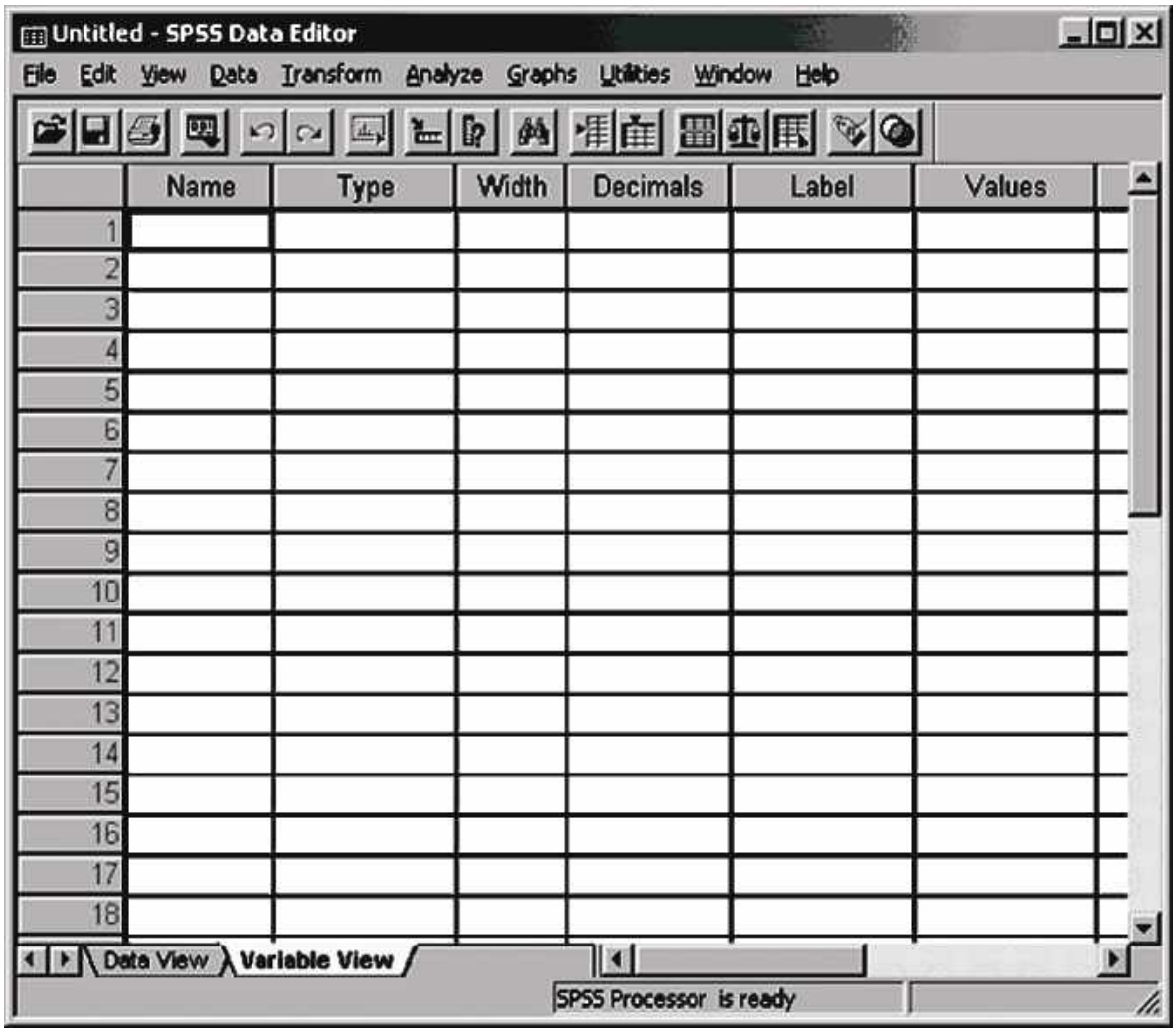


Рис. 2.3. Редактор даних: огляд змінних

Послідовність дій при визначенні змінних.

1. У стовпці Name (Ім’я) ввести ім'я змінної.

Імена змінних мають відповідати таким вимогам:

– ім'я має починатися з букви, інші символи можуть бути будь-якими символами, крім пробілів і спеціальних символів ! ? ' , * (крапки й знаки @, #, _, $ є допустимими);

– імена змінних не можуть закінчуватися крапкою;

– треба уникати змінних, яки закінчуються символом підкреслення;

– довжина імені змінної – не більше восьми знаків;

– не можна повторювати імена змінних, кожне ім'я має бути унікальним;

– імена змінних є нечутливими до регістра, імена «СТАТЬ», «Стать» й «стать» розглядаються як одне й те саме ім'я.

Уведення імені підтверджується натисканням на клавішу <Enter> або <Tab>.

2. У стовпці Type (Тип) вказати тип змінної (числовий, строковий, дата або ін.). Щоб задати тип змінної, треба клацнути в поле Type на кнопці із трьома крапками. Настроювання підтверджується кнопкою ОК, перехід до наступного поля – клавішею <Tab>.

3. У стовпці Width (Ширина) указати максимальну кількість знаків, яку може мати значення змінної, включаючи дробову частину. Щоб перейти до установки формату стовпця, нелюхідно натиснути клавішу <Tab>. Щоб змінити формат подання змінної, перенесений з діалогу Define Variable Type, треба клацнути на кнопці ліфта. У цьому випадку вибране значення ширини підтверджується клавішею <Tab>.

4. У стовпці Decimals (Дробова частина) увести кількість знаків після коми (якщо змінну задано десятковим дробом). Збільшення або зменшення значення, яке наведено в діалозі Define Variable Type, виконується за допомогою кнопки ліфта:

5. У стовпці Label (Мітка) створити мітку змінної – коментар, що дає можливість описати змінну більш докладно. Мітка змінної може містити до 256 символів. У мітках змінних розрізняються прописні й малі літери, вони відображаються в тому вигляді, у якому були введені.

6. У стовпці Values (Значення) увести значення (альтернативні відповіді) для кожної категорії відповіді. Наприклад, для змінної «Стать» можна задати мітку «жіночий» для значення «1» і мітку «чоловічий» – для значення «2». Для цього клацніть у поле Value Labels на кнопці . Відкриється діалогове вікно Define Value Labels (Визначення міток значень) (рис. 2.4).

Мітки значень визначаються таким чином:

– спочатку введіть у поле Value (Значення) число «1», натисніть клавішу <Tab>;

– уведіть у поле Value label (Мітка значення) текст «жіночий»;

– клацніть на кнопці Add (Додати), мітку значення буде додано в список, для цієї мети можна також нажати комбінацію клавіш <Alt>+<h>;

– повторіть ці дії для значень «2» – «чоловічий» і «0» – «немає даних». Максимально допустима довжина мітки значення становить 60 знаків.

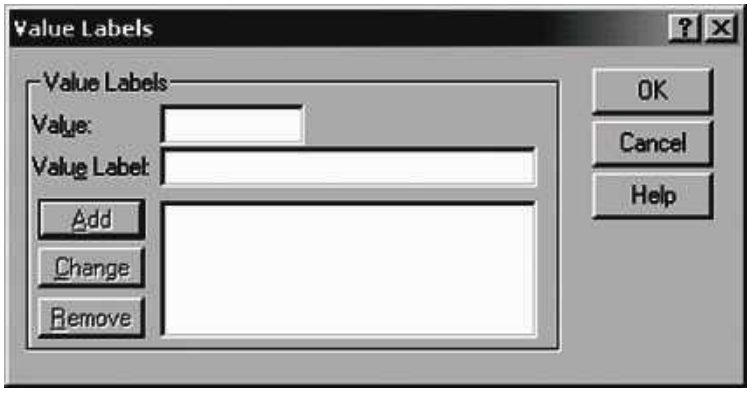


Рис. 2.4 Діалогове вікно Define Value Labels

Результат уведення значень в діалоговому вікні показано на рис. 2.5.

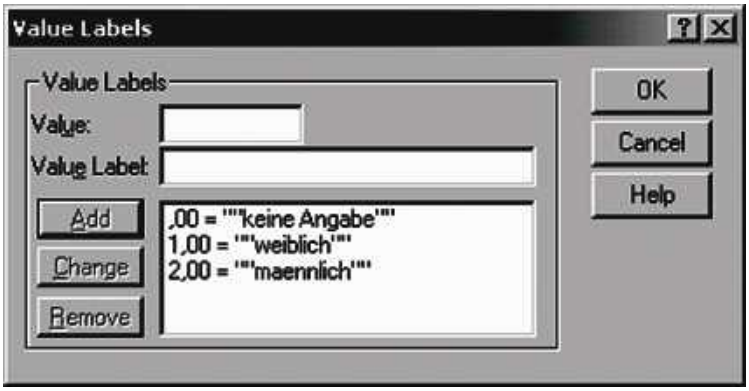


Рис. 2.5. Заповнення діалогового вікна Define Value Labels
(Визначення міток значень)

7. У стовпці Missing (пропущені значення) визначити коди для пропущених значень. У SPSS допускаються два види пропущених значень:

– які обумовлені системою (System-defined missing values): якщо в матриці даних є незаповнені чисельні комірки, система SPSS самостійно ідентифікує їх як пропущені значення (цей факт відображається в матриці даних за допомогою коми (,);

– що задаються користувачем (User-defined missing values): якщо в певних випадках у змінних немає значень, наприклад, якщо на запитання не було дано відповіді, відповідь є невідомою або існують інші причини, користувач може за допомогою кнопки Missing оголосити ці значення як пропущені. Пропущені значення можна виключити з наступних обчислень. Наприклад, пропущеним значенням, обумовленим користувачем, оголосити варіант відповіді «99» (немає даних) для змінної «Стать» (рис. 2.6).

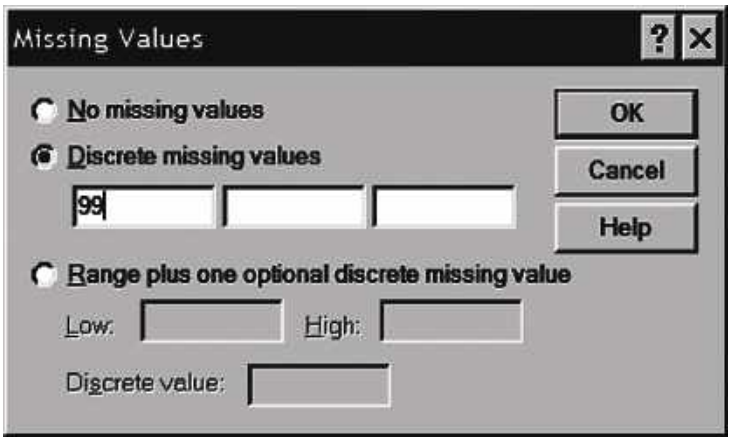


Рис. 2.6. Вікно Missing Values (Пропущені значення)

8. У стовпці Columns (Стовпці) можна, якщо забажати, змінити ширину стовпця змінної при відображенні значень (ширина стовпця змінюється у вкладці Data View).

9. У стовпці Alignment (Вирівнювання) можна задати вид вирівнювання значень, тобто визначити, як вони будуть відображатися в таблиці. Можливі види вирівнювання – Right (по правому краю), Left (по лівому краю) і Center (по центру). Щоб задати вид вирівнювання, клацніть на кнопці .

10. У стовпці Measure (Шкала виміру) можна задати шкалу змінної Scale, Ordinal або Nominal (рис. 2.7). За замовчуванням приймається метрична шкала виміру. Правда, це розходження має значення тільки при створенні інтерактивних графіків, де номінальна й порядкова шкали вимірів поєднуються в «категоріальний» тип. Якщо завантажений файл, створений в попередніх версіях SPSS, або шкала вимірів не визначається явно, SPSS спочатку автоматично припускає метричну шкалу. Однак якщо відповідна змінна має мітки значень або приймає менше 24 різних значень, то задається порядкова шкала.

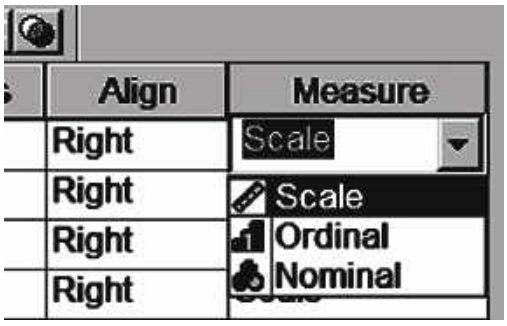


Рис. 2.7. Список вибору шкали змінної

Щоб задати тип змінної, треба перейти у стовпець Type, натиснути на кнопку, що виникла в комірці, після чого на екрані виникне діалогове вікно «Тип змінної», у якому вибирають із переліку необхідний тип (рис. 2.8).

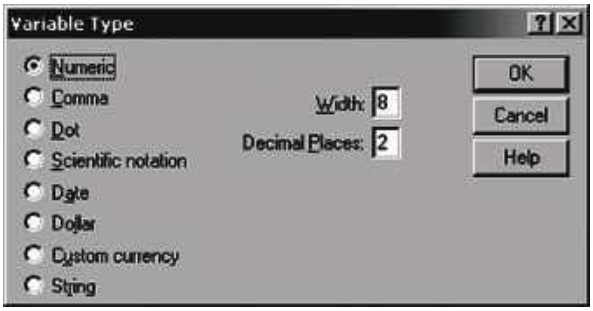


Рис. 2.8. Діалогове вікно Define Variable Type (для числової змінної)

Якщо необхідно змінити тип змінної, треба клацнути в комірці на кнопці з трьома точками (...), відкриється Діалогове вікно Define Variable Type.

У SPSS існують вісім типів змінних.

1. Numeric (Числовий) – цифри зі знаком «+» або «–» і десятковим роздільником. Знак «+» перед числом на відміну від знака «–» не відображається. У текстовому полі Length (Довжина) задається максимальна кількість знаків, включаючи позицію для десяткового роздільника. У текстовому полі Decimals (Десяткові розряди) уводиться кількість відображуваних знаків дробової частини. Тип десяткового роздільника залежить від настроювань діалогового вікна «Мова й стандарти» (Regional Settings) на панелі керування Windows. Точне значення змінної зберігається усередині програми, а Редактор даних відображає на екрані лише задану кількість десяткових розрядів. Значення, які мають більше десяткових розрядів, округлюються. Для обчислень застосовується точне значення.

2. Comma (Кома) – цифри зі знаком «+» або «–», крапкою як десятковим роздільником і однією або кількома комами як роздільниками груп розрядів. Якщо коми опускаються при введенні, вони вставляються автоматично. Довжина такої змінної дорівнює максимальній кількості знаків, включаючи десятковий роздільник і коми між групами розрядів.

3. Dot (Крапка) – цифри зі знаком «+» або «–», комою як десятковим роздільником і однією або кількома крапками як роздільниками груп розрядів. Якщо крапки опускаються при введенні, вони вставляються автоматично.

4. Scientific notation (Експонентне подання) – експонентне подання, про яке свідчить буква E або D, що міститься в числі, а також знак «плюс» або «мінус».

5. Date (Дата). Допустимі значення – дата і/(або) час. У форматах дати як роздільники між значеннями дня, місяця й числа можуть застосовуватися коса риска, дефіс, пробіл, кома або крапка. Можна вибрати один з декількох форматів дати (dd-mm-yyyy, dd-mmm-yy, mm/dd/yyyy і т.д.). Дата у форматі dd-mmm-yy відображається з роздільником-дефісом і скороченням назви місяця із трьох букв. Дата у

форматах dd/mm/yy і mm/dd/yy відображається з роздільником – косою рисою і номером місяця замість назви. Усього доступно 27 різних форматів дати й часу, які відображаються в списку, що розвертається. У форматах часу як роздільники між значеннями годин, хвилин і секунд можуть використовуватися двокрапка, крапка або пробіл.

6. Dollar (Долар) – знак долара, крапка як десятковий роздільник і коми як роздільники груп розрядів. Якщо знак долара або коми опускаються при введенні, вони вставляються автоматично.

7. Special currency (Спеціальна валюта). Користувач може задавати власні формати валюти. У поле Length уводиться максимальна кількість знаків, заданих користувачем. Позначення валюти при введенні не вказується; воно вставляється автоматично. Формати відображення валюти задаються за допомогою вкладки Currency (Валюта), що відкривається командою меню Edit (Виправлення) Options (Параметри).

8. String (Строкова) – рядок символів. До допустимих значень належать букви, цифри й спеціальні символи. Розрізняють короткі й довгі строкові змінні. Короткі строкові змінні можуть містити не більше восьми знаків. У довгих строкові змінні значення доповнюються пробілами до максимальної довжини.

Заповнені вкладки Data View (Перегляд даних) и Variable View (Перегляд зміних) вікна редактора даних показано на рис. 2.9, 2.10.

Jeans-Shara.sav - SPSS Data Editor

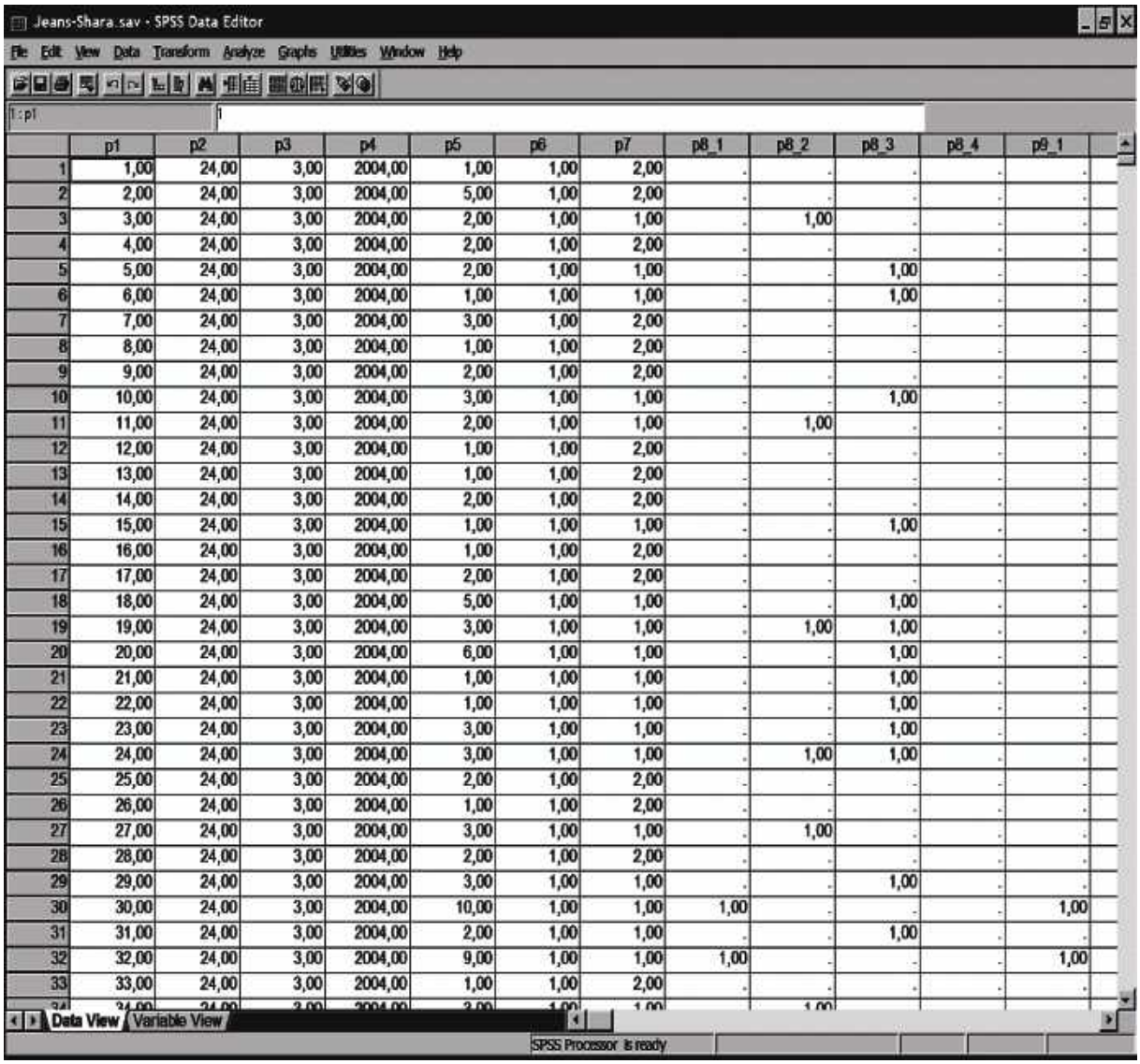
File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Value
1	p1	Numeric	8	2	Номер анкеты	None
2	p2	Numeric	8	2	Дата (день)	None
3	p3	Numeric	8	2	Дата (месяц)	None
4	p4	Numeric	8	2	Дата (год)	None
5	p5	Numeric	8	2	Продолжительность инт	None
6	p6	Numeric	8	2	Город	{1,00, Кие
7	p7	Numeric	8	2	C1. Пользуетесь ли Вы	{1,00, Да}
8	p8_1	Numeric	8	2	C2. Абонентом какого оп	{,00, Нет}
9	p8_2	Numeric	8	2	C2. Абонентом какого оп	{,00, Нет}
10	p8_3	Numeric	8	2	C2. Абонентом какого оп	{,00, Нет}
11	p8_4	Numeric	8	2	C2. Абонентом какого оп	{,00, Нет}
12	p9_1	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
13	p9_2	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
14	p9_3	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
15	p9_4	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
16	p9_5	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
17	p9_6	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
18	p9_7	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}
19	p9_8	Numeric	8	2	C3. А у Вас: ? ТИП ПАКЕ	{,00, Нет}

Data View | Variable View

SPSS Processor is ready

Рис. 2.9. Змінні на вкладці Variable View вікна редактора даних



	p1	p2	p3	p4	p5	p6	p7	p8_1	p8_2	p8_3	p8_4	p9_1
1	1,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
2	2,00	24,00	3,00	2004,00	5,00	1,00	2,00	.	.	.	.	.
3	3,00	24,00	3,00	2004,00	2,00	1,00	1,00	.	1,00	.	.	.
4	4,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
5	5,00	24,00	3,00	2004,00	2,00	1,00	1,00	.	.	1,00	.	.
6	6,00	24,00	3,00	2004,00	1,00	1,00	1,00	.	.	1,00	.	.
7	7,00	24,00	3,00	2004,00	3,00	1,00	2,00	.	.	.	.	.
8	8,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
9	9,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
10	10,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	.	1,00	.	.
11	11,00	24,00	3,00	2004,00	2,00	1,00	1,00	.	1,00	.	.	.
12	12,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
13	13,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
14	14,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
15	15,00	24,00	3,00	2004,00	1,00	1,00	1,00	.	.	1,00	.	.
16	16,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
17	17,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
18	18,00	24,00	3,00	2004,00	5,00	1,00	1,00	.	.	1,00	.	.
19	19,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	1,00	1,00	.	.
20	20,00	24,00	3,00	2004,00	6,00	1,00	1,00	.	.	1,00	.	.
21	21,00	24,00	3,00	2004,00	1,00	1,00	1,00	.	.	1,00	.	.
22	22,00	24,00	3,00	2004,00	1,00	1,00	1,00	.	.	1,00	.	.
23	23,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	.	1,00	.	.
24	24,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	1,00	1,00	.	.
25	25,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
26	26,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.
27	27,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	1,00	.	.	.
28	28,00	24,00	3,00	2004,00	2,00	1,00	2,00	.	.	.	.	.
29	29,00	24,00	3,00	2004,00	3,00	1,00	1,00	.	.	1,00	.	.
30	30,00	24,00	3,00	2004,00	10,00	1,00	1,00	1,00	.	.	.	1,00
31	31,00	24,00	3,00	2004,00	2,00	1,00	1,00	.	.	1,00	.	.
32	32,00	24,00	3,00	2004,00	9,00	1,00	1,00	1,00	.	.	.	1,00
33	33,00	24,00	3,00	2004,00	1,00	1,00	2,00	.	.	.	.	.

Рис. 2.10. Дані на вкладці Data View (Перегляд даних) вікна редактора даних

2.3.5. Вікно виведення і його редагування

Результати зроблених розрахунків по черзі виникатимуть у вікні перегляду, тобто згідно з установками кожен наступний результат розрахунку буде поміщатися в кінець вікна.

Вікно перегляду складається з двох частин. У лівій частині знаходиться ієрархія (перегляд змісту) результатів; у праву частину поміщають таблиці з результатами розрахунків і побудовані графіки. Ширину цих частин вікна можна змінювати перетяганням розділової межі за допомогою миші.

Результати кожної виконаної статистичної процедури, а також графічне виведення, відображаються у вікні перегляду у вигляді бло-

ка, причому кожен блок є окремим об'єктом (рис. 2.11). В ієрархії кожен блок озаглавлюється відповідним ім'ям процедури, перед яким встановлюється значок блока. Цьому значку передує невеликий чотирикутник, в якому спочатку ставиться знак «мінус». Усередині кожного блока спочатку йде заголовок і примітки, далі – перелічення елементів блока, яким теж передують відповідні символи. Завдяки такій конструкції ієрархії об'єктів можна здійснювати пошук необхідних елементів, переставляти їх місцями, копіювати, видаляти і т. д.

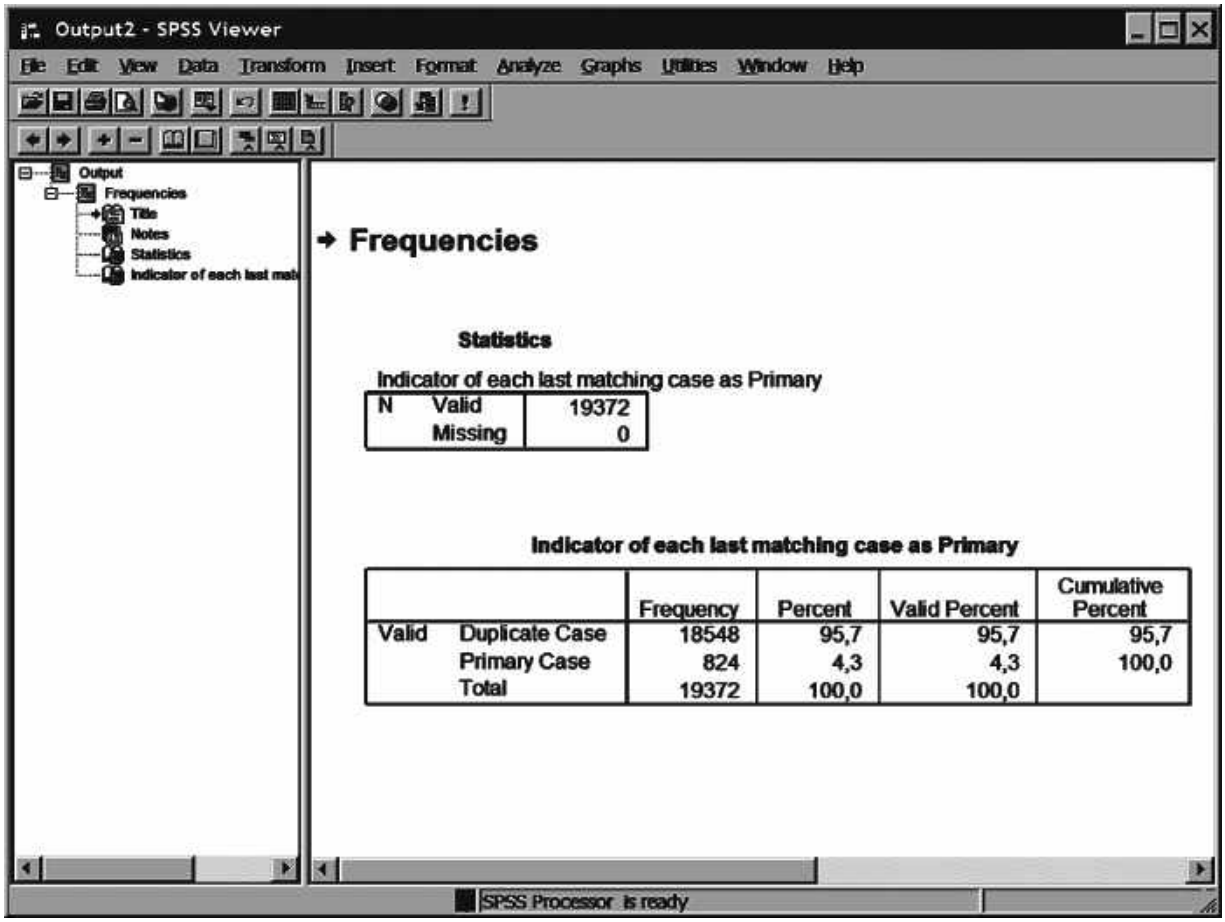


Рис. 2.11. Вікно перегляду у вигляді частотної таблиці

Пошук у вікні перегляду

Щоб побачити в області виведення необхідний об'єкт або елемент, не треба прокручувати усе вікно перегляду. Щоб потрапити в необхідне місце, треба клацнути на відповідному символі в ієрархії.

Видалення у вікні перегляду

Щоб видалити деякі елементи результатів розрахунків, необхідно клацнути на відповідному символі й вибрати в меню Edit (Правка) Delete (Видалити). Можна також просто натиснути на клавіатурі клавішу <Delete>.

Прихований режим

Замість того щоб видаляти частини блоків, можна на деякий час їх «приховати». Вони стають невидимими на екрані й під час друку.

Щоб приховати частину результатів, необхідно клацнути двічі на відповідному символі в ієрархії або виділити необхідний елемент одним клацанням з наступним вибором меню View (Вид) Hide (Приховати).

Якщо необхідно знову зробити елемент видимим, треба повторно клацнути двічі на значку або виділити його одним клацанням з наступним вибором меню View (Вид) Show (Показати)

Якщо ж необхідно приховати цілий блок, що містить усе виведення окремої процедури, треба клацнути на маленькому квадратику ліворуч від значка блока. При цьому знак «мінус» в квадратику перетвориться на знак «плюс» і ця процедура разом з усім її вмістом зникне.

Можна також виділити значок блока й зробити вибір меню View (Вид) Collapse (Згорнути).

Блок можна знову зробити видимим з допомогою повторного клацання на квадратику, при цьому знак «плюс» знову буде замінено на знак «мінус». Можна також клацанням виділити значок блока й вибрати в меню View (Вид) Expand (Розвернути).

Перестановка у вікні перегляду

Для переміщення деякої частини результатів на інше місце треба виділити відповідний значок (якщо необхідно, то значок блока) і, утримуючи натиснутою ліву кнопку миші, перемістити його до того елемента, після якого необхідно розташувати ці результати або блок.

Альтернативна можливість переміщення елементів полягає у виділенні значка, що відповідає необхідній частині інформації з наступним вибором меню Edit (Правка) Cut (Вирізувати).

Потім виділити значок, позаду якого необхідно вставити вирізаний елемент і вибрати в меню Edit (Правка) Paste After (Вставити потім).

Копіювання у вікні перегляду

Щоб скопіювати яку-небудь частину інформації в інше місце (при цьому зберігти її на попередньому місці), необхідно клацнути на значку, що відповідає необхідному елементу або блоку, не відпускаючи кнопку миші, натиснути на клавіатурі клавішу <Ctrl> і перетягти значок до того елемента, після якого має бути вставлено копійовуваний елемент.

Можна також клацнути на значку копійовуваного елемента і вибрати в меню опції Edit (Правка) Copy (Копіювати).

Потім клацнути на значку елемента, після якого має бути вставлено копійовуваний елемент, і вибрати в меню Edit (Правка) Paste After (Вставити потім).

Виведення приміток

Під час читання результатів розрахунків дуже допомагають примітки. У них знаходиться інформація про відповідний файл і загальні установки програми. За замовчуванням ці примітки спочатку є прихованими, але їх можна зробити видимими, якщо, наприклад, двічі клацнути на значку примітки (Notes).

Змінення розміру і типу шрифту ієрархічного списку

Щоб змінити розмір знаків і тип шрифту в ієрархічному списку, вибрати в меню View (Вид) Outline Size (Розмір знаків) і відповідно View (Вид) Outline Font (Шрифт знаків), після чого виникне можливість вибору серед трьох розмірів (Small (Дрібний), Medium (Середній), Large (Великий)) і великої кількості шрифтів.

Редагування мобільних таблиць

Уже було розглянуто, як за допомогою ієрархічного списку у вікні перегляду можна управляти виведенням елементів результатів розрахунків.

Багато елементів результатів розрахунків наведено у вигляді мобільних таблиць. Це нова форма таблиць, яка дає можливість змінювати місцями рядки, стовпці й шари так, щоб результати можна було б оцінити з різних точок зору. Хорошим прикладом їх застосування можуть бути, передусім, таблиці спряженості.

Зазвичай SPSS створює таблиці автоматично, самостійно вирішуючи, які показники розмістити в рядках, які – в стовпцях. Не завжди зручно покладатися на вибір програми, тому для таких випадків SPSS має можливість редагування таблиць.

Щоб дізнатися про можливість редагування, яке надає техніка мобільних таблиць, треба клацнути двічі на цій таблиці, після чого буде активовано редактор мобільних таблиць (рис. 2.12).

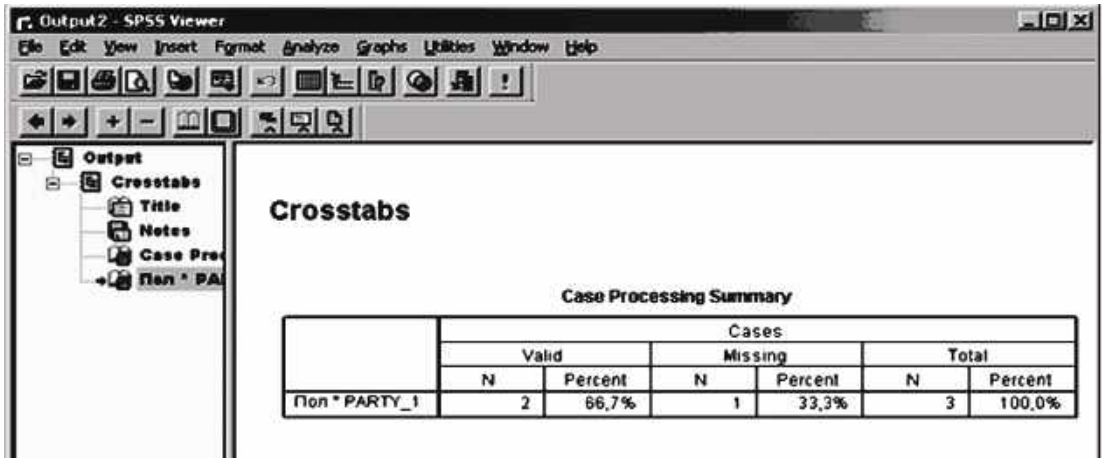


Рис. 2.12. Редактор мобільних таблиць

Вибрати в меню Pivot (Мобільна таблиця) Pivoting Trays (Поля обертання). Відкриється вікно Pivoting Trays (рис. 2.13), що містить три панелі, позначені як Layer (Шар), Row (Рядок) і Column (Стовпець). На панелі рядків розташовано два значки, а на панелі стовпців – один. Для того, щоб отримати інформацію про призначення цих значків, треба пройтися по них покажчиком, ненадовго затримуючи його над значками, після чого буде виведено мітки відповідних змінних.

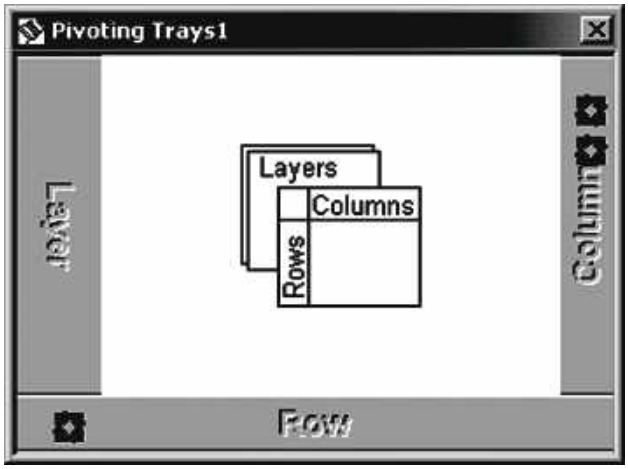


Рис. 2.13. Вікно Pivoting Trays (Поля обертання)

За допомогою цих значків можна змінити структуру таблиці. Клацнути, наприклад, на значку Statistics панелі рядків і перетягнути його мишею за значок, що знаходиться на панелі стовпців. Після цього відсоткові показники відобразяться в колонках таблиці.

Щоб вийти з редактора мобільних таблиць, необхідно клацнути в якій-небудь точці за межами виділеної таблиці.

Уведення даних

Дані в SPSS можна вводити в будь-якому порядку – за спостереженнями (анкетам) або за змінними (ознаками), інакше кажучи, за рядками (змінними) або за стовпцями (об’єктами). Безумовно, результати соціологічних опитувань зручніше вводити за рядками (анкетами). При цьому варто пам'ятати:

– дані вводяться в активну комірку (яку виділено жирною рамкою, а ім'я змінної й номер рядка відображаються в лівому верхньому куті редактора даних);

– значення даних не записуються, доки не буде натиснуто клавішу Enter або не буде здійснено перехід до іншої комірки;

– дані цифрового типу можна вводити, не визначаючи тип змінної;

– дані всіх інших типів можна вводити після визначення типу змінної;

– коли вводиться значення в порожній стовпець, SPSS автоматично створює нову змінну й присвоює їй ім'я.

Щоб увести числові дані, необхідно:

а) вибрати комірку у редакторі даних на аркуші Data View;

б) увести число (значення), що буде відображатися редактором даних у поле редактора комірки (аналогічно Excel);

в) нажати клавішу Enter або вибрати іншу комірку.

Щоб увести нечислові дані, треба:

а) двічі клацнути мишею на ім’я змінної у верхній частині редактора даних або де-небудь у колонці цієї змінної, при цьому відразу перейти на сторінку Variable View до опису вибраної змінної;

б) визначити змінну;

в) повернутися на сторінку Data View;

г) увести дані в стовпець нової змінної.

Редагування даних

Іноді виникає необхідність змінити дані, які введено в файл, тому редактор даних SPSS дає такі можливості.

1. Змінити вміст комірки, для цього клацнути на комірці, вміст якої необхідно відредагувати, увести нове значення, потім перейти до будь-якої сусідньої клітинки з допомогою клавіш або клавіш зі стрілками.

2. Вставити новий об’єкт (рядок). Щоб вставити рядок між уже набраними, слід вибрати будь-яку комірку в рядку нижче від позиції, де необхідно вставити новий рядок, і клацнути на кнопці Insert Cases (Вставка об’єкта), внаслідок чого буде вставлено рядок, заповнений системними пропущеними значеннями. Рекомендується спочатку визначити всі змінні (на сторінці Variable View), а потім уводити дані.

Якщо було надруковано символ, що не допускається заданим типом змінної, редактор даних подає тихий звуковий сигнал і не вводить цей символ до комірки.

Для строкових змінних не допускаються символи за межами заданої ширини.

Для числових змінних цілі значення, що перевищують задану ширину (Columns), можуть бути введені (відображаються на екрані), але редактор даних покаже зірочки для позначення того, що значення перевищує задану ширину.

3. Вставити нову змінну. Щоб вставити нову змінну між двома сусідніми, слід клацнути на правій, а потім на кнопці Insert Variable (Вставка змінної), після чого буде вставлено порожній стовпчик.

4. Копіювати й вирізати значення даних. Виділити необхідні осередки, в меню Edit (Редагування) вибрати команду Copy (Копіювання) або Cut (Вирізання).

Якщо тип змінної, котра копіюється, не збігається з типом, заданим у тих комірках, куди копіюється, то SPSS зробить спробу перетворити значення, що копіюються.

Якщо перетворення є неможливим, то SPSS вставить системне пропущене значення в комірку призначення. Тому рекомендується перед копіюванням переконатися, що типи змінних збігаються. Якщо ж вони не збігаються, треба перевизначити типи комірок призначення.

5. Вставити комірки. Комірки, яки раніше скопіювали, можна вставити командою Paste (Вставка).

Виявлення помилок уведення

Найточнійший метод перевірки даних (тобто значень усіх змінних) на помилки під час введення полягає в тому, щоб командами меню Analyze (Аналіз), Reports (Звіти), Case summaries... (Зведення спостережень) вивести їх список і зрівняти кожне значення з оригіналом (наприклад, анкетою). Однак цей спосіб потребує дуже багато часу, особливо при великому обсязі даних. Тому зважитися на проведення такої нудної й стомлюючої роботи можна тільки в тих випадках, коли обсяг даних є обмеженим, в інших рекомендується проводити частотний аналіз значень змінних. Для цього призначено команди меню Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies... (Частоти). Результати цього аналізу при уважному розглядуванні дають можливість виявити недопустимі значення. Наприклад, якщо змінна містить дані про зріст у сантиметрах, то значення 384, що виявляється під час частотного аналізу, явно свідчить про те, що в даних є помилка. Після проведення частотного аналізу це значення можна відшукати у файлі даних і виправити. Отже, вивчаючи частотні таблиці особливу увагу треба звертати на максимальне й мінімальне значення. Однак якщо замість віку 65 років було уведено, наприклад, значення 56, то з допомогою частотної таблиці цю помилку виявити неможливо.

Змістовий аналіз даних можна провести шляхом створення таблиць спряженості. Наприклад, якщо дані взято з анкети, у якій було питання про родиний стан (неодруженний/незаміжня, одружений/заміжня, удівець/удова, розведений/розведена, то, побудувавши таблицю спряженості для цього питання й питання типу: «Якщо у Вас є родина, то чи прийнято проводити відпустку окремо?», легко можна виявити, чи відповіли на нього тільки одружені респонденти.

За допомогою описаних і подібних способів можна виявити велику кількість помилок уведення. Навіть якщо спостережень кілька тисяч, то навіть одне суперечливе значення завдає шкоди дослідженню: створюється враження, що роботу зі збору про підготовку інформації виконано поверхово.

2.4. Частотний аналіз

Першим етапом статистичного аналізу даних, як правило, є частотний аналіз. Команда Frequencies (Частоти) є найпростішою й найпоширенішою. Дія команди зводиться до підрахунку кількості об'єктів у кожній категорії змінної. Це має назву розподілу частот за категоріями змінної. Результат, що виводиться, для кожної категорії містить мітку значень змінної, саме значення змінної, частоту, відсоток і накопичений відсоток від загальної частоти.

2.4.1. Частотні таблиці

Розглянемо будування частотних таблиць.

1. Завантажити файл даних, вибравши команди меню File (Файл) Open.. (Відкрити..). Виникне діалог Open File (Відкрити файл).

2. Вибрати потрібний файл і підтвердити вибір кнопкою Open (Відкрити). Файл виникне в Редакторі даних.

3. Вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти). Виникне діалогове вікно Frequencies (рис. 2.14).

4. Кнопкою з трикутником перенести змінну, для якої треба обчислити розподіл частот, в список вихідних змінних і підтвердити операцію кнопкою ОК.

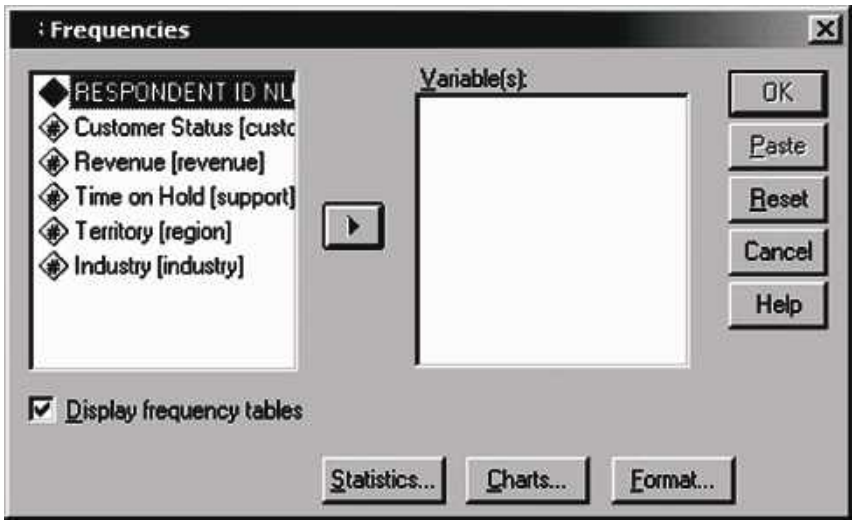


Рис. 2.14. Діалогове вікно Frequencies (Частоти)

Результати виникнуть у вікні перегляду результатів у вигляді частотної таблиці. Перед самою частотною таблицею виводиться невелика таблиця з оглядом допустимих значень і таких, яких немає.

Кожен рядок частотної таблиці описує одне можливе значення. Рядок з позначкою «немає даних» являє собою спостереження, в яких

не було дано ніякої відповіді. Перший стовпець містить мітки окремих значень (наприклад, для статі – «чоловічий», «жіночий»). У другому стовпці під заголовком Frequency (Частота) наведено частоту кожного з варіантів. У третьому стовпці Percent показано відсоткову частоту кожної відповіді. Відсоткова частота відповідає відношенню кожного з варіантів відповіді до загальної кількості опитуваних, включаючи втрачені значення. У четвертому стовпці Valid Percent дано допустиме процентне значення. При визначенні цього значення втрачені дані вилучаються. Останній стовпець містить накопичені (кумулятивні) відсотки Cumulative Percent – суму процентних частот допустимих відповідей. В останньому рядку міститься сума усіх стовпців (Усього).

2.4.2. Виведення статистичних характеристик

Щоб отримати описову статистику числових змінних, можна клацнути в діалозі Frequencies на кнопці Statistics.. (Статистика). Відкриється діалогове вікно Frequencies : Statistics (Частоти: Статистика) (рис. 2.15).

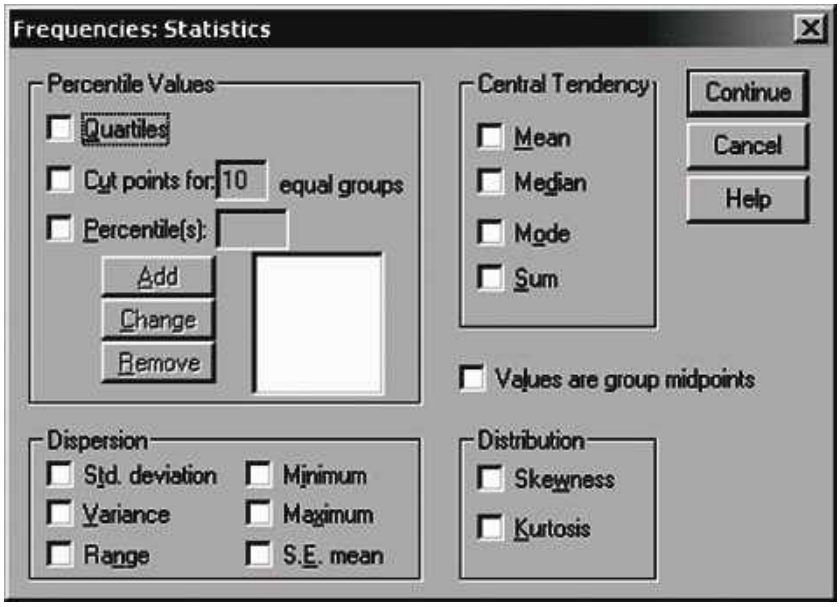


Рис. 2.15. Діалогове вікно Frequencies: Statistics

У групі Percentile Values (Значення процентилів) можна вибрати такі варіанти:

– Quartiks (Квартилі) – буде показано перший, другий і третій квартилі. Перший квартиль (Q_1) – це точка на шкалі виміряних значень, нижче (лівіше) за яку розміщено 25 % виміряних значень. Другий квартиль (Q_2) – це точка, нижче за яку розміщено 50 % виміряних значень. Другий квартиль також має назву медіани. Третій квартиль (Q_3) – це точка на шкалі виміряних значень, нижче за яку розміщено 75 %

значень. Якщо дані є тільки у вигляді порядкового відношення, то як міра розкиду використовується міжквартильная широта. Вона визначається як

$$Q = \frac{Q_3 - Q_1}{2};$$

– Cut points (Точки розділу) – будуть обчислені значення процентилів, що поділяють вибірку на групи спостережень, які мають однакову ширину, тобто містять одну й ту саму кількість виміряних значень. За замовчуванням пропонується 10 груп. Якщо задати, наприклад, 4 групи, то будуть показані квартилі, тобто квартилі відповідають процентилям 25, 50 і 75. Видно, що кількість показуваних процентилів на одиницю менша від заданої кількості груп;

– Percentile (Процентилі) – маються на увазі значення процентилів, що визначає користувач. Увести значення процентиля в межах від 0 до 100 і клацнути на кнопці Add (Додати). Повторити ці дії для усіх бажаних значень процентилів. Значення в порядку зростання будуть показані в списку. Наприклад, якщо ввести значення 25, 50 і 75, то отримаємо квартилі. Можна задавати будь-які значення процентилів, наприклад, 37 і 83. У першому випадку (37) буде показано значення вибраної змінної, нижче за яке лежать 37 % значень, а в другому випадку (83) – значення, нижче за яке лежать 83 % значень.

У групі Dispersion (Розкид) можна вибрати такі способи розкиду:

– Std. deviation (Стандартне відхилення) – це міра розкиду виміряних величин, яка дорівнює квадратному корню з дисперсії. В інтервалі шириною, що дорівнює подвоєному стандартному відхиленню, яке відкладено по обидва боки від середнього значення, розміщено майже 67 % усіх значень вибірки, що підпорядковуються нормальному розподілу;

– Variance (Дисперсія) – це квадрат стандартного відхилення і, отже, ця характеристика також є мірою розкиду виміряних величин, яка визначається як сума квадратів відхилень усіх виміряних значень від їхнього середньоарифметичного значення, поділена на кількість вимірів мінус одиниця;

– Range (Розмах) – це різниця між найбільшим (максимумом) і найменшим (мінімумом) значеннями;

– Minimum (Мінімум) – найменше значення;

– Maximum (Максимум) – найбільше значення;

– S.E. mean (Стандартна помилка) – це стандартна помилка середнього значення. В інтервалі шириною, що дорівнює подвоєній стандартній помилці, відкладеному навколо середнього значення, розміщено середнє значення генеральної сукупності з імовірністю близько

67 %. Стандартна помилка визначається як стандартне відхилення, що поділено на квадратний корінь з обсягу вибірки.

Зазвичай заходами розкиду змінних, таких, що належать до інтервальної шкали і підпорядковуються нормальному розподілу, є стандартне відхилення й стандартна помилка. Як було зазначено вище, стандартне відхилення дає можливість задати діапазон розкиду окремих значень. В одному діапазоні стандартного відхилення (що охоплює ширину стандартного відхилення в обидва боки від середнього значення) розміщено близько 67 % значень, в діапазоні подвоєного стандартного відхилення – близько 95 %, а в діапазоні потрійного стандартного відхилення – близько 99 % значень.

У групі Central Tendency (Середні) можна вибрати такі характеристики:

– Mean (Середнє значення) – це арифметичне середнє виміряних значень, яке визначається як сума значень, що ділиться на їх кількість. Наприклад, якщо є 12 виміряних значень і їх сума становить 600, то середнє значення $x = 600 : 12 = 50$;

– Median (Медіана) – це точка на шкалі виміряних значень, вище й нижче за яку лежить по половині усіх виміряних значень;

– Mode (Мода) – це значення, яке найчастіше трапляється у вибірці; якщо одна й та сама найбільша частота трапляється у декількох значень, то вибирається найменша з них;

– Sum (Сума) – сума усіх значень.

У групі Distribution (Розподіл) можна вибрати такі способи несиметричного розподілу:

– Skewness (Коефіцієнт асиметрії) – це міра відхилення розподілу частоти від симетричного розподілу, тобто такого, коли на однаковому віддаленні від середнього значення по обидва боки вибірки даних розміщено однакову кількість значень. Якщо спостереження підпорядковуються нормальному розподілу, то асиметрія дорівнює нулю. Для перевірки на нормальний розподіл можна застосовувати таке правило: якщо асиметрія значно відрізняється від нуля, то гіпотезу про те, що дані взято з нормально розподіленої генеральної сукупності, слід відкинути. Якщо вершину асиметричного розподілу зсунуто до менших значень, то говорять про додатну асиметрію, у протилежному випадку – про від'ємну;

– Kurtosis (Коефіцієнт варіації або ексцес) свідчить про те, що розподіл є пологим (при великому значенні коефіцієнта) або крутим. Коефіцієнт варіації дорівнює нулю, якщо спостереження підпорядковуються нормальному розподілу. Тому для перевірки на нормальний розподіл можна застосовувати ще одне правило: якщо коефіцієнт варіації значно відрізняється від нуля, то гіпотезу про те, що дані взято з нормально розподіленої генеральної сукупності, слід відкинути.

Як правило, для змінних, що належать до інтервальної шкали і підпорядковуються нормальному розподілу, як основну характеристику використовують середнє значення, а як міру розкиду – стандартне відхилення або стандартну помилку, для порядкових або інтервальних змінних, що не підпорядковуються нормальному розподілу, – відповідно медіану або перший і третій квартилі. Для змінних, яки належать до номінальної шкали, не можна дати інших значущих характеристик, окрім моди.

У діалозі є ще один прапорець: Values are group midpoints (Значення є середніми точками груп). Якщо встановити цей прапорець, то при обчисленні медіани й інших значень процентилів оцінки цих характеристик визначатимуться для концентрованих даних.

Для того щоб для досліджуваної змінної визначити такі характеристики, як середнє значення, медіана, мода, квартилі, стандартне відхилення, дисперсія, розмах, мінімум, максимум, стандартна помилка, асиметрія йексцес, треба зробити таке:

– вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти);

– у діалозі Frequencies клацнути на кнопці Reset (Скидання), щоб відмінити попередні настройки;

– перенести змінну в список вихідних змінних;

– клацнути на кнопці Statistics (Статистика);

– у діалозі Frequencies: Statistics встановити прапорці бажаних характеристик і клацнути на кнопці Continue (Продовжити) для повернння в діалог Frequencies;

– у діалозі Frequencies деактивувати опцію Display frequency tables (Показувати частотні таблиці); клацнути на кнопці ОК.

У вікні перегляду виникнуть результати бажаних характеристик.

2.4.3. Формати частотних таблиць

До форматування частотних таблиць належить сортування за спадною частотою.

Завантажити файл. Вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти).

Перенести змінну f, частоту якої розраховуємо, в список вихідних змінних.

Клацнути на кнопці Format... Відкриється діалогове вікно Frequencies: Format (Частоти: Формат) (рис. 2.16).

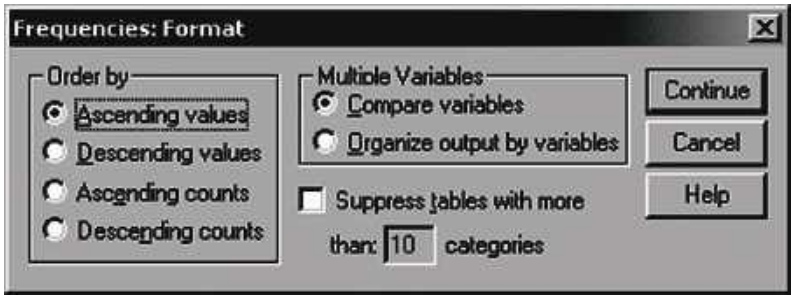


Рис. 2.16. Діалогове вікно Frequencies : Format

У групі Order by (Сортувати за) можна вибрати порядок, в якому будуть відображені значення в частотній таблиці. Можливі такі варіанти:

– Ascending values (За збільшенням значень) – дані сортуються за збільшенням значень; це настроювання за замовчуванням;

– Descending values (За убуванням значень) – дані сортуються за убуванням значень.

Вибрати варіант Descending counts.

Підтвердити вибір кнопкою Continue (Продовжити).

Клацнути на кнопці ОК, щоб почати обчислення. Отримаємо частотну таблицю, в якій частоти відсортовани за убуванням значень.

– Ascending counts (За збільшенням частот) – дані сортуються за збільшенням частот;

– Descending counts (За зменьшенням частот) – категорії сортуються за зменьшенням частот.

Крім того, прапорець Suppress tables – with more than … categories (Не виводити таблиці більш ніж ... категоріями) дає можливість уникнути виведення довгих частотних таблиць.

2.4.4. Графічне подання результатів частотного розподілу

Результати частотного розподілу можна подати графічно у вигляді стовпчастих діаграм, гістограм тощо.

Стовпчасті діаграми

Розглянемо створення стовпчастої діаграми для частотного розподілу досліджуваної зміної *f*. Для цього необхідно зробити таке:

– вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти);

– перенести змінну *f* в список вихідних змінних;

– клацнути на кнопці Charts ... (Діаграми), відкриється діалогове вікно Frequencies: Charts (Частоти: Діаграми) (рис. 2.17);

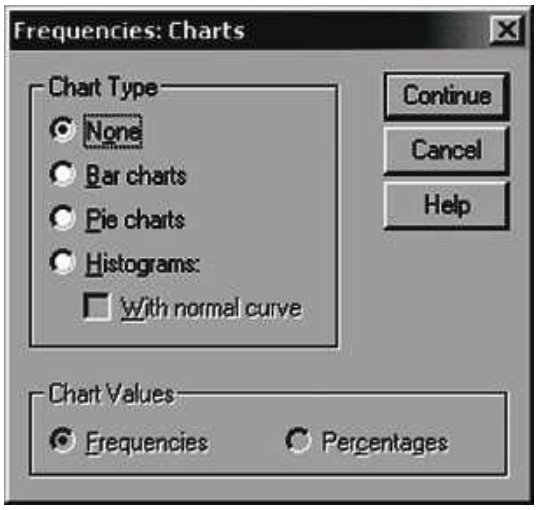


Рис. 2.17. Діалогове вікно Frequencies : Charts

– вибрати в групі Chart Type (Тип діаграми) пункт Bar charts (Стовпчаста діаграма), а в групі Chart Values (Значення діаграми) – пункт Percentages (Відсотки); підтвердити вибір кнопкою Continue (Продовжити), після чого повернутися в діалог Frequencies;

– у діалоговому вікні Frequencies зняти прапорець Display frequency tables (Показувати частотні таблиці); клацнути на кнопці ОК, діаграму буде показано у вікні перегляду (рис. 2.18).

Рис. 2.18. Стовпчаста діаграма в засобі перегляду

Можливе удосконалення вигляду цієї діаграми.

Щоб почати редагування, треба двічі клацнути в області стовпчастої діаграми, яку буде показано в редакторі діаграм.

На панелі інструментів редактора діаграм клацнути на символі міток стовпців. Відкриється діалогове вікно Bar Label Style (Стиль мі-

ток стовпців). Вибрати пункт Framed (У рамці), клацнути на кнопці Apply all (Застосувати для усіх) і потім на Close (Закрити). На кожному стовпці виникне напис з його відсотковим значенням.

Клацнути мишею на будь-якому із стовпців. На верхній стороні кожного стовпця виникне по два маленькі чорні квадрати. Це означає, що області стовпців є готовими для редагування.

Клацнути мишею на символі зразка заливки. Відкриється діалогове вікно Fill Patterns (Зразки заливки) (рис. 2.19).

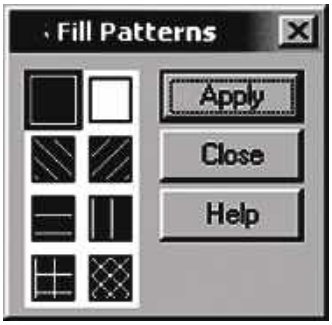


Рис. 2.19. Діалогове вікно Fill Patterns

Вибрати у вікні відповідний зразок заливки, підтвердити вибір кнопкою Apply (Застосувати) і закрити діалогове вікно.

Стовпці будуть заповнені вибраною заливкою.

Клацнути мишею на символі виду стовпців.

Вибрати пункт Drop shadow (Тінь), клацнути на кнопці Apply all (Застосувати для всіх) і потім на Close (Закрити).

Двічі клацнути на заголовку діаграми Fachbereich. Відкриється діалогове вікно Titles (рис. 2.20).

Рис. 2.20. Діалогове вікно Tirles

Змінити заголовок на потрібний і закрити діалог кнопкою ОК.

У меню Chart (Діаграма) установити прапорець Outer Frame (Зовнішня рамка). Закрити редактор діаграм; графік, що вийшов, показано на рис. 2.21.

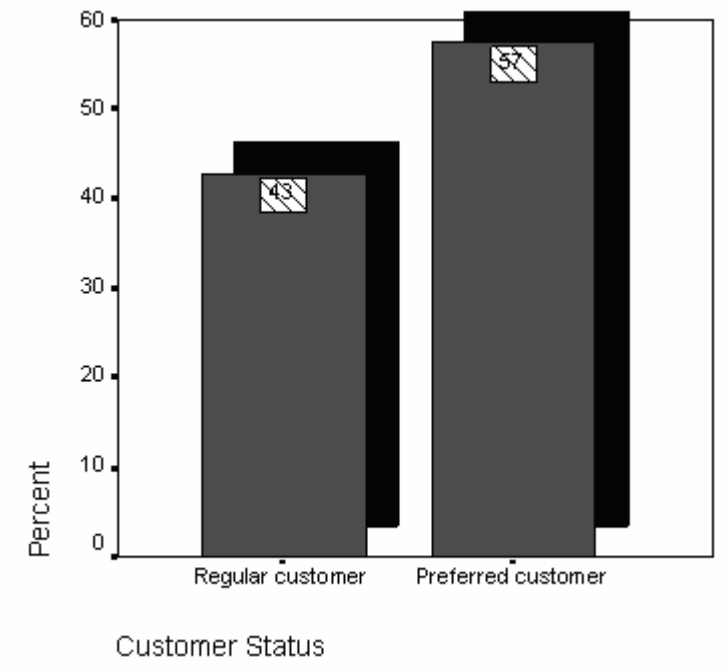


Рис. 2.21. Відредагована діаграма

Гістограми

Розглянемо візуальне подання частотного аналізу у вигляді гістограми. Процесс створеня гістограми схожий на процесс будування стовпчастих діаграм.

Вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти).

Клацнути на кнопці Reset (Скидання), щоб установити стандартні настройки.

Перенести змінну *f* в список вихідних змінних.

Клацнути на кнопці Charts… (Діаграми). У діалоговому вікні Frequencies: Charts вибрати пункт Histograms (Гістограма). Установити прапорець With normal curve (З кривою нормального розподілу), клацнути на кнопці Continue.

У діалоговому вікні Frequencies зняти прапорець Display frequency tables (Показувати частотні таблиці). Клацнути на кнопці ОК, гістограму буде показано у вікні перегляду (рис. 2.22).

Частоти на гістограмі позначено колонками, які на відміну від стовпчастої діаграми не ізольовані, а межують одна з одною. Відображуються також стандартне відхилення, середнє значення й загальна кількість спостережень *N*. Крім того, показано криву нормального розподілу.

Щоб відредагувати гістограму, треба двічі клацнути на області гістограми – відкриється редактор діаграм, в якому можна надати

гістограмі бажаного вигляду. Графік відобразиться в редакторі діаграм.

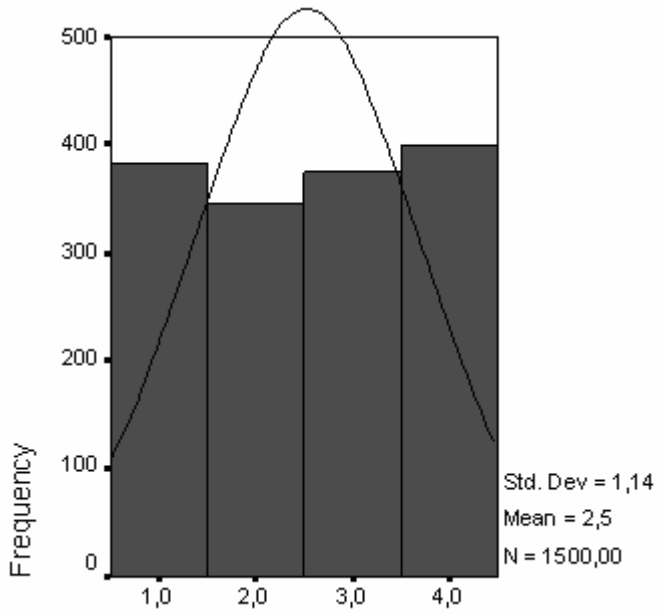


Рис. 2.22. Гістограма

2.5. Відбір даних

Відбір даних – це вибір спостережень за певними критеріями, наприклад, при опитуванні студентів – тільки студенток, що вивчають економіку й фінанси. Після цього всі обчислення проводитимуться тільки з цими відібраними спостереженнями.

Для цього в SPSS існує три принципові можливості:

– вибір спостережень за певною умовою (логічним виразом);

– витягання випадкової вибірки спостережень з файла даних;

– поділ спостережень на групи відповідно до значень однієї або декількох змінних.

2.5.1. Вибір спостережень

Проведемо частотний аналіз змінної з прикладу employment (заняття). При цьому будемо враховувати тільки респондентів-жінок. Для цього необхідно зробити таке:

– завантажити файл в Редактор даних;

– вибрати в меню команди Data (Дані) Select Cases (Вибрати спостереження). Відкриється діалогове вікно Select Cases (рис. 2.23). За замовчуванням у цьому діалозі вибрано пункт All cases (Усі спостереження);

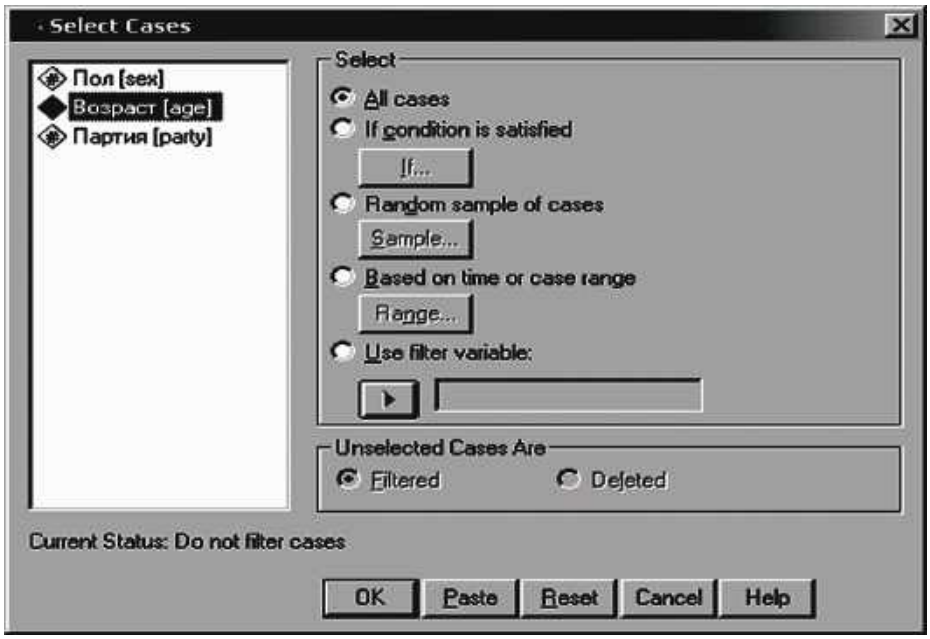


Рис. 2.23. Діалогове вікно Select Cases

– вибрати пункт If condition is satisfied (Якщо виконується умова) і клацнути на кнопці If ... (Якщо). Відкриється діалогове вікно Select Cases : If (рис. 2.24).

Рис. 2.24. Діалогове вікно Select Cases: If

Це діалогове вікно поділено на такі частини:

– список початкових змінних – містить змінні, такі, що містяться у відкритому файлі даних;

– редактор умов – тут записується логічний вираз, за яким мають бути відібрані спостереження; на цей момент редактор умов поки що є порожнім;

– кнопка з трикутником – дає можливість перенести змінну зі списку початкових змінних в редактор умов;

– клавіатура – містить цифри, а також арифметичні, логічні оператори й оператори відношення. З нею можна працювати як зі звичайним калькулятором. Якщо клацнути на якій-небудь кнопці мишею, відповідний знак, наприклад +, *, 7, буде скопійовано в редактор умов;

– список функцій – містить близько 140 функцій, кожну з яких можна скопіювати в редактор умов подвійним клацанням.

2.5.1.1. Класифікація операторів

Оператори поділяються на арифметичні, логічні й оператори відношення.

Арифметичні оператори застосовують в так званих арифметичних виразах (математичних формулах), які під час відбору даних мають лише другорядне значення. Арифметичні оператори завжди можна використовувати в логічних виразах, проте це зустрічається нечасто. Вирішальну роль ці оператори відіграють під час модифікації даних. Логічні оператори й оператори відношення застосовують виключно в логічних виразах.

Оператори відношення

Відношення – це логічний вираз, в якому два значення порівнюються один з одним за допомогою оператора відношення. В операторах відношення значення змінної порівнюються з яким-небудь числовим значенням (константою). Для будування логічних виразів застосовують оператори відношення (табл. 2.2).

Оператори можна ввести в редактор умов, клацнувши в діалоговому вікні на кнопці з відповідним знаком або ввівши з клавіатури альтернативний текст. Наприклад, замість можна ввести NE або <>.

Таблиця 2.2

Знак на кнопці	Альтернативний текст	Значення (укр./англ.)
<	LT	менше (less than)
>	GT	більше (greater than)
≤	LE	менше або дорівнює (less than or equal to)
≥	GE	більше або дорівнює (greater than or equal to)
=	EQ	дорівнює (equal to)
≠	NE или <>	не дорівнює (not equal to)

Логічні оператори

Для будування умовних виразів можна застосовувати логічні оператори, які наведено в табл. 2.3.

Таблиця 2.3

Знак на кнопці	Альтернативний текст	Значення	Пріоритет
&	AND	Логічне І	2
|	OR	Логічне АБО	3
~	NOT	Логічне НІ	1

Логічні оператори AND і OR зв'язують два відношення, а логічний оператор NOT змінює значення істинності умовного виразу на протилежне.

2.5.1.2. Логічні й строкові функції

Важлива частина діалогового вікна Select Cases: If – це список функцій, який містить безліч математичних функцій, велика частина з яких, проте, має відношення тільки до модифікації даних (розрахунку нових змінних). Розглянемо логічні й строкові функції.

Логічні функції

У SPSS реалізовано дві логічні функції:

1. Функція RANGE (variable, begin, end) – повертає значення 1, або true, якщо значення змінної лежить в діапазоні між заданими початковим і кінцевим значеннями. Змінна може мати як числовий, так і строковий тип. RANGE (alter, 18, 22) повертає значення 1, тобто true, якщо значення змінної alter лежить між 18 і 22 включно. Можна задавати декілька діапазонів, наприклад, RANGE (alter, 1,17, 63, 99). У цьому випадку функція повертає true, якщо значення змінної alter лежить між 1 або 17 або між 63 і 99 включно. У функції RANGE можна також використовувати змінні строкового типу, наприклад, RANGE (name, A, Mzzzzzz). Тоді функція повертатиме 1 для імен, що починаються з букв від A до M включно. Якщо ім'я починається з іншої букви, функція поверне 0.

2. Функція ANY (variable, val1, val2, val3,..) – повертає значення 1, або true, якщо значення змінної (значення першого аргументу) збігається принаймні з одним зі значень, вказаних в наступному списку параметрів (val1, val2, val3, ..), інакше повертається значення 0, або false. Перший елемент, як правило, – це змінна чисельного або символьного типу. Приклади: ANY (jahr, 1991, 1992, 1993, 1994) повертає true, якщо значення змінної jahr дорівнює 1991, 1992, 1993 або 1994. ANY (name, Lena, Piter, Jon) повертає значення true, або 1, у тих випадках, коли змінна name містить значення Lena, Piter або Jon. В усіх інших випадках повертається значення 0.

Строкові функції

Із загальної кількості 18 строкових функцій розглянемо три найважливіших.

1. Функція SUBSTR (variable, begin, length) витягає певну частину з рядка, повертає підрядок або окремий символ. Наприклад, якщо строкова змінна name містить значення Piter, то наступний виклик функції

SUBSTR (name, 1, 2)

поверне значення Pi. Тут зі змінної name витягаються два знаки (третій аргумент) починаючи з першої позиції (другий аргумент).

Вираз

SUBSTR (name, 1, 2) = Ma

буде істинним для значень змінної Maus або Marshall. Порівнюючи з рядками, замість подвійних лапок (= "Ma") можна також застосовувати прості (= 'Ma').

2. Функція UPCASE (argument) перетворить малі букви на великі. Як аргумент можна задавати рядок або змінну символьного типу. UPCASE (name) повертає значення ANNA, якщо змінна name має значення Anna.

3. Функція LOWER (argument) перетворить великі букви на малі.

Як параметр можна задавати рядок або змінну символьного типу.

Наприклад, LOWER (name) повертає значення anna, якщо змінна name має значення ANNA або Anna.

Функції переносяться в редактор умов таким чином:

– помістити курсор на місце в умовному виразі, на якому має бути функція;

– двічі клацнути на функції в списку функцій або виділити функцію і клацнути на кнопці з трикутником біля списку функцій.

Функцію буде вставлено у вираз. Замість аргументів в цій функції стоятимуть знаки запитання. Кількість знаків запитання означає мінімальну кількість аргументів, яку слід вставити. Відредагувати функцію можна таким чином:

– виділити знаки запитання у вставленій функції;

– замінити їх відповідними аргументами, імена змінних для аргументів можна перенести зі списку початкових змінних.

При будуванні логічних виразів необхідно дотримуватися пріоритетів (табл. 2.4).

Таблиця 2.4

Пріоритет	Оператор/функція	Значення
1	0	Оператор скобок
2	Функції	Різні значення
3	<	Менше
	≤	Менше або дорівнює
	>	Більше
	≥	Більше або дорівнює
	=	Дорівнює
	≠	Не дорівнює
4	—	Логічне НІ
5	^	Логічне І
6	I	Логічне АБО

2.5.1.3. Уведення умовного виразу

Для введення умовного виразу необхідно виконати такі дії:

– перенести змінну в редактор умов, двічі клацнувши на ній або виділивши її і клацнувши на кнопці з трикутником;

– увести умовний вираз для змінної (вигляд діалогового вікна показано на рис. 2.25).

Рис. 2.25. Умова в редакторі умов

– підтвердити вибір кнопкою Continue (Продовжити), це дасть можливість повернутися в діалог Select Cases (проте тепер в діалоговому вікні виникла умова);

– клацнути на кнопці ОК, внаслідок чого можна знову опинитися в редакторі даних.

Примітка. Вибрані опції відповідають такому командному синтаксису:

SELECT IF умовний вираз
EXECUTE .

Тепер фільтрацію спостережень увімкнено. Про те, що відбір, заданий за допомогою діалогових вікон, є здійсненим, свідчить повідомлення Filter on (Фільтр включено), яке виникає в рядку стану в нижній частині вікна SPSS. Система створює змінну filter_S. Це – числова змінна з довжиною один байт. Вона має такі мітки значень: 0 = Not Selected (Не вибрано), 1 = Selected (Вибрано), оскільки нуль означає хибність (false), а одиниця – істину (true). В усіх наступних операціях враховуватимуться тільки спостереження, для яких значення цієї змінної дорівнює одиниці, тобто ті, для яких виконується вибрана умова. Зверніть увагу, що фільтр діє й при інших статистичних процедурах. Команда SELECT IF або відповідні налаштування в діалогових вікнах фільтрують спостереження постійно, тобто доти, доки фільтр не буде видалено або деактивовано. Щоб видалити фільтр, треба зробити таке:

– клацнути на імені змінної filter_$, увесь стовпець буде виділено;

– натиснути клавішу <Backspace>, змінну фільтра буде видалено.

Якщо не видаляти фільтр, а лише тимчасово деактівизувати його, необхідно виконати такі дії:

– вибрати в меню команди Data (Дані) Select Cases ... (Вибрати спостереження);

– у діалоговому вікні Select Cases клацнути на кнопці All cases (Усі спостереження), умову фільтра буде деактивовано, проте змінна filter_S збережеться; у будь-який момент її можна буде активізувати знову.

Тимчасовий фільтр можна ввести тільки вручну в редакторі синтаксису SPSS (через діалогові вікна цього зробити неможливо).

2.5.2. Витягання випадкової вибірки

При великій кількості спостережень для заощадження часу може бути корисним використовувати невелику випадкову вибірку. Щоб витягнути випадкову вибірку із сукупності всіх спостережень, необхідно виконати такі дії:

– вибрати в меню команди Data (Дані) Select Cases (Вибрати спостереження);

вибрати пункт Random sample of cases (Випадкова вибірка), а потім клацнути на кнопці Sample... (Вибірка). Відкриється діалогове вікно Select Cases: Random Sample (Вибрати спостереження: Випадкова вибірка) (рис. 2.26).

У групі Sample Size (Розмір вибірки) можна вибрати один з таких способів визначення об'єму вибірки:

– Approximately (Приблизно) – користувач може вказати тут відсоткове значення. SPSS створить випадкову вибірку з обсягом, що при-

близно відповідає вказаному відсотку спостережень;

– Exactly (Точно) – користувач має вказати тут точну кількість спостережень у випадковій вибірці. Крім того, потрібно задати кількість спостережень, з яких буде витягнуто вибірку. Друге число не має перевищувати загальної кількості спостережень у файлі даних. Для кожної випадкової вибірки генератор випадкових чисел SPSS використовує нове початкове значення. Таким чином, кожного разу при зверненні до цього діалогу створюється нова вибірка спостережень, відмінна від попередніх.

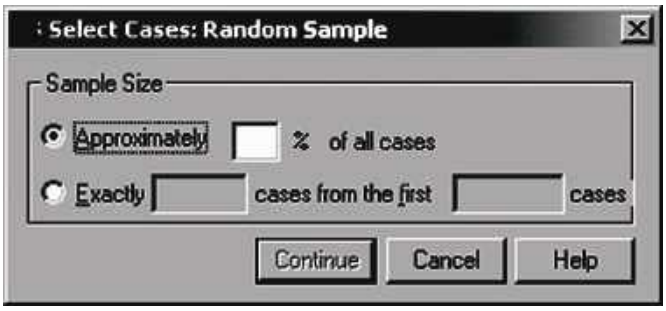


Рис. 2.26. Діалогове вікно Select Cases: Random Sample

Якщо необхідно, щоб випадкова вибірка повторювалася, потрібно задати початкове значення самостійно. Для цього виберіть в меню команди Transform (Перетворити) Random Number Seed... (Встановити початкове положення генератора випадкових чисел). Відкриється діалогове вікно Random Number Seed (рис. 2.27).

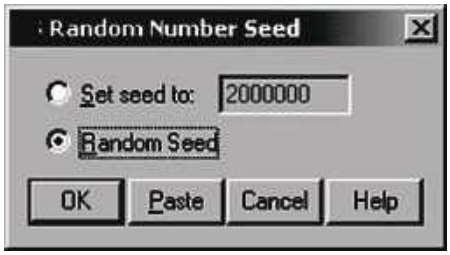


Рис. 2.27. Діалогове вікно Random Number Seed

Початкове значення може бути будь-яким додатним цілим числом. Це значення можна задати самостійно або дати зробити це SPSS (варіант Random Seed, прийнятий за замовчуванням).

2.5.3. Сортування спостережень

Дані в SPSS можна сортувати відповідно до значень однієї або декількох змінних. Розглянемо такий приклад: необхідно упорядкувати дані файла за віком.

Для цього необхідно зробити таке:

– вибрати в меню команди Data (Дані) Sort Cases.. (Сортувати спостереження). Відкриється діалогове вікно Sort Cases (рис. 2.28). Змінні файла даних будуть відображені в списку початкових змінних;

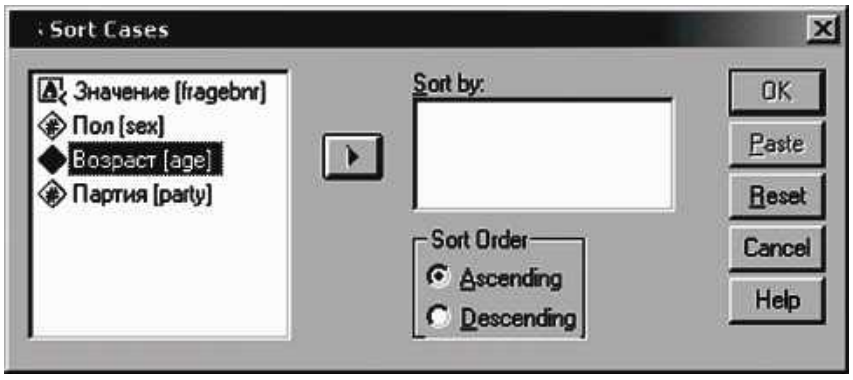


Рис. 2.28. Діалогове вікно Sort Cases

– перенести змінну, за значеннями якої треба зробити сортування, в список Sort by (Сортувати по). У групі Sort order (Порядок сортування) за замовчуванням вибрати варіант Ascending (За збільшенням). Ця опція сортує спостереження в порядку зростання значення змінного сортування, а наступна опція Descending – у порядку убування;

– підтвердити настроювання кнопкою ОК. У редакторі даних файл даних буде відсортовано за збільшенням значень вибраної змінної.

Примітка. Вибрані опції відповідають такому командному синтаксису:

SORT CASES BY змінна (A)

або, якщо потрібно сортувати за убуванням:

SORT CASES BY змінна (D).

Тут A означає ascending (зростання), а D – descending (убування).

Якщо вибрати декілька змінних сортування, їх послідовність в списку Sort by визначатиме порядок, в якому будуть відсортовані спостереження. Нехай необхідно відсортувати файл за значеннями двох змінних *var1* і *var2*. Змінна *var1* має бути першим критерієм сортування, а змінна *var2* – другим. Сортування за змінною *var1* має бути в порядку зростання, а за змінною *var2* – в порядку убування. Для цього перенести в список змінних сортування спочатку змінну *var1*, а потім – *var2*. Виділити змінну *var1* і клацнути на опції Ascending.

2.5.4. Поділ спостережень на групи

У SPSS можна виконувати аналіз даних окремо по групах. Групою в цьому контексті називають певну кількість спостережень з однаковими значеннями ознак. Щоб можна було здійснювати оброблення по групах, файл має бути відсортований за групувальними змінними. Такою змінною може бути, наприклад, змінна, яка містить ознаку

«стать». У цьому випадку всі змінні зі значенням ознаки 1 (жіночий) утворюють одну групу, а усі змінні зі значенням ознаки 2 (чоловічий) – другу групу. З кожною групою можна проводити певні операції, наприклад, виконувати частотний аналіз. При цьому частотний аналіз проводиться окремо для ознак «чоловічий» і «жіночий». У SPSS такий поділ на групи можна виконувати автоматично. Для цього необхідно зробити таке:

– завантажити файл в редактор даних;

– вибрати в меню команди Data (Дані) Split File (Розділити файл). Відкриється діалогове вікно Split File (рис. 2.29).

За замовчуванням розподіл на групи не передбачається. Якщо вибрати пункт Organize output by groups (Поділити виведення на групи), отримаємо виведення результатів по кожній групі окремо. Ці групи мають бути визначені в полі Groups based on (Групи, створені на основі) на базі відповідних змінних.

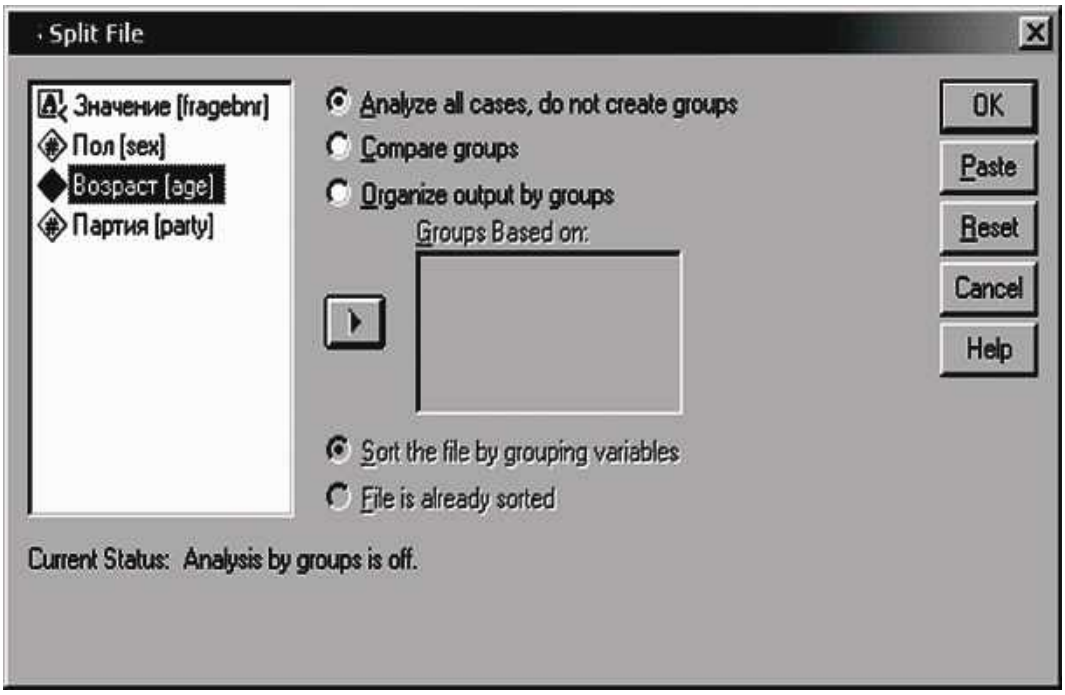


Рис. 2.29. Діалогове вікно Split File

Ще одну можливість надає опція Compare Groups (Порівняти групи). Вона організовує виведення таким чином, що можна візуально порівняти різні групи одну з одною. Але спочатку розглянемо роздільне виведення:

– вибрати опцію Organize output by groups; для роздільного виконання операцій по групах необхідно, щоб файл даних був заздалегідь відсортований за цими групувальними змінними; з цієї причини опцію Sort the file by grouping variables (Сортувати файл за групувальними змінними) вибрано за замовчуванням;

– перенести групувальну змінну *var1* в поле Groups based on; якщо вибирається декілька групувальних змінних, то послідовність, в якій вони стоять в списку, визначає порядок або пріоритет сортування;

– клацнути на кнопці ОК, файл даних буде відсортовано за змінною *var1*, тобто розбито на групи відповідно до її значень; повідомлення File split on (Розподіл файла) включено в рядок стану внизу вікна, SPSS інформує про активацію режиму розподілу.

Урахуйте, що файл даних залишиться поділеним на підгрупи, доки не буде деактивовано відповідні опції. Для цього потрібно зробити таке:

– вибрати в меню команди Data (Дані) Split File (Поділити файл);

– в діалоговому вікні Split File вибрати опцію Analyze all cases, do not create groups (Аналізувати усі спостереження, не створювати групу). Тепер розподіл файла прибрано.

2.6. Модифікація даних

Для проведення аналізу часто буває необхідним виконати перетворення даних. На основі спочатку зібраних даних можна створити нові змінні й змінити кодування. Такі перетворення називають модифікацією даних.

У SPSS існує багато можливостей для модифікації даних. До найважливішого з них належать:

– обчислення нових змінних шляхом використання різних арифметичних виразів (математичних формул);

– підрахунок частоти появи певних значень;

– перекодування значень;

– обчислення нових змінних під час виконання певної умови;

– агрегування даних;

– рангові перетворення;

– обчислення ваг спостережень.

2.6.1. Обчислення нових змінних

Шляхом обчислень у SPSS можна утворити нові змінні й додати їх у файл даних. Для цього треба зробити таке:

1. Завантажити файл у редактор даних.

2. Вибрати в меню команди Transform (Перетворити), Compute (Обчислити). Відкриється діалогове вікно Compute Variable (Обчислити змінну) (рис. 2.30);

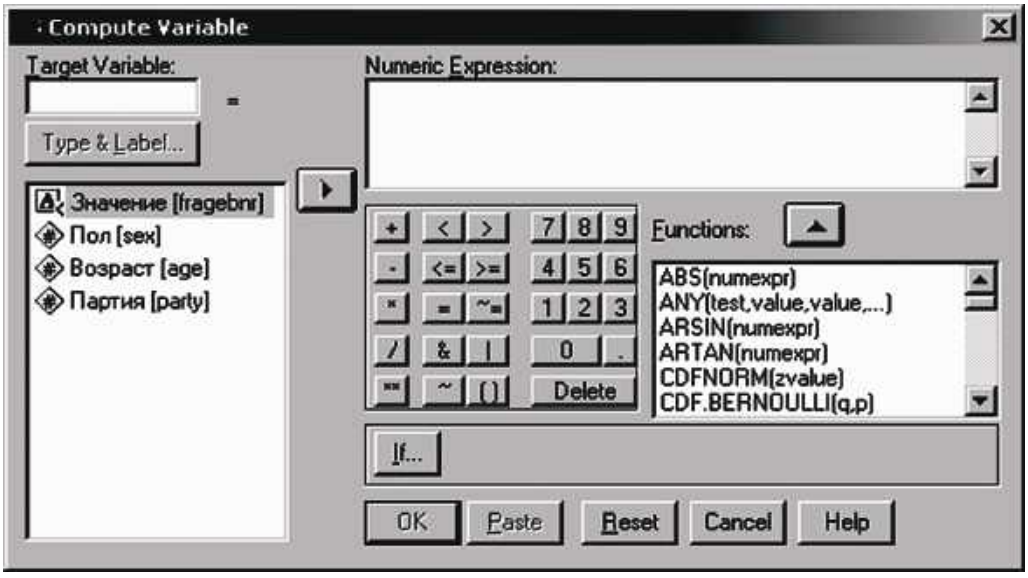


Рис. 2.30. Діалогове вікно Compute Variable

У полі Target Variable (Вихідна змінна) вказується ім'я змінної, якій присвоюється обчислене значення. Як вихідна змінна може бути вже існуюча або нова змінна. У поле Numeric Expression (Числовий вираз) уводиться вираз, який застосується для визначення вихідної змінної. У цьому виразі можуть використовуватися імена існуючих змінних, константи, арифметичні оператори й функції. Формулу можна ввести або вручну, або використовуючи список змінних і клавіатуру діалогового вікна. Кнопка із трикутником дає можливість копіювати в поле формули імена змінних, а кнопки клавіатури – вставляти цифри й знаки.

3. Клацнути на кнопці Type&Label... (Тип і мітка); відкриється діалогове вікно Compute Variable: Type and Label (Обчислити змінну: Тип і мітка) (рис. 2.31). Тут можна задати мітку для нової змінної. У поле Label увести текст коментарю для нової змінної, клацнути на кнопці Continue.

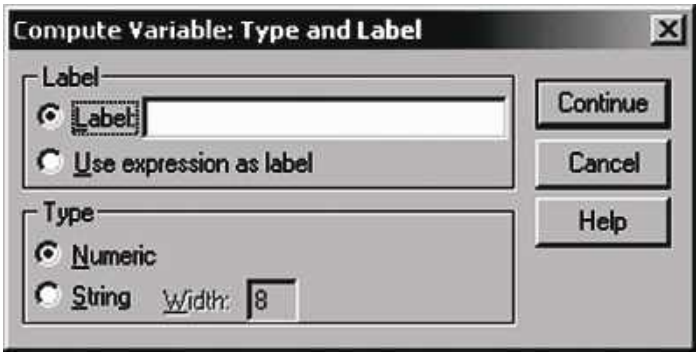


Рис. 2.31. Діалогове вікно Compute Variable: Type and Label

4. У діалоговому вікні Compute Variable клацнути на кнопці ОК.

Команда EXECUTE зчитує дані й виконує попередні команди перетворення. У файл даних додається нова змінна. Тепер її, як і інші

змінні, можна застосовувати для обчислень. Для SPSS немає різниці, введено значення змінних через редактор даних чи обчислено їх за формулою.

Замість слова «формула» будемо використовувати надалі поняття «числовий вираз». При формулюванні таких числових виразів потрібно дотримуватися певних правил, які наведено ніжче.

2.6.1.1. Формулювання числових виразів

Для будування числових виразів можна застосовувати п’ять арифметичних операторів (табл. 2.5).

За допомогою арифметичних операторів у числових (арифметичних) виразах можна задавати такі основні дії, як додавання й вирахування.

Таблиця 2.5

Знак	Арифметична дія
+	Додавання
–	Віднімання
*	Множення
/	Ділення
**	Піднесення до степеня

Структура виразів може бути складною, тому слід враховувати пріоритети арифметичних операторів.

В арифметичних виразах можуть брати участь змінні, константи й функції.

2.6.1.2. Функції

Функції можна поділити на такі класи:

– арифметичні функції;
– статистичні функції;
– функції дати й часу;
– функції оброблення відсутніх значень;
– функції добування значень спостережень;
– статистичні функції розподілу;
– функції генерації випадкових чисел.

Параметрами функцій можуть бути змінні, константи або вирази. Параметри беруть у круглі дужки; кілька параметрів відділяються один від одного комами, наприклад, SUM (5, 8, 10). Функція SUM обчислює суму трьох параметрів і повертає значення 23.

Арифметичні функції:

– ABS (x) – повертає абсолютне значення;

– RND (x) – округляє до найближчого цілого числа. Якщо змінна x має значення 3,6, то RND (x) повертає 4;

– TRUNC (x) – відкидає дробову частину значення; округлення не відбувається. Якщо змінна x має значення 3,9, TRUNC (x) повертає 3;

– MOD (x, m) – повертає остачу від ділення першого аргументу (x) на другий (m). Якщо змінна *x* має значення 1994, MOD (x, 100) повертає 94;

– SQRT (x) – повертає квадратний корінь. Якщо змінна *x* має значення 9, SQRT (x) повертає значення 3;

– EXP (x) – показова функція;

– LG10 (x) – десятковий логарифм;

– LN (x) – натуральний логарифм;

– ARSIN (x) – арксинус;

– ARTAN (x) – арктангенс;

– SIN (x) – синус;

– COS (x) – косинус.

У тригонометричних функціях аргументи задаються в радіанах.

Статистичні функції:

– SUM (x1, x2,...) – повертає суму значень допустимих аргументів. SUM (x1, x2, x3) повертає суму значень трьох змінних;

– MEAN (x1, x2,...) – повертає середнє арифметичне допустимих аргументів. MEAN (42, 19, 29) повертає значення 30;

– SD (x1, x2,...) – повертає стандартне відхилення значень допустимих аргументів;

– VARIANCE (x1, x1,...) – повертає дисперсію значень допустимих аргументів;

– CFVAR (x1, x1,...) – повертає коефіцієнт варіації значень допустимих аргументів;

– MIN (x1, x1,...) – повертає найменше зі значень допустимих аргументів;

– MAX (x1, x1,...) – повертає найбільше зі значень допустимих аргументів.

Статистичні функції можуть мати будь-яку кількість параметрів. Функції SUM, MEAN, MIN і MAX потребують хоча б одного допустимого аргументу, функціям SD, VARIANCE і CFVAR – двох. Інші аргументи можуть містити відсутні значення. Якщо цю властивість, прийняту за замовчуванням, потрібно деактивувати, то до імені функції через крапку додають кількість необхідних аргументів, наприклад MEAN.10. У цьому випадку значення функції обчислюється тільки тоді, коли існує хоча б зазначена кількість аргументів (у цьому прикладі 10).

Функції дати й часу. У SPSS часто в різних цілях використовують дату й час. Для введення даних цього типу в редакторі даних SPSS надає декілька різних форматів, які можна переглянути в діалоговому вікні Variable Type (Тип змінної).

Рекомендовано використовувати загальноприйнятий формат дати: зазначення числа місяця двома цифрами, місяця – також двома цифрами й року – чотирма цифрами через крапку: dd.mm.yyyy.

Економії місця за рахунок відкидання двох перших цифр року останнім часом, як відомо, приділяється багато уваги. При зазначенні року двома цифрами як столітній діапазон у SPSS прийнято термін з 1931 по 2030 р., отже, рік 28 інтерпретується як 2028, а 32 – як 1932. У меню Edit (Виправлення) Options... (Параметри...) на вкладці Data (Дані) користувач може самостійно задати столітній діапазон.

Якщо число або місяць можна записати однією цифрою, їх не потрібно доповнювати спереду нулями. Таким чином, зазначення дати в таких форматах буде допустимим: 20.6.1998; 13.12.1887; 5.2.1997.

Комп'ютер помічає суперечливість у зазначенні дати при введенні. Наприклад, дату 29.2.1997 не буде прийнято в комірку.

Для часу рекомендовано формат hh:mm:ss, тобто одна або дві цифри для годин, хвилин і секунд через двокрапку. За відсутності секунд можна також застосовувати формат hh:mm.

Приклади: 23:34:55; 8:5:12; 12:17:5; 12:47 8:12.

Дату й час, уведені в будь-якому вигляді, SPSS перетворить у внутрішній формат. Для дати це – секунди що пройшли з 0 годин 15.10 1582 р. (моменту введення григоріанського календаря) до 0 годин заданого дня; для часу – кількість секунд з 0 годин до заданого моменту часу.

У принципі можна також зберігати число, місяць, рік, години, хвилини й секунди в окремих змінних і визначати дату або час у внутрішньому форматі за допомогою відповідних функцій.

Усього в SPSS є 25 різних функцій для роботи з датою й часом, найважливіші з яких наведено в табл. 2.6.

Таблиця 2.6

Функції	Дія функції
XDATE.MDAY(arg)	Виділяє з дати число
XDATE.MONTH(arg)	Виділяє з дати місяць
XDATE.YEAR(arg)	Виділяє з дати рік
XDATE.WKDAY(arg)	Номер дня тижня (1 =' неділя', ..., 7 = 'субота')
XDATE.JDAY(arg)	Номер дня в році
XDATE.QUARTER(arg)	Номер кварталу в році
XDATE.WEEK(arg)	Номер тижня в році
XDATE.TDAY(arg)	Кількість днів, починаючи з 15.10.1582
XDATE.DATE(arg)	Кількість секунд, починаючи з 15.10.1582
DATE.DMY(d,m,y)	Перетворює дані числа місяця, місяця й року у внутрішню дату

Закінчення табл. 2.6

Функції	Дія функції
DATE.MOYR(m.y)	Перетворює дані місяця й року у внутрішню дату
YRMODA(y,m,d)	Перетворює дані року, місяця й числа місяця (строго в наведеній послідовності) у кількість днів, починаючи з 15.10.1582
XDATE.TIME(arg)	Кількість секунд, починаючи з 0 годин
TIME.HMS(h,m,s)	Перетворює дані годин, хвилин і секунд у секунди

Функції дати й часу застосовують найчастіше у випадках, коли потрібно обчислити проміжок між двома датами або моментами часу. Наприклад, якщо є дві дати, записані в змінних *datum1* і *datum2*, тривалість проміжку між ними в днях можна розрахувати за такою формулою:

```
COMPUTE tage=XDATE.TDAY(datum2) – XDATE.TDAY(datum2).
EXECUTE.
```

Функції оброблення пропущених значень:

– VALUE (variable) – оголошує недійсним користувацьке пропущене значення;

– MISSING (variable) – повертає значення 1 (або true), якщо змінна містить користувацьке або системне пропущене значення;

– SYSMIS (variable) – повертає значення 1 (або true), якщо змінна містить системне пропущене значення;

– NMISS (variable,variable,...) – повертає кількість пропущених значень у списку змінних;

– NVALID (variable,variable,...) – повертає кількість допустимих значень у списку змінних.

Функція добування значень спостережень LAG (variable,n) – повертає значення відповідної змінної за *n* спостережень до поточного. Так, наприклад, LAG (variable, 1) дає можливість одержати значення змінної в попередньому випадку.

Статистичні функції розподілу. У SPSS реалізовано в сукупності 20 статистичних функцій роз оділу. Ці функції обчислюють значення ймовірності для таких розподілів: ß-розподіл, розподіл Коші, хі-квадрат, експонентний розподіл, Г-розподіл, F-розподіл, розподіл Лапласа, логістичний, логарифмічно нормальний, нормальний розподіл, розподіл Парето, розподіл Стьюдента, рівномірний розподіл, розподіл Вейбулла (неперервні функції), а також розподіл Бернуллі, біноміальний, геометричний, гіпергеометричний, негативно-біноміальний розподіли і розподіл Пуассона (дискретні функції). Для 14 неперервних функцій розподілу існують відповідні обернені функції.

Так, наприклад, функція CDF.T(t,df) повертає ймовірність помилки *p* для заданого значення *t* функції розподілу Стьюдента з числом сту-

пенів свободи *df*. Функція IDF. T(p,df) повертає значення *t* для заданої ймовірності помилки *p* і числа ступенів свободи *df*.

Функції генерації випадкових чисел. У SPSS реалізовано в сукупності 24 функції генерації випадкових чисел, у тому числі для 20 вбудованих статистичних функцій розподілу; наприклад, функція RV.T(df) повертає випадкові числа, що підпорядковуються розподілу Стьюдента при *df* ступенях свободи. Функція UNIFORM (x) генерує рівномірно розподілені випадкові величини, що знаходяться в інтервалі від 0 до 1, а її аргумент задає початкове значення для генератора випадкових чисел.

2.6.2. Підрахунок частоти появи певних значень

У SPSS є можливість підрахувати кількість появи того самого значення або значень для певної змінної.

Для цього необхідно:

– завантажити файл з даними в редактор даних;

– вибрати в меню команди Transform (Перетворити) Count... (Підрахувати); відкриється діалогове вікно Count Occurences of Values within Cases (Підрахувати кількість значень у спостереженнях) (рис. 2.32).

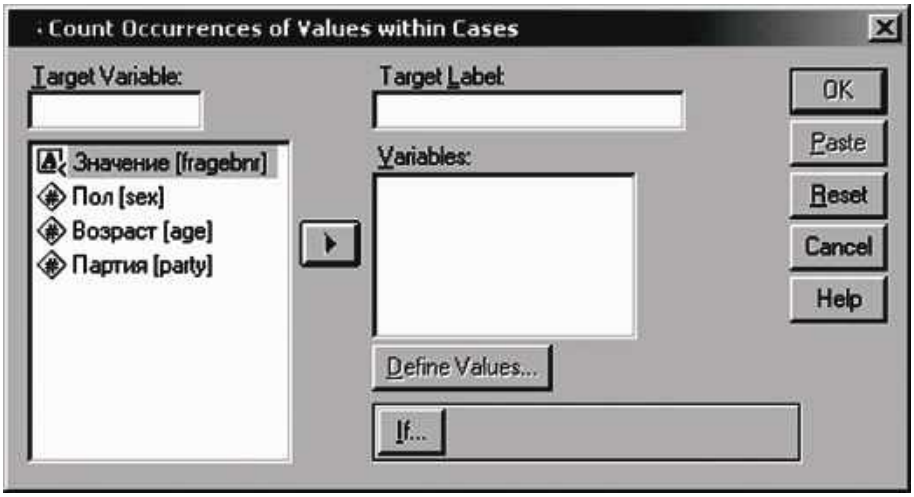


Рис. 2.32. Діалогове вікно Count Occurences of Values wirhin Cases

Це діалогове вікно поділено на такі частини:

– Target variable (Вихідна змінна) – у полі Target variable вказується ім'я змінної, у якій будуть утримуватися підраховані значення;

– Target Label (Мітка) – у полі Target Label вказується мітка для вихідної змінної;

– Variables (Змінні) – цей список містить змінні, вибрані зі списку вихідних змінних, що зберігаються у файлі даних, для яких потрібно під-

рахувати певні значення. Список не може одночасно містити числові й строкові змінні.

Виділити в списку вихідних змінних такі, для яких треба підраховувати кількість однакових значень.

Присвоїти вихідній змінній ім'я й мітку – коментар.

Клацнути на кнопці Define values... (Визначити значення). Відкриється діалогове вікно Count Values within Cases: Values to Count (Підрахувати значення в спостереженнях: які значення?) (рис. 2.33)

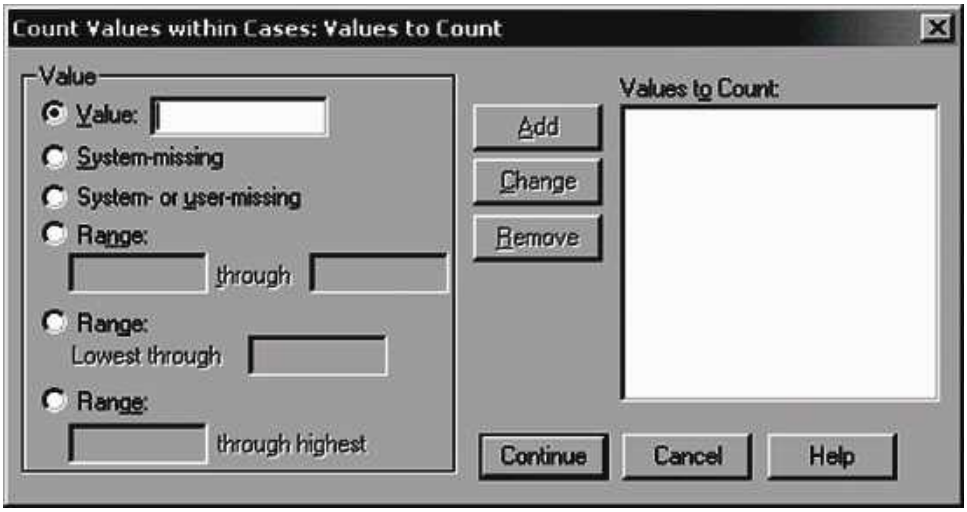


Рис. 2.33. Діалогове вікно Count Values within Cases:values to Count

Це діалогове вікно призначене для знаходження підраховуваних значень. Можна задати окреме значення, діапазон або поєднання того й іншого. У групі Value (Значення) можна вибрати один з таких варіантів:

– Value – уводиться окреме значення, частоту якого необхідно підрахувати;

– System missing (Системне пропущене) – підраховується кількість появ системного пропущеного значення; у списку Values to count (Підраховувані значення) воно відображається як SYSMIS; для строкових змінних цей варіант не застосовується;

– System or user-missing (Користувацькі або системні пропущені) – якщо вибрати цей варіант, буде підраховано кількість появ усіх пропущених значень, як системних, так і користувальницьких; у списку Values to count ці значення відображаються як MISSING;

– Range through (Діапазон) – підраховується кількість значень, що знаходяться в певному діапазоні; цей варіант також не застосовується для строкових змінних;

– Range: Lowest through (Діапазон: від найменшого до) – підраховується кількість значень, що знаходяться в діапазоні від найменшого спостережуваного до зазначеного; цей варіант не застосовується для

строкових змінних;

– Range: through highest (Діапазон: до найбільшого) – підраховується кількість значень, що знаходяться в діапазоні від зазначеного до найбільшого спостережуваного; цей варіант не застосовується для строкових змінних.

Якщо потрібно підрахувати повторюваність декількох значень, то треба клацнути після вибору опції на кнопці Add (Додати). У цьому випадку буде підраховано частоту повторень кожного значення, що є у списку Values to count.

Задати окреме значення 1 і клацнути на кнопці Add.

Підтвердити введення кнопкою Continue, а потім – ОК. У файл даних буде додано змінну результату, що містить кількість однакових значень.

2.6.3. Перекодування значень

Спочатку зібрані дані можна перекодувати за допомогою засобів SPSS. Перекодування числових даних необхідно, наприклад, тоді, коли початкове різноманіття вихідних даних не потрібне для наступного аналізу. У цьому випадку перекодування означає зменшення обсягу оброблюваної інформації. Перекодування даних можна виконати вручну або автоматично.

2.6.3.1. Ручне перекодування

Перекодування треба виконувати таким чином:

– завантажити файл з даними в редактор даних;

– вибрати в меню команди Transform (Перетворити) Recode (Перекодувати). Можна зберігати перекодовані значення в тій самій змінній або перенести їх в іншу (рис. 2.34). Якщо буде проведено перекодування в попередній змінній, усі її попередні значення буде стерто;

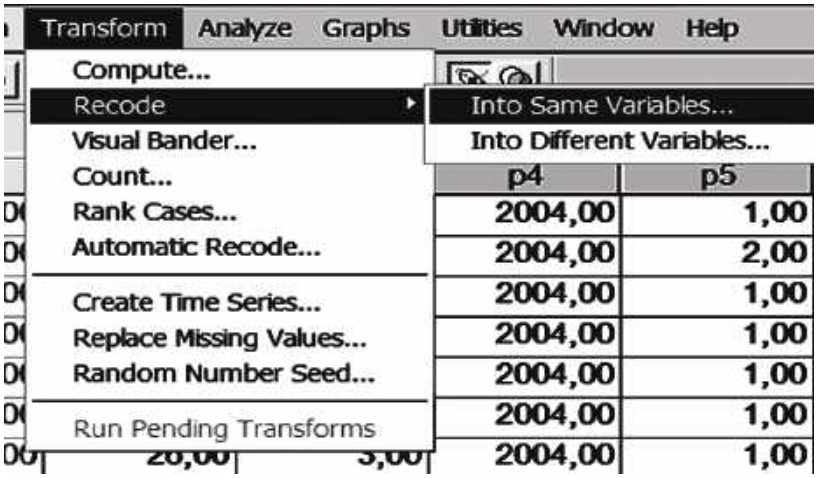


Рис. 2.34. Список команд Transform (Перетворити) Recode (Перекодувати)

– вибрати у підменю пункт Into Different Variables... (В інші змінні). Відкриється діалогове вікно Recede into Different Variables (Перекодувати в інші змінні) (рис. 2.35).

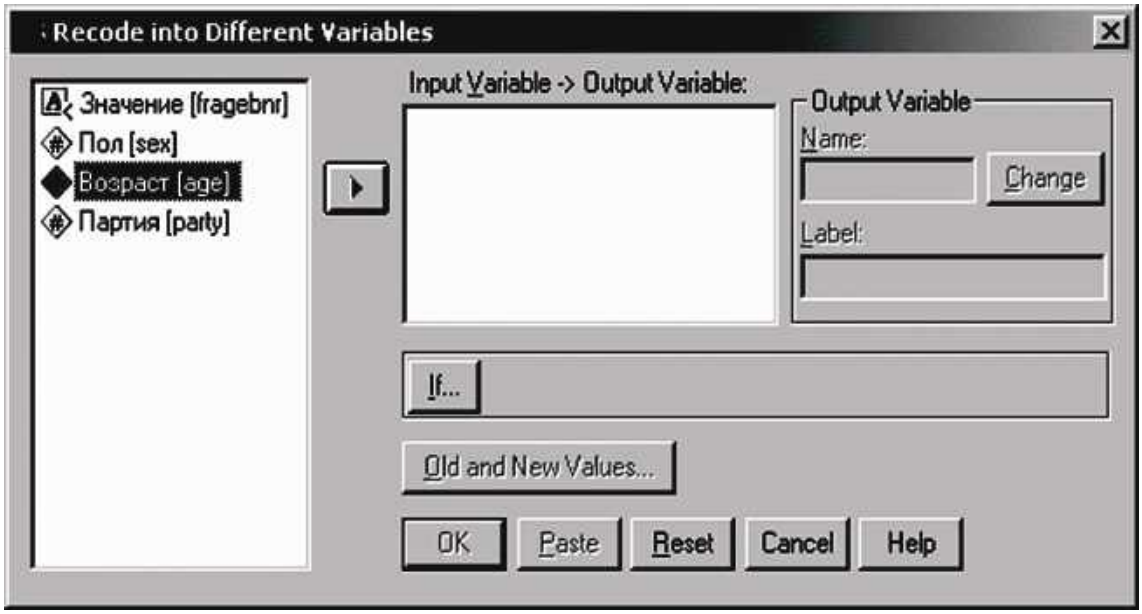


Рис. 2.35. Діалогове вікно Recede into Different Variables

Список вихідних змінних містить змінні файла даних. Тут можна вибрати одну або декілька змінних для перекодування. Якщо вибираються декілька змінних, усі вони мають бути одного типу.

Перенести змінну *x* у поле Input Variable > Output Variable (Вхідна змінна > Вихідна змінна). Знак запитання, доданий у поле, свідчить про те, що треба задати ім'я вихідної змінної.

Увести в поле Name (Ім'я) текст *x1*. Клацнути на кнопці Change (Змінити). Знак запитання в полі Input Variable > Output Variable буде замінено на *x1*.

Увести в поле Label позначення: «Коментар». Підтвердити введення, клацнувши на Change.

Щоб установити значення, які варто перекодувати, треба клацнути на кнопці Old and New Values (Старі й нові значення). Відкриється діалогове вікно Recede into Different Variables: Old and New Values.

Для здійснення кожного перекодування треба вказати значення або діапазон вхідної змінної й відповідне значення вихідної змінної. Перекодування завершується клацанням на кнопці Add.

Це діалогове вікно (рис. 2.36) поділено на декілька частин.

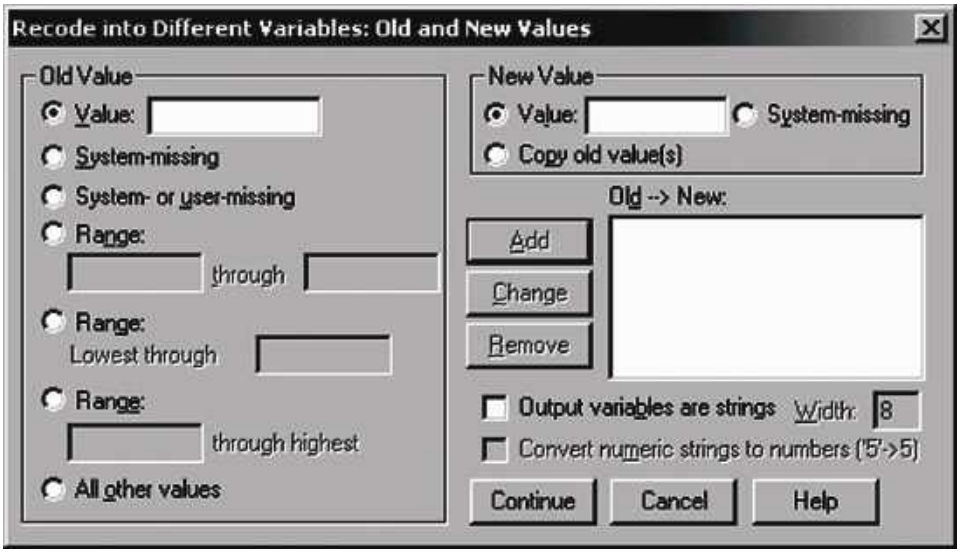


Рис. 2.36. Діалогове вікно Recede into Different Variables: Old and New Values

У групі Old Value (Попереднє значення) можна вибрати один з таких варіантів:

– Value – уводиться окреме значення;

– System missing (Системне пропущене значення) – за допомогою цієї опції значення вхідної змінної позначається як системне пропущене. Його позначено в списку значень змінних як SYSMIS; такий варіант не застосовується для строкових змінних;

– System or user-missing (Користувацьке або системне пропущене значення) – цю опцію призначено для позначення всіх користувацьких або системних пропущених значень; у списку значень змінних користувальницькі пропущені значення відображаються як MISSING;

– Range through (Дипазон) – тут можна задати замкнутий інтервал значень, який не застосовується для строкових змінних;

– Range: Lowest through (Діапазон: від найменшого до) – у цьому випадку буде перекодовано всі значення від найменшого спостережуваного до зазначеного; цей варіант не застосовується для строкових змінних;

– Range: through highest (Діапазон: до найбільшого) – у цьому випадку буде перекодовано всі значення від зазначеного до найбільшого спостережуваного; цей варіант не застосовується для строкових змінних;

– All other values (Усі інші значення) – ця опція стосується всіх іще не відмічених значень; у списку значень змінних вони відображаються як ELSE.

У групі New Value (Нове значення) можна вибрати один з таких варіантів:

– Value – уводиться нове значення;

– System missing (Системне відсутнє значення) – цю опцію призначено для відмічання значення вихідної змінної як системного відсутнього значення. Значення виникає в списку значень змінних у вигляді SYSM1S; цей варіант не застосовується для строкових змінних;

– Copy old value(s) (Копіювати попередні значення) – значення вхідної змінної збережуться без змінень;

Якщо нові вихідні змінні є строковими, слід установити прапорець Output variables are strings (Вихідні змінні є рядками). Тепер треба виконати такі дії:

– увести попередні й нові значення, при цьому попереднє значення треба ввести в поле Value у групі Old Value, нове значення – у поле Value у групі New Value і клацнути на кнопці Add;

– щоб перекодувати попередні значення, треба вибрати опцію All other values, увести нуль у поле Value у групі New Value і клацнути на кнопці Add;

– клацнути на кнопці Continue, а потім – на ОК, нову змінну *x1* додано у файл;

– у редакторі даних двічі клацнути на *x1*, щоб перейти в редактор виду змінних. Установити такі параметри: тип змінної, ширина, десяткові розряди, мітки значень, оголосити пропущене значення.

На закінчення виконати частотний аналіз змінної *x1*.

2.6.3.2. Автоматичне перекодування

Для перетворення значень числових або строкових змінних у неперервну послідовність цілих чисел у SPSS реалізовано можливість автоматичного перекодування. Як приклад розглянемо автоматичне перекодування строкової змінної в числову.

Завантажити файл.

У редакторі даних відобразяться значення строкової змінної *s*, яка складається не більш ніж із двадцяти символів.

Вибрати в меню команди Transform (Перетворити) Automatic Recode... (Автоматичне перекодування)

Відкриється діалогове вікно Automatic Recede (рис. 2.37).

Перенести строкову змінну в поле Variable > New Name (Змінна > Нове ім'я). У текстове поле під ним увести нове ім'я, наприклад s1, і клацнути на кнопці New Name (Нове ім'я).

Клацнути на кнопці ОК.

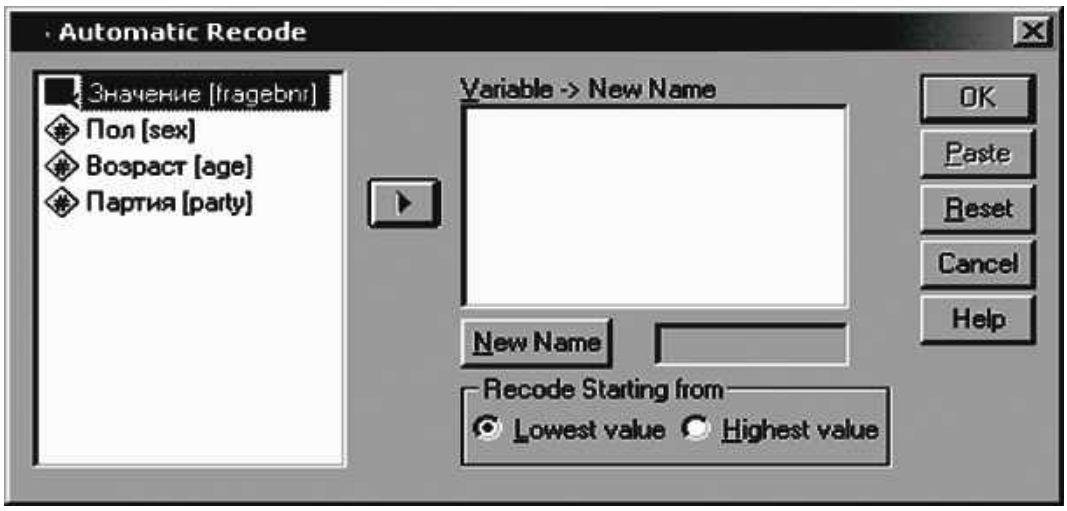


Рис. 2.37. Діалогове вікно Automatic Recede

У вікні перегляду буде відображено таблицю відповідності. Різним значенням строкової змінної *s*, поданим за абеткою, поставлено у відповідність неперервну послідовність натуральних чисел від 1 до 58; ці числові значення зберігаються в змінній *s1*. Колишні строкові значення стали мітками значень цієї змінної.

2.7. Статистичні характеристики

2.7.1. Обчислення статистичних характеристик

Статистичні характеристики обчислюються в основному для змінних, яки належать до інтервальної шкали. Для цього використовуються такі команди меню:

1) Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Descriptives, (Описова статистика);

2) Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Frequencies (Частоти);

3) Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Explore (Досліджувати);

4) Analyze (Аналіз) Reports (Звіти) Case summaries (Підсумки за спостереженнями).

У табл. 2.7 наведено огляд характеристик, що розраховуються в SPSS. У меню Descriptives... можна також провести стандартизацію змінних (*z*-перетворення).

Статистичні характеристики, які задаються в меню Case summaries, можна також обчислити окремо за категоріями групувальних змінних, що належать до номінальної або порядкової шкали.

Таблиця 2.7

Характеристика	Descriptives	Frequencies	Explore	Case summaries
Середнє знчення	X	X	X	X
Сума	X	X		X
Медіана		X	X	X
Групова медіана		X		X
Квартиль		X		
Процентиль		X	X	
Мода		X		
Стандартне відхилення	X	X	X	X
Стандартна помилка	X	X	X	X
Дисперсія	X	X	X	X
Мінімум	X	X	X	X
Максимум	X	X	X	X
Розмах	X	X	X	X
Міжквартильна широта			X	
Ексцес (варіація)	X	X	X	X
Асиметрія	X	X	X	X
Стандартна помилка ексцесу	X	X	X	X
Стандартна помилка асиметрії	X	X	X	X
Довірчий інтервал			X	
Гармонійне середнє				X
Геометричне середнє				X
М-оцінка (Хампеля)			X	
Викид			X	
Усічене середнє			X	

2.7.2. Описова статистика

Для ознайомлення з характеристиками описової статистики розглянемо змінну *age*, що відображає вік.

Завантажити файл і вибрати команди меню Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Descriptives (Описова статистика). Відкриється діалогове вікно Descriptives (рис. 2.38).

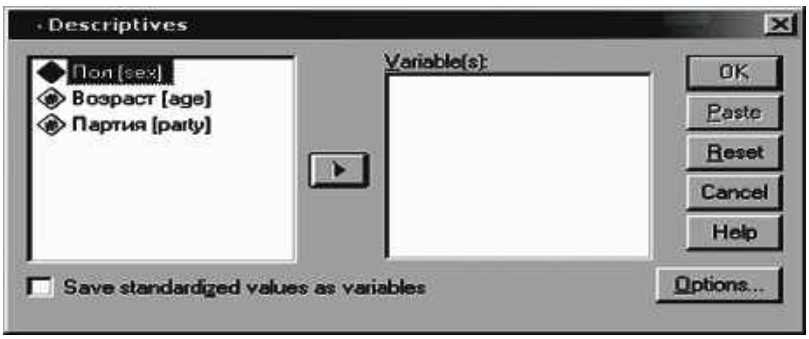


Рис. 2.38. Діалогове вікно Descriptives

Перенести змінну *age* у список змінних, які тестуються, і клацнути на кнопці Options (Параметри).

Тут можна задати обчислення статистичних характеристик (рис. 2.39).

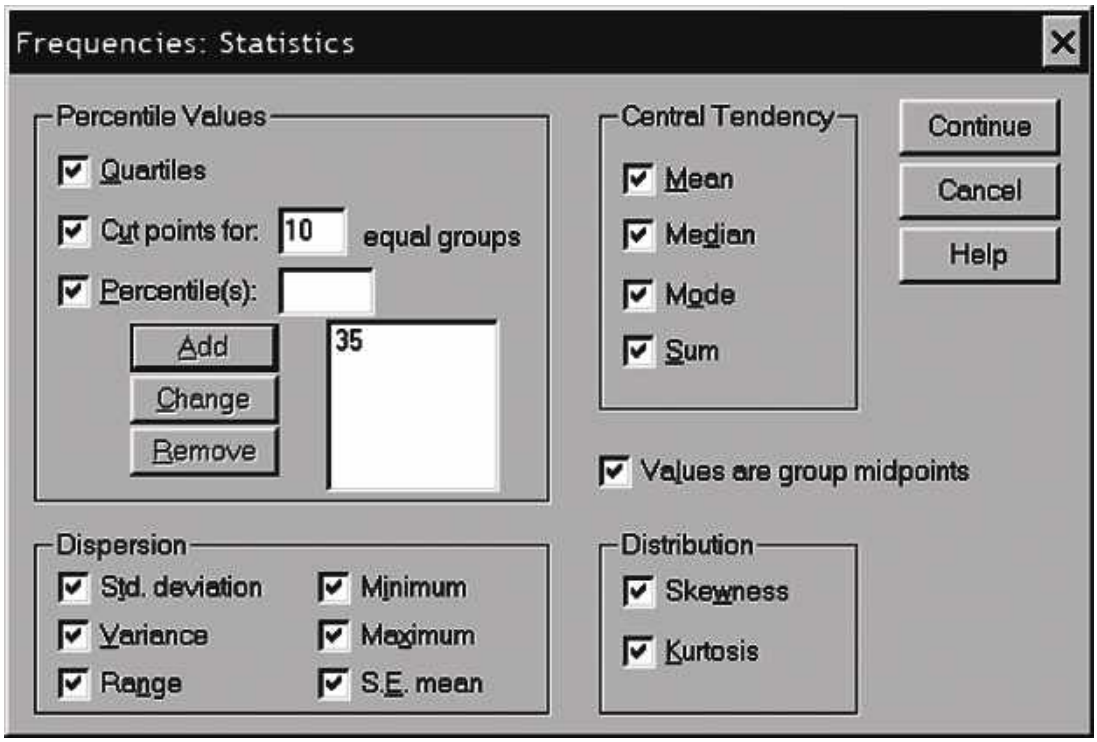


Рис. 2.39. Параметри вікна Frequencies: Statistics

Установити прапорці для виведення, наприклад, таких характеристик: Mean (Середнє значення), Minimum (Мінімум), Maximum (Максимум) і S.E. mean (Стандартна помилка).

Якщо аналізується декілька змінних, можна задати послідовність висновку:

– у порядку зростання середніх значень;

– у порядку убування середніх значень;

– за алфавітом (за іменами змінних);

– відповідно до списку вибраних цільових змінних.

За замовчуванням вибрано останній варіант. Якщо є тільки одна змінна, порядок не має значення.

Позначивши бажані характеристики, клацнути на кнопці Continue... (Далі). У головному діалоговому вікні зазначити, щоб стандартизовані значення було збережено в новій змінній відкритого файла даних, для чого встановити прапорець Save standardized values as variables.

Запустити обчислення, клацнувши на кнопці ОК. Результат буде показано у вікні перегляду у вигляді таблиці Descriptive Statistics (Описова статистика).

2.7.3. Зведення спостережень

Розглянемо обчислення статистичних характеристик.

Завантажити файл і вибрати команди меню Analyze (Аналіз) Reports (Звіти) Case summaries (Зведення спостережень).

Відкриється діалогове вікно Summarize Cases (Вивести зведення спостережень) (рис. 2.40).

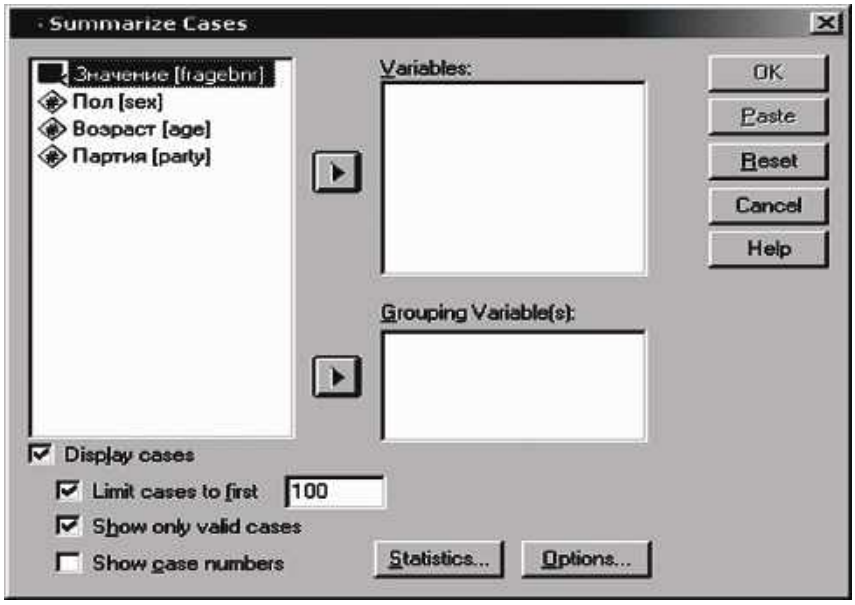


Рис. 2.40. Діалогове вікно Summarize Cases (Зведення спостережень)

Перенести змінну *age* в правий список і зняти прапорець Display Cases (Показувати спостереження).

Клацнути на кнопці Statistics…(Статистика). Відкриється діалогове вікно Summary Report: Statistics (Зведення: Статистика) (рис. 2.41).

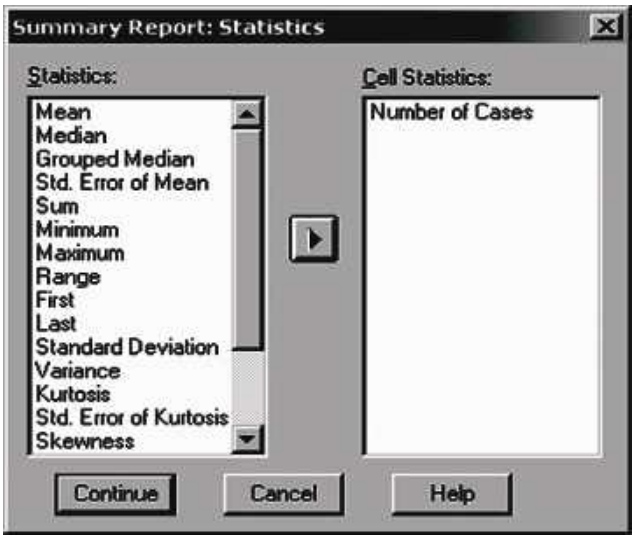


Рис. 2.41. Діалогове вікно Summary Report: Statistics (Зведення: Статистика)

Вибрати у списку обчислення середнього значення (Mean), медіани (Median), мінімального (Minimum) і максимального (Maximum) значень.

Кнопка Options... дає можливість задати заголовок для зведеної таблиці й спосіб оброблення пропущених значень.

У вікні перегляду Summarize Cases (Зведення спостережень) буде показано результати обчислення статистичних характеристик.

Описові характеристики можна також обчислити роздільно по категоріях групувальної змінної, наприклад, для жіночої й чоловічої статі.

2.8. Таблиці спряженості

Досі було розглянуто тільки окремі змінні. Було проведено частотний аналіз, а також описано окремі змінні статистичними характеристиками, такими, як мінімум, максимум і середнє значення. Методи аналізу такого роду називають одновимірними. Перейдемо до двовимірного аналізу й з'ясуємо питання, чи існує взаємозв'язок між двома і більше змінними.

У SPSS є велика кількість різноманітних процедур, за допомогою яких можна провести аналіз зв'язку між двома змінними. Зв'язок між неметричними змінними, тобто такими, що належать до номінальної або порядкової шкали з невеликою кількістю категорій, краще за все подати у формі таблиць спряженості. Для цієї мети в SPSS реалізовано тест %2, при якому перевіряється, чи є значуща відмінність між спостережуваними й очікуваними частотами. Крім того, існує можливість розрахунку різних заходів зв'язаності.

Таблиці спряженості призначено для опису двох або більше номінальних (категоріальних) змінних. Приклади номінальних змінних: стать (чоловіча, жіноча), клас (А, Б, В), відповідь (так, ні) і т. д. Таблиці зв'язаності не застосовують до неперервних змінних, проте їх можна розбити на інтервали.

Для роботи з таблицями спряженості в програмі SPSS використовується команда Crosstabs (Таблиці спряженості).

2.8.1. Створення таблиць спряженості

Для створення таблиць спряженості й обчислення міри зв'язаності на їх основі, вибрати в меню Analyze (Аналіз) команди Descriptive Statistics (Дескриптивні статистики) Crosstabs (Таблиці спряженості). Відкриється діалогове вікно Crosstabs (рис. 2.42).

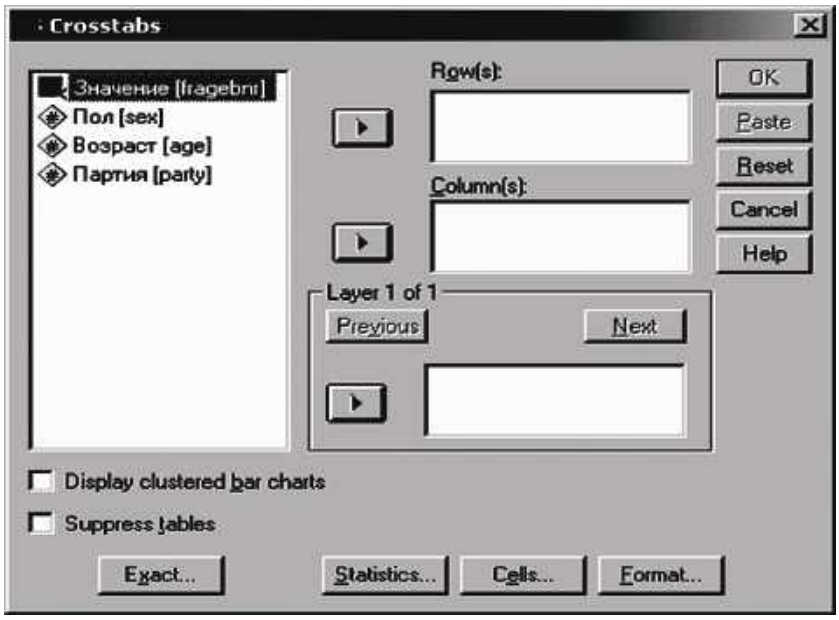


Рис. 2.42. Діалогове вікно Crosstabs (Таблиці спряженості)

Список початкових змінних містить змінні відкритого файла даних. Тут можна вибрати змінні для рядків і стовпців таблиці спряженості. Для кожного поєднання двох змінних буде створено таблицю спряженості. Наприклад, якщо в списку рядків (Rows) знаходиться три змінні, а в списку стовпців (Columns) – дві, то ми отримаємо 3*2 = 6 таблиць спряженості.

Для будування таблиці спряженості зі змінних *var1* і *var2* необхідно діяти таким чином:

– перенести змінну var1 у список рядків, а змінну *var2* – у список стовпців за допомогою кнопок зі стрілками;

– клацнути на ОК, і буде створено таблицю спряженостів стандартному форматі, яку можна проглянути у вікні перегляду (табл. 2.8).

Таблиця 2.8

Стать	Заняття				Total
	Студенти	Пенсіонери	Робітники	Викладачі	
Жіноча	16	18	9	1	44
Чоловіча	3	22	32	5	62
Total	19	40	41	6	106

Розділ Layer 1 of 1 (Шар 1 з 1) діалогового вікна дає можливість побудувати таблицю спряженості для трьох і більше змінних.

При клацанні на кнопці Cells Crosstabs: Cell Display (таблиці спряженості: комірки) відкривається діалогове вікно, що дає можливість управляти інформацією в комірках.

Для детальнішого дослідження залежності між змінними треба буде точно відповісти на такі запитання:

– чи існує залежність взагалі;

– що можна сказати про інтенсивність цієї залежності;

– що можна сказати про напрям і характер цієї залежності?

Ретельніше досліджувати існування залежності дає можливість обчислення значень очікуваних частот. Щоб визначити ці значення, необхідно виконати такі дії:

– вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Crosstabs (Таблиці спряженості). У списку рядків у нас має стояти змінна *var1*, а в списку стовпців – змінна *var2*;

– клацнути на кнопці Cells... (Комірки); відкриється діалогове вікно Crosstabs: Cell Display (Таблиці спряженості: Відображення комірок) (рис. 2.43).

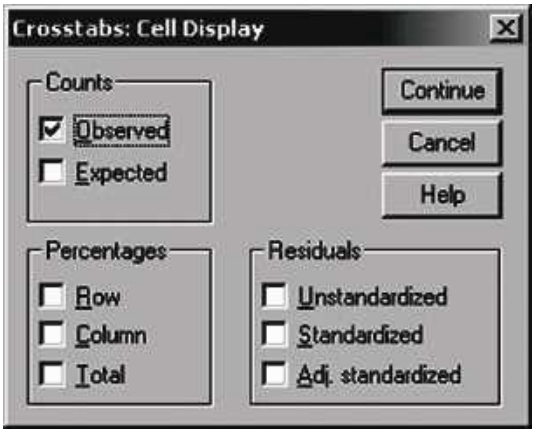


Рис. 2.43. Діалогове вікно Crosstabs: Cell Display (Таблиці спряженості: Відображення комірок)

За замовчуванням в елементах таблиці спряженості відображаються тільки спостережувані значення частот. У групі Counts (Частоти) можна вибрати один або більше таких варіантів відображення:

– Observed (Спостережувані) – відображатимуться спостережувані частоти, це – налаштування за замовчуванням;

– Expected (Очікувані) – якщо встановити цей прапорець, відображатимуться очікувані частоти, які обчислюються як добуток сум відповідного рядка і стовпця, поділений на загальну суму частот.

2.8.2. Формати таблиць спряженості

Можна змінити порядок сортування змінних рядків в таблиці спряженості, клацнувши в діалоговому вікні Crosstabs на кнопці Format... (Формат). Відкриється діалогове вікно Crosstabs: Table Format (Таблиці спряженості: Формат таблиці) (рис. 2.44).

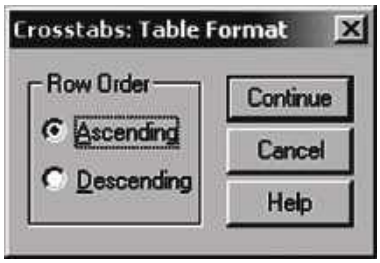


Рис. 2.44. Діалогове вікно Crosstabs: Table Format

У групі Row Order (Порядок рядків) можна вибрати один з таких варіантів сортування значень:

– Ascending (За збільшенням) – значення змінних рядків відображуються в порядку зростання від найменшого до найбільшого. Це настроювання за замовчуванням;

– Descending (За убуванням) – значення змінних рядків відображуються в порядку убування від найбільшого до найменшого.

2.8.3. Графічне подання таблиць спряженості

Щоб зробити наочнішими дані, що містяться в таблицях спряженості, їх можна подати візуально. Для цього необхідно:

– вибрати в меню команди Graphs (Графіки) Bar (Стовпчасті) Відкриється діалогове вікно Bar Charts (Стовпчасті діаграми);

– вибрати пункт Clustered (Згруповані), залишити пропоновану за замовчуванням опцію Summaries for groups of cases (Зведення категорій змінної) і клацнути на кнопці Define (Визначити); відкриється діалогове вікно Define Clustered Bar: Summaries for groups of cases (Визначити стовпчасту діаграму: Зведення категорій змінної) (рис. 2.45);

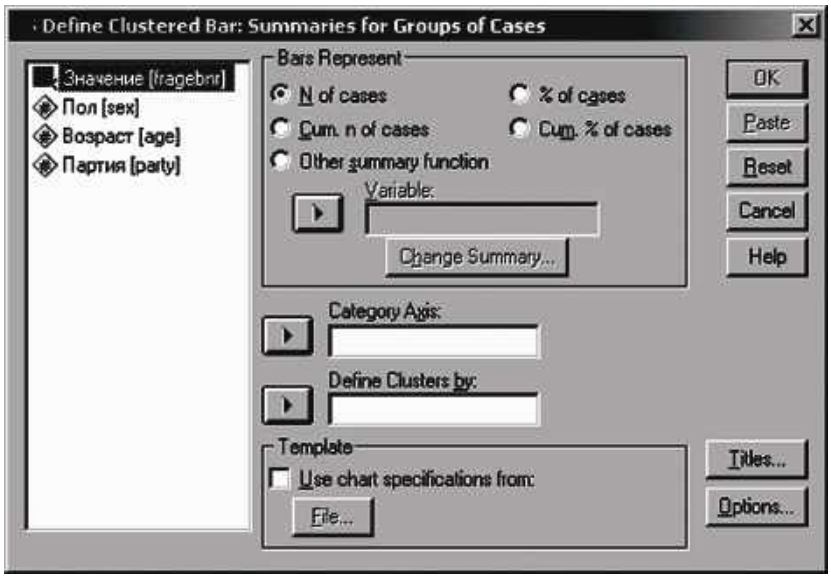


Рис. 2.45. Діалогове вікно Define Clustered Bar: Summaries for groups of cases

– вибрати пункт % of cases (% спостережень);

– перенести змінну *var1* в поле Category Axis (Вісь категорій), а змінну *var2* – в поле Define Clusters by (Визначити групи);

– клацнути на кнопці Titles... (Заголовки), відкриється діалогове вікно Titles (рис. 2.46);

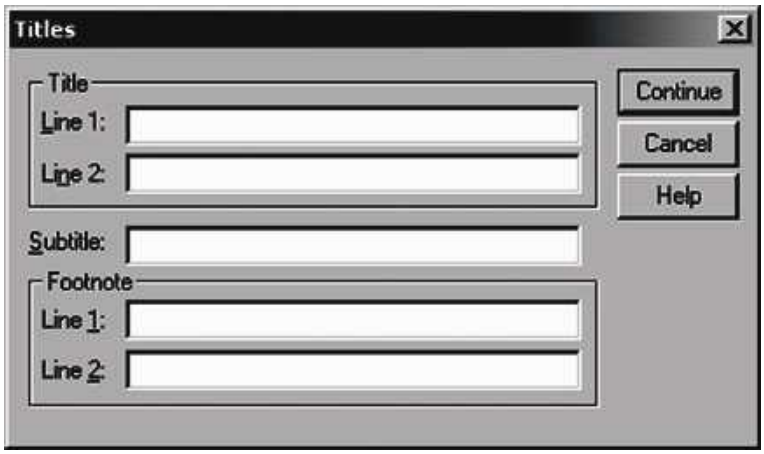


Рис. 2.46. Діалогове вікно Titles

– в полі Line 1 (Рядок 1) ввести заголовки й підзаголовки таблиці, підтвердити введення кнопкою Continue;

– клацнути на кнопці Options… (Параметри), відкриється діалогове вікно Options (рис. 2.47);

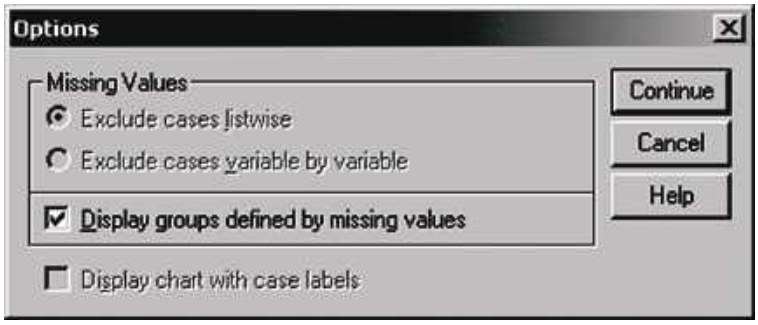


Рис. 2.47. Діалогове вікно Options

– зняти в ньому прапорець Display groups defined by missing values (Відображати групи, які створено пропущеними значеннями);

– клацнути на кнопці Continue, а потім на ОК, у вікні перегляду виникне графік;

– двічі клацнути на цьому графіку – відкриється редактор діаграм, в якому його можна правити;

– вибрати в меню команди Format (Формат) Bar Label Style (Стиль міток стовпців), відкриється діалогове вікно Bar Label Style;

– вибрати пункт Framed (У рамках), клацнути на кнопці Apply all (Застосувати для усіх) і потім на Close (Закрити);

– клацнути на одному із стовпців, що відображає значення зміної

var1, стовпці буде виділено; це можна визначити за маленькими чорними квадратиками на кутах стовпців;

– вибрати в меню команди Format (Формат) Color (Колір), відкриється діалогове вікно Colors (Кольори); тут можна змінити стандартний колір стовпців, а також колір їхніх контурів;

– клацнути на сірому полі, а потім на кнопках Apply (Застосувати) і Close (Закрити);

– на закінчення викликати команди меню Chart (Діаграма) Outer Frame (Зовнішня рамка).

Вийде графічне подання таблиці спряженості (рис. 2.48).

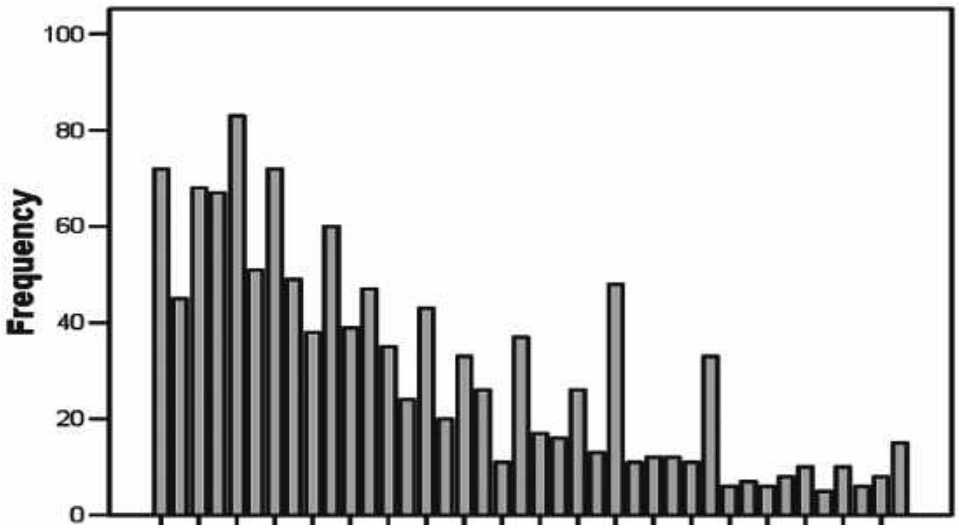


Рис. 2.48. Графічне подання: стовпчаста діаграма

2.8.4. Статистичні критерії для таблиць спряженості

Щоб отримати статистичні критерії для таблиць спряженості, треба клацнути на кнопці Statistics... (Статистика) у діалоговому вікні Crosstabs. Відкриється діалогове вікно Crosstabs : Statistics (рис. 2.49).

Прапорці в цьому діалоговому вікні дають можливість вибрати один або декілька критеріїв:

– тест хі-квадрат (X^2);
– кореляції;
– заходи зв'язаності для змінних, що належать до номінальної шкали;
– заходи зв'язаності для змінних, що належать до порядкової шкали;
– заходи зв'язаності для змінних, що належать до інтервальної шкали;
– коефіцієнт Каппа (k);
– міра ризику;
– тест Мак-Немара;
– статистики Кохрана і Мантеля–Хэнзеля.

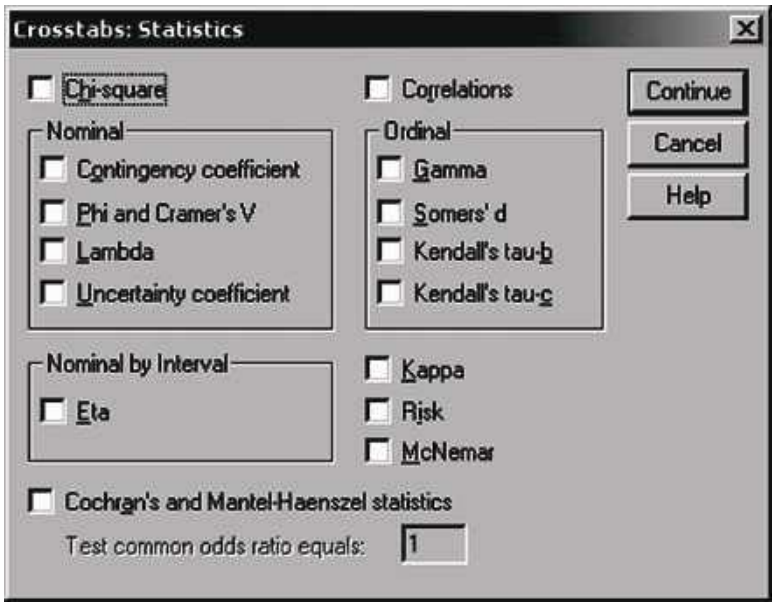


Рис. 2.49. Діалогове вікно Crosstabs: Statistics

2.8.4.1. Тест хі-квадрат (X^2)

При проведенні тесту хі-квадрат перевіряється взаємна незалежність двох змінних таблиці спряженості й завдяки цьому побічно з'ясовується залежність обох змінних. Дві змінні вважаються взаємно незалежними, якщо спостережувані частоти (*f0*) в осередках збігаються з очікуваними частотами (*fe*).

Для того, щоб провести тест хі-квадрат за допомогою SPSS, треба виконати такі дії:

– вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Crosstabs (Таблиці спряженості);

– кнопкою Reset (Скидання) видалити можливі настройки;

– перенести змінну *var1* в список рядків, а змінну *var2* – в список стовпців;

– клацнути на кнопці Cells (Комірки). У діалоговому вікні встановити окрім пропонованого за замовчуванням прапорця Observed ще й прапорці Expected і Standardized; підтвердити вибір кнопкою Continue;

– клацнути на кнопці Statistics (Статистика), відкриється описане вище діалогове вікно Crosstabs: Statistics;

– встановити прапорець Chi-square (Хі-квадрат), клацнути на кнопці, відкриється описане вище діалогове вікно Crosstabs: Statistics; у цьому вікні можна вибрати статистичні критерії і міри незалежності;

– встановити прапорець Chi-square (Хі-квадрат), клацнути на кнопці Continue, а в головному діалоговому вікні – на ОК, отримаємо таблицю спряженості.

2.8.4.2. Коефіцієнти кореляції

Досі з'ясовувався лише сам факт існування статистичної залежності між двома ознаками. Далі спробуємо з'ясувати, які висновки можна зробити про силу або слабкість цієї залежності, а також про її вид і спрямованість. Критерії кількісної оцінки залежності між змінними називають коефіцієнтами кореляції або заходами зв'язаності.

Як коефіцієнт кореляції між змінними, що належать до порядкової шкали, застосовується коефіцієнт Спирмена, а для змінних, що належать до інтервальної шкали, – коефіцієнт кореляції Пірсона (момент додатків). При цьому слід враховувати, що кожну дихотомічну змінну, тобто таку змінну, що належить до номінальної шкали і має дві категорії, можна розглядати як порядкову.

Спершу перевіримо, чи існує кореляція між двома змінними *var1* і *var2*. При цьому врахуємо, що дихотомічну змінну *var1* можна вважати порядковою. Виконаємо такі дії:

– вибрати в меню команди Analyze (Аналіз) Descriptive Statistics (Дескриптивні статистики) Crosstabs (Таблиці спряженості);

– перенести змінну *var1* в список рядків, а змінну *var2* – в список стовпців;

– клацнути на кнопці Statistics... (Статистика). У діалозі Crosstabs: Statistics встановити прапорець Correlations (Кореляції), підтвердити вибір кнопкою Continue;

– в діалозі Crosstabs відмовітися від виведення таблиць, встановивши прапорець Supress tables (Приховати таблиці), клацнути на кнопці ОК.

Буде обчислено коефіцієнти кореляції Спірмена й Пірсона, а також проведено перевірку їх значущості.

Виходячи з таблиці результатів, можна зробити висновки, чи існує кореляція між змінними, слабка вона чи ні (Виведення про силу залежності), змінні корелюють позитивно чи негативно (Виведення про напрям залежності).

2.8.4.3. Заходи зв'язаності для змінних з номінальною шкалою

Коефіцієнт кореляції не можна застосовувати як характеристику залежності між змінними, якщо ці змінні належать до номінальної шкали й мають більше двох категорій, тому що між їхніми кодуваннями неможливо встановити порядкового відношення і, отже, вони не можуть бути розташовані в певному порядку.

Найкращим засобом для аналізу таких залежностей уважається тест хі-квадрат, після якого за необхідності можна провести аналіз спостережуваних і очікуваних частот, а також нормованих залишків.

Проте і в цьому випадку також здійснювалися спроби розробити критерії кількісної оцінки міри зв'язаності двох змінних, поставлених у взаємну відповідність. Ці критерії показують міру взаємної залежності або незалежності двох змінних, що належать до номінальної шкали, причому значення 0 відповідає повній незалежності змінних, а 1 – їхній максимальній залежності. Заходи зв'язаності не можуть мати негативних значень, оскільки за відсутності порядкового відношення не можна дати відповіді на запитання про напрям залежності.

2.9. Порівняння середніх

Порівняння середніх значень різних вибірок належить до найбільш часто застосовуваних методів статистичного аналізу. При цьому завжди має бути з'ясованим питання, чи можна пояснити наявне розходження середніх значень статистичними коливаннями, чи ні. В останньому випадку говорять про значуще розходження.

При порівнянні середніх значень вибірок передбачається, що обидві вибірки підпорядковуються нормальному розподілу. Якщо це не так, то обчислюються медіани й для порівняння вибірок використовується непараметричний тест.

При порівнянні середніх значень вибірок виділяють чотири різні тестові ситуації:

– порівняння двох незалежних вибірок;
– порівняння двох залежних (спарених) вибірок;
– порівняння понад дві незалежні вибірки;
– порівняння понад дві залежні вибірки.

У цих ситуаціях застосовуються відповідно такі статистичні тести:

– t-тест для незалежних вибірок (тест Стьюдента);
– t-тест для залежних вибірок;
– однофакторний дисперсійний аналіз;
– однофакторний дисперсійний аналіз із повторними вимірами.

Перші три з цих тестів викликаються за допомогою меню Analyze (Аналіз) Compare Means (Порівняння середніх).

Щоб провести однофакторний дисперсійний аналіз із повторними вимірами (тестова ситуація, що дуже часто трапляється), треба викликати команду меню Analyze (Аналіз) General Linear Model (Загальна лінійна модель) Repeated Measures (Повторні виміри).

Спочатку розглянемо тести, виклик яких відбувається з допомогою пункту меню команди Analyze (Аналіз) Compare Means (Порівняння середніх).

У підменю є, зокрема, t-тест для незалежних вибірок (Independent-Samples T Test), t-тест для парних вибірок (Paired-Samples T Test) та однофакторний дисперсійний аналіз (ANOVA) для порівняння декількох незалежних вибірок (One-Way ANOVA).

Ще один тест, включений у це підменю, – t-тест випадкової вибірки, який використовується для порівняння середніх із заданим значенням (One-Sample T Test). У підпункті меню Means (Середні) обчислюються середні значення роздільно за категоріями групувальної змінної; тут також можна перевірити існування значущого розходження з допомогою однофакторного дисперсійного аналізу. Щодо цього цей підпункт надає менше можливостей, ніж підпункт One-Way ANOVA.

Для порівняння двох незалежних вибірок вибрати у підменю команду Independent–Samples T Test (t-тест для незалежних вибірок).

Відкриється діалогове вікно Independent-Samples T Test (рис. 2.50).

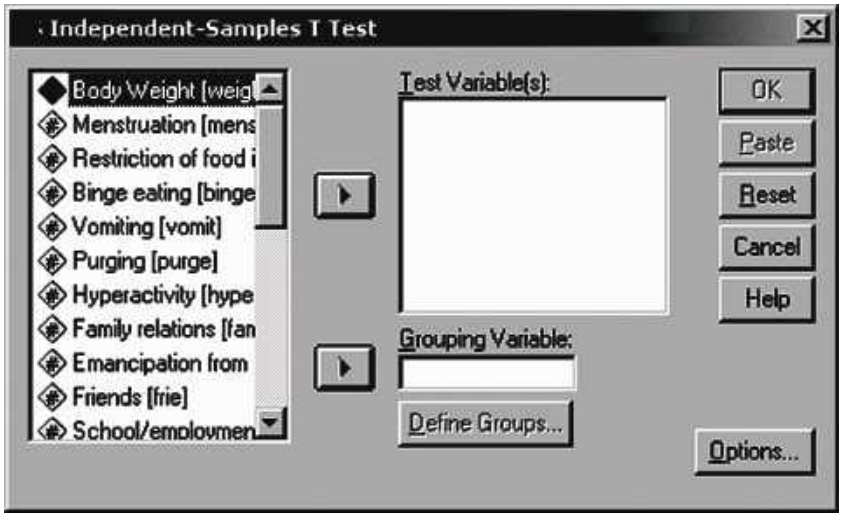


Рис. 2.50. Діалогове вікно Independent-Samples T Test

Запустити t-тест, клацнувши на ОК. У вікні перегляду виникнуть результати, які містять:

– кількість спостережень, середні значення, стандартні відхилення й стандартні помилки середніх в обох групах;

– результати тесту Левена на рівність дисперсій.

Як правило, гіпотеза про рівність (гомогенність) дисперсій не приймається, якщо тест Левена дає значення $p < 0,05$ (гетерогенність дисперсій). Для випадків як гомогенності (рівності), так і гетерогенності (нерівності) виводяться такі характеристики:

– результати t-тесту: значення розподілу t, кількість ступенів свободи *df*, імовірність помилки *p* (під позначенням «Значущість (2-стороння)»);

– різниця середніх значень, її стандартна помилка й довірчий інтервал.

2.10. Кореляції

У цьому підрозділі досліджено зв'язок (кореляцію) між двома змінними. Розрахунки таких двовимірних критеріїв взаємозв'язку ґрунтуються на формуванні парних значень, які утворюються з розглянутих залежних вибірок.

Визначати силу зв'язку можна за допомогою деякого критерію взаємозв'язку – коефіцієнта кореляції *r*.

Для графічного подання подібного зв'язку можна використовувати прямокутну систему координат з осями, які відповідають обом змінним. Кожна пара значень маркірується з допомогою певного символу. Такий графік, що має назву «діаграма розсіювання», для двох залежних змінних можна побудувати шляхом виклику меню Graphs (Графіки) Scatter plots (Діаграми розсіювання) (рис. 2.51).

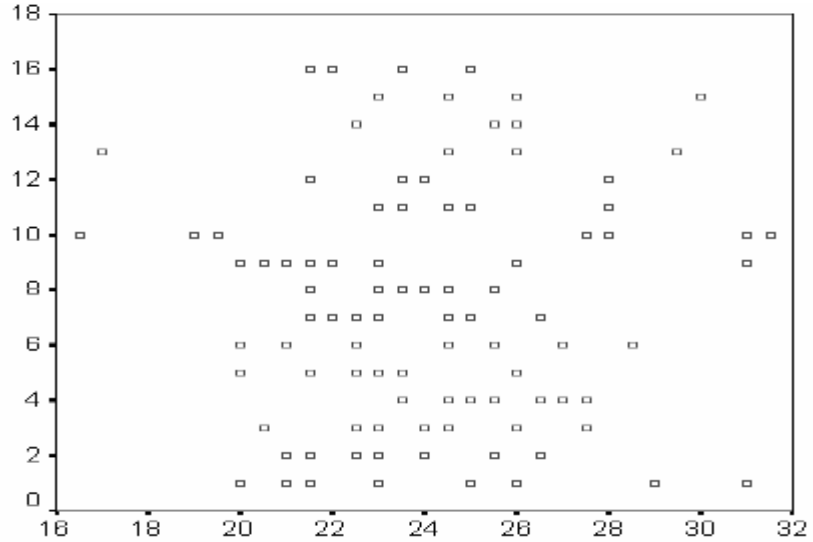


Рис. 2.51. Діаграми розсіювання

Метод обчислення коефіцієнта кореляції залежить від виду шкали, до якої належать змінні:

– для змінних, які належать до інтервальної і номінальної шкал розраховують коефіцієнт кореляції Пірсона (кореляція моментів добутків);

– якщо одна із двох змінних має порядкову шкалу або не є нормально розподіленою – рангова кореляція за Спірманом;

– якщо одна із двох змінних є дихотомічною – крапкова дворядна кореляція, однак цієї можливості в SPSS немає, замість цього можна застосувати розрахунок рангової кореляції;

– якщо обидві змінні є дихотомічними – чотирипольова кореляція, цей вид кореляції розраховується в SPSS на основі визначення мір відстані й мір подібності;

– якщо обидві змінні недихотомічні, то розрахунок коефіцієнта кореляції не позбавлений сенсу тільки тоді, коли зв'язок між ними є лінійним (односпрямованим). Якщо зв'язок, наприклад, U-подібний (неоднозначний), то коефіцієнт кореляції є непридатним для використання як міра сили зв'язку: його значення прямує до нуля.

2.10.1. Коефіцієнт кореляції Пірсона

Розглянемо на основі даних розрахунок коефіцієнта кореляції Пірсона попарно для змінних c_{10}, c_{11}, c_{12} і c_{13} (тобто сформуємо для цих змінних кореляційну матрицю).

Відкрити файл даних.

Вибрати в меню Analyze (Аналіз) Correlate (Кореляція) Bivariate (Парні). Виникне діалогове вікно Bivariate Correlations (Парні кореляції) (рис. 2.52).

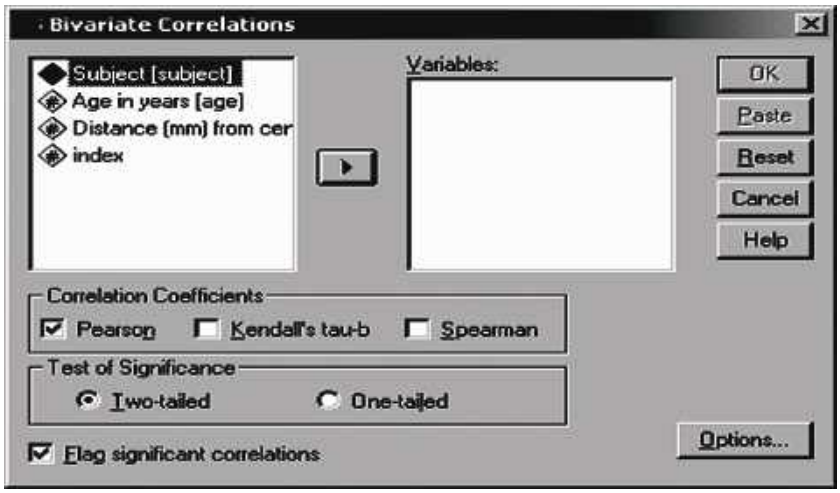


Рис. 2.52. Діалогове вікно Bivariate Correlations (Двовимірні кореляції)

Змінні c_{10}, c_{11}, c_{13} і c_{14} перенести по черзі в поле змінних, що тестуються. Розрахунок коефіцієнта кореляції Пірсона є попередньою установкою, так само, як і двостороння перевірка значущості й маркування значущих кореляцій.

Почати розрахунок шляхом натискання кнопки ОК.

Отримані результати містять: кореляційний коефіцієнт Пірсона *r*, кількість використаних пар значень змінних і ймовірність помилки *p*, що відповідає припущенню про ненульову кореляцію.

З допомогою клацання на кнопці Options... (Опції) можна організувати розрахунок середнього значення й стандартного відхилення для двох змінних. Додатково можуть виводитися відхилення добутків моментів (значень чисельника формули для коефіцієнта кореляції) та елементи коваріаційної матриці (чисельник, поділений на $n - 1$).

2.10.2. Рангові коефіцієнти кореляції за Спірманом і Кендалом

Для змінних, що належать до порядкової шкали, або таких, що не підпорядковуються нормальному розподілу, а також для змінних, що належать до інтервальної шкали, замість коефіцієнта Пірсона розраховується рангова кореляція за Спірманом. Для цього окремим значенням змінних присвоюються рангові місця, які згодом обробляються з допомогою відповідних формул.

Щоб виявити рангову кореляцію, необхідно убрати у діалоговому вікні Bivariate Correlations (Парні кореляції) мітку для розрахунку кореляції за Пірсоном, установлену за замовчуванням. Замість цього активувати розрахунок кореляції Спірмана.

Коефіцієнти рангової кореляції є досить близькими до відповідних значень коефіцієнтів Пірсона (вихідні змінні мають нормальний розподіл). Ще одним варіантом рангових коефіцієнтів кореляції є коефіцієнти Кендала (tb Кендала), розрахунок яких можна викликати в діалоговому вікні Bivariate Correlations (Парні кореляції). У цьому методі одна змінна подається у вигляді монотонної послідовності в порядку зростання величин; іншій змінній присвоюються відповідні рангові місця. Кількість інверсій (порушень монотонності порівняно з першим рядом) використовується у формулі для кореляційних коефіцієнтів. Застосування коефіцієнта Кендала є кращим, якщо у вихідних даних трапляються розкиди.

2.10.3. Часткова кореляція

Якщо, наприклад, досліджувати досить велику сукупність чоловіків і зіставити розмір їхнього взуття з рівнем освіченості, то між цими двома змінними можна помітити хоч і невелику, але в той же час значущу кореляцію. Ця кореляція може стати прикладом так званої помилкової кореляції. Тут статистично значущий коефіцієнт кореляції є не проявом деякого причинного зв'язку між двома розглянутими змінними, а більшою мірою обумовлений деякою третьою змінною. Для цього прикладу такою змінною може бути зріст. З одного боку існує деяка незначна кореляція між зростом і рівнем освіченості, а з іншого боку – цілком з'ясовний і логічний зв'язок між зростом і розміром взуття. Разом ці дві кореляції призводять до згаданої помилкової кореляції. Для виключення однієї такої спотворювальної змінної необхідно розрахувати так звану часткову кореляцію.

Якщо присвоїти корелювальним змінним індекси 1 і 2, а спотворювальній змінній – індекс 3 і попарно розрахувати кореляційний коефі-

цієнт (Пірсона) r_{12}, r_{13} і r_{23} , то для часткових кореляційних коефіцієнтів одержимо

$$r_{123} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{13}^2)}}.$$

Розрахунок часткових кореляційних коефіцієнтів

Відкрити файл.

Вибрати в меню Analyse (Аналіз) Correlate (Кореляція) Partial (Часткова).

Відкриється діалогове вікно Partial Correlations (Часткові кореляції) (рис. 2.53).

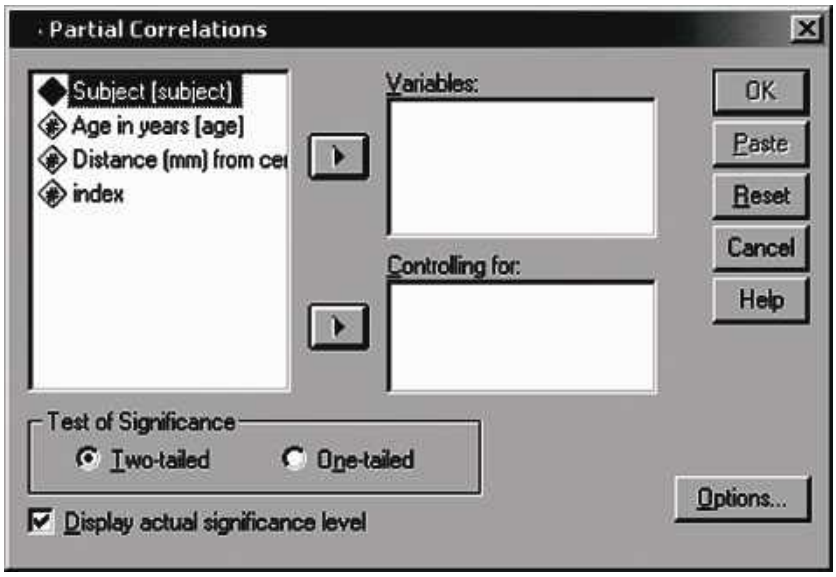


Рис. 2.53. Діалогове вікно Partial Correlations (Часткові кореляції)

Перенести змінні c_{10}, c_{11} у поле ознак, а змінну c_{12} – у поле контрольних змінних і залишити попередню установку для двостороннього тесту значущості.

Клацанням на кнопці Options... (Опції) нарівні з традиційним обробленням пропущених значень можна організувати розрахунок середнього значення, стандартного відхилення й виведення «кореляцій нульового порядку» (тобто простих кореляційних коефіцієнтів).

У випадку однієї спотворювальної змінної, як у наведеному прикладі, можливий розрахунок часткової кореляції першого порядку. За наявності декількох спотворювальних змінних SPSS видає кореляції вищих порядків.

Почати розрахунок клацанням на кнопці ОК.

Результати містять: частковий кореляційний коефіцієнт, число ступенів свободи (кількість спостережень мінус 3) і рівень значущості.

2.11. Регресійний аналіз

Якщо розрахунок кореляції характеризує силу зв'язку між двома змінними, то регресійний аналіз призначено для визначення виду цього зв'язку й дає можливість для прогнозування значення однієї (залежної) змінної, відштовхуючись від значення іншої (незалежної).

Щоб викликати регресійний аналіз в SPSS, вибрати у меню Analyze (Аналіз) Regression (Регресія) (рис. 2.54).

Відкриється відповідне підменю.

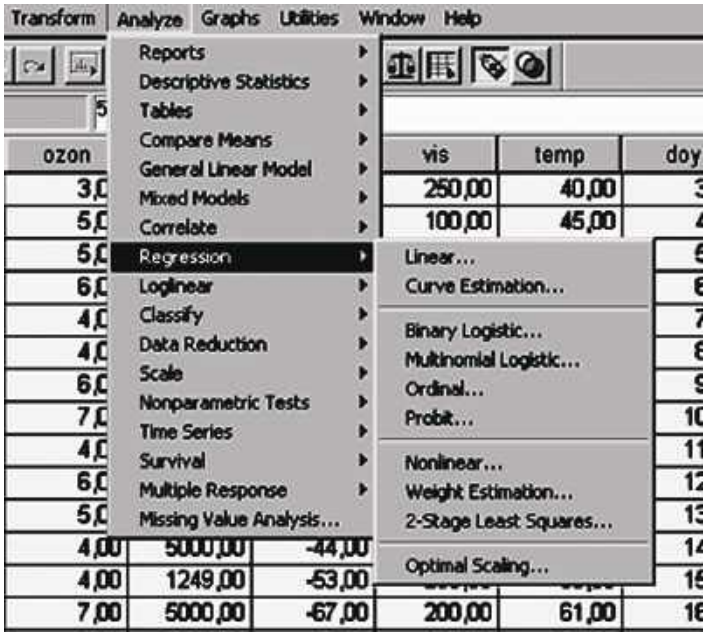


Рис. 2.54. Допоміжне меню Regression (Регресія)

При вивченні лінійного регресійного аналізу буде проведено розходження між простим (одна незалежна змінна) і множинним (декілька незалежних змінних) аналізами. Власне кажучи, ніяких принципових відмінностей між цими видами регресії немає, однак проста лінійна регресія є найпростішою й застосовується частіше від усіх інших видів.

Для проведення лінійного регресійного аналізу залежна змінна повинна мати інтервальну (або порядкову) шкалу. Якщо залежна змінна є категоріальною, але має більше двох категорій, то тут придатним методом буде мультиномільна логістична регресія. Нововведенням у 10-й версії SPSS є порядкова регресія, яку можна використовувати, коли залежні змінні належать до порядкової шкали. І нарешті, можна аналізувати й нелінійні зв'язки між змінними, які належать до інтервальної шкали. Для цього призначено метод нелінійної регресії.

2.11.1. Проста лінійна регресія

Цей вид регресії найкраще підходить для того, щоб продемонструвати основні принципи регресійного аналізу. Можна легко помітити очевидний зв'язок: обидві змінні розвиваються в одному напрямку, і множина крапок, що відповідають спостережуваним значенням показників, явно концентрується (за деякими виключеннями) поблизу прямої (прямої регресії).

При проведенні простої лінійної регресії основною задачею є визначення параметрів *b* і *a*. Оптимальним рішенням цієї задачі є така пряма, для якої сума квадратів вертикальних відстаней до окремих точок даних є мінімальною.

Після визначення цих параметрів, можна спрогнозувати показник, що виникне через деякий час.

Розрахунок рівняння лінійної регресії

Відкрити файл.

Вибрати в меню Analyze (Аналіз) Regression (Регресія) Linear (Лінійна). Виникне діалогове вікно Linear Regression (Лінійна регресія) (рис. 2.55).

Перенести залежну змінну в поле для залежних змінних і присвоїти змінній статус незалежної змінної.

Нічого більше не міняючи, почати розрахунок натисканням ОК.

Виведення основних результатів здіснюється у таблицю «Coefficients (Коефіцієнти)». У таблиці виводяться коефіцієнт регресії *b* і зсув по осі ординат *a* під ім'ям "константа".

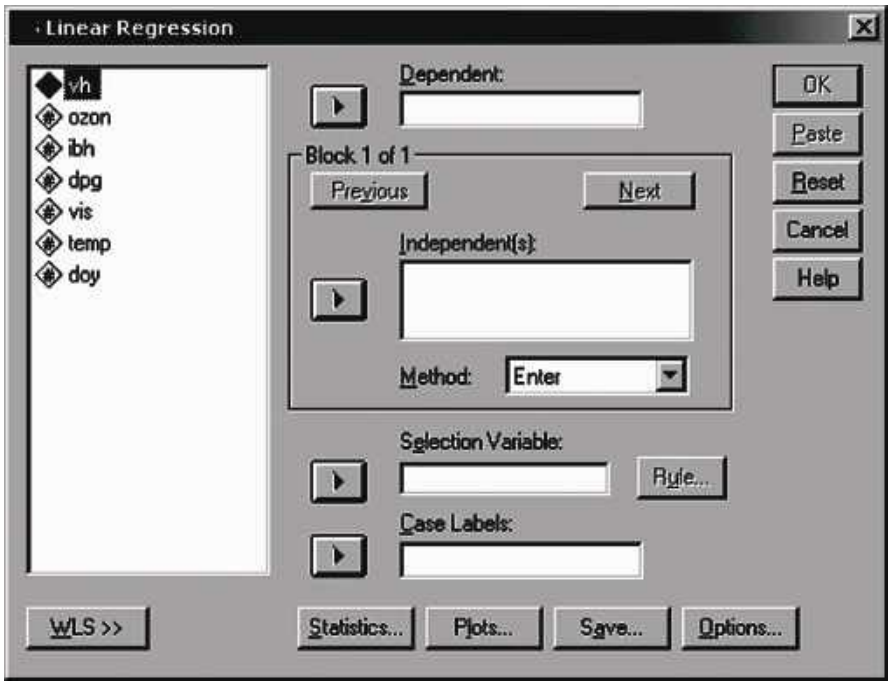


Рис.2.55. Діалогове вікно Linear Regression (Лінійна регресія)

У простому лінійному регресійному аналізі квадратний корінь із коефіцієнта детермінації, що позначено «*R*», дорівнює кореляційному коефіцієнту Пірсона. При множинному аналізі ця величина менш наочна, ніж сам коефіцієнт детермінації. Величина «зміщений R-квадрат» завжди менша, ніж величина «незміщений R-квадрат». За наявності великої кількості незалежних змінних міра визначеності коригується у бік зменшення. Принципове питання про те, чи може взагалі наявний зв'язок між змінними розглядатися як лінійний, простіше й наочніше за все вирішувати, дивлячись на відповідну діаграму розсіювання. Крім того, на користь гіпотези про лінійний зв'язок говорить також високий рівень дисперсії, яка описується рівнянням регресії.

Стандартизовані прогнозовані значення й стандартизовані залишки можна подати у вигляді графіка, який можна одержати, якщо через кнопку Plots...(Графіки) зайти у відповідне діалогове вікно й задати в ньому параметри *ZRESID і *ZPRED як змінні, що відображаються по осях *y* и *x* відповідно. У випадку лінійної регресії залишки розподіляються випадково по обох сторонах від горизонтальної нульової лінії.

2.11.1.1. Збереження нових змінних

Численні допоміжні значення, що розраховуються під час будування рівняння регресії, можна зберегти як змінні й використовувати в подальших розрахунках.

Для цього в діалоговому вікні Linear Regression (Лінійна регресія) клацнути на кнопці Save (Зберегти). Відкриється діалогове вікно Linear Regression: Save (Лінійна регресія: Збереження).

Цікавими тут є опції Standardized (Стандартизовані значення) і Unstandardized (Нестандартизовані значення), які знаходяться під рубрикою Predicted values (Прогнозовані величини опції). При виборі опції «Нестандартизовані значення» будуть розраховуватися значення *y*, що відповідають рівнянню регресії. При виборі опції «Стандартизовані значення» прогнозована величина нормалізується. SPSS автоматично присвоює нове ім'я кожній новоствореній змінній, незалежно від того, розраховуються прогнозовані значення, відстані, прогнозовані інтервали, залишки чи які-небудь інші важливі статистичні характеристики. Нестандартизованим значенням SPSS присвоює імена *pre_1* (predicted value), *pre_2* і т.д., а стандартизованим – *zpr_l*.

Клацнути у діалоговому вікні Linear Regression: Save (Лінійна регресія: Збереження) у поле Predicted values (Прогнозовані значення) на опції Unstandardized (Нестандартизовані значення).

Підтвердити натисканням Continue (Далі) і по закінченні – ОК.

У редакторі даних було створено нову змінну під ім'ям *pre_1* і додано в кінець списку змінних у файлі. Ця змінна містить нестандартизоване прогнозоване значення.

SPSS використовує в розрахунках більш точні значення, ніж ті, які виводяться у вікні перегляду результатів.

2.11.1.2. Будування регресійної прямої

Щоб на діаграмі розсіювання зобразити регресійну пряму, потрібно провести такі дії:

– вибрати в меню опції Graphs (Графіки) Scatter plots (Діаграми розсіювання), відкриється діалогове вікно Scatter plot (Діаграма розсіювання) (рис. 2.56);

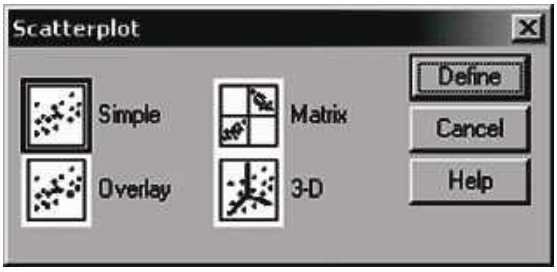


Рис. 2.56. Діалогове вікно Scatter plots... (Діаграма розсіювання)

– у діалоговому вікні Scatter plots (Діаграма розсіювання) залишіти попередню установку Simple (Проста) і клацнути на кнопці Define (Визначити). Відкриється діалогове вікно Simple Scatter plot (Проста діаграма розсіювання) (рис. 2.57);

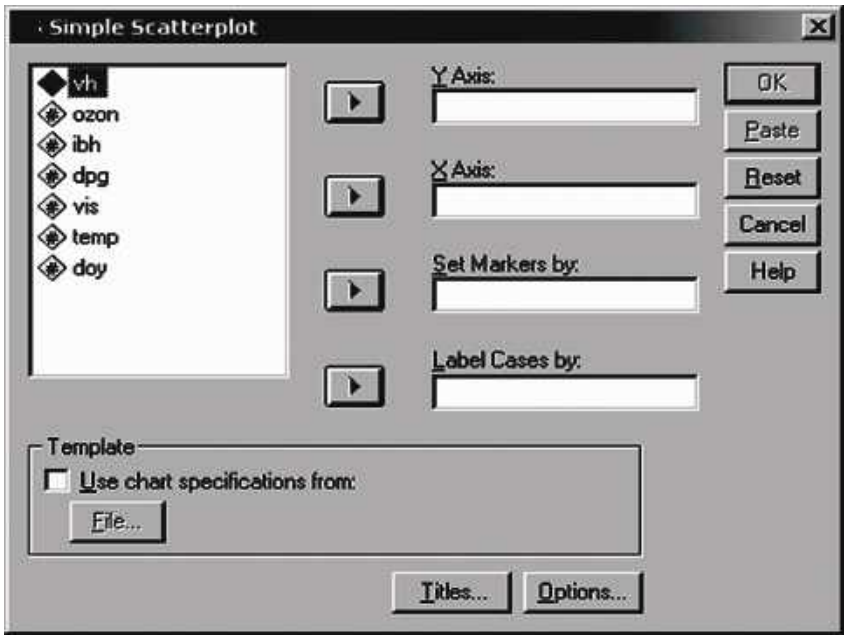


Рис. 2.57. Діалогове вікно Simple Scatterplot (Проста діаграма розсіювання)

– перенести залежну змінну в поле осі *Y*, а незалежну – в поле осі *X*;

– підтвердити клацанням на ОК; у вікні перегляду результатів виникне діаграма розсіювання (рис. 2.58);

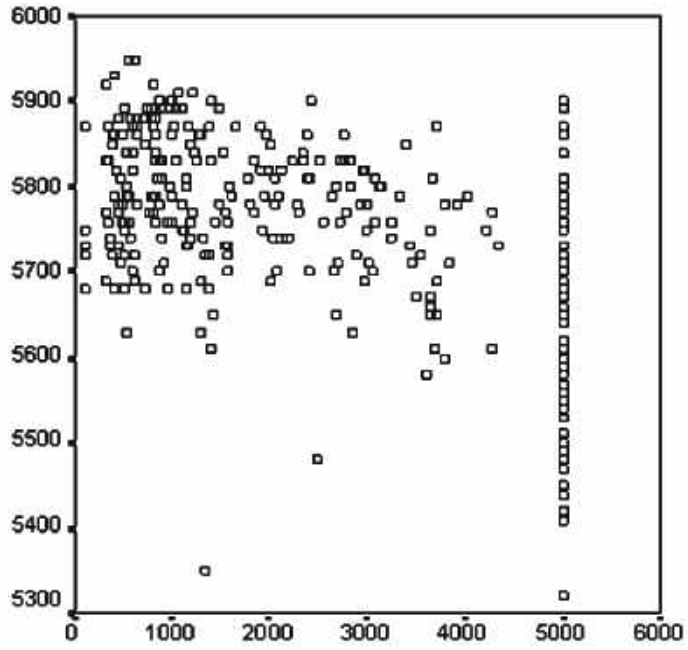


Рис. 2.58. Діаграма розсіювання у вікні перегляду

– клацнути двічі на цьому графіку, щоб перенести його в редактор діаграм;

– вибрати в редакторі діаграм меню Chart (Діаграма) Options (Опції); відкриється діалогове вікно Scatterplot Options (Опції для діаграми розсіювання) (рис. 2.59);

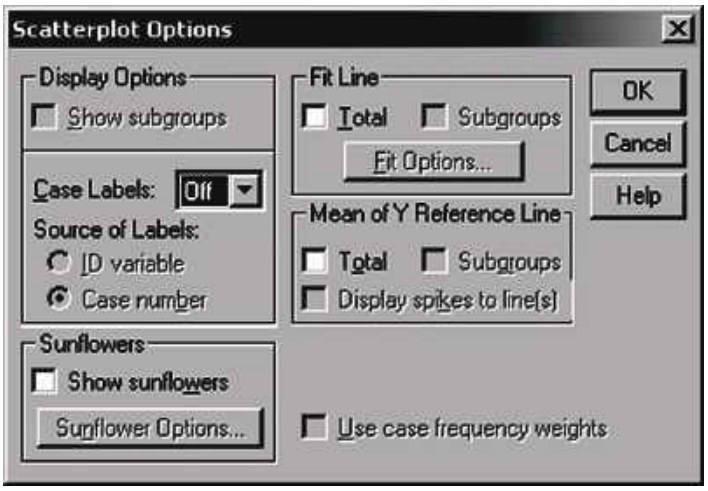


Рис. 2.59. Діалогове вікно Scatterplot Options
(Опції для діаграми розсіювання)

– у рубриці Fit Line (Наближена крива) поставити прапорець напроти опції Total (Повністю для всього файла даних) і клацнути на

кнопці Fit Options (Опції для наближення), відкриється діалогове вікно Scatterplot Options: Fit Line (Опції для діаграми розсіювання: наближена крива) (рис. 2.60);

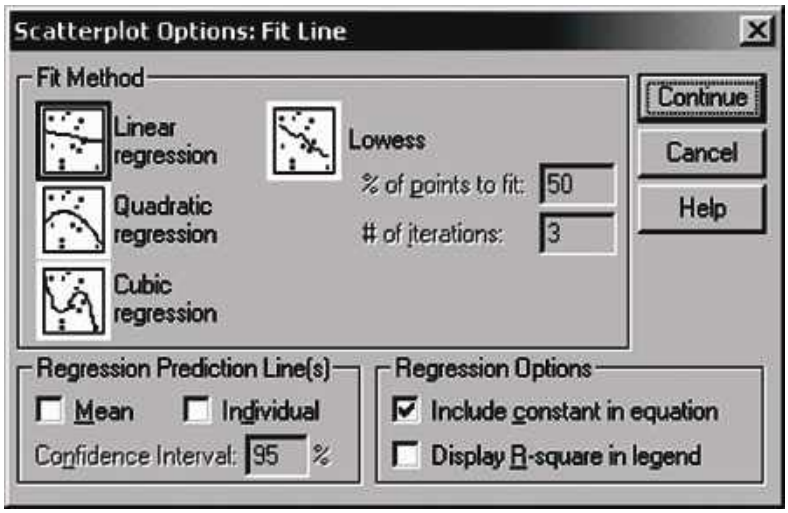


Рис. 2.60. Діалогове вікно Scatterplot Options: Fit Line (Опції для діаграми розсіювання)

– підтвердити попередню установку Linear Regression (Лінійна регресія) клацанням Continue (Далі) і потім на ОК;

– закрити редактор діаграм і клацнути один раз де-небудь поза графіком.

Тепер у діаграмі розсіювання відображається регресійна пряма (рис. 2.61).

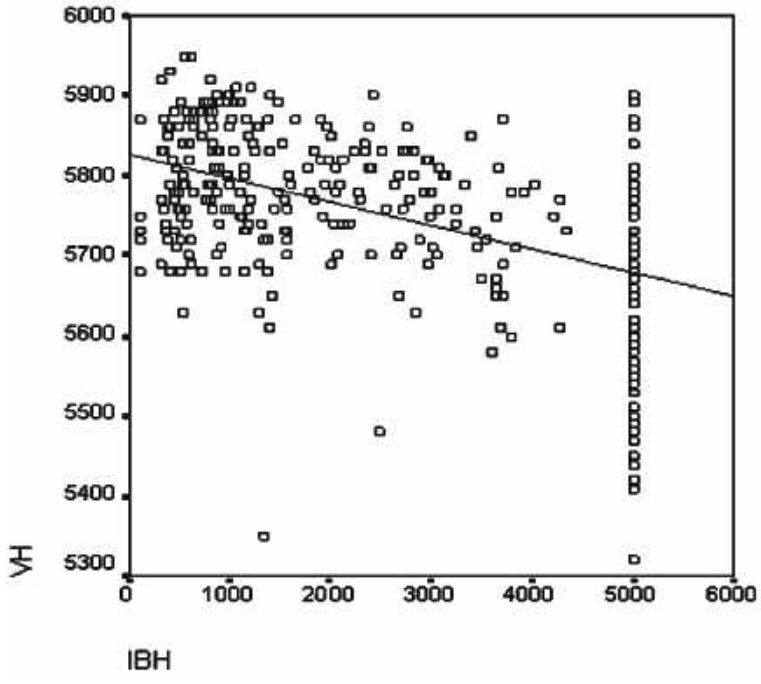


Рис. 2.61. Діаграма розсіювання з регресійної прямої

Щоб відкоригувати ось *X,* треба провести такі дії:

– двічі клацнути на графіку, і в меню редактора діаграм вибрати опції Chart (Діаграма) Axis (Осі), відкриється діалогове вікно Axis Selection (Вибір осі) (рис. 2.62);

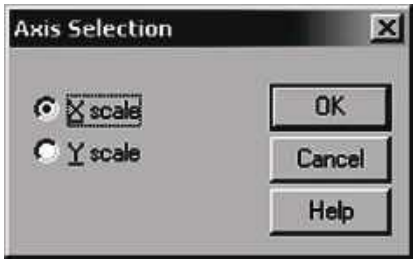


Рис. 2.62. Діалогове вікно Axis Selection (Вибір осі)

– підтвердити попередній вибір осі X натисканням кнопки ОК, відкриється діалогове вікно X-Scale Axis (Вісь X) (рис. 2.63);

– у полі, що редагується, Displayed (Відображуваний) у рубриці Range (Діапазон) змінити мінімальне значення на нуль;

– підтвердити натисканням на ОК;

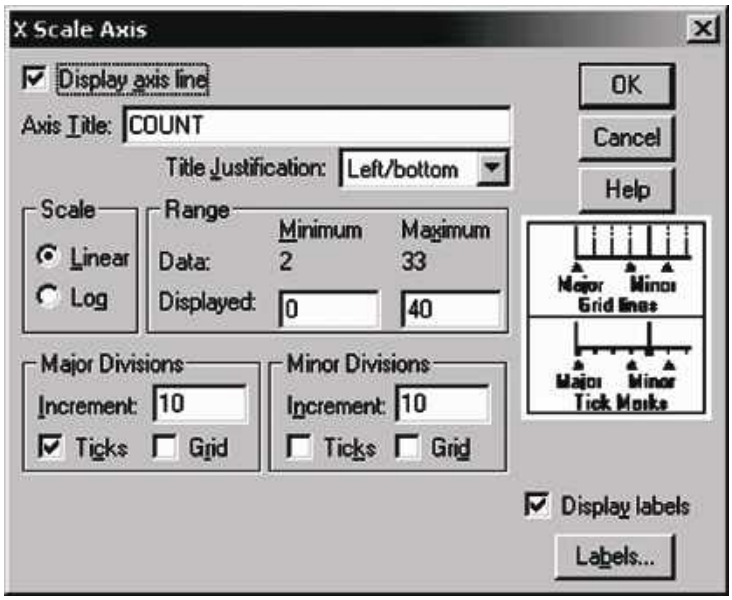


Рис. 2.63. Діалогове вікно X-Scale Axis (Вісь *X*)

– в меню редактора діаграм знову вибрати опції Chart... (Діаграма) Axis... (Осі);

– в діалоговому вікні Axis Selection (Вибір осі) активувати опцію *Y* Scale (Вісь *Y*); відкриється діалогове вікно Y-Scale Axis (Вісь *Y*);

– у рубриці Range (Діапазон) у полі Displayed (Відображуваний), що редагується, змінити мінімальне значення на нуль;

– підтвердити натисканням на ОК.

У вікні перегляду виникне відкоригована діаграма розсіювання (рис. 2.64).

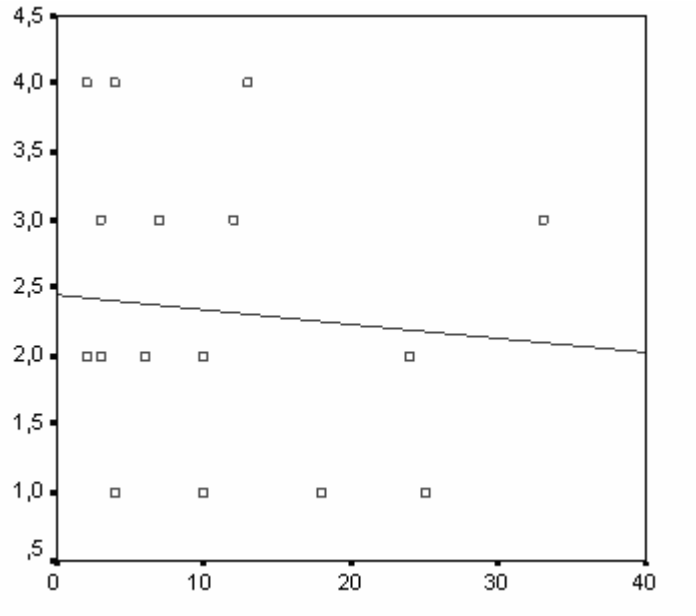


Рис. 2.64. Діаграма розсіювання з регресійної прямої після коригування осей

2.11.2. Множинна лінійна регресія

Взагалі в регресійному аналізі розглядають декілька незалежних змінних. Це, звичайно ж, завдає шкоди наочності одержуваних результатів, тому що подібні множинні зв'язки зрештою стає неможливо подати графічно.

У випадку множинного регресійного аналізу мова йде про необхідність оцінювання коефіцієнтів рівняння:

$$Y = b_1x_1 + b_2x_2 + ... + b_nx_n + a,$$

де n – кількість незалежних змінних, позначених як x_1, … ,x_n,

a – деяка константа.

Змінні, оголошені незалежними, можуть самі корелювати між собою. Цей факт необхідно обов'язково враховувати при визначенні коефіцієнтів рівняння регресії для того, щоб уникнути помилкових кореляцій.

Для проведення множинного регресійного аналізу необхідно:

– відкрити файл;

– вибрати в меню Analyze (Аналіз) Regression (Регресія) Linear (Лінійна);

– помістити залежну змінну в поле для залежних змінних, оголосити незалежні змінні незалежними.

Для множинного аналізу з декількома незалежними змінними не рекомендується залишати метод включення всіх змінних, установлений за замовчуванням. Цей метод відповідає одночасному обробленню всіх незалежних змінних, вибраних для аналізу, і тому його можна рекомендувати для використання тільки у випадку простого аналізу з однією незалежною змінною. Для множинного аналізу слід вибрати один з покрокових методів.

При прямому методі незалежні змінні, які мають найбільші коефіцієнти часткової кореляції із залежною змінною покроково додаються до регресійного рівняння.

При зворотному методі починають з результату, що містить усі незалежні змінні й потім виключають незалежні змінні з найменшими частковими кореляційними коефіцієнтами, доки відповідний регресійний коефіцієнт не виявляється незначущим (у цьому випадку рівень значущості дорівнює 0,1).

Найпоширенішим є покроковий метод, аналогічний прямому. Однак після кожного кроку змінні, які використовуються в цей момент, досліджуються за зворотним методом. При покроковому методі можуть задаватися блоки незалежних змінних; у цьому випадку задані на одному кроці блоки обробляються спільно.

При виборі покрокового методу не треба вживати блокову форму уведення даних, не задавати більше ніяких додаткових розрахунків і почати обчислення натисканням кнопки ОК, після чого отримаємо зведену таблицю моделі, яка містить R Square (Коефициент детермінації), Adjusted R Square (Скоригований R-Квадрат), Std. Error of the Estimate (Стандартну помилку оцінки).

Для кожного кроку здійснюється виведення коефіцієнтів множинної регресії, міри визначеності, зміщеної міри визначеності й стандартної помилки.

До зазначених результатів пошагово приєднують результати розрахунку дисперсії. Так само покроково здіснюється виведення відповідних коефіцієнтів регресії.

Крім того, для кожного кроку аналізуються вилучені змінні.

З допомогою відповідних опцій можна вивести велику кількість додаткових статистичних характеристик і графіків. Можна також створити багато додаткових змінних і додати їх у вихідний файл даних.

Важливим моментом є аналіз відхилень спостережуваних значень від теоретично очікуваних. Залишки мають виникати випадково (тобто не систематично) і підпорядковуватися нормальному розподілу. Це можна перевірити, якщо з допомогою кнопки Charts... (Діаграми) побудувати гістограму відхилень (рис. 2.65).

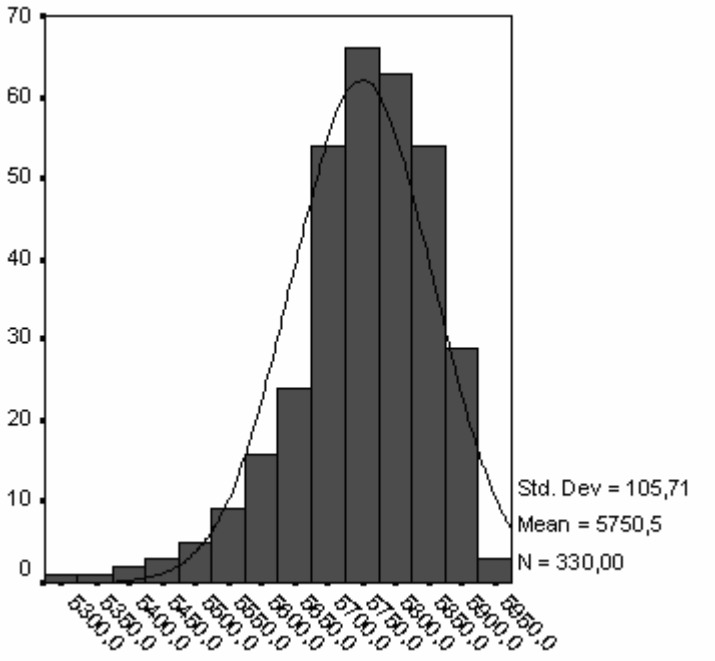


Рис. 2.65. Гістограма залишків

Перевірку на наявність систематичних зв'язків між відхиленнями сусідніх випадків (що, однак, є доречним тільки за наявності так званих даних з поздовжнім перетином) можна зробити з допомогою тесту Дарбіна–Ватсона (Durbin–Watson) на автокореляцію. Цей тест обчислює коефіцієнт, який має значення від 0 до 4. Якщо значення цього коефіцієнта дорівнює майже 2, то це означає, що автокореляції немає. Тест Дарбіна–Ватсона можна активувати через кнопку Statistics (Статистичні характеристики).

Ще однією додатковою можливістю є завдання змінної відбору в діалоговому вікні Linear Regression (Лінійна регресія). Тут з допомогою кнопки Rule... (Правило), використовуючи виборчу ознаку, у діалоговому вікні Linear Regression: Define Selection Rule (Лінійна регресія: уведення умови відбору) можна сформулювати умову, що буде обмежувати кількість випадків, долучених до аналізу.

2.11.3. Нелінійна регресія

Багато зв'язків за своєю природою, тобто в реальному житті, є або строго лінійними, або їх можна звести до лінійного виду.

Нелінійні зв'язки, які з допомогою відповідних трансформацій можна перевести в лінійний зв'язок, називають лінійними по суті (Intrinsically Linear Model). Можливість перекладу в лінійну модель потрібно використовувати завжди, тому що в цьому випадку параметри регресії обчислюються безпосередньо, а не визначаються з допомогою ітерацій.

Загального універсального методу визначення параметрів нелінійного зв'язку, на жаль, не існує, тому описана нижче послідовність дій може бути тільки прикладом.

Відкрийти файл.

Вибрати у меню Analyze (Аналіз) Regression (Регресія) Nonlinear (Нелінійна).

У діалоговому вікні Nonlinear Regression (Нелінійна регресія) перенести змінну-результат у поле для залежних змінних.

Клацнути на поле Model Expression (Модельне вираження) і внести у нього наступну формулу нелінійного рівняння. При введенні формули можна використовувати клавіатуру, що знаходиться в діалоговому вікні (рис. 2.66).

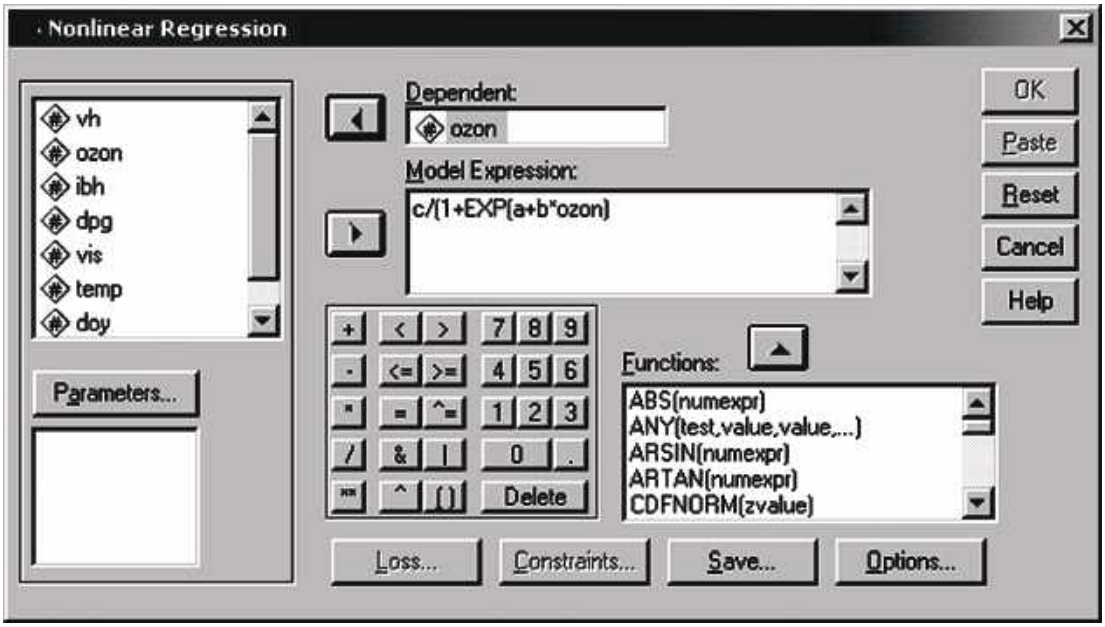


Рис. 2.66. Діалогове вікно Nonlinear Regression (Нелінійна регресія).

Задати початкові значення параметрів нелінійної регресії, яку необхідно визначити. Для цього клацнути на кнопці Parameter... (Параметр). Одержуємо діалогове вікно, у якому можна задавати початкові значення.

Указати в полі імен ім'я першого параметра, наприклад *a*, потім клацнути у поле Starting value (Початкове значення), увести початкове значення для *a* і клацнути на Add (Додати).

Зробити те саме з іншими параметрами.

Покинути діалогове вікно натисканням кнопки «Далі».

Клацнути на кнопці Save (Зберегти). Відмітити в діалоговому вікні Nonlinear Regression: Save New Variables (Нелінійна регресія: Зберегти нові змінні) параметри: Predicted Values (Прогнозовані значення) і Residuals (Залишки). Таким чином, буде створено нові змінні, які міс-

тять обчислені значення параметрів.

Почати розрахунок натисканням кнопки ОК.

На екрані виникнуть результати, причому можна помітити, що виведення відбувається не у вигляді звичних сучасних таблиць, а спочатку протоколюється процес.

Якщо необхідно візуально зрівняти розраховані значення зі спостережуваними, то можна з допомогою меню Graph (Графіки) Scatter plots (Діаграми розсіювання) побудувати багатошарову діаграму розсіювання (Staggered), на якій навести незалежні змінні.

Відповідно до попередніх установок при розрахунку нелінійної регресії відбувається мінімізація суми квадратів відхилень. З допомогою кнопки Loss...(Решта) можна задати яку-небудь іншу мінімізуючу функцію. Далі з допомогою кнопки Constraints... (Обмеження) можна відкрити вікно, у якому задати обмеження для обумовлених параметрів нелінійної регресії.

2.12. Дисперсійний аналіз

2.12.1. Дисперсійний аналіз

З допомогою дисперсійного аналізу досліджують вплив однієї або декількох незалежних змінних на одну залежну (одновимірний аналіз) або на декілька залежних змінних (багатовимірний аналіз). У звичайному випадку незалежні змінні приймають тільки дискретні значення (і належать до номінальної або порядкової шкали); у цій ситуації також говорять про факторний аналіз. Якщо ж незалежні змінні належать до інтервальной шкали або до шкали відношень, то їх називають коваріаціями, а відповідний аналіз – коваріаційним.

У межах дисперсійного аналізу SPSS пропонує множину можливостей. По-перше, потрібно відзначити, що дисперсійний аналіз може виконуватися в межах двох підходів:

– з допомогою традиційного класичного методу за Фишером (Fisher);

– з допомогою методу узагальненої лінійної моделі.

Перший підхід зводиться до розкладання за методом найменших квадратів (МНК); в однофакторному випадку сукупна дисперсія всіх спостережуваних значень розкладається на дисперсію всередині окремих груп і дисперсію між групами. В основі узагальненої лінійної моделі, навпаки, лежить кореляційний або регресійний аналіз.

Після відкриття відповідного файла дисперсійний аналіз можна викликати з допомогою меню Analyze (Аналіз) General Linear Model (Загальна лінійна модель). Відкриється допоміжне меню (рис. 2.67)

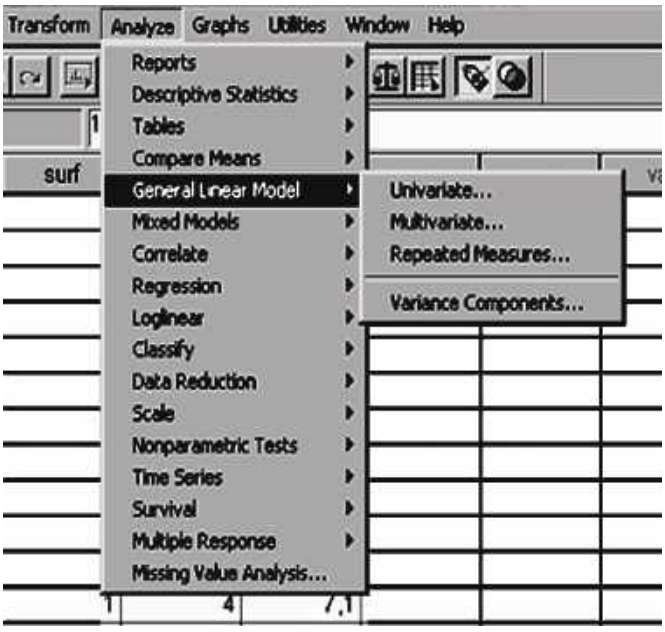


Рис. 2.67. Допоміжне меню General Linear Model
(Загальна лінійна модель)

З використанням усіх без винятку можливостей, пропонованих в діалоговому вікні, припускається проведення розрахунків на основі загальної лінійної моделі. За допомогою цього меню можна провести:

– одновимірний дисперсійний аналіз (Univariate);

– багатовимірний дисперсійний аналіз (Multivariate);

– багатовимірний дисперсійний аналіз із урахуванням повторних вимірів (Repeated Measures);

– розрахунок компонентів дисперсії (Variance Components).

Можливе також проведення дисперсійного аналізу за традиційним класичним методом Фішера. Однак такий аналіз можна виконати тільки за рахунок використання програмного синтаксису (процедура ANOVA).

З допомогою декількох прикладів зробимо загальний огляд і викладемо зауваження для основних ситуацій, до яких належать:

– одновимірний аналіз;

– коваріаційний аналіз;

– багатовимірний аналіз.

2.12.2. Одновимірний дисперсійний аналіз

Однофакторний дисперсійний аналіз (без повторних вимірів і з повторними вимірами) уже розглядався, тому відразу звернемося до багатофакторного дисперсійного аналізу.

Одновимірний дисперсійний (загальний багатофакторний) аналіз. Дослідимо вплив двох факторів на результуючу величину показника, з яких один розділено на дві категорії (наприклад, стать – жіноча й чоловіча), а другий – на три. Комбінації цих двох факторів утворять у цілому шість груп, що випробовуються (які також називають комірками). Кількість спостережень, що належать до окремих комірок, є неоднаковою.

Відкрити файл.

Вибрати у меню Analyze (Аналіз) General Linear Model (Загальна лінійна модель) Univariate (Одновимірна) Відкриється діалогове вікно Univariate (Одновимірна) (рис. 2.68).

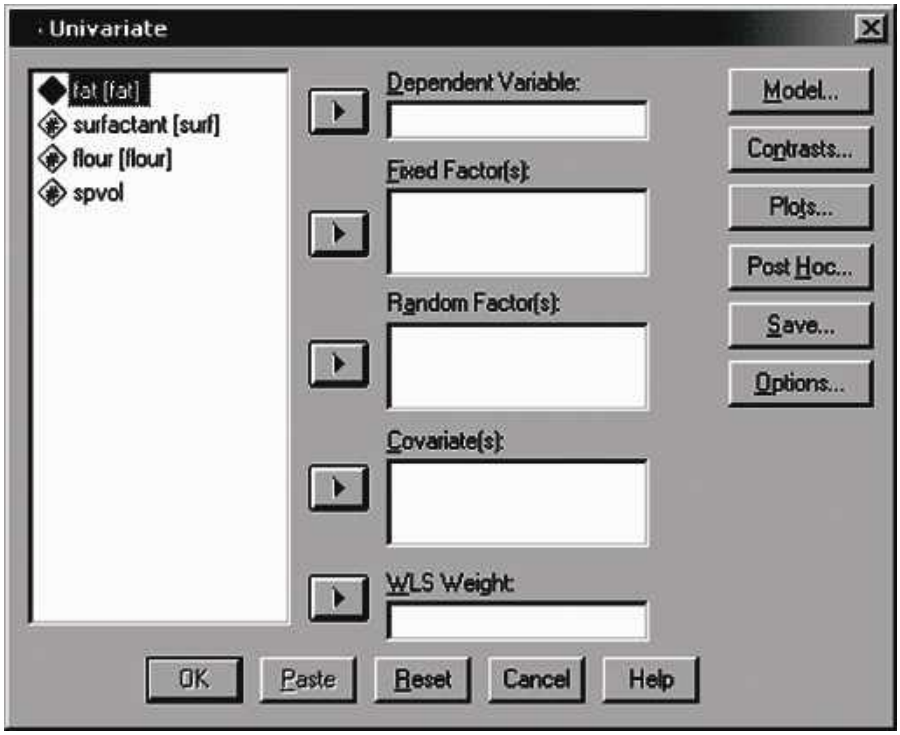


Рис. 2.68. Діалогове вікно Univariate (Одновимірна)

Перенести змінну в поле залежних змінних, а дві змінні-фактори – у поле фіксованих факторів.

Поняття «фіксовані» й «випадкові» фактори потребують додаткового пояснення. Фіксованими факторами, або факторами з фіксованими ефектами, називають такі фактори, які охоплюють усі можливі класифікаційні шари однієї незалежної змінної, наприклад, стать чоловіча – жіноча або освіта початкова – середня – вища. Однак якщо шари (підпопуляції) фактора вибираються випадково з нескінченної множини можливих підпопуляцій факторів, що мають назву генеральної популяції, то говорять про фактори з випадковими ефектами. У цьому випадку є доречним компонентний аналіз, тобто розрахунок так званих компонентів дисперсії.

Клацнути по кнопці Model... (Модель).

Відкриється діалогове вікно Univariate: Model (Одновимірна: Модель) (рис. 2.69).

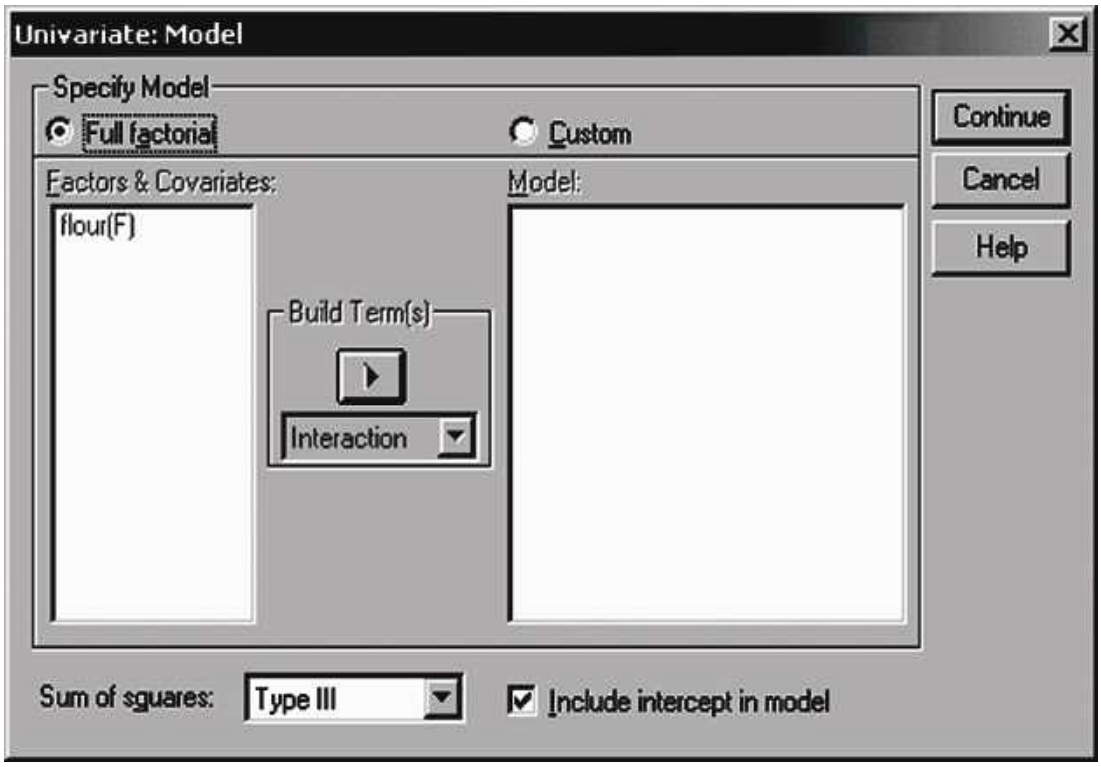


Рис. 2.69. Діалогове вікно Univariate: Model (Одновимірна: Модель)

Модель дисперсійного аналізу – це математичне співвідношення, у якому кожна змінна наведена у вигляді суми середнього значення й помилки. За замовчуванням установлено повнофакторну модель (Full factorial). У цій моделі середнє значення кожного спостереження наведено у вигляді генерального середнього й суми внеску всіх головних «ефектів» (факторів впливу). Також виконується розрахунок усіх взаємодій між факторами. Альтернативою є можливість вибору окремих взаємодій факторів впливу, що здійснюється з допомогою активування опції Custom (Користувацький режим). Таким самим чином мають бути відібрані й взаємодії з коваріаціями.

Для формування сум квадратів для МНК існує чотири різних підходи (чотири типи, позначених за допомогою римських цифр I, II, III і IV), за замовчуванням установлено тип III.

Залишити у цьому вікні всі установки за замовчуванням і покинути діалогове вікно натисканням кнопки Continue (Далі).

Клацнути на вимикачі Options (Опції)

Відкриється діалогове вікно Univariate: Options (Одновимірна: Опції) (рис. 2.70)

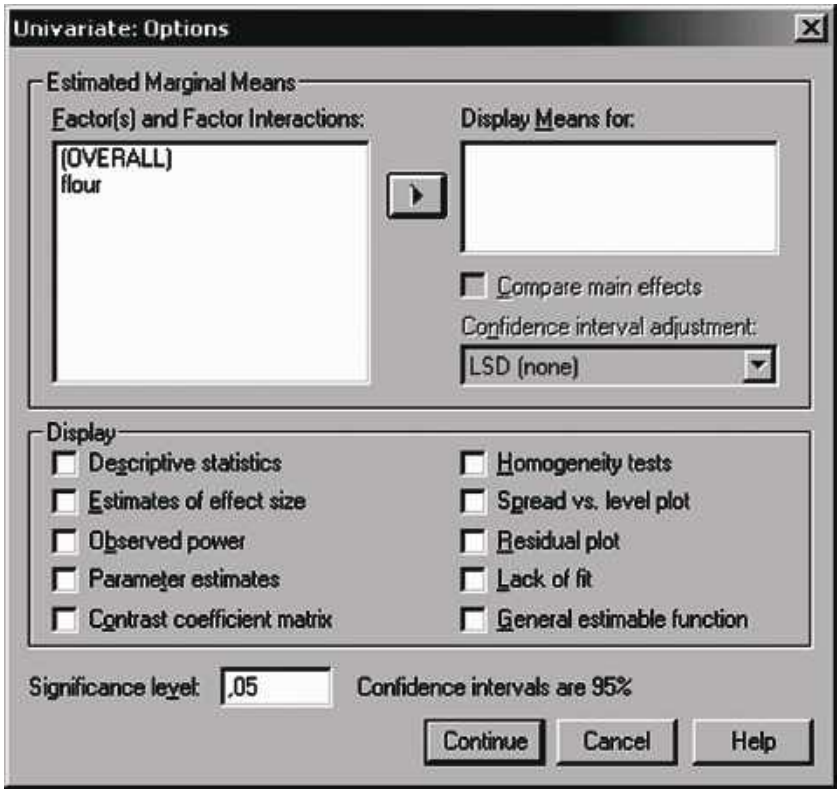


Рис. 2.70. Діалогове вікно Univariate: Options (Одновимірна: Опції)

Перенести OVERALL (У цілому) і обидві змінні-фактори в поле Display means for (Показати середні значення для); у цьому випадку як результати буде виведено середні значення й стандартна помилка для сукупної вибірки (OVERALL) і для всіх шарів за обома факторами. Середні значення для комбінацій взаємодії на цьому етапі розраховуються тільки для неповнофакторних моделей.

Активувати Descriptive Statistics (Дескриптивні статистики); з допоїмогою цієї опції виводяться середнє значення, стандартні відхилення й кількість спостережень у всіх комірках.

Активувати потім опцію Homogeneity tests (Тести на однорідність). У такий спосіб активується перевірка однорідності дисперсії. Покинути діалогове вікно натисканням Continue (Далі).

За допомогою вимикача Plots... (Діаграми) відкрити діалогове вікно Univariate: Profile Plots (Одновимірна: Профільні діаграми) (рис. 2.71).

У випадку профільних діаграм мова йде про графічне подання середніх значень шарів обраних факторів у вигляді лінійчатих діаграм. При цьому шари другого фактора відповідно можуть бути використані для відображення другої лінії. У такий спосіб можна наочно зобразити взаємодію між двома факторами.

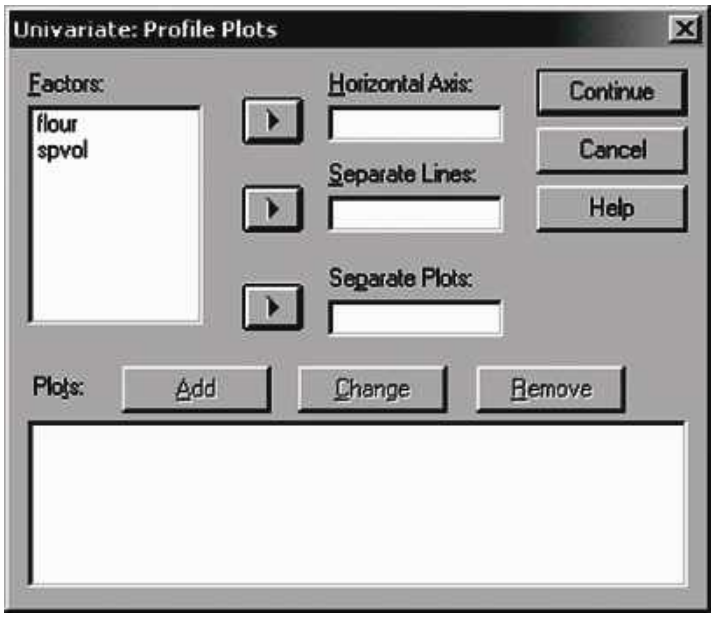


Рис. 2.71. Діалогове вікно Univariate: Profile Plots
(Одновимірна: Профільні діаграми)

Помістити першу змінну-фактор у поле Horizontal Axis (Горизонтальна вісь), а другу – в поле Separate Lines (Окремі лінії). Також можна вказувати додаткову змінну й у поле Separate Plots (Окремі графіки); тоді для окремих шарів цієї змінної будуть побудовані окремі діаграми.

Клацнути на вимикачі Add (Додати) і покинути діалогове вікно натисканням Continue (Далі).

На закінчення клацнути на вимикачі Post Hoc... (Додатковий тест). Відкриється діалогове вікно Univariate: Post Hoc Multiple Comparisons for Observed Means (Одновимірна: Додатково-множинні порівняння для спостережуваних середніх значень). Виникне можливість вибрати один або декілька з вісімнадцяти тестів, необхідних для проведення додаткового порівняння окремих шарів вибраних факторів. Однак це має сенс тільки для факторів з більш ніж двома шарами.

Помістити змінну-фактор у поле Post Hoc Tests for… (Додаткові тести для…).

Активувати тест Шеффе (Scheffe). Тепер діалогове вікно має такий вигляд, як зображено на рис. 2.72.

Покинути діалогове вікно натисканням Continue (Далі).

Тепер можна визначити контрасти й для кожного спостереження зберегти деякі статистичні характеристики як нові змінні. Почати розрахунок натисканням ОК.

У вікні спочатку виникне зведена таблиця, озаглавлена як Between-Subjects Factors (Міжсуб‘єктні фактори). Потім буде виведен-

ня середніх значень, стандартних відхилень і кількості спостережень для окремих комірок, а також результати тесту на однорідність.

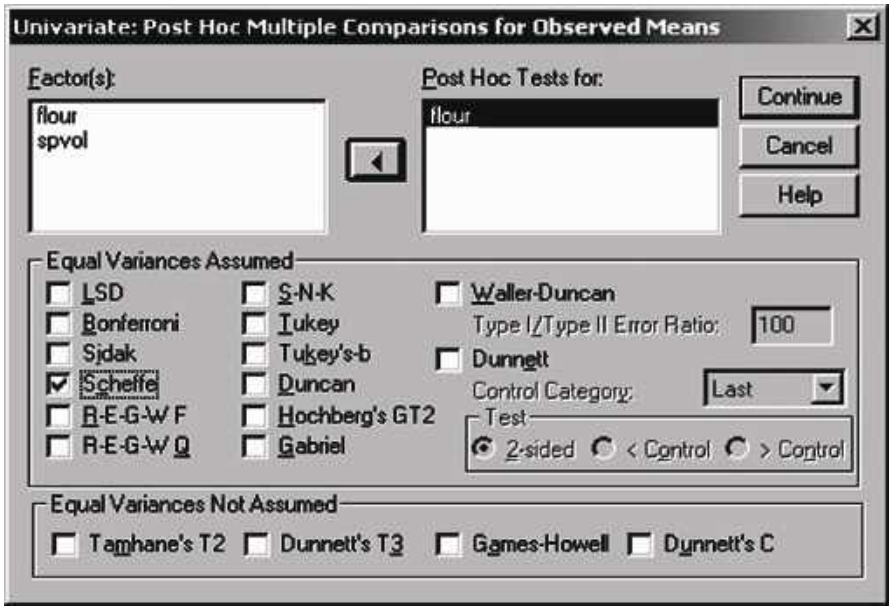


Рис. 2.72. Діалогове вікно Univariate: Post Hoc Multiple Comparisons for Observed Means (Одновимірна: Додатково-багаторазові порівняння для спостережуваних середніх значень)

Завершує виведення результатів профільна діаграма, у якій наведено лінійну діаграму однієї зміної, наприклад, віку окремо для кожної категорії другої змінної.

Одновимірний дисперсійний аналіз за методом Фішера (Fisher)

Проведемо аналіз з допомогою традиційного «класичного» методу Фішера. Оскільки, починаючи з 8.0 версії програми, цей вид аналізу вже не виводиться в діалогове вікно, то доведеться скористатися програмним синтаксисом (процедура ANOVA).

Відкрити файл.

Вибрати в меню File (Файл) New (Новий) Syntax (Синтаксис). Набрати в поле редактора синтаксису команду

```
ANOVA VARIABLES=ml BY geschl (1,2) alter (1,3)
  /STATISTICS MCA MEAN
  /METHOD EXPERIM.
```

У SPSS пропонується три методи для розкладання квадратів відхилення в МНК для випадку, коли обсяг окремих комірок (кількості спостережень, що належать до даної комірки) неоднаковий. При такому «незбалансованому компонуванні», що часто виникає при «неплановних» (неекспериментальних) дослідженнях, без подальшого оброблення неможна до загальної суми додавати суми квадратів окремих ефектів. Можна вибрати один із таких методів оброблення:

– UNIQUE – внесок кожного з факторів впливу розглядається одночасно; кожний з них розраховується за умови збереження постійного значення всіх інших. Оскільки в цьому випадку можна зробити неявне припущення про можливе існування причинного зв'язку між факторами, то цей варіант варто вибирати тоді, коли не повинне проводитися вагове порівняння значеннь окремих факторів. Цей метод установлюється за замовчуванням;

– HIERARCHICAL – черговість розрахунку ефектів визначається черговістю вибраних факторів. Цей метод варто застосовувати тоді, коли можна заздалегідь припустити ієрархічну впорядкованість факторів;

– EXPERIMENTAL – ефекти обробляються в такий послідовності: ефекти коваріацій, головні ефекти, взаємодії в порядку зростання. При розрахунку одного ефекта виконується обчислення всіх попередніх ефектів і ефектів, що знаходяться на тому ж рівні.

При однакових обсягах комірок («ортогональне компонування») усі три методи дають однакові результати.

Використовуючи допоміжну команди STATISTICS, можна організувати виведення таких даних:

– Mean – виводяться середні значення й кількість спостережень для сукупної популяції, окремих шарів фактора й кожної комірки. Однак якщо вибирається метод UNIQUE для розкладання суми квадратів у МНК, то ця опція стає недоступною;

– MCA (Множинний класифікаційний аналіз) – з допомогою спеціальних коефіцієнтів η і β відображається сила зв'язку між окремим фактором і залежною змінною. Це є доречним, якщо не спостерігається ніяких значущих взаємодій. Виведення результатів MCA є недоступним за умови вибору методу UNIQUE.

Запустити команду ANOVA на виконання клацанням на знаку Run Current (Запустити синтаксис).

Після звичайної зведеної таблиці оброблюваних спостережень спочатку виводяться середні значення й частоти (відповідні результати виведення тут не наводяться), потім зведення дисперсійного аналізу із сумами квадратів, ступенями свободи, середніми значеннями сум квадратів і т.д.

Обидва коефіцієнти η і β є мірою сили зв'язку (кореляції) між відповідним фактором і залежними змінними. Коефіцієнт β має приватну природу й характеризує силу зв'язку за відсутності впливів з боку інших факторів. Значна відмінність коефіцієнтів η і β один від одного свідчить про наявність взаємозв'язку між факторами. І, нарешті, величина «R Squared» («R-квадрат») свідчить про той ступінь відхилення від сукупної дисперсії, який можна пояснити головними ефектами.

2.13. Факторний аналіз

Факторний аналіз – це процедура, з допомогою якої велику кількість змінних стосовно наявних спостережень зводять до меншої кількості незалежних впливових величин, які називають факторами. При цьому в один фактор поєднуються змінні, що сильно корелюють між собою. Змінні з різних факторів слабко корелюють між собою. Таким чином, метою факторного аналізу є знаходження таких комплексних факторів, які більш повно пояснюють спостережувані зв'язки між наявними змінними.

На першому кроці процедури факторного аналізу відбувається стандартизація заданих значень змінних (*z*–перетворення); потім з допомогою стандартизованих значень розраховують кореляційні коефіцієнти Пірсона між розглянутими змінними.

Вихідним елементом для подальших розрахунків є кореляційна матриця. Для побудованої кореляційної матриці визначаються так звані власні значення й відповідні їм власні вектори. Щоб їх визначити, використовують оцінні значення діагональних елементів матриці (так звані відносні дисперсії простих факторів).

Власні значення сортуються в порядку убування, для чого зазвичай відбирається стільки факторів, скільки є власних значень, що перевершують за величиною одиницю. Власні вектори, що відповідають цим власним значенням, утворять фактори. Елементи власних векторів одержали назву факторного навантаження. Їх можна розуміти як коефіцієнти кореляції між відповідними змінними й факторами. Для розв’язання такої задачі визначення факторів було розроблено численні методи, з яких найбільш часто вживається метод визначення головних факторів (компонентів).

Описані вище кроки розрахунку ще не дають однозначного розв’язку задачі визначення факторів. Ґрунтуючись на геометричному зображенні цієї задачі, пошук однозначного розв’язку називають задачею обертання факторів. І тут є велика кількість методів, з яких найбільш часто вживається ортогональне обертання за так званим методом варімакса. Факторні навантаження оберненої матриці можуть розглядатися як результат виконання процедури факторного аналізу. Крім того, на основі значень цих навантажень необхідно спробувати дати тлумачення окремим факторам.

Якщо фактори знайдено й витлумачено, то на останньому кроці факторного аналізу окремим спостереженням можна присвоїти значення цих факторів, так звані факторні значення. У такий спосіб для

кожного спостереження значення великої кількості змінних можна перевести в значення невеликої кількості факторів.

Розглянемо процедуру факторного аналізу.

Відкрити файл.

Вибрати в меню Analyze (Аналіз) Data Reduction (Скорочення обсягу даних) Factor (Факторний аналіз).

Відкриється діалогове вікно Factor Analysis (Факторний аналіз).

Змінні помістити в поле тестованих змінних.

Після клацання по кнопці Descriptive Statistics (Дескриптивні статистики) залишити виведення первинних результатів, які містять первинні відносні дисперсії простих факторів, власні значення й відсоткові частки поясненої дисперсії. Досить часто буває необхідним також виведення одновимірних статистик і кореляційних коефіцієнтів.

З допомогою кнопки Extraction... (Відбір) можна вибрати метод відбору; залишити аналіз головних компонентів, установлений за замовчуванням. Кількість відібраних у цьому випадку факторів прирівнюється до кількості власних значень, що перевищують одиницю. Також є можливість власноручно зазначити цю кількість.

Вимикач Rotation (Обертання) дає можливість вибрати метод обертання. Активувати метод варімакса й залишити активованим виведення оберненої матриці факторів. Далі можна організувати виведення факторних навантажень у графічному вигляді, у якому перші три фактори буде подано в тривимірному просторі; у випадку наявності тільки двох факторів у шарі наводиться тільки одне зображення.

Якщо необхідно знайти значення факторів і зберегти їх у вигляді додаткових змінних, треба задати вимикач Scores (Значення) і відзначити Save as variables (Зберегти як змінні). За замовчуванням установлено регресійний метод. Вимикач Options (Опції) призначено для оброблення пропущених значень. Тут забезпечується можливість замінити пропущені значення середніми значеннями відповідних змінних.

Для проведення розрахунків клацнути на ОК.

У вікні перегляду виникнуть результати.

Тут треба спробувати пояснити відібрані фактори. Для цього олівцем в кожному рядку оберненої факторної матриці відмітити факторне навантаження, що має найбільше абсолютне значення.

Як уже було сказано, ці факторні навантаження слід розуміти як кореляційні коефіцієнти між змінними й факторами. У більшості випадків включення окремої змінної до одного фактора, що здійснюється на основі коефіцієнтів кореляції, є однозначним. У виняткових випадках змінна може належати до двох факторів одночасно. Можуть бути також і змінні, якими не можна навантажити жоден з відібраних факторів.

2.14. Кластерний аналіз

З допомогою кластерного аналізу й заздалегідь заданих змінних формуються групи спостережень. Під спостереженнями тут розуміються окремі особи (респонденти) або будь-які інші об'єкти. Члени однієї групи (одного кластера) повинні мати однакові прояви змінних, а члени різних груп – різни.

Разом з кластеризацією спостережень у SPSS передбачено кластеризацію змінних. Тут на основі заданих спостережень утворюються групи змінних. Майже те саме робить і факторний аналіз.

Головна схожість між кластерним і факторним аналізами полягає в тому, що обидва призначені для переходу від початкової сукупності безлічі змінних (або об'єктів) до істотно меншої кількості чинників (кластерів). Однак реалізація статистичних процедур та інтерпретація результатів для двох типів аналізів є різними.

Етапи кластерного аналізу:

1. Вибір змінних-критеріїв для кластеризации.
2. Вибір способу виміру відстані між об'єктами або кластерами (спочатку вважається, що кожен об'єкт відповідає одному кластеру).
3. Формування кластерів двома методами: методом злиття і методом дроблення.
4. Інтерпретація результатів.

2.14.1. Принцип кластерного аналізу

Для розгляду принципу кластерного аналізу виберемо спочатку дуже простий приклад.

Наведемо дві змінні *var1* й *var2* за допомогою простої діаграми розсіяння.

Вибрати в меню Graphs (Графіки) Scalier (Діаграма розсіяння).

Першу змінну помістити в поле осі *x*, а другу – в полі осі *y*.

Через кнопку Options... (Опції) активувати опцію Display Chart with case labels (Показувати графік з мітками спостережень).

Отримаємо діаграму розсіяння, показану на рис. 2.73.

На діаграмі точки утворюють чотири окремі виразні угрупування, два з яких – у нижній половині діаграми і два – у верхній. Отже, змінні, які досліджуються, явно розпадаються на чотири різні кластери за деяким показником.

Показники, які за значеннями двох розглянутих змінних є однаковими, належать до одного кластера; показники, що знаходяться в різних кластерах, не є однаковими. Вирішальним критерієм для ви-

значення однаклвості й відмінності двох показників є відстань між точками на діаграмі розсіяння, що відповідають цим показникам.

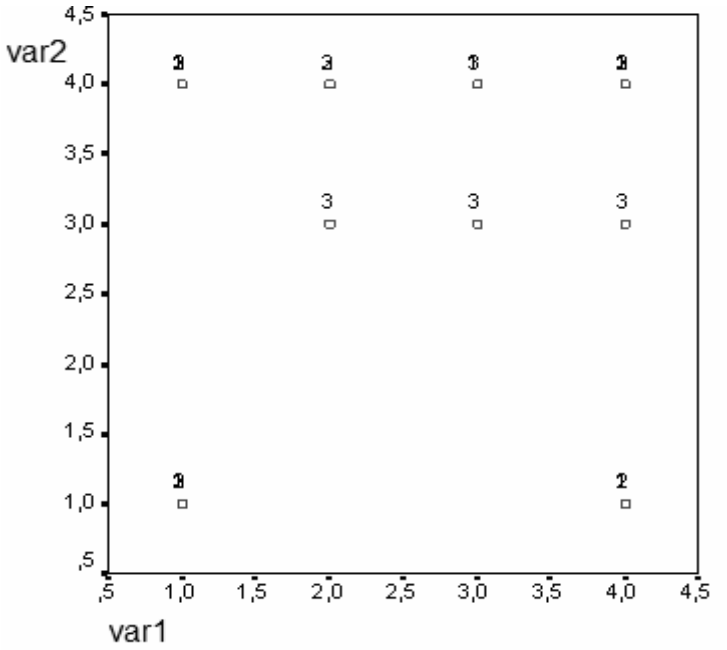


Рис. 2.73. Діаграма розсіяння двох змінних *var1* й *var2*

Найпоширенішою мірою для визначення відстані між двома точками на площині, яку утворено координатними осями *x* і *y*, є міра Евкліда:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

де x_1 і x_2 – координати першої точки, y_1 і y_2 – координати другої точки.

Таким чином, з допомогою діаграми розсіяння для двох змінних було проведено найпростіший кластерний аналіз. Було вибрано такий вид графічного подання, з допомогою якого можна було б виразно розпізнати групування в кластери (чотири в цьому випадку).

На жаль, така виразна картина відношень між змінними трапляється дуже рідко. По-перше, структури кластерів, якщо взагалі такі є, не так чітко розділені, особливо за наявності великої кількості спостережень. Скоріше навпаки, кластери є розмитими й навіть проникають один в одного. По-друге, кластерний аналіз проводиться, як правило, не з двома, а з набагато більшою кількістю змінних.

При кластерному аналізі з трьома змінними можна ввести ще одну вісь (вісь *z*) і розглядати розміщення спостережень, а також проводити обчислення відстані за формулою міри Евкліда в тривимірному просторі.

За наявності більш ніж трьох змінних визначення відстані між дво-

ма точками *x* і *y* в будь-якому *n*-вимірному просторі для математиків не викликає особливих утруднень. Формула Евкліда в таких випадках набуває вигляду

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Разом з мірою відстані Евкліда, у SPSS пропонуються й інші дистанційні заходи, а також заходи подібності. Отже, кластерний аналіз можна проводити не лише зі змінними, що належать до інтервальної шкали, але й, наприклад, із дихотомічними змінними. У такій ситуації застосовуються вже інші дистанційні заходи і заходи подібності.

У програмі SPSS реалізуються три методи кластерного аналізу:

1) двоетапний (Two-Step);
2) *k*-середніх (*K*-means);
3) ієрархічний (Hierarchical).

При проведенні кластерного аналізу окремі кластери можуть формуватися з допомогою покрокового злиття, для якого існує низка різних методів, які можна задіяти, якщо пройти через меню Analyze (Аналіз) Classify (Класифікувати). Їх поміщено в цьому меню під іменами Hierarchical Cluster (Ієрархічний кластер) і *K*-Means Cluster (Кластерний аналіз методом *k*-середніх).

2.14.2. Ієрархічний кластерний аналіз

В ієрархічних методах кожне спостереження утворює спочатку свій окремий кластер. На першому кроці два сусідні кластери об'єднуються в один; цей процес може тривати доти, доки не залишаться тільки два кластери. У методі, який у SPSS встановлено за замовчуванням (Between-groups linkage (Зв'язок між групами)), відстань між кластерами є середнім значенням усіх відстаней між усіма можливими парами точок з обох кластерів.

2.14.2.1. Ієрархічний кластерний аналіз із двома змінними

Вибрати в меню Analyze (Аналіз) Classify (Класифікувати) Hierarchical Cluster (Ієрархічний кластерний аналіз). Виникне діалогове вікно Hierarchical Cluster Analysis (рис. 2.74).

Змінні, що досліджуються, помістити в поле тестованих змінних, а текстову змінну-показник – в поле з ім'ям Label cases by: (Найменування (мітки) спостережень:).

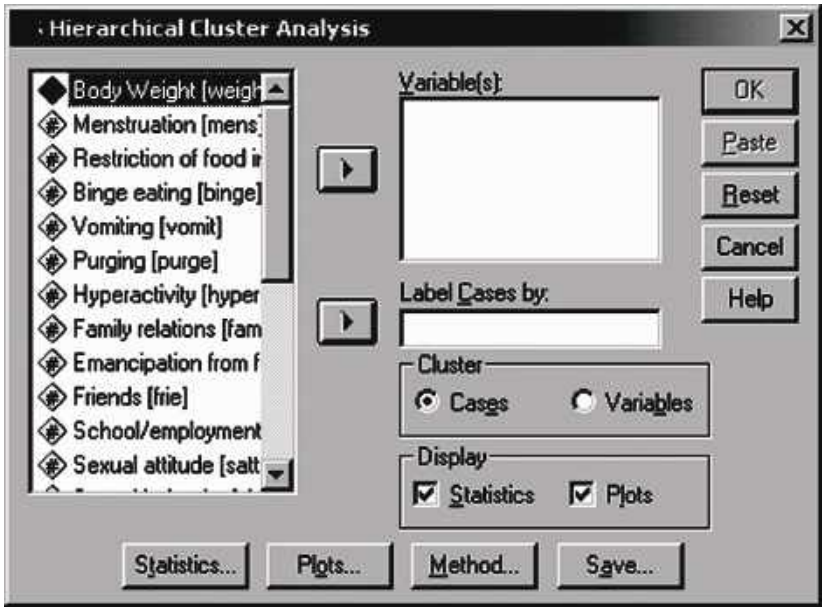


Рис. 2.74. Діалогове вікно Hierarchical Cluster Analysis (Ієрархічний кластерний аналіз)

Клацанням по вимикачу Statistics… (Статистики) відкрити діалогове вікно Hierarchical Cluster Analysis: Statistics (Ієрархічний кластерний аналіз: Статистики) і разом з виведенням послідовності злиття (Agglomeration schedule) активувати виведення показника належності до кластера для кожного спостереження.

Повернувшись в головне діалогове вікно, клацнути по вимикачу Plots… (Діаграми). Активувати опцію виведення деревоподібної діаграми (Dendrogram) і з допомогою опції None (Ні) відмінити виведення накопичувальної діаграми.

З допомогою кнопки Method... (Метод) можна вибрати метод утворення кластерів, а також метод розрахунку дистанційної міри і міри подібності відповідно.

У SPSS пропонується сім різних методів об'єднання. Метод Between-groups linkage (Зв'язок між групами) встановлюється за замовчуванням.

Дистанційні міри і міри подібності залежать від виду змінних, таких, що беруть участь в аналізі, тобто вибір міри залежить від типу змінної і шкали, до якої вона належить: інтервальна змінна, частоти або бінарні (дихотомічні) дані. Якщо дані належать до інтервальної шкали, то для них за замовчуванням як дистанційну міру встановлюють квадрат відстані (Squared Euclidean distance) Евкліда.

Залишити попередні установки, і в полі Transform Values (Перетворювати значення) встановити *z*-претворення (стандартизацію) значень.

Повернутися назад в головне діалогове вікно і почати розрахунок натисненням кнопки ОК.

Після звичайного загального статистичного зведення підсумків за спостереженнями у вікні перегляду спочатку наводиться огляд належності, з якого можна з'ясувати черговість будування кластерів, а також їхню оптимальну кількість. За двома колонками, розташованими під загальною шапкою Cluster Combined (Об'єднання в кластери), можна побачити, які спостереження було об'єднано на першому кроці, тому що вони максимально однакові і віддалені один від одного на дуже малу відстань. Далі показано, яке об'єднання спостережень відбувається на наступному кроці й т. д.

Щоб визначити, яку кількість кластерів слід було б вважати оптимальною, вирішальне значення має показник, що виводиться під заголовком «коефіцієнт». Під цим коефіцієнтом мається на увазі відстань між двома кластерами, яку визначено на основі вибраної дистанційної міри з урахуванням передбаченого перетворення значень. У цьому випадку – це квадрат відстані Евкліда, визначений з використанням стандартизованных значень. На цьому етапі, де міра відстані між двома кластерами збільшується стрибкоподібно, процес об'єднання в нові кластери необхідно зупинити, оскільки інакше були б об'єднані всі кластери, що знаходяться на відносно великій відстані один від одного. Оптимальною вважається кількість кластерів, що дорівнює різниці кількості спостережень і кількості кроків, після якої коефіцієнт збільшується стрибкоподібно.

Нарешті наводиться дендрограма, яка візуалізує процес злиття, наведений в оглядовій таблиці порядку агломерації у вигляді деревоподібної структури. З її допомогою ідентифікуються об'єднані кластери й значення коефіцієнтів на кожному кроці. При цьому відображаються не початкові значення коефіцієнтів, а значення, зведені до шкали від 0 до 25. Кластери, що утворюються внаслідок злиття, відображаються горизонтальними пунктирними лініями.

2.14.2.2. Ієрархічний кластерний аналіз з більш ніж двома змінними

Розглянемо приклад з області кадрової політики деякого підприємства. Вісімнадцять претендентів пройшли десять різних тестів в кадровому відділі. Максимальна оцінка, яку можна було отримати на кожному з тестів, становить 10 балів. Список тестів наведено в табл. 2.9.

Таблиця 2.9

Номер тесту	Предмет тесту
1	Пам'ять на числа
2	Математичні завдання
3	Винахідливість під час прямого діалогу
4	Тест на складання алгоритмів
5	Упевненість під час виступу
6	Командний дух
7	Винахідливість
8	Співпраця
9	Визнання в колективі
10	Сила переконання

Результати тесту зберігаються у файлі в змінних *t1 – t10*. У файлі знаходиться також і текстова змінна для характеристики того, кого тестують. З використанням результатів тесту відповідності проведемо кластерний аналіз, метою якого є виявлення груп кандидатів, схожих за своїми якостями.

Відкрити файл.

Вибрати в меню Analyze (Аналіз) Classify (Класифікувати) Hierarchical Cluster (Ієрархічний кластерний аналіз).

У діалоговому вікні Hierarchical Cluster Analysis (Ієрархічний кластерний аналіз) змінні *t1 – t10* помістити в поле тестованих змінних, а текстову змінну name (ім'я) використовувати для позначення (маркірування) спостережень.

Спершу має бути достатньо виведення оглядової таблиці порядку агломерації (табл. 2.10); не робити більше запиту на які-небудь дані й деактивувати виведення діаграм. Оскільки всі змінні в цьому прикладі мають однакові межі значень, стандартизація змінних є зайвою.

Значний стрибок коефіцієнта спостерігається після 14-го кроку; це означає, що для даних, що містять 18 спостережень, оптимальним є рішення з чотирма кластерами.

Після визначення оптимальної кількості кластерів організувати для кожного спостереження Виведення інформації про належність до кластера.

Для цього знову відкрити діалогове вікно Hierarchical Cluster Analysis (Ієрархічний кластерний аналіз) і клацнути по вимикачу Statistics (Статистики). У розділі Cluster Membership (Належність до кластера) активувати опцію Single solution (Одне рішення) і вказати бажану кількість кластерів (чотири).

Інформацію про належність кожного спостереження до певного кластера можна зберегти в новій змінній.

Таблиця 2.10

Порядок агломерації						
Stage (Крок)	Cluster Combined (Об'єднання в кластери)		Coefficients (Коефіцієнти)	Stage Cluster First Appears (Крок, на якому кластер вини-кає вперше)		Next Stage (Наступний крок)
	Cluster 1 (Кластер1)	Cluster 2 (Кластер 2)		Cluster 1 (Кластер 1)	Cluster 2 (Кластер 2)	
1	1	4	0,000	0	0	6
1	14	18	2,000	0	0	4
3	12	15	2,000	0	0	6
4	9	14	2,000	0	2	8
5	2	10	2,000	0	0	13
6	1	12	3,000	1	3	15
7	13	16	4,000	0	0	12
8	9	11	4,000	4	0	11
9	5	7	5,000	0	0	14
10	6	17	6,000	0	0	13
11	3	9	6,000	0	8	15
12	8	13	7,000	0	7	14
13	2	6	7,500	5	10	16
14	5	8	12,833	9	12	16
15	1	3	194,000	6	11	17
16	2	5	198,500	13	14	17
17	1	2	219,407	15	16	0

Активувати вимикач Save (Зберегти), активувати опцію Single solution (Одне рішення) і для зазначення бажаної кількості кластерів увести цифру 4. Тепер окрім таблиці порядку агломерації для кожного спостереження виводитиметься й інформація про належність до кластера.

Якщо розглянути дані в редакторі даних, то можна побачити, що додалася змінна *clu4_l*, що вказує на кластерну належність кожного спостереження і яку можна використати для розрахунку кластерного профілю.

Вибрати в меню Analyze (Аналіз) Compare Means (Порівняти середні значення) Means (Середні значення)

Змінним *t1 – t10* присвоїти статус залежних змінних, а змінній *clu4_1* – статус незалежної змінної й почати розрахунок. Як результати розрахунку виводяться середні значення й стандартні відхилення підсумків десяти тестів для чотирьох кластерів. Для зручності помістити середні значення в окрему табл. 2.11.

Таблиця 2.11

Предмет тесту	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Пам'ять на числа	10,00	10,00	4,20	4,80
Математичні завдання	10,00	10,00	4,80	4,40
Винахідливість під час прямого діалогу	9,00	4,25	10,00	4,00
Тест на складання алгоритмів	10,00	10,00	4,40	4,00
Упевненість під час виступу	10,00	4,75	10,00	4,20
Командний дух	9,50	4,50	4,40	10,00
Винахідливість	9,25	3,75	10,00	4,40
Співпраця	9,75	4,25	4,00	10,00
Визнання в колективі	10,00	4,25	3,80	10,00
Сила переконання	9,50	4,25	10,00	5,00

Ті люді, які тестуються, що мають дуже хороші показники в усіх тестах, належать до першого кластера. Це ті конкурсанти, які напевно пройшли б на завершальний відбірковий тур. У другий кластер включено тих, хто має хороші показники з математичних тестів (пам'ять на числа, математичні завдання, тест на складання алгоритмів), але із слабкими оцінками в соціальній компетентності й упевненості під час виступів. До третього кластера належать ті, хто впевнено себе почуває під час виступу, але мають слабкі показники з математичних тестів і соціальної компетентності. Нарешті, до четвертого кластера належать люди з високим рівнем соціальної компетентності, але із слабкими результатами з тестів на розв’язання математичних завдань і на силу переконання.

2.14.3. Кластерний аналіз при великій кількості спостережень (кластерний аналіз методом *k*-середніх)

Ієрархічні методи об'єднання, хоча й точні, але трудомісткі: на кожному кроці необхідно вибудовувати дистанційну матрицю для усіх поточних кластерів. Тому за наявності великої кількості спостережень застосовують інші методи. Недолік цих методів полягає в тому, що тут необхідно заздалегідь задавати кількість кластерів, а не отримувати це як результат в ієрархічному аналізі. Цю проблему можна здолати проведенням ієрархічного аналізу з випадково відібраною вибіркою спостережень і, таким чином, визначити оптимальну кількість кластерів.

Якщо кількість кластерів вказати заздалегідь, то виникає проблема визначення початкових значень центрів кластерів. Їх також можна взяти із заздалегідь проведеного ієрархічного аналізу, в якому для кожного спостереження розраховують середні значення змінних, що

використовувалися під час аналізу, а потім в певній формі зберігають їх у деякому файлі. Цей файл можна потім прочитати методом, який застосовується для оброблення великої кількості спостережень. Якщо немає бажання проходити весь цей довгий шлях, то можна скористатися методом, пропонованим для цього спостереження програмою SPSS.

Якщо кількість кластерів *k*, яку необхідно отримати внаслідок об'єднання, задано заздалегідь, то перші *k* спостережень, що містяться у файлі, використовуються як перші кластери. На наступних кроках кластерний центр замінюється спостереженням, якщо найменша відстань від нього до кластерного центру більша на відстань між двома найближчими кластерами. За цим правилом замінюється той кластерний центр, який є найближчим до цього спостереження. Таким чином, одержуємо новий набір початкових кластерних центрів. Для завершення кроку процедури розраховується нове положення центрів кластерів, а спостереження перерозподіляються між кластерами зі зміненими центрами. Цей ітераційний процес триває доти, доки кластерні центри не перестануть змінювати своє положення або доки не буде досягнуто максимальної кількості ітерацій.

Спершу рекомендується скоротити кількість змінних з допомогою факторного аналізу.

Відкрити файл.

Вибрати в меню Analyze (Аналіз) Data Reduction (Перетворення даних) Factor (Факторний аналіз).

Змінні *v1a – v5b* занести до списку цільових змінних.

Через вимикач Extraction (Відбір) деактивувати виведення неповерненого факторного рішення.

Через вимикач Rotation (Обертання) для здійснення обертання активувати метод варімакса.

Клацнувши по вимикачу Options… (Опції) в розділі Coefficient Display Format активувати Sorted by Size (Відсортовані за розміром). Потім активувати опцію Suppress absolute values less then: (Не виводити абсолютні значення, менші, ніж:) і ввести значення, наприклад 40.

На закінчення клацнути по вимикачу Scores (Значення), щоб значення чинників зберегти у вигляді нових змінних.

Після розрахунку було відібрано чотири чинники і додано у файл чотири змінні від (*fac1_1* до *fac4_1*), які й відображують ці чотири чинники. Серед результатів є обернена факторна матриця (таблиця), з якої видно, що відібрані чинники можна розташувати в певній смисловій послідовності (за убуванням значущості).

Тепер використаємо збережені значення цих чотирьох чинників для проведення кластерного аналізу. Якщо кількість спостережень є

занадто великою для ієрархічного кластерного аналізу, виберемо метод аналізу кластерних центрів.

Присвоїти змінним *fac1_1 – fac4_1* мітки.

Вибрати в меню Analyze (Аналіз) Classify (Класифікувати) K-Means Cluster (Кластерний аналіз методом *k*-середніх).

Відкриється діалогове вікно K-Means Cluster Analysis (Кластерний аналіз методом *k*-середніх) (рис. 2.75).

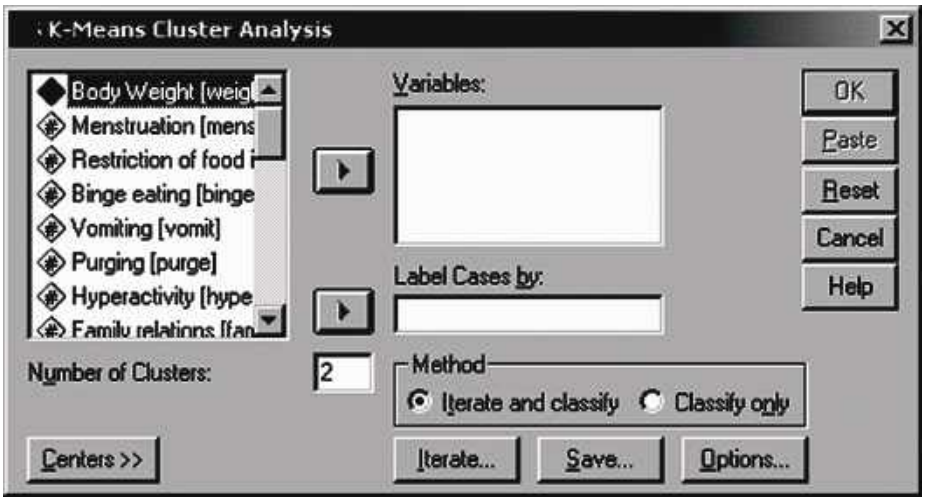


Рис. 2.75. Діалогове вікно K-Means Cluster Analysis

Змінні від *fac1_1* до *fac4_1* помістити в поле тестованих змінних. Тепер необхідно вказати кількість кластерів. За краще було б спершу провести ієрархічний кластерний аналіз для довільно вибраних спостережень і кількість кластерів, яку було одержано, взяти за оптимальну.

Розглянемо чотири кластери. Увести цифру «4» в поле Number of Clusters (Кількість кластерів).

Через вимикач Iterate (Ітерації) вказати кількість ітерацій, що дорівнює 99; установлена за замовчуванням кількість ітерацій, що дорівнює 10, виявилася б недостатньою.

Клацнути по вимикачу Save... (Зберегти), щоб з допомогою додаткових змінних зафіксувати належність спостережень до кластера.

Клацнути на ОК, щоб почати розрахунок.

Спочатку наводяться первинні кластерні центри й узагальнені дані ітераційного процесу (30 ітерацій); потім виводяться остаточні кластерні центри й інформація про кількість спостережень.

При оцінюванні кластерних центрів слід в першу чергу звернути увагу на те, що розглядаються середні значення чинників, які знаходяться в межах приблизно від –3 до +3. До того ж потрібно пам'ятати, що відповідно до кодування відповідей (1 = відмінно, 5 = абсолютно не використовую) велике негативне значення чинника означає велику

міру його прояву, тобто сигналізує про високу компетентність, і навпаки, велике позитивне значення чинника означає невелику міру його прояву.

На закінчення виводяться показники кількості спостережень, що належать до кожного з кластерів.

До початкового файла було додано змінну *qc1_1*, що відбиває належність до певного кластера. Цю змінну можна використовувати для виявлення можливих зв'язків між кластерною належністю і статтю, віком або професією.

2.15. Стандартні графіки

У SPSS є велика кількість різних графіків, які можна побудувати як з допомогою процедур меню графіків, так і різноманітних процедур меню статистик.

Кожний створений графік виникає у вікні перегляду разом з іншими таблицями. Для будування графіка, як правило, виявляється достатньо після вибору типу графіка вказати необхідні змінні, на основі яких його буде побудовано за раніше заданою схемою. Для редагування графіка треба двічі клацнути на якій-небудь точці в межах графіка. Після цього виникне множина можливостей для додаткового редагування.

Починаючи з 8-ї версії, у SPSS нарівні зі створенням традиційних стандартних графіків існує можливість створювати й інтерактивні. Стандартні графіки будуються з допомогою численних процедур статистичного меню або меню графіків. Однак у меню графіків додано ще одну позицію – Interactive, що відкриває ще одне власне меню, яке призначено для будування так званих інтерактивних графіків, що дають досить широку палітру нових можливостей.

Крім того, що є можливість зручно змінювати окремі стильові елементи графіків і перетворювати змінні, які використуються для будування графіка, за допомогою інтерактивних графіків можна одночасно будувати декілька графіків для окремих категорій додаткових змінних.

Під час розроблення графічного подання діаграм можна помітити, що на практиці існують дві різні вихідні ситуації. Найчастіше трапляється ситуація, коли додатково до результатів статистичного аналізу, що зберігаються у файлі даних SPSS, необхідно побудувати й графічне подання цих результатів. Наприклад, необхідно подати досліджувані частоти у вигляді лінійної діаграми. У цьому випадку комп'ютер сам з допомогою відповідних розрахунків знаходить частоти, необхідні для будування стовпців діаграми.

Зовсім іншу ситуацію можна спостерігати, якщо є вже підраховані й оброблені дані. Такий випадок виникає, якщо, наприклад, взяти з га-

зети інформацію про щоденний видобуток нафти країнами, що входять в OPEC, і подати ці дані у вигляді лінійної діаграми. За наявності таких готових даних треба розуміти, як їх подати у файлі.

Якщо клацнути у списку меню на Graphs (Графіки), то буде видно меню з варіантами графіків.

Установки за замовчуванням задають різні кольори, у які зафарбовуються елементи графіків (наприклад, маркери, сегменти), і лінії, що полегшує розуміння діаграми й покращує її презентабельність. Якщо ж необхідно надрукувати графік на принтері або подати його в інших формах, то здебільшого використовувати кольорові графіки не рекомендується. У цих випадках різні поверхні можна позначити з допомогою різних штриховок, а лінії – з допомогою різних видів ліній.

Ці властивості можна змінити, якщо вибрати у меню Edit (Виправлення) Options (Параметри) і в діалоговому вікні Options (Параметри) клацнути на Charts (Діаграми).

У розділі Fill Patterns and Line Styles (Заливання візерунком і стиль ліній) замість опції Cycle through colors, then patterns (Спочатку переглянути кольори, потім візерунки) активувати опцію Cycle through patterns (Переглянути візерунки).

У методах будування графіків через вимикач Titles... (Заголовки) можна присвоїти діаграмі іншу назву, через вимикач Options (Параметри) вибрати метод оброблення пропущених значень і в поле Template (Шаблон) з допомогою активування Use chart specifications from: (Установки діаграми взяти з:) завантажити установки для будування графіка з інших файлів.

2.15.1. Стовпчасті діаграми

Стовпчасті діаграми застосовують, як правило, у таких ситуаціях:

– відображення частот змінних, що належать до номінальної або порядкової шкали;

– відображення середніх значень, сум або інших показників послідовних змінних (тобто змінних, що належать до інтервальної шкали або до шкали відношень);

– відображення змінних, згрупованих за категоріями змінних з номінальною або порядковою шкалою або тимчасової залежності.

Для будування стовпчастої діаграми після відкриття відповідного файла SPSS вибрати в меню Graphs (Графіки) Bar... (Стовпчасті).

Відкриється діалогове вікно Bar Charts (Стовпчасті діаграми) (рис. 2.76).

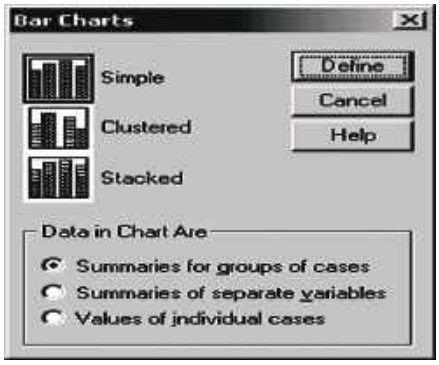


Рис. 2.76. Діалогове вікно Bar Charts (Стовпчасті діаграми)

Можна вибрати просту, кластеризовану (кластерну) або зістиковану стовпчасту діаграму. Дані, які відображено в цих діаграмах, можна задати як категорії однієї змінної, як різні змінні або як значення окремих спостережень.

2.15.1.1. Прості стовпчасті діаграми

Відкрити файл із даними, що досліджуються.

Для будування стовпчастої діаграми для відсоткових показників частот необхідно:

– клацнути на області Simple (Проста) і залишити попередню установку Summaries for groups of cases (Оброблення категорій однієї змінної);

– клацнути по кнопці Define (Визначити); відкриється відповідне діалогове вікно (рис. 2.77);

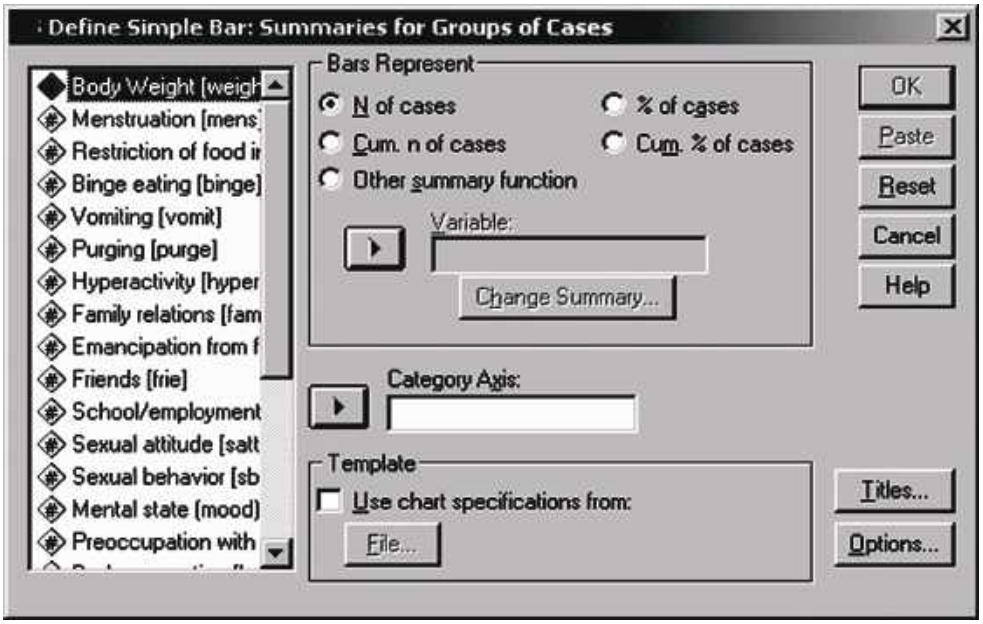


Рис. 2.77. Діалогове вікно Define Simple Bar: Summaries for groups of cases (Проста Стовпчаста діаграма: Оброблення категорій однієї змінної)

– у поле Category Axis: (Вісь категорій) увести змінну, для якої побудувати стовпчасту діаграму, активувати % of cases (% спостережень), якщо змінна має відсоткові показники, і, пройшовши вимикач Titles... (Заголовок), увести заголовок для діаграми;

– клацнути на ОК.

Буде побудовано графік, показаний на рис. 2.78.

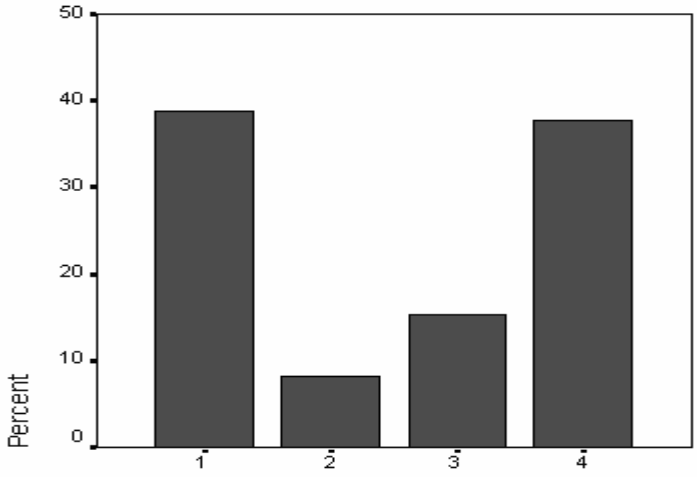


Рис. 2.78. Проста стовпчаста діаграма (Категорії однієї змінної)

Для подання в графічному вигляді значень окремих змінних:

– в діалоговому вікні Bar Charts (Стовпчасті діаграми) активувати Summaries of separate variables (Оброблення окремих змінних); після натискання вимикача Define (Визначити) відкриється відповідне діалогове вікно (рис. 2.79);

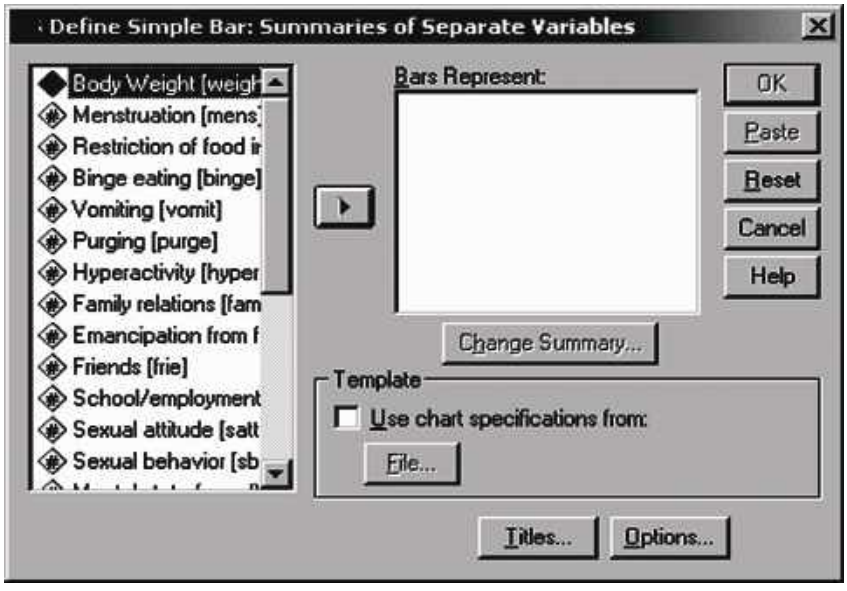


Рис. 2.79. Діалогове вікно Define Simple Bar: Summaries of separate variables (Будування простої стовпчастої діаграми: Оброблення окремих змінних)

– у поле Bars Represent (Значення стовпців) по черзі внести окремі змінні, для яких будується діаграма, і залишити встановлену за замовчуванням функцію Mean of values (Середні значення);

– клацнувши по вимикачу Titles... (Заголовок), увести заголовок діаграми;

– клацнути на ОК.

Буде побудовано графік, показаний на рис. 2.80.

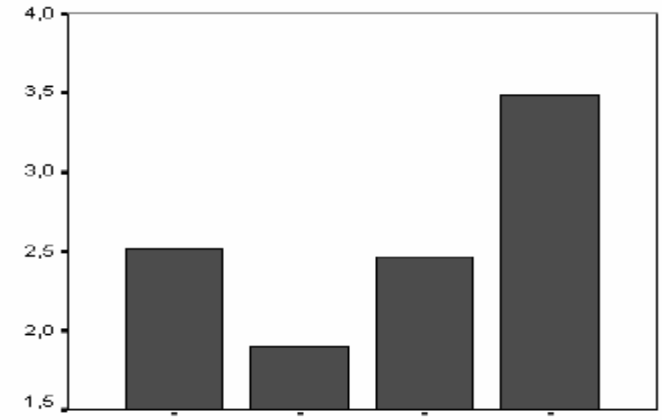


Рис. 2.80. Проста стовпчаста діаграма (Окремі змінні)

Діаграми можна підкоригувати в редакторі діаграм.

Якщо необхідно вибрати функцію, відмінну від установленої за замовчуванням Mean of values (Середні значення), то треба клацнути на одній зі змінних у списку й потім на вимикачі Change Summary... (Змінити метод оброблення).

Відкриється діалогове вікно з переліком функцій (рис. 2.81).

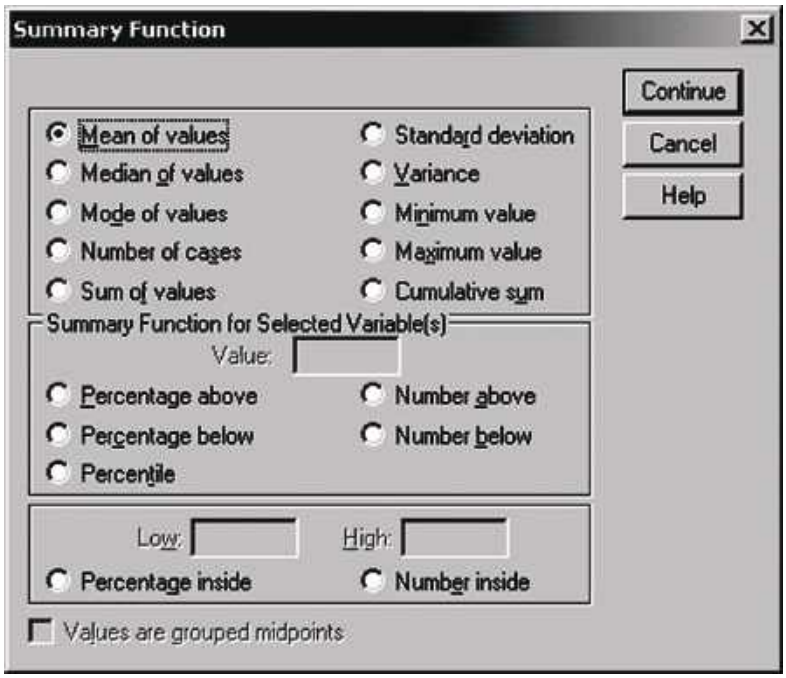


Рис. 2.81. Діалогове вікно Summary Function (Обробна функція).

Це діалогове вікно виникає тільки для стовпчастих, лінійних, кругових діаграм і діаграм з областями, причому не кожна з функцій, що знаходиться тут, є придатною для всіх видів діаграм.

Якщо для наявних даних необхідно відобразити медіани або інші процентилі, то треба активувати опцію Values are grouped midpoints (Значення є згрупованими середніми точками).

Подамо готові дані у вигляді стовпчастої діаграми:

Відкрити файл.

У діалоговому вікні Bar Charts (Стовпчасті діаграми) активувати опцію Values of individual cases (Значення окремих спостережень.

Після натискання вимикача Define (Визначити) відкриється відповідне діалогове вікно (рис. 2.82).

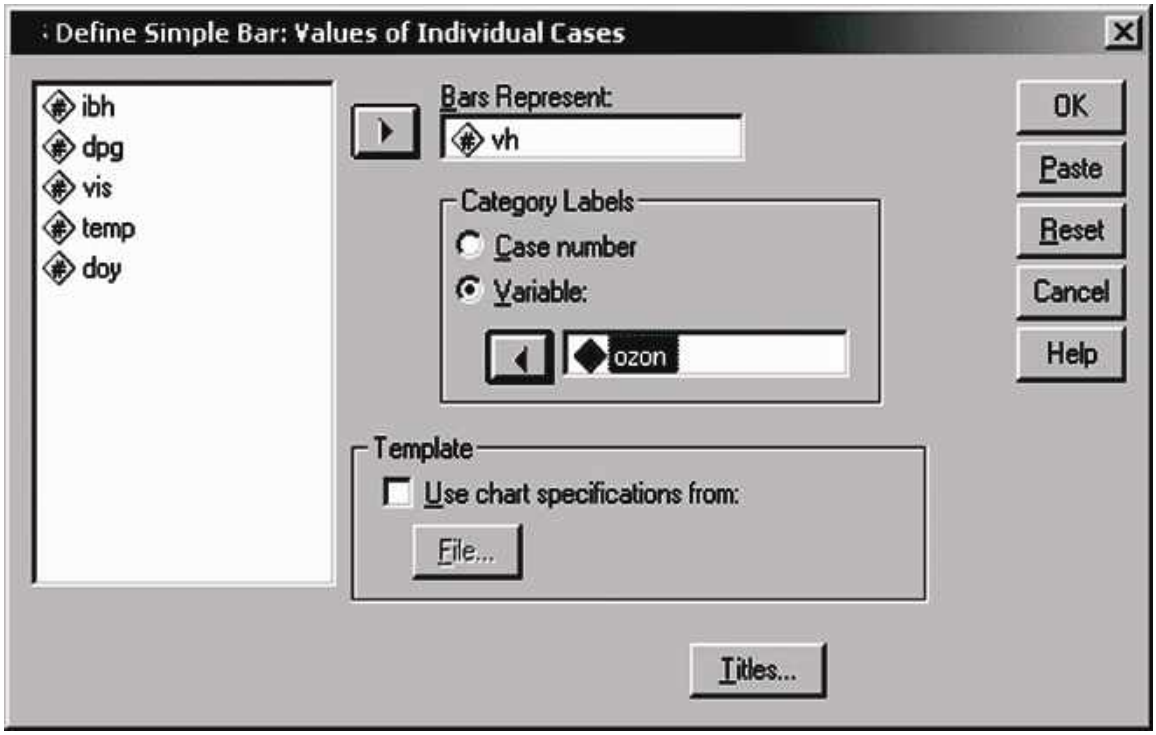


Рис. 2.82. Діалогове вікно Define Simple Bar: Values of individual cases (Будування простої стовпчастої діаграми: Значення окремих випадків)

У поле Bars Represent (Значення стовпців) внести числову змінну; у групі Category Labels (Мітки категорій) активувати Variable: (Змінна) і внести другу змінну.

Клацнувши по вимикачу Titles... (Заголовок), увести заголовок діаграми й клацнути на ОК.

Графік буде мати такий вигляд, як показано на рис. 2.83.

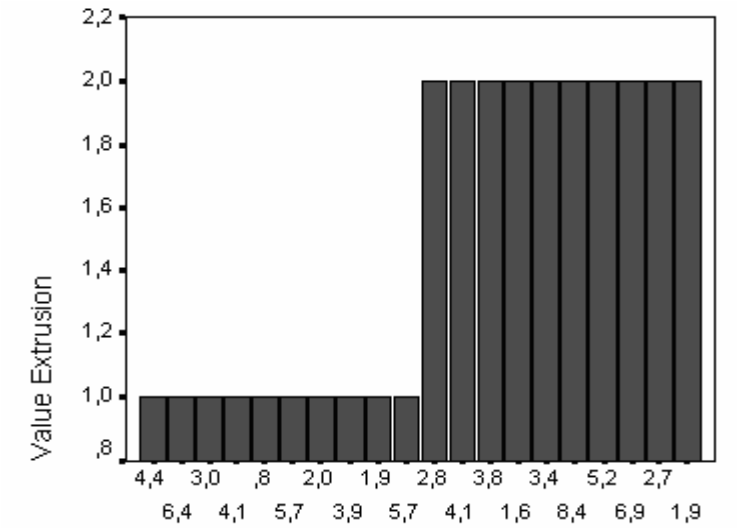


Рис. 2.83. Проста стовпчаста діаграма (Значення окремих випадків)

2.15.1.2. Кластеризовані стовпчасті діаграми

Відкрити файл.

У діалоговому вікні Bar Charts (Стовпчасті діаграми) клацнути на області Clustered (Кластеризована); активувати опцію, що установлюється за замовчуванням, Summaries for groups of cases (Оброблення категорій однієї змінної).

Клацнути на кнопці Define (Визначити); відкриється головне діалогове вікно, що зображено на рис. 2.84.

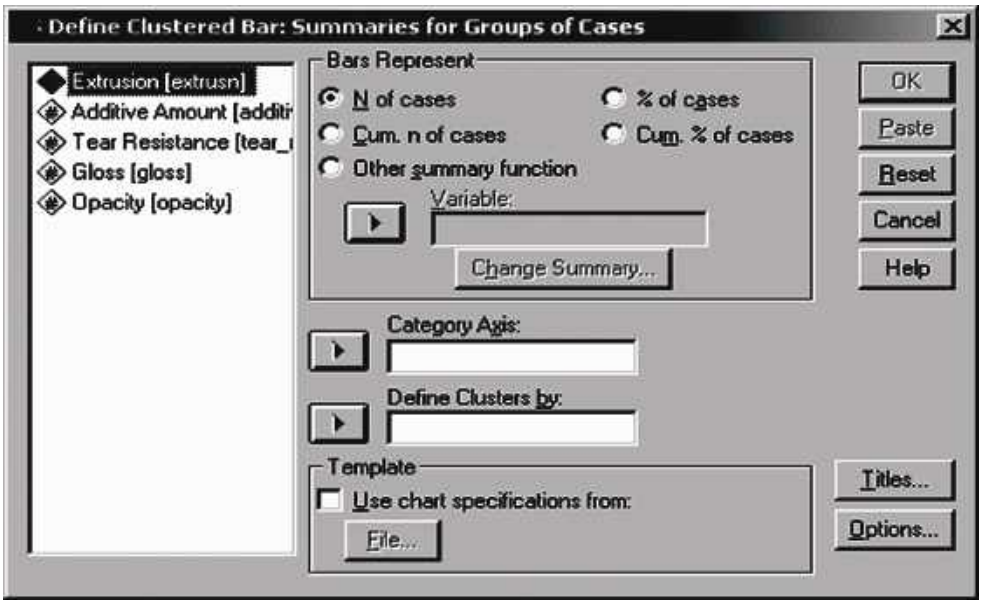


Рис. 2.84. Діалогове вікно Define Clustered Bar: Summaries for groups of cases (Будування групованої діаграми: Оброблення категорій однієї змінної)

У поле Category Axis: (Вісь категорій) увести першу змінну, у поле Define Clusters by: (Створити групи за допомогою:) – другу змінну. Активувати % of cases (% спостережень).

Клацнувши по вимикачу Titles... (Заголовок), увести заголовок для діаграми й почати будування діаграми клацанням на ОК (рис. 2.85).

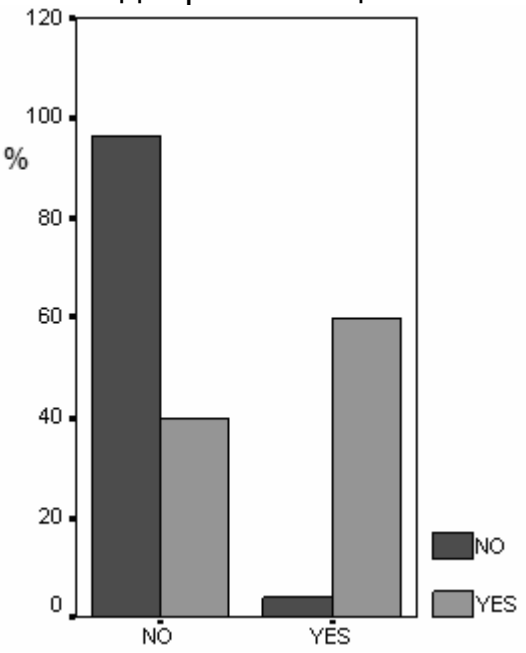


Рис. 2.85. Групована стовпчаста діаграма

2.15.1.3. Зістиковані діаграми

Як правило, зістиковану стовпчасту діаграму застосовують тоді, коли у стовпцях відображено частоти, які мають бути поділені за допомогою деякої зовнішньої змінної. У цьому випадку і вигляд сумарних частот надається користувачеві інакше, ніж вигляд кластеризованої стовпчастої діаграми.

Відкрити файл, що містить дані.

Необхідно відобразити в графічному вигляді розподіл частот окремо для кожної категорії змінної, наприклад, для статі – жіночої й чоловічої.

У діалоговому вікні Bar Charts (Стовпчасті діаграми) клацнути на області Stacked (Зістикована) і активувати опцію, яку установлюють за замовчуванням, Summaries for groups of cases (Оброблення категорій однієї змінної) (рис. 2.86). Клацанням по кнопці Define (Визначити) відкрити відповідне діалогове вікно.

У поле Category Axis: (Вісь категорій) увести першу змінну, а в поле Define Stacks by: (Створити штабелі з допомогою:) увести змінну, яка має декілька категорій. Залишити установку за замовчуванням *N* of cases (Кількість спостережень).

Клацнувши по вимикачу Titles... (Заголовок), увести відповідний заголовок.

Якщо є пропущені значення, які відповідно до установок за замовчуванням будуть оброблятися як окремі категорії, то для того щоб заборонити цю дію, треба клацнути на вимикачі Options... (Параметри) і вибрати оцінку для опції Display groups defined by missing values (Пропущені значення відображати як категорії).

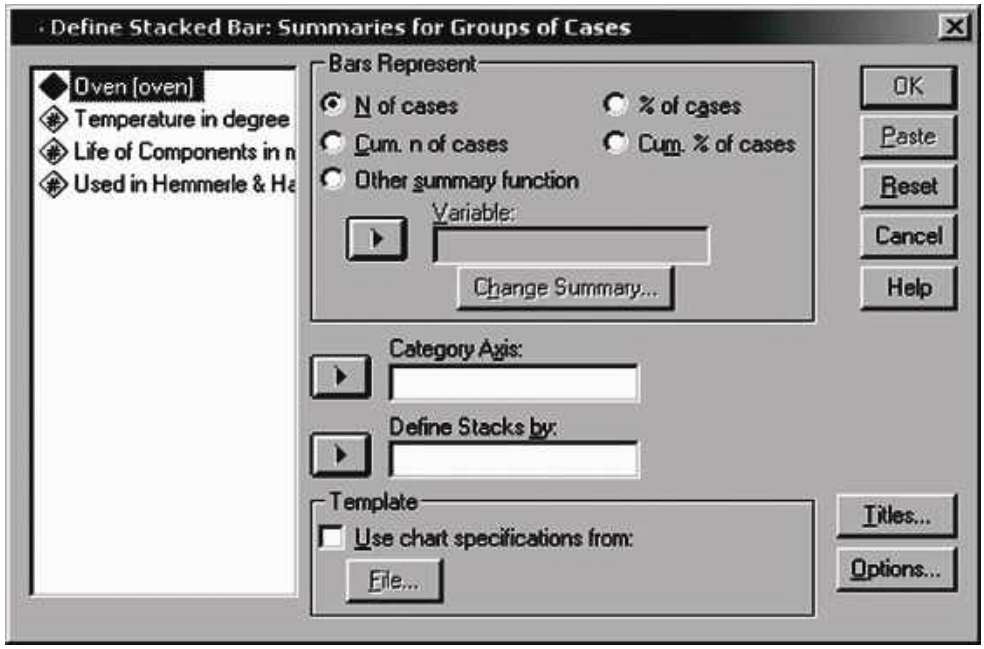


Рис. 2.86. Діалогове вікно Define Stacked Bar: Summaries for groups of cases (Будування штабельної діаграми: Оброблення категорій однієї змінної)

Повернувшись у діалогове вікно Define Stacked Bar: Summaries for groups of cases (Будування зістикованої діаграми: Оброблення категорій однієї змінної) клацанням на ОК почати будування діаграми (рис. 2.87).

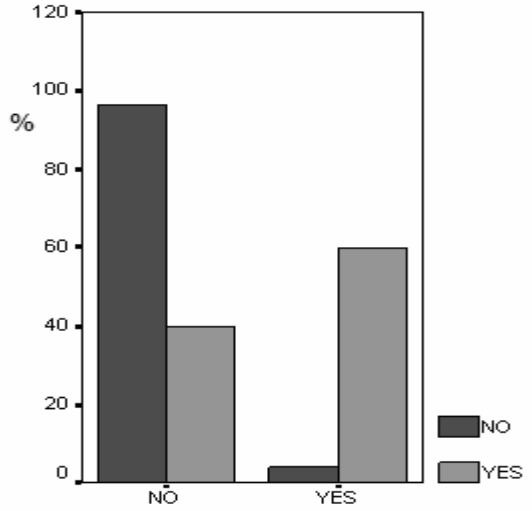


Рис. 2.87. Зістикована стовпчаста діаграма

2.15.2. Лінійні діаграми

Лінійну діаграму замість стовпчастої слід вибирати тоді, коли необхідно відобразити велику кількість стовпців, а також коли стовпці розташовано в певній послідовності. Як правило, це тимчасова послідовність.

Для будування лінійної діаграми після відкриття відповідного файла SPSS вибрати в меню: Graphs (Графіки) Line (Лінійні). Відкриється діалогове вікно Line Charts (Лінійні діаграми) (рис. 2.88).

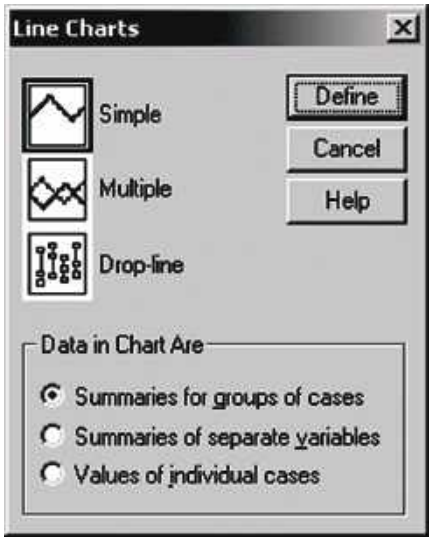


Рис. 2.88. Діалогове вікно Line Charts (Лінійні діаграми)

Можна побудувати просту, складну й зв'язану лінійні діаграми. Як і для стовпчастих діаграм дані, що відображено в цих діаграмах, можна задати як категорії однієї змінної, як різні змінні або як значення окремих спостережень.

2.15.2.1. Прості лінійні діаграми

Відкрити файл і переглянути його вміст у редакторі даних.

У діалоговому вікні Line Charts (Лінійні діаграми) клацнути на області Simple (Проста) і залишити опцію Summaries for groups of cases (Оброблення категорій однієї змінної), що установлюється за замовчуванням.

Після клацання по вимикачу Define (Визначити) відкриється відповідне діалогове вікно (рис. 2.89).

У поле Category Axis: (Вісь категорій) увести першу змінну. У групі Line Represent (Значення ліній) активувати Other summary function (Інша обробна функція) і в поле, що виникло, увести другу змінну. Замість установленої за замовчуванням функції Mean of values (Середні

значення), клацнувши по вимикачу Change Summary...(Змінити метод оброблення), відмітити функцію суми значень (Sum of values) (яка в цьому випадку, щоправда, дає той самий ефект).

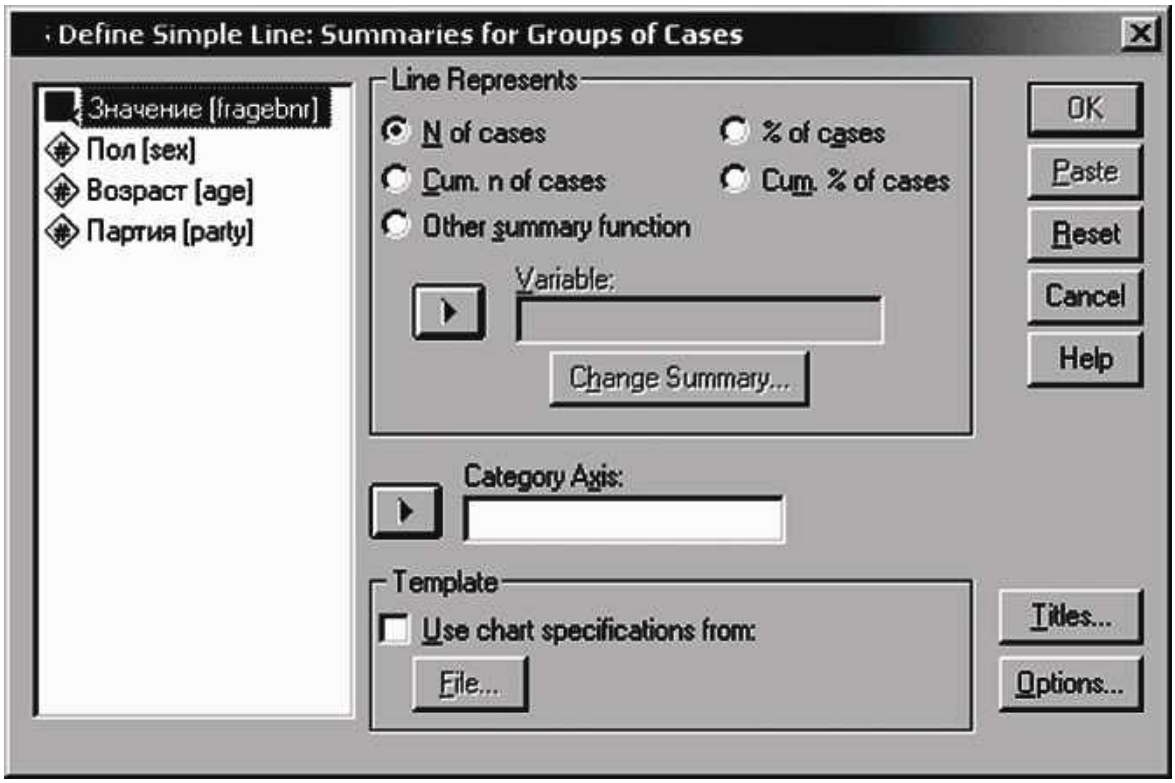


Рис. 2.89. Діалогове вікно Define Simple Line: Summaries for Groups of Cases (Будування простої лінійної діаграми: Оброблення категорій однієї змінної)

За допомогою вимикача Titles... (Заголовок) увести відповіний заголовок.

Почати будування діаграми клацанням на ОК (рис. 2.90).

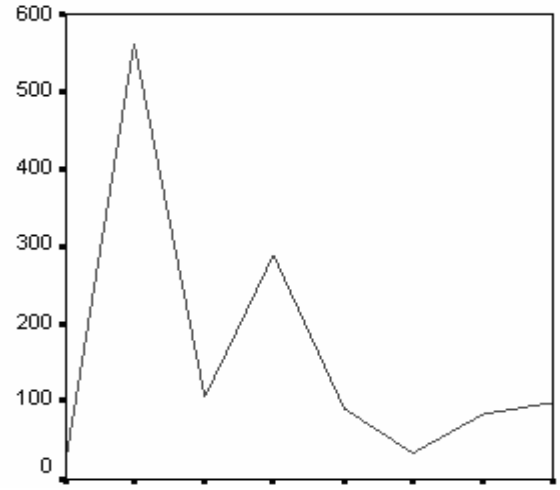


Рис. 2.90. Лінійна діаграма

2.15.2.2. Складні лінійні діаграми

Відкрити файл і переглянути його вміст у редакторі даних.

У діалоговому вікні Line Charts (Лінійні діаграми) клацнути на області Multiple (Складна) і активувати опцію Summaries of separate variables (Оброблення окремих змінних).

Після клацання по вимикачу Define (Визначити) відкриється відповідне діалогове вікно (рис. 2.91).

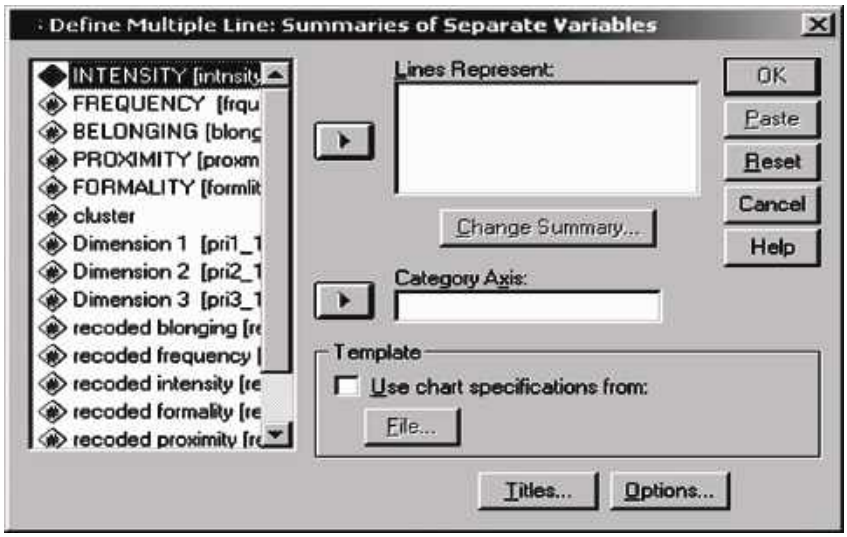


Рис. 2.91. Діалогове вікно Define Multiple Line: Summaries of Separate Variables (Будування складної лінійної діаграми)

У поле Category Axis: (Вісь категорій) увести першу змінну. У поле Line Represent (Значення ліній) по черзі ввести змінні, що залежать від першої зміної; замість установленої за замовчуванням функції Mean of values (Середні значення) з допомогою вимикача Change Summary (Змінити метод оброблення) відмітити функцію суми значень (Sum of values).

Після клацання по вимикачу Titles... (Заголовок) увести відповіний заголовок і почати будування діаграми клацанням на ОК (рис. 2.92).

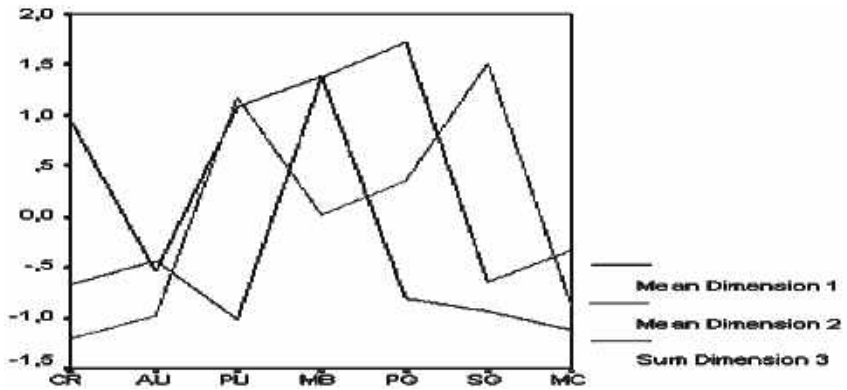


Рис. 2.92. Складна лінійна діаграма

2.15.2.3. Зв'язані лінійні діаграми

Це різновид складної лінійної діаграми, у якому точки даних позначено різними символами й з'єднані вертикальним зв'язком.

У діалоговому вікні Line Charts (Лінійні діаграми) клацнути на області Drop-line (Зв'язані лінії).

Далі зробити те саме, що й у попередньому пункті.

Побудована діаграма буде відповідати показаній на рис. 2.93.

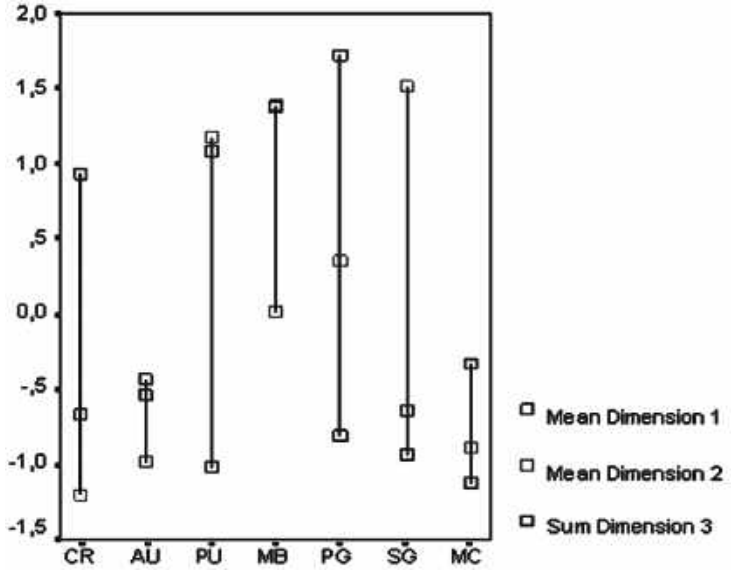


Рис. 2.93. Зв'язана лінійна діаграма

2.15.3. Діаграми з областями

Діаграми з областями є різновидом лінійної діаграми, у якій області, що знаходяться під лініями, зафарбовуються, завдяки чому графік виглядає більш наочним.

Для будування діаграми з областями після відкриття необхідного файла SPSS вибрати ву меню Graphs (Графіки) Area (З областями). Відкриється діалогове вікно Area Charts (Діаграми з областями) (рис. 2.94).

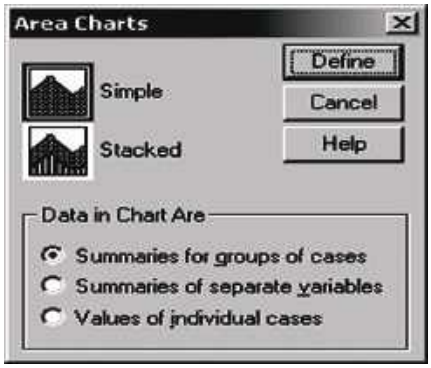


Рис. 2.94. Діалогове вікно Area Charts (Діаграми з областями)

Можна побудувати просту або зістиковану діаграму з областями. Дані, що відображено в цих діаграмах, можна задати як категорії однієї змінної, як різні змінні або як значення окремих спостережень.

Проста діаграма з областями

Наведена ніжче табл. 2.12 містить інформацію про виробництво велосипедів з 1996 р. до 2002 р. Виробничі показники розбито додатково на дані про збут усередині країни й експорт.

Таблиця 2.12

Рік	Кількість, млн шт.		
	Виробництво	Усередині країни	Експорт
1996	4,00	3,14	0,86
1997	3,74	3,01	0,73
1998	3,88	3,14	0,74
1999	4,40	3,67	0,73
2000	4,81	4,08	0,73
2001	4,91	4,35	0,56
2002	4,55	4,10	0,45

Ці дані построчно збережено в змінних *o1* (рік), *o2* (загальний обсяг виробництва), *o3* (усередині країни) та *export* (експорт).

Відкрити файл і переглянути його вміст у вікні редактора даних.

Дані про сукупне виробництво подамо у вигляді простої діаграми з областями.

У діалоговому вікні Area Charts (Діаграми з областями) клацнути на області Simple (Проста) і залишити опцію Summaries for groups of cases (Оброблення категорій однієї змінної), що установлюється за замовчуванням.

Після клацання по вимикачу Define (Визначити) відкриється головне діалогове вікно (рис. 2.95).

У поле Category Axis: (Вісь категорій) увести змінну *o1* і в групі Area Represents (Значення областей) установити маркер біля Other summary function (Інша обробна функція). У поле, що виникло, увести змінну *o2* і залишити функцію Mean of values (Середні значення), що встановлюється за замовчуванням.

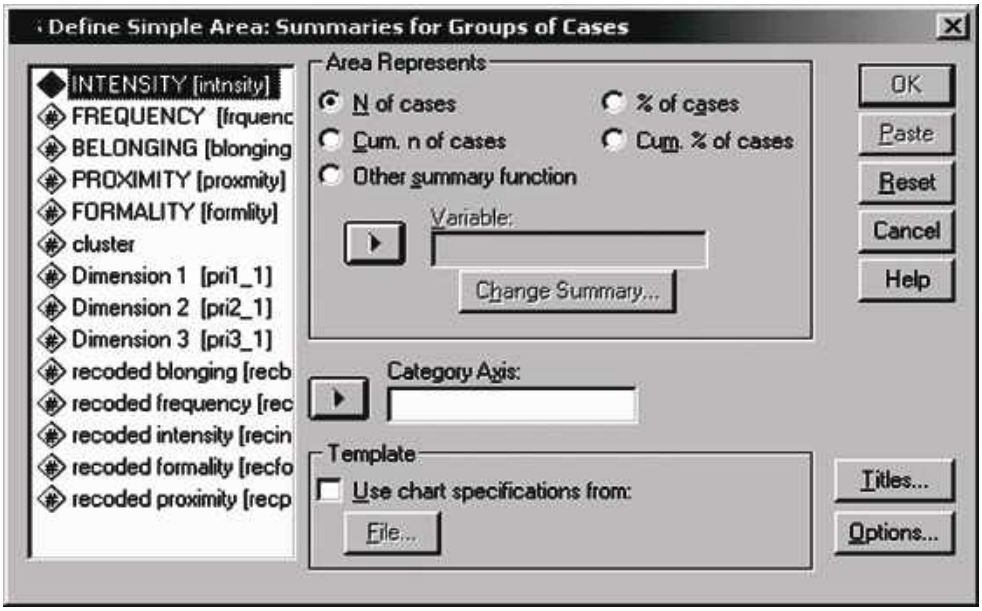


Рис. 2.95. Діалогове вікно Define Simple Area: Summaries for Groups of Cases (Будування простої діаграми з областями: Оброблення категорій однієї змінної)

З допомогою вимикача Titles... (Заголовок) увести відповідний заголовок і почати будування діаграми клацанням на ОК.

Отримаємо діаграму з областями (рис. 2.96).

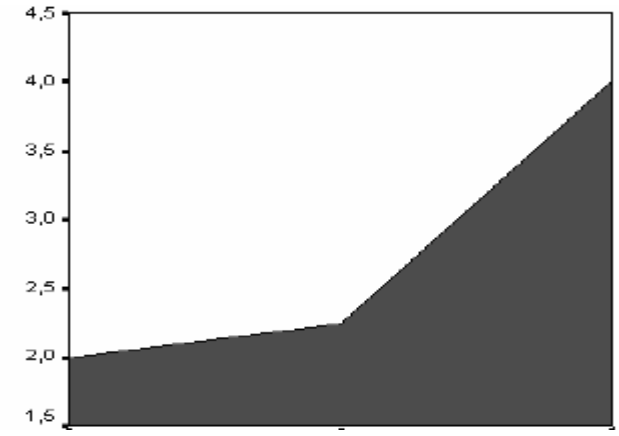


Рис. 2.96. Діаграма з областями

2.15.4. Кругові діаграми

Подання даних у вигляді кругових діаграм слід вибирати тоді, коли частоти або значення змінних можна скласти разом і ця сума відповідатиме ста відсоткам.

Відобразимо з допомогою кругової діаграми частоти деякої змінної.

Відкрити файл і вибрати в меню Graphs (Графіки) Pie (Кругові). Відкриється діалогове вікно Pie Charts (Кругові діаграми) (рис. 2.97).

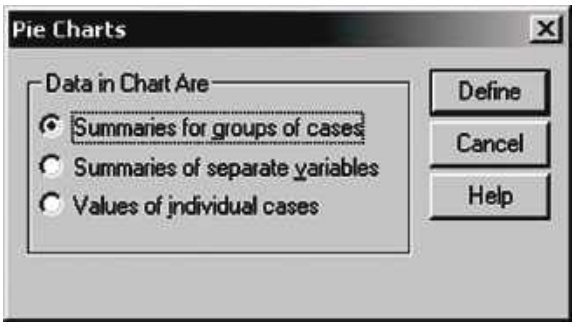


Рис. 2.97. Діалогове вікно Pie Charts (Кругові діаграми)

Залишити опцію Summaries for groups of cases (Оброблення категорій однією змінною), установлену за замовчуванням, і клацанням на кнопці Define (Визначити) відкрити діалогове вікно (рис. 2.98).

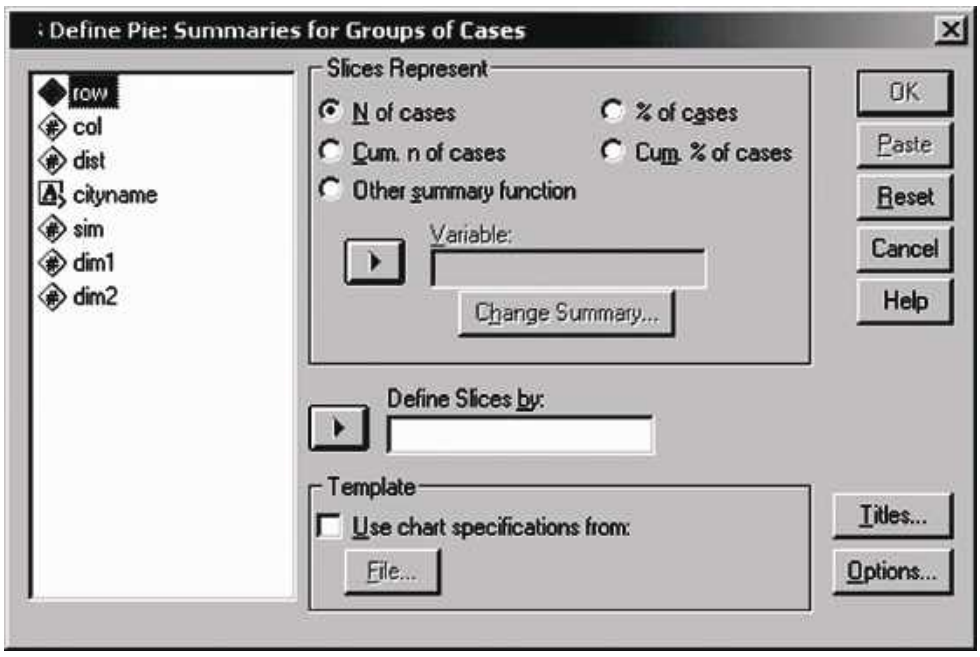


Рис. 2.98. Діалогове вікно Define Pie : Summaries for Groups of Cases (Будування кругової діаграми : Оброблення категорій однією змінною)

У полі Define slices by: (Створити сектори з допомогою:) увести змінну *psyche*.

Клацнути на вимикачі Options... (Параметри) і прибрати маркер з опції Display groups defined by missing values (Пропущені значення відображати як категорії).

З допомогою вимикача Titles... (Заголовок) увести відповідний заголовок і почати будування діаграми клацанням на ОК (рис. 2.99).

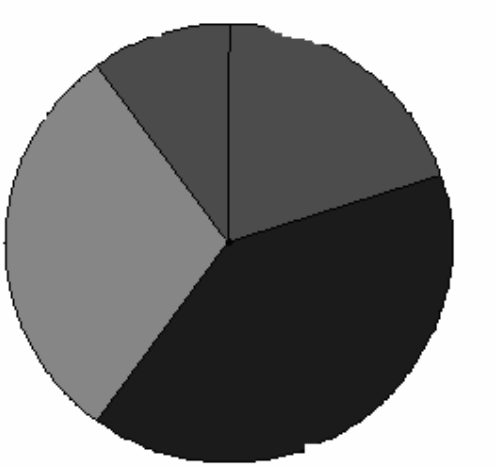


Рис. 2.99. Кругова діаграма

2.15.5. Лінійні діаграми різниць

З допомогою цієї діаграми можна подати взаємне змінення значень двох змінних, причому обидві результуючі криві можуть перетинатися. Цей перетин якраз і можна наочно подати з допомогою лінійних діаграм різниць.

Відкрити файл, у якому зберігаються необхідні дані.

У діалоговому вікні High-Low Charts (Діаграми максимуму-мінімуму) клацнути на області Difference Line (Лінія різниць). Установити мітку біля опції Summaries of separate variables (Оброблення окремих змінних).

Натисненням вимикача Define (Визначити) відкрити діалогове вікно (рис. 2.100).

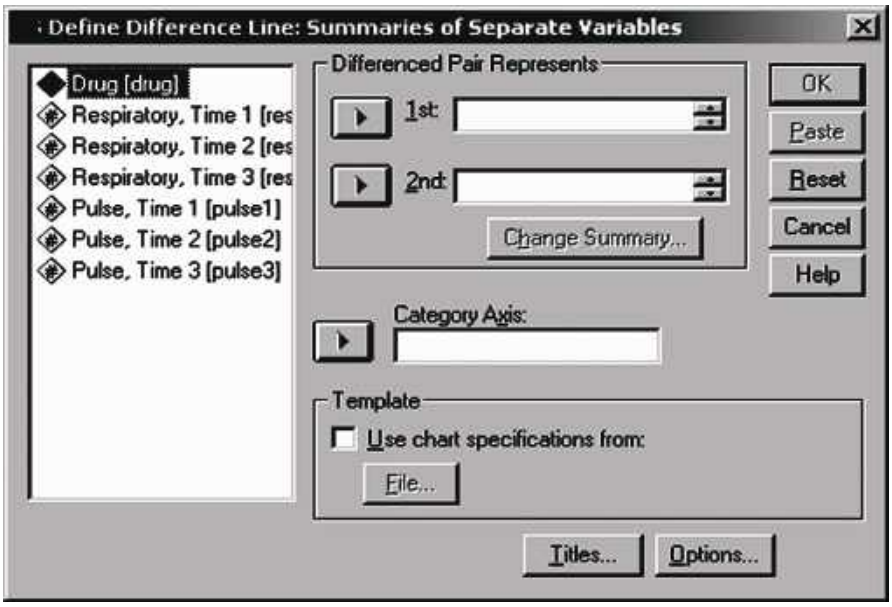


Рис. 2.100. Діалогове вікно Define Difference Line : Summaries of Separate Variables (Будування лінійної діаграми різниць)

У полі Category Axis : (Вісь категорій) увести незалежну змінну, і в групі Differenced Pair Represents (Значення різницевих пар) в поля 1 і 2 ввести змінні, різниця значень яких досліджується. Активувати функцію суми (Sum of values) з допомогою кнопки Change Summary (Змінити процедуру оброблення).

З допомогою вимикача Titles... (Заголовок), увести відповідний заголовок.

Почати будування діаграми клацанням на ОК (рис. 2.101).

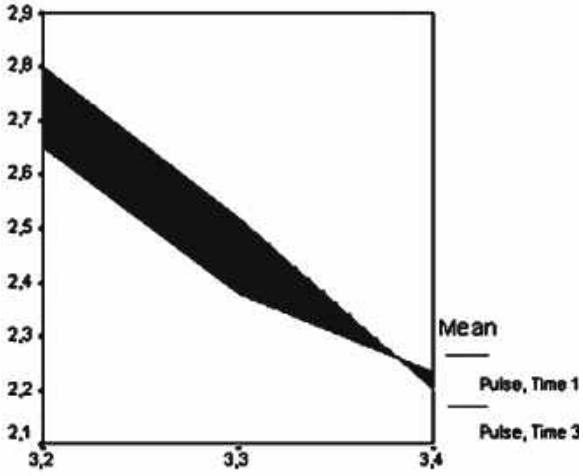


Рис. 2.101. Лінійна діаграма різниць

2.15.6. Коробчасті діаграми

З допомогою коробчастої діаграми можна відобразити медіану й обидва квартиля, мінімальні й максимальні значення, а також пропущені й екстремальні значення. Ці діаграми можна побудувати під час попереднього дослідження даних або через меню графіків.

Коробчаста діаграма складається з прямокутника, що займає простір від першого до третього квартиля (тобто від 25-го до 75-го процентиля). Лінія усередині цього прямокутника відповідає медіані. Крім того, на коробчастій діаграмі відмічається максимальне й мінімальне значення, якщо тільки вони не є викидами. Значення, віддалені від меж більш ніж на три довжини побудованого прямокутника (екстремальні значення), позначаються на діаграмі зірочками. Значення, віддалені більш ніж на півтори довжини прямокутника, позначаються кружками.

Після відкриття необхідного файла SPSS вибрати в меню Graphs (Графіки) Boxplot (Коробчасті діаграми). Відкриється діалогове вікно Boxplot (рис. 2.102).

Можна вибрати просту або кластеризовану діаграму, причому дані можна подати у вигляді категорій однієї змінної або різних змінних.

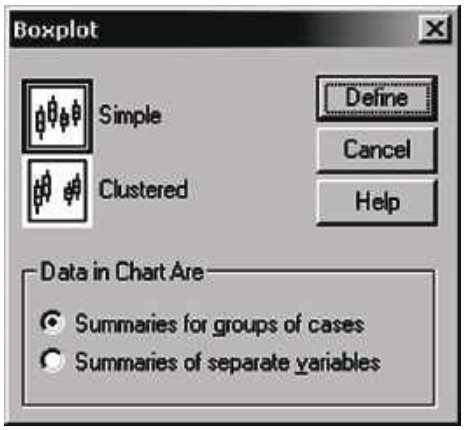


Рис. 2.102. Діалогове вікно Boxplot (Коробчаста діаграма)

У діалоговому вікні Boxplot (Коробчаста діаграма) клацнути на області Simple (Проста) і залишити опцію Summaries for groups of cases (Оброблення категорій однією змінною), що встановлюється за замовчуванням.

Клацанням по вимикачу Define (Визначити) відкрити головне діалогове вікно, в якому в полі Category Axis: (Вісь категорій) увести незалежну змінну, а в полі Variable: (Змінна) – залежну змінну. Якщо ввести яку-небудь змінну в поле Label Cases by: (Мітки спостережень), то її мітки значень будуть використані для позначення пропущених і екстремальних значень.

Почати будування діаграми клацанням на ОК (рис. 2.103).

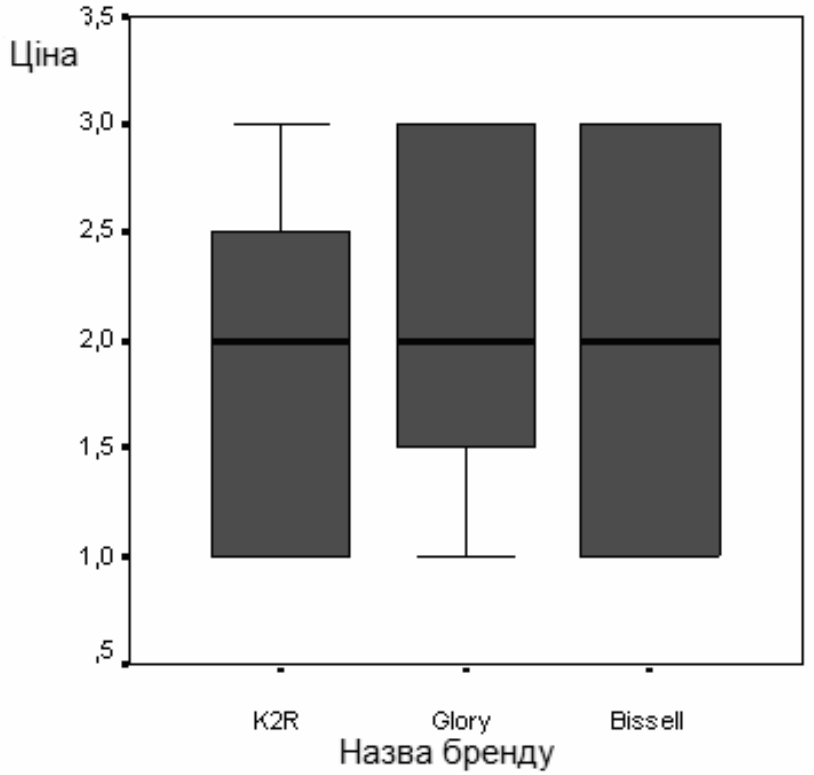


Рис. 2.103. Коробчаста діаграма (Категорії однієї змінної)

2.15.7. Діаграми розсіяння

Діаграма розсіяння в графічному вигляді відображує відношення між двома змінними, які щонайменше належать до інтервальної шкали.

Щоб побудувати діаграму розсіяння, після відкриття необхідного файла SPSS вибрати в меню Graphs (Графіки) Scatter (Розсіяння) (рис. 2.104).

Відкриється діалогове вікно Scatterplot (Діаграма розсіяння).

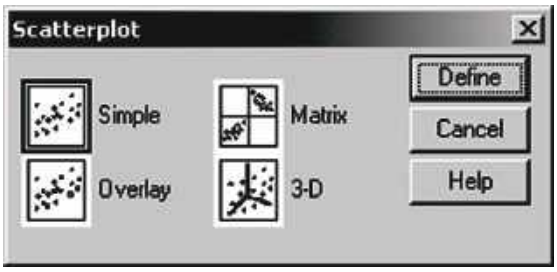


Рис. 2.104. Діалогове вікно Scatterplot (Діаграма розсіяння)

2.15.7.1. Прості діаграми розсіяння

Відкрити файл.

У діалоговому вікні Scatterplot (Діаграма розсіяння) клацнути на області Simple (Проста).

Клацанням по вимикачу Define (Визначити) відкрити відповідне діалогове вікно (рис. 2.105).

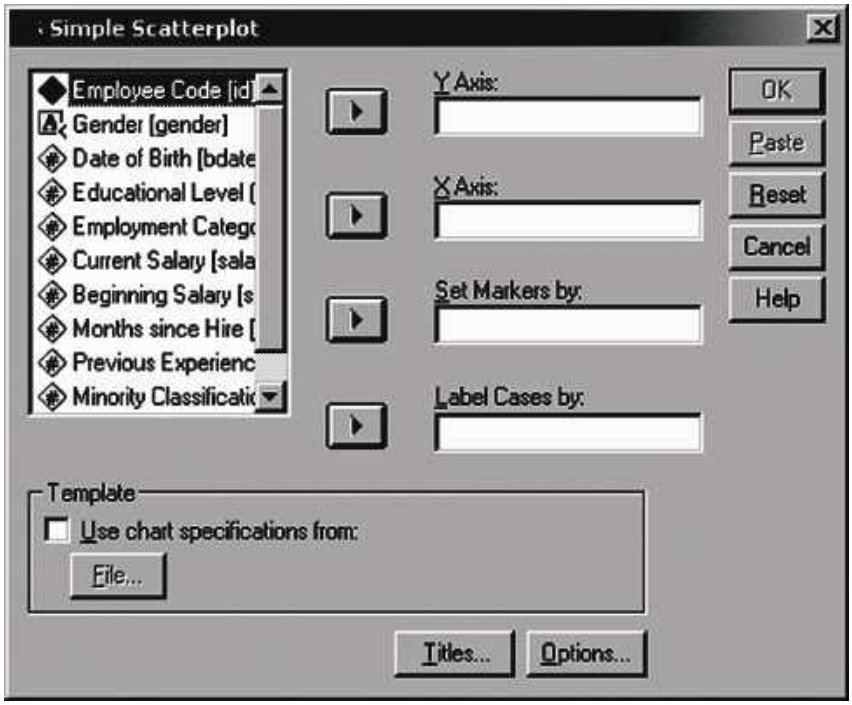


Рис. 2.105. Діалогове вікно Simple Scatterplot (Проста діаграма розсіяння)

Залежну змінну із списку початкових змінних перенести в поле осі *Y*, а незалежну – в полі осі *X*.

Якщо помістити яку-небудь змінну в поле Set Markers by: (Установити маркери для:), то згідно з належністю до цієї змінної окремі точки значень на діаграмі будуть зафарбованими в інший колір або помічені з допомогою якого-небудь примітного маркувального символу.

Помістити змінну-коментар в поле, яке передбачено для опису спостережень (Label Cases by: (Мітки спостережень)). Значення цієї змінної буде поміщено в діаграмі розсіяння поблизу відповідної точки даних.

З цією метою клацнути по вимикачу Options… (Параметри) і в діалоговому вікні, що виникло, активувати опцію Display chart with case labels (Показати графік з мітками спостережень).

Клацнувши по вимикачу Titles.. (Заголовок), увести відповідний заголовок і почати будування діаграми клацанням на ОК (рис. 2.106).

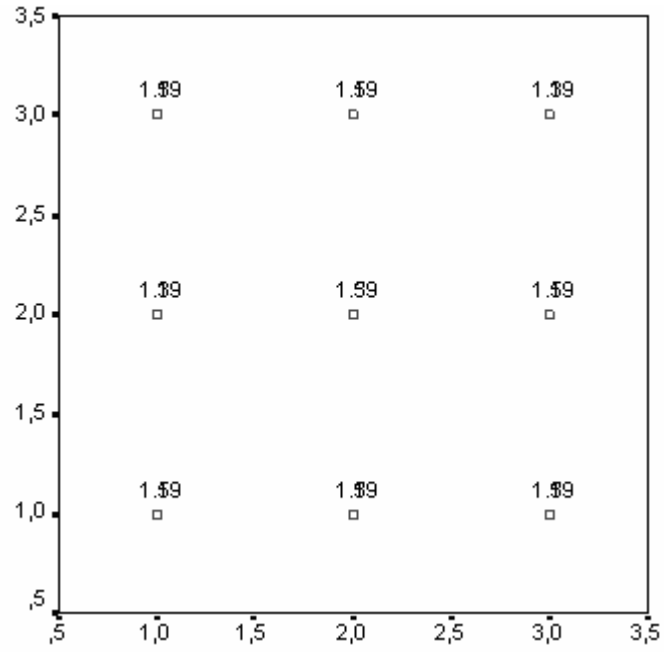


Рис. 2.106. Проста діаграма розсіяння з мітками частот

Велика кількість міток спостережень призводить до зменшення наочності графіка, тому можна рекомендувати залишити їх тільки для вибраних точок.

Як альтернативу можна взяти позначення мітками тільки найбільш характерних точок (рис. 2.107).

Побудувати діаграму наново.

Через вимикач параметрів прибрати маркер опції Display chart with case labels (Показати графік з мітками спостережень).

Тепер міток на графіку не буде.

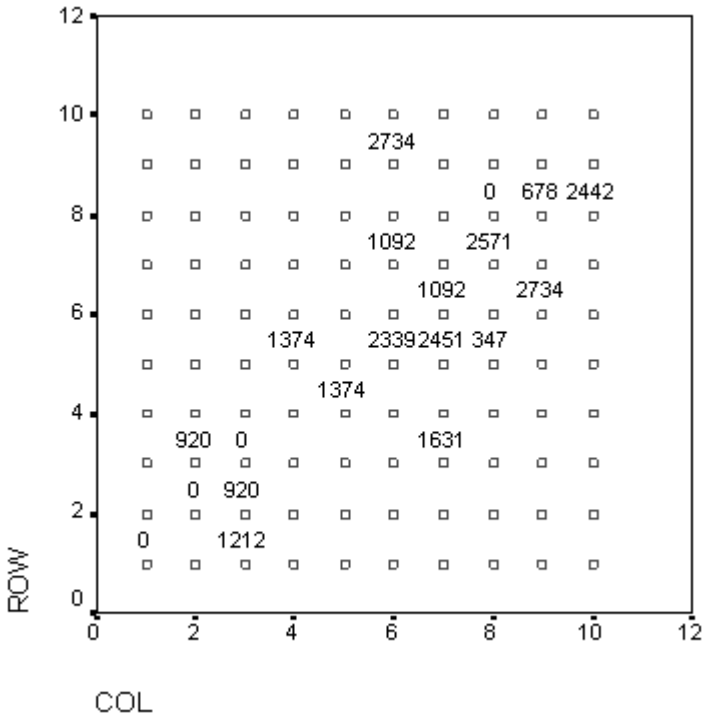


Рис. 2.107. Проста діаграма розсіяння з вибірковими мітками частот

Подвійним клацанням помістити графік в редактор діаграм.

Одним клацанням по символу вибору точок перейти в режим вибору точок. Тепер з допомогою курсора для виділення точок можна вибрати окремі точки на діаграмі розсіяння й позначити їх мітками.

Якщо декілька точок знаходяться дуже близько одна до одної, то буде показано список міток, з якого можна вибрати необхідну.

Числові показники для будь-якої точки, що знаходиться на діаграмі розсіяння, також можна переглянути в редакторі даних.

Для цього з допомогою курсора для виділення точок вибрати необхідну і в списку команд клацнути на кнопці переходу в редактор даних.

Виникне редактор даних. Зміни даних, що заносяться до редактора даних, безпосередньо не впливають на вже побудовану діаграму розсіяння.

2.15.7.2. Матричні діаграми розсіяння

Цей метод застосовують для відображення декількох діаграм розсіяння на одному графіку.

У діалоговому вікні Scatterplot (Діаграма розсіяння) клацнути на області Matrix (Матриця).

Клацанням на вимикачі Define (Визначити) відкрити відповідне діалогове вікно (рис. 2.108).

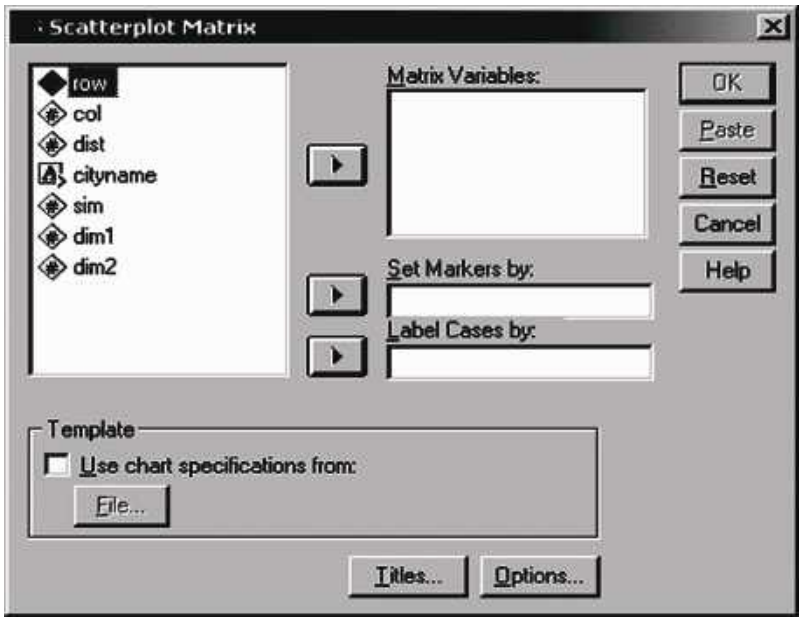


Рис. 2.108. Діалогове вікно Scatterplot Matrix (Матрична діаграма розсіяння)

Якщо необхідно попарно зв'язати одне з одним значення трьох змінних, треба розглянуті змінні по черзі перенести в поле, що передбачено для матричних змінних.

Почати будування діаграми клацанням на ОК (рис. 2.109).

Кількість рядків і стовпців в матричній діаграмі відповідає кількості змінних. Кожна комірка є діаграмою розсіяння для однієї пари змінних. Діагональні комірки містять мітки змінних, що знаходяться у відповідних комірках матриці.

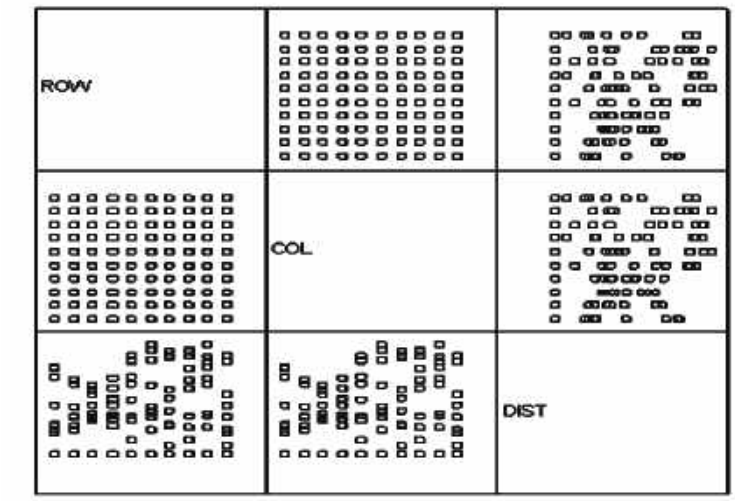


Рис. 2.109. Матрична діаграма розсіяння

У матричних діаграмах розсіяння можна задати маркірування для деякої змінної, організувати виведення міток спостережень, а також відображення будь-якої іншої необхідної інформації; можна також організувати будування різних ліній регресії.

2.15.7.3. Накладені діаграми розсіяння

На одному графіку можна подати декілька діаграм розсіяння.

Для цього в діалоговому вікні Scatterplot (Діаграма розсіяння) клацнути на області Overlay (Накладення) і потім на кнопці Define (Визначити).

У діалоговому вікні, що виникло, можна задати відповідні пари змінних, які мають бути подані разом. Значення, що належать до відповідної пари, на діаграмі будуть відмічені одним певним маркіруванням.

Цей метод можна застосовувати тільки тоді, коли йдеться про змінні з одними й тими самими областями значень.

2.15.7.4. Тривимірні діаграми розсіяння

Ці діаграми будуються на основі значень трьох змінних і тому мають три осі.

На осі *y* відкладають висоту розташування, на осі *x* – горизонтальне розташування, на осі *z* – глибину розташування кожної точки.

У діалоговому вікні Scatterplot (Діаграма розсіяння) клацнути на області 3D (3-вимірна).

Клацанням по вимикачу Define (Визначити) відкрити відповідне діалогове вікно (рис. 2.110).

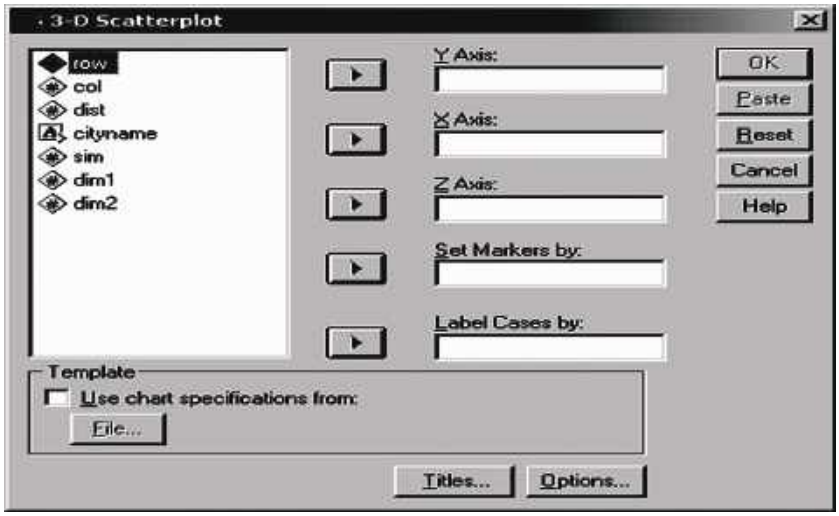


Рис. 2.110. Діалогове вікно 3-D Scanerplot (Тривимірна діаграма розсіяння)

Перенести по черзі змінні, що досліджуються, зі списку початкових змінних у поля, що належать осям *y, x* і *z*.

Почати будування діаграми клацанням на ОК (рис. 2.111).

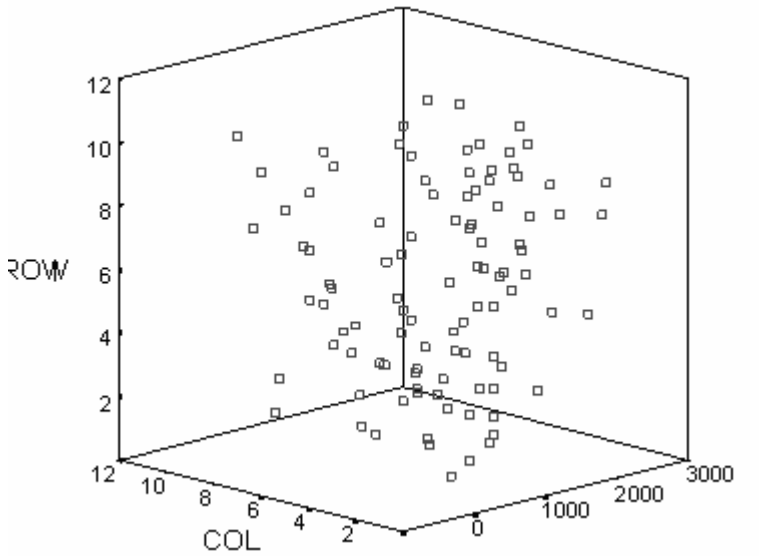


Рис. 2.111. Тривимірна діаграма розсіяння

2.16. Гістограми

Гістограма вже кілька разів розглядалася в попередніх розділах.

Щоб побудувати гістограму, після відкриття необхідного файла SPSS вибрати в меню Graphs (Графіки) Histogram (Гістограма).

Відкриється діалогове вікно Histogram (рис. 2.112).

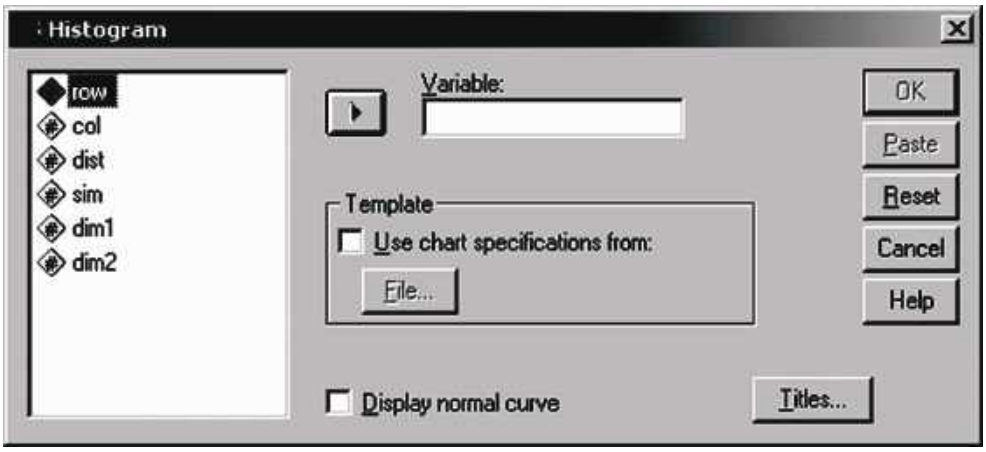


Рис. 2.112. Діалогове вікно Histogram (Гістограма)

З допомогою гістограми можна наочно відобразити розподіл змінних, що належать до інтервальної шкали.

Відкрити файл.

Помістити змінну в поле змінних і активувати виведення кривої нормального розподілу.

Почати будування гістограми клацанням на ОК (рис. 2.113).

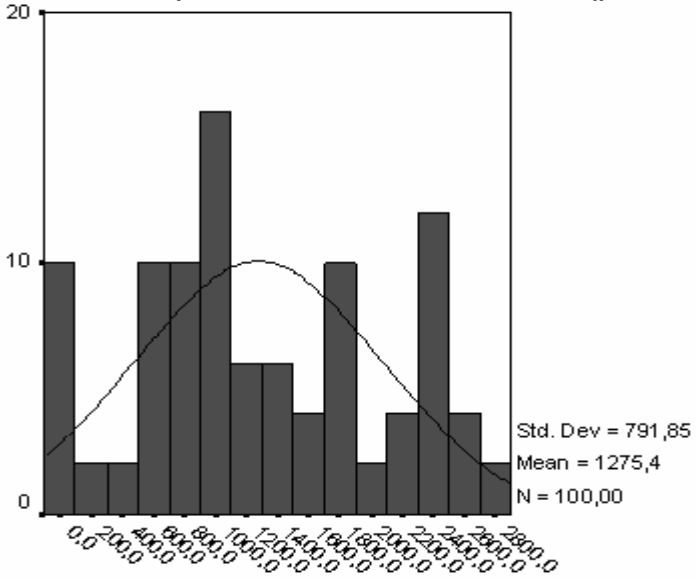


Рис. 2.113. Гістограма з кривою нормального розподілу

Щоб з'ясувати, чи значно відрізняється отриманий розподіл від нормального, недостатньо тільки зовнішнього вигляду гістограми, необхідно перевірити його з допомогою спеціального статистичного тесту.

2.16. Діаграми Парето

Діаграма Парето є стовпчастою діаграмою, у якій стовпці розташовуються в порядку убування, а додаткова крива може вказувати на сукупну частоту для наведених категорій. При цьому при підсумовуванні окремих стовпців за заданим правилом має одержуватися деяка підсумкова величина, що має певний сенс.

Щоб побудувати діаграму Парето, після відкриття необхідного файла SPSS вибрати в меню Graphs (Графіки) Pareto (Парето).

Відкриється відповідне діалогове вікно (рис. 2.114).

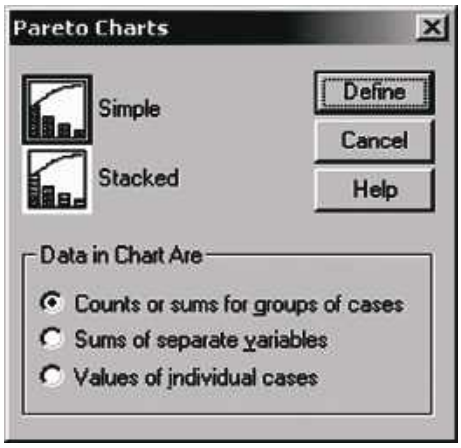


Рис. 2.114. Діалогове вікно Pareto Charts (Діаграми Парето)

Можна побудувати просту або зістиковану діаграму Парето, причому й тут існує три варіанти подання даних.

Відкрити файл, у якому збережено дані.

У діалоговому вікні Pareto Charts (Діаграми Парето) клацнути на області Simple (Проста) і залишити опцію Counts or sums for groups of cases (Частоти або суми категорій однієї змінної), яку встановлено за замовчуванням.

Натисненням вимикача Define (Визначити) відкрити діалогове вікно.

У полі Category Axis : (Вісь категорій) увести першу змінну. У групі Bars Represent (Значення стовпців) поставити маркер поряд з опцією варіанта вибору Sums of variable: (Суми змінних) і перевести залежну змінну в поле, що виникло. Відображення сукупної (кумулятивної) кривої встановлюється за замовчуванням.

З допомогою вимикача Titles... (Заголовок) увести відповідний заголовок.

Клацанням на ОК почати будування діаграми (рис. 2.115).

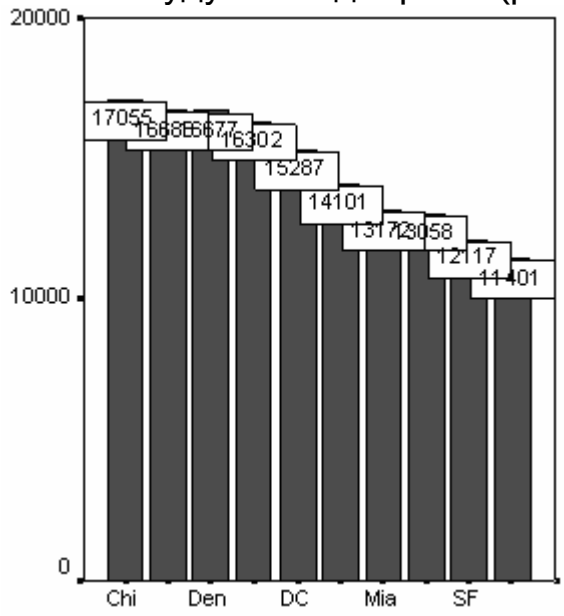


Рис. 2.115. Діаграма Парето (без сукупної кривої)

2.18. Основи редагування графіків

Будування графіків відбувається з допомогою великої кількості процедур меню статистик із меню графіків. Усі графіки, побудовані таким чином, потрапляють відразу у вікно перегляду. Немає проміжного збереження, що існувало аж до шостої версії SPSS.

Навіть при будуванні перших графіків (тепер в SPSS вони, як правило, мають назву діаграм) можна не турбуватися про їхній зовнішній вигляд, оскільки починають діяти відповідні установки за замовчуванням.

Якщо до того ж додати деякі найменування (заголовок, підзаголовок, виноски), то цього вигляду буде вже цілком достатньо, щоб графіки можна було використовувати в більшості практичних ситуацій.

Якщо графікам необхідно надати більш наочного й презентабельного вигляду або ж провести певні коригування (наприклад, якщо мітки змінних занадто довгі), то графік слід перенести в редактор діаграм. Для цього у вікні перегляду двічі клацнути в будь-якому місці в області діаграми.

У редакторі діаграм над графіком можна виконувати такі дії:

– коригувати (або змінити);
– зберегти графік в якому-небудь іншому графічному форматі;
– зберегти як зразок для інших графіків;
– копіювати в буфер обміну Windows.

2.19. Редактор діаграм

Для того, щоб графік можна було змінити (допрацювати, редагувати), його треба помістити в редактор діаграм. Це відбудеться після подвійного клацання на якій-небудь точці в області діаграми, що знаходиться у вікні перегляду. Тоді редактор діаграм буде мати вигляд, показаний на рис. 2.116.

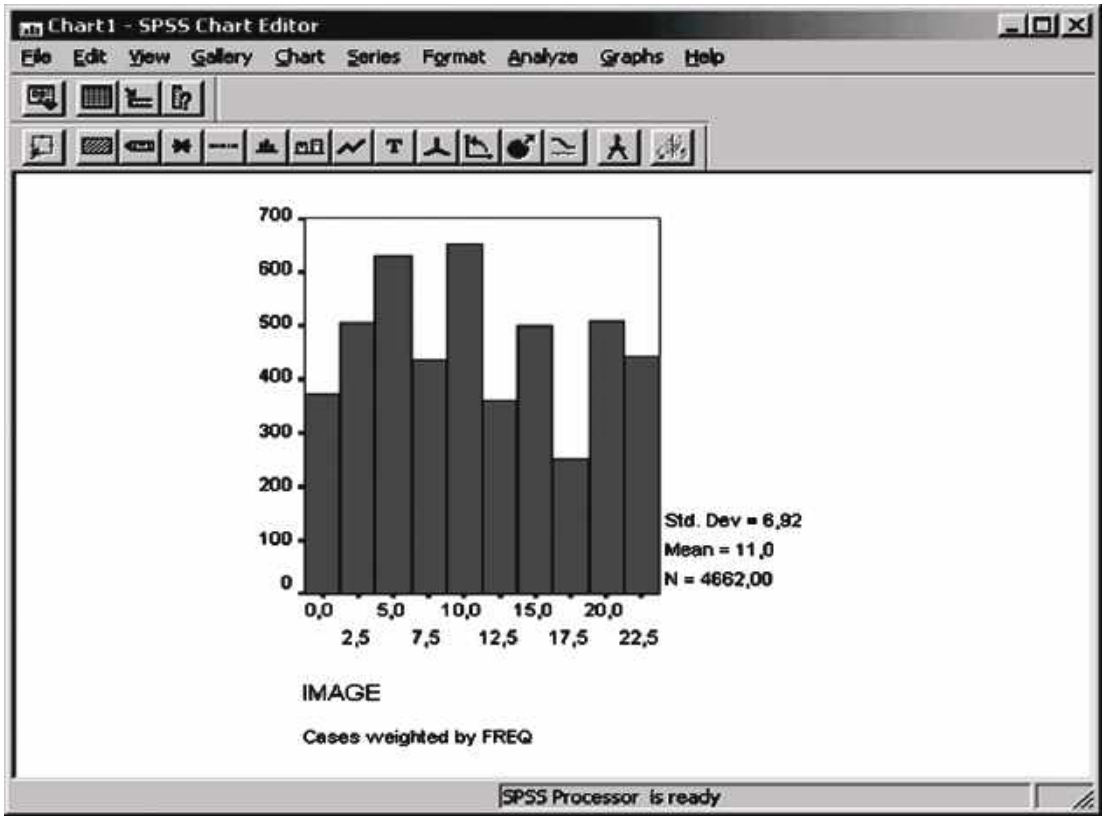


Рис. 2.116. Редактор діаграм

У верхній частині редактора діаграм є меню й дві панелі інструментів. Якщо пройти курсором по кнопках панелей інструментів, не натискаючи їх, то можна короткий опис. З допомогою кнопок верхньої панелі інструментів можна отримати інформацію про діалогові поля, які було заповнено в останніх побудованих діаграмах, перейти в редактор даних, у ньому перейти до необхідного спостереження, а також отримати інформацію про окремі змінні.

Кнопки, що розташовані на другій панелі інструментів, призначено переважно для виклику форматувальних меню. Статистичні, графічні меню й меню допомоги є вже відомими, і тому тут вони розглядатися не будуть.

Існують таки команди меню:

– File (Файл) – з допомогою цього меню побудовану діаграму можна зберегти, вивести на друк або скопіювати властивості з деякого графіка-зразка;

– Edit (Правка) – з його допомогою можна скопіювати графік в буфер обміну або змінити установки графіка;

– View (Вид) – в меню можна ввімкнути або вимкнути рядок стану й управляти панелями інструментів;

– Gallery (Галерея) – з допомогою цього меню можна вибрати інший тип графіка для відображення даних; причому в списку можна побачити деякі додаткові типи графіків, які ще не було розглянуто, наприклад змішані діаграми, діаграми зв'язувальних ліній і розділені кругові діаграми;

– Chart (Діаграми) – меню призначено для змінення зовнішнього вигляду діаграми й елементів її опису; пункти меню Options.. (Параметри), Axis.. (Осі) і Bar Spacing.. (Відстань між стовпцями) є специфічними для поточного типу діаграми. Після вибору цих опцій відкриваються відповідні діалогові вікна, зміст яких говорить сам за себе;

– Series (Ряди) – з допомогою цього меню можна змінювати подання даних, тобто стовпці на лінії або інші види графічного подання;

– Format (Формат) – якщо клацнути на цій кнопці, то отримаємо список меню (рис. 2.117).

Більшість пунктів цього меню виведено на другу панель інструментів. Замість того, щоб відкривати меню, можна просто клацнути на кнопці з відповідним символом на панелі інструментів.

З допомогою кнопки Point Id (Виділення точок)  можна змінювати режими відображення точок на діаграмі.

З допомогою кнопки Fill Pattern (Заливка візерунком)  відкриється діалогове меню, в якому можна вибрати необхідний рисунок з во-

сьми зразків заливки для зафарбовування замкнутих контурів, таких, як стовпці, області під лініями й області заднього плану.

Потрібний об'єкт виділяється клацанням на його полі. Після цього на кутах об'єкта мають виникнути маркери корекції.

Вибрати необхідний тип заливки, і клацанням на кнопці Apply (Застосувати) присвоїти його вибраному об'єкту.

Заливка білим кольором є прозорою. Цей вид заливки слід вибирати тоді, коли деяка послідовність даних має бути показана в іншій послідовності.

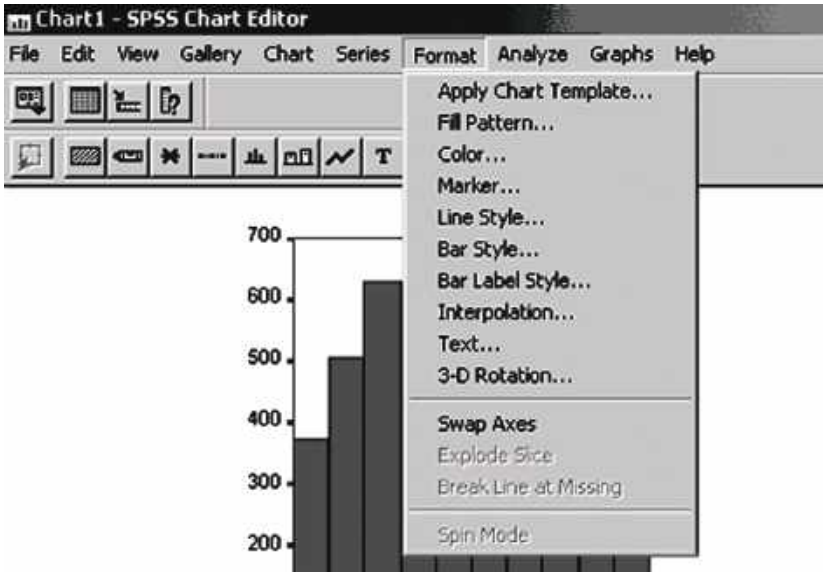


Рис. 2.117. Меню Format (Формат)

Для змінення кольору об'єкта графіка (елемента подання даних або тексту) виділити цей об'єкт і вибрати кнопку Color (Колір) . Відкриється палітра з шістнадцятьма різними кольорами. Якщо цього не вистачає, можна відкрити ще одну, значно більшу палітру.

Вибираючи опції Fill (Заливка) і Border (Рамка), можна змінити колір об'єкта або рамки (контура) виділеного об'єкта.

Вибрати одну з двох наявних опцій. З допомогою Apply (Застосувати) колір буде перенесено на виділений об'єкт.

Щоб розширити наявну палітру кольорів, треба клацнути на кнопці Edit (Правка); після цього можна створити додаткові або призначені для користувача кольори.

Якщо поточній палітрі треба присвоїти статус палітри за замовчуванням, то клацнути на вимикачі Save as Default (Зберегти як палітру за замовчуванням).

Кнопка Marker (Маркер) відкриває палітру з 28 різних маркерів для позначення розташування точки даних на лінійних діаграмах, діа-

грамах з областями й діаграмах розсіяння. Можна також установити один із чотирьох заздалегідь установлених розмірів маркерів.

Для змінення виду подання точок або рядів даних виділити спочатку необхідний елемент з допомогою клацання на графіку. Після цього на виділеному об'єкті виникнуть чорні маркери корекції.

У групі Style (Стиль) вибрати необхідне маркірування.

У групі Size (Розмір) активувати одну з опцій заздалегідь установлених розмірів маркерів. На екрані різниця між розмірами маркерів, що відображаються, є незначною, але під час друку її буде досить добре помітно.

З допомогою Apply (Застосувати) присвоїти виділеному ряду даних маркери з вибраними властивостями. Якщо натиснути кнопку Apply All (Застосувати для всіх), то вибраний тип маркірування буде присвоєно всім послідовностям даних.

Якщо змінення мають стосовуватися тільки розміру маркерів, але не стилю маркірування, то слід деактивувати опцію Apply style (Застосувати стиль).

Якщо змінення мають стосовуватися тільки стилю подання маркерів, але не розміру, то слід деактивувати опцію Apply size (Застосувати розмір).

Маркери на лінійних діаграмах і діаграмах з областями стають видимими тільки у тому випадку, якщо їх виведення буде задано в діалоговому вікні Interpolation (Інтерполяція). Це діалогове вікно викликається з меню Format (Формат). Маркери не можна задати для зображення точок гістограм і стовпчастих діаграм.

У пункті меню Line Style (Лінії)  на вибір пропонуються чотири типи ліній і чотири товщини для цих ліній.

На графіку клацанням виділити лінію, яку необхідно змінити. Після цього на об'єкті виникнуть маркери корекції.

У групі Style (Стиль) вибрати тип лінії.

У групі Weight (Товщина) присвоїти необхідну товщину вибраному типу лінії.

Після клацання на кнопці Apply (Застосувати) вибрану конфігурацію лінії буде присвоєно активному об'єкту. Ця кнопка залишається неактивною, якщо виділено дані, які не можна показати на графіку з допомогою лінії або елемента, що містить лінії (рамки, осі).

Опцію Bar Style (Стовпці)  призначено для змінення подання стовпців у графіках. Деякі типи стовпців не можна застосовувати для гістограм.

У програмі пропонується декілька типів стовпців. Якщо вибрано стовпці з тінню (Drop shadow) або 3D-ефектами (3D-effect), то для них

додатково ще можна встановити й товщину (Depth). Ця опція управляє товщиною сторін і верхнього торця стовпця. Товщина при цьому наводиться у відсотках від ширини стовпця. При додатних значеннях параметра Depth (Товщина) добитися ефекту можна, починаючи з правого боку стовпця, як показано на рисунках відповідних опцій, а при від'ємних значеннях – з лівого боку стовпця.

Якщо натиснути кнопку Apply All (Застосувати для всіх), то встановлені властивості будуть застосовані до усіх стовпців. Ця кнопка стає активною тільки тоді, коли в редакторі діаграм знаходиться Стовпчаста діаграма або інтервальна Стовпчаста діаграма.

З допомогою кнопкі Bar Label Style (Мітки столбцов)  програма пропонує три варіанти ідентифікації стовпців з допомогою числових значень.

Якщо вибрано один із стилів оформлення числового значення (окрім None), то на кожному стовпці виникає числове значення, що відповідає висоті цього стовпця. Для стовпчастої діаграми з областями мітки стовпців ставляться зверху й знизу кожного стовпця. Три опції, наведені в діалоговому вікні Bar Label Styles (Мітки стовпців), визначають зовнішній вигляд мітки на стовпці. Якщо треба застосувати темні кольори або візерунки, то рекомендується вибирати опцію Framed (У рамці), числове значення в рамці буде краще читатися.

Якщо натиснути кнопку Apply All (Застосувати для всіх), то встановлені властивості мітки будуть застосовані до усіх стовпців. Ця кнопка стає активною тільки тоді, коли в редакторі діаграм знаходиться стовпчаста діаграма, інтервальна стовпчаста діаграма або гістограма.

У діалоговому вікні Interpolation (Інтерполяція)  задаються різні можливості й методи для з'єднання точок даних.

Ця опція може застосовуватися для діаграм з областями, лінійних діаграм, лінійних діаграм різниць, для послідовностей середніх значень в діаграмах величини помилки, для завершальних показників на діаграмах максимальних і мінімальних значень, а також для діаграм розсіяння (виключаючи 3D-діаграми розсіяння).

На графіку клацанням виділити лінію або послідовність даних. Після цього на кожному об'єкті виникнуть маркери корекції.

У групі Line Interpolation (Вид інтерполяційної лінії) вибрати один із методів з'єднання точок з допомогою деякої кривої. Якщо в SPSS необхідно розрахувати регресійну пряму для діаграми розсіяння, то треба вибрати в меню Chart (Діаграми) пункт Options (Параметри).

Якщо натиснути кнопку Apply All (Застосувати для всіх), інтерполяцію буде застосовано до всіх послідовностей даних. З допомогою Apply (Застосувати) інтерполяцію буде застосовано тільки до об'єктів,

виділених у цей момент. Якщо було виділено дані, які не можна відобразити на графіку з допомогою лінії, кнопка Apply (Застосувати) стає неактивною.

Якщо активувати опцію Display markers (Показати маркери), то для кожної точки виділеної кривої буде відображено маркірування. Тип маркера можна вибрати з допомогою опції Marker (Маркер), що знаходиться в меню Format (Формат).

Існують такі види інтерполяції:

– None (Відсутній) – при виборі цієї опції з'єднання між точками немає;

– Straight (Пряма) – точки послідовно з'єднуються прямою лінією в тому порядку, в якому вони знаходяться у файлі даних.

У списку Steps (Кроки) можна вибрати один із альтернативних методів будування ступінчастої інтерполяції. Ці методи відповідають кроковим функціям, у яких точки даних з'єднуються з лівих боків, у центрах або з правих боків кроків залежно від того, чи було вибрано опцію Left step (Лівий крок), Center step (Центральний крок) або Right step (Правий крок). Кроки між собою з'єднуються вертикальними відрізками.

У списку Jump (Стрибок) можна вибрати один з методів стрибкоподібної інтерполяції. Стрибкоподібні методи будуються так само, як і покрокові, але в них немає вертикального з'єднання. Залежно від вибору Left jump (Стрибок ліворуч) Center jump (Стрибок по центру) або Right jump (Стрибок праворуч) точки даних лежатимуть з лівого боку, посередині або з правого боку горизонтальних відрізків.

У списку Spline (Сплайн) можна вибрати один з методів з'єднання точок даних з допомогою кривої:

– при виборі опції Spline (Сплайн) для з'єднання точок даних між собою будуються кубічні сплайни;

– при виборі опції 3rd-order Lagrange (Лагранж 3-го порядку) здійснюється інтерполяція, коли крива апроксимується поліномом третього порядку, що будується на основі чотирьох послідовних точок даних;

– при виборі опції 5rd-order Lagrange (Лагранж 5-го порядку) здійснюється інтерполяція, коли крива апроксимується поліномом п'ятого порядку, що будується на основі шести послідовних точок даних.

Опція Text (Текст)  надає можливість змінити шрифт і розмір текстових елементів.

Спочатку одним клацанням виділяють текст на графіку. Після цього на тексті виникають мітки корекції.

У групі Font (Шрифт) вибирають необхідний тип шрифту, а в групі Size (Розмір) – необхідний розмір. Розмір шрифту (кегль) виражається в точках.

Після клацання на кнопці Apply (Застосувати) вибрані властивості буде перенесено на виділений об'єкт. Ця кнопка стає активною тільки тоді, коли виділено текстовий об'єкт.

3D-Rotation (3D-обертання)  – це один з двох методів, з використанням яких можна повертати 3D-диаграму розсіяння. З допомогою перемикачів на лівій стороні діалогового вікна діаграму можна повертати вперед або назад відносно осей *X, Y* і *Z*.

Рисунки на перемикачах вказують на вісь і напрям повороту. Можна повертати систему координат, коротко клацаючи на відповідних перемикачах або утримуючи натиснутою кнопку миші. Поворот, що задається таким чином, відображається на спрощеній схемі, де зображено три осі; ця схема знаходиться в центрі діалогового вікна.

Якщо активовано опцію Show tripod (Показати триногу), то буде показано триногу, лінії якого проходять через центр області будування діаграми паралельно осям. Активування триноги особливо рекомендується тоді, коли необхідно прослідкувати поворот осей при вимкненому обрамленні тривимірного графіка.

Поворот виділеної діаграми відбувається з допомогою кнопки Apply (Застосувати).

Графік буде повернено тільки тоді, коли до нього буде застосовано заданий поворот. Під час операції повороту застосування яких-небудь інших команд стає неможливим.

З допомогою кнопки Swap Axes (Змінення осей)  у двовимірному графіку можна поміняти місцями вертикальну й горизонтальну осі.

Щоб висунути сегмент кругової діаграми, треба виділити його й натиснути кнопку Explode Slice (Висунути сегмент)  .

Кнопка Break Lines at Missing (Розірвати лінію в місці відсутнього значення)  розриває лінії на лінійній діаграмі за наявності значення, якого не було раніше

Кнопка Chart options (Параметри графіка)  пропонує вибір додаткових параметрів для стовпчастих і лінійних діаграм, а також діаграм з областями. У лінійних діаграмах можна також розділити лінії за категоріями.

При активуванні опції Change scale to 100 % (Перевести масштаб у відсотки) точки цих стовпчастих діаграм і частотних діаграм з областями переводяться у відсоткові показники і відображаються як відсоткові частки. Якщо діаграма, що редагується, є стовпчастою, то стовпці будуть автоматично штабельовані. Якщо на діаграмі, що редагується, стовпець або область відображає тільки один ряд даних, то ця опція залишається недосяжною. Ця опція також непридатна у разі, якщо діаграма відображає функцію накопичувальної суми.

У групі Line Options (Параметри лінії) пропонуються ще дві можливості оброблення лінійних діаграм:

– опція Connect markers within categories (З'єднати маркери усередині категорій) з'єднує маркери, які належать до одних і тих самих категорій, але лежать на різних кривих; ця опція може застосовуватися для діаграм, на яких наведено щонайменше дві криві, і не впливати на поточний статус інтерполяції або маркірування кривих;

– опція Display projection (Показати проекцію) дає можливість виділити деяку категорію, що проектується. Категорії, які розташовано праворуч від проектованої категорії, відображаються інакше.

Якщо на діаграмі у вигляді стовпців наведено щонайменше два ряди даних, то з допомогою групи Bar Type (Тип стовпців) її можна перетворити на кластеризовану або зістиковану діаграму. Якщо активовано опцію Change scale to 100 % (Перевести масштаб у відсотки), то група Bar Type (Тип стовпців) стає недоступною.

Кнопка Set/exit spin mode (Увімкнути/вимкнути режим повороту)  робить можливим безпосередній поворот 3D-діаграми розсіяння у вікні редактора діаграм; але тут під час повороту діаграма дещо спрощується.

Повертати діаграму вперед і назад відносно осей *X, Y* і *Z* можна з допомогою кнопок з відповідними символами в лівій частині діалогового вікна.

Символи на кнопках повороту вказують на осі й напрям обертання. Можна повертати область координат покроково з допомогою коротких клацань або неперервно, утримуючи кнопку миші натиснутою. Виконуваний таким чином поворот відображається з допомогою системи трьох осей у центрі вікна редактора діаграм.

БІБЛІОГРАФІЧНИЙ СПИСОК

Кулінич О. І. Теорія статистики / О.І. Кулінич. – К. : Вища шк., 1992. – 135 с.

Общая теория статистики / под ред. А.А. Спириной, О.Э. Башиной. – М. : Финансы и статистика, 1996. – 296 с.

Статистика *:* підруч. для вузів / за ред. А. В. Головача. – К. : Вища шк., 1993. – 623 с.

Теорія статистики : навч. посіб. / П. Г. Вашків, П. І. Пастер, В.П. Сторожук, Є.І. Ткач. – К. : Либідь, 2001. – 320 с.

Наследов А.Д. SPSS 15: профессиональный статистический анализ данных / А. Д. Наследов. – СПб. : Питер, 2008. – 416 с.

Бююль А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цёфель. – СПб. : ООО «ДиаСофтЮП», 2005. – 608 с.

ЗМІСТ

Навчальне видання

Петрик Валерія Леонідівна
Голованова Майя Анатоліївна,
Мельніков Сергій Михайлович

СТАТИСТИКА В SPSS

Редактор О.Ф. Серьожкіна

Зв. план, 2010
Підписано до друку 30.12.2010
Формат 60 × 84$^{1}/_{16}$. Папір офс. № 2. Офс. друк
Ум. друк. арк. 9,6. Обл.-вид. арк. 10,75. Наклад 100 прим.
Замовлення 456. Ціна вільна

Національний аерокосмічний університет ім. М.Є. Жуковського
«Харківський авіаційний інститут»
61070, Харків-70, вул. Чкалова, 17
http://www.khai.edu
Видавничий центр «ХАІ»
61070, Харків-70, вул. Чкалова, 17
izdat@khai.edu