А. П. Горбачик
С. А. Сальнікова

Аналіз даних соціологічних досліджень засобами SPSS

Волинський національний університет імені Лесі Українки
Інститут соціальних наук

А. П. Горбачик, С. А. Сальнікова

АНАЛІЗ ДАНИХ СОЦІОЛОГІЧНИХ ДОСЛІДЖЕНЬ ЗАСОБАМИ SPSS

Навчально-методичний посібник

Редакційно-видавничий відділ "Вежа"
Волинського національного університету
імені Лесі Українки
Луцьк – 2008

УДК 303.717 (072)
ББК 60.6я7
Г 67

Рекомендовано до друку вченою радою
Волинського національного університету імені Лесі Українки
(протокол № 6 від 31.01.2008 р.)

Рецензенти:
Кондратик Л. Й. *– доктор філософських наук, професор Волинського національного університету імені Лесі Українки;*
Жулькевська О. В. *– кандидат соціологічних наук, доцент Київського національного університету імені Т. Г. Шевченка.*

Горбачик А. П., Сальнікова С. А.
Г 67 **Аналіз даних** соціологічних досліджень засобами SPSS: Навч. посіб.– Луцьк: РВВ "Вежа" Волин. нац. ун-ту ім. Лесі Українки, 2008.– 164 с.
ISBN 978-966-600-334-1

У навчально-методичному посібнику представлені теоретичні основи та практичні прийоми аналізу даних емпіричних соціологічних досліджень у середовищі відомого статистичного пакету аналізу даних SPSS. Теоретичний матеріал супроводжується практичними завданнями, орієнтованими на вивчення студентами навчального курсу "Математичні методи в соціології". Поданий матеріал також може бути використаний як допоміжний під час вивчення навчальних курсів, дотичних за напрямом до соціології. Методичні рекомендації щодо використання SPSS стануть у нагоді не лише студентам-соціологам, а й студентам і фахівцям, котрі працюють в інших галузях (зокрема психологам, маркетологам), а також аспірантам, викладачам та всім, хто цікавиться обробкою й аналізом емпіричних даних.

УДК 303.717 (072)
ББК 60.6я7

ISBN 978-966-600-334-1

ЗМІСТ

ВСТУП

Уміння аналізувати дані емпіричних досліджень є обов'язковою складовою частиною сучасної фахової підготовки соціологів. Цей навчальний посібник орієнтований насамперед на студентів 2–4 курсів ВНЗ, що вивчають соціологію, але може зацікавити також соціологів-практиків, психологів, економістів, демографів і всіх тих, хто має необхідність аналізувати дані емпіричних досліджень або просто бажає набути навичок читання та розуміння соціологічних звітів і публікацій. Видання охоплює основні традиційні методи аналізу даних, які широко використовуються у повсякденній практиці соціологів – табулювання (побудова та аналіз таблиць одновимірних та двовимірних розподілів ознак), обчислення мір центральної тенденції та варіації, побудова та аналіз графіків, перевірка статистичних гіпотез (значущість різниць відсотків, середніх тощо), кореляційний аналіз та аналіз рівняння регресії.

Подано математичні поняття, що не входять до курсу математики в середній школі, але необхідні для розуміння та ефективного засвоєння матеріалу. Однак для цього потрібно, аби читач володів математичними знаннями в обсязі стандартного вступного курсу вищої математики та математичної статистики у ВНЗ і практичними навичками самостійної роботи з комп'ютером.

Книга призначена як для використання в навчальному процесі, так і для самостійного вивчення методів та прийомів аналізу соціологічних даних із використанням комп'ютера. Ефективність освоєння читачем посібника досягається і шляхом паралельного вивчення курсу з аналізу соціологічних даних.

Важливою умовою засвоєння матеріалу є постійний контакт читача з комп'ютером (робота під керівництвом викладача або самостійна) та виконання практичних завдань, поданих у кінці кожного розділу. Завдання ґрунтуються на фрагментах реальних соціологічних досліджень, зокрема: поштового опитування "Якість життя киян", проведеного відділенням соціології Інституту філософії Національної академії наук України у 1989 році, керівник проекту – В. І. Паніотто; опитування "Людина та суспільство: думки та оцінки киян", проведеного Інститутом соціології Національної академії наук України у 1991 році, автори програми – Є. І. Головаха, Н. В. Паніна, М. М. Чурилов. Обидва опитування проводилися російською мовою. Відповідні файли даних kiev89.sav та kiev91.sav з електронного банку даних ІС

НАНУ були надані авторам цієї книги для розробки прикладів та практичних завдань у навчальному курсі з використання математичних методів у соціології.

Одним із лідерів (як за можливостями, так і за поширеністю у світі) серед сучасних пакетів статистичного аналізу даних є пакет програм SPSS. Цей пакет з успіхом використовується не тільки у дослідницьких організаціях, а й у багатьох університетах світу, і, по суті, є фактичним стандартом для спеціалістів зі статистичного аналізу соціальних даних. Саме тому книга присвячена викладу основ комп'ютерного аналізу емпіричних соціологічних даних саме в середовищі SPSS. Усі приклади та завдання для самостійної роботи подаються в термінах саме цього пакету.

Навчальний посібник не замінює собою книжки з математичних методів аналізу даних та документацію з SPSS. Однак інформація, що міститься у ньому, достатня для практичного застосування математичних методів та пакету програм для розв'язування конкретних задач аналізу, що розглядаються в книзі. Саме орієнтація на розв'язок конкретних задач із конкретної предметної області є методичною особливістю видання.

РОЗДІЛ 1. ЗАГАЛЬНА ІНФОРМАЦІЯ ЩОДО РОБОТИ У СЕРЕДОВИЩІ SPSS

1.1. Основні поняття та загальні угоди

Назва SPSS утворена як абревіатура назви англійською мовою Statistical Package for the Social Science (статистичний пакет для соціальної науки). Пакет розвивається понад 40 років і став фактичним стандартом для науковців, що займаються емпіричними соціальними дослідженнями. Існує багато версій SPSS для різних типів обчислювальних машин та різних операційних систем. У 2006 р. була випущена версія SPSS 15.0 для роботи в середовищі операційної системи MS Windows. Нові версії пакету більш надійно працюють у середовищі нових версій операційної системи MS Windows, мають більше графічних можливостей, містять нові операції з аналізу даних. Інформація, вміщена в цих методичних матеріалах, значною мірою стосується лише тієї частини SPSS для MS Windows, що є незмінною, починаючи із версії 7. Подані у тексті приклади стосуються версії 10.0, але без змін можуть бути перенесені і на більш старші версії пакету.

Дані, з якими працює SPSS, мають за логічну структуру прямокутну таблицю (див. рис. 1.2). Кожен рядок цієї таблиці відповідає одному об'єкту, щодо якого збиралася інформація, і такий рядок називають спостереженням (англ. мовою *case*). Кожен стовпчик цієї таблиці відповідає одній ознаці, і такий стовпчик називають змінною (англ. мовою *variable*). Якщо йдеться про дані соціологічного опитування, то рядок відповідає респонденту, а змінна – питанню опитувальника[1]. Інакше кажучи, один рядок містить відповіді одного респондента на всі питання анкети дослідження, а один стовпчик – відповіді всіх респондентів на одне питання. В більшості випадків упорядкованість рядків у таблиці не впливає на результат статистичного аналізу. За типом значень змінні поділяються на числові та рядкові. Ми надалі будемо вести мову про числові (англ. мовою *numeric*) змінні.

Будь-який файл даних у стандартному для SPSS форматі "системного файлу" (англ. мовою *SPSS system file*) має розширення імені *.sav та містить як власне дані, так і додаткову описову інформацію, яку називають словником (англ. мовою *dictionary*). Словник

[1] Дані також, як правило, включають змінні з інформацією, яку вносить не респондент, а інтерв'юер (номер інтерв'ю, час початку та закінчення інтерв'ю тощо).

містить імена змінних, мітки змінних, мітки окремих значень, формати представлення значень тощо.

Кожна змінна обов'язково має ім'я (англ. мовою *variable name*). Звернутися до змінної можна лише за іменем. В одному файлі даних не може бути двох змінних з однаковими іменами. Ім'я змінної має довжину не більше ніж 8 знаків. У запису імені змінної можна використовувати латинські літери (від A до Z), цифри (від 0 до 9) та деякі спеціальні знаки (зокрема крапку, підкреслення, знаки @, #, $). Першим знаком у запису імені має бути латинська літера або знак @. Крапка не може бути останнім знаком у запису імені. Великі та малі латинські літери у запису імен не розрізняються. Використовувати літери українського алфавіту, пробіли, знак коми та більшість інших знаків розділення під час запису імен не можна. Ім'я змінної не повинно збігатися з зарезервованими ключовими словами мови запису програм SPSS: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.

Кожна змінна може мати мітку (англ. мовою *variable label*). Мітка змінної – це текст, максимальна допустима довжина якого залежить від версії пакету, але принаймні є не меншою ніж 120 знаків. Для запису цього тексту можна використовувати будь-які знаки, включаючи літери українського алфавіту, пробіли, розділові знаки тощо. Пакет SPSS не аналізує мітки змінних, і ці мітки ніяк не впливають на результати обчислень. Однак чіткі та зрозумілі мітки змінних дуже полегшують роботу з великим файлом даних, а також полегшують сприйняття дослідником результатів обчислень. Наявність добре сформульованих міток змінних є ознакою професіоналізму дослідника, що готував файл даних. Коли файл містить результати опитування, то часто як мітки змінних використовують скорочені формулювання відповідних питань анкети.

Окремі значення дискретної змінної можуть мати мітки (англ. мовою *value label*). Аналогічно до міток змінної мітка значення є текстом, максимально допустима довжина якого залежить від версії пакету, але принаймні є не меншою ніж 60 знаків. Для запису цього тексту можна використовувати будь-які знаки, включаючи літери українського алфавіту, пробіли, розділові знаки тощо. Пакет SPSS не аналізує мітки значень, і ці мітки не мають ніякого впливу на результати обчислень. Однак чіткі та зрозумілі мітки змінних дуже полегшують роботу із великим файлом даних, а також полегшують сприйняття дослідником результатів обчислень. Коли файл містить

результати опитування, то часто як мітки значень використовують скорочені формулювання варіантів відповідей на питання анкети, що їх пропонує респонденту автор опитувальника.

Із певних причин для деяких спостережень значення змінної може бути невідомим. Наприклад, респондент, що не має телевізора, не дав нам відповіді на питання стосовно вечірнього випуску новин, оскільки це питання цьому респонденту просто не ставилося. В такому випадку говорять про відсутнє значення (англ. мовою *missing value*). Більшість операцій пакету SPSS мають свої особливості у роботі з відсутніми значеннями, і з цими особливостями ми ознайомимося під час розгляду відповідних операцій. Розрізняють два види відсутніх значень: системне відсутнє значення (англ. мовою *system missing value*) та значення, що об'явлене користувачем як відсутнє (англ. мовою *user missing value*). Системне відсутнє значення є універсальним (тобто може бути значенням будь-якої змінної будь-якого типу) і, як правило, результатом некоректних обчислень або перетворень (зокрема ділення на нуль, добування кореня квадратного із від'ємного числа тощо). Якщо значенням змінної є системне відсутнє значення, то у відповідній клітині електронної таблиці редактора даних відображається лише крапка або кома (залежно від того, який із цих двох знаків використовується для відділення дробової частини числа від цілої). У випадку ж, коли дослідник передбачив у своєму вимірювальному інструменті можливість відсутності значення (наприклад альтернативу "відмова відповідати" як варіант відповіді на певне питання) і закодував такі відсутні значення певним числовим кодом, то є можливість об'явити конкретні коди для цієї змінної кодами відсутніх значень, і операції пакету SPSS будуть відповідно інтерпретувати ці коди. Саме в такому випадку ми маємо справу із другим типом відсутніх значень. Пакет SPSS дає змогу для кожної змінної описати або не більше трьох конкретних значень, що мають бути інтерпретовані операціями пакету як відсутні значення, або ж одне значення та один інтервал значень, що також інтерпретуються як відсутні.

Основні угоди про імена файлів для роботи із SPSS є такими:

– файли з даними у форматі пакету SPSS (*SPSS system file*) мають стандартне розширення SAV;

– файли з послідовностями команд (програмами), записаними мовою SPSS (*syntax file*), мають стандартне розширення SPS;

– файли з результатами обчислень пакету SPSS (*output file*) мають стандартне розширення SPO.

1.2. Встановлення та запуск SPSS

Пакет SPSS розповсюджується записаним на компакт-диску і має бути перед початком використання встановлений на комп'ютері. Стандартна процедура встановлення передбачає розміщення основної частини пакету в підкаталозі SPSS стандартного каталогу Program Files. Головний файл пакету, що його потрібно запускати, міститься саме в цьому підкаталозі і має ім'я spsswin.exe. Процедура встановлення "зв'язує" розширення SAV, SPS та SPO із файлом spsswin.exe. Це означає, якщо двічі клацнути на файлі, що має одне з цих трьох розширень, то пакет SPSS запускається та читає відповідний файл.

Існує кілька способів запуску пакету SPSS:

- знайти потрібний файл з даними (наприклад kiev89.sav) та двічі клацнути на ньому лівою кнопкою миші;
- як і більшість встановлених пакетів, SPSS можна запустити через головне меню запуску:

Пуск ⇒ Программы ⇒ SPSS for Windows;

- у тому підкаталозі, де встановлений SPSS, потрібно знайти файл spsswin.exe та двічі клацнути на ньому лівою кнопкою миші;
- знайти на робочому столі ярлик із піктограмою, зображеною праворуч, та двічі клацнути на ній лівою кнопкою миші. (*Зауваження*. У різних версіях пакету піктограма може мати дещо інший вигляд).

Якщо на робочому столі немає відповідного ярлика, то його можна досить легко створити. Для цього потрібно виконати такі дії:

– клацніть правою кнопкою миші на вільному місці робочого столу Windows та оберіть у контекстному меню, що з'явилося, **Создать ⇒ Ярлык**;

– введіть у діалоговому вікні, що з'явилося, повне ім'я (включаючи і повний шлях) файлу запуску пакету. Якщо пакет встановлений стандартно, то йдеться про "C:\ProgramFiles\SPSS\ spsswin.exe". Можна також скористатися для пошуку місцезнаходження цього файлу кнопкою **[Обзор]**. Ця кнопка відкриває структуру каталогів, у якій потрібно знайти основний файл запуску пакету spsswin.exe. Потім натисніть кнопку **[Далее]**;

– у вікні **Выбор названия программы**, що відкривається, знайдіть поле *Укажите название ярлыка* та введіть відповідний текст (наприклад "SPSS"). Потім натисніть кнопку **[Готово]**.

У всіх вказаних вище випадках запуску пакету, окрім першого, після старту пакету SPSS одразу з'являється вікно:

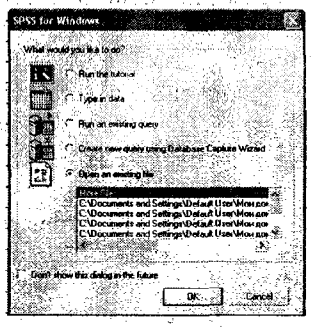


Рис. 1.1. *Стартове меню*[2].

У цьому вікні можна обрати один з п'яти варіантів початку сеансу роботи із SPSS, а саме:

- **Run the tutorial** – запустити підручник, в якому викладені основи роботи з програмою.[3]
- **Type in data** – відкрити пусте вікно редактора даних для створення нового файлу даних.
- **Run the existing query** – завантажити раніше сформований запит на імпорт у програму даних в іншому форматі.
- **Create new query using Database Capture Wizard** – створити новий запит на імпорт у програму даних в іншому форматі із використання відповідного майстра.
- **Open an existing file** – обрати для роботи файл даних із переліку "останніх використаних" (тобто тих, із якими вже працювали на цьому комп'ютері за допомогою SPSS).

Якщо поставити т. зв. галку біля **Don't show this dialog in the future**, то при подальших запусках SPSS відповідне вікно більше не з'являтиметься, а буде відкриватися пусте (тобто без даних для аналізу) вікно редактора даних.

---

[2] Вікна, зображені на цьому та інших рисунках, можуть дещо різнитися залежно від встановленої на вашому комп'ютері версії програми SPSS, але така різниця є незначною і не обмежує Ваших можливостей у подальшому процесі обробки та аналізу даних.

[3] Вбудований підручник є англомовним у нерусифікованих версіях.

1.3. Робочі вікна SPSS

Як і в будь-якій програмі, що працює в середовищі MS Windows, робота в пакеті SPSS ведеться через вікна. Їх є три різних типи:

1. Вікно редактора даних, що має назву **Data Edito.**
2. Вікно результатів обчислень, що має назву **Output**.
3. Вікно програм мовою SPSS, що має назву **Syntax**.

Вікна всіх трьох типів можуть бути відкритими одночасно. Інформація у вікні Data Editor після завершення сеансу роботи може бути збережена у файлі даних із розширенням SAV. Інформація у вікні Output після завершення сеансу роботи може бути збережена у файлі даних із розширенням SPO. Інформація у вікні Syntax після завершення сеансу роботи може бути збережена у файлі даних із розширенням SPS. Вікон Syntax та Output може бути відкрито декілька. Вікно Data Editor можна відкрити тільки одне. Це означає, що у будь-який момент часу можна працювати лише з одним файлом даних, але до нього можна застосовувати кілька програм, записаних мовою SPSS, та результати обчислень можна виводити в одне з декількох відкритих вікон із результатами.

Переключатись між вікнами можна стандартно – або через панель завдань, або через пункт Windows головного меню пакету. Після завершення сеансу роботи пакет пропонує зберегти спочатку інформацію з вікон Syntax та Output, а потім файл даних із вікна Data Editor.

1.3.1. Редактор даних (Data Editor)

Пусте (тобто за відсутності даних для аналізу) вікно редактора даних **Data Editor** має такий вигляд:

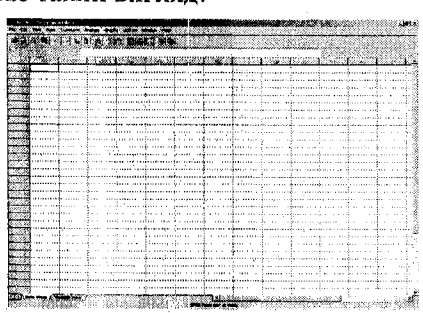


Рис. 1.2. *Вікно редактора даних* ***Data Editor*** *пакету SPSS.*

У більшості випадків саме вікно редактора даних відкривається після запуску SPSS. Це вікно має такі ж стандартні елементи, що й вікно офісних електронних таблиць Excel:

- на верхній частині рамки вікна містяться системна кнопка, назва відкритого файлу даних та стандартні кнопки мінімізації, максимізації та закриття вікна;
- у верхній частині вікна розташований рядок головного меню пакету;
- під рядком головного меню розміщується панель інструментів;
- під панеллю інструментів є рядок, у лівій частині якого вказана адреса виділеної клітини, а в правій частині – вміст цієї клітини;
- основну частину вікна редактора даних займає електронна таблиця;
- вікно має горизонтальну та вертикальну смуги прокрутки;
- на нижній частині рамки вікна вміщено рядок статусу, в якому виводяться певні повідомлення процедур пакету (наприклад кількість опрацьованих спостережень) та інформація про стан ряду важливих поточних параметрів роботи (зокрема інформація про те, чи встановлений фільтр – індикатор *Filter On*, чи зваженими є дані – індикатор *Weight On* тощо).

1.3.2. Вікно результатів обчислень (Output)

Результати обчислень подаються у вікні Output у вигляді дерева. Вікно Output є розділеним на дві частини. Ліва частина вікна містить стандартне для MS Windows графічне зображення деревоподібної структури, що є зручним для швидкого пересування по результатах обчислень. Вузли цього дерева розмічені назвами процедур, які застосовувалися до даних, а також заголовками таблиць та інших структурних елементів результатів обчислень. Самі результати є листям цього дерева і вміщені у правій частині вікна Output. Загалом робота у вікні Output схожа на роботу в програмі *Проводник*, лише там ми переглядаємо папки і файли, а тут – процедури та результати застосування цих процедур.

В усьому іншому вікно результату є схожим на вікно редактора даних. Можливості редагування у вікні результату ми розглянемо пізніше, коли навчимося застосовувати до даних принаймні найпростіші процедури статистичного аналізу.

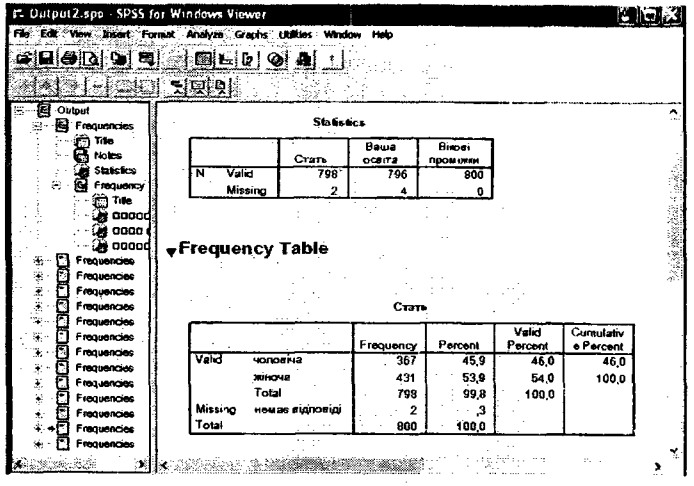


Рис. 1.3. *Вікно результатів обчислень Output.*

Зауважимо лише, що вертикальну межу, яка розділяє ліву (із деревом) та праву (із результатами обчислень) частини вікна результатів, можна зсувати і саме так *змінювати розміри* частин вікна. Для цього потрібно поставити на межу курсор миші та, коли він набере вигляду двоспрямованої стрілки, "протягнути" курсор у потрібному напрямі (вліво або вправо).

1.3.3. Послідовність операцій (Syntax)

Редактор **Syntax** використовують для введення і запуску на виконання команд SPSS, тобто програмування мовою SPSS методів обробки та аналізу даних згідно з чітко прописаними синтаксичними командами. Послідовність операцій (програма мовою SPSS) записується як звичайний текст. У відповідне вікно можна зчитати файл із програмою (нагадаємо, такий файл має мати розширення SPS) або ж у такому вікні можна безпосередньо вводити і редагувати текст програми мовою SPSS. Під час роботи у вікні **Syntax** можна використовувати всі стандартні для MS Windows засоби та прийоми редагування (як у стандартному для Windows текстовому редакторі Notepad).

Отже, вводити команди можна безпосередньо у вікні редактора синтаксису (за допомогою клавіатури), вибравши в головному меню

File ⇒ New ⇒ Syntax,

або просто перенести виставлені параметри діалогових вікон конкретних процедур аналізу за допомогою кнопки **[Paste]** (*Вставить*), яка міститься в тому ж діалоговому вікні[4].

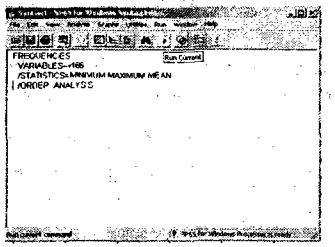


Рис. 1.4. *Вікно редактора синтаксису* ***Syntax.***

На панелі інструментів вікна **Syntax** є кнопка **[Run Current]** (*Выполнить текущую команду*), позначена трикутником (див. рис. 1.4), спрямованим вправо. Для того, щоб виконати певну послідовність операцій, потрібно виділити стандартними засобами Windows відповідну частину тексту програми (тобто одну або кілька операцій) та натиснути кнопку **[Run Current]**. Отже, програмний файл SPSS може містити не одну команду. Запустити їх на виконання можна поодинці або ж одночасно всі, після чого відкривається вікно **Output** з результатами замовлених процедур. Так можна скласти певний "сценарій" для того, щоб автоматизувати виконання деяких задач. Причому таких "сценаріїв" можна створити кілька, створені файли синтаксису після завершення сеансу роботи варто зберегти (процедура збереження стандартна) та використовувати у подальшій роботі. Збережені файли, як вказувалося раніше, матимуть розширення SPS.

Власне писати програми мовою SPSS ми будемо пізніше – під час виконання конкретних практичних завдань. А зараз на прикладі, зображеному на рис. 1.4, запишемо правила створення командного синтаксису, якими користуватимемося надалі. Отже, у вікні синтаксису на рис. 4 запрограмована на виконання процедури FREQUENCIES (*Частоти*)[5] для VARIABLES=v166 (змінної "вік"), для цієї змінної замовлено також статистики STATISTICS = мінімум, максимум та середнє.

[4] Реалізацію різних процедур аналізу послідовно розглядатимемо, починаючи з третього розділу.

[5] Інший спосіб виконання цієї процедури буде показано у п.3.1 розділу 3.

Елементи кожної команди програмної мови SPSS поділяють на такі категорії:

- *Команда* – це власне інструкція, яка керує процесом роботи SPSS.
- *Допоміжна команда* – додаткова інструкція до команди SPSS. До однієї команди може входити кілька допоміжних команд.
- *Специфікації* – дані, які доповнюють команду або допоміжні команди. У специфікаціях можуть бути ключові слова, цифри, арифметичні операції, імена змінних і спеціальні розділові знаки.
- *Ключові слова* – це слова, які мають у мові програмування SPSS чітко визначене значення.

У прикладі, який ми розглядаємо, маємо: FREQUENCIES – команда; VARIABLES, STATISTICS, ORDER – допоміжні команди, після знака рівності у цих командах йдуть різні специфікації, серед яких MINIMUM, MAXIMUM, MEAN, ANALYSIS – ключові слова.

Створюючи та редагуючи командний синтаксис, потрібно дотримуватися таких правил:

– кожна команда починається з нового рядка, після запису синтаксису однієї команди в кінці ставиться крапка. Тобто одна команда – це одне речення;

– команда може містити будь-яку кількість рядків, але кожен наступний рядок однієї команди повинен починатися як мінімум з одиничного пробілу;

– рядок із командним синтаксисом не повинен перевищувати 80 символів;

– перед кожною (окрім, можливо, першої) допоміжною командою повинен стояти знак у вигляді похилої лінії /;

– перейти на новий рядок або ввести один пробіл можна перед і після знака похилої лінії /, дужок, арифметичних операторів або між іменами змінних;

– для ідентифікації міток можна використовувати текст, який записується в лапках і повинен бути в одному рядку;

– у специфікаціях числового формату як десятковий розділювач повинна використовуватися крапка (.), незалежно від встановлених опцій операційної системи Windows[6];

[6] Про те, як вказати символ крапки (.) як десяткового розділювача, вказано в п.1.5 цього ж розділу.

– вводити синтаксис можна будь-яким регістром: програма не розрізняє великі та малі літери.

1.4. Огляд пунктів головного меню пакету

Як це є стандартним для програм, що працюють у середовищі MS Windows, усі операції пакету SPSS організовані у вигляді горизонтального меню. Коротко розглянемо основні пункти цього меню.

- **File** – тут зосереджені стандартні операції читання та запису файлів, експорту та імпорту даних, друкування тощо. Основні операції з цього пункту меню:
  - **New** ⇒
    - очистити вікно редактора даних **New ⇒ Data**,
    - відкрити нове вікно для введення та редагування програм **New ⇒ Syntax**,
    - відкрити нове вікно для виведення результатів обчислень **New ⇒ Output**.
  - **Open** ⇒
    - прочитати файл з даними **Open ⇒ Data**. Операція дає змогу читати файли даних не тільки у форматі системного файлу SPSS (із розширенням SAV), а й у багатьох інших форматах,
    - прочитати текстовий файл із послідовністю команд SPSS **Open ⇒ Syntax**,
    - прочитати файл із результатами раніше виконаних обчислень **Open ⇒ Output**.
  - **Close** – закрити активне вікно. Проте необхідно зазначити, що вікно редактора даних **Data Editor** закрити не можна.
  - **Save** – записати вміст активного вікна (дані, результати обчислень, тексти програм) у файл на диск під тим іменем, з яким файл був у активному вікні прочитаний.
  - **Save as** – записати вміст активного вікна у файл на диск під іншим іменем або в інший каталог.
  - **Print** – видрукувати вміст активного вікна.
  - **Stop Processor** – терміново зупинити виконання обчислень.

- **Exit** – завершити сеанс роботи із пакетом. За необхідності, під час завершення роботи програма SPSS може пропонувати зберегти на диску вміст окремих вікон.

▪ **Edit** – стандартні операції редагування, зокрема такі:

- **Cut** – вирізати виділений текст або комірки з даними у буфер обміну MS Windows.
- **Copy, Copy Objects** – копіювати виділений текст або об'єкт у буфер обміну MS Windows.
- **Past** – вставити текст або об'єкт з буфера обміну MS Windows.
- **Clear** – видалити виділений текст, об'єкт або очистити діапазон комірок.
- **Options** – зміна параметрів роботи пакету SPSS (докладніше дивись нижче).

▪ **View** – група операцій зміни зовнішнього вигляду робочих вікон SPSS, настроювання панелі інструментів, шрифтів тощо. Зокрема, цей пункт головного (горизонтального) меню містить такі операції та параметри зовнішнього вигляду робочих вікон:

- **Status bar** (*Рядок стану*) – параметр, що визначає, чи потрібно або не потрібно показувати рядок стану (рядок на нижній частині рамки вікна, що містить інформацію про ряд параметрів роботи пакету).
- **Toolbars...** (*Панель символів*) – операція, що дає змогу редагувати панель інструментів, виносити на панель кнопки із операціями, що часто застосовуються.
- **Fonts...** (*Шрифти*) – операція, що дає змогу обрати потрібний для роботи шрифт, встановити розмір літер тощо.
- **Grid lines** (*Лінії сітки*) – параметр, що визначає, чи потрібно або не потрібно показувати лінії, які обмежують комірки електронної таблиці редактора даних.
- **Value labels** (*Мітки значень*) – параметр, що визначає, чи потрібно виводити в комірках електронної таблиці редактора даних безпосередньо значення, чи потрібно виводити відповідні мітки значень.

▪ **Data** – група операцій роботи із файлами даних. Зокрема, тут містяться операції, що дають змогу впорядковувати об'єкти за зна-

ченнями однієї або декількох змінних (**Sort cases...**), об'єднувати декілька файлів даних у один (**Merge files**), створювати файл даних, що характеризують визначені групи об'єктів спостереження (**Aggregate...**), відбирати об'єкти для аналізу за вказаною умовою або випадково (**Select cases...**), зважувати файл даних з метою ремонтування вибірки або моделювання вибірки із певними необхідними параметрами (**Weight cases...**).

▪ **Transform** – група операцій зміни значень змінних та обчислення нових змінних у файлі даних. Зокрема, тут містяться операції, що дають змогу створювати у файлі даних нові змінні шляхом обчислення за певними формулами (**Compute...**) та шляхом перекодування (**Recode**).

▪ **Analyze** – операції статистичного аналізу даних. Саме задля виконання цих операцій і створений пакет SPSS. Операції з інших пунктів меню є необхідними для роботи, але все ж таки в певному сенсі допоміжними, спрямованими на підготовку даних для аналізу. Перелік операцій, включених до цього пункту меню, залежить від того, які модулі пакету SPSS встановлені.

▪ **Graphs** – операції створення на основі файлу даних та редагування графіків різного типу (гістограм, полігонів, секторних діаграм тощо), що характеризують розподіл окремих змінних, сумісний розподіл кількох змінних, особливості характеристик змінних тощо.

▪ **Utilities** – допоміжні операції. Зокрема, тут містяться операції, що дають змогу переглянути всю інформацію із словника файлу даних, що описує певну змінну (**Variables...**), а також операції створення та використання множин змінних (Define Sets... та Use Sets...), які є особливо зручними під час аналізу груп змістовно пов'язаних між собою дихотомічних змінних (зокрема груп дихотомічних змінних, що містять відповідь на питання, яке допускає можливість респонденту обрати більше ніж один варіант відповіді).

▪ **Windows** – стандартні для програм, що працюють у середовищі MS Windows, операції управління вікнами програми.

▪ **Help** – довідкова інформація про пакет SPSS, про правила роботи із пакетом, про реалізовані в пакеті методи аналізу даних, про правила запису програм мовою SPSS тощо.

Зауваження. Процедура збереження файлів у програмі SPSS стандартна, але особливість збереження *нового!* файлу даних полягає в тому, що, задавши ім'я, потрібно *обов'язково!* вказати каталог (папку), в яку потрібно зберегти файл даних, інакше за загальною угодою програма збереже новостворений файл у каталог SPSS, що міститься у Program Files; з часом кількість файлів у папці SPSS збільшиться, їх переміщення або видалення може призвести до некоректної роботи самої програми.

1.5. Основні параметри SPSS

Наступна інформація не буде достатньо зрозумілою для початківців, але корисною в міру набуття ними практичних навичок роботи у програмі SPSS. До викладеного в цьому пункті матеріалу можна звертатися на будь-якому етапі аналізу даних.

Для зміни системних параметрів програми SPSS виберіть у меню

Edit (*Правка*) ⇒ **Options...** (*Параметри*).

Відкриється діалогове вікно **Options** (див. рис. 1.5), в якому є десять вкладок. Змінювати системні настройки простому користувачу небажано, а тому розглядатимемо лише ті опції у вкладках, які потрібні для спрощеної роботи в SPSS і які не призведуть до збоїв у роботі програми.

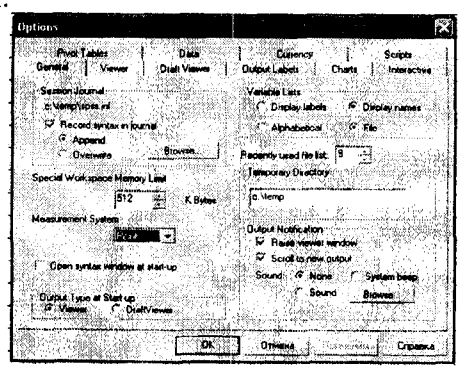


Рис. 1.5. *Діалогове вікно* ***Options****, вкладка* ***General.***

▪ **General** (*Загальне*): тут важливою є група **Variable Lists** (*Списки змінних*). У ній можна задати тип сортування списку змінних. За загальною угодою змінні виводяться в порядку їх уведення в робочому файлі (**File**), але можна задати сортування в алфавітному порядку (**Alphabetical**). Також важливо задати те, що саме вказувати

в усіх діалогових вікнах, – мітки змінних (**Display labels**) чи імена змінних (**Display names**), за загальною угодою виводяться імена змінних.

У групі **Session Journal** вказано шлях до файлу, в якому ведеться журнал синтаксису (запис відбувається, якщо не знімати галочку у вікні **Record syntax in journal** згідно з загальною угодою).

▪ **Viewer** (*Вікно перегляду*): у цій вкладці встановлюють тип і розмір шрифту заголовків (**Title Font**) і тексту (**Text Output Font**), що відображаються у вікні перегляду результатів **Output**, а також задають розміри сторінки (**Text Output Page Size**).

▪ **Draft Viewer** (*Вікно текстового режиму*): на цій вкладці присутні різні установки зовнішнього вигляду таблиць і тексту.

▪ **Output Labels** (*Позначення вивідних значень*): у вікні, зображеному на рис. 1.6, є дві групи. У першій для позначення категорій змінної можна вибрати значення змінної (**Value**) або мітку значення (**Labels**) (опція за загальною угодою), чи обидва варіанти одночасно (**Value and Labels**). У другій групі аналогічне позначення вивідних значень можна вибрати для мобільних таблиць (**Pivot Table Labeling**).

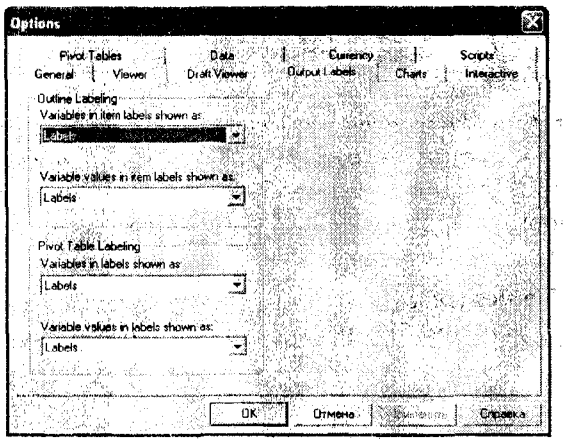


Рис. 1.6. *Діалогове вікно* ***Options****:* ***Output Labels.***

▪ **Charts** (*Діаграми*): у розділі **Current Setting** (*Зміна оформлення*), крім шрифту (**Font**), можна задати відображення різних стовпчиків, ліній, областей графіка різними кольорами (**Cycle through colors, then patterns**) (опція за загальною угодою) або за допомогою різного

штрихування і відповідних типів ліній (**Cycle through patterns**), що є важливим у випадку внесення графіків до звіту, друкований варіант якого виконується у чорно-білому варіанті. Можна також керувати компонуванням рамки (рамка зовні (**Outer**) чи всередині (**Inner**)) і організовувати відображення координатної сітки (**Grid Lines**): замовити додаткові лінії по порядковій осі (**Scale axis**) і/або по осі категорій (**Category axis**).

Ця вкладка є важливою у випадку, коли графічне подання деяких результатів виконується у програмі SPSS і виведення наочностей різниться згідно з програмою соціологічного дослідження.

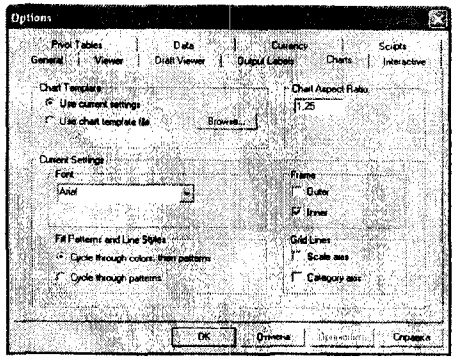


Рис. 1.7. *Діалогове вікно* ***Options****: Charts.*

- **Interactive** (*Інтерактивний режим*): вкладку використовують для вибору параметрів інтерактивних графіків, можна задати, наприклад, деякий зразок. Якщо побудовані діаграми роздруковують, потрібен чорно-білий режим, для цього зручним є зразок *Grayscale.clo* (*Відтінки сірого*).

- **Pivot Tables** (*Мобільні таблиці*): тут можна вибрати зовнішній вигляд (компоновку) мобільних таблиць у вікні перегляду результатів **Output**.

- **Data** (*Дані*): у цій вкладці (див. рис 1.8) можна змінити формат подання змінних у редакторі даних **Data Editor** (за загальною угодою – це 8 позицій (**Width**) з двома знаками після коми (**Decimal Places**)). Для зазначення року двома останніми цифрами можна додатково вказати століття. Якщо активувати автоматичну опцію (**Automatic**), то століття буде розраховуватися в межах від 1937 до 2036. Інакше – як звичайно (**Custom**).

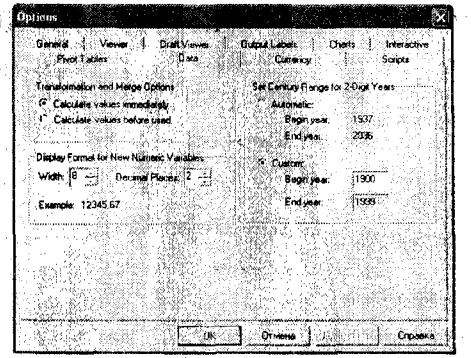


Рис. 1.8. *Діалогове вікно* ***Options****: Data.*

- **Currency** (*Грошова одиниця*): тут можна вибрати грошовий формат, розділювач цілої і десяткової частин числа: або крапку (**Period**) (за загальною угодою), або кому (**Comma**).

- **Scripts** (*Сценарії*): для активації автоматичних сценаріїв.

Зауваження. Зміна системних опцій набуває чинності після повторного завантаження програми SPSS.

Контрольні запитання

1. Як запускається програма SPSS? Вкажіть усі можливі способи запуску.
2. Назвіть основні типи вікон пакету SPSS.
3. Чи можливим є одночасно аналізувати дані з кількох файлів?
4. Охарактеризуйте логічну структуру файлу даних. Дайте визначення змінної та спостереження.
5. Яку інформацію містить словник файлу даних?
6. Сформулюйте правила запису імен змінних.
7. Що таке "відсутні значення" у файлі даних? Які типи відсутніх значень Вам відомі?
8. Для чого використовуються мітки змінних? Якими є правила запису мітки змінної?
9. Для чого використовуються мітки значень? Якими є правила запису мітки значень?

РОЗДІЛ 2. СТВОРЕННЯ ТА РЕДАГУВАННЯ ФАЙЛУ ДАНИХ

2.1. Підготовка словника файлу даних

Створення нового файлу даних розпочинається зі створення словника. Дослідник, що планує введення файлу даних, має визначити кількість змінних у майбутньому файлі даних та задати для кожної змінної необхідну описову інформацію. Інакше кажучи, ще до початку введення самих даних необхідно попередньо сформувати словник майбутнього файлу. Введенням даних (перенесенням інформації із заповнених інтерв'юерами або респондентами паперових анкет до комп'ютера) можуть займатися оператори, необізнані із метою подальшого аналізу даних та із завданнями дослідження. В той же час створенням словника має займатися сам дослідник, який розуміє специфіку майбутнього аналізу даних та обізнаний із відповідними можливостями SPSS.

Нагадаємо, що дані, з якими працює SPSS, мають структуру прямокутної таблиці (див. рис. 1.2 пункту 1.3.1). Перед початком формування структури файлу потрібно:

- визначити кількість стовпчиків таблиці (тобто кількість змінних у майбутньому файлі);
- визначити впорядкованість змінних у файлі та дати коротку назву (ім'я) кожній змінній;
- сформулювати мітку для кожної змінної (текст, що пояснює сенс відповідної змінної). Якщо створюється файл із даними опитування, то як мітки часто використовують скорочені або навіть повні формулювання відповідних питань анкети;
- визначити тип шкали змінної. Для дискретної змінної визначити перелік можливих значень, кодування та мітки цих значень;
- визначити кодування для відсутніх значень, якщо в цьому є потреба.

Є кілька різних можливостей введення словника нового файлу даних. Найбільш ефективною, на нашу думку, є така послідовність дій:

➢ у нижній частині вікна редактора даних є дві закладки – **Data View** (*Перегляд даних*) та **Variable View** (*Перегляд змінних*). Обираємо закладку **Variable View** і починаємо створювати та редагувати перелік змінних файлу;

➢ кожен рядок переліку містить інформацію про одну змінну. Змінні у перелік вносимо у попередньо визначеному порядку;

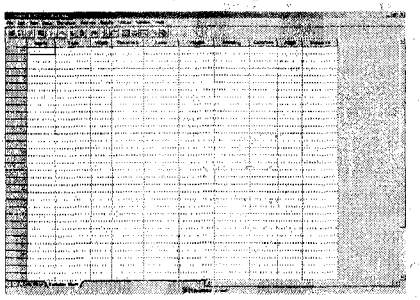


Рис. 2.1. *Закладка **Variable View**.*

➢ у стовпчик **Name** (*Ім'я*) заносимо ім'я змінної. Нагадаємо, що в запису імені змінної можна використовувати латинські літери (великі та малі літери не розрізняються), цифри та лише деякі спеціальні знаки (зокрема такі, як точка ".", підкреслення "_", позначення грошей "$", знак "@"); не можна використовувати літери української абетки; ім'я має починатися із будь-якої латинської літери і не може закінчуватися точкою; довжина імені має бути не більше 8 знаків.

Якщо кількість змінних надто велика, їх можна йменувати як математичні змінні: v1, v2, v3, v4_1, v4_2, v4_3, v5, … , тобто кожна змінна має вигляд vi, де i – номер питання в анкеті; якщо ж питання "розщеплюється" на кілька змінних, то матимемо вигляд vi_j, де j – кількість змінних i-го питання. Наприклад, питання № 4 ділиться на три змінні. Така ситуація трапляється тоді, коли на одне питання пропонується вибрати не одну відповідь. Хоча залежно від цілей дослідження та застосування того чи іншого виду аналізу даних питання з можливістю обрання одночасно кількох альтернатив[7] кодуються по-різному. А оскільки програма SPSS не розрізняє такий тип номінальної шкали, то маємо такі варіанти:

1. Кожен із варіантів відповідей розглядається як окрема дихотомічна змінна типу vi_j, яка набуває значення "0", якщо варіант не обраний, і "1" – варіант обраний. Цей спосіб кодування дає змогу отримати точну і детальну інформацію, але кількість змінних при цьому значно зростає.

[7] Шкали такого типу називаються номінальними із сумісними альтернативами.

2. Кожному з варіантів відповідей приписується число натурального ряду від 1 до n, де n – кількість альтернатив; у значення змінної вносяться ті цифри, варіанти відповідей яких були обрані. Отримуємо одну змінну типу v*i*. Наприклад, значення змінної "235" означає, що із запропонованих варіантів респондент обрав 2-гу, 3-тю та 5-ту альтернативи. Такий спосіб спрощує аналіз даних, але ускладнює роботу з результатами обробки і звітом. Крім того, кількість варіантів відповідей не повинна бути надто великою;

- у стовпчик **Label** (*Мітка*) заносимо мітку змінної – текст, що не впливає на результати аналізу, але пояснює сенс змінної та поліпшує орієнтацію у великому файлі даних та полегшує сприйняття результатів обчислень. У запису мітки можна використовувати будь-які знаки, включаючи пробіли та літери української абетки;
- для дискретної змінної у стовпчику **Values** (*Значення*) визначаємо перелік можливих значень та мітки цих значень. Для цього потрібно клацнути мишею у стовпчику **Values** рядка відповідної змінної, у правій частині клітини з'являється три крапки "...", потрібно клацнути на цих трьох крапках і з'являється вікно **Value Labels** (*Мітки значень*). У поля **Value** (*Значення*) та **Label** (*Мітка*) цього вікна вносимо відповідно значення (певне число) та мітку (текст, правила запису якого не відрізняються від правил запису мітки змінної) і, користуючись кнопкою **[Add]** (*Додати*), додаємо мітку та значення до списку. Кнопки **[Change]** (*Змінити*) та **[Remove]** (*Видалити*) використовують для редагування списку можливих значень дискретної змінної;

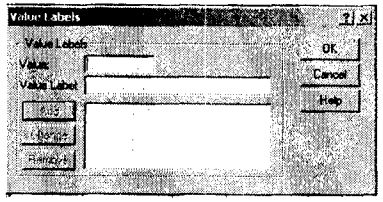


Рис. 2.2. *Вікно* ***Value Labels.***

- визначаємо у стовпчику **Missing** (*Відсутній*) кодування для відсутніх значень. Для цього потрібно клацнути мишею у стовпчику **Missing** рядка відповідної змінної, у правій частині клітини з'являється три крапки "...", потрібно клацнути на цих трьох крапках і з'являється вікно **Missing Values** (*Відсутні значення*).

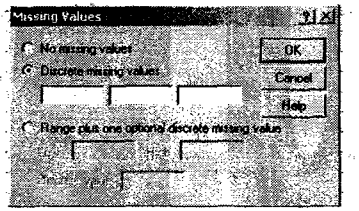


Рис. 2.3. *Вікно визначення пропущених значень* ***Missing Values.***

Є три можливості визначити інформацію про кодування відсутніх значень певної змінної. Якщо обрати **No missing values** (*Немає відсутніх значень*), то для відповідної змінної як відсутнього значення можна буде використовувати лише системне відсутнє значення (*system missing value*). Якщо обрати **Discrete missing values** (*Дискретні відсутні значення*), то з'являється можливість ввести одне, два або три значення, що будуть під час аналізу розглядатися як коди відсутніх значень відповідної змінної. Якщо ж обрати **Range plus one optional discrete missing value** (*Інтервал та одне додаткове дискретне відсутнє значення*), то з'являється можливість вказати інтервал значень (від **Low** (*Нижня границя*) до **High** (*Верхня границя*) інтервалу), що будуть розглядатися для цієї змінної як коди відсутніх значень. На додаток до інтервалу можна вказати ще одне значення (*Discrete value*), що також буде інтерпретуватися як код відсутнього значення;

- у стовпчиках **Type** (*Тип* змінної), **Width** (*Ширина*, тобто кількість знаків у представленні значень змінної), **Decimals** (*Десяткові*, тобто кількість знаків після десяткової крапки або коми у представленні значень змінної), **Columns** (*Стовпчики*, тобто ширина стовпчика для представлення значень змінної), **Align** (*Вирівнювання* інформації у стовпчику із значеннями змінної), **Measure** (*Шкала* вимірювання змінної) можна залишити ту інформацію, що з'являється за загальною угодою. Стандартно, за загальною угодою, змінні визначаються як числові (значення ***Numeric*** у полі ***Type***), представлені як "притиснуті" у стовпчику вправо (значення ***Right*** у полі ***Align***) у фіксованому форматі з восьми знаків (значення **8** у полі **Width**) та з двома цифрами після десяткової крапки (значення **2** у полі **Decimals**)[8].

[8] Як змінити параметри, виставлені за загальною угодою, ми розглядали у пункті 1.5 розділу 1.

Окремо варто розглянути стовпчик **Measure**, в якому потрібно зазначити рівень вимірювання шкали відповідної змінної. Є можливість вказати один із трьох можливих рівнів вимірювання:

- **Scale** (*Шкала*), числова неперервна змінна, виміряна на рівні шкали інтервалів або шкали відносин, до таких змінних можливо застосовувати будь-які арифметичні операції;
- **Ordinal** (*Порядкова*), змінна, виміряна на рівні шкали рангів (на рівні порядкової шкали);
- **Nominal** (*Номінальна*), змінна, виміряна на рівні шкали найменувань (на рівні номінальної шкали або шкали категорій).

Рівень вимірювання визначає перелік арифметичних операцій, що їх можна застосовувати до змінних, виміряних на відповідному рівні. Необхідно зазначити, що SPSS майже не контролює відповідність між вказаним у полі **Measure** рівнем вимірювання змінної та можливістю застосовувати до цієї змінної ті чи інші математичні операції та перетворення.

Після створення нової змінної у заголовку відповідного стовпчика електронної таблиці даних закладки **Data View** (*Перегляд даних*) вікна редактора даних **Data Editor** з'являється ім'я створеної змінної.

Для того, щоб переглянути у повному обсязі інформацію про всі змінні відкритого робочого файлу, необхідно послідовно обрати в головному горизонтальному меню **Utilities** (*Допоміжні засоби*) ⇒ **Variables...** (*Змінні*).

На рис. 2.4 показана закладка Data View редактора даних Data Editor для відкритого робочого файлу даних kiev91, з яким ми будемо працювати надалі. У зображеному на рисунку фрагменті словника виділений рядок під номером 254, який містить інформацію про змінну з іменем v174. Ця змінна містить інформацію про заробітну плату респондента (у радянських рублях), має мітку "Каков размер Вашей заработной платы?" (поле **Label**) і числовий тип (значення **Numeric** у полі **Type**), представлена у форматі з восьми знаків (значення "8" у стовпчику **Width**), з яких 2 знаки після десяткової крапки або коми (значення "2" у стовпчику **Decimals**). Стовпчик для представлення значень змінної матиме ширину 8 знаків (поле **Columns**) та значення змінної у стовпчиках вирівнюються по правому краю (значення Right у полі Align).

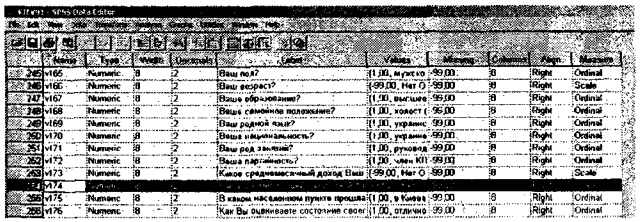


Рис. 2.4. *Фрагмент вікна зі словником файлу даних kiev91.*

Для дискретних змінних у полі **Value Labels** потрібно вказати перелік можливих значень та мітки для цих значень. Для того, щоб зробити це, потрібно клацнути у відповідній клітині мишею, відкрити вікно для введення та редагування списку значень дискретної змінної та міток цих значень. Заробітна плата розглядається як неперервна змінна, тому в цьому випадку вводити можливі значення та відповідні мітки потреби немає. Як код для відсутнього значення, описаного користувачем (*user missing value*), використано значення -99. Саме цей код є у полі **Missing**. Оскільки заробітна плата розглядається як неперервна числова ознака, то у полі **Measure** вказаний рівень вимірювання **Scale**.

2.2. Редагування таблиці даних

Часто виникає необхідність внести зміни у створену таблицю даних. Основні можливості щодо редагування даних представлені в таких опціях головного меню програми SPSS, як **Edit** (*Правка*), **Data** (*Дані*), **Transform** (*Трансформація*). У цьому пункті розглянемо лише першу та частково другу опції, можливості інших пунктів меню виділено в окремі параграфи.

Отже, опція **Edit** (див. пункт 1.3) містить стандартні можливості редагування виділеного масиву даних, а також однієї або кількох змінних (спостережень). Виділити об'єкт редагування можна так:

1. Курсор встановити в потрібну клітинку і, утримуючи клавішу *Shift*, за допомогою стрілок розтягнути виділення до потрібного розміру.
2. Курсор встановити на потрібну клітинку і, утримуючи ліву кнопку миші, розтягнути виділення до потрібного розміру.
3. Для виділення змінної натиснути правою кнопкою миші на назві змінної, для виділення спостереження – на його номері; якщо

потрібно виділити діапазон змінних (спостережень), то виділити назву першого і, утримуючи клавішу *Shift* або ліву кнопку миші, виділити в потрібному напрямі.

Зауваження. У програмі SPSS виділити діапазон несуміжних змінних або спостережень *неможливо*!

Виділений діапазон (даних, змінних або спостережень) можна:

- перемістити: **Cut** (*Вирізати*) ⇒ поставити курсор у позицію вставки і ⇒ **Past** (*Вставити*);
- скопіювати: **Copy** (*Копіювати*) ⇒ поставити курсор у позицію вставки і ⇒ **Past** (*Вставити*);
- знищити: **Clear** (*Очистити*).

Зауваження. Застосовуючи команду **Clear** до діапазону змінних або спостережень, програма *знищує!* виділені змінні (тобто повністю видаляє їх з аналізу); діапазон даних після цього стає порожнім, але самі клітини залишаються.

Ще одне не менш важливе *зауваження*.

Команда **Edit** ⇒ **Undo** (*Відмінити*) повертає лише *останню*! операцію з даними (хоча існують такі операції, відмінити які *неможливо*!).

У головному меню навпроти кожної команди вказано комбінацію клавіш, яка є стандартною, тобто такою ж, як і в MS Word, MS Excel.

Опція **Data** містить такі можливості щодо редагування даних:

- **Define Variable...** (*Створити нову змінну*) – див. пункт 2.1.
- **Define Dates...** (*Визначити формат часової змінної*) – застосовують до змінних, які містять інформацію про час (наприклад дата, години і ін.).
- **Templates...** (*Задати шаблони змінних*).
- **Insert Variable** (*Вставити змінну*).
- **Insert Case** (*Вставити спостереження*).
- **Go to Case** (*Перейти до спостереження*).

- **Sort Cases...** (*Сортувати спостереження*).
- **Transpose...** (*Транспортувати дані*).

- **Merge Files** (*Приєднати файли*).
- **Aggregate** (*Агрегація даних*).

- **Split File** (*Розщепити файли*).
- **Select Cases...** (*Вибрати спостереження*).
- **Weight Cases...** (*Зважити спостереження*).

Продовжимо розгляд виділених блоків (опції **Merge Files**, **Select Cases...** та **Weight Cases...** розглянемо в окремих параграфах).

Команда **Data ⇒ Templates...** дає змогу створити деякі шаблони для постійних однотипних змінних (використання готових шаблонів значно прискорює роботу зі створення змінних). Вікно створення шаблону змінних має такий вигляд (див. рис. 2.5).

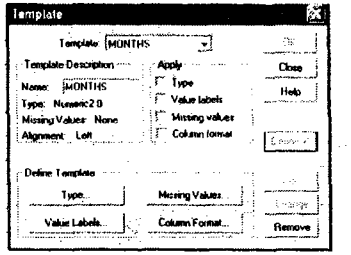


Рис. 2.5. *Вікно створення шаблону змінних.*

Блок **Define Template** містить такі ж чотири кнопки, як вікно створення нової змінної, їх потрібно заповнити. Якщо під час завантаження вікна **Template** цього блоку не виявиться, то слід натиснути кнопку **[Define>>]** (*Визначити*), що праворуч. Після внесення всіх параметрів у вікно **Name** вводять ім'я шаблону і натискають **[Add]** (*Додати*) і **[Ok]** – шаблон готовий.

Щоб використати готові шаблони, потрібно:

- або, виділивши будь-яку клітинку нової змінної, виконати команду **Data ⇒ Templates...**,
- або викликати контекстне меню на заголовку стовпчика і вибрати зі списку опцію **Templates...**.

Потім у віконечку, що з'явиться (див. рис. 2.5), вибрати зі списку створених у розділі **Template** ім'я нового шаблону, а у блоці **Apply** поставити галочки поряд з тими параметрами, які будуть властиві новій змінній, і натиснути **[Ok]** для завершення операції.

Команда **Data ⇒ Insert Variable** дає змогу вставити нову змінну *зліва*! від виділеної.

Щоб прописати нову змінну всередині вже створених, можна також виконати такі дії: виділити змінну, перед якою потрібно вставити пустий стовпчик, викликати контекстне меню (натиснути праву клавішу миші) і вибрати зі списку опцію **Insert Variable**.

Команда **Data ⇒ Insert Case** дає змогу вставити нове спостереження *вище*! від виділеного.

Аналогічно до попередньої команди виділити спостереження, перед яким потрібно вставити чистий рядок, викликати контекстне меню (натиснути праву клавішу миші) і вибрати зі списку опцію **Insert Case**.

Щойно вставлений стовпчик матиме стандартний заголовок var00001, а після доданого спостереження нумерація рядків перерозподілиться і знову утворюватиме натуральний ряд.

Команда **Data ⇒ Go to Case** дає змогу, вказавши у віконечку **Case Number** конкретний номер спостереження, до якого слід перейти, швидко рухатися по достатньо великому файлу даних.

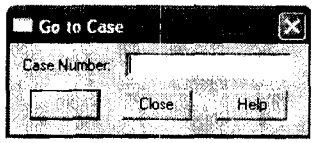


Рис. 2.6. *Вікно переходу до зазначеного номера спостереження.*

Для сортування даних використовують опцію **Data ⇒ Sort Cases...**.

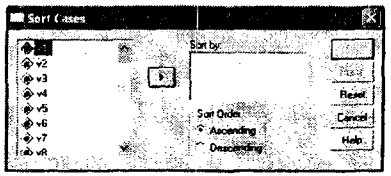


Рис. 2.7. *Вікно сортування даних.*

Зліва у вікні сортування даних міститься перелік усіх змінних, за допомогою стрілки перекидаємо в поле **Sort by:** (*Сортувати за*) одну або кілька змінних, за якими всі інші будуть відсортовані. Після цього у блоці **Sort Order** (*Порядок сортування*) визначити порядок сортування, обравши один із двох варіантів:

- **Ascending** – у порядку зростання, тобто від найменшого значення відібраної змінної до її найбільшого значення;
- **Descending** – у порядку спадання.

Натиснути кнопку **Ok** для завершення операції.

У випадку кількох змінних програма виконає послідовне сортування: спочатку дані будуть відсортовані за першою внесеною в поле **Sort by:** змінною, потім – за другою, за третьою і т. д.

У деяких випадках, наприклад, застосовуючи процедури багатовимірного статистичного аналізу (кластерного, факторного і ін.), виникає необхідність транспонувати, тобто "перевернути", матрицю даних.

Для транспонування даних використовують опцію **Data ⇒ Transpose...**. У лівому полі вікна транспонування даних (див. рис. 2.8) міститься перелік усіх змінних; за допомогою кнопки-вказівника в поле **Variable(s):** (Змінні) переміщають ті змінні, які потрібно транспонувати, причому *текстові*! змінні переміщають у поле **Name Variable:** (*Назва змінної*), за відсутності змінної типу *string* це поле може бути незаповненим. Після відбору змінних натискаємо **[Ok]**, і програма переверне матрицю, а також у вікні виведення результатів Output з'явиться звіт Log про успішність або неуспішність процедури транспонування.

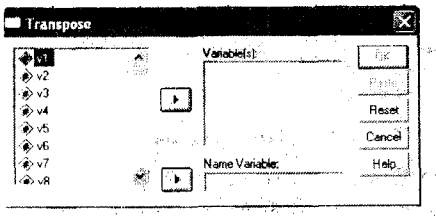
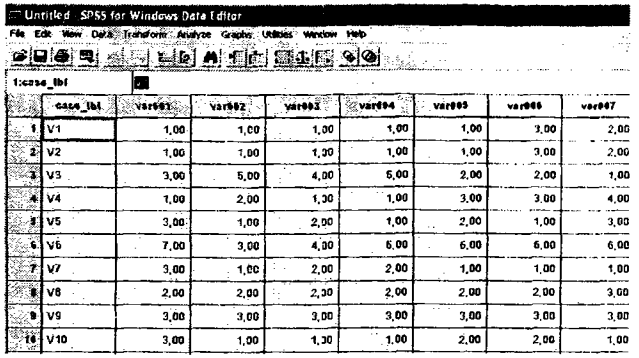


Рис. 2.8. *Вікно транспонування даних.*

Транспонована матриця – це новий файл даних (див. рис. 2.9), який необхідно назвати та зберегти (при цьому попередній залишиться без змін). Як результат транспонування перша змінна, що була перенесена у поле **Variable(s):**, стає першим спостереженням (рядком) нового файлу даних, друга змінна – другим спостереженням і т. д. У новому файлі перша змінна автоматично отримує назву *case_lbl* і містить старі назви змінних (до транспонування). Нові змінні (спостереження вихідного файлу) автоматично отримують назви *var001*, *var002* і т. д. Зрозуміло, що для зручності нові змінні можна перейменувати та відредагувати.



	case_lbl	var001	var002	var003	var004	var005	var006	var007
1	V1	1,00	1,00	1,00	1,00	1,00	3,00	2,00
2	V2	1,00	1,00	1,30	1,00	1,00	3,00	2,00
3	V3	3,00	5,00	4,00	5,00	2,00	2,00	1,00
4	V4	1,00	2,00	1,30	1,00	3,00	3,00	4,00
5	V5	3,00	1,00	2,00	1,00	2,00	1,00	3,00
6	V6	7,00	3,00	4,00	6,00	6,00	6,00	6,00
7	V7	3,00	1,00	2,00	2,00	1,00	1,00	1,00
8	V8	2,00	2,00	2,30	2,00	2,00	2,00	3,00
9	V9	3,00	3,00	3,00	3,00	3,00	3,00	3,00
10	V10	3,00	1,00	1,30	1,00	2,00	2,00	1,00

Рис. 2.9. *Новий файл транспонованих даних.*

Опція **Data ⇒ Aggregate** дає змогу стиснути файл даних за рахунок групування та об'єднання початкових значень за певними правилами.

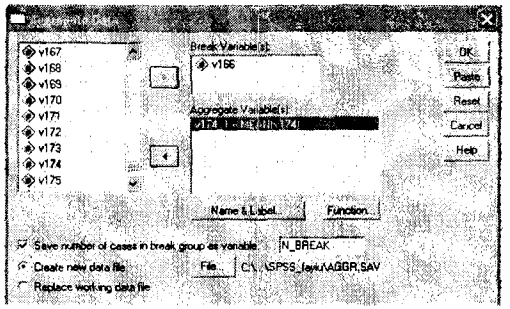


Рис. 2.10. *Вікно агрегації даних.*

Спочатку в поле **Break Variable(s):** (*Відкинуті змінні*) переносять одну або кілька змінних, які не будуть агрегованими (тобто дані в них залишаться без змін), але стануть основою для об'єднання даних в інших змінних.

Потім у поле **Aggregate Variable(s):** (*Агреговані змінні*) потрібно перемістити одну або кілька змінних, дані в яких слід об'єднати. Після цього у віконечку з'явиться вираз, наприклад (див. рис. 2.10), v174_1=MEAN(v174), де v174_1 – назва нової змінної, а MEAN – функція агрегації. За допомогою кнопки **[Name&Label...]** (*Назва та мітка*) можна вказати іншу назву змінної, а за допомогою кнопки **[Function...]** (*Функція*) вибрати іншу функцію агрегації, при цьому з'явиться вікно (див. рис. 2.11), в якому матимемо можливість вибрати:

- середнє арифметичне (*Mean of values*) – задане за загальною угодою;
- перше значення (*First value*);
- останнє значення (*Last value*);
- кількість пропущених або незважених спостережень (*Number of cases*);
- стандартне відхилення (*Standard deviation*);
- мінімальне значення (*Minimum value*);
- максимальне значення (*Maximum value*);
- суму значень (*Sum of values*).

Крім того, можна підрахувати кількість відсотків, які мають значення вище (*Percentage above*) або нижче (*Percentage below*) вказаного числа (*Value*), міститься всередині (*Percentage inside*) або зовні (*Percentage outside*) заданого числового інтервалу (*Low, High*).

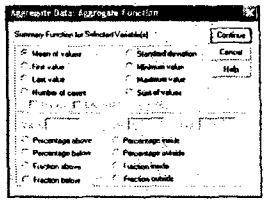


Рис. 2.11. *Функції агрегації.*

За бажанням, можна зберегти в окремій змінній кількість спостережень, які були задіяні в агрегації даних із кожної з категорій. За загальною угодою це змінна **N_BREAK**; для цього необхідно встановити галочку в опції **Save number of cases in break group as variable** (*Зберегти кількість спостережень у категоріях відкинутої змінної*).

Результатом процедури агрегації даних, як і процедури транспонування, буде новий файл даних – за загальною угодою має назву **AGGR.SAV**. Змінити назву файлу та вказати каталог, в який слід зберегти новий файл, можна за допомогою кнопки **[File...]**.

Якщо ви бажаєте відразу перейти до роботи у новому файлі, виділіть опцію **Replace working data file** (*Замінити відкритий файл даних*), при цьому попередній файл даних збережеться. Якщо ж ви хочете продовжити роботу з основним файлом, то нічого змінювати не потрібно: за загальною угодою програма вже виділила параметр **Create new data file** (*Створити новий файл даних*).

Кнопка **[Ok]** запускає процедуру агрегації даних на виконання.

Опція **Data ⇒ Split File** дає змогу поділити спостереження на групи (під групою розуміють визначену кількість спостережень з однаковими значеннями певної ознаки) для одночасного виконання обробки та аналізу даних у різних групах. Змінна, за якою всі спостереження ділять на групи, називається *групуючою*. Наприклад, такою змінною може бути змінна "Стать", тоді всі змінні зі значенням ознаки 1 (чоловіча) утворюватимуть одну групу, а зі значенням 2 (жіноча) – іншу; при цьому з кожною групою можна буде виконувати якісь операції.

За загальною угодою в аналізі даних беруть участь усі спостереження (*Analyze all cases, do not create groups*); але, обираючи пункт **Organize output by groups** (*Поділити виведення на групи*), ми отримаємо виведення результатів із кожної групи окремо (в різних таблицях), а виділяючи пункт **Compare groups** (*Порівняти групи*), отримаємо виведення результатів так, щоб візуально їх можна було порівнювати (в одній таблиці). В обох випадках у поле **Groups based on** (*Групи, створені на основі*) потрібно ввести групуючу змінну, і для подальшого виконання обробки даних файл повинен бути відсортований за цією, тому варто залишити виділеною опцію **Sort the file by grouping variables** (*Сортувати за групуючими змінними*), встановлену за загальною угодою.

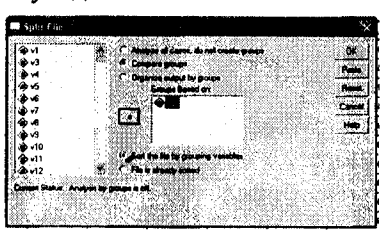


Рис. 2.12. *Вікно операції розщеплення файлу.*

Після виконання усіх необхідних операцій із різними групами розщеплення файлу скасовують, для цього в діалоговому вікні **Split File** знову оберіть опцію **Analyze all cases, do not create groups** (*Аналізувати всі спостереження, не створювати групи*) і натисніть **[Ок]**.

2.3. Імпорт та експорт даних у середовищі SPSS

Соціологічні дослідження є, як правило, доволі масштабними: великий обсяг вибірки, або велика кількість змінних, або і одне, і друге. Тому дані в програму вносить не одна людина, відповідно, на

різних комп'ютерах, а отже, в різних файлах. Хоча дійсних причин для імпорту даних є багато. Наприклад, доволі часто дослідникам потрібно аналізувати дані, що були внесені в іншу програму, можливо, менш потужну. Як відомо, програма Excel також дає змогу обробляти дані соціологічних досліджень та проводити нескладний аналіз, але можливості має доволі обмежені, тому, експортувавши такі дані з програми Excel у статистичний пакет SPSS, соціолог може скористатися більш потужними статистико-математичними процедурами аналізу даних.

Щоб приєднати до одного файлу формату *.sav інший файл даних формату *.sav, потрібно виконати одну з команд:

1) **Data ⇒ Merge Files ⇒ Add Cases...** – для приєднання спостережень або

2) **Data ⇒ Merge Files ⇒Add Variables...** – для приєднання змінних.

Операцію **Add Cases...** використовують для об'єднання двох файлів з однаковими змінними, але різною кількістю спостережень. Наприклад, анкети одного дослідження для скорішого введення даних у SPSS були розподілені між 6-ма особами по 150, 160, 180, 190, 200 та 220 відповідно; як результат, матимемо шість файлів з однаковими змінними, але різними за кількістю спостережень.

Виконавши порядок команд 1), потрібно вказати файл, спостереження з якого будуть приєднані. В результаті з'явиться вікно приєднання спостережень (рис. 2.13).

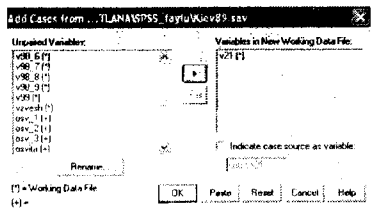


Рис. 2.13. *Вікно операції приєднання спостережень.*

Зліва у вікні **Unpaired Variables:** (*Непарні змінні*) міститься перелік усіх змінних з обох файлів, назви яких не збігаються, причому зірочкою (*) позначені назви змінних з основного файлу, а значком (+) – з додаткового.

У віконечко справа **Variables in New Working Data File** (*Змінні в новому робочому файлі даних*) переносять змінні, які ввійдуть до нового файлу даних.

Непарні змінні з'являються, наприклад, тоді, коли одні й ті ж змінні прописують різні особи, як результат, різниця в один символ у назві дає дві різні назви однієї змінної. Цього можна уникнути, якщо:

1. Надати новій змінній ім'я, що складатиметься з об'єднаних назв змінних: H_o&$H_д$, де H_o – назва змінної з основного файлу, $H_д$ – з додаткового. Для цього треба, утримуючи клавішу *Control*, клацнути лівою кнопкою миші на змінних, які потрібно об'єднати. Недоліком такого варіанту є надто громіздка назва нової змінної.

2. Надати обом змінним однакові назви. Для цього потрібно скористатися кнопкою **[Rename...]** (*Перейменувати*), причому перейменовують змінну з додаткового файлу, надавши їй таке ім'я, як у основної. Після цього повторюють операцію об'єднання згідно з пунктом 1.

Для завершення процедури об'єднання натискаємо кнопку **[Ok]**.

Операцію **Add Variables...** використовують для об'єднання двох файлів з однаковими спостереженнями, але різними змінними. Приклад: опитування однієї й тієї ж групи респондентів з певним часовим інтервалом.

Процедура об'єднання змінних аналогічна до попередньої процедури додавання спостережень. Для коректного виконання цієї операції необхідно пам'ятати, що:

- по-перше, кількість спостережень повинна бути однакова,
- по-друге, всі спостереження повинні бути відсортовані в однаковому порядку.

Згідно з попереднім прикладом дані можна об'єднати лише, якщо:

- кількість респондентів в обох опитуваннях однакова;
- це одні й ті ж респонденти;
- порядок введення однаковий (наприклад за алфавітом або згідно з номером анкети).

Крім того, імпортувати у програму SPSS можна дані іншого формату. Продемонструємо це на прикладі приєднання файлу, створеного в MS Excel.

Щоб приєднати до файлу даних формату *.sav файл іншого формату, наприклад формату *.xls, потрібно виконати команду **File ⇒ Open Data ...**, у вікні, що з'явиться, вибрати формат файлу (наприклад *.xls), потім сам файл, що імпортується, та натиснути **[Ok]**. Після цього програма відкриє вікно імпорту файлу **Opening File Options** (див. рис. 2.14).

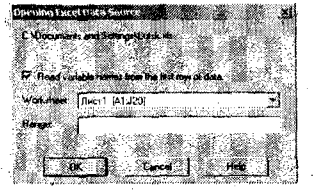


Рис. 2.14. *Вікно імпорту файлу даних* ***Opening File Options.***

У вікні **Range** (*Імпортувати діапазон*) вказують координати лівої верхньої та правої нижньої точок імпортованого прямокутного фрагменту даних (наприклад A7:C235), якщо імпортується частина таблиці даних. У випадку імпорту всієї таблиці ця графа не заповнюється.

Якщо поставити позначку в опції **Read variable name** (*Читати назви змінних*), то програма SPSS перенесе верхній текстовий рядок таблиці як назву змінних, при цьому текстові надписи будуть скорочені до 8 символів. Інакше назви змінних будуть стандартними var00001, var00002, ... У будь-якому випадку доведеться прописувати всі змінні (тобто задавати тип, мітки, відсутні значення й ін.).

Після внесення змін (або не внесення) натискаємо **[Ok]**, як результат, програма SPSS імпортує дані та представить у вікні виведення результатів Output звіт (*Log*) про успішність або неуспішність процедури перенесення даних.

Зауваження. Доповнити робочий файл SPSS даними із зовнішнього файлу можна, крім того, скориставшись буфером обміну. Є такі варіанти:

1. Із зовнішнього файлу Excel,
2. Із зовнішнього файлу SPSS.

У першому випадку потрібно: відкрити обидва файли; у файлі Excel стандартним способом виділити потрібний блок даних, скопіювати його в буфер обміну; вікно SPSS розгорнути на вкладці Data View, встановити курсор у ліву верхню клітину і вставити дані з буфера обміну.

Особливість другого випадку полягає в тому, що і робочий, і зовнішній файли SPSS не можуть бути відкриті одночасно, тому спочатку треба відкрити зовнішній файл; у ньому стандартним способом виділити потрібний блок даних, скопіювати його в буфер обміну; закрити зовнішній файл; відкрити робочий файл; встановити курсор у ліву верхню клітину і вставити дані з буфера обміну.

Контрольні запитання

1. Що таке словник файлу даних?
2. Чи можливим є варіант застосування одного словника до різних файлів даних?
3. У чому полягає специфіка введення даних у форматі *.sav?
4. На якому етапі комп'ютерної обробки інформації можна редагувати дані? Переваги та недоліки цього процесу в середовищі SPSS.
5. Дані одного дослідження вводилися кількома операторами на різних комп'ютерах. Як можна введену таким способом інформацію об'єднати в один масив?
6. Які ще операції можна проводити з файлами даних типу *.sav?
7. Чи існують обмеження при імпорті/експорті даних у середовищі SPSS?

Практичні завдання

1. Прописати змінні за анкетою. Внести 20 спостережень. Файл даних зберегти під назвою, що є власним прізвищем, англомовними літерами.
2. Створити шаблони для змінних стать, вік, освіта.

РОЗДІЛ 3. АНАЛІЗ РОЗПОДІЛУ ОДНІЄЇ ЗМІННОЇ

3.1. Побудова та аналіз одновимірної таблиці частот і відсотків

Структуру сукупності об'єктів із точки зору певної однієї дискретної ознаки досить добре вивчати шляхом аналізу таблиці, в якій для кожного з можливих значень ознаки записано, скільки разів зустрічаються в сукупності об'єкти, що мають відповідне значення. Таку таблицю називають таблицею одновимірного розподілу, або ж одновимірною таблицею частот та відсотків. Таблиця має стільки рядків, скільки є категорій у дискретній ознаці.

Для побудови таблиці частот та відсотків у пакеті SPSS обираємо послідовно

Analyze ⇒ Descriptive statistics ⇒ Frequencies...

(*Аналіз ⇒ Описові статистики ⇒ Частоти*)

Розкривається вікно параметрів

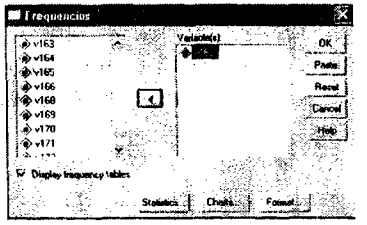


Рис. 3.1. *Вікно параметрів операції* ***Frequencies***

На прикладі вікна параметрів операції Frequencies розглянемо основні структурні компоненти вікон різних статистичних процедур SPSS.

- У вікні зліва є перелік усіх змінних файлу даних. Залежно від встановлених значень певних параметрів роботи пакету (**Edit ⇒ Options**, закладка **General**, група **Variable Lists**) у вікно виводяться або імена змінних (**Display names**), або ж мітки змінних (**Display labels**). Крім того, змінні у переліку у вікні можуть бути впорядковані за іменами (**Alphabetical**) або ж у такому порядку, як вони розташовані у словнику файлу даних (**File**). Перелік змінних – це список зліва всіх змінних у файлі даних. Перед іменем кожної змінної стоїть піктограма, що дає змогу легко визначити тип змінної (числова чи рядкова). Всі

змінні у вікні на рис. 3.1 є числовими (**Numeric**). Користуючись стандартними засобами MS Windows (ліва кнопка миші, можливо із одночасним натисканням та утримуванням клавіші *Shift* або ж клавіші *Ctrl*), у лівому вікні можна виділити одну змінну або групу змінних, до яких має бути застосована відповідна операція. У випадку операції **Frequencies** йдеться про ті змінні, для яких необхідно побудувати таблицю частот та відсотків, обчислити значення різних характеристик розподілу. Імена цих змінних мають бути перенесені у праве вікно, що має заголовок **Variable(s)**. Для цього потрібно натиснути кнопку, позначену спрямованим у правий бік трикутником, що міститься між цими двома вікнами.

- У розташованому у правій частині вікні **Variable(s)** формується перелік тих змінних, до яких має бути застосована операція. Якщо є потреба прибрати одну або декілька змінних із цього списку, то необхідно виділити відповідну змінну (або змінні) та натиснути позначену спрямованим у лівий бік трикутником кнопку, яка є між двома вікнами.
- Уздовж правої границі вікна вертикально розташовані п'ять стандартних кнопок. Такі кнопки є у вікні параметрів кожної з операцій SPSS.
  - Кнопка **[Ok]** – закриває вікно параметрів операції та розпочинає виконання операції з заданими параметрами.
  - Кнопка **[Paste]** – розкриває вікно для запису послідовності операцій мовою SPSS (вікно Syntax) та виводить у це вікно текст команди, що відповідає обраній операції з установленими параметрами. Це дає можливість писати програми мовою SPSS, навіть не знаючи деталей синтаксису цієї мови.
  - Кнопка **[Reset]** – встановлює для всіх параметрів операції значення за загальною угодою, (тобто повертає вікно параметрів операції до "початкового стану").
  - Кнопка **[Cancel]** – відміняє всі зміни, зроблені у вікні параметрів із моменту останнього його відкриття, та закриває вікно параметрів.
  - Кнопка **[Help]** – відкриває вікно з інформацією, що стосується змісту та параметрів відповідної операції SPSS.
- У нижній частині вікна горизонтально розташовані кнопки, що відкривають вікна параметрів операції **Frequencies...**, а саме:

  - Кнопка **[Statistics...]** – дає змогу визначити, які характеристики розподілу та показники необхідно обчислити і вивести у вікно результатів.
  - Кнопка **[Charts...]** – дає змогу визначити, які графіки необхідно побудувати та вивести у вікно результатів.
  - Кнопка **[Format...]** – визначає структуру представлення (формат) таблиць та показників, що виводяться у вікно результатів.

Кнопки допоміжних діалогових вікон мають три крапки після назви.

Параметри операції **Frequencies...**, що задаються з використанням кнопок **[Statistics...]** та **[Charts...]**, вивчимо пізніше. Зараз розглянемо можливості керування зовнішнім виглядом представлення результатів обчислень операції **Frequencies...**. Натискання на кнопку **[Format...]** розкриває відповідне вікно, зображене на рис. 3.2.

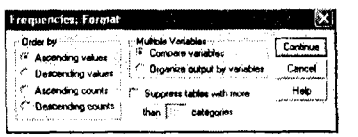


Рис. 3.2. *Вікно параметрів представлення результатів операції Frequencies.*

- У групі **Order by** можна обрати один із чотирьох способів упорядкування рядків одновимірної таблиці частот та відсотків:
  - **Ascending values** – рядки таблиці впорядковуються за зростанням (збільшенням, від меншого до більшого) значень кодів категорій, що відповідають рядкам; саме так упорядковуються рядки таблиці за загальною угодою;
  - **Descending values** – рядки таблиці впорядковуються за спаданням (зменшенням, від більшого до меншого) значень кодів категорій, що відповідають рядкам;
  - **Ascending counts** – рядки таблиці впорядковуються за зростанням частот категорій, що відповідають рядкам;
  - **Descending counts** – рядки таблиці впорядковуються за спаданням частот категорій, що відповідають рядкам.
- Група **Multiple Variables** визначає, як потрібно представляти характеристики розподілів та інформацію про таблиці у випадку, коли операція **Frequencies...** застосовується для декількох змінних:

- **Compare variables** – інформація про таблиці та характеристики розподілів для всіх змінних виводиться в одну таблицю; такий формат може бути зручним для порівняння характеристик декількох змінних;
- **Organize Output by Variables** – інформація про таблицю та характеристики розподілу для кожної змінної виводиться окремо.

- Позначка поруч із **Suppress tables with more than __ categories** дає змогу уникнути виведення довгих (із кількістю категорій, що перевищує задане значення) таблиць.

Отже, для того, щоб побудувати таблицю частот та відсотків для однієї або ж декількох дискретних змінних, необхідно виконати такі дії:

- сформувати у вікні **Variable(s)** перелік змінних, для яких потрібно побудувати одновимірні таблиці;
- перевірити, щоб стояла позначка біля **Display frequency tables** (*Виводити таблиці частот*);
- за необхідності зміни формату представлення таблиць натиснути кнопку **[Format...]**, встановити необхідні параметри представлення таблиць, натиснути кнопку **[Continue]**;
- натиснути кнопку **[OK]**.

Результат обчислень операції **Frequencies...** виводиться у вікно результатів **Output**. Розглянемо структуру результату на прикладі однієї змінної (див. рис. 3.3 – результат обчислень для змінної v167, що має мітку "Ваша освіта", з файлу даних kiev91). Для однієї змінної результат роботи операції складається з двох таблиць.

Перша таблиця має заголовок **Statistics** і містить інформацію про кількість об'єктів (**N**), представлену у вигляді двох складників – кількість об'єктів, для яких значення змінної є відомим (**Valid**), та кількість об'єктів із відсутнім значенням змінної (**Missing**).

Друга таблиця, власне, і є потрібною нам таблицею частот та відсотків. Над таблицею виводиться мітка відповідної змінної. Перший (лівий) стовпчик таблиці містить або мітки значень (якщо для значень вказані мітки у словнику файлу даних), або ж безпосередньо самі значення змінної. Другий стовпчик таблиці має назву **Frequency** і містить частоти категорій змінної. Третій стовпчик таблиці має назву **Percent** і містить для кожної категорії змінної відсоток, обчислений відносно загальної кількості об'єктів у файлі даних. Четвертий стовпчик таблиці (із назвою **Valid Percent**) також містить для кожної

категорії відсоток, але в цьому стовпчику є відсотки, обчислені відносно кількості об'єктів, що мають дійсне (відмінне від відсутнього) значення (так звані "дійсні" відсотки). П'ятий, останній, стовпчик таблиці містить накопичений (кумулятивний) відсоток – до значення "дійсного" відсотка кожної категорії додають значення "дійсних" відсотків усіх тих категорій, що розташовані в таблиці вище.

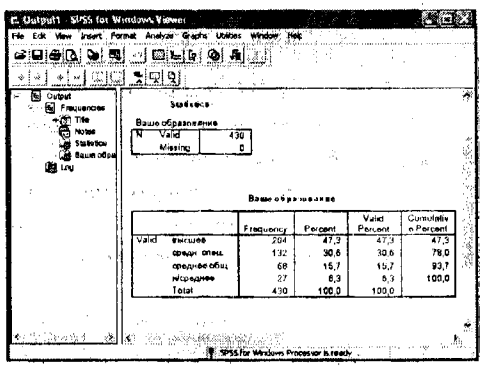


Рис. 3.3. *Одновимірна таблиця для змінної v167.*

Рядок таблиці із міткою **Missing** містить частоти та відсотки для відсутніх значень (тобто відсутні значення розглядаються як певні категорії змінної). Рядок таблиці із міткою **Total** містить "контрольні суми" усіх значень у відповідних стовпчиках.

Для того, щоб перенести таблицю до тексту звіту, необхідно виконати таку послідовність дій:

- встановити курсор миші на потрібній таблиці;
- клацнути правою кнопкою миші та у контекстному меню, що з'являється, обрати операцію **Copy**. Ця операція копіює таблицю у буфер обміну MS Windows;
- відкрити текст звіту. Як правило, його готують за допомогою текстового редактора MS Word;
- встановити курсор миші у потрібному місці у тексті, натиснути праву кнопку миші та у контекстному меню, що з'являється, обрати операцію **Paste**. У тексті має з'явитися таблиця;
- відформатувати таблицю, що з'явилася, засобами MS Word.

Необхідно зазначити, що одновимірні таблиці, як правило, потребують форматування у тексті звіту або іншої публікації. Кожна таблиця повинна мати заголовок та власний номер. Це необхідно для

того, щоб на таблицю була можливість посилатися із різних частин тексту. Якщо таблиця ілюструє результати опитування, то необхідно вказати у заголовку таблиці повний текст питання, яке ставилося респонденту, та у самій таблиці вказати повні тексти варіантів відповіді на це питання, що пропонувалися респонденту для обрання. У публікаціях, як правило, використовують таблиці, що містять або частоти та один тип відсотків, або ж навіть просто один стовпчик відсотків. Які саме відсотки презентувати (відсотки до загальної кількості об'єктів або "дійсні" відсотки) – вирішує дослідник. Якщо в таблиці презентуються тільки відсотки, то обов'язково чітко має бути вказана загальна кількість об'єктів, для яких побудована таблиця. Якщо таблиця презентує розподіл для певної частини об'єктів (наприклад лише для респондентів Західного регіону або лише для жінок із вищою освітою), то необхідно чітко вказувати умову відбору та обсяг (кількість об'єктів) цієї частини.

Одновимірні таблиці є дуже поширеною формою презентації результатів емпіричних соціологічних досліджень, вони є важливими та необхідними насамперед для опису емпіричних даних, часто ілюструють висновки дослідників. У текстах необхідно уникати простого "переказування" таблиць, концентруватися на особливостях розподілу (його несиметричності або ж, навпаки, симетричності тощо), на головних тенденціях розподілу, на неоднорідності розподілу тощо.

3.2. Обчислення мір центральної тенденції та варіації

Міри центральної тенденції – це узагальнюючі характеристики розподілу деякої ознаки у сукупності об'єктів. Щоб такий показник дійсно відображав тенденцію (закономірність), необхідно, щоб він застосовувався до досить однорідної сукупності. Тому для аналізу важливими є також міри варіації, котрі характеризують ступінь того, наскільки сильно змінюються (варіюють) значення ознаки у сукупності об'єктів.

Для кількісних змінних (метрична шкала) як міра центральної тенденції часто використовують середнє арифметичне (яке називають просто "середнє"). Середнє обчислюється за формулою $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, де N – кількість об'єктів у сукупності, x_i – значення кількісної змінної X у об'єкта з номером i.

Для порядкових змінних як міри центральної тенденції використовують медіану – центральний елемент упорядкованого ряду значень у об'єктів сукупності. Головна властивість медіани полягає в тому, що половина об'єктів сукупності мають значення змінної, що не перевищує медіану, а половина – значення, що є не меншими, ніж медіана. Медіана є більш стійкою, порівняно із середнім, стосовно наявності у сукупності невеликої кількості нетипових дуже великих або ж дуже малих значень. Тому для дуже неоднорідних кількісних змінних як міри центральної тенденції також використовують медіану. Якщо розподіл змінної є симетричним, то медіана і середнє збігаються.

Для номінальних змінних як міри центральної тенденції часто використовують моду – значення змінної з найбільшою частотою. Використовувати моду доцільно для дискретних змінних із невеликою кількістю можливих значень (невеликою кількістю категорій).

Для кількісних змінних як міри варіації використовують дисперсію. Оцінка дисперсії обчислюється за формулою $D = \sum_{i=1}^{N} \frac{(x_i - \overline{X})^2}{N-1}$, де x_i – значення i-тої ознаки, $\overline{X}$ – середня арифметична N значень змінної. Певну складність у використанні дисперсії становить те, що розмірність дисперсії є квадратом розмірності змінної (тобто, якщо змінна виміряна в кілограмах, то дисперсія цієї змінної вимірюється у "кілограмах в квадраті"). Тому часто як міру варіації розглядають стандартне відхилення, що обчислюється як корінь квадратний із дисперсії $S = \sqrt{D}$ і вимірюється в тих самих одиницях вимірювання, що й сама змінна. Якщо дисперсія або стандартне відхилення дорівнює нулю, то всі об'єкти мають однакове значення змінної, а отже, ця змінна не має варіації. Чим більшим є значення дисперсії або ж стандартного відхилення, тим більш неоднорідною за даною змінною є сукупність. Для порівняння варіації у декількох різних сукупностях використовують коефіцієнт варіації, який є безрозмірною величиною і обчислюється як відношення стандартного відхилення до середнього $C = S/\overline{X}$. Вважають, що сукупність є достатньо однорідною за деякою змінною X, якщо відповідний коефіцієнт варіації не перевищує $0{,}3 \div 0{,}4$.

Як характеристику варіації порядкової змінної можна розглядати квантильні розподіли. Квантилі – це показники, що ділять упорядкований ряд значень змінної у сукупності на n рівних частин. Якщо

$n=4$, то відповідні показники називаються квартилями. Всього є три квартилі Q_1, Q_2, Q_3 (див. рис. 3.4).

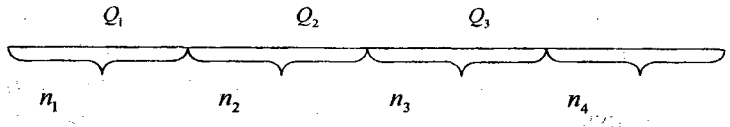


Рис. 3.4. *Графічне представлення квартильного розподілу.*

Різницю між третім та першим квартилями $\Delta Q = Q_3 - Q_1$ також розглядають як міру варіації. Зауважимо:

- 25 % об'єктів мають значення змінної, що не перевищують Q_1;
- 25 % об'єктів мають значення змінної, що не менше ніж Q_3;
- 50 % об'єктів мають значення змінної між Q_1 та Q_3;
- квартиль Q_2 – це медіана.

Якщо $n=10$, то відповідні показники називаються децилями. Усього є дев'ять децилів, і в такому випадку ведуть мову про дециль-ний розподіл. Якщо ж $n=100$, то говорять про процентилі. Усього процентилів 99, але часто обчислюють 5-й та 95-й процентилі, на числовому інтервалі між якими є 90 % усіх значень змінної в об'єктів сукупності.

Різні процедури SPSS обчислюють міри центральної тенденції та міри варіації для вказаних змінних. Ми розглянемо обчислення цих показників за допомогою операції **Frequencies...** . Для виклику операції послідовно обираємо в горизонтальному меню **Analyze** ⇒ **Descriptive statistics** ⇒ **Frequencies...** . Потім у вікні **Variable(s):** формуємо перелік тих змінних, для яких нам потрібно обчислити міри центральної тенденції та міри варіації. Натискаємо кнопку **[Statistics...]**, і відкривається вікно параметрів (див. рис. 3.5).

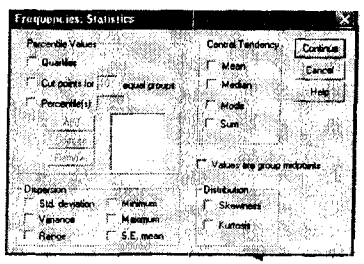


Рис. 3.5. *Вікно статистик.*

Саме в цьому вікні потрібно вказати те, які характеристики одновимірного розподілу потрібно обчислити та вивести у вікно результатів. Під час вибору характеристик дослідник обов'язково має враховувати тип шкали відповідної змінної.

У групі **Central Tendency** перераховані міри центральної тенденції, які може обчислювати SPSS. Зокрема, пакет дає змогу обчислювати середнє арифметичне (**Mean**), медіану (**Median**), моду (**Mode**) та суму всіх значень (**Sum**).

У групі **Percentile Values** можна вказати на необхідність обчислити та вивести у вікно результатів:

- квартилі (**Quartiles**);
- квантилі (**Cut point for __ equal groups**) – необхідно вказати, на скільки однакових груп потрібно розділити впорядкований ряд значень змінної; стандартно, за загальною угодою кількість груп дорівнює 10;
- процентилі (**Percentile(s)**) – користуючись кнопками **[Add]**, **[Change]** та **[Remove]**, у відповідному вікні потрібно сформувати перелік саме тих процентилів, які необхідно вивести.

У групі **Dispersion** перераховані показники варіації, які може обчислювати SPSS. Зокрема, пакет дає змогу обчислювати стандартне відхилення (**Std. deviation**), дисперсію (**Variance**), варіаційний розмах (**Range**), різницю між найбільшим та найменшим значеннями змінної, найменше (**Minimum**) та найбільше (**Maximum**) значення змінної.

У групі **Distribution** вказані ще дві важливі характеристики розподілу:

- Асиметрія (**Skewness**) – міра відхилення емпіричного розподілу від розподілу симетричного. Якщо асиметрія дорівнює 0 (або близька до 0), то емпіричний розподіл є симетричним (близьким до симетричного). Чим більше показник асиметрії відрізняється від 0, тим більш асиметричним є емпіричний розподіл. Говорять про позитивну (праву) та негативну (ліву) асиметрію.
- Ексцес (**Kurtosis**) – характеристика гостроти вершини емпіричного унімодального розподілу порівняно із розподілом нормальним. Якщо ексцес дорівнює 0, то гострота вершини емпіричного розподілу є такою ж, як і у нормального розподілу. Якщо ексцес є від'ємним, то говорять про "плосковершинний" розподіл (вершина емпіричного розподілу має більш плоску форму, ніж у відповідного

нормального). Якщо ж ексцес є додатним, то говорять про "гостровершинний" розподіл (вершина емпіричного розподілу є більш гострою порівняно із вершиною відповідного нормального розподілу).

Приклад. Розглянемо обчислення та інтерпретацію основних числових характеристик змінної v174, що містить інформацію про заробітну плату респондента (параметри цієї змінної у словнику файлу даних kiev91 ми вже розглядали у пункті 2.1).

Для того, щоб обчислити характеристики одновимірного розподілу неперервної числової змінної v174, послідовно виконуємо такі дії:

- обираємо в головному горизонтальному меню **Analyze** ⇒ **Descriptive statistics** ⇒ **Frequencies...**;
- у вікно **Variable(s):** заносимо змінну v174 (див рис. 3.6);

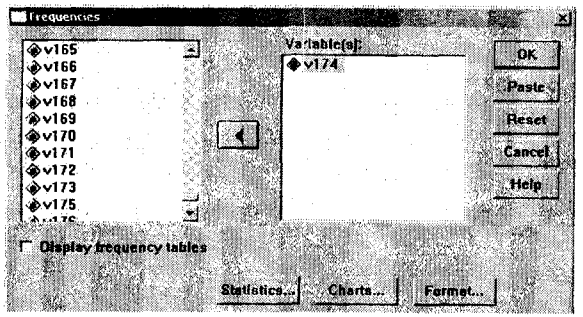


Рис. 3.6. *Вікно параметрів операції Frequencies для змінної v174.*

- натискаємо кнопку **[Statistics...]** (див рис. 3.6), відкривається вікно замовлення статистик (див. рис. 3.7), в якому ми замовляємо обчислення:
  - децилів – у групі **Percentile Values** (*Процентильні значення*) ставимо позначку біля **Cup points for ___ equal groups** (*Точки поділу для ___ однакових груп*) та залишаємо кількість 10 (точки поділу сукупності на 10 рівних частин задають децильні значення);
  - процентилів – у групі **Percentile Values** ставимо позначку біля **Percentile(s):** (*Процентилі*) та, використовуючи кнопку **[Add]**, замовляємо 5-й, 50-й та 95-й процентилі;
  - мір центральної тенденції – у групі **Central Tendency** (*Центральна тенденція*) ставимо позначки біля **Mean** (*Середнє*), **Median** (*Медіана*) та **Mode** (*Мода*);

  - мір варіації – у групі **Dispersion** (*Варіація*) ставимо позначки біля **Std. Deviation** (*Стандартне відхилення*), **Variance** (*Дисперсія*), **Minimum** (*Мінімум*) та **Maximum** (*Максимум*);
  - характеристик для порівняння емпіричного розподілу з розподілом нормальним – у групі **Distribution** (*Розподіл*) ставимо позначку біля **Skewness** (*Асиметрія* або *Ухил*) та **Kurtosis** (*Ексцес*);
- натискаємо кнопку **[Continue]**;
- відмовляємося від побудови двовимірних таблиць частот та відсотків – знімаємо встановлену за загальною угодою позначку біля **Display frequency tables** (*Виводити таблиці частот*);
- натискаємо кнопку **[OK]**.

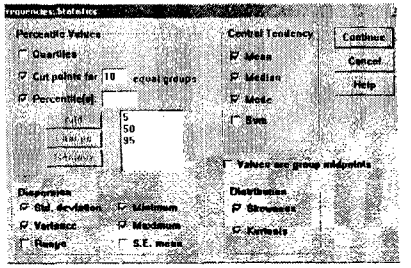


Рис. 3.7. *Вікно замовлених статистик для змінної v174.*

Прокоментуємо результат обчислень, що виводиться у вікно Output (див. рис. 3.7).

Відповідь на питання про розмір заробітної плати (стипендії, пенсії) дали 409 респондентів, 22 респонденти не дали відповідь, тому файл даних містить 22 відсутніх значення. Середня зарплата респондентів дослідження становить 305,36 руб. Значення 300 руб. зустрічається у відповідях респондентів частіше, ніж інші значення, тому є модальним. Сукупність опитаних респондентів є досить неоднорідною з точки зору заробітної плати (мінімальне значення дорівнює 0, максимальне значення дорівнює 3300 руб., значення стандартного відхилення дорівнює 225,19 руб. і становить приблизно 74 % від середнього, адже $^{225,17}/_{305,36} \approx 0,74$). За умов такої неоднорідності часто як міру центральної тенденції розподілу використовують медіану, яка менше, ніж середнє арифметичне залежить від екстремальних значень на кінцях упорядкованого ряду значень ознаки. У нашому розподілі

медіана дорівнює 260,00. Це означає, що половина опитаних отримує зарплату, що не перевищує 260 руб., а друга половина – зарплату, що є не меншою, ніж 260 руб.

Statistics

Каков размер Вашей заработной платы (степендии, пенсии)? (руб)

N	Valid	409
	Missing	22
Mean		305,3619
Median		260,0000
Mode		300,00
Std. Deviation		225,1859
Variance		50708,70
Skewness		6,573
Std. Error of Skewness		,121
Kurtosis		78,083
Std. Error of Kurtosis		,241
Minimum		,00
Maximum		3300,00
Percentiles	5	100,0000
	10	135,0000
	20	167,0000
	30	200,0000
	40	230,0000
	50	260,0000
	60	300,0000
	70	350,0000
	80	400,0000
	90	500,0000
	95	645,0000

Рис. 3.8. *Таблиця замовлених статистик для змінної v174.*

Значення медіани менше від значення середньої арифметичної, а отже, розподіл значень змінної має праву асиметрію (ухил вправо). Це підтверджує також позитивне значення показника асиметрії +6,5). Позитивне значення показника ексцесу +78,1 свідчить про гостровершинність розподілу (велика кількість респондентів, зарплатня яких близька до середньої).

Тепер проаналізуємо варіацію ознаки. Зазначимо, що оскільки ми не виділяли групи респондентів за зайнятістю (хто працює і хто не працює), за освітою, за професією, то важко очікувати, що розподіл заробітної плати буде однорідним або ж близьким до такого. Розподіл характеризується досить великим значенням дисперсії 50708,70 та, відповідно, великим значенням стандартного відхилення 225,1859.

Ми вже зазначили вище, що коефіцієнт варіації ознаки (відношення стандартного відхилення до середнього) має досить велике значення 0,74, що свідчить про доволі велику неоднорідність розподілу ознаки. З аналізу значень замовлених нами процентилів видно, що 5 % опитаних мають зарплату менше 100 руб., а інші 5 % мають зарплату більше 645 руб. Основна частина, 90 % респондентів, мають зарплату в межах від 100 до 645 руб. Зазначимо, що, значення 50-го процентиля та медіани збігаються не випадково, адже це дві назви для одного й того ж показника. Отже, загальний висновок стосовно варіації полягає в тому, що у всієї сукупності опитаних заробітна плата є досить неоднорідною. Для подальшого аналізу доцільно виділити окремі групи (за зайнятістю, освітою, професією, віком тощо), більш однорідних за заробітною платою.

Контрольні запитання

1. З якою метою будують одновимірну таблицю частот та відсотків?
2. Чим відсотки, обчислені до загальної кількості об'єктів (Percent), відрізняються від відсотків, обчислених до кількості дійсних значень змінної (Valid Percent)?
3. Що можна сказати про суму всіх частот (стовпчик Frequency), суму всіх відсотків (стовпчик Percent) та суму всіх "дійсних" відсотків (стовпчик Valid Percent) у таблиці?
4. Як для певної окремої категорії змінної співвідносяться між собою відсоток (стовпчик Percent) та "дійсний" відсоток (стовпчик Valid Percent)?
5. Які міри центральної тенденції обчислює операція **Frequencies...**?
6. Які міри варіації обчислює операція **Frequencies...**?
7. Які міри центральної тенденції можна обчислювати для кількісних змінних (метрична шкала)?
8. Як співвідносяться між собою процентилі, квартилі, децилі та медіана?

Практичні завдання

1. Побудувати одновимірні таблиці для всіх змінних файлу даних, створеного на попередньому занятті.
2. Файл kiev91.sav містить результати дослідження Інституту соціології НАН України, проведеного в м. Києві у червні 1991 р. (опитування проводилося російською мовою). У питаннях із номерами від 40 до 50 (відповідні змінні у файлі даних – v40–v50) киянам

було запропоновано оцінити прийнятність різних способів виходу з економічної кризи. Серед цих способів є такі, що орієнтовані на ринкові реформи, і такі, що не орієнтовані на ринкові реформи. На основі відповідей на питання 40–50 (на основі аналізу відповідних одновимірних таблиць) зробіть висновок про ставлення киян до ринкових реформ у червні 1991 року.

3. Для змінної v3 (файл даних kiev91) побудувати таблицю частот та відсотків, моду, середнє, стандартне відхилення.

Звернути увагу на те, що альтернатива "важко сказати" із кодом 6 не дає можливості інтерпретувати шкалу змінної v3 як порядкову або ж як "квазіметричну". Визначити для змінної v3 значення "6" як відсутнє. Знову обчислити середнє та стандартне відхилення для v3. Порівняти результат із попереднім. Звернути увагу на те, що оскільки середнє обчислюється тільки для тих анкет, де є визначена відповідь (де значення змінної відмінне від відсутнього значення), у другому випадку анкети із варіантом "важко сказати" в обчисленнях середнього не враховуються.

4. Для аналізу змінної v173 не виводьте таблиці, обчисліть медіану, середнє, дисперсію, стандартне відхилення, 5-й та 95-й процентилі, квартилі. Прокоментуйте результати обчислень.

5. На основі аналізу мір центральної тенденції та мір варіації змінних v40–v50 проаналізуйте згоду киян із різними можливими діями влади, спрямованими на вихід з економічної кризи у червні 1991 року.

РОЗДІЛ 4. АНАЛІЗ ЗВ'ЯЗКУ МІЖ ДВОМА ЗМІННИМИ

Від аналізу розподілу однієї окремо виділеної змінної перейдемо до аналізу двох змінних. Будемо розглядати задачу вивчення спільного розподілу двох змінних (побудова та аналіз двовимірних таблиць частот та відсотків, двовимірних таблиць групових середніх, діаграм розсіяння тощо) та задачу вивчення зв'язку між двома змінними.

Вивчення зв'язку між змінними, обчислення кількісних характеристик такого зв'язку (так званих коефіцієнтів зв'язку, кореляції, асоціації тощо) має відповідати тій чи іншій формальній моделі зв'язку. В свою чергу, вибір формальної моделі зв'язку значною мірою залежить від рівня вимірювання (тип шкали кожної із змінних) та від властивостей емпіричних даних, що піддаються аналізу (зокрема дискретною чи неперервною є кожна із змінних).

Ми розглянемо найбільш часто вживані у практичному аналізі емпіричних соціологічних даних моделі зв'язку між двома змінними та відповідні цим моделям характеристики (коефіцієнти) зв'язку. Зокрема розглянемо:

- модель зв'язку як відмінності від статистичної незалежності (коефіцієнт χ^2, коефіцієнт Крамера, коефіцієнт φ^2);
- модель зв'язку як можливості поліпшувати передбачення значення однієї ознаки на основі значення другої ознаки (коефіцієнт λ Гудмана);
- модель зв'язку як сумісної варіації (узгодженої зміни) значень двох ознак (коефіцієнт кореляції Пірсона);
- модель зв'язку як сумісної варіації (узгодженої зміни) рангів двох ознак (коефіцієнт рангової кореляції Спірмена).

4.1. Дві дискретні змінні. Двовимірна таблиця частот і відсотків

Сумісний розподіл двох дискретних змінних із невеликою кількістю категорій зручно подавити у вигляді двовимірної (плоскої) таблиці частот та відсотків. Така таблиця має стільки рядків, скільки є категорій у першій змінній, та стільки стовпчиків, скільки є категорій у другій змінній. Клітина такої таблиці, що є на перетині рядка з номером i та стовпчика з номером j, містить кількість об'єктів, що за першою змінною мають значенням категорію i та одночасно за другою змінною мають значенням категорію j. У таблиці 4.1 зображена

двовимірна таблиця частот сумісного розподілу двох змінних – змінної "Стать", що має дві категорії, та змінної "Освіта", що має чотири категорії. Перший зверху рядок та перший зліва стовпчик таблиці містять назви категорій відповідних змінних. Сама ж таблиця частот містить два стовпчики та чотири рядки. Кожна клітина таблиці містить кількість об'єктів із певною комбінацією значень за цими двома змінними (так звана *клітинкова частота*). Наприклад, легко бачити, що в даних, для яких побудована таблиця, є інформація про 95 чоловіків із вищою освітою (клітина у верхньому лівому куті таблиці), 80 жінок із середньою спеціальною освітою і т. д.

Таблиця 4.1

Двовимірна таблиця частот сумісного розподілу змінних "Стать" та "Освіта"

Освіта	Чоловік	Жінка
Вища	95	109
Середня спеціальна	51	80
Середня загальна	35	33
Незакінчена середня та нижче	10	17

До таблиці клітинкових частот додають ще один стовпчик, що містить суми частот у кожному рядку (так званий *маргінальний стовпчик*), та ще один рядок, що містить суми частот у кожному стовпчику (так званий *маргінальний рядок*). Маргінальний стовпчик та маргінальний рядок, як правило, помічають у таблиці словом "Всього". Для уможливлення порівняння розподілів у окремих рядках та/або у стовпчиках таблиці для кожної клітини обчислюють відсотки відносно суми частот у рядку, де є відповідна клітина (так званий *відсоток у рядку*) та/або відносно суми частот у відповідному стовпчику (так званий *відсоток у стовпчику*).

Таблиця 4.2

Двовимірна таблиця частот та відсотків для сумісного розподілу змінних "Стать" та "Освіта"

Освіта	Чоловік	Жінка	Всього
1	2	3	4
Вища	46,6 % 95 49,7 %	53,4 % 109 45,6 %	204 47, 4 %

Закінчення таблиці 4.2

1	2	3	4
Середня спеціальна	38,9 % 51 26,7 %	61,1 % 80 33,5 %	131 30,5 %
Середня загальна	51.5% 35 18,3%	48.5% 33 13,8%	68 15,8 %
Незакінчена середня та нижче	37,0% 10 5,2 %	63,0 % 17 7,1 %	27 6,3 %
Всього	191 44,4 %	239 55,6 %	430

З таблиці 4,2 бачимо, що 95 чоловіків із вищою освітою (клітина у верхньому лівому куті таблиці) становлять 46,6 % від загальної кількості респондентів із вищою освітою (обчислений для цієї клітини відсоток у рядку) та 49,7 % від загальної кількості чоловіків (відповідний відсоток у стовпчику). Сума всіх відсотків у рядку для кожного рядка дорівнює 100 %. Сума всіх відсотків у стовпчику для кожного стовпчика дорівнює 100 %. Маргінальний стовпчик та маргінальний рядок є одновимірними розподілами для відповідних змінних. До таблиці включаються тільки ті об'єкти, що мають визначені значення двох змінних, для яких побудована таблиця (інакше кажучи, об'єкти із *відсутніми значеннями* хоча б за однією з двох змінних до таблиці не вносять).

Зауваження. У таблицях, що включають до публікацій та презентацій, не рекомендують "перенавантажувати" клітини великою кількістю чисел. Як правило, до клітини включають або тільки один із відсотків, або клітинкову частоту та один із відсотків. Вважається, що краще використовувати відсоток у стовпчику.

У пакеті SPSS для побудови двовимірної таблиці частот та відсотків послідовно обираємо

Analyze ⇒ Descriptive statistics ⇒ Crosstabs ...

(*Аналіз ⇒ Описові статистики ⇒ Таблиці спряженості...*)

Розкривається вікно параметрів **Crosstabs** (див. рис. 4.1).

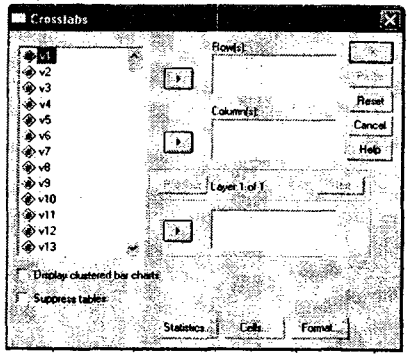


Рис. 4.1. *Вікно параметрів операції* ***Crosstabs.***

У лівій частині вікна параметрів вміщено перелік усіх змінних файлу даних. У правій частині вгорі є два вікна невеликого розміру із заголовками **Row(s)** та **Column(s)**. У вікні **Row(s)** необхідно сформувати перелік змінних, категорії яких виділяють рядки двовимірних таблиць, що ми їх будуємо (*як переносити імена змінних із загального списку до відповідного вікна, дивись у п.1 розділу 3*). У вікні **Column(s)** вказуємо перелік тих змінних, категорії яких виділяють стовпчики цих двовимірних таблиць. Операція **Crosstabs** будує двовимірні таблиці для всіх комбінацій із двох змінних, одна з яких зі списку **Row(s)**, а друга зі списку **Column(s)**. Кожна така пара змінних визначає одну двовимірну таблицю.

Операція **Crosstabs** дає змогу також будувати таблиці розмірності більше ніж 2 (тривимірні, чотиривимірні тощо). Відповідні контрольні змінні (одну або декілька) потрібно вказати у вікні **Layer ... of ...** (*Рівні ... з ...*). Багатовимірна таблиця подається у вікні результатів **Output** як кілька двовимірних таблиць, побудованих для двох зазначених у параметрах **Row(s)** та **Column(s)** при фіксованих значеннях вказаних у параметрі **Layer** контрольних змінних. Наприклад, якщо, будуючи двовимірну таблицю для змінних "Стать" та "Освіта", як контрольну змінну вказати змінну "Місце проживання", що має дві категорії – "місто" та "село", то операція **Crosstabs** побудує дві таблиці сумісного розподілу статі та освіти – одну для респондентів із міста, а другу – для респондентів із села.

Як приклад розглянемо двовимірну таблицю спільного розподілу освіти респондентів (змінна v167) та їх ставлення до посилення в країні контролю за виконанням трудової дисципліни, включаючи

перевірку громадян на вулицях у робочий час (v133). Для цього виконуємо такі дії:

- відкриваємо вікно параметрів операції: послідовно обираємо в головному меню **Analyze ⇒ Descriptive statistics ⇒ Crosstabs ...;**
- переносимо до списку **Row(s)** ім'я змінної v133;
- переносимо до списку **Column(s)** ім'я змінної v167;
- натискаємо кнопку **OK**.

Як видно з таблиць 4.3, 4.4, результат роботи операції **Crosstabs** складається з двох частин. У лівому стовпчику **Valid** першої частини результату ми бачимо, що двовимірна таблиця буде побудована для 422 респондентів, які складають 97,9 % від загальної кількості опитаних. Ці респонденти дали відповідь на обидва питання, які нас цікавлять (про освіту та про ставлення до підсилення контролю за дисципліною). Не відповіли принаймні на одне з цих двох питань (середній стовпчик **Missing**) 9 респондентів (2,1 % від загальної кількості опитаних). Ці респонденти не включаються до двовимірної таблиці. У цілому ж було опитано 431 респондент (правий стовпчик Total).

Таблиця 4.3

Інформація про кількість спостережень[9]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Усиление контроля за дисциплиной* Ваше образование	422	97,9 %	9	2,1 %	431	100,0 %

[9] Таблиці 4.3, 4.4 та 4.5 є перенесеними з програми SPPP (файл kiev.91) результатами обчислень процедури, яка розглядається, а тому мітки змінних та мітки значень подані російською мовою, адже саме цією мовою вони прописані у представлених нам для обробки та аналізу файлах kiev.89 та kiev.91.

Таблиця 4.4

Двовимірна таблиця

Усиление контроля за дисциплиной * Ваше образование Crosstabulation

Count

		Ваше образование				Total
		высшее	Средн. спец.	сред нее общ.	н/сред-нее	
Усиление контроля за дисциплиной	не согл.	150	73	39	7	269
	не знаю	11	7	7	5	30
	согласен	42	49	18	14	123
Total		203	129	64	26	422

Друга частина результату – це, власне, двовимірна таблиця частот. З таблиці видно, що 150 респондентів із вищою освітою не погоджуються з думкою про те, що підсилення контролю за трудовою дисципліною спрямовано на захист їхніх інтересів (клітина у верхньому лівому куті таблиці). У той же час є 14 респондентів із незакінченою середньою освітою, які погоджуються із такою думкою (клітина на перетині рядка "погоджуюсь" та стовпчика "н/середня"). Всього ми маємо 203 респонденти із вищою освітою (див. у рядку *Total*) та 123 респонденти, що погоджуються із зазначеною вище думкою (див. у стовпчику *Total*).

Ми бачимо, що є 42 респонденти з вищою освітою та 14 респондентів із незакінченою середньою освітою, що позитивно ставляться до підсилення контролю за виконанням трудової дисципліни. На перший погляд здається, що респонденти із вищою освітою більше підтримують контроль за дисципліною, ніж респонденти із незакінченою середньою освітою. Проте група респондентів із незакінченою середньою освітою є набагато менш чисельною порівняно з групою респондентів із вищою освітою (26 та 203 відповідно), тому і кількість незгодних у цій групі є меншою порівняно з кількістю незгодних у групі респондентів із вищою освітою. Для вирішення питання про те, яка з двох освітніх груп більше підтримує посилення дисципліни, важливим є не абсолютна, відносна (щодо розміру групи) кількість симпатиків підсилення контролю за трудовою дисципліною. Інакше кажучи, нам потрібно знати, яку частку або який відсоток становлять ті, хто підтримує підсилення контролю за трудовою дисципліною у відповідній освітній групі.

За загальною угодою процедура **Crosstabs** обчислює лише клітинкові частоти. Для того, щоб замовити обчислення відсотків, потрібно натиснути кнопку [**Cells...**]. Розкриється вікно параметрів (див. рис. 4.2), які дозволяють задати вміст кожної клітини двовимірної таблиці.

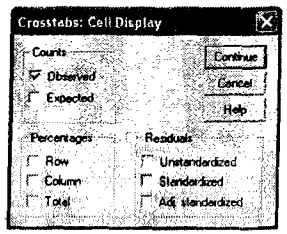


Рис. 4.2. *Параметри відображення вмісту клітин двовимірної таблиці.*

Для кожної клітини двовимірної таблиці можна обчислити та вивести такі показники:

I. Група Counts –

- клітинкова частота, що ми її дійсно спостерігаємо в емпіричних даних (параметр **Observed**);
- клітинкова частота, що ми її могли б очікувати у випадку статистичної незалежності між змінними, для яких побудована таблиця (параметр **Expected**); більш детально про статистичну незалежність буде у пункті 4.2.

II. Група Percentages –

- відсоток, що його становить клітинкова частота відносно суми частот усіх клітин у відповідному рядку (параметр **Row**);
- відсоток, що його становить клітинкова частота відносно суми частот усіх клітин у відповідному стовпчику (параметр **Column**);
- відсоток, що його становить клітинкова частота відносно суми частот усіх клітин таблиці (параметр **Total**).

III. Група Residuals – відмінність клітинкової частоти (дійсної, що ми її спостерігаємо у емпіричних даних) від очікуваної клітинкової частоти (що ми її могли б очікувати у випадку статистичної незалежності двох змінних, для яких побудована таблиця);

- є можливість обчислити та вивести ненормовану, нормовану та уточнену нормовану різницю (відповідно, параметри **Unstandardized**, **Standardized** та **Adj. Standardized** у групі).

Ці показники інформують дослідника про те, для яких категорій об'єктів дійсні частоти найбільше відрізняються від очікуваних.

Якщо, будуючи двовимірну таблицю для змінних v167 та v133, додатково замовити обчислення відсотків у стовпчику (натиснути кнопку **[Cells...]** та обрати параметр **Column** у групі **Percentages**), то отримаємо певний результат (див. таблицю 4.5).

Таблиця 4.5

Двовимірна таблиця для V167 та V133 із обчисленням відсотків у стовпчику

Усиление контроля трудовой дисциплины, включая проверку гражд * Ваше образование? Crosstabulation

			Ваше образование?				Total
			высшее, незаконченное высшее (3-4 курса вуза)	среднее специальное, среднее профессионально-техническое (ПТ)	среднее общее (10-11 классов)	9 классов и меньше	
1	2	3	4	5	6	7	8
Усиление контроля трудовой дисциплины, включая проверку граждан	не согласен	Count	150	73	39	7	269
		% within Ваше образование?	73,9 %	56,6 %	60,9 %	26,9 %	63,7 %

Закінчення таблиці 4.5

1	2	3	4	5	6	7	8
	не знаю	Count	11	7	7	5	30
		% within Ваше образование?	5,4 %	5,4 %	10,9 %	9,2 %	7,1 %
	согласен	Count	42	49	18	14	123
		% within Ваше образование?	20,7 %	38,0 %	28,1 %	53,8%	29,1 %
Total		Count	203	129	64	26	422
		% within Ваше образование?	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %

З цієї таблиці одразу видно, що твердження про зростання підтримки підсилення контролю за трудовою дисципліною зі зменшенням освіти є необґрунтованим. Дійсно, адже в групі респондентів із вищою освітою 20,7 % опитаних позитивно оцінюють підсилення контролю за трудовою дисципліною, в той час як серед респондентів із незакінченою середньою освітою такі можливі дії підтримують майже 54 % опитаних.

Натискання на кнопку **[Format...]** розкриває вікно (див. рис. 4.3), де є можливість обрати один із двох способів упорядкування рядків двовимірної таблиці: за збільшенням значень числових кодів альтернатив (значення **Ascending** у групі **Row Order**, таке впорядкування використовується за загальною угодою) або за їх зменшенням (значення **Descending** у групі **Row Order**).

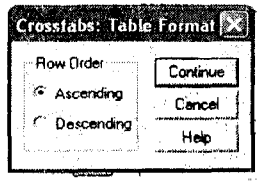


Рис. 4.3. *Вікно вибору способу впорядкування рядків таблиці.*

У дослідника є також можливість не виводити двовимірні таблиці частот та відсотків і обмежитися лише обчисленням показників, які характеризують зв'язок між двома змінними (про такі показники йтиметься далі, у пункті 4.2). Для цього потрібно обрати параметр

Suppress tables під списком змінних у основному вікні операції **Crosstabs...** (див. рис. 4.1).

Крім того, є можливість на додаток до двовимірної таблиці частот та відсотків вивести ще й її графічне зображення у вигляді згрупованої стовпчикової діаграми. Для цього потрібно обрати параметр **Display clustered bar charts** під списком змінних у основному вікні операції **Crosstabs...** (див. рис. 4.1). Детальніше побудову різних графіків ми будемо розглядати у розділі 5.

4.2. Міри зв'язку для двох дискретних змінних

Операція **Crosstabs...** дає змогу не тільки будувати двовимірні таблиці частот та відсотків, а й обчислювати показники зв'язку між двома змінними, що відповідають моделям зв'язку. Для того, щоб замовити обчислення необхідних показників зв'язку, потрібно натиснути кнопку **Statistics**. Відкривається вікно, в якому дослідник може обрати один або декілька необхідних йому показників (див. рис. 4.4).

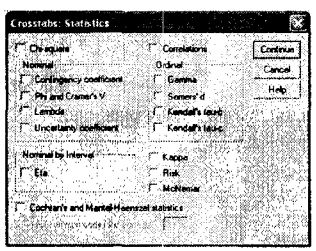


Рис. 4.4. *Вікно показників зв'язку між двома змінними операції* ***Crosstabs...***

У цьому розділі розглянемо деякі з найбільш часто вживаних показників для двох дискретних змінних, що їх обчислює операція **Crosstabs...**

Коефіцієнт χ^2 (читається "хі-квадрат", часто говорять "хі-квадрат Пірсона" – на честь визначного англійського статистика Карла Пірсона; параметр **Chi-square** у вікні **Statistics**) може набувати будь-які позитивні значення і є мірою відмінності побудованої на емпіричних даних (як результат спостережень) двовимірної таблиці частот від гіпотетичної двовимірної таблиці частот (тобто такої таблиці частот, яка мала би бути за умови, що дві змінні є статистично незалежними). Модель зв'язку полягає в тому, що зв'язок розглядається як відмінність від статистичної незалежності. За допомогою коефіцієнта χ^2, що як випадкова величина має статистичний

розподіл χ^2, перевіряється статистична гіпотеза про відмінність між емпіричною та гіпотетичною таблицями. Коефіцієнт χ^2 характеризується також кількістю ступенів волі (англійською мовою *degree of freedom*, часто позначають як **df**), що обчислюється як добуток "кількості рядків у таблиці мінус одиниця" на "кількість стовпчиків у таблиці мінус одиниця". Відкинути гіпотезу про відсутність значної відмінності між емпіричною та гіпотетичною таблицями означає прийняти рішення про наявність зв'язку між двома змінними. Оскільки ми маємо справу із вибірковими даними, то кожне таке рішення має статистичний характер, а отже, може бути помилковим і тому супроводжується ймовірністю такої помилки. У цьому випадку ми на основі значення коефіцієнта χ^2 та кількості ступенів волі, керуючись знаннями статистичного розподілу χ^2, приймаємо рішення про відмінність статистичної незалежності між двома змінними, а отже (відповідно до нашої моделі зв'язку), про наявність між цими змінними статистичного зв'язку. Ймовірність припуститися помилки в результаті прийняття такого рішення називається значущістю (англійською мовою *significance*, часто позначають як Sig.) і ця ймовірність має бути невеликою. Прийнято вважати (це певна угода між дослідниками), що значущість не повинна сильно перевищувати значення 0,05. Інакше кажучи, ми приймаємо рішення про наявність зв'язку між змінними за умови, що значущість не перевищує або ж "не дуже перевищує" значення 0.05[10].

У попередньому розділі ми розглянули побудову двовимірної таблиці для змінних v133 та v167. Якщо обчислити коефіцієнт χ^2 для цієї таблиці (натиснути кнопку [**Statistics...**], поставити позначку біля параметра **Chi-square**), то додатково отримаємо таблицю 4.6.

Таблиця 4.6

Значення критерію χ^2 для двовимірної таблиці v167 на v133

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	31.851(a)	6	.000
Likelihood Ratio	30.651	6	.000
Linear-by-Linear Association	17.553	1	.000
N of Valid Cases	422		

a 2 cells (16.7%) have expected count less than 5. The minimum expected count is 1.85.

[10] У деяких випадках значущість 0.06 або навіть 0.08 теж вважають за прийнятну.

Нас у цій таблиці цікавить лише перший рядок. Коефіцієнт χ^2 Пірсона[11] (*Pearson Chi-Square*) дорівнює 31.851 (стовпчик *Value*), кількість ступенів волі таблиці дорівнює 6 (стовпчик *df*; у побудованій нами двовимірній таблиці частот та відсотків є три рядки та чотири стовпчики, див. таблицю 4.5) і рівень значущості є меншим ніж 0.001 (округлене до трьох знаків після десяткової крапки значення ймовірності помилки представлено в стовпчику *Asymp. Sig. (2-sided)* як 0.000). Отже, з дуже невеликою ймовірністю помилитися (ця ймовірність менше ніж 0.001) ми можемо стверджувати, що змінні v133 та v167 не є статистично незалежними, тобто наші емпіричні дані свідчать, що між цими двома змінними існує статистичний зв'язок.

Отже, ми маємо деяке ***правило*** для прийняття рішення про можливу наявність зв'язку між двома змінними. Це правило часто називають критерієм χ^2, або тестом χ^2. Для того, щоб результат застосування тесту χ^2 був надійним, необхідно, щоб кількість об'єктів, для яких ми будуємо таблицю, була не менше 100, і щоб у кожній клітині таблиці очікувана частота була не менше ніж 5 (тобто, як ще інколи говорять, у таблиці не було погано наповнених клітин). Саме тому під таблицею з інформацією про χ^2 SPSS сповіщає про кількість таких погано наповнених клітин. У нашому випадку таких клітин у таблиці 2 і мінімальне значення очікуваної частоти дорівнює 1.85.

Критерій χ^2 пов'язаний із моделлю зв'язку як відмінність від статистичної незалежності. Цей показник може набувати будь-які позитивні значення, і тому його не можна використовувати для оцінювання сили зв'язку. У випадку, коли коефіцієнт χ^2 значущий, є сенс обчислювати та інтерпретувати такі показники зв'язку, як коефіцієнт Крамера та коефіцієнт спряженості Пірсона.

Для того, щоб обчислити коефіцієнт спряженості Пірсона, потрібно у вікні **Statistics** процедури **Crosstabs** у групі **Nominal** (*Номінальна*) поставити позначку біля **Contingency coefficient** (*Коефіцієнт контингенції*). Коефіцієнт спряженості Пірсона обчислюється за формулою $c = \sqrt{\frac{\chi^2}{\chi^2 + N}}$ і змінюється в інтервалі від 0 до 1 (значення 1 показник ніколи не набуває). Коефіцієнт є симетричним (не спрямованим) показником. Значення 0 інтерпретується як відсутність зв'язку між змінними (як статистична незалежність). Чим більшим є

[11] К. Пірсон (1857–1936) – визначний англійський статистик.

значення коефіцієнта, тим сильнішим є зв'язок. Звичайно, інтерпретувати коефіцієнт є сенс лише у випадку, якщо відповідний χ^2 є значущим на потрібному нам рівні.

Для того, щоб обчислити коефіцієнт асоціації Крамера, потрібно у вікні **Statistics** процедури **Crosstabs** у групі **Nominal** (*Номінальна*) поставити позначку біля **Phi and Cramer's V** (*Фі та коефіцієнт Крамера V*). Для таблиць розмірності 2×2 операція **Crosstabs** обчислює коефіцієнт "Фі" за формулою $\varphi = \sqrt{\frac{\chi^2}{N}}$. Чим більшим є значення φ, тим сильнішим є зв'язок. Для таблиць, у яких кількість рядків та/або стовпчиків є більшою ніж 2, обчислюється коефіцієнт Крамера за формулою $V = \sqrt{\frac{\chi^2}{N \cdot (k-1)}}$, де k – мінімум із кількості стовпчиків та рядків таблиці. Коефіцієнт Крамера набуває значень від 0 до 1. Значення 0 інтерпретується як відсутність зв'язку (статистична незалежність). Значення 1 інтерпретується як повний (максимально сильний) зв'язок. Обидва ці коефіцієнти є сенс розглядати та інтерпретувати лише тоді, коли відповідний χ^2 є значущим на потрібному нам рівні.

Коефіцієнт λ Гудмана (читається "лямбда") пов'язаний із моделлю зв'язку між двома дискретними номінальними змінними як можливість передбачати (прогнозувати) значення однієї змінної за значенням другої змінної. Такий зв'язок є спрямованим. Виділяється незалежна змінна та залежна змінна. Модель передбачає розгляд можливості передбачати значення залежної змінної на основі відомого значення незалежної змінної. Чим більше знання незалежної змінної зменшує помилку в передбаченні значення залежної змінної, тим сильнішим є зв'язок. Оскільки зв'язок розглядається як спрямований, то відповідні коефіцієнти є несиметричними.

Для того, щоб обчислити коефіцієнт λ Гудмана, потрібно у вікні **Statistics** процедури **Crosstabs** у групі **Nominal** (*Номінальна*) поставити позначку біля **Lambda**. Процедура **Crosstabs** обчислює обидва коефіцієнти λ – один для випадку, коли вказана в параметрі **Column(s)** змінна розглядається як незалежна, і другий для випадку, коли ця змінна розглядається як залежна. Дослідник обирає потрібний коефіцієнт відповідно до логіки аналізу. Коефіцієнт λ набуває значень в інтервалі від 0 до 1. Значення коефіцієнта λ інтерпретується як частка зменшення помилки передбачення за-

лежної змінної в результаті застосування знання про значення незалежної змінної. Значення $\lambda_{yx}=0$ інтерпретується як відсутність впливу незалежної ознаки x на залежну ознаку y (як відсутність зв'язку між x та y), оскільки знання x ніяк не зменшує помилку передбачення y. Значення $\lambda_{yx}=1$ інтерпретується як повний вплив x на y, оскільки знання x зменшує помилку у передбаченні y на 100 % (інакше кажучи, на основі знання значення x ми можемо абсолютно точно визначити значення y). Проміжні значення між 0 та 1 також мають відповідну інтерпретацію. Так, наприклад, $\lambda_{yx}=0{,}3$ інтерпретується як наявність помірного впливу x на y, оскільки знання значення x зменшує помилку у передбаченні y на 30 %.

Приклад. З'ясуємо на даних емпіричного соціологічного дослідження, чи існував у 1991 р. у місті Києві зв'язок між статтю та освітою.

У нашому файлі даних є інформація про стать респондента (змінна V165) та про рівень його або ж її освіти (змінна v167). Ми маємо дві дискретні змінні. Зв'язок будемо розглядати як відмінність від статистичної незалежності. Для вирішення поставленого питання обчислимо показник χ^2 та коефіцієнт Крамера.

З цією метою послідовно обираємо **Analyze ⇒ Descriptive statistics ⇒ Crosstabs ...** . У вікні, що розкривається (див. рис. 4.1), заносимо змінну v167 до поля **Row(s)** та змінну v165 до поля **Column(s)**. Натискаємо кнопку [**Statistics**] та у вікні, що розкривається (див. рис. 4.4), ставимо позначку біля **Chi-square** та біля **Phi and Cramer's V** у групі **Nominal**. Натискаємо кнопку [**Continue**] та закриваємо вікно **Statistics**. Оскільки нас цікавлять лише коефіцієнти, то ставимо позначку біля **Suppress tables** (*Не виводити таблиці*). Після цього натискаємо кнопку [**Ok**] і отримуємо результат (див. таблицю 4.7 і таблицю 4.8).

Таблиця 4.7

Значення коефіцієнта χ^2 для двовимірної таблиці v165 на v167

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3,945[a]	3	,267
Likelihood Ratio	3,960	3	,266
Linear-by-Linear Association	,142	1	,706
N of Valid Cases	430		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 11,99.

Таблиця 4.8

Значення коефіцієнта Крамера для двовимірної таблиці v165 на v167

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,096	,267
	Cramer's V	,096	,267
N of Valid Cases		430	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Коефіцієнт χ^2 має значення 3.945 і є незначущим, оскільки вказане в стовпчику **Asimp. Sig. (2-sided)** значення дорівнює 0.267, а нам для статистично обґрунтованого твердження про наявність зв'язку потрібно, щоб це значення не перевищувало 0.05. Інакше кажучи, емпіричні дані не дають підстав стверджувати, що між статтю та освітою існує зв'язок. Незначущість χ^2 має наслідком також те, що значення коефіцієнта Крамера 0,096 взагалі не інтерпретується. У цілому ж відсутність зв'язку в цьому випадку можна інтерпретувати як рівність доступу до освіти та рівність здібностей до набуття освіти чоловіків та жінок.

Вивчаючи зв'язок між двома кількісними змінними (вимірюваними за шкалою інтервалів або ж відношень), доцільно використовувати модель зв'язку як коваріації (сумісної варіації) значень двох змінних.

Для оцінювання сили та типу лінійного зв'язку в межах такої моделі часто використовують коефіцієнт кореляції Пірсона, який часто позначають літерою r. Цей показник є симетричним (у парі змінних не виділяють залежну та незалежну змінні), застосовується для двох кількісних змінних та набуває значень в інтервалі від -1 до 1. Знак коефіцієнта кореляції Пірсона інтерпретується як тип зв'язку. Додатний коефіцієнт свідчить про позитивний, або прямий, зв'язок, а від'ємний коефіцієнт – про негативний, або зворотний, зв'язок. Наявність прямого зв'язку свідчить про те, що збільшення значення однієї ознаки пов'язано зі збільшенням значення другої та, відповідно, зменшення пов'язано зі зменшенням. Водночас наявність зворотного зв'язку означає, що збільшення значення однієї ознаки пов'язано зі зменшенням значення другої та, відповідно, зменшення пов'язано зі збільшенням.

Абсолютне значення коефіцієнта кореляції Пірсона інтерпретується як сила лінійного зв'язку. Якщо $r_{xy}=0$, то говорять про відсутність лінійного зв'язку між x та y. У випадку $r_{xy}=+1$ говорять про прямий функціональний лінійний зв'язок між x та y. У випадку ж, коли $r_{xy}=-1$, то ми маємо справу зі зворотним функціональним лінійним зв'язком. Зазвичай коефіцієнт кореляції Пірсона набуває проміжних значень між 0 та 1. Питання про те, чи є хоч якийсь зв'язок між двома змінними (тобто чи відрізняється відповідний коефіцієнт від нуля), може бути статистично з'ясовано на емпіричних даних. Для того, щоб можна було говорити про наявність лінійного зв'язку між змінними, потрібно, щоб рівень значущості відповідного коефіцієнта кореляції не перевищував 0.05.

Для того, щоб обчислити коефіцієнт кореляції Пірсона, потрібно у вікні **Statistics** процедури **Crosstabs** поставити позначку поруч із **Correlations**.

Разом із коефіцієнтом кореляції Пірсона процедура **Crosstabs** також обчислює та виводить відповідний коефіцієнт рангової кореляції Спірмена. Цей показник позначають літерою ρ, застосовують для двох порядкових змінних і пов'язаний він із моделлю зв'язку як коваріації рангів. Коефіцієнт рангової кореляції Спірмена, як і коефіцієнт кореляції Пірсона, набуває значення в інтервалі від -1 до +1. Значення +1 інтерпретується як однакове впорядкування рангів значень двох ознак. Значення -1 інтерпретується, відповідно, як зворотне (обернене) впорядкування рангів значень двох ознак. Значення 0 інтерпретується як відсутність зв'язку між порядком рангів значень двох ознак.

Для двох рангових змінних є можливість за допомогою SPSS обчислити ряд коефіцієнтів, серед яких найбільш часто використовують коефіцієнт Кендела. Його позначають літерою τ (читається "тау"), і він набуває значень від -1 (максимально можливий негативний зв'язок) до +1 (максимально можливий позитивний зв'язок). Пакет SPSS обчислює дві версії цього показника, які позначають, відповідно, τ_b та τ_c. Для того, щоб обчислити коефіцієнт рангової кореляції Кандела, потрібно у вікні **Statistics** процедури **Crosstabs** поставити у групі **Ordinal** позначку біля **Kendall's tau-b** та/або біля **Kendall's tau-c**.

4.3. Дискретна та неперервна змінні. Таблиця групових середніх

У випадку, коли одна змінна – дискретна, а інша – кількісна неперервна, побудова двовимірної таблиці за допомогою **Crosstabs** є неефективною. За таких обставин більш зручним для вивчення сумісного розподілу та зв'язку є побудова таблиці групових середніх. У такій таблиці дискретна змінна виділяє категорії, в межах яких виконується усереднення та обчислення узагальнюючих характеристик для неперервної змінної.

Для побудови таблиць групових середніх у SPSS потрібно послідовно обрати:

Analyze ⇒ Compare Means ⇒ Means...

(*Аналіз ⇒ Порівняння середніх ⇒ Середні...*)

Відкривається вікно параметрів операції **Means** (див. рис. 4.5).

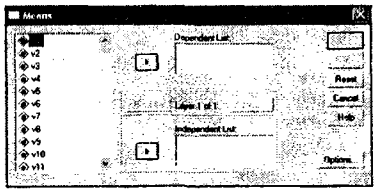


Рис. 4.5. *Діалогове вікно* ***Means.***

Поле **Dependent List** (*Список залежних*) має містити список залежних змінних. Ці змінні повинні бути кількісними, бо їх значення будуть усереднюватися.

Поле **Independent List** (*Список незалежних*) має містити список незалежних дискретних змінних. Категорії незалежної змінної виділяють групи, у яких виконується усереднення залежної змінної.

Якщо натиснути на кнопку **[Options...]** (*Опції*), то розкривається вікно, яке дає змогу визначити структуру (перелік стовпчиків) таблиці (див. рис. 4.6). У переліку **Statistics** можна обрати та перенести до переліку **Cell Statistics** потрібні для представлення показники. Зокрема, є можливість показати у таблиці групових середніх:

- **Mean** – середнє арифметичне в межах групи;
- **Number of Cases** – кількість спостережень у групі;
- **Standard Deviation** – стандартне відхилення у межах групи;
- **Variance** – дисперсія в межах групи та інші показники.

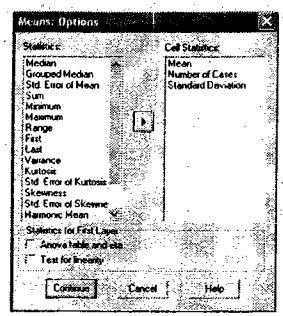


Рис. 4.6. *Вікно вобору показників* ***Means: Options.***

4.4. Міри зв'язку між дискретною та неперервною змінними

Для оцінювання зв'язку між дискретною та неперервною змінними є можливість обчислити кореляційне відношення (коефіцієнт η, читається "ета"). Дискретна змінна розглядається як фактор (незалежна змінна), що впливає на неперервну змінну (на залежну змінну). Коефіцієнт η є несиметричним і набуває значень від 0 (відсутність зв'язку, відсутність впливу) до 1 (повний зв'язок, повний вплив).

Для того, щоб обчислити кореляційне відношення η, потрібно у вікні **Statistics** процедури **Crosstabs** поставити в групі **Nominal by Interval** (*Номінальна на інтервальну*) позначку поруч із **Eta**. Процедура **Crosstabs** обчислює обидва коефіцієнти η – один для випадку, коли вказана в параметрі **Column(s)** змінна розглядається як незалежна, і другий для випадку, коли ця змінна розглядається як залежна. Дослідник обирає потрібний коефіцієнт за логікою аналізу.

4.5. Дві неперервні змінні. Матриця парних та часткових кореляцій

Пакет SPSS дає змогу будувати матрицю коефіцієнтів кореляції Пірсона для списку змінних. Для цього потрібно послідовно обрати

Analyze ⇒ Correlate ⇒ Bivariate
(*Аналіз ⇒ Кореляція ⇒ Парні*)

У вікні параметрів, що відкривається (див. рис. 4.7), у полі **Variables:** потрібно сформувати список змінних, для яких необхідно побудувати квадратну матрицю парних кореляцій. Є можливість обрати один із трьох видів коефіцієнтів кореляції. Для цього потрібно

у групі **Correlation coefficients** (*Коефіцієнти кореляції*) поставити позначку біля **Pearson** (*Коефіцієнт кореляції Пірсона* r), **Kendall tau-b** (*Коефіцієнт рангової кореляції Кандела* τ_b) або **Spearman** (*Коефіцієнт рангової кореляції Спірмена* ρ).

За загальною угодою програма обчислюватиме коефіцієнт кореляції Пірсона (*Pearson*), виводитиме для нього значення двосторонньої значущості (*Two-tailed*), а також відмітить двома зірочками значущі коефіцієнти (*Flag significant correlations*).

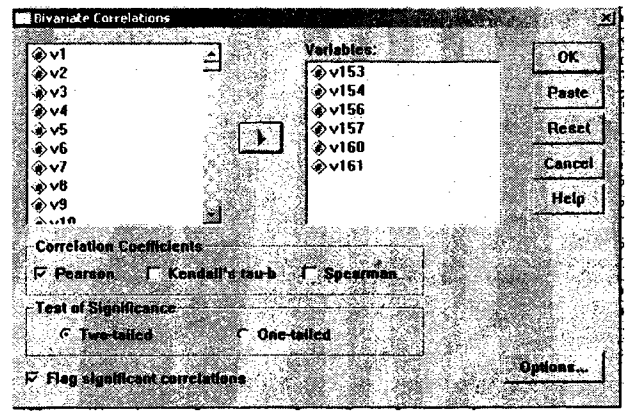


Рис. 4.7. *Вікно параметрів обчислення парних кореляцій*

Приклад. У дослідженні 1991 р. киянам пропонувалося оцінити свій власний ступінь довіри до деяких активних на той час українських політиків. Оцінки довіри до політиків, перерахованих за прізвищем в алфавітному порядку, містяться у групі змінних від v151 до v164. Дивимося кодування цих змінних (для цього послідовно обираємо Utilities ⇒ Variables) і бачимо, що кодами 5 та 6 закодовані варіанти відповіді "Не знаю такого політика" та "Важко сказати" відповідно. Будемо вважати, що на основі таких відповідей не можна скласти враження про ступінь довіри респондента до політика, а отже, коди 5 та 6 визначимо як коди відсутніх значень, задані користувачем (*user missing value*). Перші чотири варіанти відповіді змістовно впорядковані за збільшенням недовіри до політичного діяча і, відповідно, їм присвоєні коди, що збільшуються (послідовні цілі числа від 1 до 4). Оскільки коефіцієнт кореляції Пірсона є досить стійким до нерівномірності відстані між окремими пунктами шкали, то його використовують досить часто для таких, як ми маємо в цьому

випадку, порядкових шкал. Отже, розглянемо кореляції між довірою киян до різних політичних діячів. Позитивну кореляцію довіри до двох політиків будемо інтерпретувати як прояв схожості цих політиків в оцінках респондентів. Відповідно, негативну кореляцію будемо інтерпретувати як прояв протиставлення політиків в оцінках респондентів.

Побудуємо матрицю парних коефіцієнтів кореляції для змінних v153, v154, v156, v157, v160 та v161. Для цього:

- послідовно обираємо в головному меню **Analyze ⇒ Correlate ⇒ Bivariate;**
- заносимо у вікно **Variables:** перелік потрібних нам змінних (див. рис. 4.7);
- натискаємо кнопку [**Options...**], відкривається вікно параметрів (див. рис. 4.8);

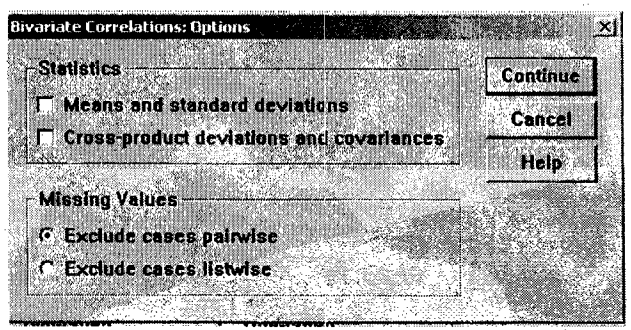


Рис. 4.8. *Вікно параметрів операції Correlate.*

- обираємо у групі **Missing Values** (*Відсутні значення*) один із двох можливих способів роботи з відсутніми значеннями – **Exclude cases pairwise** (*Видаляти з обчислень спостереження для кожної пари змінних окремо*) або **Exclude cases listwise** (*Видаляти з обчислень спостереження для всього списку змінних*);
- натискаємо кнопку [**Ok**].

У таблицях 4.9 та 4.10 подано результати обчислення кореляційної матриці для шести змінних та двох різних способів роботи із відсутніми значеннями.

Таблиця 4.9

Матриця парних кореляцій рівня довіри до політиків. Відсутні значення обробляються для кожної пари змінних окремо (метод pairwise)

Correlations

		В какой степени Вы доверяете И.Драчу?	В какой степени Вы доверяете Л.Кравчуку?	В какой степени Вы доверяете А.Морозу?	В какой степени Вы доверяете Л.Плющу?	В какой степени Вы доверяете С.Хмаре?	В какой степени Вы доверяете В.Черновол у?
В какой степени Вы доверяете И.Драчу?	Pearson Correlation	1,000	-,195**	,086	-,140*	,556**	,710**
	Sig. (2-tailed)	.	,002	,383	,041	,000	,000
	N	254	244	106	213	224	202
В какой степени Вы доверяете Л.Кравчуку?	Pearson Correlation	-,195**	1,000	,233*	,514**	-,323**	-,293**
	Sig. (2-tailed)	,002	.	,014	,000	,000	,000
	N	244	330	110	238	259	227
В какой степени Вы доверяете А.Морозу?	Pearson Correlation	,086	,233*	1,000	,352**	,010	-,077
	Sig. (2-tailed)	,383	,014	.	,000	,919	,452
	N	106	110	114	108	102	97
В какой степени Вы доверяете Л.Плющу?	Pearson Correlation	-,140*	,514**	,352**	1,000	-,094	-,158*
	Sig. (2-tailed)	,041	,000	,000	.	,165	,028
	N	213	238	108	245	221	195
В какой степени Вы доверяете С.Хмаре?	Pearson Correlation	,556**	,323**	,010	-,094	1,000	,762**
	Sig. (2-tailed)	,000	,000	,919	,165	.	,000
	N	224	259	102	221	281	226
В какой степени Вы доверяете В.Черноволу?	Pearson Correlation	,710**	-,293**	-,077	-,158*	,762**	1,000
	Sig. (2-tailed)	,000	,000	,452	,028	,000	.
	N	202	227	97	195	226	241

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

*. Correlation is significant at the 0.05 level (2-tailed).

Таблиця 4.10

Матриця парних кореляцій рівня довіри до політиків. Відсутні значення обробляються списком (метод listwise)

Correlations[a]

		В какой степени Вы доверяете И.Драчу?	В какой степени Вы доверяете Л.Кравчуку?	В какой степени Вы доверяете А.Морозу?	В какой степени Вы доверяете Л.Плющу?	В какой степени Вы доверяете С.Хмаре?	В какой степени Вы доверяете В.Черновол у?
В какой степени Вы доверяете И.Драчу?	Pearson Correlation	1,000	-,203	,040	-,017	,659**	,755**
	Sig. (2-tailed)	.	,059	,715	,875	,000	,000
В какой степени Вы доверяете Л.Кравчуку?	Pearson Correlation	-,203	1,000	,270*	,553**	-,240*	-,183
	Sig. (2-tailed)	,059	.	,011	,000	,025	,091
В какой степени Вы доверяете А.Морозу?	Pearson Correlation	,040	,270*	1,000	,378**	,042	-,049
	Sig. (2-tailed)	,715	,011	.	,000	,698	,654
В какой степени Вы доверяете Л.Плющу?	Pearson Correlation	-,017	,553**	,378**	1,000	,058	,053
	Sig. (2-tailed)	,875	,000	,000	.	,591	,625
В какой степени Вы доверяете С.Хмаре?	Pearson Correlation	,659**	-,240*	,042	,058	1,000	,832**
	Sig. (2-tailed)	,000	,025	,698	,591	.	,000
В какой степени Вы доверяете В.Черноволу?	Pearson Correlation	,755**	-,183	-,049	,053	,832**	1,000
	Sig. (2-tailed)	,000	,091	,654	,625	,000	.

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

*. Correlation is significant at the 0.05 level (2-tailed).

a. Listwise N=87

Обидві кореляційні матриці є квадратними. На діагоналі стоїть значення 1 (оскільки кожна змінна повністю та позитивно корелює сама із собою). Матриці є симетричними відносно головної діагоналі. Всі значення в кореляційній матриці не перевищують за модулем 1 (оскільки таким є обмеження значень коефіцієнта кореляції Пірсона).

У першій матриці (робота з відсутніми значеннями методом pairwise) кожен коефіцієнт кореляції обчислюється для всіх тих спостережень, які містять реальні (тобто відмінні від відсутніх) значення у двох відповідних змінних. Тому в кожній клітині матриці, крім власне значення коефіцієнта кореляції (верхнє число) та значущості (середнє число), міститься також інформація про кількість спостережень, для яких обчислений відповідний коефіцієнт кореляції. Як бачимо, кількість відсутніх значень (а до відсутніх значень ми зарахували й відмову від відповіді і незнання політика) є досить значною. Так, найбільше респондентів дали відповідь стосовно довіри до Л. Кравчука (330 респондентів із 431), найменше – стосовно довіри до О. Мороза (114 респондентів із 431).

Водночас під час побудови другої матриці використана інша стратегія роботи з відсутніми значеннями (метод listwise). Якщо у певного спостереження хоча б в одній із шести змінних міститься відсутнє значення, це спостереження виключається з обчислення всіх коефіцієнтів матриці. Перевагою такого підходу є те, що всі коефіцієнти обчислюються для однієї й тієї ж самої сукупності респондентів. Недоліком є те, що значна кількість спостережень вилучається з аналізу. Так, у нашому прикладі друга матриця обчислена лише для 87 респондентів (оскільки тільки 87 респондентів дали змістовну відповідь про свою довіру для всіх шести політичних діячів).

Ті коефіцієнти кореляції, які є значущими (тобто дійсно відрізняються від нуля), характеризуються рівнем значущості (середнє число в клітині), що не перевищує 0.05. Такі клітини помічені однією або двома зірочками.

Високопозитивно корелюють між собою довіра до І. Драча, С. Хмари та В. Чорновола. Також позитивно корелюють між собою довіра до Л. Кравчука та до І. Плюща. З іншого боку, довіра до В. Чорновола негативно корелює з довірою до Л. Кравчука та до І. Плюща. Отже, матриця парних коефіцієнтів кореляції демонструє, що у сприйнятті опитаних респондентів шість політичних діячів створюють дві групи – це "подібні між собою" І. Драч, С. Хмара та В. Чорновіл (перша група, досить висока позитивна кореляція між всіма парами) та

відмінні від них (негативна кореляція з політиками із першої групи), але теж подібні між собою (досить висока позитивна кореляція) Л. Кравчук та І. Плющ. Дещо особливу позицію між цими двома групами займає О. Мороз, який досить високопозитивно корелює із другою групою, але водночас не є протиставленим першій групі (кореляція з першою групою не є негативною, вона просто відсутня).

У деяких випадках виникає потреба обчислити часткові коефіцієнти кореляції Пірсона. Наприклад, мова може йти про необхідність обчислення кореляції між освітою працівника та його заробітною платою, не враховуючи стаж роботи. Інколи знаходять часткові коефіцієнти кореляції для підтвердження (або спростування) кореляції між двома змінними, яку спостерігають із використанням процедури **Crosstabs**. Для цього вводять третю змінну, яка корелює з обома нашими змінними і, можливо, цим спричинює кореляцію між ними. Якщо у побудованій матриці коефіцієнти часткових кореляцій будуть значущі – спостережуваний зв'язок дійсно існує, якщо ж не значущі – то має місце так звана *фальшива кореляція* між двома змінними, спричинена впливом третьої.

У пакеті SPSS для обчислення матриці часткових коефіцієнтів кореляції потрібно послідовно обрати:

Analyze ⇒ Correlate ⇒ Partial

(*Аналіз ⇒ Кореляція ⇒ Часткова*)

У вікні параметрів, що відкривається (див. рис. 4.9), у полі **Variables:** потрібно сформувати список змінних, для яких необхідно побудувати квадратну матрицю коефіцієнтів кореляцій, а у полі **Controlling for:** потрібно сформувати список тих змінних, які при цьому потрібно статистично контролювати.

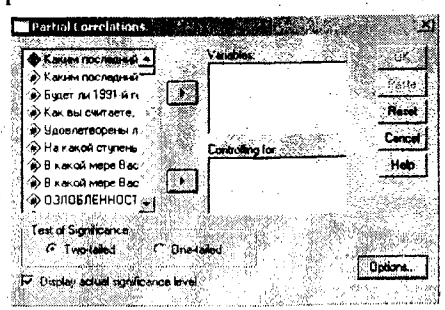


Рис. 4.9. *Вікно параметрів обчислення часткових кореляцій*

Результатом замовлення цієї процедури буде стандартна таблиця часткових кореляційних коефіцієнтів (*Partial correlation coefficients*). І якщо вказані там коефіцієнти є значущими, то її можна інтерпретувати аналогічно до коефіцієнтів парних кореляцій.

Зауваження. І матриця парних кореляцій, і матриця часткових кореляцій є квадратними матрицями, симетричними відносно головної діагоналі (оскільки всі коефіцієнти кореляції, для яких будуються матриці, є симетричними). На головній діагоналі матриці кореляцій стоять значення 1 (оскільки кожна змінна має повну кореляцію із самою собою), всі інші значення в матриці за абсолютним значенням є не більшими ніж значення 1.

Контрольні запитання

1. Назвіть основні структурні елементи двовимірної таблиці частот та відсотків.
2. Які клітинкові показники можна обчислювати за допомогою операції Crosstabs?
3. Якими є обмеження застосування критерію χ^2?
4. Перерахуйте коефіцієнти зв'язку між двома змінними, що їх може обчислювати процедура Crosstabs для порядкових та для кількісних змінних.
5. Чому коефіцієнт спряженості Пірсона не набуває значення 1?
6. Перерахуйте коефіцієнти зв'язку між двома змінними, що їх може обчислювати процедура Crosstabs для номінальних змінних.

Практичні завдання

1. Пояснити структуру таблиці для прикладу таблиці v1 та v167. *Зауваження.* Необхідно замовити відсотки як у рядку, так і у стовпчику.
2. Пояснити структуру двовимірної таблиці на прикладі таблиці для змінних v1 та v167 при контролюванні v165.
3. Побудувати таблицю для змінних v167 та v133 і на основі аналізу відсотків (порівняння розподілів у стовпчиках та/або рядках таблиці) зробити висновок про наявність впливу освіти (v167) на підтримку радикальних мір посилення трудової дисципліни в країні (v133).
4. Використовуючи змінні

v5 – "Чи задоволені Ви своїм становищем у суспільстві?",
v7 – "Чи задоволені Ви тим, що отримуєте від суспільства?",

v8 – "Чи задоволені Ви тим, що віддаєте суспільству?",
v165 – "Стать",
v166 – "Вік респондента",
v167 – "Освіта",
v168 – "Сімейний стан",

знайдіть відповіді на такі питання:

1. Який відсоток чоловіків взагалі задоволені своїм становищем у суспільстві?
2. Який відсоток серед жінок становлять розлучені?
3. Скільки чоловіків мають вищу освіту?
4. Який середній вік респондентів із середньою освітою?
5. Кого більше серед тих, хто задоволений одержаним від суспільства, чоловіків чи жінок?
6. Який середній вік розлучених?
4. Пояснити структуру таблиці групових середніх на прикладі таблиці для змінних v173 (залежна) та v167 (незалежна).

РОЗДІЛ 5. НАОЧНЕ ПРЕДСТАВЛЕННЯ РЕЗУЛЬТАТІВ. ГРАФІКИ

5.1. Побудова графіків

Графічне представлення даних дослідження та результатів дослідження може не тільки зробити більш привабливим текст публікації або ж презентацію на семінарі, а й часто допомагає безпосередньо у процесі аналізу. Пакет програм SPSS має розвинуті можливості побудови графіків різного типу та містить досить потужний графічний редактор, що дає змогу наочно подавати результати аналізу даних. Деякі операції, що ми їх розглядали у попередніх розділах, дозволяють будувати певні графіки. Так, наприклад, операція **Frequencies** дає змогу представити одновимірний розподіл частот та відсотків не тільки у вигляді таблиці, а й у вигляді графіка. Для цього потрібно послідовно обрати **Analyze** ⇒ **Descriptive Statistics** ⇒ **Frequencies**, потім натиснути кнопку **[Charts...]** (*Графіки*) та у вікні параметрів, що з'явиться (див. рис. 5.1), обрати один із трьох *типів графіка* (група **Charts Type**):

- **Bar charts** (*Стовпчикові діаграми*);
- **Pie charts** (*Кругові діаграми*);
- **Histograms** (*Гістограми*), можливо із додатково накладеною кривою нормального розподілу із відповідними параметрами **With normal curve** (*З нормальною кривою*).

Далі потрібно у групі **Chart Values** (*Представлені на графіку значення*) обрати, які саме показники, **Frequencies** (*Частоти*) чи **Percentages** (*Відсотки*), мають бути на графіку.

За загальною угодою операція **Frequencies** графіки не продукує (опція **None** у групі **Charts Type**).

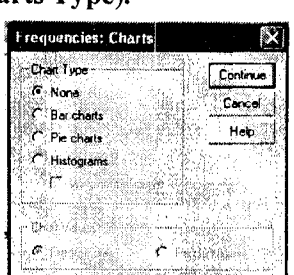


Рис. 5.1. *Вікно параметрів графіків операції* ***Frequencies.***

Операція побудови двовимірних таблиць **Crosstabs...** може також представити частотні розподіли у кожному рядку двовимірної таблиці у вигляді стовпчикової діаграми. Для цього потрібно у головному вікні параметрів операції **Crosstabs...** (див. рис. 4.1) поставити позначку біля **Display clustered bar charts** (*Виводити згруповані стовпчикові діаграми*).

Проте основні графічні можливості SPSS є доступними через пункт **Graphs** головного горизонтального меню (див. рис. 5.2). Зауважимо, що всі замовлені для побудови графіки виводяться так само, як і таблиці або окремі показники, у вікно результатів Output.

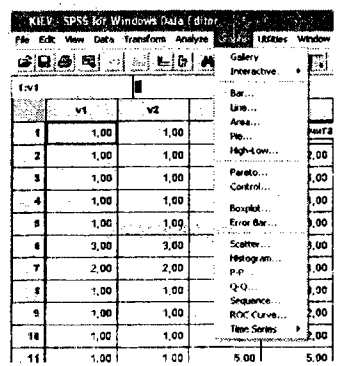


Рис. 5.2. *Пункт Graphs головного горизонтального меню.*

Пункт головного меню Graphs містить багато різновидів графіків. Зразки більшості з них є у пункті **Gallery** (*Галерея*). Не будемо розглядати всі доступні для використання види графіків. Розглянемо лише ті з них, що використовуються дослідниками найчастіше:

▪ **Bar...** (*Стовпчикові діаграми*). Такі графіки використовують, як правило, для:

- представлення одновимірного розподілу частот або відсотків (тоді обирають **Simple** (*Простий*) графік); зауважимо, що для номінальних змінних доцільно стовпчики переставляти так, щоб вони були впорядковані за зростанням або спаданням поданих на графіку значень;
- представлення середніх значень, сум або інших характеристик однієї або кількох кількісних змінних;
- у випадку представлення на одному графіку водночас кількох розподілів відповідні стовпчики можна виводити як **Clustered** (*Згруповані*) або ж як **Stacked** (*Розташовані один над іншим*).

Зауважимо, що стовпчикові діаграми доцільно використовувати для змінних із невеликою кількістю категорій.

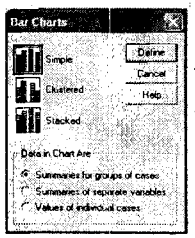


Рис. 5.3. *Вікно параметрів стовпчикових діаграм* ***Bar Charts.***

- **Line...** (*Лінія*). Відрізки прямої лінії з'єднують точки на площині, що відповідають поданим на графіку значенням. Такий графік часто називають "полігон". Як і під час побудови стовпчикових діаграм, значення номінальних змінних (коди категорій) доцільно переставляти так, щоб точки на графіку були впорядковані за зростанням або спаданням значень на графіку.

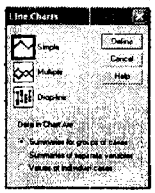


Рис. 5.4. *Вікно параметрів діаграм із ліній* ***Line Charts.***

Діаграми з ліній є більш зручними, порівняно зі стовпчиковими, у тих випадках, коли необхідно на графіку представити одну або кілька змінних із досить великою кількістю значень або ж коли ці значення є певними чином впорядковані (наприклад упорядковані у часі).

- **Area...** (*Область*). Такі діаграми схожі на діаграми з ліній, але розташовані під лініями області і зафарбовуються, що робить графік у деяких випадках більш наочним.

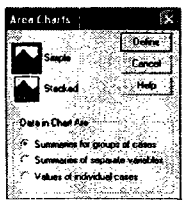


Рис. 5.5. *Вікно параметрів діаграм з областями* ***Area Charts.***

- **Pie...** (*Кругові діаграми*). Доцільно представляти у вигляді кругової діаграми відсотки одновимірного розподілу дискретної змінної. В такому випадку весь круг становить 100 %, а кожна категорія є сектором, розмір якого пропорційний відсотку (частці) цієї категорії у загальному розподілі. Зазначимо, що кругова діаграма є дійсно наочною тоді, коли кількість секторів (кількість категорій змінної) не є дуже великою (не перевищує 10).

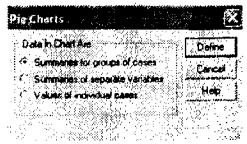


Рис. 5.6. *Вікно параметрів кругової діаграми* ***Pie Charts.***

Те, що саме відображати на графіку, визначається шляхом вибору в групі **Data in Chart Area** (*Дані, що зображені графічно*). Це може бути **Summaries for groups of cases** (*Узагальнююча інформація про групи спостережень*), **Summaries of separate variables** (*Узагальнююча інформація про окремі змінні*), **Values of individual cases** (*Значення для окремих спостережень*). У випадку, коли на полігоні або на діаграмі з областями представлені водночас кілька розподілів, бажано, якщо є така можливість, переставляти точки так, щоб лінії або не перетиналися, або мали лише одну точку перетину. В іншому випадку (якщо лінії на графіку перетинаються кілька разів) графік може виглядати як занадто заплутаний.

- **Pareto...** (*Діаграми Парето*).

Діаграма Парето – це стовпчикова діаграма, в якій стовпчики розміщуються в порядку спадання, із додатковою кривою для представлення кумулятивної (накопиченої) частоти для подання категорій. Звичайно, накопичення має бути таким, щоб результат мав певний сенс.

- **Histogram...** (*Гістограма*).

Гістограми використовують переважно для представлення одновимірних розподілів кількісних змінних. Під час побудови графіка діапазон можливих значень кількісної змінної розбивається на інтервали.

Вибір типу графіка, що виразно представляє певні особливості даних або певні тенденції у даних, є своєрідним мистецтвом. Невдало обраний тип графіка може не допомогти, а навпаки, зашкодити

презентації результатів аналізу. Крім того, важливим також є навіть і спосіб використання графіка. Зокрема, вказані за загальною угодою кольори різних елементів графіка та типи ліній є такими, що покращують зображення графіка на екрані комп'ютера. Проте, якщо графік потрібно надрукувати на чорно-білому принтері, то кольори подаються як певні відтінки сірого, і різні за кольором сегменти або лінії можуть стати такими, що не розрізняються. В такому випадку різні елементи графіка краще виділяти не кольором, а заповненням (наприклад різним штрихуванням), а різні лінії – за допомогою різноманітних типів ліній (наприклад жирна, пунктирна тощо).

Загальні угоди для вибору кольорів та заповнень можна змінити. Для цього потрібно послідовно обрати в головному горизонтальному меню програми **Edit** (*Редагувати*) ⇒ **Options** (*Параметри*) закладку **Charts** (*Графіки*) і в розділі **Fill Patterns and Line Styles** (*Зразки заповнень та стиль ліній*) замість опції **Cycle through colors, then patterns** (*Насамперед використовувати кольори, потім заповнення*) активізувати опцію **Cycle through patterns** (*Використовувати заповнення*). Тоді для виділення елементів графіка будуть використані передусім не різні кольори, а різні заповнення, що, наприклад, покращить представлення графіка у відтінках сірого кольору у надрукованому на звичайному лазерному принтері звіті.

Сформулюємо деякі, дуже загальні, рекомендації до процесу побудови графіків.

- Не використовувати в одній публікації або презентації велику кількість різних графіків, особливо якщо йдеться про однотипні дані.
- Назви та коментарі на графіку повинні бути чіткими, короткими, зрозумілими та мати однаковий стиль для всіх графіків у публікації або презентації.
- Потрібно чітко зрозуміти, що саме спочатку має порівнюватися на графіку (лінійні розміри зображень, площі окремих елементів, відмінності між розмірами та/або площами) і, відповідно, відтворити на графіку потрібні пропорції.
- Користуватися опорними лініями, що фіксують певні значення на осях графіка.
- Для кожної з осей графіка ретельно продумати шкалу та визначити масштаб представлення.
- Не перенавантажувати графік великою кількістю категорій.

- Обережно ставитися до введення ефекту "глибини" у представленні стовпчиків, секторів кругових діаграм, виділених на графіку областей тощо (так званих "3-D" графіків).
- Ретельно обирати кольори та заповнення для окремих елементів графіка. Пам'ятати про чорно-біле представлення графіка.

Приклад. Розглянемо послідовність побудови графіка на прикладі побудови простої стовпчикової діаграми для змінної v167 (освіта респондента).

Отже, обираємо послідовно в головному горизонтальному меню **Graphs** ⇒ **Bar...**; у вікні, що з'явиться (див. рис. 5.3), обираємо тип графіка Simple (*Простий*) та залишаємо встановлену за загальною угодою опцію **Summaries for groups of cases**, оскільки на графіку має бути зображений розподіл тільки однієї змінної. Кнопка **[Define]** (*Визначити*) розкриває вікно, в якому можна задати додаткові параметри нашого графіка (див. рис. 5.7). Це вікно параметрів виглядає приблизно однаково для всіх стандартних графіків.

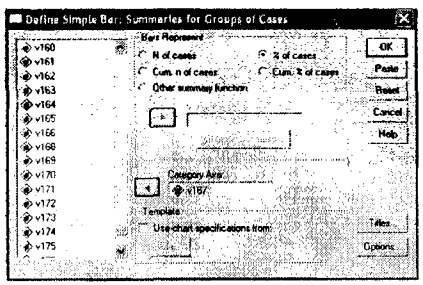


Рис. 5.7. *Вікно параметрів під час побудови простої стовпчикової діаграми для однієї змінної.*

У групі **Bars Represent** (*Стовпчики представляють*) нам потрібно обрати, що саме буде представлено висотою стовпчика на графіку:

- **N of cases** – частота, тобто кількість спостережень,
- **Cum. n of cases** – накопичена (кумулятивна) частота спостережень,
- **% of cases** – відсоток спостережень (ми оберемо саме цей варіант для нашого графіка),
- **Cum. % of cases** – кумулятивний відсоток спостережень,
- **Other summary function** – значення певної узагальнюючої функції (наприклад середнє арифметичне, сума тощо) для вказаної змінної. Ім'я змінної, до якої і має бути застосована узагальнююча

функція, вводиться у поле **Variable** (*Змінна*). За загальною угодою як функцію узагальнення використовують **MEAN** (*Середнє арифметичне*) (див. рис. 5.8). Для зміни функції потрібно натиснути кнопку **[Change Summary...]** (*Змінити функцію узагальнення*), яка відкриває діалогове вікно з переліком функцій[12].

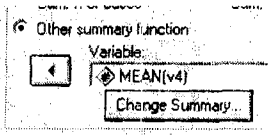


Рис. 5.8. *Фрагмент вікна, зображеного на рис. 5.7.*

У поле **Category Axis:** (*Вісь категорій*) потрібно внести ім'я змінної, для категорій якої має бути побудований графік. Ми заносимо в це поле змінну v167.

Використовуючи кнопку **[Title...]** (*Заголовки*), визначаємо для майбутнього графіка заголовок "Освіта" (див. рис. 5.9).

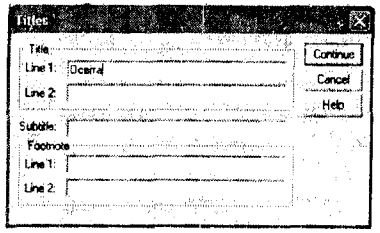


Рис. 5.9. *Вікно для введення заголовків графіка Titles...*

У вікні, зображеному на рис. 5.7, натискаємо кнопку **[Ok]** і запускаємо операцію побудови визначеного нами графіка. Побудований графік з'являється у вікні результатів **Output**.

У більшості випадків побудований графік потрібно редагувати – змінити кольори, написи, шрифти тощо. Для того, щоб графік перенести у редактор **Chart**, потрібно двічі клацнути лівою кнопкою миші у будь-якому місці графіка (див. рис. 5.10).

[12] Як узагальнюючі можна використовувати більшість зі згаданих раніше функцій. Зазначимо, однак, що обирати функцію можна лише для стовпчикової діаграми, полігона, кругової діаграми та діаграми з областями, причому не кожна із функцій, що їх пропонує програма, може бути використана для всіх видів графіків.

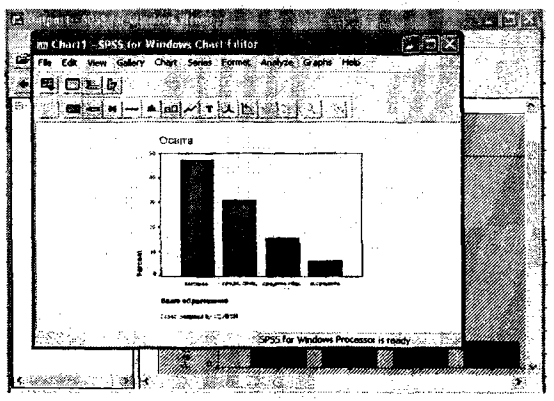


Рис. 5.10. *Редактор графічного редактора* ***Chart1*** *на фоні вікна результатів* ***Output1.***

У вікні параметрів графіка (див. рис. 5.7) за допомогою кнопки **[Titles...]** (*Заголовки*) можна змінити заголовок (назву) графіка, за допомогою кнопки **[Options...]** (*Параметри*) обрати спосіб роботи із відсутніми значеннями і в групі **Template** (*Шаблон*) обрати **Use charts specification from:** (*Використати параметри графіка з:*) і використати параметри з іншого графіка окремого файлу.

5.2. Основи редагування графіків

Розбиратися з усіма можливостями редагування графіків у SPSS потрібно в процесі реальної практичної роботи. Проте побудова графіків не є дуже складною, оскільки для більшості параметрів існують стандартні загальні угоди для значень.

Як уже зазначалося вище, побудовані графіки зазвичай потрібно редагувати – відкоригувати занадто довгу мітку змінної, зробити пояснення (зноску) до певного елементу тощо. Редагування здійснюється засобами графічного редактора **Chart**. Для запуску цього редактора потрібно двічі клацнути лівою кнопкою миші на будь-якій частині графіка у вікні результатів **Output**.

Вікно графічного редактора **Chart** (див. рис. 5.10) містить меню операцій та дві панелі інструментів. Якщо підвести курсор миші до кнопки на панелі інструментів, то можна отримати короткий опис цієї кнопки.

Розглянемо основні, найбільш часто вживані, операції редагування графіків, що їх можна знайти у пункті меню **Chart** (див. рис. 5.11).

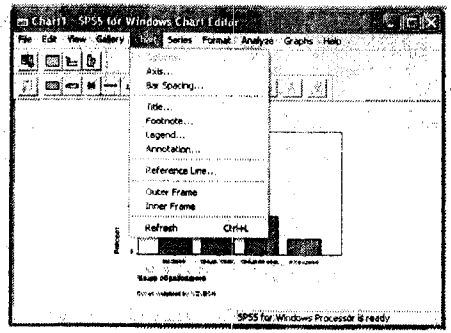


Рис. 5.11. *Меню* ***Chart.***

Зовнішній вигляд та структура вікон параметрів, що відкриваються у пунктах меню **Options**... (*Параметри*), **Axis**... (*Осі*) і **Bar Spacing**... (*Відстань між стовпчиками*), залежать від типу графіка, що редагується. Крім того, є можливість додати нову або відредагувати вже створену назву графіка (пункт **Title**... (*Заголовок*)), внести пояснення (пункт **Footnote**... (*Зноска*)), додати опис окремих категорій (пункт **Legend**... (*Напис*)), доповнити графік анотацією (пункт **Annotation**... (*Анотація*)), провести опорні лінії (пункт **Reference Line**... (*Опорна лінія*)), помістити графік у зовнішню (пункт Outer Frame (*Зовнішня рамка*)) та/або внутрішню (пункт **Inner Frame** (*Внутрішня рамка*)) рамки.

Меню **Format** (див. рис. 5.12) містить операції, більшість із яких виведена у вигляді кнопок на другу панель інструментів (див. рис. 5.13), що має назву панелі інструментів форматування та редагування.

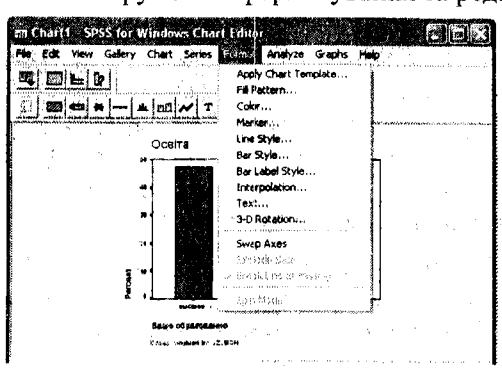


Рис. 5.12. *Меню* ***Format.***

Щоб відредагувати якусь частину графіка, потрібно:

– виділити той об'єкт, який має бути відредагованим (по краях об'єкта з'являються маркери корекції),

– натиснути потрібну кнопку чи обрати відповідний пункт меню,

– внести необхідні зміни у графік зміни та підтвердити їх кнопкою **[Apply]** (*Застосувати*).

Коротко розглянемо основні можливості операцій, зосереджених у вигляді кнопок на панелі форматування та редагування.

Рис. 5.13. *Панель форматування та редагування*

Point Id (*Виділення точок*)

Кнопка (або пункт меню) дає змогу змінювати режими відображення точок на діаграмі розсіювання.

Fill Pattern (*Зразки заповнень*)

Кнопка (або пункт меню) дає змогу обрати необхідний із восьми зразків для заповнення певних частин графіка: стовпчиків, областей під лініями тощо.

Color (*Колір*)

Кнопка (або пункт меню), відкриваючи палітру з кольорами (їх шістнадцять за загальною угодою), дає змогу змінити колір частини графіка або текстового напису. Опції **Fill** (*Заповнення*) та **Border** (*Границя*) дозволяють змінити колір як об'єкта, так і його контуру.

Marker (*Маркер*)

Кнопка відкриває палітру з 28 різних маркерів для позначення положення точки даних на діаграмах з ліній, діаграмах з областями і діаграмах розсіювання. Можна також встановити один із чотирьох передбачених розмірів маркерів.

У групі **Style** (*Стиль*) обирають необхідне маркування, а в групі **Size** (*Розмір*) – розмір маркерів[13]. Є можливість визначати різні стилі та розмір для окремих груп маркерів на графіку. Внесені зміни потрібно підтверджувати натисканням на кнопку **[Apply]**.

[13] На екрані відмінності у розмірах маркерів виглядають незначними, але після друку ця різниця може бути добре помітною.

Маркери на діаграмах з областями стають видимими лише у випадку, якщо їх виведення буде задане в діалоговому вікні **Interpolation** (*Інтерполяція*) меню **Format** (див. далі).

Зауваження. Маркери не можуть бути задані для зображення точок гістограм і стовпчикових діаграм.

Line Style (*Стиль лінії*)

Кнопка дає змогу обрати в групі **Style** (*Стиль*) один із чотирьох типів лінії, а в групі **Weight** (*Товщина*) – один із чотирьох можливих варіантів товщини лінії.

Bar Style (*Стиль стовпчиків*)

Кнопка дозволяє задати тип для графіків, що містять стовпчики: **Drop shadow** (*Із тінню*) чи **3D-effect** (*Із ефектом глибини, третьої розмірності*), у пункті **Depth** (*Глибина*) встановити глибину стовпчика. Глибина вказується у відсотках стосовно ширини стовпчика, знак вказує напрям ефекту тіні: знак "+" – тінь праворуч від стовпчика, знак "-" – тінь ліворуч від стовпчика.

Bar Label Style (*Стиль міток стовпчиків*)

Програма пропонує три стилі оформлення написів, що характеризують висоту кожного стовпчика. Зауважимо, що під час застосування темних кольорів та заповнень для написів числових значень краще обирати опцію **Framed** (*В рамці*), вони краще читатимуться.

Interpolation (*Інтерполяція*)

Кнопка для різного типу діаграм із ліній дає змогу визначити способи сполучення точок даних. Якщо ж додатково обрати опцію **Display markers** (*Показати маркери*), то для кожної точки виділеної кривої буде відображене маркування, обране за допомогою опції **Marker** меню **Format** або відповідної кнопки на панелі інструментів.

У групі **Line Interpolation** (*Вид інтерполяційної лінії*) обирають один із методів сполучення (інтерполяції) точок за допомогою деякої кривої.

Text (*Текст*)

Для текстових елементів графіка ця кнопка дає змогу: в групі **Font** (*Шрифт*) змінити тип шрифта, а в групі **Size** (*Розмір*) – його розмір.

3D-Rotation (*3D-обертання*)

Кнопка дозволяє обрати та застосовувати один із двох методів обертання тривимірних діаграм розсіювання. За допомогою перемикачів на лівій стороні діалогового вікна діаграму можна обертати вперед чи назад відносно осей X, Y і Z.

Обертається виділена діаграма натисканням на кнопку **[Apply]**. Під час обертання застосування будь-яких інших операцій стає неможливим.

Swap Axes (*Зміна осей*)

Кнопка дозволяє поміняти місцями вертикальну та горизонтальну осі плоского (двовимірного) графіка.

Explode Slice (*Виділити сегмент*)

Кнопка дозволяє виділити сегмент кругової діаграми.

Break Lines at Missing (*Розірвати лінію в місці відсутнього значення*).

Set/exit spin mode (*Включити/виключити режим обертання*).

5.3. Редагування мобільних таблиць

Результати обчислень для всіх операцій відображаються у вікні перегляду результатів **Output** у вигляді блоку, кожен елемент якого є окремим об'єктом. Більшість елементів результатів розрахунків представлені у вигляді так званих мобільних таблиць. Отримані таблиці можна відразу вставляти у звіт, але часто їх попередньо потрібно відредагувати – додати заголовок, внести зміни до текстових тощо.

Процес редагування таблиць у SPSS схожий на відповідний процес у відомому текстовому процесорі MS Word. Для внесення змін до зовнішнього вигляду таблиць та їх редагування в SPSS існують такі можливості:

- вибір зовнішнього вигляду таблиці з бібліотеки таблиць,
- зміна формату всієї таблиці,
- зміна формату клітин таблиці,
- зміна текстових написів у таблиці,
- доповнення таблиці поясненнями,
- доповнення таблиці зносками,
- введення назви об'єкта та додаткового тексту.

Щоб відредагувати таблицю, потрібно перенести її в редактор мобільних таблиць, тобто двічі клацнути на потрібній таблиці.

Вибір зовнішнього вигляду таблиці. Щоб обрати інший зовнішній вигляд таблиці, потрібно послідовно вибрати в головному меню

Format (Формат) ⇒ **Table Looks...** (Зовнішній вигляд таблиць)

З'явиться діалогове вікно **Table Looks...**, в якому можна здійснити вибір серед більш ніж п'ятдесяти різних варіантів зовнішнього вигляду (дизайну) таблиць. Якщо обрати певний варіант (наприклад *Avant-garde*) та натиснути кнопку **[Ok]**, то обраний дизайн буде застосований до таблиці, що редагується.

Кнопка **[Edit Look]** (*Редагувати дизайн*) відкриває допоміжне діалогове вікно **Table Properties** (*Властивості таблиці*), в якому можна додатково змінити окремі елементи компонування таблиці. Відредагований дизайн можна зберегти за допомогою команд **Save Look** (Зберегти дизайн) та **Save as...** (Зберегти як).

Зміна властивостей таблиці відбувається за допомогою операцій меню

Format (*Формат*) ⇒ **Table Properties** (*Властивості таблиці*).

У вікні параметрів, що розкривається, є закладки, в яких можна змінити представлення деяких даних, зноски, формати клітин і різновиди рамок.

Для окремих областей таблиці можна змінити шрифт, застосовуючи послідовність

Format (*Формат*) ⇒ **Font...** (*Шрифт*).

Встановити однакову ширину для всіх клітин таблиці можна, застосовуючи послідовність

Format (*Формат*) ⇒ **Set Data Cell Widths...**
(*Встановити ширину клітин даних*).

Зміна властивостей клітин таблиці. Можна змінювати не тільки дизайн усієї таблиці, а й формат окремих її клітин. Для цього потрібно виділити необхідну клітину (клацнути на ній лівою кнопкою миші) і послідовно обрати в меню

Format (*Формат*) ⇒ **Cell Properties** (*Властивості клітини*).

У діалоговому вікні **Cell Properties** є чотири вкладки, які дають змогу обрати необхідний формат чисел, спосіб вирівнювання даних у клітинці, поля і відтінок. У полі **Sample** (*Зразок*) наводиться зразок напису з урахуванням відповідних змін.

Для того, щоб **змінити текст у таблиці,** необхідно потрібний текст виділити (двічі клацнути правою кнопкою миші на клітині з текстом) та відредагувати. Після завершення редагування потрібно натиснути клавішу *Enter*. Так можна відредагувати текст у будь-якій частині таблиці.

Для того, щоб **доповнити таблицю поясненнями,** необхідно послідовно обрати в меню

Insert (*Вставка*) ⇒ **Captions** (*Підпис*).

Під таблицею з'явиться рамка з текстом **Table Captions** (*Підпис таблиці*) всередині. Потрібно двічі клацнути правою кнопкою миші на цьому тексті, з'явиться курсор, і після цього можна ввести необхідний текст пояснення.

Доповнення таблиці зносками. Будь-який текст у таблиці можна доповнити зносками. Для цього потрібно виділити обраний текст (двічі клацнути на ньому правою кнопкою миші) і послідовно обрати в меню редактора мобільних таблиць

Insert (*Вставка*) ⇒ **Footnote** (*Зноска*).

У рамці, що з'явилася, потрібно двічі клацнути на слові **Footnote** правою кнопкою миші, з'явиться курсор, і після цього можна ввести необхідний текст зноски.

Зноски з'являються під таблицею та нумеруються маркерами у вигляді літер зменшеного розміру (для першої зноски використовується літера *a*). Щоб змінити маркер, виділіть зноску і оберіть в меню

Insert ⇒ **Footnote Marker...** (*Маркер зноски*).

Оберіть **Special marker** (*Особливий маркер*) і введіть, наприклад, цифру 1.

Введення назви об'єкта і додаткового тексту. Щоб додати назву або деякий текст, виділіть відповідний об'єкт (заголовок, таблицю, графік тощо), після якого потрібно ввести підзаголовок або текст, потім оберіть у меню

Insert (*Вставка*) ⇒ **New Title** (*Нова назва*)

або, відповідно,

Insert (Вставка) ⇒ **New Text** (*Новий текст*).

Після цього двічі клацнути на новому об'єкті та ввести необхідну назву або текст.

Якщо потрібний текст міститься у текстовому файлі, то потрібно обрати в меню

Insert (*Вставка*) ⇒ **Text File...** (*Текстовий файл*).

У діалоговому вікні, що з'явиться, потрібно вказати ім'я потрібного файлу.

Зауваження. Всі зміни, що стосуються зовнішнього вигляду таблиць, виконуються лише в редакторі мобільних таблиць. Для цього потрібно двічі клацнути на потрібній таблиці. Інакше зробити зміни неможливо, бо тільки при активованому редакторі мобільних таблиць змінюється головне меню.

Примітка. Будь-яку мобільну таблицю можна скопіювати в іншу програму (Word, Excel), для цього потрібно встановити курсор миші на таблицю, клацнути лівою кнопкою миші та у контекстному меню, що з'явилося, обрати **Copy**. Таблиця копіюється у буфер (*clipboard*) Windows, звідки її завжди можна дістати та вставити у середовищі будь-якої програми.

Зокрема, так таблицю можна вставити у звіт, текст якого підготовлений за допомогою MS Word, або ж побудувати на даних тавлиці графік у середовищі електронних таблиць MS Excel.

Контрольні запитання

1. Як можна відредагувати таблицю або діаграму у вікні перегляду результатів?
2. Назвіть основні типи діаграм.
3. Чи можна за даними таблиці з SPSS побудувати діаграму в програмі Excel?
4. Чи є різниця між графічними редакторами статистичного пакету SPSS та програми Excel?
5. Як Ви думаєте, чому таблиці в SPSS мають назву мобільних?
6. Чи обов'язковою є процедура редагування мобільних таблиць?
7. Які можливості імпорту/експорту таблиць та графіків у середовищі SPSS?

Практичні завдання

1. Побудуйте графік для змінної v167 (рівень освіти респондента), як показано у п.5.1. Відредагуйте заголовок. Задайте назви категорій. Зробіть на графіку посилання на дослідження, дані якого використано. Відформатуйте графік так, щоб його можна було роздрукувати.
2. Представте графічно результати виконання завдання 3 (Розділ 4) у програмі Excel, попередньо виконавши експорт мобільних таблиць зі статистичного пакету SPSS у табличний процесор Excel.
3. Відредагуйте одну з таблиць (завдання 3 з Розділу 4) так, щоб її можна було вставити як об'єкт у документ звіту?

РОЗДІЛ 6. ВІДБІР ОБ'ЄКТІВ ДЛЯ АНАЛІЗУ (ПОБУДОВА ФІЛЬТРІВ)

Досить часто із загального масиву даних потрібно відібрати для аналізу певну групу об'єктів (спостережень). Наприклад, із масиву даних загальноукраїнського опитування потрібно відібрати респондентів із Західного регіону, щоб з'ясувати саме для них зв'язок між матеріальним станом та зовнішньополітичними орієнтаціями. Для реалізації такого відбору потрібно сформулювати певне правило, керуючись яким комп'ютерна програма в процесі побудови таблиць, побудови графіків, обчислень показників буде вирішувати для кожного спостереження, включати чи не включати це спостереження до аналізу. В SPSS таке правило відбору називають *фільтром*. У процесі роботи із даними дослідник може визначати та встановлювати різні, потрібні йому для аналізу, фільтри. Якщо фільтр встановлений, то робота виконується не з усім масивом даних, а лише з тією частиною, що пройшла крізь фільтр (так би мовити відфільтрованими спостереженнями).

Для здійснення відбору спостережень у пакеті SPSS послідовно обираємо

Data ⇒ Select Cases...

(*Дані*) ⇒ (*Відібрати спостереження*).

Розкривається вікно параметрів

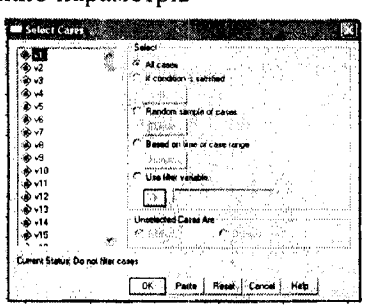


Рис. 6.1. *Вікно параметрів операції* ***Select Cases.***

У групі **Select** є можливість обрати один із п'яти можливих варіантів роботи із файлом даних:

- **All cases** – аналізувати всі спостереження;
- **If condition is satisfied** – відібрати для аналізу спостереження, що задовольняють певну логічну умову (див. пункт 6.1);

- **Random sample of cases** – відібрати для аналізу певну частину спостережень випадково, застосовуючи датчик рівномірно розподілених випадкових чисел (див. пункт 6.2);
- **Based on time or case range** – відібрати для аналізу спостереження на основі номерів рядків у редакторі даних або на основі часу здійснення спостереження (якщо у спостереженнях вказані дата та час і йдеться про часовий ряд). При цьому у вікні **Range...** (*Діапазон*) вказують номер першого (*First Case*) і останнього (*Last Case*) спостереження; всі спостереження, що не ввійшли у вказаний діапазон, будуть відфільтровані[14];
- **Use filter variable** – відібрати для аналізу спостереження за допомогою заздалегідь обчисленої фільтруючої змінної (до аналізу включаються лише ті спостереження, для яких значення цієї змінної відмінне від 0 та не є відсутнім значенням, див. пункт 6.3).

Під вікном списку змінних міститься інформація про поточний стан (**Current Status**) файлу даних – застосовується (**Filter Cases by values of ...**) чи не застосовується (**Do not filter cases**) до файлу фільтр. Якщо до файлу застосований фільтр, то на нижній рамці вікна (у рядку статусу) з'явиться помітка **Filter On**.

У групі **Unselected Cases Are** (*Не відібрані спостереження*) можна обрати один із двох можливих варіантів роботи із тими спостереженнями, які не пройшли через фільтр та не відібрані для аналізу:

- **Filtered** – невідібрані спостереження залишаються у робочому файлі, але не аналізуються. У вікні редактора даних номери рядків для таких спостережень перекреслюються. Якщо підключити до роботи з файлом новий фільтр, то цей фільтр буде застосований для всього робочого файлу. Якщо обрати в групі **Select** варіант **All Cases**, то далі буде здійснюватися аналіз усього робочого файлу;

- **Deleted** – невідібрані спостереження видаляються з робочого файлу. Кількість спостережень у робочому файлі змінюється. Повернути в подальшому до робочого файлу видалені таким способом спостереження не можна. Якщо підключити до роботи з файлом новий фільтр, то цей фільтр буде застосований до вже зміненого

[14] Такий відбір, по-перше, не дуже часто застосовують, а по-друге, він є доволі нескладним, а тому не виноситься для розгляду окремим пунктом.

робочого файлу. Якщо обрати в групі **Select** варіант **All Cases**, то далі буде здійснюватися аналіз не всього, а вже зміненого робочого файлу. Перш ніж скористатися таким відбором, варто зберегти копію файлу даних. Якщо невідібрані спостереження видалені з файлу, то позначка **Filter On** у рядку статусу не з'являється.

Щоб **повернутися до аналізу всіх спостережень у робочому файлі даних**, тобто зняти фільтр, потрібно в групі **Select** обрати варіант **All cases**.

Розглянемо три основні способи відбору спостережень більш детально.

6.1. Відбір за умовою. Правила запису логічних умов

У процесі аналізу дуже часто виникає потреба відбирати спостереження за умовою. Наприклад, за результатами загальноукраїнського опитування потрібно подивитися середню заробітну плату не для всіх респондентів, а лише для жінок. Або ж потрібно подивитися розподіл довіри до уряду країни не для всіх респондентів, а лише для молодих людей (у віці до 35 років) зі Східного регіону країни.

Для того, щоб відібрати спостереження для аналізу, потрібно сформувати умову відбору, тобто записати логічний вираз, який набуває значення "істина" для тих спостережень, які мають бути включені до аналізу, і значення "хиба" для всіх інших змінних. Отже, якщо у наших даних змінна v165 містить інформацію про стать респондента і жінки закодовані числом 2, то логічний вираз (v165=2) набуває значення "істина" тільки для анкет жінок, а отже, фільтр із таким логічним виразом буде відбирати з файлу даних для аналізу лише анкети жінок.

Для того, щоб здійснити відбір спостережень за умовою, необхідно в групі **Select** (*Відбір*) обрати **If condition is satisfied** (*Якщо задовольняється умова*) та натиснути кнопку **[If ...]**. Розкривається вікно введення логічної умови (див. рис. 6.2).

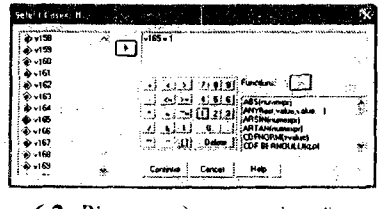


Рис. 6.2. *Вікно введення логічної умови.*

У верхній правій частині цього вікна розміщено поле для введення логічного виразу. Логічний вираз можна ввести як з клавіатури, так і за допомогою миші з використанням списку змінних (у вікні зліва), кнопок арифметичних операцій і чисел (блок, схожий на калькулятор, є у центральній частині вікна) і списку функцій (у вікні правіше кнопок).

Після введення логічної умови потрібно натиснути кнопку **[Continue]**. Комп'ютерна програма перевіряє синтаксичну правильність запису логічної умови і у випадку, якщо немає помилок у запису, закриває вікно введення логічної умови. Якщо після цього у вікні параметрів операції **Select Cases** натиснути кнопку **[Ok]**, то відповідний фільтр буде підключений. Комп'ютер перевіряє вказану умову для всіх спостережень і результати перевірки записує у змінну зі стандартним іменем **filter_$** (не потрібно використовувати таке ім'я для інших змінних): ця змінна набуває значення 0 (мітка значення **Not Selected**) для тих спостережень, що не були відібрані, та значення 1 (мітка значення **Selected**) для відібраних спостережень. Як мітка змінної використовується запис відповідного логічного виразу. Надалі аналізуються лише ті спостереження, для яких логічна умова набуває значення "істина" (тобто для тих, для яких filter_$=1). Фільтр діє доти, доки не буде знову виконаний оператор **Select Cases** із вибором варіанту **All cases**.

Під час створення нового фільтру значення змінної **filter_$** будуть змінені відповідно до нової умови, тобто кожен наступний результат відбору даних відобразиться в тій же змінній **filter_$**. Але якщо ж змінити ім'я цієї змінної (замість **filter_$** вказати якесь інше ім'я, як це зробити – див. розділ 2), то потім, навіть наступного дня, цю змінну можна буде використати як фільтруючу (обрати **Use filter variable** та вказати ім'я цієї змінної).

Правила запису логічних виразів у SPSS схожі на правила запису логічних виразів у поширених мовах програмування (Basic, Pascal тощо). Логічні вирази будуються з елементарних арифметичних відношень ("рівності" =; "нерівності" <> або ~=; "менше" <; "більше" >; "менше або дорівнює" <=; "більше або дорівнює" >=), предикатів (функцій, що набувають значення 0 – "хиба" або 1 – "істина") та логічних операцій ("заперечення" not або ~; "або" or або |; "та" and або &).

Правила обчислення логічних виразів такі: спочатку обчислюються арифметичні вирази (що є частиною арифметичних відношень), потім арифметичні відношення, а вже потім логічні операції (спочатку "заперечення", потім "та", потім "або"). Операції з однаковим пріоритетом обчислюються зліва-направо. Для зміни порядку обчислень використовують круглі дужки.

Арифметичні вирази, як і в мовах програмування, будуються з числових констант, імен змінних, знаків арифметичних операцій ("додавання" +, "віднімання" -, "множення" *, "ділення" /, "піднесення до степеня" **), арифметичних функцій та круглих дужок. Порядок обчислення арифметичних виразів є звичним: спочатку обчислюються арифметичні функції, потім піднесення до степеня, потім множення та ділення, потім додавання та віднімання. Операції з однаковим пріоритетом обчислюються зліва направо. Для зміни порядку обчислень використовують круглі дужки.

Кількість арифметичних функцій, які можна використовувати в запису виразів, є значною. Ми зазначимо лише ті, що найбільш часто зустрічаються.

Таблиця 6.1

Деякі арифметичні функції

ABS(вираз)	модуль (абсолютне значення)
EXP(вираз)	експонента
LG10(вираз)	логарифм з основою 10
LN(вираз)	натуральний логарифм
MAX(знач1, знач2, ...)	найбільше з послідовності значень
MIN(знач1, знач2, ...)	найменше з послідовності значень
SUM(вираз1, вираз2, ...)	сума значень виразів (тих, що мають визначені, відмінні від відсутніх, значення)
RND(вираз)	округлення значення виразу
TRUNC(вираз)	відкидання дробової частини виразу
VALUE(змінна)	значення змінної, включаючи числове значення описаного відсутнього значення (user missing value)

Крім того, є можливість використовувати досить значну кількість логічних предикатів. Ми зазначимо лише ті, що найчастіше зустрічаються.

Таблиця 6.2

Деякі логічні функції

ANY(вираз, знач1, знач2, ...)	"істина", якщо значення виразу дорівнює хоча б одному значенню зі списку
RANGE(вираз, н1, в1, н2, в2, ...)	"істина", якщо значення виразу належить принаймні одному з інтервалів ([н1,в1], або [н2,в2] тощо)
MISSING(змінна)	"істина", якщо значенням змінної є системне (system missing) або описане (user missing) відсутнє значення
SYSMIS(змінна)	"істина", якщо значенням змінної є системне (system missing) відсутнє значення

Приклади

1. Побудуємо логічний вираз, що відбирає в нашому файлі даних анкети жінок української національності. Стать респондента закодована у змінній v165, жінки – значенням 2, національність респондента – у змінній v170, українці значенням 1. Отже, нам потрібно відібрати для аналізу спостереження, в яких v165=2 і в той же час v170=1. Логічний вираз для відповідного фільтру має такий вигляд:

(v165=2) and (v170=1).

2. Побудуємо логічний вираз для фільтру, що відбирає для подальшого аналізу анкети чоловіків (код 1 змінної v165) із середньою (код 3 змінної v167) та середньою спеціальною (код 2 змінної v167) освітою. Фільтр має відбирати тих чоловіків, що мають або середню, або середню спеціальну освіту. Відповідний логічний вираз набуде такого вигляду:

(v165=1) and ((v167=2) or (v167=3)).

Зверніть увагу, що ми додали додаткові дужки, для того, щоб змінити порядок обчислення виразу. Якби ми записали цей вираз без дужок

(v165=1) and (v167=2) or (v167=3),

то б отримали фільтр, що відбирає чоловіків із середньою спеціальною освітою та всіх респондентів (і чоловіків, і жінок), що мають середню освіту.

6.2. Випадковий відбір

Необхідність відібрати випадкову частину загальної кількості спостережень часто пов'язана або із моделюванням, або із вивченням стійкості певних показників. Пакет SPSS використовує для такого відбору рівномірно розподілений датчик випадкових чисел, а тому кожне зі спостережень у файлі даних має однакову ймовірність потрапити до вибірки. Так здійснюється імітація одноступеневої випадкової вибірки заданого обсягу із генеральної сукупності, якою є всі спостереження, що містяться у файлі даних.

Для того, щоб здійснити випадковий відбір, необхідно в групі **Select** обрати **Random sample of cases** (*Випадкова вибірка спостережень*) і натиснути кнопку **[Sample...]**. Розкривається вікно параметрів випадкового відбору (див. рис. 6.3).

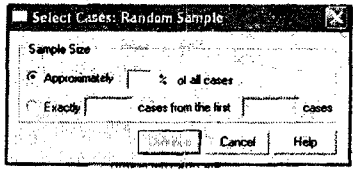


Рис. 6.3. *Вікно параметрів випадкового відбору.*

Є дві можливості здійснити випадковий відбір:

- **Approximately *p*% of all cases** – відібрати приблизно *p* відсотків спостережень з усього файлу даних (приблизно, оскільки відбирається ціла кількість спостережень і відповідний відсоток може дещо відрізнятися від вказаного);
- **Exactly *k* cases from the first *n* cases** – відібрати *k* спостережень із перших *n* спостережень файлу даних.

Необхідно пам'ятати, що повторне застосування випадкового відбору дає інший, порівняно із першим застосуванням, результат.

6.3. Використання фільтруючих змінних

Для того, щоб скористатися таким способом відбору, потрібно заздалегідь підготувати змінну, яка і буде використовуватися для відбирання (фільтруюча змінна). Всі спостереження, для яких ця змінна набуває значення 0 або відсутнє значення, "відфільтровуються" (тобто виключаються з аналізу). Всі інші спостереження вважаються такими, що пройшли фільтр, а тому беруть участь у

подальшому аналізі. Фільтруюча змінна може бути підготовлена різними засобами. Один із способів створення фільтруючої змінної полягає в тому, щоб змінити стандартне ім'я filter_$ змінної, що створюється у процесі відбору спостережень за умовою (див. 6.1). Так, наприклад, якщо після застосування відбору за логічним виразом (v165=2) and (v170=1) змінити стандартне ім'я змінної filter_$ на WOMANUKR, то пізніше за необхідності відібрати для аналізу анкети жінок української національності можна не записувати знову відповідний логічний вираз, а просто підключити змінну WOMANUKR як фільтруючу зміну. Інакше кажучи, можна створити у файлі даних кілька різних фільтруючих змінних, які відбирають для аналізу потрібні частини загальної сукупності спостережень, і при необхідності просто підключати ці змінні як фільтруючі.

Для того, щоб підключити фільтруючу змінну, потрібно в групі **Select** обрати **Use filter variable** та перенести у відповідне поле ім'я тієї змінної, яка має бути використана для відбору (див. рис. 6.1). Нагадаємо, що ті спостереження, для яких ця змінна набуває значення 0, будуть виключені з аналізу (номери рядків перекреслені), доки діє фільтр (є позначка **Filter On** у рядку статусу).

Контрольні запитання

1. Які є способи відбору спостережень для аналізу у процедурі **Select Cases**?
2. В якому порядку застосовуються арифметичні відношення, арифметичні та логічні операції при обчисленні логічних виразів?
3. Як зберегти для повторного застосування фільтр, що здійснює відбір спостережень за умовою?
4. Для написання курсової роботи на основі аналізу результатів загальноукраїнського опитування потрібна лише частина цих даних, що стосується Західного регіону. Як можна отримати відповідний файл даних, що не містить непотрібних для аналізу спостережень?

Практичні завдання

1. Побудувати таблицю одновимірного розподілу за сімейним станом (змінна v168) для чоловіків (змінна v165) у віці (змінна v166) від 32 до 60 років.
2. Який відсоток задоволених своїм становищем у суспільстві (змінна v5) серед тих респондентів у віці від 35 до 50 років, які задоволені тим, що вони отримують від суспільства (змінна v7).

3. Вважаючи, що у змінній v5 задоволеність респондентів своїм становищем у суспільстві виміряна в балах, порівняйте середній бал задоволеності (змінна v5) чоловіків та жінок із вищою освітою (змінна v167).

4. Побудуйте двовимірну таблицю "стать" на "сімейний стан" для випадково відібраної половини від загальної кількості респондентів у віці від 25 до 50 років з освітою, не вище від середньої. Виконайте цю операцію (включаючи і відбирання спостережень, і побудову таблиці) двічі і порівняйте результати.

РОЗДІЛ 7. ОБЧИСЛЕННЯ НОВИХ ЗМІННИХ У ФАЙЛІ ДАНИХ

У процесі аналізу часто виникає необхідність різних перетворень даних. Зокрема, може йтися про:

- побудову деякого інтегрального показника на основі групи змінних. Прикладом є середній бал студента за сесію, який можна обчислити як середнє арифметичне отриманих студентом оцінок на іспитах;
- перехід до інших одиниць вимірювання. Наприклад, за необхідності порівняння із даними досліджень у європейських країнах може виникнути потреба представити місячний дохід у сім'ї на одну особу не у гривнях, а у євро;
- підрахунок частоти вибору конкретних значень у групі змінних. Наприклад, можна підрахувати кількість відмінних оцінок, отриманих студентом за 4 роки, і розглядати цю кількість як певний показник, що характеризує старанність студента протягом усього терміну навчання;
- перехід до аналогічного за змістом, але більш зручного для конкретного аналізу показника. Наприклад, перехід від року народження до віку респондента;
- розбити діапазон значень неперервної змінної на певні інтервали. Наприклад, розбити діапазон значень змінної "Вік респондента" на три інтервали – молодь (до 35 років включно), люди середнього віку (від 36 до 55 років включно) та літні люди (56 років та старші);
- здійснити певне потрібне для подальшого аналізу групування категорій дискретної змінної. Наприклад, об'єднати кілька професій у категорію "робочі професії".

Пакет SPSS має різні інструменти для обчислення нових змінних. Зокрема, ми більш детально розглянемо операції **Compute** (*Обчислити*), **Recode** (*Перекодувати*) та **Count** (*Підрахувати*) у застосуванні вирішення перерахованих нами задач. Зазначимо також, що всі перетворення за допомогою цих операцій можна виконувати як для усього файлу даних, так і для певної його частини, тобто для відібраної за певною логічною умовою категорії респондентів. Для того, щоб відібрати спостереження для подальших перетворень, потрібно скористатися кнопкою **[If...]** (*Якщо*), яка є стандартною для всіх трьох операцій. Правила запису логічних виразів (умов) для відбору об'єктів ми вже розглядали (див. пункт 6.1).

Зауваження. Всі перетворення (створення нових змінних чи зміна значень існуючих змінних) виконуються в робочому файлі даних. Для того, щоб *зберегти внесені у файл зміни* для подальшої роботи, необхідно записати змінений робочий файл на диск.

7.1. Створення нових змінних шляхом арифметичних обчислень

Для створення нових змінних шляхом застосування арифметичних операцій до вже існуючих у файлі змінних у пакеті SPSS потрібно застосувати операцію **Compute**. Ця операція дозволяє або створити в робочому файлі даних нову змінну, або змінити значення вже існуючої у файлі даних змінної за певною математичною формулою. Для того, щоб скористатися цією операцією, необхідно послідовно обрати

Transform ⇒ Compute Variable

(*Перетворення ⇒ Обчислити змінну*)

Розкривається вікно параметрів операції (див. рис 7.1).

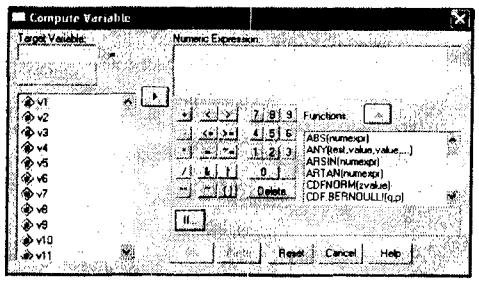


Рис. 7.1. *Вікно параметрів операції* ***Compute Variable.***

У поле **Target Variable** (*Цільова змінна*) необхідно ввести ім'я змінної:

- якщо вводиться ім'я існуючої змінної, то будуть змінені значення існуючої змінної;
- якщо вводиться нове ім'я, то буде створена в робочому файлі нова змінна під введеним іменем.

Після введення імені активується кнопка **[Type&Label....]**, яка дає змогу вказати параметри нової обчислюваної змінної. Зокрема, є можливість ввести як мітку змінної певний текст (обрати варіант **Label** та ввести текст мітки змінної) або використати як мітку змінної

ту формулу, за якою будуть виконуватися обчислення (обрати варіант **Use expression as label**).

У поле **Numeric expression** потрібно ввести формулу (арифметичний вираз), за якою має обчислюватися нова змінна. Під час введення формули можна використовувати список змінних, список предикатів і функцій, а також спеціальну намальовану на екрані клавіатуру, що зовні нагадує "калькулятор" (з такою клавіатурою ми вже мали справу, коли розглядали введення умов під час створення фільтрів у розділі 6).

Якщо обчислення потрібно виконати не для всіх спостережень, а лише для тих, що задовольняють певну умову, потрібно натиснути кнопку **[If...]** (*Якщо*), обрати варіант **Include if case satisfies condition** (*Включати, якщо спостереження задовольняє умову*), ввести умову (використовуючи список змінних, список предикатів та функцій, клавіатуру "калькулятора") та натиснути кнопку **[Continue]**. Після завершення введення параметрів для виконання операції **Compute** потрібно натиснути кнопку **[Ok]**.

Під час створення у робочому файлі даних нової змінної (або зміни значень існуючої змінної) операція **Compute** перевіряє для кожного спостереження вказану логічну умову. Для тих спостережень, для яких умова є істинною, значення виразу обчислюється та записується у вказану змінну. Для тих спостережень, для яких умова є хибною, або залишається попереднє значення змінної (якщо змінюються значення існуючої змінної), або записується системне відсутнє значення (*system missing value*).

Операція **Compute** застосовується для числових або для коротких рядкових змінних. Тип результату обчислення виразу має бути узгодженим із типом змінної. Файл даних, з яким ми працюємо, містить виключно числові змінні.

Якщо для обчислення виразу або певної його частини не достатньо інформації (деякі необхідні для обчислень змінні мають відсутнє значення), то результатом є системне відсутнє значення. Виняток складають деякі випадки, перераховані нижче:

0 * відсутнє = 0	0 ** відсутнє = 0
0 / відсутнє = 0	відсутнє ** 0 = 1

Щоб описані відсутні значення можна було використовувати в обчисленнях, необхідно застосовувати функцію VALUE. Системні відсутні значення в такий спосіб використовувати в обчисленнях не можна.

Правила запису арифметичних і логічних виразів, а також перелік функцій, що можуть бути використані, ми вже розглядали у пункті 1 розділу 6.

Розглянемо застосування операції **Compute** до вирішення конкретних задач, що виникають у процесі аналізу даних.

7.1.1. Створення індексів

Індексом називають об'єднання кількох окремих змінних, що є індикаторами якогось складного стану чи явища, в єдиний показник, числове значення якого характеризує досліджуваний складний стан. Прикладами є коефіцієнт інтелекту IQ, що часто використовують психологи, або ж відомий у нас в країні інтегральний індекс соціального самопочуття, сконструйований Є. Головахою та Н. Паніною. Загалом побудова індексу є складною процедурою, яка не вичерпується застосуванням лише операції **Compute**. Крім того, процедура побудови індексу безпосередньо залежить від рівня вимірювання тих змінних, на основі яких будується сукупний показник. Ми розглянемо один із можливих способів побудови індексу шляхом обчислення середнього арифметичного значень групи змінних (так званий адитивний індекс). Звичайно, всі змінні, що їх ми будемо застосовувати для обчислення середнього, повинні бути такими, щоб арифметичні операції мали для них сенс.

Приклад. У опитуванні киян 1991 р. було одинадцять запитань, в яких респондентам пропонувалося обрати один із семи варіантів соціальної дистанції,[15] до людей певної національності (до певної національної групи). Відповіді на ці питання містяться у групі змінних від v72 до v82. На основі відповідей про соціальну дистанцію можна обчислити низку різних індексів. Побудуємо один із таких індексів.

Кодування кожної з 11 змінних є таким, що число 1 відповідає найменшій дистанції (респондент погоджується допустити представника конкретної національності як члена власної сім'ї), а значення 7 – найбільшій дистанції (респондент вважає, що представникам цієї національності потрібно заборонити навіть відвідувати Київ). Середнє

[15] Шкала, за якою вимірювалася соціальна дистанція, була розроблена відомим американським соціологом та соціальним психологом Е. Богардусом та адаптована для застосування в Україні відомим українським соціологом Н. Паніною.

кожної такої змінної інтерпретують як показник дистанційованості киян від певної національної групи (середній бал за такою семибальною шкалою можна обчислити, якщо замовити середнє арифметичне Mean у операції Frequencies (див. розділ 3.2), причому таких індексів можна обчислити стільки, скільки є змінних). А якщо обчислити середнє значення дистанції за всіма національностями, крім українців, то отримаємо результат, який Н. Паніна інтерпретує як інтегральний індекс національної дистанційованості (ІІНД).

Для того, щоб здійснити побудову такого індексу, виконаємо певну послідовність дій:

- обираємо послідовно **Transform ⇒ Compute**,
- у вікні **Compute Variable** виконуємо такі дії (див. рис. 7.2):
  - у поле **Target Variable:** вводимо ім'я нової змінної **Index_nd**;
  - натискаємо кнопку **[Type&Label....]** і у вікні, що відкрилося, ставимо позначку біля **Label** та вводимо мітку нової змінної "Індекс національної дистанційованості";
  - залишаємо встановлений за загальною угодою числовий тип змінної;

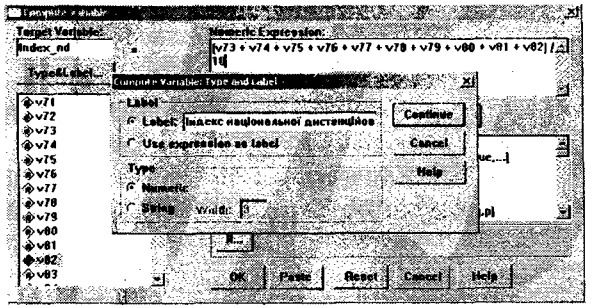


Рис. 7.2. *Параметри побудови інтегрального індексу національної дистанційованості (ІІНД).*

  - натискаємо кнопку **[Continue]**;
  - у поле **Numeric expression** вводимо формулу $(v73 + v74 + v75 + v76 + v77 + v78 + v79 + v80 + v81 + v82)/10$, яка задає обчислення значення середнього арифметичного змінних v73-v82 для кожного спостереження окремо (змінну v72 ми не розглядаємо, оскільки в цій змінній містяться відповіді про соціальну дистанцію до українців);

– оскільки ми обчислюємо індекс для всіх респондентів за однією й тією ж самою формулою, то ми не натискаємо кнопку **[If...]** і не встановлюємо умов для відбору спостережень;
– натискаємо кнопку **[Ok]**.

Нова змінна під іменем **Index_nd** з'явиться в кінці списку всіх змінних у файлі даних Data View (див. рис. 7.7) та у словнику файлу даних Variable View (див. рис. 7.3).

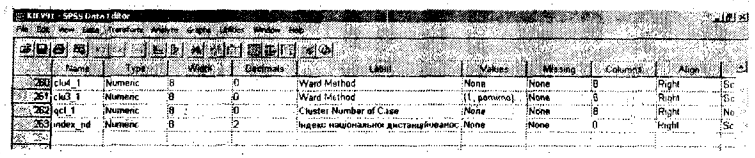


Рис. 7.3. *Нова змінна* ***Index_nd*** *у словнику файлу даних.*

Значення нової змінної є інтегральним індексом національної дистанційованості кожного респондента. Обчислимо середнє значення цього індексу для всієї вибірки. Для цього застосуємо до нової змінної **Index_nd** операцію **Frequencies**, у вікні статистик замовимо обчислення середнього арифметичного та відмовимося від виведення таблиці частот та відсотків. Результат таких обчислень на рис. 7.4.

Statistics

Індекс національної дистанційованості

N	Valid	365
	Missing	66
Mean		3,7381

Рис. 7.4. *Значення інтегрального індексу національної дистанційованості у м. Києві в 1991 році.*

Значення індексу становить 3,7, згідно зі шкалою соціальної дистанції це означає, що кияни у 1991 р. дистанціювалися від представників інших національностей (мається на увазі не українців) як від колег по роботі. Зверніть увагу на те, що 66 респондентів не дали відповідь принаймні на одне з десяти питань, і тому індекс для них не обчислений.

7.1.2. Стандартизація кількісної змінної

Як відомо, не можна порівнювати дані з різними одиницями вимірювання. Якщо ж виникає така потреба, то є принаймні два способи вирішення цього, а саме:

– можна перейти до однакових одиниць вимірювання,
– спробувати взагалі позбутися одиниць вимірювання.

У першому випадку можна скористатися операцією **Compute**. Перехід до іншої одиниці вимірювання, як правило, здійснюється шляхом певних лінійних перетворень. Нехай, наприклад, нам потрібно виражений у гривнях розмір доходу на одну особу в сім'ї респондента (змінна Inc_UAH) представити у євро. Припустимо, що поточний курс складає 7,4 гривні за 1 євро. Застосовуємо операцію **Compute** і обчислюємо нову змінну Inc_EUR за формулою *Inc_UAH*/7.4.

Один із способів "звільнення" змінної від одиниць вимірювання має назву стандартизації, яка полягає в тому, що висхідна змінна "зсувається" на значення свого середнього і нормується своїм стандартним відхиленням. Інакше кажучи, у нової (отриманої в результаті стандартизації) змінної за нуль береться середнє висхідної змінної і одиницею вимірювання є стандартне відхилення висхідної змінної. Здійснюється стандартизація змінної x за формулою $z_{x_i} = \frac{x_i - \bar{x}}{S_x}$, де z_{x_i} – значення після стандартизації, x_i – значення до стандартизації, $\bar{x}$ – середнє арифметичне, S_x – стандартне відхилення.

Приклад. Розглянемо процедуру стандартизації змінної v174, що містить інформацію про заробітну плату респондента. Для цієї змінної ми вже обчислили основні числові характеристики її розподілу (див. приклад у кінці розділу 3). Середнє значення змінної v174 дорівнює 305.36, стандартне відхилення цієї змінної дорівнює 225.19 (див. рис. 3.8). Отже значення стандартизованої змінної обчислюватиметься за формулою (v174 – 305.36)/225.19.

З метою стандартизації змінної v174 послідовно виконаємо кроки у вікні **Compute Variable** операції **Compute**:

– у поле **Target Variable:** вводимо ім'я нової змінної **Std_v174**;
– натискаємо кнопку **[Type&Label....]** і у вікні, що відкривається, обираємо опцію **Use expression as label**, тоді як мітка нової змінної буде використаний числовий вираз, за яким ця нова змінна буде обчислена; залишаємо встановлений за загальною угодою числовий тип змінної; натискаємо кнопку **[Continue]**;
– у поле **Numeric expression** вводимо формулу (v174 – 305.36)/225.19;
– натискаємо кнопку **[Ok]**.

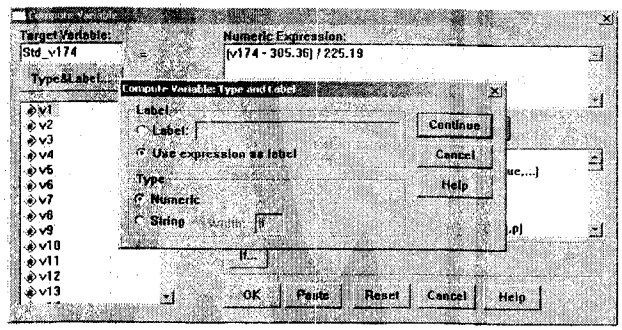
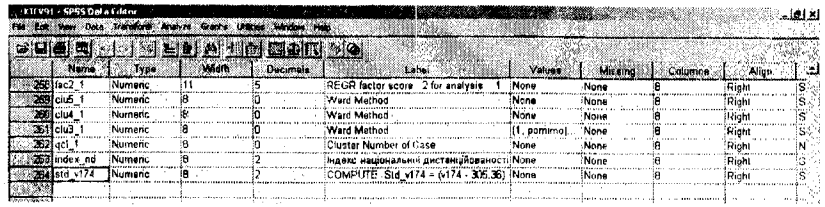


Рис. 7.5. *Вікно операції* ***Compute*** *зі значеннями параметрів для стандартизації змінної v174.*

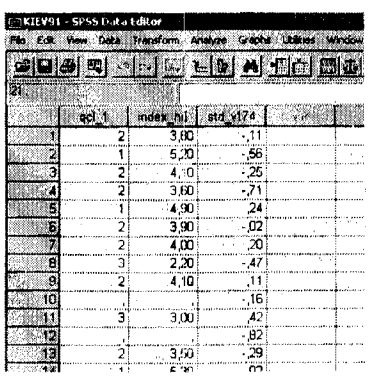
Нова змінна std_v174 з'являється в кінці списку змінних робочого файлу. На рис. 7.6 можна побачити, який вигляд має мітка цієї стандартизованої змінної.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	
358	fac2_1	Numeric	11	5	REGR factor score 2 for analysis 1	None	None	8	Right	S
359	clu5_1	Numeric	8	0	Ward Method	None	None	8	Right	S
360	clu4_1	Numeric	8	0	Ward Method	None	None	8	Right	S
361	clu3_1	Numeric	8	0	Ward Method	[1, pomimo]	None	8	Right	S
362	qcl_1	Numeric	8	0	Cluster Number of Case	None	None	8	Right	N
363	index_nd	Numeric	8	2	Індекс національної дистанційованості	None	None	8	Right	S
364	std_v174	Numeric	8	2	COMPUTE Std_v174 = (v174 - 305.36)	None	None	8	Right	S

Рис. 7.6. *Нова змінна* ***Std_v174*** *у словнику робочого файлу даних*

У нової змінної Ind_v174 (див. рис. 7.7), як і у будь-якої стандартизованої змінної, середнє дорівнює 0, а стандартне відхилення дорівнює 1.



	qcl_1	index_nd	std_v174
1	2	3,80	-,11
2	1	5,20	,56
3	2	4,10	-,25
4	2	3,60	-,71
5	1	4,90	,24
6	2	3,90	-,02
7	2	4,00	,20
8	3	2,20	-,47
9	2	4,10	,11
10	.	.	,16
11	3	3,00	-,42
12	.	.	,82
13	2	3,50	-,29

Рис. 7.7. *Нові змінні* ***Index_nd*** *та* ***Std_v174*** *у файлі даних.*

7.2. Підрахунок частоти вибору конкретних значень у групі змінних

Операція **Count** дає змогу для кожного спостереження підраховувати, скільки разів зустрічається одне або кілька конкретних значень у групі змінних і результат такого підрахунку записати у нову змінну. Наприклад, нехай за чотири роки кожен студент здав 32 іспити, і ми маємо результати складання всіх цих іспитів студентами 4-го курсу, що закінчують своє навчання. Для кожного студента маємо підраховувати, скільки він або вона мають відмінних оцінок (закодованих числом 5), і створити відповідну змінну. Ця змінна буде набувати значення в інтервалі від 0 (для тих студентів, хто за чотири роки жодного разу не отримав п'ятірки на іспиті) до 32 (для "повних відмінників").

Щоб здійснити побудову нової змінної шляхом підрахунку того, скільки разів одне або кілька значень зустрічаються у групі змінних, потрібно послідовно обрати

Transform ⇒ Count

(*Перетворення ⇒ Підрахунок*).

З'явиться вікно **Count Occurrences of Values within Cases** (*Підрахунок кількості значень у спостереженнях*), як на рис. 7.8.

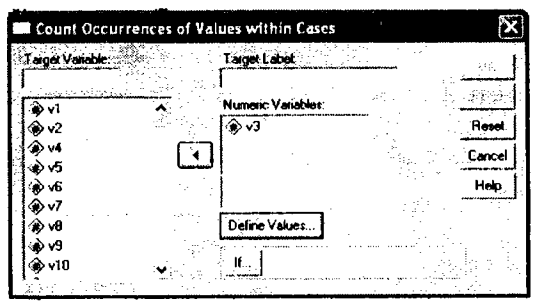


Рис. 7.8. *Вікно підрахунку кількості значень у спостереженнях.*

Далі необхідно послідовно виконати такі дії:

- у полі **Target Variable** (*Результуюча змінна*) ввести ім'я нової змінної та у полі **Target Label** (*Мітка результуючої змінної*) ввести текст відповідної мітки;
- у полі **Numeric Variables:** сформувати список змінних, для яких будуть виконуватися підрахунки;

- якщо підрахунки потрібно виконати не для всіх спостережень, а лише для тих, що відповідають певній умові, то потрібно натиснути кнопку **[If ...]**, обрати опцію **Include if case satisfies condition** (*Включати спостереження, що задовольняють умову*), ввести необхідну умову і натиснути кнопку **[Continue]**;
- натиснути кнопку **[Define Values...]** (*Визначити значення*) і сформувати у групі **Values to Count** (*Значення для підрахунку*) перелік тих значень, що будуть підраховуватися;
- натиснути **[Ok]**.

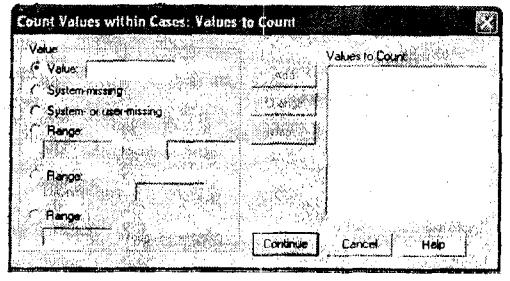


Рис. 7.9. *Вікно визначення значень у спостереженнях, які потрібно підрахувати.*

У вікні **Count Values within Cases: Values to Count** (*Підрахувати значення у спостереженнях: значення для підрахунку*) необхідно визначити, які саме значення потрібно підраховувати (див. рис. 7.9). Такими значеннями можуть бути:

- або конкретні значення, які потрібно вводити в поле **Value**: (*Значення*);
- або системне відсутнє значення; з цією метою потрібно обрати опцію **System missing**;
- або будь-яке відсутнє значення (системне відсутнє значення чи відсутнє значення, визначене користувачем); з цією метою потрібно обрати опцію **System- or user-missing**;
- або значення із певного діапазону, що визначається правою та лівою границями; з цією метою потрібно обрати опцію **Range ___through___** (*Діапазон від ___ до ___*) та ввести відповідні границі діапазону; цю опцію не можна застосовувати до рядкових змінних;

- або значення, що є не меншими ніж певне задане; з цією метою потрібно обрати опцію **Range: Lowest through___** (*Діапазон від найменшого до ___*) та ввести відповідну верхню границю відкритого зліва діапазону значень; цю опцію не можна застосовувати до рядкових змінних;
- або значення, що не перевищують певне задане; з цією метою потрібно обрати опцію **Range: ___ through highest** (*Діапазон від ___ до найбільшого*) та ввести відповідну нижню границю відкритого справа діапазону значень; цю опцію не можна застосовувати до рядкових змінних.

7.3. Перекодування значень

Якщо немає потреби обчислювати за формулою, що включає одну або декілька змінних, а потрібно лише змінити кодування змінної, більш зручним інструментом, порівняно із операцією Compute, є операція **Recode** (*Перекодувати*). Вона має два варіанти реалізації:

- **Recode into Same Variables** (*Перекодувати до тієї ж самої змінної*), який дає змогу змінити значення вже існуючої змінної;
- **Recode into Different variables** (*Перекодувати до іншої змінної*), який дозволяє не змінювати ту змінну, що перекодовують, а записати результати перекодування до нової змінної.

Для запису результатів перекодування до тієї ж змінної, що перекодовується, потрібно послідовно обрати

Transform ⇒ Recode ⇒ into Same Variables.

Розкривається вікно параметрів цієї операції (див. рис. 7.10).

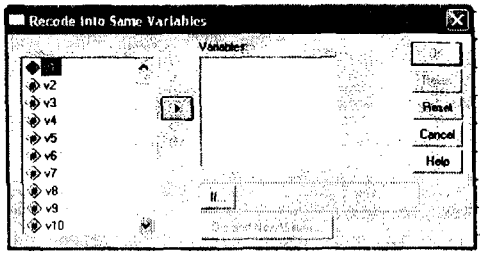


Рис. 7.10. *Вікно параметрів операції* ***Recode into Same Variables.***

Далі потрібно:

- сформувати у групі **Variables:** список тих змінних, які потрібно перекодувати; всі змінні зі списку будуть перекодовані за однією і тією ж схемою;
- якщо перекодовувати потрібно значення не у всіх спостереженнях, необхідно натиснути кнопку **[If...]**, обрати опцію **Include if case satisfies condition** (*Включати спостереження, що відповідають умові*), ввести умову і натиснути кнопку **[Continue]**;
- натиснути кнопку **[Old and New Values ...]** (*Старі та нові значення*), сформувати схему перекодування та натиснути кнопку **[Continue]**;
- натиснути кнопку **[Ok]**.

Схема перекодування формується у вікні, що розкривається кнопкою **[Old and New Values ...]**, як послідовність пар виду

старе_значення -> нове_значення,

де старе_значення – це окреме значення або інтервал значень,
нове_значення – це окреме значення.

Для запису результатів перекодування до іншої (нової) змінної потрібно послідовно обрати

Transform ⇒ Recode ⇒ into Different Variables.

Розкривається вікно параметрів цієї операції (див. рис. 7.11).

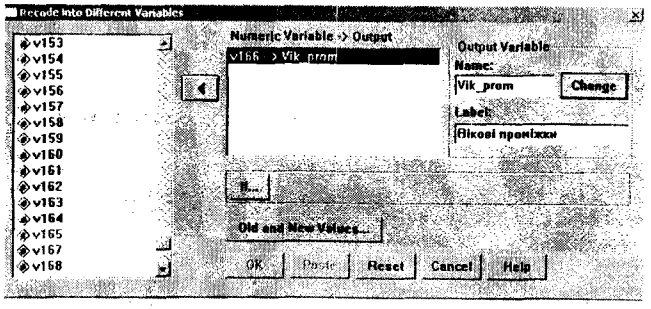


Рис. 7.11. *Вікно параметрів операції* ***Recode into Different Variables.***

Далі потрібно:

- у групі **Numeric Variable -> Output** сформувати список пар змінних виду

існуюча_змінна -> нова_змінна.

Ім'я існуючої змінної обирається зі списку всіх змінних робочого файлу. Для введення відповідної нової змінної потрібно в полі **Name** ввести ім'я, в полі **Label** ввести мітку та натиснути кнопку **[Change]** (*Змінити*);

- якщо перекодовувати потрібно значення не у всіх спостереженнях, то натиснути кнопку **[If...]**, обрати опцію **Include if case satisfies condition,** ввести потрібну умову і натиснути кнопку **[Continue]**; для тих спостережень, що не відповідають умові, нова змінна отримає системне відсутнє значення (*system missing value*);
- натиснути кнопку **[Old and New Values ...]**, сформувати схему перекодування та натиснути кнопку **[Continue]**;
- натиснути кнопку **[Ok]**.

Кнопка **[Old and New Values ...]** розкриває відповідне вікно (див. рис. 7.12), яке ділиться навпіл. У лівій частині, позначеній **Old Value** (*Старе значення*), потрібно вказати значення, що перекодовується, а у правій частині, яка позначена **New Value** (*Нове значення*), потрібно вказати результат перекодування.

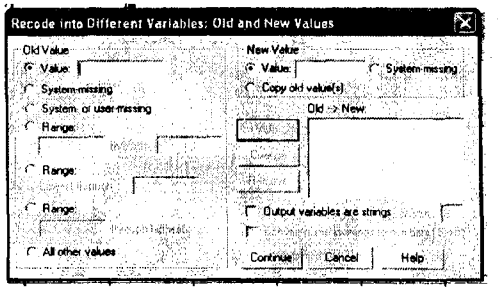


Рис. 7.12. *Вікно для формування схеми перекодування.*

Ліва частина вікна, позначена Old Value, має структуру, подібну до структури вікна операції Count, зображеного на рис. 7.9 та описаного у розділі 7.2. Відмінність полягає лише в тому, що тепер з'явилася ще одна додаткова можливість – обрати **All other values** (*Усі інші значення*) і визначити, як потрібно вчинити з усіма іншими, тобто не врахованими сформованою схемою перекодування, значеннями.

У правій частині вікна, позначеній New Value, є опція **Copy old value(s)** (*Копіювати старі значення*), яка дає змогу не змінювати (тобто просто копіювати) діапазон значень.

7.3.1. Розбиття діапазону значень неперервної змінної на інтервали

Необхідність розбити діапазон значень неперервної змінної на інтервали виникає досить часто. Прикладом може бути виділення кількох вікових когорт респондентів (розбиття на інтервали неперервної змінної "Вік"), виділення груп респондентів за розмірами прибутків (розбиття на інтервали неперервної змінної "Прибуток в сім'ї на одну особу") тощо. Таке розбиття на інтервали зручно виконувати з використанням операції **Recode**.

Приклад. Нехай під час аналізу даних опитування киян 1991 р. виникла необхідність виділити три вікові групи – молодь (до 35 років включно), люди середнього віку (від 36 до 55 років включно) та літні люди (56 років та старше). У нашому файлі даних інформація про вік респондента міститься у змінній v166. Обчислимо нову змінну Vik_prom шляхом перекодування існуючої змінної v166. Ця нова змінна буде дорівнювати 1 для молодих респондентів, 2 – для респондентів середнього віку, 3 – для респондентів літнього віку.

Послідовно обираємо **Transform** ⇒ **Recode** ⇒ **into Different Variables** та у вікні, що з'явилося (див. рис. 7.11), виконуємо такі дії:

- заносимо змінну v166 у поле **Numeric Variable -> Output**;
- вводимо ім'я нової змінної **Vik_prom** у поле **Name;**
- вводимо текст мітки "Вікові проміжки" у поле **Label;**
- натискаємо кнопку **[Change]**.

Оскільки нам потрібно здійснити перекодування для всіх спостережень у файлі, ми не будемо використовувати кнопку **[If...]**).

Натискаємо кнопку **[Old and New Values ...]** та формуємо у вікні, що розкрилося (див. рис. 7.13), схему перекодування. З цією метою:

- у лівій частині вікна **Old value** виділяємо опцію **Range Lowest through** і вводимо значення 35 для правої границі відкритого інтервалу; у правій частині вікна **New value** обираємо опцію **Value**, вводимо значення 1 та натискаємо кнопку **[Add]**;
- у лівій частині вікна **Old value** виділяємо опцію **Range ___through___** і вводимо значення 36 та 55 для лівої та для правої границь інтервалу; у правій частині вікна **New value** обираємо опцію **Value**, вводимо значення 2 та натискаємо кнопку **[Add]**;
- у лівій частині вікна **Old value** виділяємо опцію **Range ___through highest** і вводимо значення 56 для лівої границі відкритого

інтервалу; у правій частині вікна **New value** обираємо опцію **Value**, вводимо значення 3 та натискаємо кнопку **[Add]**;
- перекодуємо відсутні значення у відсутні, для чого у лівій частині вікна **Old value** виділяємо опцію **System- or user-missing;** у правій частині вікна **New value** обираємо опцію System missing та натискаємо кнопку **[Add]**;
- натискаємо кнопку **[Continue]**, закривається вікно параметрів, натискаємо кнопку **[Ok]**.

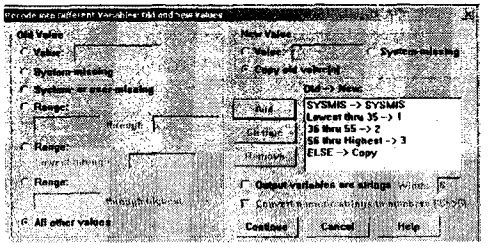


Рис. 7.13. *Вікно для формування схеми перекодування змінної v166.*

Нова змінна **Vik_prom** з'явиться у файлі даних (див. рис. 7.16), а також у словнику (див. рис. 7.17). Для більш зручної роботи бажано визначити мітки значень нової змінної.

7.3.2. Групування категорій дискретної змінної

Групування категорій дискретної змінної також зручно виконувати за допомогою операції **Recode**.

Приклад. У нашому файлі даних опитування киян 1991 р. є змінна v171, яка містить інформацію про рід занять респондента. Побудуємо нову змінну, яка розділяє респондентів на дві категорії – керівники різного рівня (такий рід занять закодований у змінній v171 значеннями 1 та 2) та всіх інших (інші значення змінної v171). Побудуємо нову змінну v171_1, яка буде мати значення 1 для керівників різного рівня та значення 0 для інших респондентів. Ця дихотомічна змінна виділяє категорію керівників із загальної множини респондентів.

Послідовно обираємо **Transform** ⇒ **Recode** ⇒ **into Different Variables** та у вікні, що відкривається (див. рис. 7.14), виконуємо такі дії:

- переносимо ім'я змінної v171 у поле **Numeric Variable -> Output**;
- вводимо ім'я нової змінної v171_1 у поле **Name**;
- вводимо текст мітки "Керівники різних рівнів" у поле **Label**;
- натиснемо кнопку **[Change]**.

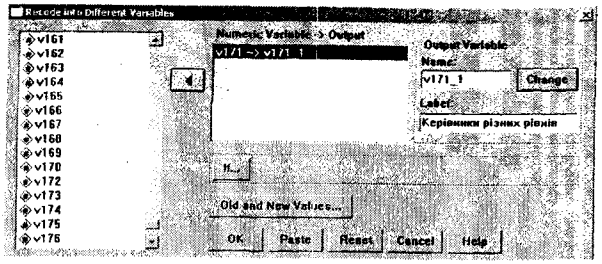


Рис. 7.14. *Параметри операції* ***Recode into Different Variables*** *для перекодування змінної v171.*

Натискаємо кнопку **[Old and New Values ...]** та у вікні, що відкривається (див. рис. 7.15), формуємо схему:

```
1 -> 1
2 -> 1
Missing -> SYSMIS
ELSE -> 0
```

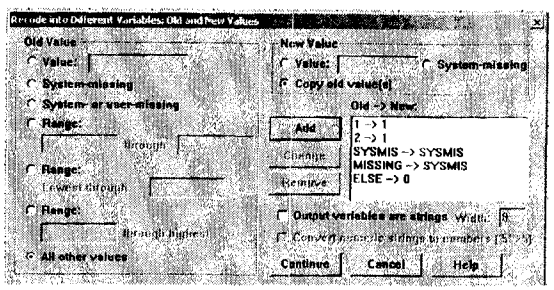


Рис. 7.15. *Вікно для формування схеми перекодування змінної v171.*

Нова змінна v171_1 з'явиться у файлі даних (див. рис. 7.16), а також у словнику (див. рис. 7.17). Для більш зручної роботи бажано визначити мітки значень нової змінної.

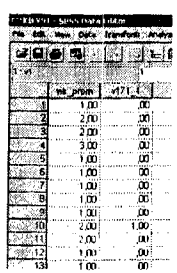


Рис. 7.16. *Нові змінні* ***Vik_prom*** *та* ***v171_1*** *у файлі даних.*

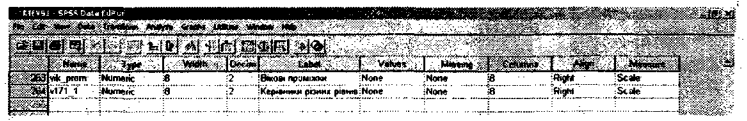


Рис. 7.17. *Нові змінні* ***Vik_prom*** *та* ***_v171_1*** *у словнику файлу даних.*

Зауваження. Для того, щоб змінну можна було використовувати під час подальшого аналізу даних (у наступних сеансах роботи), потрібно обов'язково записати робочий файл даних на диск.

Контрольні запитання

1. Які команди для обчислення нових змінних Ви знаєте?
2. В яких випадках виникає потреба розбити діапазон значень неперервної змінної на інтервали? Наведіть приклади.
3. В яких випадках виникає потреба здійснити перекодування категорій дискретної змінної? Наведіть приклади.
4. Чим відрізняється зміна значень існуючої змінної від перекодування до нової змінної?
5. Як формується схема перекодування змінної?
6. В операції Recode є можливість відбирати за умовою спостереження, значення яких будуть перекодовуватися (кнопка **[If...]**). Які значення має нова змінна (при перекодуванні до нової змінної) для спостережень, що не задовольняють цю умову?

Практичні завдання

Використовуючи такі змінні файлу даних опитування киян у 1991 р., як

v5	Чи задоволені Ви своїм становищем у суспільстві?
v7	Чи задоволені Ви тим, що отримуєте від суспільства?
v8	Чи задоволені Ви тим, що віддаєте суспільству?
v165	Стать респондента
v166	Вік респондента
v167	Освіта респондента
v168	Сімейний стан респондента,

зробіть такі операції:

1. Побудуйте нову змінну Rik_Nar, значенням якої є рік народження респондента. Обчисліть середній та медіанний рік народження респондента.
2. Обчисліть відсоток чоловіків, що мають вік у інтервалі від 35 до 43 років, та відсоток жінок, що мають вік до 25 років.

Зауваження. Побудуйте нову змінну, що розбиває вік респондента на інтервали.

3. Вважаючи, що загальна задоволеність життям респондента (нова змінна zadovol) може бути обчислена як середнє арифметичне змінних v5, v7 та v8, обчисліть параметри розподілу загальної задоволеності життям (середнє, мода, медіана, стандартне відхилення, коефіцієнт варіації).

4. Побудуйте нову змінну simya, що набуває значення 3 для заміжніх жінок, значення 2 – для одружених чоловіків і значення 0 – для всіх інших. Побудуйте одновимірний розподіл для цієї змінної.

РОЗДІЛ 8. РЕМОНТУВАННЯ ВИБІРКИ. ЗВАЖУВАННЯ

Навіть якщо вибірка є правильно спланованою, та під час практичної реалізації плану вибірки часто виникають різноманітні проблеми. Зокрема, "недосяжність" певних потенційних респондентів може призводити до певних викривлень у структурі вибірки. Досить часто після завершення етапу збору інформації виконують процедуру ремонтування вибірки. Метою цієї процедури є відтворення у вибірці відомих із зовнішніх надійних джерел інформації характеристик генеральної сукупності. Такими зовнішніми джерелами інформації часто є або дані державної статистики, або ж дані інших досліджень.

Пакет SPSS дає можливість дослідникові ремонтувати вибірку методом зважування. Ідея методу зважування полягає в тому, що кожному спостереженню присвоюється певне позитивне число, що розглядається як ваговий коефіцієнт цього спостереження (або говорять просто про вагу спостереження) у загальній вибірці. Працюючи зі зваженою вибіркою, комп'ютерна програма оперує не кількостями спостережень, а сумами вагових коефіцієнтів цих спостережень. Зокрема зазначимо, що навіть незважена вибірка може розглядатися як така, в якій кожне спостереження має вагу 1.

Технічно процес зважування виглядає дуже просто. Потрібно вказати певну змінну, яка під час подальшого аналізу буде розглядатися як вагова. Значення цієї змінної для кожного із спостережень розглядається як ваговий коефіцієнт цього спостереження у загальній вибірці. Якщо для певного спостереження значенням вагової змінної є 0 або від'ємне значення, або відсутнє значення, то таке спостереження в результаті зважування виключається з аналізу.

Для виконання зважування файлу даних потрібно виконати таку послідовність дій:

- обрати в головному меню

Data ⇒ Weight Cases ...

(*Дані ⇒ Зважити спостереження*).

Розкривається вікно параметрів операції **Weight Cases ...** (див. рис. 8.1);

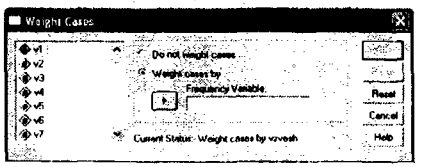


Рис. 8.1. *Вікно параметрів операції* ***Weight Cases.***

- обрати опцію **Weight cases by** (*Зважити спостереження за допомогою*) та занести в поле **Frequency Variable** (*Частотна змінна*) ім'я тієї змінної, яка буде використовуватися як вагова;
- натиснути кнопку [**Ok**].

На рамці вікна редактора даних SPSS має з'явитися повідомлення **Weight On** (*Вага підключена*).

Для того, щоб відключити вагу та повернутися до аналізу незваженої вибірки, потрібно:

- послідовно обрати в головному меню **Data ⇒ Weight Cases ...**;
- обрати опцію **Do not weight cases**;
- натиснути кнопку [**Ok**].

На рамці вікна редактора даних SPSS має зникнути повідомлення **Weight On**.

Зауваження. Якщо файл даних записаний на диск як зважений, то наступного разу після відкриття для роботи він буде розглядатися також як зважений.

Процедура зважування спостережень проста. Цей етап виконується в два кроки:

- визначити ваги даних;
- обчислити змінну, значеннями якої будуть відповідні значення ваг.

Обчислення нової змінної подавалося у розділі 7. Розглянемо процес обчислення вагової змінної.

Приклад. Ми продовжуємо працювати із даними дослідження, проведеного у місті Києві у 1991 р. Нехай, за даними державної статистики, нам відомі такі пропорції:

- за статтю – 46 % становлять чоловіки і 54 % жінки;
- за освітою у чоловіків – 42,2 % вища освіта, 29,6 % середня спеціальна освіта, 22,2 % середня загальна освіта, 6,1 % незакінчена середня освіта;
- за освітою у жінок – 47, 4 % вища освіта, 30,4 % середня спеціальна освіта, 18,1 % середня загальна освіта, 4,1 % незакінчена середня.

Наша мета полягає в тому, щоб обчислити такі ваги, які б відтворювали в нашому файлі даних зазначені пропорції за статтю та освітою.

Для того, щоб з'ясувати реальний розподіл за статтю та освітою у файлі kiev91, побудуємо двовимірну таблицю для змінних v165 (стать) та v167 (освіта). Результат обчислень операції Crosstabs... представлений нижче у таблиці 8.2.

Потрібно зважити вибірку дослідження. Маємо 431 об'єкт, але у одного респондента стать не вказана. Тому у зваженій вибірці буде 430 об'єктів (див. табл. 8.1).

Таблиця 8.1

Розподіл кількості спостережень

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Ваше образование * Ваш пол	430	99.8%	1	.2%	431	100.0%

Таблиця 8.2

Спільний розподіл статі та освіти до зважування

Ваше образование * Ваш пол Crosstabulation

			Ваш пол		Total
			мужск.	женск.	мужск.
Ваше образование	высшее	Count	95	109	204
		% within Ваш пол	49.7 %	45.6 %	47.4 %
	средн. спец.	Count	51	80	131
		% within Ваш пол	26.7 %	33.5 %	30.5 %
	среднее общ.	Count	35	33	68
		% within Ваш пол	18.3 %	13.8 %	15.8 %
	н/среднее	Count	10	17	27
		% within Ваш пол	5.2 %	7.1 %	6.3 %
Total		Count	191	239	430
		% within Ваш пол	100.0 %	100.0 %	100.0 %

Всього у файлі даних 431 спостереження. Проте в одному спостереженні не вказана стать респондента, отже таблиця побудована для 430 спостережень, і після зважування ми будемо мати також 430 спостережень (сума всіх ваг для всіх спостережень має дорівнювати 430). Спостереження із незазначеною статтю респондента в результаті зважування буде просто викинуто з аналізу.

У побудованій двовимірній таблиці є 8 клітин. Отже, маємо 8 груп спостережень, і для кожної такої групи треба вказати формулу для обчислення ваги (всі спостереження з однієї групи будуть мати однакову вагу). Розглянемо верхню ліву клітину таблиці. Вона містить дані про чоловіків із вищою освітою. Спостереження, що потрапили до цієї клітини, можна відібрати умовою (v165=1 and v167=1). Загальний обсяг вибірки після зважування має бути 430 і 46 % мають становити чоловіки. Інакше кажучи, після зважування у вибірці має бути 430×0.46 чоловіків. Серед них 42,2 % – чоловіки із вищою освітою. Отже, після зважування у нас має бути $(430 \times 0{,}46) \times 0.422$ чоловіків із вищою освітою. Саме така сума ваг спостережень, що стосуються чоловіків із вищою освітою. Проте у нашому файлі даних є 95 чоловіків із вищою освітою. Отже, кожному зі спостережень, що стосуються цієї групи (тобто спостережень, що відповідають умові (v165=1 and v167=1)) потрібно присвоїти вагу, що обчислюється за формулою

$$\frac{430 \times 0.46 \times 0.422}{95}.$$

Використовуючи поданий у таблиці 8.1 сумісний розподіл двох ознак, а також міркування, подібні до викладених у попередньому абзаці, легко побачити, що:

- для чоловіків із середньою спеціальною освітою (умова відбору (v165=1 and v167=2)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.46 \times 0.296}{51};$$
- для чоловіків із середньою загальною освітою (умова відбору (v165=1 and v167=3)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.46 \times 0.222}{35};$$
- для чоловіків із незакінченою середньою освітою (умова відбору (v165=1 and v167=4)) ваговий коефіцієнт потрібно обчислювати за формулою

$$\frac{430 \times 0.46 \times 0.061}{10};$$

- для жінок із вищою освітою (умова відбору (v165=2 and v167=1)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.54 \times 0.474}{109};$$
- для жінок із середньою спеціальною освітою (умова відбору (v165=2 and v167=2)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.54 \times 0.304}{80};$$
- для жінок із середньою загальною освітою (умова відбору (v165=2 and v167=3)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.54 \times 0.181}{33};$$
- для жінок із незакінченою середньою освітою (умова відбору (v165=2 and v167=4)) ваговий коефіцієнт потрібно обчислювати за формулою
$$\frac{430 \times 0.54 \times 0.041}{17}.$$

Якщо тепер обчислити за цими формулами значення певної змінної (категорії цієї змінної матимуть відповідно вісім різних значень) і вказати цю змінну як вагову, то потрібні нам пропорції за статтю та освітою будуть відтворені.

Зауваження 1. Якщо ваги є не цілими числами (а, як правило, так воно завжди і є), то в результаті округлень може виникнути ситуація, коли сума частот на 1 або 2 об'єкти не збігається із маргінальними значенням суми у відповідному рядку. Те ж саме і відносно сум частот у стовпчиках та загальної суми частот у таблиці. Такі розбіжності не впливають на значення статистичних коефіцієнтів та на висновки аналізу, але можуть справляти неприємне враження на тих, хто не знайомий зі специфікою зважування нецілими значеннями.

Зауваження 2. Значення вагової змінної не повинні бути дуже великими. Ті групи спостережень, для яких обчислене значення ваги є більшим ніж 1, у зваженій вибірці будуть "штучно збільшені". Ті ж групи, для яких ваговий коефіцієнт буде менше 1, після зважування зменшуватимуться. Занадто велике штучне, шляхом зважування, збільшення певної групи є небажаним. Не потрібно компенсувати зважуванням невдале планування вибірки або ж її погану реалізацію. Бажано, щоб значення ваг не перевищували $\sqrt{2}$.

Контрольні запитання

1. З якою метою застосовують процедуру зважування?
2. У чому, на Вашу думку, полягають особливості виконання операції зважування у SPSS?
3. Що таке "вагова змінна" і якою є її роль в операції зважування?
4. З яких міркувань обчислюються значення ваг?
5. Чи змінюються частотні розподіли змінних після зважування?

Практичні завдання

1. Виконайте зважування файлу даних за формулами, наведеними в розділі. Порівняйте двовимірний розподіл змінних v165 і v167 до зважування та після зважування. Чи відтворені після зважування потрібні нам пропорції двовимірного розподілу статі та освіти?

2. Будемо вважати, що у нашому файлі даних є результати опитування населення, старшого 16 років у місті N. Нехай, за даними міської адміністрації, відомо, що в місті проживає 1 160 000 жінок, старших 16 років, і 840 000 чоловіків, старших 16 років. Зважте файл даних так, щоб кількість спостережень у файлі не змінилася, і в той же час розподіл за ознакою стать (змінна v165) у вибірці відповідав розподілу у генеральній сукупності. Обчисліть у зваженому файлі середню задоволеність власним становищем (змінна v5) та зробіть одновимірний розподіл за сімейним станом (змінна v168). Порівняйте ці розподіли з результатами для незваженого файлу.

3. З метою порівняння результатів дослідження, дані якого є в нашому файлі, з результатами аналогічного дослідження, проведеного в минулому році, зважте файл даних так, щоб розподіл за статтю відповідав генеральній сукупності (див. вище), а обсяг вибірки у зваженому файлі був таким, як і в минулорічному дослідженні – 450 спостережень. Обчисліть у зваженому файлі середню задоволеність власним становищем (змінна v5) та зробіть одновимірний розподіл за сімейним станом. Порівняйте з результатами у незваженому файлі.

РОЗДІЛ 9. СТАТИСТИЧНІ ВИСНОВКИ

Статистичні висновки – це твердження про невідомі параметри генеральної сукупності, зроблені на основі параметрів вибірки. Такі твердження мають імовірнісний характер. Їх поділяють на три види: точкове статистичне оцінювання, інтервальне статистичне оцінювання та перевірка статистичних гіпотез.

Метою точкового оцінювання є пошук вибіркового показника, найбільш близького за значенням до параметра генеральної сукупності, що оцінюється. Як правило, точковими оцінками параметрів генеральної сукупності є значення відповідних параметрів вибірки, в деяких випадках обчислених за дещо зміненими формулами.

Метою інтервального оцінювання є побудова такого інтервалу на числовій осі, всередині якого з достатньо високою ймовірністю міститься значення параметра генеральної сукупності, що оцінюється. Інтервал називають довірчим інтервалом, а ймовірність – довірчою ймовірністю.

Статистична гіпотеза формулюється для параметрів генеральної сукупності. Перевірка статистичної гіпотези полягає у побудові на основі вибіркових параметрів певного правила, яке дозволяє з невеликою ймовірністю припуститися помилки і або прийняти, або відхилити сформульовану гіпотезу.

9.1. Інтервальне оцінювання

Дуже часто у практичній роботі зі статистичного аналізу емпіричних даних виникає потреба будувати довірчі інтервали для відсотка (або, що є тим самим, для частки) об'єктів із певною властивістю у генеральній сукупності та для середнього значення певної змінної. Як стандартні значення довірчої ймовірності прийнято використовувати значення 0.95 або 0.99.

Розглянемо процедуру побудови довірчого інтервалу для відсотка (частки) об'єктів із певною властивістю.

Якщо з нескінченної генеральної сукупності[16] зроблена проста випадкова вибірка обсягу n і у цій вибірці частка об'єктів із властивістю X, що нас цікавить, становить $\hat{p}$, то за умови що $n\cdot\hat{p}>5$ (тобто є принаймні 6 об'єктів, що мають властивість X) та $n(1-\hat{p})>5$

[16] Нагадаємо, що генеральну сукупність можна розглядати як нескінченну, якщо її розмір принаймні в 100 разів більший ніж розмір вибірки.

(тобто є принаймні 6 об'єктів, що не мають властивість X) довірчий інтервал для частки об'єктів із властивістю X у генеральній сукупності обчислюється за формулою $\hat{p} \pm z\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$, де z – критичне значення нормального розподілу з параметрами (0,1) для необхідного значення довірчої ймовірності[17].

Розглянемо приклади побудови довірчого інтервалу для частки (відсотка) об'єктів із певною властивістю.

Приклад 1. Використовуючи дані опитування киян у 1991 р., побудуємо довірчий інтервал для частки тих, хто вважав, що останній рік був кращий, ніж попередній. Відповіді респондентів вибіркового опитування на відповідне питання є у змінній v2 файлу даних.

Застосовуємо операцію **Frequencies...** до змінної v2, будуємо одновимірну таблицю частот та відсотків і бачимо, що всього на поставлене питання відповіли 426 респондентів і з них 2,1 % обрали варіант відповіді "кращий, ніж попередній". Отже, відсоток оптимістично налаштованих респондентів у вибірці становить 2,1 % (обсяг вибірки). Нас цікавить, яким є відсоток оптимістично налаштованих респондентів у генеральній сукупності.

Будуємо довірчий інтервал для довірчої ймовірності 0.95. Маємо обсяг вибірки $n = 426$, частка потрібної нам властивості у вибірці $\hat{p} = 0{,}021$. Оскільки $n\cdot\hat{p} = 426\cdot 0{,}021 > 5$ та $n\cdot(1-\hat{p}) = 426\cdot(1-0{,}021) > 5$, то для довірчої ймовірності 0.95 маємо такий довірчий інтервал для частки об'єктів із потрібною нам властивістю:

$$0.021 \pm 1.96\sqrt{\frac{0.021\cdot(1-0.021)}{426}} = 0.021 \pm 0.014 = [0.007\ 0.035].$$

Інакше кажучи, з ймовірністю 0.95 можна стверджувати, що в генеральній сукупності (а для цього дослідження генеральною сукупністю було населення міста Києва у віці 18 років та старше) відсоток тих, хто оцінює минулий рік як кращий ніж попередній, міститься в інтервалі від 0,7 % до 3,5 %.

Зробимо важливі зауваження.

Зауваження 1. Ширина довірчого інтервалу є тим більшою, чим більшою є довірча ймовірність. Інтервал для довірчої ймовірності 0.99 включає в себе аналогічний інтервал для довірчої ймовірності 0.95.

[17] Для значення довірчої ймовірності 0.95 використовуємо $z = 1{,}96$, для значення довірчої ймовірності 0.99 використовуємо $z = 2{,}58$.

Зауваження 2. Ширина довірчого інтервалу є тим меншою, чим більшим є обсяг вибірки. У процесі аналізу великих за обсягом вибірок довірчі інтервали є більш вузькими, ніж при роботі із невеликими вибірками.

Зауваження 3. Вираз $\hat{p}\cdot(1-\hat{p})$ має найбільше значення у випадку $\hat{p} = 0.5$. Це означає, що найширшим буде інтервал для значення частки, близького до 0.5 (тобто для значення відсотка, близького до 50 %). При зростанні значення частки від 0.5 (при зменшенні значення частки від 0.5) ширина довірчого інтервалу зменшується.

Зауваження 4. У випадку, коли значення $\hat{p}$ є дуже близьким до 0 або ж до 1 (тобто у вибірці немає/майже немає об'єктів із потрібною нам властивістю, або ж у вибірці всі/майже всі об'єкти мають потрібну нам властивість), потрібно застосовувати для оцінювання довірчого інтервалу інші формули.

Тепер розглянемо побудову довірчого інтервалу для середнього значення змінної.

Якщо з нескінченної генеральної сукупності зроблена проста випадкова вибірка обсягу n і для цієї вибірки обчислені середнє X_B та стандартне відхилення S_b, то довірчий інтервал для генерального середнього $\bar{X}_Г$ має вигляд $X_B \pm t_{n-1}\frac{S_B}{\sqrt{n}}$, де t_{n-1} є критичним значенням розподілу Стьюдента з $(n-1)$ ступенями волі для обраного значення довірчої ймовірності. Нагадаємо, що для великої кількості ступенів волі (зокрема 300 та більше) розподіл Стьюдента добре апроксимується нормальним розподілом із параметрами (0,1). Отже, в нашому випадку, якщо обсяг вибірки є досить великим (зокрема більше ніж 300), то, будуючи довірчий інтервал для середнього замість критичного значення розподілу Стьюдента, можна використовувати критичні значення нормального розподілу. Зауважимо також, що значення виразу $\frac{S_B}{\sqrt{n}}$ часто називають стандартною помилкою середнього.

Розглянемо приклад побудови довірчого інтервалу для середнього значення змінної.

Приклад 2. Використовуючи дані опитування киян у 1991 році, побудуємо довірчий інтервал для місячного прибутку на одну особу в сім'ї (змінна v173, одиниця вимірювання – радянський рубль).

Застосовуємо процедуру **Frequencies...** до змінної v173 та замовляємо обчислення середнього (параметр **Mean** у групі **Central tendency**), стандартного відхилення та стандартної помилки середнього (параметри **Std.deviation** та **S.E.mean** відповідно). Результат обчислень свідчить про те, що на питання про розмір місячного прибутку відповіли 418 респондентів, середнє значення розміру місячного прибутку дорівнює 222.28, стандартне відхилення дорівнює 104.29 і стандартна помилка середнього дорівнює 5.1. Бачимо, що сукупність є досить неоднорідною за прибутками на одну особу в сім'ї, оскільки коефіцієнт варіації є досить великим (маємо $104.29/222.28 \approx 0.47$). Ми могли б застосувати вказану вище просту формулу обчислення стандартної помилки середнього й отримали б те ж саме значення, що нам обчислила комп'ютерна програма (дійсно, маємо $104.29/\sqrt{418} \approx 5.1$).

Оскільки обсяг нашої вибірки є досить великим, то потрібне нам критичне значення розподілу Стьюдента з 417 ступенями волі збігається з відповідним критичним значенням нормального розподілу з параметрами (0,1). Таким чином, для довірчої ймовірності 0.95 потрібний нам довірчий інтервал буде мати такий вигляд

$$222.28 \pm 1.96 \times 5.1 = 222.28 \pm 10 = [212.28,\ 232.28].$$

Отже, за результатами нашого вибіркового опитування з довірчою ймовірністю 0.95 ми можемо стверджувати, що у 1991 р. середнє значення місячного прибутку на одну особу в місті Києві було в інтервалі від 212 руб. до 232 руб.

Побудову довірчого інтервалу для середнього в SPSS також здійснює операція **Explore...** (*Розвідувати*). Побудуємо довірчий інтервал для середнього значення змінної v173 і порівняємо результат із отриманими раніше значеннями. Послідовно обираємо

Analyze ⇒ Descriptive Statistics ⇒ Explore...

Розкривається вікно параметрів (див. рис. 9.1).

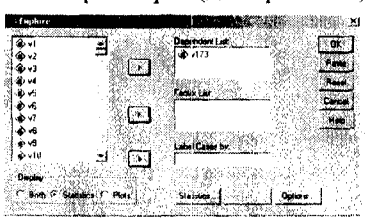


Рис. 9.1. *Вікно параметрів процедури* ***Explore.***

Заносимо у список **Dependent List** (*Список залежних*) ім'я v173, обираємо в групі **Display** (*Виводити*) опцію **Statistics** (*Статистики*) та натискаємо кнопку **[OK]**. Результат обчислень подано на рис. 9.2.

Descriptives

			Statistic	Std. Error
Каков среднемесячный доход Вашей семьи на одного человека? (	Mean		222,2775	5,1012
	95% Confidence Interval for Mean	Lower Bound	212,2503	
		Upper Bound	232,3047	
	5% Trimmed Mean		214,0984	
	Median		200,0000	
	Variance		10877,218	
	Std. Deviation		104,2939	
	Minimum		50,00	
	Maximum		800,00	
	Range		750,00	
	Interquartile Range		120,7500	
	Skewness		1,490	,119
	Kurtosis		3,999	,238

Рис. 9.2. *Результат обчислень операції* ***Explore...*** *для змінної v173.*

Бачимо, що для довірчої ймовірності 0.95 довірчий інтервал для середнього значення v173 (**95 % Confidence Interval for Mean**) міститься між лівою границею (**Lower Bound**) 212.25 та правою границею (**Upper Bound**) 232.30. Цей результат досить близько збігається із побудованим нами раніше на основі округлених значень довірчим інтервалом [212.28; 232.28].

Зауважимо, що під час побудови довірчого інтервалу для середнього в операції **Explore...** є можливість встановлювати необхідний рівень довірчої ймовірності. Для цього потрібно натиснути кнопку **[Statistics]** та встановити необхідне значення довірчої ймовірності (у відсотках) в полі **Confidence Interval for Mean**.

9.2. Перевірка статистичних гіпотез

Статистична гіпотеза є твердженням про невідомі параметри генеральної сукупності. Перевірка статистичної гіпотези полягає у вирішенні того, приймати чи відхилити гіпотезу на основі відомих значень параметрів вибірки. Прийняття або відхилення статистичної гіпотези супроводжується ризиком припуститися помилки. Ймовірність відхилити гіпотезу за умов, якщо насправді вона є правильною, називають ймовірністю помилки першого роду. Таку ймовірність називають ще значущістю. На практиці правило відхилення гіпотези формулюють так, щоб значущість не перевищувала або 0.05, або 0.01.

Звичайно, не будь-яку статистичну гіпотезу можна перевірити. Для того, щоб на основі значень параметрів однієї вибірки мати можливість побудувати таке правило прийняття/відхилення конкретної статистичної гіпотези, що гарантує не перевищення певного рівня ймовірності помилки, потрібно знати певні (які саме, це залежить від формулювання гіпотези) вибіркові розподіли.

У розділі 4 ми розглядали операцію **Crosstabs...**. Для з'ясування питання про наявність зв'язку між двома дискретними змінними операція **Crosstabs...** обчислює коефіцієнт χ^2 та оцінює його значущість (див. рис. 9.2).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	78,741[a]	76	,396
Likelihood Ratio	99,930	76	,034
Linear-by-Linear Association	3,280	1	,070
N of Valid Cases	417		

a. 138 cells (89,6%) have expected count less than 5. The minimum expected count is ,44

Рис. 9.2. *Результат обчислення коефіцієнта χ^2 операцією* ***Crosstab...***

Ми зазначали, що говорити про наявність зв'язку між змінними доцільно тоді, коли значущість коефіцієнта χ^2 не перевищує 0.05. Тепер можемо пояснити це правило більш детально. Операція **Crosstabs...** на основі значення коефіцієнта χ^2 перевіряє статистичну гіпотезу про те, що емпіричний двовимірний розподіл, який ми аналізуємо, збігається із розподілом статистичної незалежності двох змінних. Відкидання цієї гіпотези рівнозначне твердженню про те, що дві змінні в емпіричних даних не є статистично незалежними, а отже, між ними існує кореляційний зв'язок. Значущість – це ймовірність помилки, яка полягає в тому, що ми відкидаємо певну гіпотезу, коли ця гіпотеза насправді є правильною. В нашому випадку значущість – це ймовірність помилки, яка полягає у твердженні, що зв'язок між змінними є, коли насправді зв'язку немає. Ця ймовірність має бути невеликою. Вважається, що прийнятною є ймовірність, яка не перевищує або є 0.05, або 0.01.

Операція **Crosstabs...** дозволяє також обчислювати і коефіцієнт кореляції Пірсона для двох змінних. Для кожного такого коефіцієнта кореляції перевіряється гіпотеза про його рівність нулю. Ймовірність

помилки першого роду під час перевірки такої гіпотези оцінюється і виводиться як "значущість коефіцієнта кореляції". Інакше кажучи, значущість – це ймовірність того, що ми помилково заявляємо про те, що у генеральній сукупності коефіцієнт відрізняється від нуля, а насправді у генеральній сукупності кореляції немає (коефіцієнт дорівнює нулю). Отже, якщо коефіцієнт кореляції Пірсона є значущим, то статистично фіксується наявність лінійного кореляційного зв'язку між двома ознаками.

Операція **Correlate...**, яку ми також розглядали у розділі 4, дає змогу будувати матрицю коефіцієнтів кореляції Пірсона (так звану кореляційну матрицю) і для кожного коефіцієнта у матриці визначає його рівень значущості. Процедура **Regression** побудови та оцінювання рівняння лінійної регресії, яку ми будемо розглядати у наступному розділі, перевіряє гіпотезу про рівність нулю для кожного з коефіцієнтів побудованого рівняння. Крім того, багато інших статистичних процедур використовують перевірку статистичних гіпотез для оцінки результатів своєї роботи, для прийняття рішень про необхідність продовження обчислень тощо.

Зауваження 1. Незначущі на потрібному для дослідника рівні значущості[18] коефіцієнти не інтерпретуються.

Розглянемо, як здійснюється перевірка ряду стандартних процедур стосовно середнього значення змінної.

Під час аналізу може виникнути потреба з'ясувати, чи відрізняється в генеральній сукупності середнє певної змінної від деякої константи, від значення, визначеного із джерела інформації, зовнішнього щодо емпіричних даних, які аналізуються (з даних державної статистики, з нормативних документів, із результатів інших досліджень тощо).

Приклад 1. Припустимо (це суто гіпотетичне припущення), що прожитковий мінімум в Україні у 1991 р. становив 240 рублів. Джерелом такої інформації можуть бути дані державної статистики. Використовуючи дані опитування киян у цьому ж році, з'ясуємо, чи відрізнявся середній місячний прибутку на одну особу в Києві (змінна v173) від 240 рублів.

Формулюємо гіпотезу, яка підлягає перевірці:

$$H_0 : \overline{v173} - 240 = 0 .$$

[18] Стандартно використовують один із двох рівнів значущості – або 0.05, або 0.01.

Для перевірки гіпотези використовуємо процедуру, що називається t-test (перевірка, яка ґрунтується на розподілі Стьюдента, тобто на t-розподілі) для однієї вибірки. Послідовно обираємо

Analyze ⇒ Compare Means ⇒ One-sample T Test...

(*Аналіз ⇒ Порівняння середніх ⇒ t-test для однієї вибірки*).

Відкриється вікно параметрів процедури **One-sample T Test** (див. рис. 9.3). У поле **Test Variable(s)** (*Змінні, що перевіряються*) заносимо ім'я змінної, для якої сформульована гіпотеза про рівність середнього (у нашому випадку це змінна v173), а в поле **Test Value** (*Контрольне значення*) записуємо значення, з яким порівнюємо середнє значення змінної (у нашому випадку це число 240). Потім натискаємо кнопку **[Ok]**.

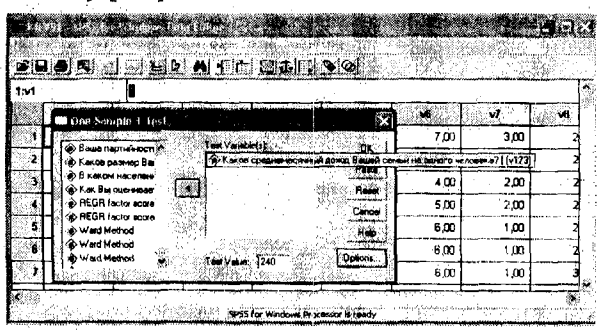


Рис. 9.3. *Вікно T Test-у для однієї вибірки.*

Результат обчислень зображений на рис. 9.4. У колонці *Sig. (2-tailed)* вказана ймовірність помилки першого роду (значущість) під час перевірки статистичної гіпотези про рівність середнього числа 240. Ця ймовірність є малою (дорівнює 0.001), значно меншою, ніж стандартний рівень 0.05, а отже, гіпотеза про рівність нулю різниці між середнім значенням змінної v173 та числом 240, може бути відкинута.

One-Sample Test

	Test Value = 240					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Каков среднемесячный доход Вашей семьи на одного человека? (	-3,474	417	,001	-17,7225	-27,7497	-7,6953

Рис. 9.4. *Результат застосування T Test-у для однієї вибірки.*

Більше того, ця різниця є негативною і дорівнює -17.72 (див. стовпчик *Mean Difference* на рис. 9.4) і 95 % довірчий інтервал для цієї різниці є таким [-27.75, -7.70]. Отже, на основі даних нашого опитування можна з ймовірністю 0.95 зробити висновок про те, що киянам у середньому не вистачає на одну особу від 7 руб. 70 коп. до 27 руб. 75 коп. Нагадаємо, що наш висновок ґрунтується на гіпотетичному (не реальному) значенні прожиткового мінімуму.

Перевірка статистичної гіпотези про рівність середніх значень певної змінної у двох незалежних вибірках здійснюється шляхом застосування процедури t-test для двох незалежних вибірок.

Приклад 2. Використовуючи дані опитування киян у 1991 р., перевіримо гіпотезу про рівність середньої заробітної плати чоловіків та жінок у місті Києві. Заробітна плата, виміряна у радянських рублях, закодована у змінній v174. Стать респондента закодована у змінній v165 (чоловіки закодовані числом 1, жінки – числом 2).

Розділимо спочатку нашу вибірку на дві частини (на чоловіків та жінок) і порівняємо середнє значення змінної v174 (заробітна плата) у цих двох групах. Будемо вважати, що ми маємо дві генеральні сукупності – чоловіки міста Києва та жінки міста Києва, а також те, що нами зроблені дві випадкові вибірки – одна з першої генеральної сукупності (частина наших даних, що стосується чоловіків), а друга – з другої генеральної сукупності (частина наших даних, що стосується жінок). Формулюємо гіпотезу про рівність двох генеральних середніх (тобто про те, що середня заробітна плата чоловіків не відрізняється в місті Києві від середньої заробітної плати жінок) і за допомогою процедури t-test для двох незалежних вибірок намагаємося спростувати (відкинути) цю гіпотезу.

Послідовно обираємо

Analyze ⇒ Compare Means ⇒ Independent-Samples T Test...

(*Аналіз ⇒ Порівняння середніх ⇒ t-test для незалежних вибірок*).

Відкривається вікно параметрів операції **Independent Samples T Test** (див. рис. 9.5).

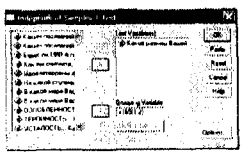


Рис. 9.5. *Вікно параметрів операції t-test для двох незалежних вибірок.*

У поле **Test Variable(s)** заносимо ім'я відповідної змінної (у цьому випадку це змінна v174, що містить розмір заробітної плати респондента). У поле **Grouping Variable:** (*Групувальна змінна*) заносимо змінну v165 (стать респондента), натискаємо кнопку **[Define Groups...]** (*Визначити групи*), яка відкриває вікно параметрів (див. рис. 9.6).

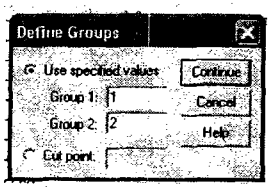


Рис. 9.6. *Вікно* ***Define Groups***

Є дві можливості виділити дві групи для порівняння. Перша можливість полягає в тому, що ми обираємо **Use specifies values** (*Використати відповідні значення*) та вказуємо два конкретні значення групувальної змінної (поля **Group 1** та **Group 2**). Об'єкти, у яких групувальна змінна дорівнює першому значенню, утворюють першу групу. Об'єкти, у яких ця змінна дорівнює другому значенню, утворюють відповідно другу групу для порівняння. Друга можливість полягає в тому, що ми обираємо **Cut point** (*Точка поділу*) та у відповідне поле вносимо значення, що і буде точкою поділу. Першу групу складають об'єкти, у яких значення групувальної змінної не менше ніж точка поділу. Другу групу, відповідно, об'єкти, у яких ця змінна має значення, що менше ніж точка поділу.

Ми обираємо **Use specified values**, заносимо в поле **Group 1** значення 1, а у поле **Group 2,** відповідно, значення 2. Отже, в нашому випадку першу групу складуть чоловіки, а другу – жінки. Натискаємо кнопку **[Continue]**, вікно визначення груп закривається. Натискаємо кнопку **[OK]**.

Результат перевірки гіпотези про рівність заробітної плати чоловіків та жінок у місті Києві у 1991 р. можна побачити на рис. 9.7.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Каков размер Вашей заработной платы (стипендии,пенсии)? (руб	Equal variances assumed	21,935	,000	7,063	406	,000	149,9377	21,2284	108,2064	191,6691
	Equal variances not assumed			6,480	219,083	,000	149,9377	23,1396	104,3330	195,5425

Рис. 9.7. *Результат порівняння середніх у двох незалежних вибірках.*

Зверніть увагу, що комп'ютерна програма перевіряє гіпотезу про рівність двічі – у припущенні про рівність у двох групах дисперсій змінної, що перевіряється (**Equal variances assumed**, верхній рядок таблиці), та без такого припущення (**Equal variances not assumed**, нижній рядок таблиці). Оскільки у нас немає відповідних припущень, ми використовуємо для прийняття рішення стосовно гіпотези нижній рядок таблиці.

У стовпчику *Sig. (2-tailed)* стоїть значущість – ймовірність помилки першого роду під час перевірки гіпотези про рівність середніх. Ця ймовірність менша ніж 0.001 (округлене значення дорівнює 0.000), а отже, гіпотеза про рівність може бути відкинута. Відмінність середніх у двох групах є позитивною (отже, в першій групі середнє більше, ніж у другій) і складає 149.94 у наших вибіркових даних, або ж характеризується 95 % довірчим інтервалом [104.33, 195.54].

Отже, на основі наших емпіричних даних можна з довірчою ймовірністю 0,95 зробити висновок, що середня заробітна плата у чоловіків є більшою, ніж у жінок, і ця відмінність є не меншою ніж 104 руб., але не більшою ніж 196 руб.

Зауважимо, що ми розглядали всю вибірку, не виділяли у вибірці працюючих респондентів, а отже, середня заробітна плата жінок виявилася дещо "заниженою" за рахунок жінок, що тимчасово не працюють і виховують вдома малих дітей. Крім того, зазначимо, що будувати довірчий інтервал для середнього можна не тільки для довірчої ймовірності 0.95. Для того, щоб змінити рівень довірчої ймовірності, потрібно скористатися кнопкою **[Options]**, розкрити відповідне вікно параметрів та ввести потрібне значення довірчої ймовірності.

Досить типовою є ситуація, коли кожен об'єкт характеризується двома вимірами, усереднені значення яких потрібно порівняти. В такому випадку доцільним є застосування так званої процедури порівняння середніх значень у двох зв'язаних (тобто залежних) вибірках.

Приклад 3. В опитуванні 1991 р. киянам пропонувалося визначити, кого вони готові допустити як представників тієї чи іншої національної групи. Пропонувалися варіанти відповіді, що характеризували різного розміру соціальні дистанції (від найменшої соціальної дистанції "членів власної родини" до найбільшої – "взагалі не пускати до міста Києва"). Ці варіанти були впорядковані (від наймен-

шої дистанції до найбільшої) і закодовані, відповідно, числами від 1 до 7. Зауважимо, що з цими даними ми вже працювали (див. приклад у розділі 7 пункту 7.1.1).

Розглянемо в такому контексті дві національні групи – росіян та білорусів. Оскільки кожному респонденту ставили обидва питання – і про росіян, і про білорусів, – то ми маємо дві змінних, які для кожного респондента містять його дистанцію до білорусів (змінна v73) та до росіян (змінна v74). Застосовуємо процедуру порівняння середніх у двох залежних вибірках. Послідовно обираємо

Analyze ⇒ Compare Means ⇒ Paired-Samples T Test...

(*Аналіз ⇒ Порівняння середніх ⇒ t-test для двох зв'язаних вибірок*).

Відкриється вікно параметрів **Paired-Samples T Test** (див. рис. 9.8). Виділяємо курсором миші та заносимо до списку **Pared variables** (*Об'єднані в пари змінні*) дві змінні – v73 та v74. Натискаємо кнопку [**OK**].

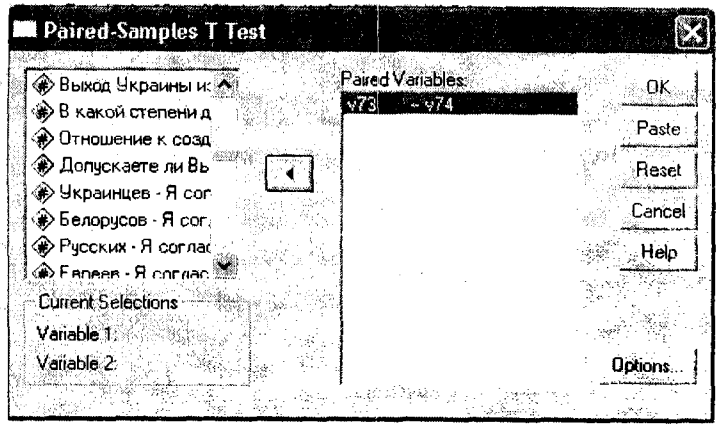


Рис. 9.8. *Вікно параметрів операції t-test для двох зв'язаних вибірок.*

Результат обчислення зображений на рис. 9.9. У стовпчику *Sig. (2-tailed)* стоїть значущість – ймовірність помилки першого роду під час перевірки гіпотези про рівність середніх. Ця ймовірність дорівнює 0.032 (менше ніж 0.05), а отже, гіпотеза про рівність соціальних дистанцій до білорусів та росіян може бути відкинута. Відмінність середніх у двох групах є позитивною (отже, у першій групі середнє більше, ніж у другій) і складає 0.133 у наших вибіркових даних, або ж характеризується 95 % довірчим інтервалом [0.011, 0.254].

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Белорусов - Я согласен допустить представителей данной нацио - Русских - Я согласен допустить представителей данной национа	,1328	1,2339	6,177E-02	1,139E-02	,2543	2,150	398	,032

Рис. 9.9. *Результат застосування процедури t-test для двох зв'язаних вибірок*

Отже, на основі наших емпіричних даних можна з довірчою ймовірністю 0.95 зробити висновок, що у киян в 1991 р. соціальна дистанція до білорусів була статистично значущо більшою, ніж соціальна дистанція до росіян, але ця відмінність була досить невеликою – не менше ніж 0.011 та не більше ніж 0.254 за загальною шкалою, що вимірює соціальну дистанцію від 1 до 7.

Необхідно зауважити, що статистична значущість не завжди означає змістовну важливість, а лише те, що емпіричні дані фіксують певну ненульову відмінність. Однак чи є ця відмінність настільки великою, щоб її можна було вважати змістовно важливою, – це інше питання, вирішення якого часто ґрунтується зовсім не на статистичних висновках.

У випадку, коли кількість груп, для яких обчислюються середні, є більшою ніж два, для перевірки гіпотези про рівність декількох середніх застосовують техніку однофакторного дисперсійного аналізу. Критерій для відкидання гіпотези про рівність кількох середніх пов'язаний із F-розподілом Фішера. Питання про відмінність між окремими груповими середніми вирішується за допомогою спеціальних методів множинного порівняння.

Приклад 4. Спробуємо на даних опитування киян у 1991 р. з'ясувати, чи впливає освіта на рівень заробітної плати. Маємо дві змінні – v167 (освіта, чотири освітні рівні) та v174 (заробітна плата респондента, виміряна в радянських рублях).

Якщо середня заробітна плата у всіх чотирьох освітніх групах є однаковою, то ми не можемо стверджувати, що освіта як фактор впливає на розмір заробітної плати. Для перевірки статистичної гіпотези про рівність чотирьох середніх скористаємося процедурою однофакторного дисперсійного аналізу. Послідовно обираємо

Analyze ⇒ Compare Means ⇒ One-Way ANOVA...

(*Аналіз ⇒ Порівняння середніх ⇒ Однофакторний дисперсійний аналіз*).

Розкривається вікно параметрів (див. рис. 9.10). До списку **Dependent List** (*Список залежних*) заносимо змінну v174 (заробітна плата), яка буде усереднюватися в межах груп за освітою. Ми маємо на меті перевірити вплив на заробітну плату такого фактору, як освіта. Тому в поле **Factor** (*Фактор*) вносимо ім'я змінної v167.

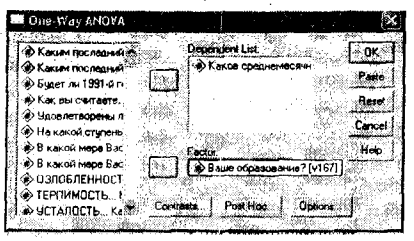


Рис. 9.10. *Вікно параметрів однофакторного дисперсійного аналізу* ***One-Way ANOVA.***

Результат застосування однофакторного дисперсійного аналізу зображений на рис. 9.11. У стовпчику *Sig.* зазначена значущість перевірки статистичної гіпотези про рівність чотирьох середніх. Оскільки вказане в цьому стовпчику значення значно менше 0.05, то ми можемо відкинути гіпотезу про рівність середньої зарплати у чотирьох освітніх групах, при цьому ймовірність припуститися помилки не тільки не перевищує 0.05, але й 0.01. Отже, і на рівні 0.05, і на рівні 0.01 ми можемо стверджувати, що чотири середні не є однаковими, а освіта як фактор впливає на розмір заробітної плати.

ANOVA

Каков размер Вашей заработной платы (стипендии,пенсии)? (руб)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1171230.815	3	390410.272	8.101	.000
Within Groups	19517917.630	405	48192.389		
Total	20689148.445	408			

Рис. 9.11. *Результат застосування однофакторного дисперсійного аналізу* ***ANOVA.***

З'ясувати питання про те, в яких власне освітніх групах спостерігається відмінність середніх, можна шляхом застосування одного з методів множинних порівнянь. Множинні порівняння застосовують лише тоді, коли F-тест у таблиці ANOVA є значущим. У випадку незначущого F-тесту всі середні не відрізняються між собою, а отже, здійснювати множинні порівняння немає сенсу.

Для того, щоб обрати один із методів множинних порівнянь, потрібно натиснути кнопку **[Post Hoc...]** і розкрити відповідне вікно параметрів (див. рис. 9.12). Пакет SPSS пропонує кілька методів множинного порівняння, об'єднаних у дві групи. Перша група включає методи, що застосовують у припущенні рівності дисперсій у всіх групах (**Equal Variances Assumed**). Друга група об'єднує методи, що не потребують такого припущення (**Equal Variances Not Assumed**).

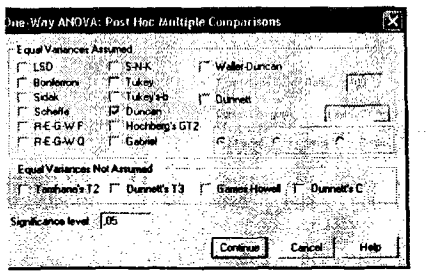


Рис. 9.12. *Вікно вибору методу множинного порівняння.*

Оберемо метод **Dunnett's T3** (він не грунтується на припущенні про рівність дисперсій у групах), залишимо рівень значущості для порівняння (**Significance Level**) рівним 0.05 і подивимося на результат такого порівняння, зображений на рис. 9.13.

Descriptives

Каков размер Вашей заработной платы (стипендии,пенсии)? (руб

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
высшее, незаконченное высшее (3-4 курса ВУЗа)	195	354.5128	198.12719	14.18818	326.5299	382.4957	.00	1500.00
среднее специальное, среднее профессионально-техническое (ПТ	123	249.7886	118.57627	10.69166	228.6234	270.9538	6.00	700.00
среднее общее (10-11 классов)	65	307.4462	395.63339	49.07228	209.4130	405.4793	90.00	3300.00
9 классов и меньше	26	194.4231	82.35444	16.15103	161.1594	227.6868	90.00	400.00
Total	409	305.3619	225.18592	11.13473	283.4733	327.2505	.00	3300.00

Multiple Comparisons

Dependent Variable: Каков размер Вашей заработной платы (стипендии,пенсии)? (руб

Dunnett T3

(I) Ваше образование?	(J) Ваше образование?	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
высшее, незаконченное высшее (3-4 курса ВУЗа)	среднее специальное, среднее профессионально-техническое (ПТ	104.72420*	17.76559	.000	57.6997	151.7487
	среднее общее (10-11 классов)	47.06667	51.08222	.927	-90.7944	184.9277
	9 классов и меньше	160.08974*	21.49792	.000	102.0284	218.1511
среднее специальное, среднее профессионально-техническое (ПТ	высшее, незаконченное высшее (3-4 курса ВУЗа)	-104.72420*	17.76559	.000	-151.7487	-57.6997
	среднее общее (10-11 классов)	-57.65754	50.22351	.821	-193.4402	78.1251
	9 классов и меньше	55.36554*	19.36924	.036	2.4088	108.3223
среднее общее (10-11 классов)	высшее, незаконченное высшее (3-4 курса ВУЗа)	-47.06667	51.08222	.927	-184.9277	90.7944
	среднее специальное, среднее профессионально-техническое (ПТ	57.65754	50.22351	.821	-78.1251	193.4402
	9 классов и меньше	113.02308	51.66183	.173	-26.3392	252.3854
9 классов и меньше	высшее, незаконченное высшее (3-4 курса ВУЗа)	-160.08974*	21.49792	.000	-218.1511	-102.0284
	среднее специальное, среднее профессионально-техническое (ПТ	-55.36554*	19.36924	.036	-108.3223	-2.4088
	среднее общее (10-11 классов)	-113.02308	51.66183	.173	-252.3854	26.3392

*. The mean difference is significant at the .05 level.

Рис. 9.13. *Результат множинного порівняння середніх значень заробітної плати у 4-х групах за освітою.*

Як видно з таблиць, наведених на рис. 9.13, група людей із середньою загальною освітою є дуже неоднорідною за заробітною платою. Стандартне відхилення значно перевищує середнє в цій групі. Це призводить до того, що довірчий інтервал для середнього в цій групі є дуже широким. Для довірчої ймовірності 0.95 у групі людей із загальною середньою освітою довірчий інтервал для заробітної плати складає приблизно від 209 до 405 руб. Друга таблиця на рис. 9.13 містить власне результати множинного порівняння. Відмінності на рівні 0.05 позначені в таблиці зірочкою. З таблиці видно, що три рівні освіти – "вища", "середня спеціальна" та "незакінчена середня" – впорядковані між собою за рівнем заробітної плати. Відмінності значення середньої заробітної плати в групах є значущими принаймні на рівні 0.05. Група людей із загальною середньою освітою є не дуже чисельною (65 респондентів) і характеризується великою неоднорідністю

заробітної плати всередині групи (стандартне відхилення навіть перевищує середнє). Це призводить до того, що довірчий інтервал для середнього значення заробітної плати в цій групі є настільки широким, що ця група статистично не відрізняється ні від однієї з трьох інших груп, виділених за освітою.

Контрольні запитання

1. Дайте визначення статистичного висновку.
2. Які різновиди статистичного висновку Ви знаєте?
3. Що є метою точкового статистичного оцінювання?
4. Сформулюйте мету інтервального статистичного оцінювання.
5. Як співвідносяться між собою довірчі інтервали для середнього, побудовані для ймовірностей 0.95 та 0.99?
6. Як впливає обсяг вибірки на ширину довірчого інтервалу для частки (відсотка)?
7. Які операції SPSS здійснюють перевірку статистичних гіпотез? Яких саме статистичних гіпотез?
8. Який висновок можна зробити з того, що відмінність між двома середніми є значущою?
9. Який висновок можна зробити з того, що коефіцієнт кореляції Пірсона є значущим?
10. Який висновок можна зробити з того, що коефіцієнт χ^2 є значущим?
11. Якщо коефіцієнт χ^2 є значущим на рівні 0.01, то який висновок можна зробити з цього про його значущість на рівні 0.05?
12. Якщо коефіцієнт кореляцій Пірсона є значущим на рівні 0.05, то який висновок можна зробити з цього про його значущість на рівні 0.01?
13. У чому полягає помилка першого роду під час перевірки статистичної гіпотези?

Практичні завдання

Працюємо із масивом Kiev91.

1. На основі аналізу розподілу змінної v37 чи можна сказати, що у 1991 р. більшість киян не схвалювала розвиток економіки у напрямку побудови конкурентного ринку товарів та послуг?
2. Побудуйте 95 % довірчий інтервал для відсотка таких киян, що вважають за необхідне встановити в країні фіксовані ціни на товари та послуги (змінна v49).

3. Побудуйте 99 % довірчий інтервал для середнього значення виміряної за шкалою Богардуса соціальної дистанції до росіян (змінна v74). Порівняйте цей інтервал з 99 % довірчим інтервалом для середнього значення соціальної дистанції до білорусів (змінна v73). Зробіть висновок про ставлення киян до росіян та про ставлення киян до білорусів у 1991 році.

4. Змінна v169 дозволяє виділити дві групи респондентів – тих, хто вважають рідною мовою українську (значення 1), та тих, хто вважають рідною мовою російську (значення 2). Порівняйте середній рівень довіри до Л. Кравчука (змінна v154) у цих двох групах. Зробіть висновок про зв'язок між рівнем довіри до цього політика та рідною мовою. Зробіть те ж саме для середньої довіри до В. Чорновола (змінна v161).

Зауваження. Зверніть увагу на кодування змінних v154 та v16, зокрема на коди 5 та 6.

5. Порівняйте середню довіру до Л. Кравчука (змінна v154) у чотирьох групах за освітою (змінна v167) та зробіть висновок про вплив освіти (як фактору) на довіру до Л. Кравчука. Зробіть те ж саме для середньої довіри до В. Чорновола (змінна v161, ознака 161).

6. Виконайте такі завдання:

1. Порівняйте рівень довіри до Л. Кравчука (змінна v154) із рівнем довіри до О. Мороза (v156). Зробіть висновок.

2. Порівняйте рівень довіри до Л. Кравчука (змінна v154) із рівнем довіри до Л. Плюща (v157). Зробіть висновок.

3. Порівняйте рівень довіри до О. Мороза (v156) із рівнем довіри до Л. Плюща (v157). Зробіть висновок.

Зробіть загальний висновок із цих трьох порівнянь.

РОЗДІЛ 10. ВИВЧЕННЯ ВПЛИВУ ФАКТОРІВ НА ЗАЛЕЖНУ ЗМІННУ. АНАЛІЗ ЛІНІЙНОЇ РЕГРЕСІЇ

Модель регресії передбачає наявність однієї залежної змінної та припущення про те, що одна або кілька незалежних змінних (їх ще часто називають факторами) мають безпосередній вплив на цю залежну змінну. Модель лінійної регресії передбачає, що зв'язок між залежною змінною та факторами є лінійним. Модель регресії не вимагає, щоб залежна змінна та фактори були виміряні в одних одиницях вимірювання, але вимагає, щоб залежна змінна була кількісною, а фактори – або кількісними, або ж дихотомічними змінними (такі дихотомічні змінні інколи ще називають фіктивними). Рівняння регресії використовують або для пояснення, або для прогнозування результату впливу факторів на залежну змінну.

У випадку одного фактору маємо справу із парною лінійною регресією. Для залежної змінної y та одного фактору x рівняння парної лінійної регресії має вигляд:

$$\bar{y} = bx + a,$$

де b – коефіцієнт регресії, а коефіцієнт a – зсув (константа рівняння). У геометричному виразі рівняння парної регресії описує пряму, яка характеризується тангенсом кута нахилу, що дорівнює b та відбиває на осі Y відрізок, довжиною a. Ця пряма проводиться так, щоб сума квадратів відхилень реальних точок від цієї прямої була мінімальною (метод найменших квадратів).

Зсув a, як правило, не інтерпретують. Коефіцієнт регресії b показує, на скільки зміниться середнє значення залежної змінної $\bar{y}$ при зміні фактору x на одну одиницю. Якщо позначити коефіцієнт кореляції між фактором x та залежною змінною y через r, то показник r^2 називається коефіцієнтом детермінації рівняння парної регресії й інтерпретується як частка дисперсії залежної змінної y, що пояснюється фактором x. Чим більшим є значення r^2, тим краще рівняння пояснює (або передбачає) поведінку залежної змінної y залежно від фактору x.

У випадку, якщо кількість факторів більше ніж один, маємо справу із множинною регресією, яка є узагальненням регресії парної. Рівняння множинної лінійної регресії таке:

$$\bar{y} = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + a,$$

де b_i – коефіцієнти регресії, а a – зсув (константа рівняння). Під час побудови рівняння множинної регресії також застосовують МНК (метод найменших квадратів).

Зсув не інтерпретують, а коефіцієнт регресії b_i показує, на скільки зміниться середнє значення залежної змінної $\bar{y}$ при зміні фактору x_i на одиницю та зафіксованих (незмінних) значеннях інших факторів. Або ж коефіцієнт регресії b_i інтерпретується як сила впливу фактору x_i на середнє значення залежної змінної $\bar{y}$. Інтерпретується не тільки значення, а й знак коефіцієнта регресії. Знак "плюс" біля коефіцієнта регресії b_i свідчить про позитивний вплив фактору x_i на середнє значення залежної змінної $\bar{y}$ (збільшення значення фактору приводить до збільшення значення залежної змінної, зменшення значення фактору, відповідно, до зменшення залежної змінної), а знак "мінус" – про негативний вплив (збільшення фактору веде до зменшення залежної змінної, а зменшення фактору – до збільшення залежної змінної).

Квадрат коефіцієнта множинної (сукупної) кореляції між залежною змінною y та факторами $x_1, x_2, \ldots x_n$ позначають R^2 і називають коефіцієнтом детермінації та інтерпретують як частку дисперсії залежної ознаки y, що пояснюється усіма факторами $x_1, x_2, \ldots x_n$ сукупно. Чим більшим є значення R^2, тим краще рівняння пояснює (або передбачає) поведінку залежної змінної y залежно від факторів $x_1, x_2, \ldots x_n$.

Перевірка гіпотези про рівність нулю коефіцієнта детермінації дає змогу з'ясувати питання про те, чи справляють хоч якийсь статистично значущий вплив на залежну змінну y фактори $x_1, x_2, \ldots x_n$. Або, інакше кажучи, чи пояснюють усі введені до моделі фактори $x_1, x_2, \ldots x_n$ хоч якусь, відмінну від нуля, частку дисперсії залежної змінної y. Якщо коефіцієнт детермінації статистично не відрізняється від нуля (тобто, якщо статистична гіпотеза про рівність коефіцієнта детермінації нулю не може бути відкинута на прийнятному для нас рівні значущості), то інтерпретувати окремі показники рівняння регресії немає сенсу.

Перевірка гіпотези про рівність нулю окремих коефіцієнтів регресії дозволяє з'ясувати, чи має вплив на залежну змінну кожний окремий фактор. Якщо коефіцієнт b_i не є значущим на потрібному рівні, то це інтерпретується як відсутність впливу фактору x_i. У більшості випадків такий фактор, з огляду на принцип економності, просто вилучається з моделі.

Значення коефіцієнта регресії b_i залежить від одиниць вимірювання відповідного фактору та залежної змінної. Це означає, що коефіцієнти регресії для окремих факторів немає сенсу порівнювати

за значенням із метою їх ранжування за силою впливу. Звільнитися від одиниць вимірювання в окремій змінній можна шляхом її стандартизації. Якщо стандартизувати всі змінні у регресійній моделі і для цих стандартизованих змінних побудувати методом МНК рівняння лінійної регресії, то ми отримаємо таке рівняння регресії у стандартних координатах:

$$\bar{Z}_y = \beta_1 Z_{x1} + \beta_2 Z_{x2} + \ldots + \beta_n Z_{xn}.$$

Важливим є те, що окремі коефіцієнти регресії β_i можна порівнювати за значенням із метою визначення, які фактори є потужнішими (мають більший вплив, більшу пояснюючу силу), а які є менш потужними.

Особливістю рівняння регресії у стандартних координатах є відсутність зсуву (константи). Стандартизація зберігає напрямок впливу фактору (знак у β_i є таким самим, як і у відповідного b_i) та не впливає на значення коефіцієнта детермінації R^2.

Комп'ютерна програма SPSS для вказаних змінних (одна залежна змінна та одна або декілька незалежних змінних) оцінює значення коефіцієнтів регресії та значення коефіцієнтів стандартизованого рівняння регресії, обчислює значення коефіцієнта детермінації, перевіряє гіпотезу про його рівність нулю та перевіряє гіпотези про рівність нулю для всіх коефіцієнтів рівняння регресії.

Для побудови рівняння множинної лінійної регресії послідовно обираємо:

Analyze ⇒ Regression ⇒ Linear...

(*Аналіз ⇒ Регресія ⇒ Лінійна...*).

Розкривається (див. рис. 10.1) вікно параметрів **Linear Regression** (*Лінійна регресія*). У цьому вікні необхідно:

- у полі **Dependent** (*Залежна*) ввести ім'я залежної змінної;
- у полі **Independent(s)** (*Незалежні*) сформувати список імен незалежних змінних (факторів);

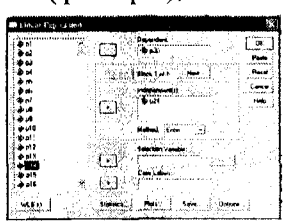


Рис. 10.1. *Вікно параметрів операції* ***Linear Regression.***

- зі списку **Method** (*Метод*) обрати метод автоматичного введення та виведення факторів до моделі. Програма пропонує кілька методів, серед яких найбільш часто вживаними є:
  - **Enter** (*Ввести*) – метод безпосереднього введення факторів до моделі; виконується побудова одного рівняння для всієї множини визначених факторів;
  - **Backward** (*Зворотний*) – метод покрокового видалення незначущих факторів із моделі; здійснюється побудова рівняння для всіх вказаних факторів, а потім виконується процедура покрокового видалення з моделі факторів, що мають малий вплив; видаляються ті фактори, які не впливають на розмір частки дисперсії залежної ознаки, що її пояснює модель;
  - **Stepwise** (*Покроковий*) – метод покрокового введення факторів до моделі; для всіх вказаних факторів перевіряється можливість їх введення до моделі; вводяться ті фактори, які значно збільшують частку дисперсії залежної ознаки, що її пояснює модель;
- натиснути кнопку **Statistics** (*Статистики*) та встановити позначку біля **Estimates** (*Оцінки*) у групі **Regression Coefficients** (*Коефіцієнти регресії*) для виведення значень коефіцієнтів регресії та зробити позначку **Model fit** (*Відповідність моделі*) для виведення коефіцієнта детермінації R^2.

Розглянемо процес побудови та аналізу лінійної регресії на конкретному прикладі. Для цього скористаємося результатами поштового опитування, проведеного під керівництвом проф. В. І. Паніотто у 1989 р. Відділенням соціології Інституту філософії НАН України в м. Києві.

Приклад. Розглянемо процес побудови та аналізу рівняння лінійної регресії на прикладі вивчення факторів, що впливають на задоволеність працею. Респондентам були поставлені питання про те, чи задоволені вони такими різними сторонами праці, як умови праці (змінна p21), зміст праці (змінна p22), режим праці (змінна p23), розмір заробітної плати (змінна p24), можливість підвищувати кваліфікацію (змінна p25), стосунками з колегами (змінна p26), стосунками з керівництвом (змінна p27) та віддаленістю роботи від житла (змінна p28). Крім того, поставлено питання про задоволеність роботою в цілому (змінна p30). Для всіх зазначених питань пропонувалося обрати один із таких варіантів відповіді: "зовсім ні" (значення 1), "скоріше ні, ніж так" (значення 2), "настільки так, наскільки і ні" (значення 3), "скоріше так, ніж ні" (значення 4), "повністю" (значення 5) та "не має значення" (значення 0). Визначимо значення 0 як відсутнє (user missing value) та будемо розглядати змінні p21–p28 та p30 як кількісні, виміряні в балах, для яких можна обчислювати коефіцієнт кореляції Пірсона та застосовувати регресійний аналіз.

У рамках регресійної моделі розглянемо вплив задоволеності різними сторонами праці (p21–p28) на задоволеність роботою в цілому (p30). Розпочнемо з парної регресії. Можна припустити, що найбільший вплив на задоволеність роботою має задоволеність розміром заробітної плати (матеріальний фактор). Будуємо рівняння парної регресії із залежною змінною p30 та одним фактором p24. Результат обчислень зображений на рис. 65.

Спочатку дивимось у стовпчик *Sign.* таблиці ANOVA. Число, що у цьому стовпчику, означає ймовірність помилки першого роду під час перевірки статистичної гіпотези про рівність нулю коефіцієнта детермінації побудованого рівняння. Перевірка здійснюється за допомогою F-критерію Фішера. Оскільки F-критерій є значущим (рівень значущості менше 0.05 і навіть менше 0.01), то фактор p24 пояснює певну (відмінну від нуля) частку дисперсії залежної змінної p30, а отже, рівняння можна інтерпретувати.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,388[a]	,151	,149	1,0400

a. Predictors: (Constant), Задоволеність розміром заробітної плати

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	102.731	1	102.731	94.987	.000[a]
	Residual	578.620	535	1.082		
	Total	681.352	536			

a. Predictors: (Constant), Задоволеність розміром заробітної плати

b. Dependent Variable: Задоволеність роботою в цілому

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,833	,094		30,008	,000
	Задоволеність розміром заробітної плати	,338	,035	,388	9,746	,000

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 10.2. *Результати побудови рівняння парної регресії для залежної змінної p30 (задоволеність роботою в цілому) та одного фактору p24 (задоволеність розміром заробітної плати).*

Далі дивимось у таблицю Model Summary (*Коротка інформація про модель*). Коефіцієнт множинної кореляції між усіма факторами та залежною змінною *R* дорівнює 0.388. Оскільки у нас множина факторів складається із одного фактора (ми працюємо з парною регресією), то в нашому випадку коефіцієнт множинної кореляції є, по суті, звичайним коефіцієнтом парної кореляції Пірсона між фактором та залежною змінною. Для коефіцієнта детермінації (*R Square*) маємо $R^2 = 0.151$. Це означає, що такий фактор, як задоволеність розміром заробітної плати (змінна p24) пояснює приблизно 15 % варіативності задоволеності роботою в цілому (змінна p30). Для одного фактору це – непоганий результат.

Необхідно зазначити, що пояснення дисперсії є одним із важливих завдань аналізу даних. Нас цікавить питання про те, чому люди не однаково задоволені своєю роботою. Що впливає на те, що одні люди більше, а інші менше задоволені своєю роботою? Ми висунули припущення, що саме задоволеність заробітною платою є одним із джерел варіативності задоволеності роботою в цілому. Ми побудували модель регресії з одним фактором і бачимо, що сама по собі задоволеність заробітною платою пояснює приблизно 15 % варіативності (дисперсії) задоволеності роботою в цілому. Приблизно на 85 % дисперсія задоволеності роботою в цілому залежить не від задоволеності заробітною платою, а від інших факторів, які не увійшли до нашої моделі.

Звертаємось до таблиці **Coefficients** (*Коефіцієнти*), яка власне і містить побудоване нами рівняння. У стовпчику **Unstandardized Coefficients B** (*Нестандартизовані коефіцієнти B*) записані коефіцієнти нестандартизованого рівняння (або, як ще інколи кажуть,

рівняння регресії у первинних координатах). Це рівняння можна записати так:

$$p_{30} = 0.338 p_{24} + 2.833.$$

У стовпчику **Standardized Coefficients Beta** (*Стандартизовані коефіцієнти Beta*) містяться коефіцієнти стандартизованого рівняння регресії. Згадаємо, що у стандартизованому рівнянні немає зсуву, тому відповідна клітина в таблиці є пустою. Коефіцієнт β для фактору p24 збігається за значенням із коефіцієнтом кореляції між фактором та залежною змінною (стовпчик *R* таблиці **Model Summary**), і це не є випадковістю. Для випадку парної регресії коефіцієнт парної кореляції і є коефіцієнтом стандартизованого рівняння регресії.

Важлива інформація міститься у стовпчику *Sign.* таблиці **Coefficients**. Тут записані значущості коефіцієнтів рівняння регресії (ймовірність помилки першого роду під час перевірки статистичної гіпотези про рівність нулю для кожного з коефіцієнтів рівняння). Оскільки коефіцієнт регресії є значущим, то фактор p24 дійсно впливає на залежну змінну p30, і відповідний коефіцієнт (його знак та значення) може інтерпретуватися. Отже, задоволеність рівнем заробітної плати має позитивний вплив на задоволеність роботою в цілому (збільшення задоволеності заробітною платою приводить до збільшення задоволеності роботою в цілому, а зменшення – відповідно, до зменшення) і збільшення задоволеності заробітною платою на один пункт приводить у середньому до збільшення задоволеності роботою в цілому на 0.338 пункти.

Тепер спробуємо протестувати інший фактор. Перевіримо, як впливає на задоволеність роботою в цілому (змінна p30) задоволеність умовами праці (змінна p21). Будуємо відповідне рівняння парної регресії. Таблиці **Model Summary** та **Coefficients** для цієї моделі зображені на рис. 10.3.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,417[a]	,174	,172	1,0327

a. Predictors: (Constant), Задоволеність умовами праці

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,496	,116		21,439	,000
	Задоволеність умовами праці	,361	,034	,417	10,674	,000

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 10.3. *Результати побудови рівняння парної регресії для залежної змінної p30 (задоволеність роботою в цілому) та одного фактору p21 (задоволеність умовами праці).*

Можна побачити, що задоволеність умовами позитивно впливає на задоволеність роботою в цілому (коефіцієнт регресії є значущим та додатнім) і рівняння регресії можна записати так

$$p_{30} = 0.361 p_{21} + 2.496 .$$

Задоволеність умовами праці як фактор описує приблизно 17,4 % дисперсії задоволеності роботою в цілому (коефіцієнт детермінації $R^2 = 0.174$). Отже, задоволеність умовами праці є потужнішим фактором (має більший вплив, описує більше дисперсії), ніж задоволеність заробітною платою.

Проте в місті Києві у 1989 р. структура робочих місць була такою, що ті, хто мав кращі умови праці, до того ж отримували й кращу заробітну плату. Інакше кажучи, є аргументи на користь припущення про те, що задоволеність заробітною платою (p24) та задоволеність умовами праці (p21) позитивно корелюють між собою. Це означає, що показник сили впливу p24 на p30 у відповідному рівнянні парної регресії частково відображає і вплив p21 на p30. З іншого боку, і показник впливу p21 на p30 частково відображає і вплив p24 на p30. Для того, щоб розглянути вплив p21 на p30 із видаленням впливу p24, потрібно здійснити статистичне контролювання p24 при оцінці впливу p21 на p30. Так само, для того, щоб розглянути вплив p24 на p30 із видаленням впливу p21, потрібно здійснити статистичне контролювання p21 при оцінці впливу p24 на p30. Для цього ми розглянемо регресійну модель із двома факторами – p24 та p21. Результати відповідних обчислень зображені на рис. 10.4.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,484[a]	,234	,231	,9935

a. Predictors: (Constant), Задоволеність розміром заробітної плати, Задоволеність умовами праці

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,216	,122		18,175	,000
	Задоволеність умовами праці	,272	,036	,313	7,528	,000
	Задоволеність розміром заробітної плати	,232	,036	,266	6,402	,000

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 10.4. *Результати побудови рівняння регресії для залежної змінної p30 та двох факторів p21 і p24.*

Звичайно, для моделі з двома факторами частка поясненої дисперсії залежної змінної не дорівнює сумі відповідних часток у двох відповідних рівняннях парної регресії, але ця частка і тепер уже два фактори пояснюють трохи більше ніж 23 % дисперсії залежної змінної. Взагалі потрібно зазначити, що введення нових факторів до моделі не погіршує пояснювальну силу рівняння (не зменшує коефіцієнт детермінації). Введення нового фактору до моделі може або не покращити модель (у випадку, якщо фактор ніяк не впливає на залежну змінну), або ж покращити її, але ніколи не зменшує пояснювальну силу моделі.

Обидва фактори в моделі є значущими. Рівняння регресії у первинних координатах можна записати так:

$$p_{30} = 0.272 p_{21} + 0.232 p_{24} + 2.216 ,$$

а відповідне рівняння у стандартизованих координатах:

$$p_{30}' = 0.313 p_{21}' + 0.266 p_{24}' .$$

Порівнюючи між собою стандартизовані коефіцієнти, можна зробити висновок, що в межах застосованої нами моделі лінійної регресії з двома факторами задоволеність умовами праці (змінна p21) має більший вплив на задоволеність роботою в цілому (змінна p30), ніж задоволеність заробітною платою (змінна p24).

Це доволі-таки цікавий сам по собі результат, який свідчить про те, що заробітна плата (у Києві в 1989 році) є в певному сенсі менш важливою, ніж умови праці. Тепер спробуємо збагатити нашу модель і введемо в рівняння всі визначені нами попередньо потенційні фактори впливу, а саме: змінні від p21 до p28. Результат оцінки такої моделі на наших емпіричних даних зображений на рис. 10.5.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,672[a]	,452	,439	,8355

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,576	,287		2,007	,046
	Задоволеність умовами праці	,117	,041	,133	2,841	,005
	Задоволеність змістом праці	,371	,046	,396	8,124	,000
	Задоволеність режимом праці	4,466E-02	,038	,052	1,184	,237
	Задоволеність розміром заробітної плати	,106	,041	,122	2,622	,009
	Задоволеність можливістю підвищувати кваліфікацію	,110	,037	,143	2,953	,003
	Задоволеність стосунками з колегами	,109	,065	,078	1,665	,097
	Задоволеність стосунками з керівництвом	3,247E-02	,046	,034	,704	,482
	Задоволеність віддаленістю роботи від житла	5,032E-03	,031	,007	,163	,870

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 68. *Результати побудови рівняння регресії для залежної змінної p30 та множини потенційних факторів p21...p28.*

Запишемо відразу рівняння регресії в стандартизованих координатах:

$$p_{30}' = 0.133p_{21}' + 0.396p_{22}' + 0.052p_{23}' + 0.122p_{24}' + 0.143p_{25}' + 0.078p_{26}' + 0.034p_{27}' + 0.007p_{28}'.$$

Вісім факторів описують досить значну частку дисперсії залежної ознаки – більш ніж 45 %.

Отже, всі разом вісім факторів досить істотно впливають на залежну змінну. Проте залишаються нез'ясованими питання про те, яким є вплив кожного конкретного фактору. Адже можливою є ситуація, коли вісім факторів мають вплив, але серед цих восьми реально впливають лише, наприклад, три, а інші п'ять впливу не мають, а отже, без особливих втрат можуть бути виключені з моделі. Принцип економності, про який ми вже згадували раніше, вимагає, щоб якість моделі досягалася якомога меншою кількістю засобів. У нашому випадку це означає, що всі фактори, які не мають значущого впливу на залежну змінну, не тільки можуть, а й мають бути видалені з моделі. Адже з двох моделей більш проста (з меншою кількістю факторів) є кращою, звичайно, за умови, що обидві моделі мають однакову або майже однакову пояснювальну силу (близькі між собою значення коефіцієнтів детермінації). З більш простою моделлю легше працювати, її легше використовувати тощо.

Для з'ясування значущості впливу кожного окремого фактору звертаємося до стовпчика *Sign.* таблиці коефіцієнтів рівняння регресії **Coefficients**. Бачимо, що фактори p28 (задоволеність віддаленістю роботи від житла), p27 (задоволеність стосунками з керівництвом) та p23 (задоволеність режимом праці) є незначущими. Незначущий, але досить близький до визначеного нами порогу значущості (а це число 0.05) фактор p26 (стосунки з колегами).

Отже, ми маємо всі підстави для того, щоб спростити модель. Видалення незначущих факторів починаємо із p28 (задоволеність віддаленістю роботи від житла). Видаляємо цей фактор і перебудовуємо модель. Результат такого спрощення зображено на рис. 10.6.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,666[a]	,443	,432	,8341

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,648	,268		2,414	,016
	Задоволеність умовами праці	,106	,040	,123	2,671	,008
	Задоволеність змістом праці	,366	,044	,394	8,374	,000
	Задоволеність режимом праці	5,000E-02	,037	,059	1,350	,178
	Задоволеність розміром заробітної плати	,109	,040	,126	2,723	,007
	Задоволеність можливістю підвищувати кваліфікацію	,112	,036	,146	3,089	,002
	Задоволеність істосунками з колегами	9,279E-02	,064	,067	1,452	,147
	Задоволеність стосунками з керівництвом	4,081E-02	,045	,044	,916	,360

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 10.6. *Результат видалення фактору p28 із загальної моделі.*

Матимемо таке рівняння регресії в стандартизованих координатах (див. рис. 10.6):

$$p_{30}' = 0.123p_{21}' + 0.394p_{22}' + 0.059p_{23}' + 0.126p_{24}' + 0.146p_{25}' + 0.067p_{26}' + 0.044p_{27}'.$$

Після видалення фактору p28 модель майже не змінила свою пояснювальну силу: частка дисперсії зменшилася (з 45,2 % до 44,3 %), але незначно. У нас все одно залишилися незначущі коефіцієнти (біля p27 та p23). Проводимо покрокове видалення цих двох факторів і отримуємо модель, що включає п'ять факторів (див. рис. 10.7).

У цьому рівнянні всі фактори мають статистично значущий вплив. Рівняння пояснює приблизно 44 % дисперсії задоволеності роботою в цілому. В стандартизованих координатах остаточно побудоване нами рівняння (економна, оптимальна модель) має такий вигляд:

$$p_{30}' = 0.125p_{21}' + 0.421p_{22}' + 0.136p_{24}' + 0.155p_{25}' + 0.093p_{26}'.$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,663[a]	,439	,432	,8290

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,733	,256		2,861	,004
	Задоволеність умовами праці	,107	,037	,125	2,859	,004
	Задоволеність змістом праці	,391	,041	,421	9,468	,000
	Задоволеність розміром заробітної плати	,116	,038	,136	3,021	,003
	Задоволеність можливістю підвищувати кваліфікацію	,118	,035	,155	3,351	,001
	Задоволеність стосунками з колегами	,126	,055	,093	2,300	,022

a. Dependent Variable: Задоволеність роботою в цілому

Рис. 10.7. *Результат видалення із загальної моделі факторів p28, p27 та p23*

Порівнюємо значення коефіцієнтів рівняння регресії в стандартизованих координатах та впорядковуємо фактори за силою їх впливу на залежну змінну: $p_{30}' = 0.421p_{22}' + 0.155p_{25}' + 0.136p_{24}' + 0.125p_{21}' + 0.093p_{26}'$.

Найбільший вплив на задоволеність роботою в цілому (p30) має задоволеність змістом праці (p22), потім задоволеність можливістю підвищувати кваліфікацію (p25), лише на третьому місці – задоволеність розміром заробітної плати (p24), на четвертому – задоволеність умовами праці (p21) і на останньому п'ятому – задоволеність стосунками з колегами (p26). Раніше ми вже визначили, що такі фактори, як задоволеність віддаленістю роботи від житла (p28), задоволеність стосунками з керівництвом (p27) та режимом праці (p23) статистично значущого впливу на задоволеність роботою в цілому в межах регресійної моделі не мають.

Отже, рішення про те, вводити чи виводити той чи інший фактор до моделі, ґрунтується на результатах перевірки значущості цього фактору, на результатах з'ясування того, чи значущим є вплив введення (виведення) фактору з моделі на оцінку якості моделі, на значення коефіцієнта детермінації R^2. Ця процедура може бути автоматизована. Відповідно, SPSS має кілька методів автоматичного введення та/або виведення факторів до моделі регресії, які відповідають тим чи іншим правилам прийняття рішень. Зокрема, найбільш часто вживаними є методи **Stepwise** (покрокове введення) та

Backward (покрокове видалення), про які ми вже згадували раніше. У більшості випадків обидва методи дають один і той самий результат. Однак є випадки, коли результати роботи цих двох методів дещо відрізняються (множина значущих факторів, відібрана одним методом, може дещо відрізнятися від факторів, відібраних другим методом). За будь-яких обставин дослідник не повинен сліпо покладатися на результати роботи автоматичних процедур і має співвідносити ці результати зі своїми теоретичними уявленнями.

Контрольні запитання

1. За якою шкалою має бути виміряна залежна змінна у моделі лінійної регресії?
2. У чому полягають основні завдання та мета регресійного аналізу?
3. Опишіть коротко основні етапи побудови рівняння регресії.
4. Як оцінюється якість побудованого рівняння регресії?
5. З якою метою будують стандартизоване рівняння регресії?
6. У чому полягає принцип економності у застосуванні його до побудови моделі лінійної регресії?

Практичні завдання

Працюємо із масивом Kiev89.

1. Здійсніть із множини потенційних факторів p21...p28 відбір значущих факторів різними методами автоматичного введення/виведення факторів до моделі (принаймні методами **Stepwise** та **Backward**) і порівняйте результати роботи різних методів.

2. В анкеті опитування міститься блок питань, пов'язаних із задоволеністю різними сторонами життя (група змінних від p30 до p52), та питання про задоволеність життям у цілому (змінна p53). Побудуйте рівняння регресії для задоволеності життям у цілому (змінна p53) від не більш ніж 4-х факторів (потенційні фактори – змінні p30...p52).

Зауваження. Вводити фактори до рівняння потрібно методом **Enter** (методом безпосереднього введення). Вибір остаточного рівняння потрібно здійснювати на основі значення коефіцієнта детермінації.

СПИСОК ЛІТЕРАТУРИ

1. Бююль Ахим, Цефель Петер. SPSS: искусство обработки информации. Platinum Edition: Пер. с нем. / Ахим Бююль, Петер Цефель.– СПб.: ООО "ДиаСофтЮП", 2005.– 608 с.

2. Калинин С. И. Компьютерная обработка данных для психологов / Под науч. ред. А. Л. Тулупьева.– Изд. 2-е.– СПб.: Речь; М., 2004.– 134 с.

3. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках.– СПб.: Питер, 2005. – 416 с.

4. Сидоренко Е. В. Методы математической обработки в психологии.– СПб., 2000.

5. Тюрин Ю. Н., Макаров А. А. Анализ данных на компьютере / Под ред. В.Э. Фигурнова.– 3-е изд., перераб. и доп.– М., 2003.– 544 с.

Навчально-методичне видання

Горбачик Андрій Петрович
Сальнікова Світлана Анатоліївна

АНАЛІЗ ДАНИХ СОЦІОЛОГІЧНИХ ДОСЛІДЖЕНЬ ЗАСОБАМИ SPSS

Навчально-методичний посібник

Редактор і коректор *Н. Я. Ярмольчук*
Верстка *Л. М. Козлюк*

Підп. до друку 08.04.2008. Формат 60x84 1/16. Папір офс. Гарн. Таймс. Друк цифровий. Обсяг 9,53 ум. друк. арк., 6,4 обл.-вид. арк. Наклад 300 пр. Зам. 2024. Редакційно-видавничий відділ "Вежа" Волинського національного університету ім. Лесі Українки (43025 м. Луцьк, просп. Волі, 13). Друк – РВВ "Вежа" ВНУ ім. Лесі Українки (43025 м. Луцьк, просп. Волі, 13). Свідоцтво Держ. комітету телебачення та радіомовлення України ДК № 3156 від 04.04.2008 р.