

Конспект лекцій, література, питання для самоконтролю

Зміст

Тема 1. Вступ до методів аналізу та комп'ютерної обробки соціологічної інформації	3
1.1. Соціальна інформація, соціологічна інформація та емпіричні дані	3
1.2. Вторинний аналіз: сутність та актуальність в сучасному комп'ютеризованому світі	4
1.3. Місце і роль процедур обробки та аналізу даних у соціологічному дослідженні	8
1.4. Поняття статистики та статистичного аналізу	11
Література до теми	13
Питання для самоконтролю	14
Тема 2. Аналіз одновимірних розподілів	15
2.1. Вигляд одновимірних розподілів та їх побудова в SPSS	15
2.2. Статистики одновимірного розподілу	21
2.3. Візуалізація одновимірних розподілів	25
2.4. Приклади застосування статистик одновимірних розподілів у вирішенні завдань соціологічного аналізу	29
Література до теми	35
Питання для самоконтролю	36
Тема 3. Відбір даних в SPSS: побудова фільтрів	37
3.1. Фільтри та відбір даних	37
3.2. Способи відбору даних в SPSS	38
3.3. Відбір анкет за визначеною умовою	40
3.4. Витяг випадкової вибірки спостережень з файлу даних	43
Література до теми	45
Питання для самоконтролю	45
Тема 4. Модифікація даних в SPSS: створення нових змінних	47
4.1. Створення нової змінної на основі однієї ознаки	47
4.2. Створення нової змінної на основі кількох ознак	50
4.4. Підрахунок зустрічальності значень у спостереженнях в SPSS	53
Література до теми	55
Питання для самоконтролю	55
Тема 5. Кореляційний аналіз та двовимірні розподіли	56
5.1. Кореляційний аналіз та кореляційна залежність	56
5.2. Аналіз двовимірних розподілів	58
5.3. Візуалізація двовимірних розподілів	66
Література до теми	70
Питання для самоконтролю	70
Тема 6. Статистичні висновки: статистичне оцінювання та перевірка гіпотез	72
6.1. Статистичний виведення і статистичні висновки	72
6.2. Статистичне оцінювання параметрів генеральних сукупностей	76
6.3. Довірчий інтервал для середнього	77
6.4. Довірчий інтервал для частки (відсотку)	82
6.5. Розрахунок довірчих інтервалів в SPSS	84

6.6. Приклади побудови довірчих інтервалів в SPSS та вручну	89
6.7. Статистична перевірка гіпотез	94
Література до теми	101
Питання для самоконтролю	101
Тема 7. Аналіз розбіжностей	103
7.1. Аналіз розбіжностей та статистична значущість розбіжностей	103
7.2. Т-тести	106
7.3. Дисперсійний аналіз	114
7.4. Непараметричні тести	123
7.5. Аналіз розбіжностей відсотків (часток)	129
Література до теми	135
Питання для самоконтролю	136
Тема 8. Факторний аналіз	138
8.1. Сутність та формальна модель факторного аналізу	138
8.2. Порядок виконання процедури факторного аналізу на прикладі пошуку чинників моральних преференцій	142
Література до теми	154
Питання для самоконтролю	154
Тема 9. Кластерний аналіз	155
9.1. Сутність кластерного аналізу	155
9.2. Реалізація ієрархічного кластерного аналізу в пакеті SPSS	158
9.3. Реалізація неієрархічного кластерного аналізу в пакеті SPSS	163
9.4. Приклад застосування кластерного аналізу в соціологічному дослідженні	167
Література до теми	175
Питання для самоконтролю	175
Тема 10. Регресійний аналіз та його застосування у соціологічних дослідженнях	176
10.1. Сутність регресійного аналізу	176
10.2. Приклад розрахунку множинної лінійної регресії в пакеті SPSS	178
10.3. Приклад застосування регресійного аналізу: визначення факторів довіри до політичних інститутів в Україні	182
10.4. Проблеми, пов'язані з можливістю застосування регресійного аналізу	184
Література до теми	185
Питання для самоконтролю	186

Тема 1. Вступ до методів аналізу та комп'ютерної обробки соціологічної інформації

1.1. Соціальна інформація, соціологічна інформація та емпіричні дані

Соціальна інформація – це будь-яка інформація, що створюється в суспільстві, тобто сукупність відомостей, знань, даних, які формуються в суспільстві та використовуються індивідами, соціальними групами, організаціями для регулювання соціальної взаємодії, суспільних відносин та процесів. Соціологічна інформація є різновидом соціальної інформації. У Законі України «Про інформацію» наведено таке визначення: «Соціологічна інформація - будь-які документовані відомості про ставлення до окремих осіб, подій, явищ, процесів, фактів тощо» (Стаття 19). Основними джерелами соціологічної інформації є задокументовані або публічно оголошені відомості, у яких відображені результати соціологічних опитувань, спостережень та досліджень. Отже, соціологічною інформацією є будь-які емпіричні дані, що містять інформацію про соціальну реальність: соціальні явища, соціальні процеси, соціальні спільноти, соціальні інститути, соціальні системи, соціальні групи та інші соціальні феномени.

Треба підкреслити, що терміни «соціологічна інформація» та «емпіричні дані» дуже близькі за змістом, проте вони не є взаємозамінними, оскільки поняття соціологічної інформації є більш широким, ніж поняття «емпіричні дані», включаючи до себе також інформацію більш високого рівня: теоретичні концепції, висновки й положення, які виступають у вигляді повідомлень, відомостей і застосовуються людьми в практичній діяльності. Таким чином, соціологічна інформація поєднує у собі *емпіричні дані* (як первинні, так й вторинні), *висновки й практичні рекомендації*, зроблені на їх основі, а також *теоретичні концепції та положення*, що перевірялися й уточнювалися на основі отриманого емпіричного матеріалу.

Звичайно соціологічну інформацію поділяють на певні різновиди за такими **критеріями**: 1) рівень узагальнення; 2) тип емпіричних даних; 3) первинність/вторинність.

За рівнем узагальнення соціологічна інформація поділяється на *емпіричні дані, висновки й практичні рекомендації, теоретичні концепції та положення*.

За типом емпіричних даних соціологічна інформація поділяється на числову та текстову. Як відомо емпіричні дані, що отримують в результаті соціологічних досліджень, можуть бути як *числовими*¹ (результати кількісних

¹ Ми спеціально застосовуємо термін «числові масиви», уникаючи словосполучень «кількісний масив», «кількісні дані», оскільки навіть дані, представлені у вигляді чисел не завжди є кількісними. В цьому контексті слід згадати шкали вимірювання, а також те, що дійсно кількісними даними є тільки ті дані, які виміряні за допомогою метричних або інтервальних шкал.

досліджень), так й *текстовими* (результати якісних досліджень). Треба підкреслити, що до самого останнього часу соціологи використовували тільки числові або текстові масиви в якості емпіричних даних. Числові масиви (квантифіковані думки респондентів) застосовувались в кількісних дослідженнях, текстові – у якісних. При цьому треба зазначити, що сьогодні йде мова про можливість застосування графічних даних (зокрема фотографій), web-контенту, мультимедіа тощо у якості емпіричних даних при проведенні соціологічних досліджень. Наразі активно розробляються відповідні нові методи аналізу таких, поки що незвичних, даних, адекватні як самим даним, так й цілям соціологічного аналізу. Мова йде про так звані цифрові методи, які виникли у зв'язку з появою «великих даних».

На сьогодні дуже актуальним є відрізнення первинної та вторинної соціологічної інформації. ***Первинна соціологічна інформація*** – дані, зібрані соціологом у результаті проведення соціологічного дослідження з метою вивчення конкретного соціального феномену, для вирішення актуальної на даний момент проблеми. Такі дані одержують у процесі анкетування, проведення інтерв'ю, спостереження або іншими способами в залежності від обраної дослідником методології. Вторинна інформація – дані, зібрані раніше для вирішення завдань, відмінних від вирішуваної в даний момент проблеми. Наприклад, дані державної статистики; результати виборів; публікації у ЗМІ; дані, одержані іншими дослідниками, можуть застосовуватися соціологом у власному дослідженні.

Вторинна соціологічна інформація є основою розвитку вторинного аналізу, який стає все більш актуальним засобом отримання соціологічного знання. Його популярність обумовлена тим, що завдяки застосуванню вторинних даних дослідник має можливість уникнути труднощів та матеріальних витрат, пов'язаних зі збором первинної соціологічної інформації.

Інформаційна цінність вторинних даних зростає при повторному застосуванні (в інших дослідницьких контекстах та у сполученні з даними інших емпіричних досліджень), а коректний синтез первинних даних власного дослідження з аналізом вторинної інформації дає можливість вийти на більш високий рівень узагальнень та висновків, одержати нетривіальні теоретичні результати.

Розвиток банків соціологічної інформації, які дозволяють одержати доступ до результатів аналізу й до вихідних масивів даних (на основі яких цей аналіз проводився), відкриває доступ до вторинних даних широкому колу дослідників, завдяки чому створюються передумови для більш комплексного та доцільного використання соціологічної інформації.

1.2. Вторинний аналіз: сутність та актуальність в сучасному комп'ютеризованому світі

Під вторинним аналізом розуміють сукупність методів і засобів отримання нового знання, що характеризуються наступними ознаками:

1. Це методи, які використовуються в разі, коли дослідники відмовляються від проведення спеціально організованого емпіричного дослідження, збору нового емпіричного матеріалу і задовольняються «старою» інформацією з раніше проведених досліджень.

2. На відміну від первинного аналізу, вторинний аналітик ставить перед собою нові дослідницькі цілі та завдання, які не були артикульовані при зборі первинних даних.

3. Основним етапом, який вирізняє первинний аналіз від вторинного є етап формування «інформаційного поля» на базі первинної інформації. Саме на цьому етапі проявляється специфіка вторинного аналізу, що дає підставу віднести його до того чи іншого типу.

4. Як правило, постановка нових дослідницьких завдань здійснюється іншими авторами, тобто зміна авторства є однією з ознак вторинного аналізу.

Вторинний аналіз – це аналіз результатів раніше проведених соціологічних досліджень, що переслідує цілі, відмінні від тих, які ставилися в цих дослідженнях.

Відповідно до тієї інформації, яка є про проведені раніше дослідженнях, можна виділити вторинний аналіз, що ґрунтується на публікаціях за підсумками досліджень, і вторинний аналіз на базі безпосередньо первинних даних раніше проведених досліджень. Другий різновид вторинного аналізу передбачає застосування більш складних методів статистичного аналізу з метою перегляду попередньо проаналізованих дослідних даних під новим кутом зору, що має розкрити нові грані аналізованого феномену.

Переваги використання вторинного аналізу полягають у його відносній дешевизні, бо досліднику не потрібно проводити власне дослідження та збирати дані, а також в можливості проведення лонгitudного, історичного або кроскультурного аналізу. Головною незручністю вторинного аналізу є те, що дослідник не має контролю над конструюванням змінних і часто обмежується лише знанням способу та обставин збору цих даних.

Вторинний аналіз дозволяє вирішувати ряд різноманітних завдань: порівняння результатів декількох досліджень, присвячених вивченню одного предмета, але проведених на різних об'єктах з метою виявлення специфіки того або іншого процесу в різних соціальних групах; агрегація результатів, отриманих при вивченні окремих соціальних спільнот, для виявлення характеристик більших спільнот; вивчення тимчасової динаміки соціальних процесів на основі використання матеріалів досліджень, проведених у різний час; порівняння ефективності різних методик збору й аналізу емпіричних даних; формування моделі вибірки одержання попередньої інформації про досліджуваний об'єкт.

Розвиток вторинного аналізу соціологічних даних є можливим при виконанні трьох умов: 1) наявність накопичених первинних даних соціологічних досліджень, оскільки саме первинні дані (емпіричні масиви) надають найцінніший матеріал для вторинного аналізу; 2) оскільки коректне використання даних будь-якого дослідження неможливе без докладної

інформації про його методичні аспекти (тобто про методи збору даних, особливості вибірки й т. п.), поряд з отриманими в дослідженні результатами необхідне цілеспрямоване нагромадження всіх цих відомостей; 3) потрібна ефективна система пошуку інформації, що необхідна для вторинного аналізу, серед всіх матеріалів, які накопичені соціологами. Таким чином, вторинний аналіз припускає наявність розвинутої системи нагромадження, зберігання, пошуку й аналізу соціологічних даних. Такою системою є банки соціологічних даних.

Зазвичай центри зберігання будь-яких даних називають архівами. Архіви даних існували задовго до появи комп'ютерів. Під архівами даних розуміють склади інформації на певних носіях, а також спеціально організовані установи, які займаються збором даних, їх зберіганням і розповсюдженням, забезпечуючи необхідну якість, наукову обґрунтованість, порівнянність інформаційних масивів. Застосування сучасної комп'ютерної техніки дозволяє не лише полегшити зберігання та копіювання великих обсягів даних, але й допомагає вирішувати питання пошуку та впорядкування наявних даних. Комп'ютерні архіви даних сьогодні перетворюються у більш функціональні центри накопичення даних, які прийнято називати **банками інформації**, хоча слово «архів» залишилося в назвах багатьох з них як більш відоме та зрозуміле широкому колу користувачів.

Банк соціологічної інформації – сукупність (1) інформації, що отримується й використовується в процесі соціологічних досліджень та (2) засобів її отримання, зберігання, переробки й розповсюдження.

Різноманітні банки соціологічної інформації існують в усьому світі при наукових та академічних закладах, університетах, великих компаніях, що займаються соціологічними дослідженнями, наприклад, наступні: Архів Даних Великобританії (UKDA) (<http://www.data-archive.ac.uk/>); Архів Даних Соціологічного Дослідницького Комітету (SSRC) (<http://www.ssrc.org/>); Архів Штайнметца (<http://www.dans.knaw.nl/>); Європейський Консорціум Політичних Досліджень (ECPR) (<http://www.essex.ac.uk/ecpr/>); Міжнародна Федерація Організацій Даних по Соціальних Науках (IFDO) (<http://www.ifdo.org/>); Міжуніверситетський Консорціум Політичних і Соціальних Досліджень (ICPSR) (<http://www.icpsr.umich.edu/>); Центр Роупера <http://www.ropercenter.uconn.edu/>).

Сьогодні помітною є тенденція об'єднання існуючих банків соціологічних даних у більш великі формації: у національні або у міжнародні.

Перші кроки для створення національного банку соціологічних даних вже зробили Київський міжнародний інститут соціології та Інститут соціології НАН України, які дали доступ до результатів своїх досліджень всім зацікавленим користувачам. У цих організаціях створено локальні банки соціологічної інформації, що відповідають міжнародним стандартам, і дозволяють подолати фрагментарність існуючих знань про соціальні процеси в Україні, а також проводити порівняльний аналіз накопичених і очікуваних

результатів емпіричних соціологічних досліджень. Банк соціологічної інформації Інституту соціології НАН України містить не тільки результати регулярно проведених опитувань на різну тематику, але й величезну базу даних соціологічного моніторингу «Українське суспільство» з 1992 року до нинішнього часу. На основі наявних ознак можна вивчати в динаміці тенденції соціальних змін, стан громадської думки, ціннісні орієнтації населення, перспективи розвитку українського суспільства тощо. У банку даних Київського міжнародного інституту соціології зберігаються результати численних омнібусних досліджень інституту та дані відкритих досліджень, які, у своїй більшості, присвячені дослідженню політичних поглядів електорату України.

У грудні 2014 року Київським міжнародним інститутом соціології і Центром «Соціальні індикатори» у співробітництві з Києво-Могилянською академією за грантом Міжнародного фонду «Відродження» створена перша черга Національного банку соціологічних даних "Київський архів" (див. рис. 1.1). Перелік організацій, що передають дані до Національного банку соціологічних даних:

1. Київський міжнародний інститут соціології <http://kiis.com.ua/>
2. Компанія TNS <http://www.tns-ua.com/>
3. Інститут соціології НАН України <http://i-soc.com.ua/institute/>
4. Київський національний університет ім. Тараса Шевченка <http://www.univ.kiev.ua/ru/>
5. Український центр економічних і політичних досліджень імені Олександра Разумкова <http://www.razumkov.org.ua/>
6. Фонд «Демократичні ініціативи» імені Ілька Кучеріва <http://www.dif.org.ua/>
7. Український інститут соціальних досліджень ім. Олександра Яременко <http://www.uisr.org.ua/>
8. Київський центр політичних досліджень и конфліктології <http://analitik.org.ua/>
9. СОЦИС <http://www.socis.kiev.ua/>

Наявність Національного банку соціологічних даних сприяє кращому розумінню соціальних процесів, що відбувається в нашій країні. Користуватися накопиченими даними можуть всі зацікавлені особи, зокрема, студенти та аспіранти, що готують курсові, дипломні та дисертаційні роботи. Користувачі можуть шукати, переглядати, аналізувати та завантажувати дані соціологічних опитувань, проведених українськими дослідницькими установами – від перших досліджень початку 1990-х до сьогодення.

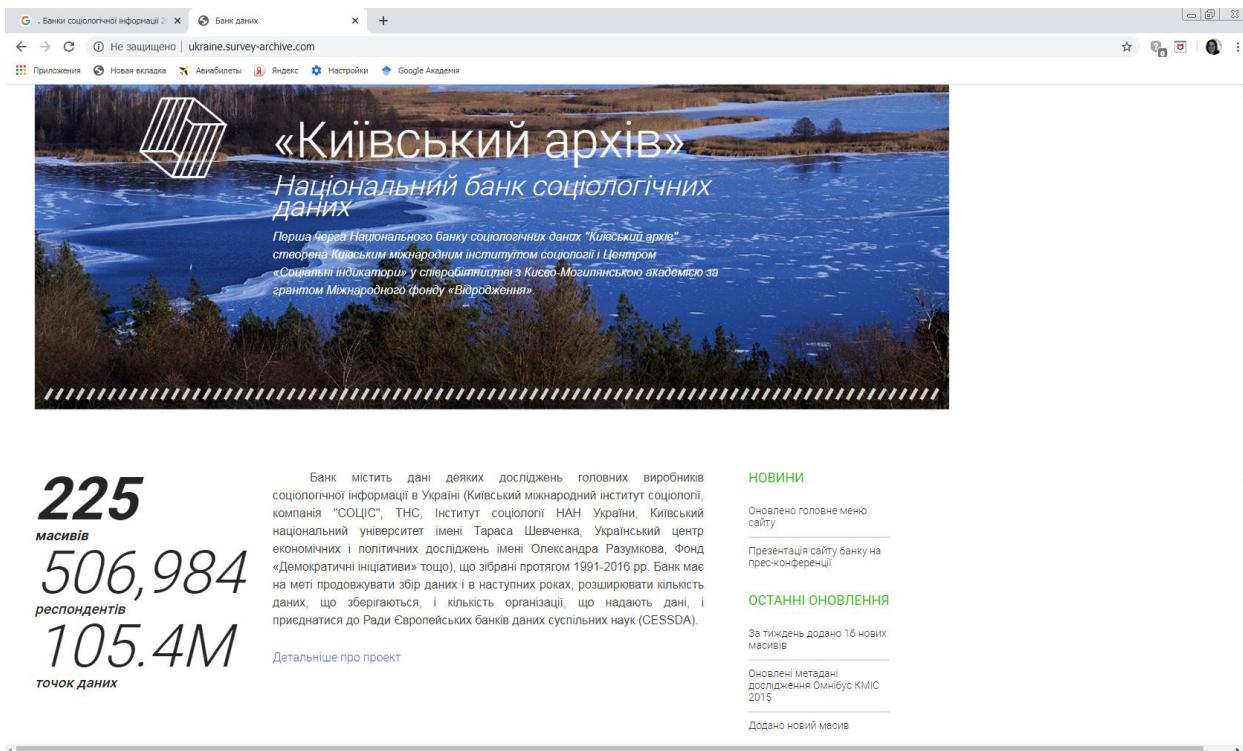


Рис. 1.1. Сайт Національного банку соціологічних даних «Київський архів»

Щоб скористатися зібраними даними необхідно зайти на сайт «Київського архіву» <http://survey-archive.com/Ukraine/>, де представлено перелік усіх досліджень, що містяться в архіві. По кожному дослідженню викладено інформацію про його ключові характеристики, мається доступ як до результатів досліджень, так й до масивів первинних даних.

1.3. Місце і роль процедур обробки та аналізу даних у соціологічному дослідженні

Соціологічне дослідження має свою структуру, що передбачає послідовність дій соціолога, які дають змогу вирішити поставлену проблему. Цю послідовність дій можна визначити як етапи соціологічного дослідження:

- 1) розробка програми дослідження;
- 2) збір емпіричних даних (польовий етап) або прийняття рішення про застосування вторинних даних;
- 3) обробка та аналіз даних, презентація отриманих результатів.

Обробка даних – сукупність технічних прийомів і методів, що дозволяють згрупувати дані (знизити їх розмірність, представити у вигляді таблиці чи візуалізувати). Наразі обробка соціологічних даних неможлива без застосування комп'ютерів і спеціалізованих пакетів програм.

Аналіз даних – сукупність дій, що реалізуються дослідником у процесі вивчення отриманих даних з метою формування певних уявлень про характер досліджуваного явища. Дослідник намагається стиснути дані, скоротити їх кількість, прагнучи втратити щонайменший обсяг корисної інформації, яка

потенційно міститься в даних дослідження. Робиться це зазвичай за допомогою математичних методів.

У процесі обробки й аналізу даних найчастіше застосовують тотожні технічні й математичні прийоми. Проте з гносеологічної точки зору ці підходи істотно відрізняються. Під час обробки даних соціолог використовує стандартний набір засобів (як правило – це одновимірні розподіли, таблиці, діаграми і графіки) для найбільш наочної демонстрації отриманих даних. У випадку вдалого підбору технічних засобів дані ніби «говорять самі за себе». Під час аналізу даних дослідник висуває певну модель соціального явища, демонструє відповідність (або протиріччя) емпіричних даних цій моделі, розробляє моделі досліджуваного соціального феномену.

Аналіз даних є ключовим етапом усього соціологічного дослідження. Під час аналізу даних відбувається безпосередня перевірка відповідності зібраної інформації тим моделям соціальних явищ, які (явно чи латентно) застосовувалися дослідниками; формулюються і перевіряються нові моделі, що відображують закономірності, втілені в зібраних даних.

У наш час обробка й аналіз соціологічних даних не мисляться без застосування комп'ютерів і спеціалізованих пакетів програм, ера «ручної» обробки емпіричного матеріалу закінчилась. Інформаційна епоха висуває нові вимоги до методів роботи з інформацією й способів витягу нових знань із емпіричних даних. Стрімкий розвиток інформаційних технологій багато в чому полегшує роботу соціолога-аналітика, але, водночас, змушує освоювати нові інструменти аналізу соціологічної інформації, зокрема, пакети програм, що дають можливість обробляти та аналізувати результати масових опитувань.

Для проведення аналізу даних соціологи мають можливість застосовувати найрізноманітніші програмні продукти, що дозволяють здійснювати статистичну обробку даних. Кількість таких програм за даними Міжнародного статистичного інституту (англ. International Statistical Institute) наразі наближається до тисячі. До них відносяться: 1) електронні таблиці (Excel, Lotus, QuattroPro тощо), які до свого складу включають стандартні методи статистичної обробки даних; 2) математичні пакети загального призначення (наприклад, MatLab, MathCad, Mathematica та ін.); 3) спеціалізовані статистичні пакети, що дозволяють застосовувати найсучасніші методи математичної статистики для обробки даних, найвідомішими з яких є SPSS, STATISTICA, STADIA, SAS, STATGRAPHICS; 4) мови програмування, особливо Python та R, які все частіше викликають зацікавленість соціологів, націлених на аналіз big data та набуття навичок у сфері data science.

Проте у світовій практиці аналізу первинної соціологічної інформації й досі лідируватиме пакет SPSS, який застосовують провідні науково-дослідні агентства для аналізу результатів опитувань.

SPSS (Statistical Package for Social Science) є найбільш поширеною програмою аналізу соціологічної інформації на міжнародному ринку соціологічних досліджень. Дана програма містить широкий спектр методів

статистичного аналізу, які дають можливість здійснити обробку первинної соціологічної інформації на будь-якому рівні: від розрахунку дескриптивних статистик до побудови складних багатовимірних моделей. Крім того, саме формат SPSS є найбільш поширеною формою обміну соціологічною інформацією з міжнародними партнерами.

Не можна обійти увагою програму OCA (Обробка Соціологічних Анкет), яка є вітчизняною розробкою. OCA компактна, зручна в роботі, легко засвоюється. Суттєвою перевагою цієї програми є її сумісність з SPSS: дані, введені в комп'ютер за допомогою програми OCA, за необхідністю легко переводяться в формат SPSS.

Комп'ютерна обробка емпіричних даних передбачає їх представлення у специфічному вигляді, що поєднує інформацію про відповіді респондентів на всі запитання анкети з інформацією про інструментарій, яким здійснювалося вимірювання.

Масив даних – матриця, рядки якої містять числа, якими закодовані відповіді кожного респондента на всі запитання анкети, а стовпці – відповіді всіх респондентів на конкретне запитання. Розмірність цієї матриці дорівнює $n \times m$, де n – кількість опитаних, m – число ознак, що виміряні за допомогою анкети.

«Паспорт» масиву даних – це текст листа опитування, що введений в особливому форматі у комп'ютер. У «паспорті» міститься інформація про всі пункти аркуша опитування (анкети): текст запитання; альтернативи відповідей на запитання; тип шкали вимірювання тощо.

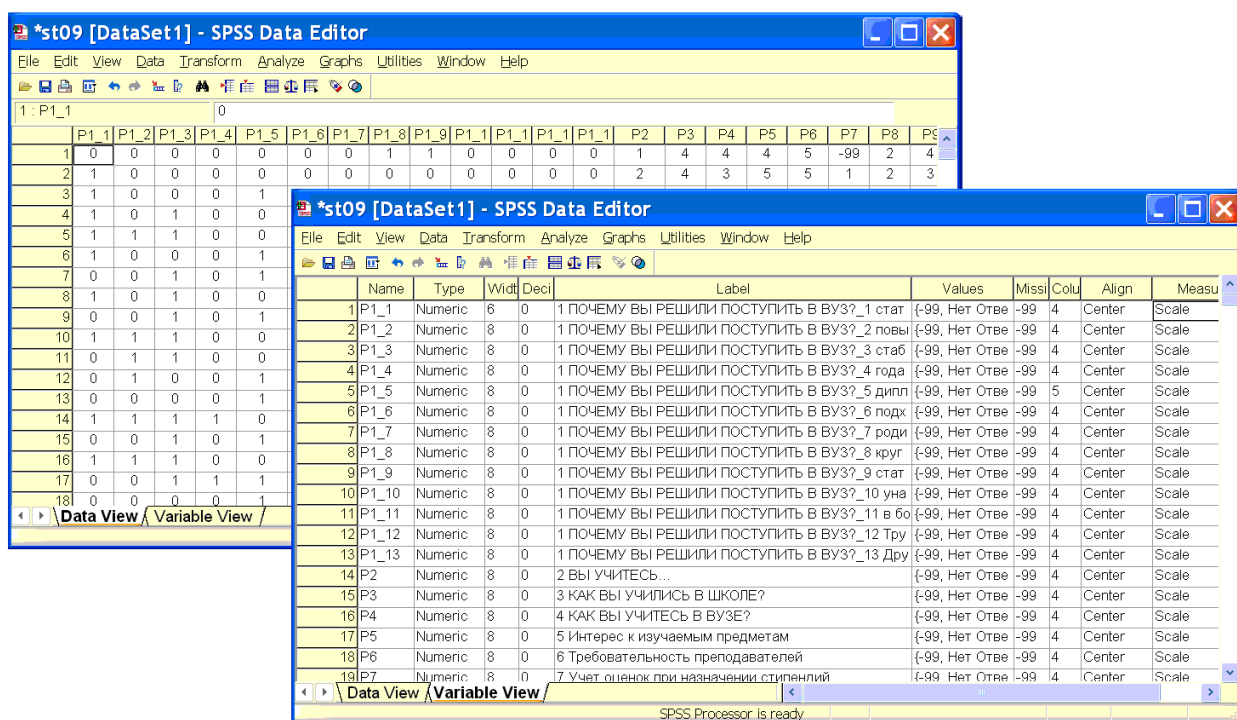


Рис. 1.2. Представлення масиву даних в пакеті SPSS: дані (Data View) та «паспорт» масиву (Variable View)

Аналіз результатів соціологічних опитувань необхідно проводити з урахуванням *методологічних принципів аналізу соціологічних даних*:

Перший принцип – необхідність перевірки адекватності моделі методу аналізу даних (тобто системи передумов і постулатів) і моделі того соціального явища, що вивчається.

Другий принцип – системний підхід, що враховує зв'язок різних етапів соціологічного дослідження. Цей принцип конкретизується у вигляді цілої низки положень, найважливішими серед яких є такі: 1) зв'язок вимірювання та аналізу результатів вимірювання; 2) залежність інтерпретації результатів застосування методу від концептуальних настанов дослідника, від поставлених перед ним цілей.

Третій принцип передбачає розгляд аналізу даних як способу їхнього існування. З появою нових даних виникають нові ідеї, підходи та методи, уточнюється розуміння досліджуваних феноменів, часто виникає необхідність повторно звернутися до вже проаналізованих даних, поглянути на них з нових позицій.

1.4. Поняття статистики та статистичного аналізу

Термін «статистика» походить від латинського «status», що означає стан речей, політичний стан. Від кореня цього слова виникли слова «stato» (держава), «statista» (статистик, знавець держави), «statistiks» (статистика — певна сума знань, зведень про державу). Спочатку, в XVIII в., коли статистика початку оформлятися в наукову дисципліну, термін «статистика» зв'язувався лише із системою опису фактів, що характеризують стан держави. При цьому важко було спрогнозувати, що згодом статистика перетвориться у науку, яка не тільки займається збором та аналізом даних, а й вивчає закономірності будь-яких масових процесів.

Сьогодні слово «статистика» використовують у кількох значеннях:

- Сукупність даних про яке-небудь явище або процес (наприклад, можна говорити про статистику виборів, статистику народжуваності, злочинів тощо).

- Галузь практичної діяльності, спрямовану на збір, обробку та аналіз статистичних даних, що відбивають явища й процеси громадського життя. Цю роботу зазвичай виконують і очолюють спеціальні державні установи (наприклад, Держкомстат України).

- Наука про методи збору, обробки, аналізу й інтерпретації даних, що характеризують масові (зокрема, суспільні) явища й процеси.

- Статистичний показник, що характеризує властивості вибіркової сукупності.

Статистика як наука про методи збору, обробки та аналізу даних, поєднує цілий комплекс спеціалізованих наукових дисциплін, у якому можна виділити такі основні напрямки:

- Методи збору даних. Основними методами збору даних є повне або вибіркоче обстеження генеральної сукупності та експеримент, методологічною основою яких є теорія вибірки й планування експерименту.

- Методи виміру. Теоретичною основою цього напрямку є загальна теорія вимірів, на базі якої розробляються спеціальні показники, використовувані певними науками, зокрема соціологією.

- Методи обробки й аналізу даних – статистичний аналіз, що поєднує теорію ймовірностей, математичну статистику та їхні додатки в різних наукових галузях – від технічних наук до соціальних.

Якщо математична статистика має теоретичне призначення, розробляє методи статистичної обробки й аналізу даних, займається обґрунтуванням і перевіркою їх валідності, ефективності, умов застосування, стійкості до порушення умов застосування, то прикладна статистика націлена на практичне застосування цих методів для аналізу результатів дослідження, наприклад, соціологічного. Отже, можна сказати, що прикладна статистика є застосуванням статистичного аналізу з метою дослідження емпіричних даних.

Статистичний аналіз – це аналіз статистичних даних за допомогою статистичних методів з метою з'ясування тих закономірностей, які можуть бути встановлені на їх основі.

Статистичні дані являють собою сукупність об'єктів і ознак (змінних), що їх характеризують; вони можуть бути отримані шляхом проведення масових опитувань, експериментів, вилучення інформації з відкритих джерел тощо.

До методів статистичного аналізу відносяться методи угруповань, визначення форми та параметрів розподілів, аналіз рядів динаміки, кореляційний, регресійний, дисперсійний, факторний аналізи, а також багато інших. Вибір методів статистичного аналізу, що будуть застосовані в конкретному дослідженні, визначається наступними чинниками: 1) типами шкал, якими вимірювались досліджувані ознаки; 2) характером наявних даних (результати вибіркового чи суцільного дослідження); 3) необхідним рівнем аналізу (описовим, пояснювальним чи прогностичним).

Статистичний аналіз поділяють на дескриптивний (описовий) та аналітичний (індуктивний) (див. рис.1.3).

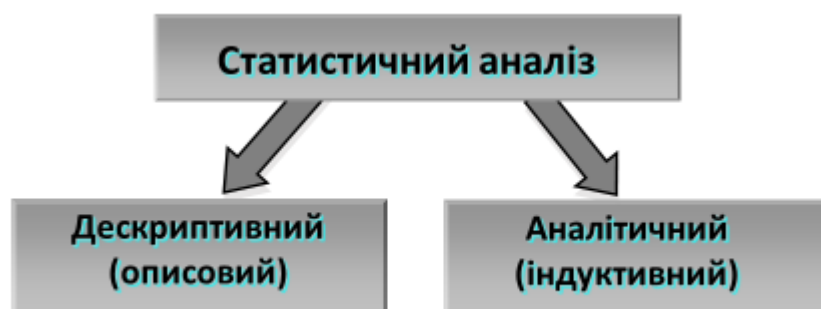


Рис. 1.3. Різновиди статистичного аналізу

Дескриптивний (описовий) аналіз має метою наочне представлення основних властивостей досліджуваних даних та виявлення притаманних їм закономірностей. Найпростіший спосіб досягнення цієї мети полягає у «стисненні», усередненні інформації, що міститься у досліджуваних даних. Найбільш поширеним різновидом дескриптивного аналізу є формування та візуалізація рядів розподілів, застосування дескриптивних статистик, які, з одного боку, висвітлюють загальне в сукупності даних (міри центральної тенденції – середнє арифметичне, медіана, мода), з другого боку, демонструють, у чому й наскільки дані розрізняються (міри варіації, наприклад, дисперсія, стандартне відхилення, інтерквартильний розмах тощо). Дескриптивний аналіз представляє собою опис результатів вибіркового дослідження без поширення їх на генеральну сукупність. Він забезпечує короткий підсумок про вибірку та про спостереження, які були зроблені.

Аналітичний (чи індуктивний) статистичний аналіз є процесом статистичного виведення, що полягає в поширенні результатів вибіркового дослідження на генеральну сукупність, тобто в отриманні статистичних висновків про генеральну сукупність на основі вибіркового дослідження.

Статистичні висновки мають ймовірнісну природу та умовно поділяються на дві групи:

1) оцінювання параметрів генеральної сукупності – точкове (знаходження конкретного числового значення шуканого параметру генеральної сукупності) або інтервальне (знаходження інтервалу, в якому із заданою ймовірністю знаходиться шуканий параметр);

2) перевірка статистичних гіпотез – отримання ймовірнісних висновків про те, що певні характеристики вибіркової сукупності (кореляції, відмінності тощо) відображають відповідні параметри генеральної сукупності. При проведенні індуктивного статистичного аналізу застосування навіть найпростіших статистичних методів (напр., аналізу рядів розподілів) передбачає додаткову процедуру – оцінювання ймовірності помилки зроблених висновків.

Статистичний аналіз є невід'ємною частиною кількісних соціологічних досліджень, він сприяє коректній перевірці дослідницьких гіпотез та створює підстави для формулювання нових припущень щодо досліджуваних феноменів. Наразі статистичний аналіз проводиться за допомогою спеціальних пакетів програм (напр., OCA чи SPSS), завдяки чому дослідник звільняється від необхідності виконання трудомістких розрахунків.

Література до теми

Бююль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2002. С. 26-43.

Закон України «Про інформацію». Верховна Рада України; Закон від 02.10.1992 № 2657-XII. URL: <http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=2657-12>.

Паніотто, В. І. *Національний банк соціологічних даних – що він дасть Україні.* URL: <http://kiis.com.ua/materials/news/2014/marketing%20conf/DataBank%202.pdf>.

Додаткова література

Головаха, Є. І. Концептуальні й організаційно-методичні засади створення «Українського соціологічного архіву і банку даних соціальних досліджень» *Соціологія: теорія, методи, маркетинг.* 2000. № 1. С. 138–151.

Горбачик, А. П. Архіви соціальних даних: цілі існування, форми роботи, проблеми створення. *Соціологія: теорія, методи, маркетинг.* 2000. №3. С. 130-144.

Горбачик, О. А. Електронні банки соціологічних даних. *Український соціум.* 2009. № 2 (29). С. 14-21.

Ковтун, Н.В. *Теорія статистики: підручник.* К.: Знання, 2012.

Мармоза, А. Т. *Теорія статистики.* 2-ге вид. перероб. та доп. К.: «Центр учбової літератури», 2013.

Толстова, Ю. Н. Принципы анализа данных в социологии. *Социология: методология, методы, математическое моделирование (4М).* 1991. №. 1. С. 51-61. URL: <https://www.jour.isras.ru/index.php/soc4m/article/view/3838>.

Хижняк, Л. М. Міфи про соціальну статистику, професійна підготовка соціологів та інформаційна безпека держави. *Вісник ХНУ імені В. Н. Каразіна. Серія «Соціологічні дослідження сучасного суспільства: методологія, теорія, методи».* – 2017. № 38. С. 92-95. URL: <https://periodicals.karazin.ua/ssms/article/view/8644/>

Питання для самоконтролю

1. Чи є масиви результатів соціологічних опитувань соціологічною інформацією?
2. Чи є відрізнєння у трактуванні «соціологічних даних», «емпіричних соціологічних даних» та «соціологічної інформації»?
3. Соціологічна інформація та соціальна інформація. В чому різниця?
4. Чому англomовному сегменті немає визначення терміну «sociological information»?
5. Що таке масив даних? Чи тотожні поняття «масив даних» та «масив анкет»?
6. Чим відрізняється обробка даних від аналізу даних?
7. У чому полягають методологічні принципи аналізу соціологічних даних?
8. Які значення слова «статистика» Ви знаєте?
9. Що таке статистичний аналіз даних?
10. Чим відрізняється дескриптивний статистичний аналіз від статистичного виведення?

Тема 2. Аналіз одновимірних розподілів

2.1. Вигляд одновимірних розподілів та їх побудова в SPSS

Перший крок аналізу результатів соціологічного опитування – опис отриманих результатів, який головним чином зводиться до розрахунку та аналізу одновимірних розподілів.

Одновимірний розподіл (варіаційний ряд) являє собою таблицю, що демонструє частоту, з якою різні варіанти відповіді на певне запитання анкети спостерігаються в наборі даних (див. рис. 2.1а). Такі розподіли називають одновимірними, оскільки вони будуються на основі однієї змінної.

Наскільки Ви задоволені своїм життям?	частота
Повністю задоволений	66
Скоріш задоволений	199
Важко відповісти	39
Скоріш не задоволений	127
Повністю не задоволений	78
Разом	509
НеВідп	1
Всього	510

Рис 2.1а. Частотний розподіл

Наскільки Ви задоволені своїм життям?	відсоток
Повністю задоволений	12,9
Скоріш задоволений	33,0
Важко відповісти	7,6
Скоріш не задоволений	24,9
Повністю не задоволений	15,3
Разом	99,8
НеВідп	,2
Всього	100,0

Рис 2.1б. Відсотковий розподіл

Рис. 2.1. Вигляд частотного одновимірного та відсоткового одновимірного розподілу ознаки «задоволеність життям»

Проте аналітики віддають перевагу аналізу відсоткових одновимірних розподілів (рис. 2.1б), оскільки вони, на відміну від частотних, дозволяють здійснювати порівняння з аналогічними даними інших опитувань, або ж порівнювати думки різних груп населення.

Саме порівняння даних є джерелом корисної інформації щодо певних соціальних процесів. Відсотки дозволяють визначити відмінності в пропорціях і проводити порівняльний аналіз. Дані, виражені у відсотках значно спрощують сприйняття, дозволяють без додаткових обчислень «побачити» картину в цілому.

У всіх пакетах статистичної обробки даних одновимірні розподіли поряд с частотами містять відсотки, а також додаткову інформацію: проценти до опитаних, кумулятивні відсотки тощо.

Побудова одновимірного розподілу в пакеті SPSS:

- Виберіть у меню команди *Analyze (Аналіз)* → *Descriptive Statistics (Дескриптивні статистики)* → *Frequencies (Частоту)*.
- З'явиться діалогове вікно *Frequencies*.
- Виберіть змінні, для яких будуть побудовані одновимірні розподіли, і перенесіть їх у перелік вихідних змінних.

- Підтвердити операцію кнопкою **ОК**.

У цьому полі вказуємо одну чи кілька змінних, для яких необхідно побудувати одновимірний розподіл.

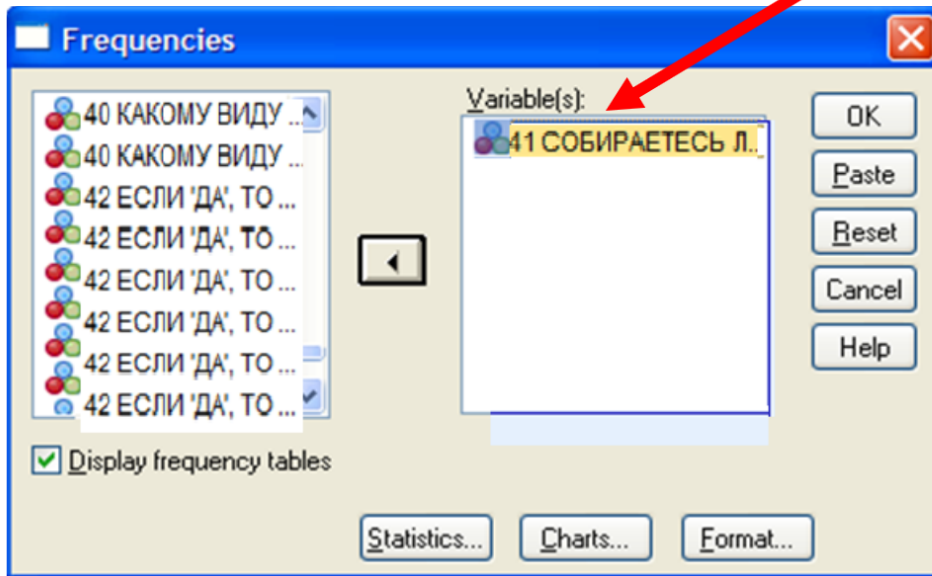


Рис. 2.2. Побудова одновимірного розподілу в пакеті SPSS

У результаті отримаємо дві таблиці: 1) *Case Summary*, що містить загальну інформацію про кількість анкет у масиві (N), про кількість респондентів, що відповіли на аналізоване запитання (Valid), а також про кількість респондентів, які не відповіли на це запитання (Missing); 2) *Frequencies* – одновимірний розподіл відповідей на запитання: «*Чи збираєтеся Ви у майбутньому вчитися в аспірантурі?*» (рис. 2.3).

41. Чи збираєтеся Ви у майбутньому вчитися в аспірантурі?					
		Частоти	Відсотки	Валідні відсотки	Кумулятивні відсотки
Валідні	1 Ні	492	16,1	16,2	16,2
	2 Скоріше ні	741	24,2	24,4	40,5
	3 Важко відповісти	926	30,3	30,4	71,0
	4 Скоріше так	497	16,3	16,3	87,3
	5 Так	386	12,6	12,7	100,0
	Усього	3042	99,5	100,0	
Пропущені	Немає відповіді	16	,5		
Загалом		3058	100,0		

Рис. 2.3. Вигляд одновимірного розподілу у пакеті SPSS

У цьому розподілі представлено наступну інформацію:

- 1) варіанти відповіді на запитання;
- 2) частоти – кількість респондентів, що обрали кожний з варіантів відповіді;
- 3) відсотки – відсотки респондентів, що обрали кожний з варіантів відповіді, які розраховані відносно всіх опитаних;

- 4) валідні відсотки – відсотки респондентів, що обрали кожний з варіантів відповіді, які розраховані відносно тих, хто відповів на запитання;
- 5) кумулятивні відсотки – накопичені відсотки.

Множинні відповіді: побудова в SPSS одновимірних розподілів для номінальних ознак із сумісними альтернативами

Питання, на які можна одночасно дати кілька відповідей зустрічаються досить часто в соціологічних анкетах (такі шкали називають номінальними з сумісними альтернативами або множинними відповідями). Для кодування й аналізу множинних відповідей у SPSS є два різних методи: метод множинної дихотомії та категоріальний метод. Найчастіше застосовують метод множинної дихотомії. Саме тому ми розглянемо лише його.

Розглянемо особливості кодування й аналізу множинних відповідей у SPSS на прикладі питання анкети, яке має такий вигляд:

1. ЧОМУ ВИ ВИРІШИЛИ ВСТУПИТИ ДО ВНЗ?

(можна позначити не більш 3-х варіантів відповідей)

1. Хотіли стати висококваліфікованим спеціалістом в обраній галузі
2. Хотіли підвищити свій соціальний статус, мати більш престижне становище у суспільстві
3. Хотіли забезпечити собі стабільний матеріальний достаток у майбутньому
4. Думали продовжити роки учнівства, безтурботного існування (не йти працювати, не служити в армії)
5. Вважали, що диплом про вищу освіту (усе одно, яку) знадобиться Вам у житті
6. Сподівалися зустріти майбутнього супутника життя
7. На цьому наполягли Ваші батьки
8. Думали, що вища освіта забезпечить Вам цікаве коло спілкування в теперішньому і майбутньому
9. Вважали, що вища освіта надасть Вам можливість стати культурною, високоосвіченою людиною
10. Хотіли успадкувати професію батьків
11. Важко відповісти.

Відповіді респондентів на питання «Чому Ви вирішили вступити до ВНЗ?» у файлі даних будуть зберігатися у вигляді 11 дихотомічних ознак (p1_1, p1_2, ... p1_11), кожна з яких представляє одну з 11 альтернатив відповіді (рис. 2.4).

	Name	Type	Width	Deci	Label	
1	p1_1	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Хотели стать	{-9
2	p1_2	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Хотели повысить	{-9
3	p1_3	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Хотели обеспечить	{-9
4	p1_4	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Думали продлить	{-9
5	p1_5	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Считали, что	{-9
6	p1_6	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Надеялись встретить	{-9
7	p1_7	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_На этом настояли	{-9
8	p1_8	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Думали что высшее	{-9
9	p1_9	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Предполагали, что	{-9
10	p1_10	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Хотели унаследовать	{-9
11	p1_11	Numeric	8	2	1. ПОЧЕМУ ВЫ РЕШИЛИ ПОСТУПИТЬ В ВУЗ_Трудно сказать	{-9
12	p2	Numeric	8	2	2. КАКОЕ УЧЕБНОЕ ЗАВЕДЕНИЕ ВИ ЗАКОНЧИЛИ ДО ПОСТУПЛЕН	{-9
13	p3	Numeric	8	2	3. ОБУЧЕНИЕ В УЧЕБНОМ ЗАВЕДЕНИИ БЫЛО ...	{-9
14	p4	Numeric	8	2	4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	{-9
15	p5	Numeric	8	2	5. ВЫ ПОСТУПИЛИ В ВУЗ ПО РЕЗУЛЬТАТАМ...	{-9
16	p6	Numeric	8	2	6. ВЫ ПОСТУПИЛИ В ВУЗ В ГОД ПОЛУЧЕНИЯ АТЕСТАТА?	{-9
17	p7_1	Numeric	8	2	7. ЕСЛИ НЕТ, ТО ПОЧЕМУ_Не прошли по конкурсу	{-9
18	p7_2	Numeric	8	2	7. ЕСЛИ НЕТ, ТО ПОЧЕМУ_Служили в армии	{-9
19	p7_3	Numeric	8	2	7. ЕСЛИ НЕТ, ТО ПОЧЕМУ_Не определились с вузом	{-9
20	p7_4	Numeric	8	2	7. ЕСЛИ НЕТ, ТО ПОЧЕМУ_По семейным обстоятельствам	{-9

Рис. 2.4. Представлення відповідей респондентів на питання «Чому Ви вирішили вступити до ВНЗ?» методом множинної дихотомії

У методі множинної дихотомії для кожної з альтернатив відповіді на запитання анкети створюється окрема змінна, яка має значення 1, якщо відповідна альтернатива вибрана респондентом, і 0 – якщо не вибрана. Кодові значення (0 та 1) при цьому вибираються довільно, однак для всіх відповідей вони повинні бути однаковими. Найчастіше використовують значення 0 і 1, оскільки в таких випадках ми отримуємо дихотомічні змінні, які є псевдометричними та дозволяють застосовувати багато методів та статистик, які неможливо застосовувати для інших номінальних.

Перш ніж обчислити одновимірний розподіл для множинних відповідей в SPSS, необхідно визначити набори змінних:

- Завантажте файл даних.
- Виберіть в меню команди *Analyze Multiple Response (Множинні відповіді)* → *Define Sets...(Визначити набори)*.
 - Відкриється діалогове вікно *Define Multiple Response Sets (Визначення наборів відповідей)*, яке показано на рисунку 2.5.
 - Виділіть в списку початкових змінних дихотомічні змінні, які відповідають одному питанню анкети (номінальна шкала із сумісними альтернативами, тобто множинна відповідь) і перенесіть їх в список *Variables in Set (набори змінних)*.

- Задайте дихотомічне кодування змінних (Опція *Dechotomies* в групі *Variables Are Coded As*). Це налаштування вибирається автоматично програмою. В полі *Counted Value* (*Враховане значення*) введіть 1.
- Дайте набору ім'я і мітку.
- Натисніть на клавішу *Add* (*Додати*), і створений набір буде внесений до списку наборів множинних відповідей (*Multiple Response Sets*).
- Натисніть на клавіші *Close* (*Закрити*), щоб закінчити процес визначення набору.

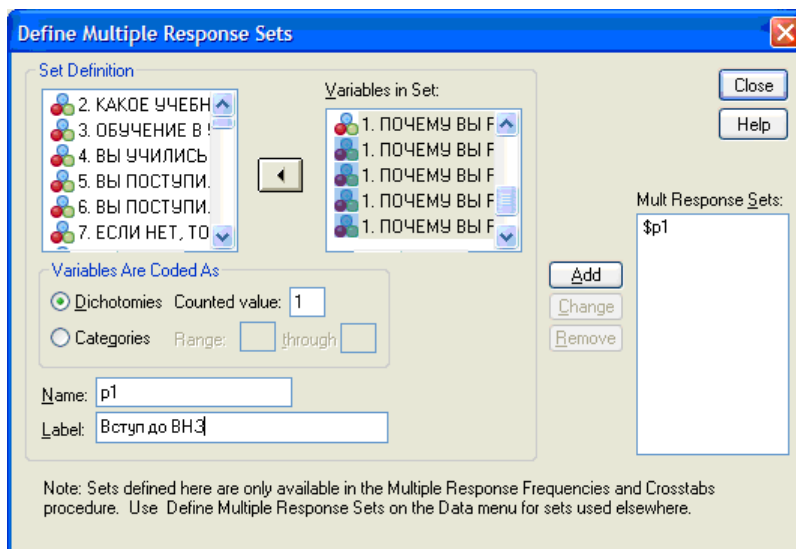


Рис.2.5. Діалогове вікно *Define Multiple Response Sets* (*Визначення множини множинних відповідей*)

Розрахунок одновимірного розподілу для дихотомічних наборів

- Щоб створити одновимірний розподіл для дихотомічного набору, виберіть команду меню *Analyze* (*Аналіз*) → *Multiple Response* (*Множинні відповіді*) → *Frequencies...* (*Частоти*).
- Відкриється діалогове вікно *Multiple Response Frequencies* (*Частоти множинних відповідей*), яке можна побачити на рисунку 2.6.
- В списку *Multiple Response Sets* (*Набір відповідей*) цього діалогового вікна відображаються попередньо визначені набори змінних.
- Перенесіть набір *\$p1* (*Вступ до ВНЗ*) у список *Table(s) for* (*Таблиці для*).
- Натисніть клавішу *OK*.

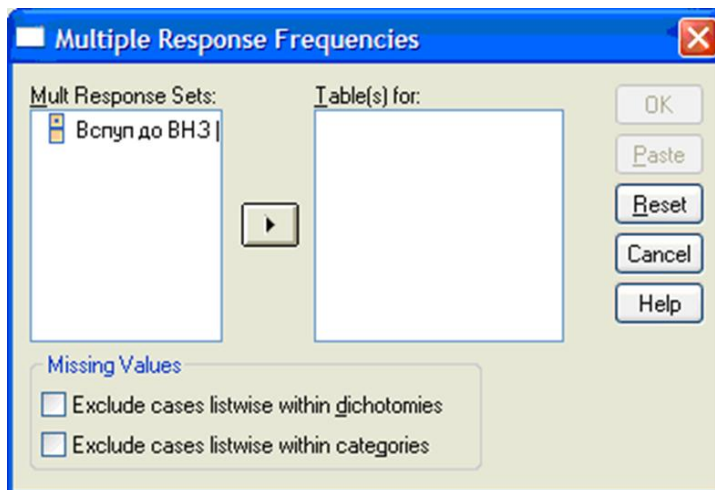


Рис. 2.6. Діалогове вікно *Multiple Response Frequencies* (Частоти множинних відповідей)

У результаті обчислень отримуємо дві таблиці: *Case Summary* та *Frequencies* (рис. 2.7).

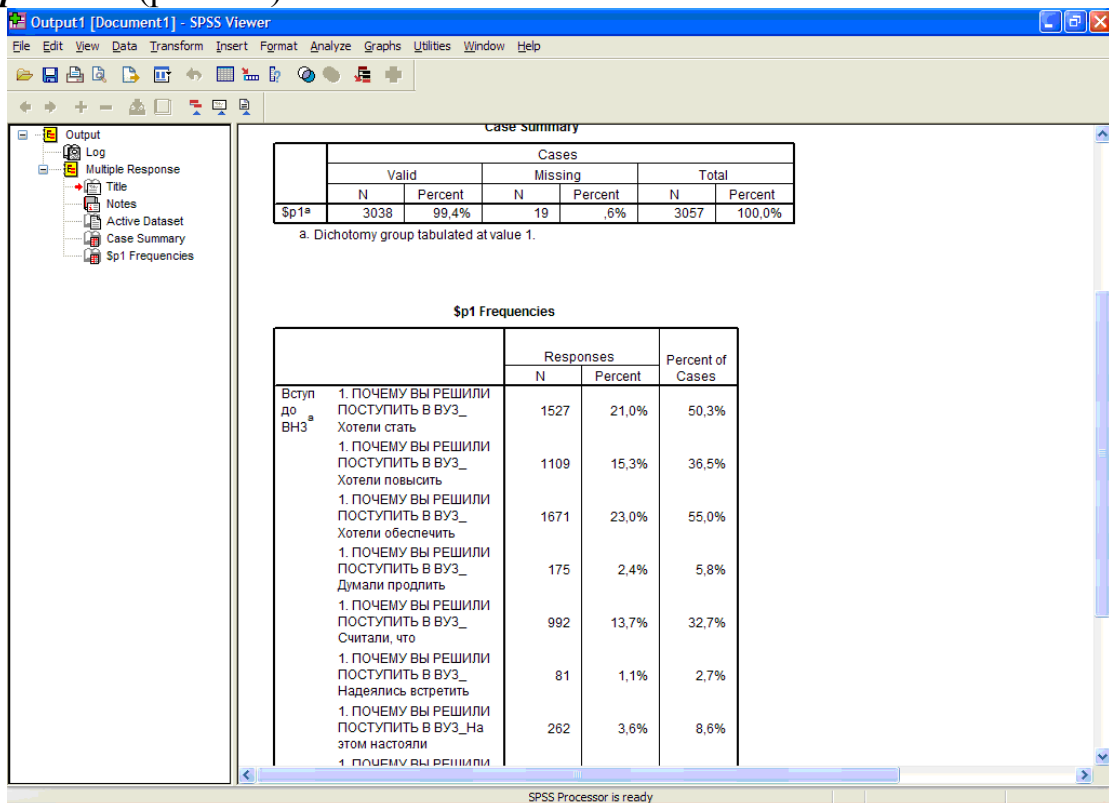


Рис. 2.7. Результати обчислень одновимірного розподілу для дихотомічного набору

Результати обчислень одновимірного розподілу для дихотомічного набору містять дві таблиці: *Case Summary* та *Frequencies*

Таблиця *Case Summary* представляє загальну інформацію:

- Кількість респондентів, що відповіли на питання «Чому Ви вирішили вступити до ВНЗ?» (Valid: N = 3038, Percent = 94,4%).
- Кількість респондентів, які не відповіли на це питання (Missing: N = 19, Percent = 0,6%). Анкета розглядується як відсутнє спостереження (Missing,

НеВідп), якщо жодна зі змінних набору не має значення, що враховується (у нашому прикладі значення "1").

- Кількість анкет (Total: N = 3057; Percent = 100%).

Основна інформація, тобто сам одномірний розподіл, міститься у таблиці *\$P1\$ Frequencies*. При інтерпретації результатів треба пам'ятати, що ознака, яка нами аналізується, виміряна за допомогою номінальної шкали з сумісними альтернативами. Саме тому у рядку **Total** вказано 7256 (кількість виборів респондентами альтернатив відповідей на питання анкети), а не 3038 (кількість анкет, що аналізуються).

У таблиці ***\$P1\$ Frequencies*** представлено два різних процентних значення. При визначенні першого з них (***Responses Percent***) спостережувана частота віднесена до загальної кількості відповідей "так" (7256), а при визначенні другого (***Percent of Cases***) - до загальної кількості спостережень (3038).

2.2. Статистики одномірного розподілу

Для номінальних та порядкових шкал у якості описових статистик використовуються в основному частотні та процентні розподіли, а для метричних та псевдометричних шкал – статистики, що дозволяють охарактеризувати середнє значення чи медіану отриманих даних.

Середні значення розраховують інколи і для порядкових шкал (у тих випадках, коли у дослідника є підстави розглядати порядкову шкалу як псевдометричну).

В таблиці нижче зазначені статистики, які слід використовувати для певних типів шкал.

Таблиця 2.1

Допустимі міри центральної тенденції відповідно до типу шкали

Тип шкали	Міри центральної тенденції
Номінальна	Мода
Порядкова	Мода, медіана
Метрична	Мода, медіана, середнє

Окремої уваги заслуговує різновид дихотомічних шкал, що має назву **фіктивні змінні (dummy variables)**.

Фіктивні змінні являють собою псевдометричні шкали, вони створюються з метою застосування до номінальних даних будь-яких кількісних методів. Відповідне перетворення носить назву дихотомізації номінальних даних. Що це означає? Замість кожної номінальної ознаки, що приймає кілька (n) значень, створюємо n нових дихотомічних ознак, які мають два значення: 1 – так, 0 – ні. Така дихотомія створюється штучно, вона не тотожна природної дихотомії. Наприклад, ознака «стать», яка приймає

значення 1 – чоловіки, 2 – жінки, не є фіктивною змінною. Якщо ми бажаємо перейти до фіктивних змінних, то трансформуємо цю ознаку у дві фіктивні:

- Чоловіки (питання в анкеті: Ви чоловік?), можливі значення: 1 – так, 0 – ні.
- Жінки (питання в анкеті: Ви жінка?), можливі значення: 1 – так, 0 – ні.

З формальної точки такі дихотомічні номінальні шкали можна розглядати як окремий випадок інтервальної шкали, яка має лише один інтервал – між 0 і 1, що зумовлює можливість застосування до таких шкал кількісних методів, які передбачають інтервальний рівень вимірювання вихідних даних. Саме цим зумовлено те, що багато відомих статистик, будучи обчисленими для фіктивних змінних мають розумну інтерпретацію, чого аж ніяк не можна сказати про інтерпретацію відповідних показників, обчислених для багатозначних номінальних шкал. Наприклад, середнє значення, розраховане для фіктивної змінної «Чоловіки» дорівнює 0,45. Отримане число інтерпретується як частка чоловіків серед опитаних.

Розрахунок в SPSS статистик одновимірного розподілу

Щоб розрахувати статистики одновимірного розподілу в пакеті SPSS необхідно виконати наступні дії:

- Виберіть у меню команди *Analyze (Аналіз) → Descriptive Statistics (Дескриптивні статистики) → Frequencies (Частоту)*.
- З'явиться діалогове вікно *Frequencies (Частоту)*.
- Натисніть кнопку *Statistics (Статистики)*.
- З'явиться діалогове вікно *Frequencies: Statistics (Частоту: Статистики)*, в якому можна задати розрахунок всіх статистик, які потрібні аналітику для проведення подальшого аналізу.

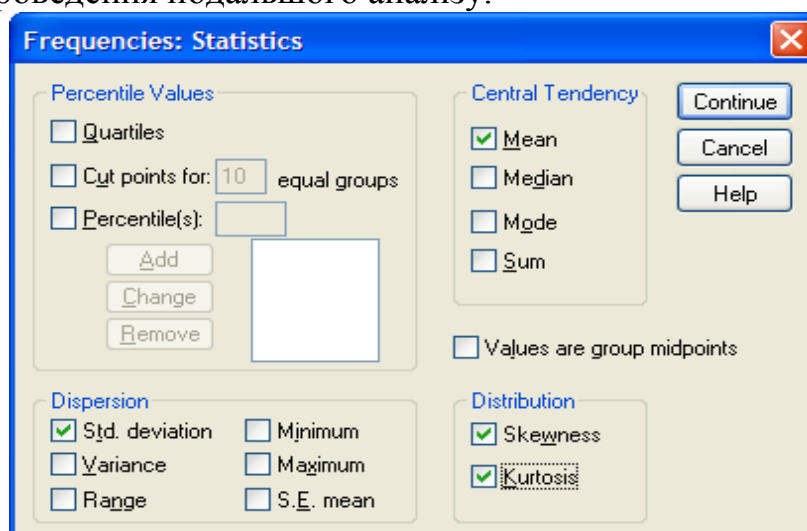


Рис. 2.8. Діалогове вікно *Frequencies: Statistics (Частоту: Статистики)*

Розглянемо статистики одновимірного розподілу, що розраховує SPSS (див. рис. 2.8).

1. Значення процентилей (*Percentile Values*).

Процентилі - це характеристики набору даних, які відображають ранги елементів масиву у вигляді чисел від 1 до 100, і є показником того, який відсоток значень знаходиться нижче певного рівня.

Наприклад, значення процентилю, що дорівнює 30, вказує, що 30% значень аналізованої кількісної змінної розташовується нижче цього рівня.

Квартилі в статистиці - це три величини, які ділять набір даних на чотири рівні частини. Q_{25} - перший квартиль (Q_{25} або Q_1 - 25% процентиль). Нижче Q_{25} розташовується 25% значень досліджуваної ознаки. Q_{50} - другий квартиль Q_2 (медіана або 50% процентиль). Нижче Q_{50} розташовується 50% значень досліджуваної ознаки. Q_{75} - третій квартиль Q_3 (75% процентиль). Нижче Q_{75} розташовується 75% значень досліджуваної ознаки.

Процентилі не часто застосовують для аналізу результатів соціологічних досліджень. Проте розрахунок квартилей буде дуже корисним у випадку, коли соціолог аналізує ознаки, виміряні порядковими шкалами, застосовуючи для стиснення інформації таку міру центральної тенденції, як медіана. Відомо, що для адекватного аналізу міри центральної тенденції необхідно аналізувати з урахуванням варіації аналізованої ознаки. Тобто значення медіани потрібно аналізувати укупі зі значення міжквартильного розмаху, який SPSS не розраховує. Проте його легко знайти, знаючи квартилі. Відомо, що міжквартильний дорівнює різниці між третім і першим квартилями: *міжквартильний розмах* = $Q_{75} - Q_{25}$, тобто знаючи значення першого та третього квартилей, ми автоматично визначаємо міжквартильний розмах.

2. Міри центральної тенденції (Central Tendency).

Мода (Mode) – значення ознаки, яке зустрічається найчастіше у наборі даних. У випадку, якщо дані згруповані та побудовано розподіл частот, модою є значення, яке має найбільшу частоту.

Медіана (Median) – середнє значення вибірки, або значення, вище та нижче якого розташовується однакова кількість спостережень. Для знаходження медіани необхідно впорядкувати дані.

Середнє (Mean) визначається як середнє арифметичне вибірки, тобто сума всіх значень вибірки, поділена на її обсяг.

Сума (Sum) – сума всіх значень змінної, що є додатковим показником для подальших розрахунків.

Міри центральної тенденції показують загальні характеристики розподілу даних за певною змінною та застосовуються з метою найбільшого стиснення інформації. Необхідність їхнього застосування зумовлена, насамперед, потребою порівнювати між собою значення певних характеристик сукупності. При застосуванні міри центральної тенденції важливо пам'ятати, що вони дійсно відображають тенденцію лише у досить однорідних групах, тобто у групах, де варіація досліджуваної ознаки не дуже велика (коефіцієнт варіації менше, ніж 33%).

3. Міри варіації (Dispersion).

Стандартне відхилення (Std. deviation) – це найбільш розповсюджений показник варіації, що демонструє відхилення від середнього значення значень

досліджуваної змінної. Стандартне відхилення є квадратним коренем із дисперсії вибірки.

Дисперсія (Variance) також, як і стандартне відхилення, є мірою розсіювання значень досліджуваної змінної відносно середнього значення. Дисперсія дорівнює квадрату стандартного відхилення.

Розмах (Range) – різниця між найбільшим та найменшим значеннями ознаки.

Крім того SPSS виводить найбільше та найменше значення, які є додатковою інформацією в контексті дослідження варіації ознаки.

Мінімум (Minimum) – мімальне значення ознаки.

Максимум (Maximum) – максимальне значення ознаки.

Стандартна похибка середнього (S.E. mean) дорівнює відношенню стандартного відхилення до квадратного кореня із обсягу вибірки. У наукових публікаціях емпіричні дані зазвичай підсумовуються з використанням середнього значення і стандартного відхилення вибірових даних або середнього значення та стандартною помилкою середнього значення. *Це інколи призводить до плутанини щодо взаємозамінності стандартного відхилення та стандартної помилки середнього значення.* У цьому контексті слід пам'ятати, що середні значення та їх стандартне відхилення є описовими статистиками, що розраховуються на основі досліджуваної вибірки. Стандартна похибка середнього характеризує процес формування випадкової вибірки. Стандартне відхилення даних вибірки являє собою опис варіації значень вимірних ознак, в той час як стандартна похибка середнього є імовірнісним твердженням про те, як розмір вибірки забезпечує кращу оцінку генеральної сукупності на основі вибіркової сукупності. Простіше кажучи, стандартна похибка вибіркового середнього є оцінкою того, наскільки далеко вибіркоче середнє, ймовірно, буде від вибіркового середнього значення, в той час як стандартне відхилення вибірки є ступінь, в якій окремі значення всередині вибірки відрізняються від вибіркового середнього. Саме стандартна похибка вибіркового середнього дозволяє зрозуміти, що збільшення розміру вибірки не завжди призводить до підвищення точності результатів.

Поряд із мірами центральної тенденції для опису даних необхідно наводити й характеристики, що описують ступінь мінливості (варіації, розсіювання) ознаки. Аналіз значень мір центральної тенденції без урахування ступеня варіації можуть дати помилкові результати. Наприклад, середнє значення можна застосовувати лише для однорідних вибірок, тобто таких, де коефіцієнт варіації менше, ніж 33%. У разі неоднорідних вибірок, коли коефіцієнт варіації більше 33%, середнє значення дає результати, які можна віднести до «нахабної брехні». Найвідомішим прикладом є усереднення прибутку найбідніших верств населення і невеликої кількості олігархів.

4. Характеристики форми розподілу (Distribution).

Skewness (асиметрія) – характеристика розподілу, яка повідомляє про наявність або відсутність симетрії даних. Якщо значення більше 0, асиметрія називається позитивною; крива розподілу зміщена вліво або в бік менших

значень. Якщо значення менше 0, асиметрія називається негативною; крива розподілу зміщена вправо або в бік більших значень.

Kurtosis (ексцес) – характеристика розподілу, яка характеризує крутість кривої розподілу. Якщо значення більше 0, ексцес називається позитивним; крива розподілу більш гостроверха, ніж нормальна крива. Якщо значення менше 0, ексцес називається негативним; крива розподілу більш полого, ніж нормальна крива.

Практично будь-які емпіричні дані тієї чи іншою мірою відхиляються від нормального розподілу ймовірностей, закону якого підкоряються розподіли випадкових величин. Але оскільки всі розрахунки, що включають значення середнього арифметичного і стандартного відхилення, засновані на теорії ймовірності, до аналітичного завдання дослідника входить оцінка (хоча б, приблизна) того, наскільки правомірно використовувати даний тип аналізу до отриманих результатів. Тому навіть на рівні дескриптивного аналізу (не кажучи вже про аналіз, націлений на поширення результатів вибіркового дослідження на генеральну сукупність), перш ніж наводити дані за їхніми середніми значеннями (середнє арифметичне та стандартне відхилення), необхідно оцінити характер форми розподілу, тобто виявити наскільки аналізований емпіричний розподіл відрізняється від нормального розподілу. Для цього використовують показники асиметрії (skewness) і ексцесу (kurtosis).

2.3. Візуалізація одновимірних розподілів

Під візуалізацією розуміють наочне подання інформації у вигляді графіків, діаграм, структурних схем, карт тощо.

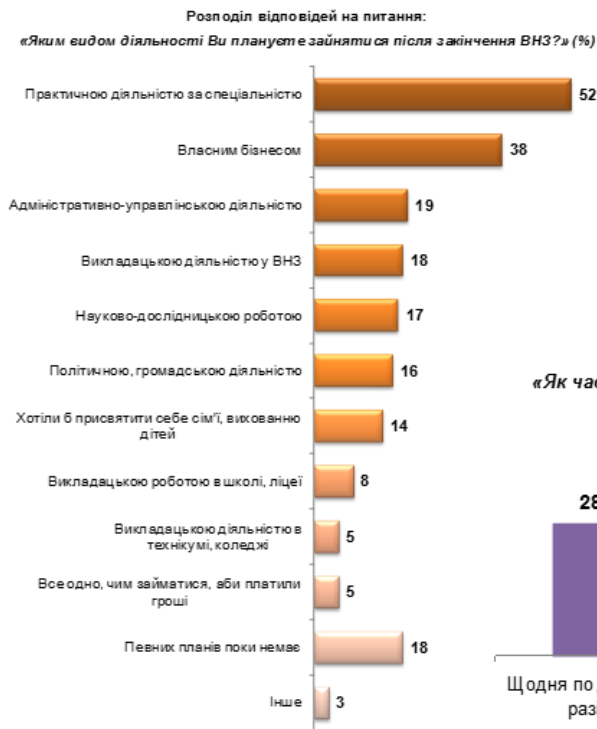
Візуалізація інформації сприяє:

- Спрощенню сприйняття матеріалу;
- Поглибленню розуміння даних;
- Концентрації уваги на найголовніших фактах/аспектах інформації;
- Виокремленню найсуттєвішого змісту даних.

Головні методи візуалізації одновимірних розподілів:

- ✓ стовпчикові діаграми;
- ✓ кругові діаграми;
- ✓ лінійні графіки.

Стовпчикова діаграма – найпоширеніший засіб візуалізації. Вона відображає кілька елементів даних у вигляді стовпців, «зростаючих» в заданому напрямку від базової лінії. Стовпчикова діаграма наочно відображає різницю у значеннях категорій, завдяки тому, що розміри стовпців пропорційні значенням відповідних елементів даних.



Масив: усі опитані перш окурники (n=1002)

Масив: студенти, які зареєстровані в соціальних мережах (n=942)

Рис. 2.9. Приклади стовпчикових діаграм

Кругові діаграми (секторні) застосовують для зображення відношення розмірів елементів, що утворюють ряд, до суми усіх елементів. У круговій діаграмі кожному елементу відповідає сектор, градусна міра якого є пропорційною величині елементу. Такий різновид діаграм доцільно застосовувати, коли необхідно зобразити складові частини цілого.

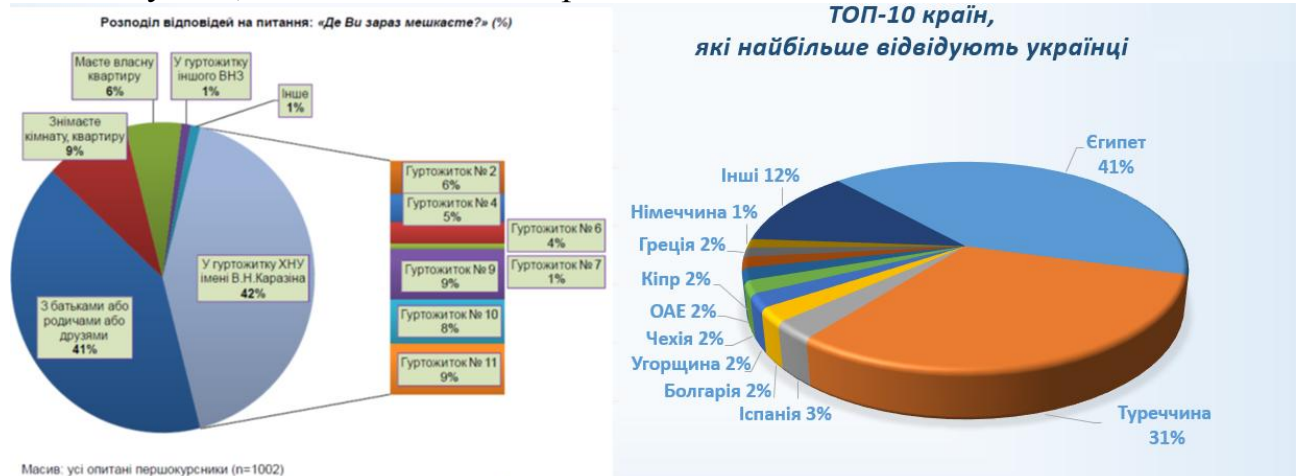


Рис. 2.10. Приклади кругових діаграм

Лінійні графіки застосовують для ілюстрації динаміки феномену, що характеризується зміною статистичних показників у часі.



Рис. 2.11. Приклад лінійного графіку

Серед інших засобів візуалізації одновимірних розподілів акцентуємо увагу на таких: 1) ящикові (коробчасті) діаграми, які ще називають ящик з вусами чи біржова діаграма; 2) картограми; 3) картодіаграми.

Діаграма «ящик з вусами» (інші назви: boxplot, коробчата діаграма, біржова діаграма) є дуже інформативним засобом візуального представлення одновимірних розподілів. Вона дозволяє одночасно зобразити шість величин, що характеризують варіаційний ряд: міжквартильний розмах (IQR - Interquartile range), мінімальне і максимальне спостережувані значення (обчислювані як різниця першого квартиля Q_1 і $1,5 \times IQR$; сума третього квартиля Q_3 і $1,5 \times IQR$), медіану, перший і третій квартили (Q_1 і Q_3 , звані ще 25-й і 75-й процентилями).

Унікальність цієї діаграми полягає в тому, що на ній не тільки представлені основні характеристики розподілу, але і міжквартильний розмах (IQR - довжина «ящика») і його асиметрія (розташування «ящика»). Крім того, на ній відображаються аномальні значення: 1) «викиди» - значення, віддалені від кордонів більш ніж на півтори довжини прямокутника (позначаються кружками); 2) екстремальні значення - значення, віддалені від кордонів більш ніж на три довжини побудованого прямокутника (позначаються на діаграмі зірочками). Межами ящика служать перший і третій квартили (тобто 25-й і 75-й процентиля відповідно), лінія всередині ящика - медіана (50-й процентиля). Кінці вусів - краї статистично значущої вибірки (без викидів) (див. рис.2.12). Таким чином, діаграма «ящик з вусами» дозволяє аналізувати і порівнювати між собою розподіл ознак цілком, а не окремі їх частини, а також демонструє відхилення емпіричного розподілу від нормального. У світовій практиці такі діаграми часто використовуються як більш інформативний аналог звичних гістограм.

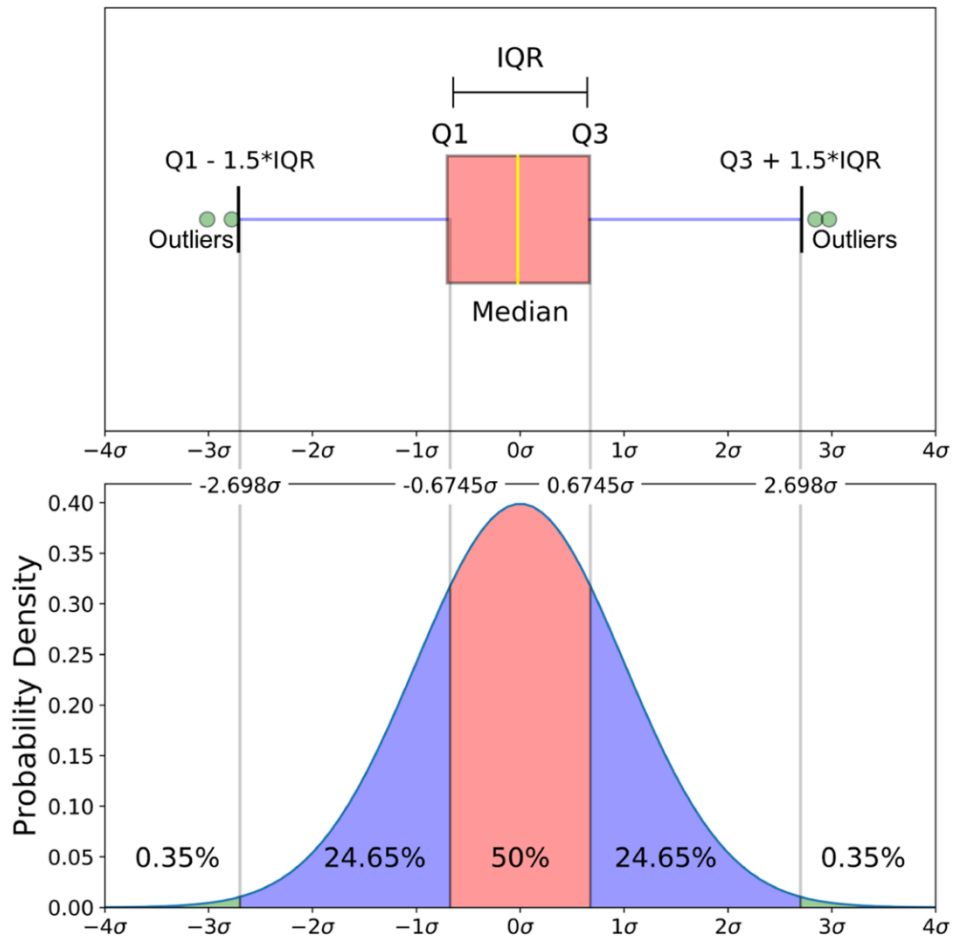


Рис. 2.12. Крива нормального розподілу та діаграма «ящик з вусами»

3. Картограма - це спосіб картографічного зображення (але не карта), що візуально демонструє інтенсивність будь-якого показника в межах території на карті (напр., щільність позашлюбної народжуваності по областям).

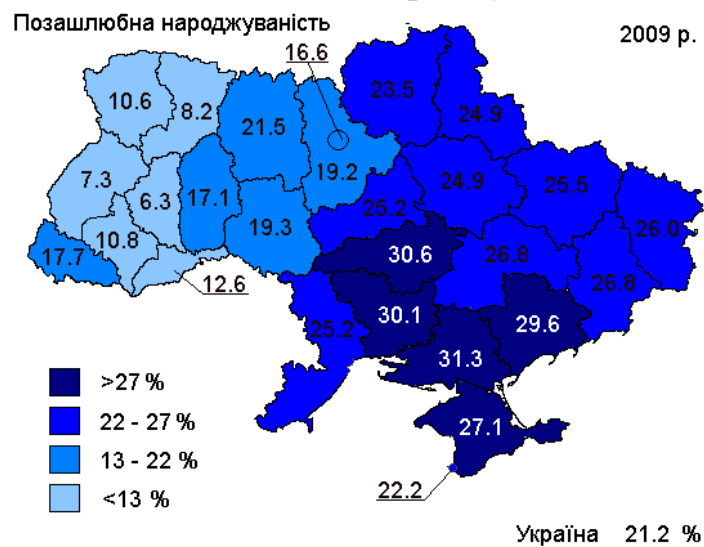


Рис. 2.13. Приклад картограми

4. Картодіаграма (карта і діаграма) – це схематична географічна карта, на якій за допомогою фігур (стовпчиків, кіл, квадратів) показано сумарну величину аналізованого явища в межах зображуваних на ній територіальних одиниць (наприклад, онкологічної захворюваності населення).

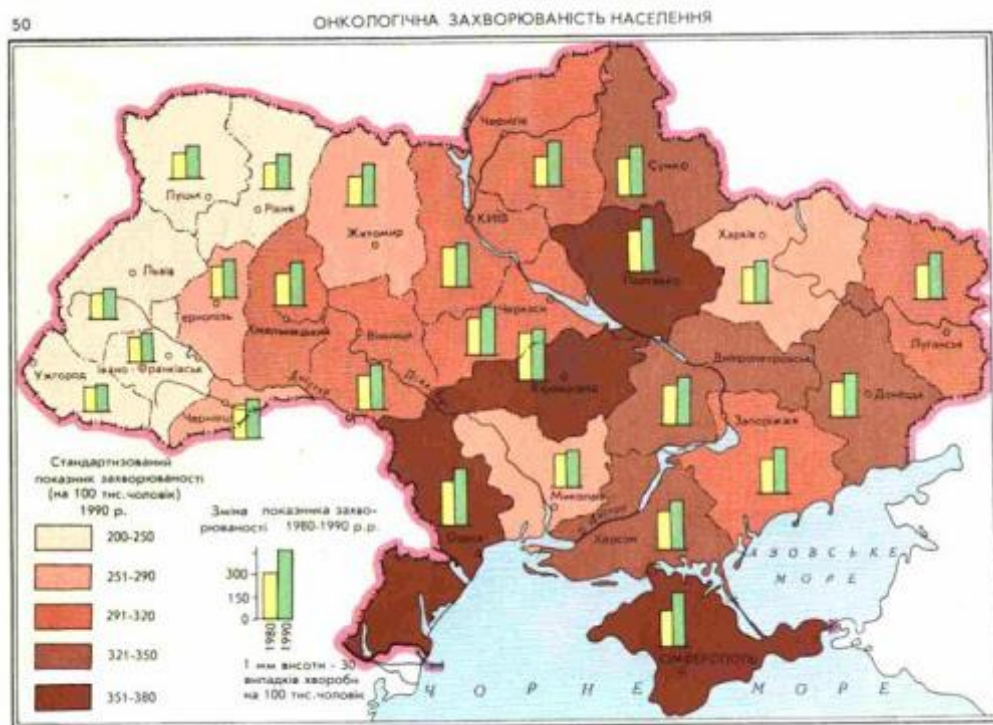


Рис. 2.14. Приклад картодіаграми

2.4. Приклади застосування статистик одновимірних розподілів у вирішенні завдань соціологічного аналізу

У процесі аналізу результатів соціологічних опитувань виникає потреба вирішувати різноманітні завдання. Розглянемо одне з них, яке є дуже розповсюдженим: завдання, пов'язане ранжуванням ознак за значеннями узагальнюючих статистик.

Завдання 1: Побудуйте ієрархію ціннісних орієнтацій першокурсників ХНУ (Масив: Univer_1_курс.sav). Емпірична база – масив даних, отриманий в результаті соціологічного дослідження студентів перших курсів Харківського національного університету імені В.Н. Каразіна, проведеного у період з 24 жовтня по 8 листопада 2011 року співробітниками науково-дослідного Інституту соціально-гуманітарних досліджень спільно з соціологічним факультетом Харківського національного університету імені В.Н. Каразіна (опитано 1002 особи методом анкетування; опитування суцільне (за винятком іноземних студентів), кількість опитаних на різних факультетах пов'язана з рівнем відвідуваності потокових занять серед студентів).

Вирішення даного завдання передбачає аналіз відповідей опитаних респондентів на 18 запитань щодо важливості певних цінностей, які увійшли до анкети: респондентів запитували «Наскільки цінними особисто для Вас є

... ?» (нижче були перелічені різноманітні цінності, які респонденти мали оцінити за наступною шкалою: **5** - дуже цінно; **4** – скоріше цінно; **3** – частково цінно, частково ні; **2** – скоріше не цінно; **1** - зовсім не цінно; **0** – важко відповісти).

Для створення ясної картини ціннісних пріоритетів, що не замутилена зайвими деталями, ми маємо перейти від аналізу 18 одновимірних розподілів до аналізу мір центральної тенденції, тобто наше завдання зводиться до розрахунку 18 чисел, які треба упорядкувати, щоб отримати ієрархію ціннісних орієнтирів.

Насамперед необхідно визначити, яку міру центральної тенденції ми будемо застосовувати у подальшому аналізі. Вибір завжди зумовлюється типами шкал, якими вимірювалися досліджувані ознаки. В нашому випадку шкали номінальні, але можуть бути трансформовані у порядкові з п'ятьма альтернативами відповіді (1-5), якщо відповідь «важко відповісти» (0) враховувати як пропущене значення. Відомо, що для порядкових шкал адекватною мірою центральної тенденції є медіана. Проте такі шкали часто розглядають, як псевдоінтервальні, для яких можна застосовувати середні значення. Отже ми можемо обирати між медіаною та середнім. Безумовно, аналіз середніх значень є більш розповсюдженим та звичним, у зв'язку з чим ми віддаємо йому перевагу. Проте застосовувати середні значення, як відомо, можна лише для однорідних вибірок (коефіцієнт варіації менше, ніж 33%). Перевірка даної умови (див. табл. 2.2) демонструє однорідність вибірки за всіма аналізованими ознаками, що дає нам можливість аналізувати середні значення та візуалізувати ціннісну ієрархію (рис. 2.14).

Таблиця 2.2

Ціннісні орієнтації першокурсників

(середні значення; інтервал від 1 – зовсім не цінно до 5 – дуже цінно)

Цінності	Середнє значення	Стд. відхилення	Коеф. варіації	Ранг
Цікава, творча робота	4,25	0,884	21	11
Матеріальне благополуччя	4,19	0,817	19	13
Гарні, добрі відносини з оточуючими людьми	4,56	0,761	17	6
Можливість приносити користь суспільству	4,06	0,908	22	14
Участь у суспільному житті, у вирішенні проблем, що стоять перед суспільством	3,59	0,985	27	18
Освіченість, знання	4,65	0,598	13	4
Особистий спокій, відсутність неприємностей	4,40	0,855	19	8
Сімейне благополуччя	4,78	0,572	12	1
Здоров'я	4,67	0,668	14	3

Повноцінний відпочинок, цікаві розваги	4,39	0,770	18	9
Високе службове і суспільне становище	3,97	0,976	25	15
Прилучення до літератури і мистецтва	3,65	1,050	29	17
Екологічна безпека	3,75	0,968	26	16
Взаємопорозуміння з батьками, старшим поколінням	4,43	0,778	18	7
Особиста свобода, незалежність у судженнях і діях	4,57	0,660	14	5
Можливість розвитку, реалізації своїх здібностей, талантів	4,69	0,551	12	2
Економічна незалежність	4,37	0,790	18	10
Побутовий комфорт	4,25	0,845	20	12

Масив: всі опитані першокурсники (n=1002)

Перш за все звернемо увагу на трійку цінностей–лідерів: сімейне благополуччя, самореалізація та здоров'я. Відзначимо, що цінності здоров'я та сім'ї незмінно, як свідчать чисельні соціологічні дослідження, посідають перші рангові місця у ціннісній ієрархії студентства.

Розглянемо кожну з цих цінностей окремо та поміркуємо, з чим це може бути пов'язано.

Перше рангове місце посідає цінність сімейного благополуччя. Загальновідомо, що сім'я - природна частина життя кожної людини, всі люди на різних етапах свого життя так чи інакше пов'язані з сім'єю. Головне призначення сім'ї - створювати умови для душевного відпочинку, бути у бурхливому океані життя тією «тихою гаванню», де людина може відпочити, знайти порозуміння та любов. Прагнення до сімейного благополуччя у цьому контексті представляє собою спосіб створення надійного тилу, що набуває особливої значущості у сучасному світі, де тенденції індивідуалізації мають явно виражені риси.



Рис. 2.15. Ціннісні орієнтації першокурсників (середні значення; інтервал від 1 – зовсім не цінно до 5 – дуже цінно)

На другому ранговому місці опинилась цінність самореалізації, яка, як відомо, є поширеною загальною мотивацією життя, що досягає піку своєї актуальності саме в період молодості, коли молода людина, в силу своїх психофізичних і ментальних особливостей, володіє найбільшою енергією і працездатністю. Цей пошук своєї цілісності і відрізняє молодість від більш пізніх етапів життя, коли самореалізація зі статусу нездійсненого бажання переходить в розряд здійснюваної потреби. В цілому, для студентської молоді пошук та реалізація своїх здібностей, талантів виступає головним завданням, яке вирішується нею безпосередньо в стінах вузу. Досвід повноцінної самореалізації у вузівському навчанні стає надалі надійною основою вибудовування життєвого шляху особистості. Наші студенти, відповідаючи на питання стосовно цінності особисто для себе можливості розвитку, реалізації своїх здібностей, талантів, оцінюючі її значущість, продемонстрували своє розуміння такої ситуації. Зокрема, важливість розвитку саме *всіх* своїх талантів, а не тільки підвищення рівня освіченості. Це можна побачити порівнюючи цю цінність з цінністю освіченості, придбання знань, яка зайняла четверте рангове місце в ціннісній ієрархії першокурсників. Цінність самореалізації оцінюється першокурсниками як дещо важливіша (середнє значення 4.69 проти 4,65).

Важливість цінності здоров'я високо оцінюється нашими респондентами. Вона посідає третє рангове місце в їх ціннісній ієрархії. Здоров'я - безцінне надбання не тільки кожної людини, а й усього

суспільства. Здоров'я допомагає нам здійснювати наші плани, успішно вирішувати основні життєві завдання, долати труднощі, а якщо доведеться, то й значні перевантаження. Відомо, що здоров'я і життя стають цінністю для людини саме тоді, коли їм починають реально загрожувати хвороби і смерть. Висока значущість цінності здоров'я, виявлена серед першокурсників, мабуть обумовлюється тим, що життя сучасної людини пов'язано з певними (так би мовити «побутовими») ризиками: екологічною кризою, що негативно впливає на здоров'я мешканців України; неякісними продуктами харчування, що заповнили полиці наших супермаркетів, інформаційним перевантаженням, яке часто призводить до нервових зривів, тощо. Недарма сьогодні модно бути здоровим. Наші першокурсники орієнтуються на ідеали здорового способу життя: не курять (82%), не пробують наркотики (89%). Мабуть вони вважають, що серед буденних ризиків, яких не уникнути сучасній людині, немає сенсу самотійно створювати собі зайві проблеми.

Не можна не відмітити також цінності, що, як довело проведене дослідження, уявляються найменш важливими для опитаних. Так, цінність «участь у суспільному житті, у вирішенні проблем, що стоять перед суспільством» (середнє значення дорівнює 3,59) посіла останнє рангове місце, а «прилучення до літератури і мистецтва» (3,65) – передостаннє.

Завдання 2: Дослідити питання щодо місця комп'ютерних ігор та спілкування у соціальних мережах у структурі дозвілля підлітків. (Масив: *Підліток2013.sav*). Емпірична база – масив даних, отриманий в результаті соціологічного дослідження «Життєвий світ підлітків Харківщини», проведеного в 2013 році кафедрою соціології Харківського національного університету імені В. Н. Каразіна (опитано 1909 осіб, з них 986 - підлітки м. Харкова, 923 - Харківської області).

У якості узагальнюючої характеристики використаємо медіану, оскільки вона є мірою центральної тенденції, що адекватна порядковому рівню вимірювання, використаному в інструментарії. Однак при цьому ми врахуємо і інші важливі параметри одновимірних розподілів: верхній та нижній квартили, мінімальні та максимальні значення кожної ознаки вибірки, а також пропущені та екстремальні значення (викиди), що дозволить визначити рангові місця конкретних дозвіллевих практик в структурі вільного часу.

Рис. 2.16 демонструє, що спілкування в соціальних мережах відноситься до щоденних дозвіллевих практик і займає друге місце (після прослуховування музичних записів) в структурі дозвілля школярів. Комп'ютерні ігри займають шосте місце і практикуються більшістю підлітків кілька разів на тиждень.

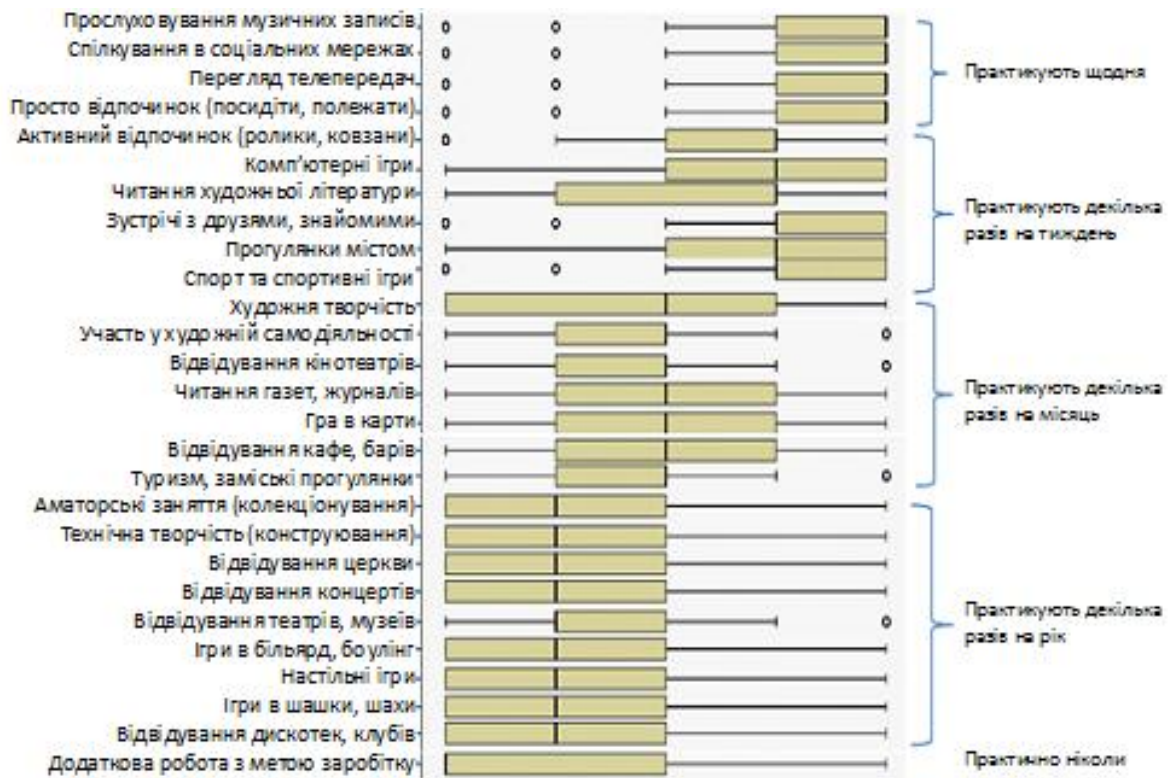


Рис. 2.16. Розподіл відповідей підлітків на запитання: «Чим і як часто Ви займаєтеся у вільний час? ...» (шкала виміру: 5 - щодня, 4 - кілька разів на тиждень, 3 - кілька разів на місяць, 2 - кілька разів на рік, 1 - практично ніколи)

Аналізуючи отримані дані, можна побачити тенденцію, яка полягає в тому, що з віком школярі все менше уваги приділяють комп'ютерним іграм (це центральна тенденція по всьому масиву), а спілкування у соціальних мережах стабільно залишається щоденною дозвільною практикою для більшості опитаних. Візуалізацію цієї тенденції ми здійснимо за допомогою діаграми «ящик з вусами».

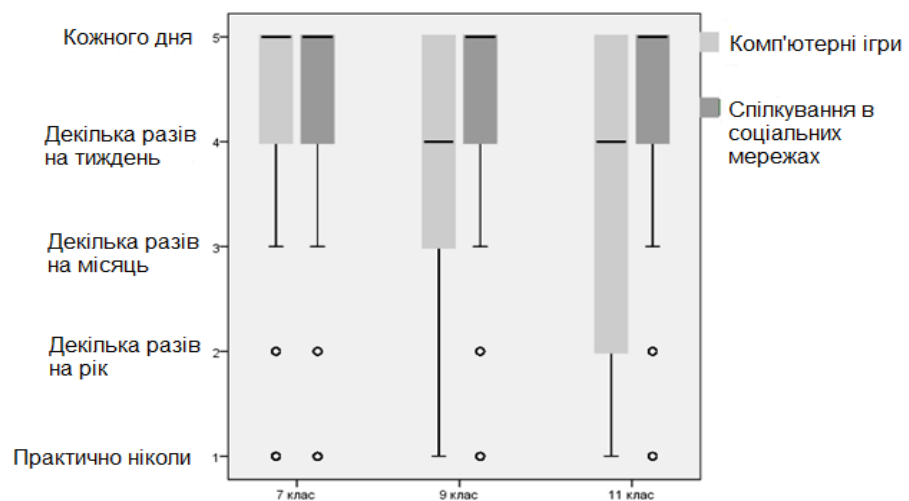


Рис. 2.17. Діаграма «ящик з вусами», що представляє розподіли відповідей учнів 7, 9 та 11 класів на запитання: «Як часто Ви граєте в комп'ютерні ігри?» і «Як часто Ви спілкуєтеся в соціальних мережах?»

Як видно на рис. 2.17, практично всі семикласники багато часу приділяють комп'ютерним іграм (медіана дорівнює 5 - грають щодня, міжквартильний розмах дорівнює 1, а такі варіанти, як «граю кілька разів на рік» і «практично ніколи не граю» відносяться до аномальних явищ). Дев'ятикласники більш різноманітні за ознакою, що характеризує їх ставлення до комп'ютерних ігор. Ми бачимо, що центральною тенденцією для них є «грати кілька разів на тиждень» (медіана = 4, міжквартильний розмах дорівнює 2), а такі варіанти, як «граю кілька разів на рік» і «практично ніколи не граю» перестають бути аномальними. Учні 11 класу ще менше часу присвячують комп'ютерним іграм. Незважаючи на те, що медіана також дорівнює 4 («кілька разів на тиждень»), міжквартильний розмах дорівнює 3, що демонструє підвищення варіативності відповідей школярів, збільшення ваги тих, хто грає в комп'ютерні ігри вкрай рідко. При цьому всі розподіли асиметричні, що демонструє поширеність у вже згадуваному масиві значень 5 («грають щодня») і 4 («кілька разів на тиждень»).

Література до теми

Бююль, А., Цёфель П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005. С. 91-103.

Крыштановский, А.О. *Анализ социологических данных с помощью пакета SPSS*. М.: ГУ ВШЭ, 2007. С. 10-38.

Паніна, Н.В. *Технологія соціологічного дослідження: Курс лекцій / 2-е видання, доповнене*. К., 2007. С. 221-240.

Паніотто В.І., Максименко В.С., Харченко Н.М. *Статистичний аналіз соціологічних даних*. К.: «КМ Академія», 2004. С. 10-64.

Толстова, Ю.Н. *Анализ социологических данных (Методология, дескриптивная статистика, изучение связей между номинальными признаками)*. М.: Научный мир, 2003. С.124-163.

Додаткова література

Желязны, Д. *Говори на языке диаграмм: Пособие по визуальным коммуникациям для руководителей / Пер. с англ.* М.: Институт комплексных стратегических исследований, 2007. 336 с.

Иванов, О.В. *Статистика. Учебный курс для социологов и менеджеров. Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода*. М.: МГУ им. Ломоносова, 2005. 187 с.

Малхотра, Нэреш К. *Маркетинговые исследования. Практическое руководство*. 3-е изд., пер. с англ. М.: Вильямс, 2002. С. 552-601.

Наследов, А.Д. *SPSS 15: профессиональный статистический анализ данных*. СПб.: Питер, 2008.

Реброва, О. Среднее или всё же медиана? *ТрВ-Наука*. 2011. № 90. С. 13. URL <http://trv-science.ru/2011/10/25/srednee-ili-vsjo-zhe-mediana/>

Татарова, Г.Г. *Методология анализа данных в социологии (введение)*. Учебник для вузов. М.: NOTA BENE, 1999.

Питання для самоконтролю

1. Що таке «одновимірний розподіл»?
2. Як будується одновимірний розподіл?
3. Які статистики одновимірного розподілу застосовуються для номінальної шкали?
4. Які статистики одновимірного розподілу застосовуються для порядкової шкали?
5. Які статистики одновимірного розподілу застосовуються для метричної шкали?
6. Визначте міжквартильний розмах для даних з таблиці, що наведена нижче.

Якою мірою Вам притаманні рішучість, готовність до ризику, підприємливість?

		Частота	Відсоток	Валідний відсоток	Кумулятивний відсоток
Валідні	1 Не притаманні	43	1,4	1,5	1,5
	2 Скоріше не притаманні	212	6,9	7,6	9,2
	3 Важко сказати	660	21,6	23,7	32,9
	4 Скоріше притаманні	1058	34,6	38,1	71,0
	5 Притаманні повною мірою	806	26,4	29,0	100,0
	Разом	2779	90,9	100,0	
Пропущені	Немає відповіді	279	9,1		
Разом		3058	100,0		

7. Яким чином можна візуалізувати одновимірний розподіл?

Тема 3. Відбір даних в SPSS: побудова фільтрів

3.1. Фільтри та відбір даних

Відбір даних – це вибір спостережень (анкет) за визначеними критеріями. Так, наприклад, при опитуванні виборців можна відібрати тільки чоловіків, що збираються голосувати за певну партію, а при опитуванні студентів – тільки студенток, які вирішили вступити до вузу тому, що сподівалися зустріти майбутнього супутника життя.

Для здійснення відбору необхідно створити певне правило (задати логічну умову), яке дозволить SPSS виділити з усього масиву даних анкети тих респондентів, що цікавлять дослідника. Після виконання операції відбору всі обчислення будуть проводитися тільки з відібраними спостереженням, що надає можливість досліднику вивчити специфічні характеристики обраної соціальної групи та побудувати її соціологічний портрет.

Логічну умову, за якою здійснюється відбір з масиву певних анкет, називають **фільтром**. Фільтр виділяє з масиву ті анкети, які задовольняють указаній умові (тобто ті, для яких логічна умова істинна).

Коротко розглянемо оператори, які застосовуються у SPSS для створення логічних виразів для відбору анкет з масиву.

Оператори поділяються на арифметичні, логічні та оператори відносин. Арифметичні оператори застосовуються в так званих арифметичних виразах (математичних формулах), що при відборі даних мають лише другорядне значення. Арифметичні оператори можна використовувати й в логічних виразах, однак це зустрічається нечасто.

При створенні фільтрів для відбору анкет головним чином застосовують логічні оператори та оператори відносин.

Оператори відносин

Відношення - це логічне вираження, у якому два значення порівнюються один з одним за допомогою оператора відносини. Частіш за все значення перемінної порівнюються з яким-небудь чисельним значенням (константою), наприклад:

$p1_6 = 1$ (вирішили вступити до вузу тому, що сподівалися зустріти майбутнього супутника життя);

$p146 > 2$ (відчувають себе представниками своєї національності частково або повною мірою).

Для побудови логічних виразів можуть застосовуватися наступні оператори відносин (див. табл. 3.1).

Таблиця 3.1

Оператори відносин

Знак на кнопці	Альтернативний текст	Значення (укр./англ.)
<	LT	Менше (less than)
>	GT	Більше (greater than)
<=	LE	Менше або дорівнює (less than or equal to)
>=	GE	Більше або дорівнює (greater than or equal to)
=	EQ	Дорівнює (equal to)
~=	NE или <>	Не дорівнює (not equal to)

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

Логічні оператори

Для побудови умовних виразів можуть застосовуватися наступні логічні оператори (див. табл. 3.2).

Таблиця 3.2

Логічні оператори

Знак на кнопці	Альтернативний текст	Значення
&	AND	Логічне І
	OR	Логічне АБО
~	NOT	Логічне НЕ

Логічні оператори AND і OR зв'язують між собою відносини, а логічний оператор NOT змінює значення істинності умовного вираження на протилежне. Між логічними операторами встановлюються наступні пріоритети:

Пріоритет	Оператор
1	NOT
2	AND
3	OR

3.2. Способи відбору даних в SPSS

Для відбору даних завантажте у редактор даних файл, що буде аналізуватися та виберіть в меню команди *Data (Дані) → Select Cases ... (Вибрати спостереження)*. Відкриється діалогове вікно *Select Cases* (див. Рис. 3.1). За замовчуванням SPSS встановлює параметр *All cases (Усі*

спостереження), що означає роботу зі всіма анкетами, які містяться у масиві. Дослідник завжди має можливість після аналізу частини масиву (відібраними за будь-якими умовами анкетами) повернутися до роботи зі всіма анкетами, якщо встановить цю опцію.

Data → *Select Cases...*

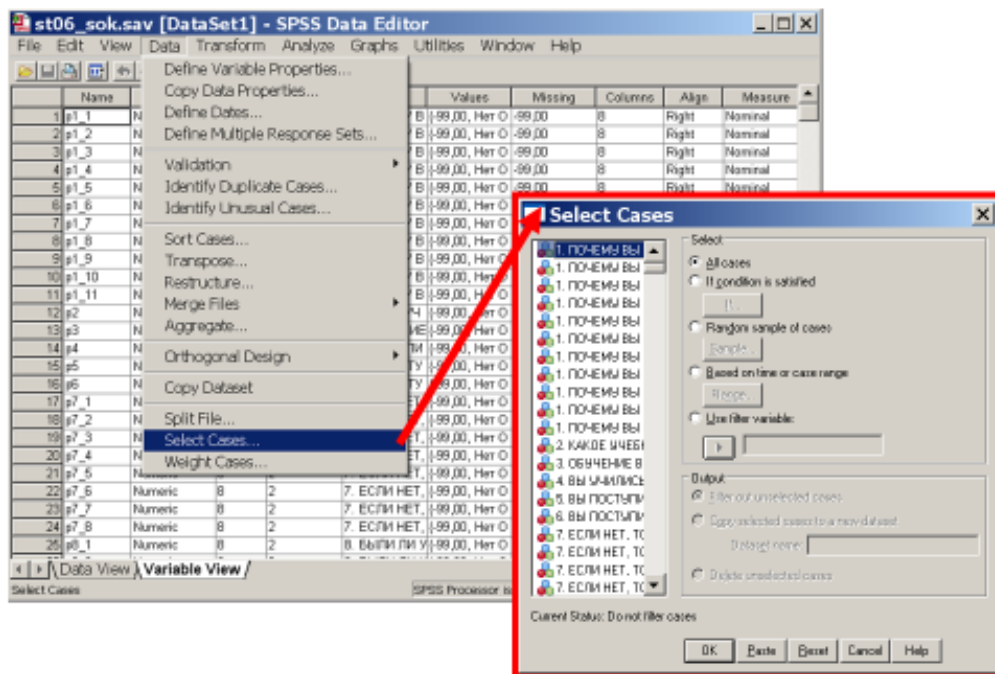


Рис. 3.1. Результат виконання команди *Data (Дані) → Select Cases ... (Вибрати спостереження)*

У діалоговому вікні *Select Cases* представлені наступні можливості відбору (див. рис. 3.1):

- *If condition is satisfied (Якщо виконується умова)*. Вибір спостережень (анкет) по визначеній умові.

- *Random Sample of cases (Випадковий відбір спостережень)*. Витяг випадкової вибірки з масиву анкет.

- *Based on time or case range (На основі часових обмежень або інтервалів спостережень)*. SPSS надає можливість відібрати інтервали по датам або часу для часових рядів, що містять перемінні у вигляді дати. Відзначимо, що потреба у застосуванні такої можливості при аналізі масивів соціологічних анкет практично ніколи не виникає.

- *Use filter variables (Застосовувати перемінну фільтру)*. Встановлення цієї опції дає аналітику можливість вказати у файлі даних перемінну, яка буде застосовуватися з метою фільтрації анкет.

При аналізі результатів соціологічних опитувань частіше за все виникає потреба у відборі анкет по певній умові, але інколи застосовують й інші можливості, зокрема, вилучення випадкової вибірки з масиву даних.

Необхідно звернути увагу на опції *Filtered* та *Deleted*, що розташовані унизу діалогового вікна *Select Cases*. Зазвичай застосовується опція *Filtered*,

яка дає можливість працювати лише з тими анкетами, які цікавлять дослідника при вивченні певного аспекту досліджуваного явища.

Опція *Deleted* дозволяє видалити з масиву даних всі анкети, які не задовольняють певній вимозі. Наприклад, якщо задати вимогу $p_{204} = 1$ (стать = чоловік) та встановити опцію *Deleted*, то з масиву будуть видалені всі анкети, які не задовольняють цій умові, тобто у масиві остануться тільки анкети чоловіків. Зрозуміло, що потрібно застосовувати цю опцію дуже обережно, оскільки можна втратити значну кількість корисних даних.

3.3. Відбір анкет за визначеною умовою

Відбір даних за певною умовою – це дуже поширене завдання, потреба вирішення якого виникає у випадках, коли соціолог прагне дослідити особливості певної групи респондентів, виділених за якимось соціально значимими критеріями (стать, вік, місце проживання, національність, професія, дохід, рівень освіти тощо).

Наприклад, нам необхідно дослідити відношення до політики чоловіків, які повною мірою відчувають себе громадянами України (масив *st06.sav*, ознака p_{147}) та вважають, що Україна у зовнішній політиці повинна передусім орієнтуватися на Росію, Білорусь та інші країни Єдиного економічного простору (ознака p_{152}). В такому разі виникає потреба вилучити з масиву анкет спостереження, які задовольняють певній умові. Щоб це зробити у SPSS необхідно завантажити файл *st06.sav* у редактор даних та виконати команду *Data (Дані) → Select Cases... (Вибрати спостереження)*. У результаті відкриється діалогове вікно *Select Cases* (див. рис. 3.2).

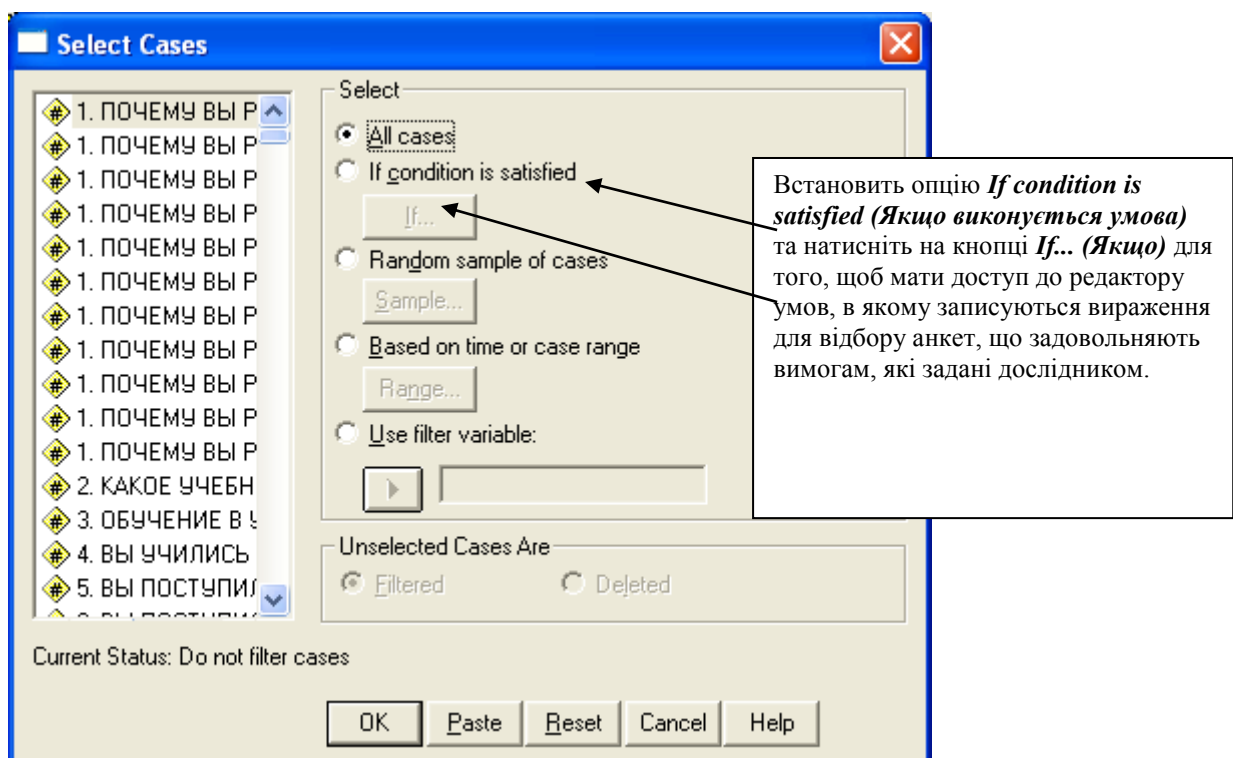


Рис. 3.2. Діалогове вікно *Select Cases*

Для вибору певних анкет з масиву даних необхідно вибрати пункт *If condition is satisfied (Якщо виконується умова)* і клацнути на кнопці *If... (Якщо)*, як показано на рисунку 3.2. Відкриється діалогове вікно *Select Cases: If* (див. рис. 3.3), у якому у полі редактора умов необхідно записати логічний вираз для відбору необхідних анкет зі всього масиву.

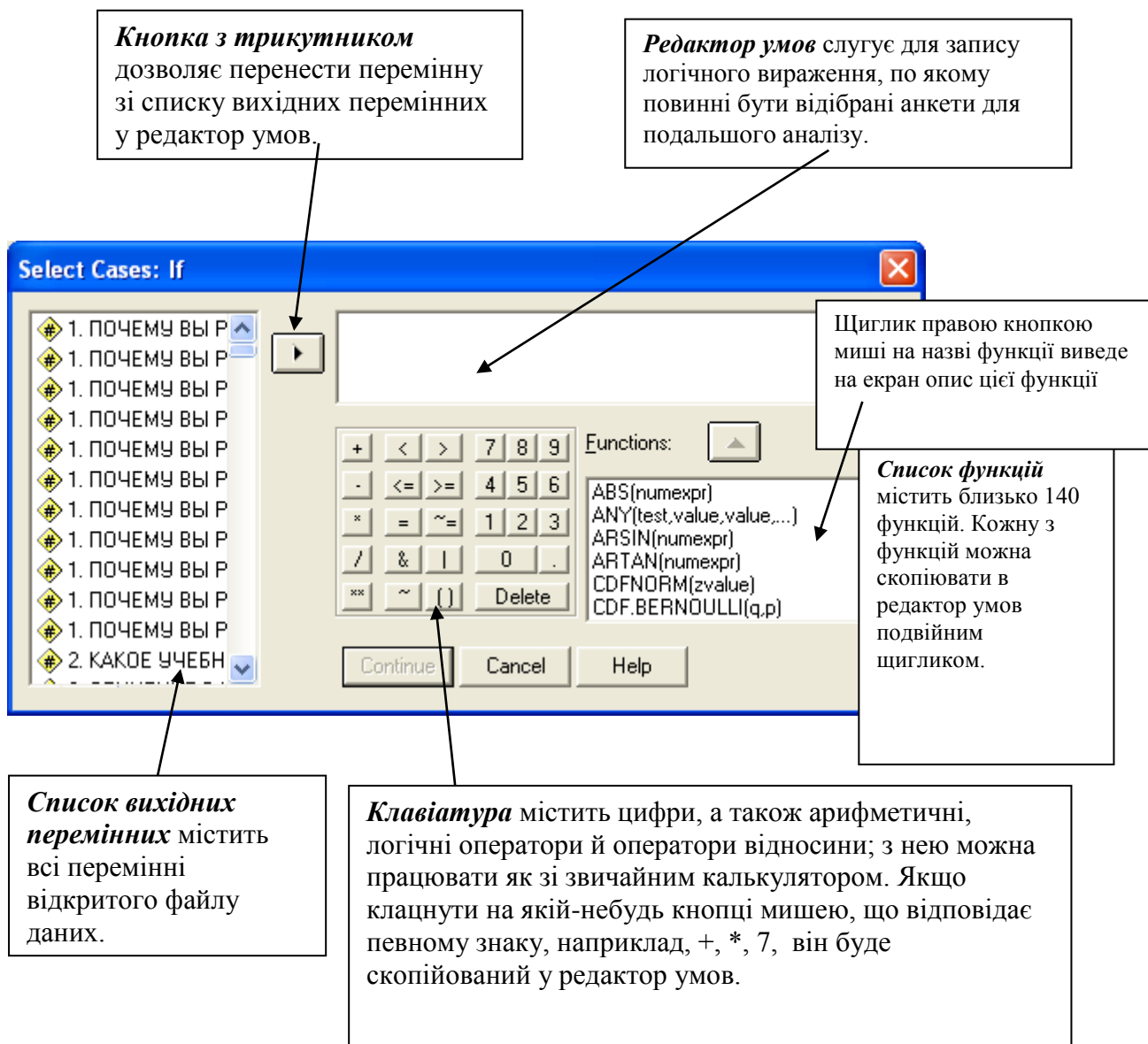


Рис. 3.3. Діалогове вікно *Select Cases: If* (Вибрати спостереження за певною логічною умовою)

У редакторі умов запишіть логічний вираз та натисніть кнопку *Continue*. У результаті Ви повернетесь у діалогове вікно *Select Cases*, в якому необхідно натиснути кнопку *OK* для того, щоб SPSS підключив вказаний фільтр. Після виконання цих дій SPSS буде обробляти лише відфільтровані анкети, тобто такі, що задовільність умові: $p204 = 1 \ \& \ p147 = 4 \ \& \ p152 = 1$.

Для відбору анкет у редакторі умов необхідно записати логічний вираз. Наприклад, для вилучення чоловіків, які повною мірою відчують себе громадянами України та при цьому вважають, що Україна у зовнішній політиці повинна передусім орієнтуватися на Росію, Білорусь та інші країни Єдиного економічного простору, логічний вираз можна записати таким чином

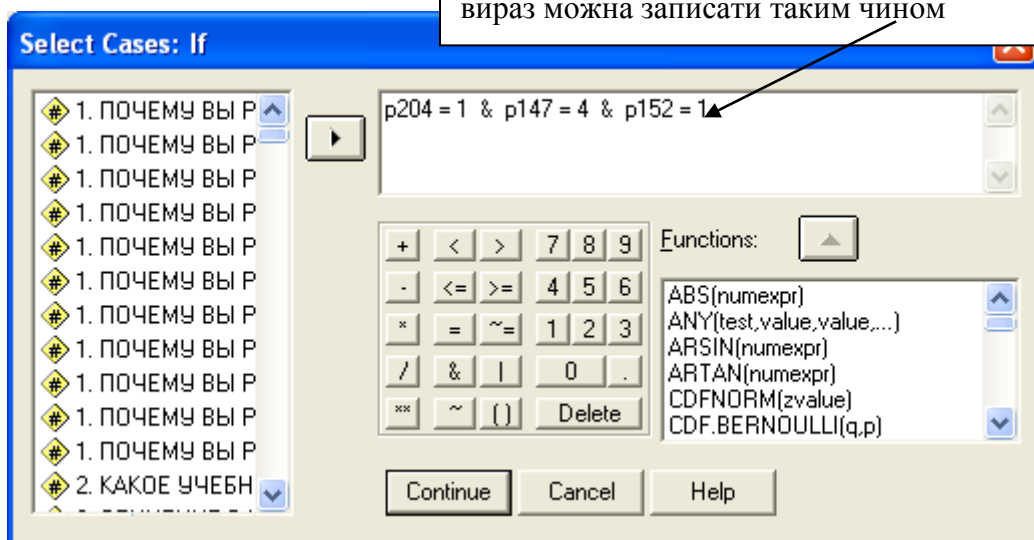


Рис. 3.4. Діалогове вікно *Select Cases: If (Вибрати спостереження: Якщо)*, у якому міститься логічний вираз для вилучення з масиву анкет, що відповідають умові, яка задана аналітиком

Побудуйте одномірний розподіл за ознакою 157. Цей розподіл, як Ви самі побачите, буде побудований лише для 156 анкет, які задовольняють умові, що була Вами задана (див. табл. 3.3). Проаналізуйте отриманий результат.

Таблиця 3.3

Одномірний розподіл відповідей на питання «157. Ваше ставлення до політики, політичних процесів у суспільстві?» серед студентів, які відібрані за умовою $p204 = 1 \ \& \ p147 = 4 \ \& \ p152 = 1$

157. Ваше ставлення до політики, політичних процесів у суспільстві?		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. Зацікавлене, постійно стежите за подіями	45	28,8	29,2	29,2
	2. Періодично цікавитесь (незалежно від подій)	81	51,9	52,6	81,8
	3. Байдужі до політики	11	7,1	7,1	89,0
	4. Принципово не цікавитесь	4	2,6	2,6	91,6

	5. Політика викликає у Вас роздратування	13	8,3	8,4	100,0
	6. Інше				
	Total	154	98,7	100,0	
Missing	Немає відповіді	2	1,3		
Total		156	100,0		

Далі проаналізуйте ставлення студентської молоді до політики, політичних процесів у суспільстві (масив st06, ознака p157). Для цього поверніться до роботи зі всім масивом (виконайте команду **Data (Дані) → Select Cases... (Вибрати спостереження)** та встановіть опцію **All cases (Усі спостереження)**. Розрахуйте одномірний розподіл за ознакою p157, у результаті Ви отримаєте наступну таблицю (див. табл. 3.4).

Таблиця 3.4

Одномірний розподіл відповідей студентів на питання «157. Ваше ставлення до політики, політичних процесів у суспільстві?»

157. Ваше ставлення до політики, політичних процесів у суспільстві?		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. Зацікавлене, постійно стежите за подіями	594	19,4	19,8	19,8
	2. Періодично цікавитесь (незалежно від подій)	1738	56,9	57,9	77,7
	3. Байдужі до політики	294	9,6	9,8	87,5
	4. Принципово не цікавитесь	111	3,6	3,7	91,2
	5. Політика викликає у Вас роздратування	257	8,4	8,6	99,8
	6. Інше	7	,2	,2	100,0
	Total	3001	98,2	100,0	
Missing	Немає відповіді	56	1,8		
Total		3057	100,0		

Порівняйте результати, що представлені у таблицях 1 та 2. Зробіть висновки.

Розміркуйте, які групи студентської молоді можуть відрізнитися від загальної маси за своїм ставленням до політики, політичних процесів у суспільстві. Висуньте свої гіпотези та запропонуйте засоби їх перевірки.

3.4. Витяг випадкової вибірки спостережень з файлу даних

Необхідність відібрати випадкову частину спостережень з масиву даних виникає найчастіше при першій попередній перевірці дослідницьких гіпотез та вивченні стійкості певних показників.

Щоб витягти випадкову вибірку із сукупності всіх спостережень, необхідно виконати наступні дії:

Виберіть у меню команди **Data (Дані) → Select Cases... (Вибрати спостереження)**. У діалоговому вікні **Select Cases**, яке з'явиться на екрані після виконання команди, виберіть пункт **Random sample of cases (Випадкова вибірка)**, як показано на рис. 3.5. Потім клацніть на кнопці **Sample... (Вибірка)**, щоб відкрити діалогове вікно **Select Cases: Random Sample (Вибрати спостереження: Випадкова вибірка)** (див. рис. 3.6).

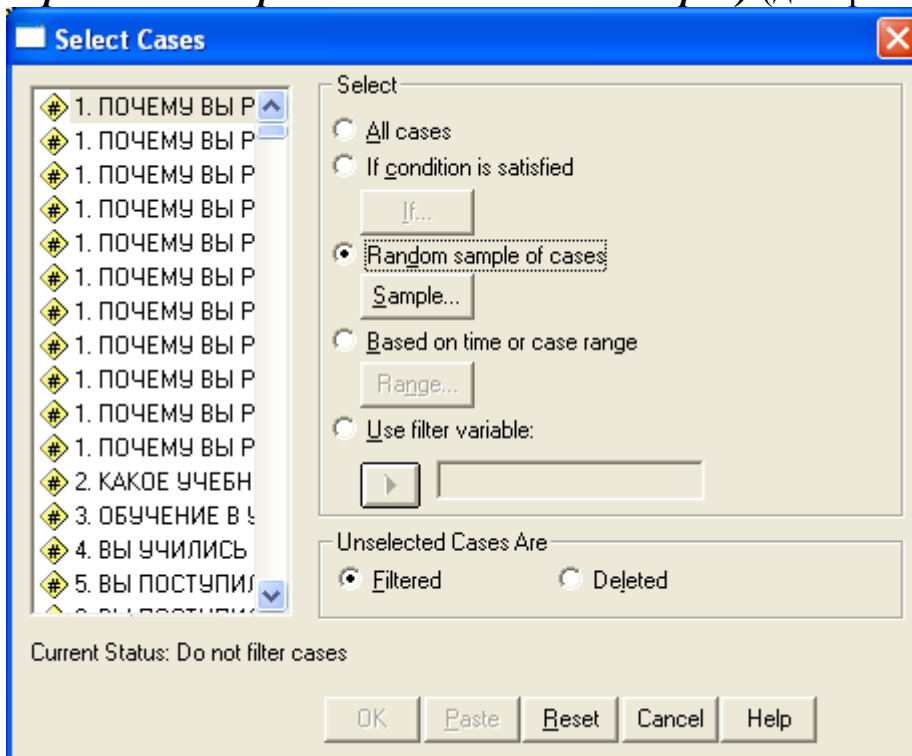
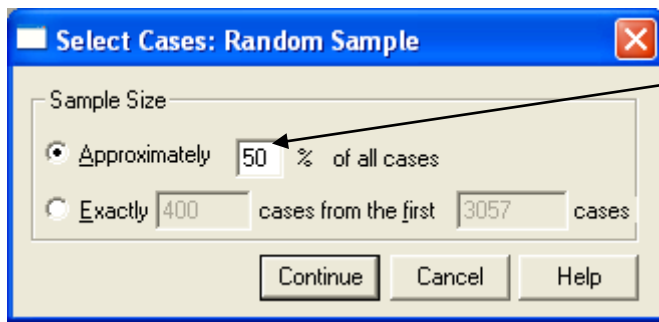


Рис. 3.5. Діалогове вікно **Select Cases** з встановленою опцією **Random Sample of cases**

Витяг випадкової вибірки з масиву анкет у SPSS можна здійснити одним з наступних способів:

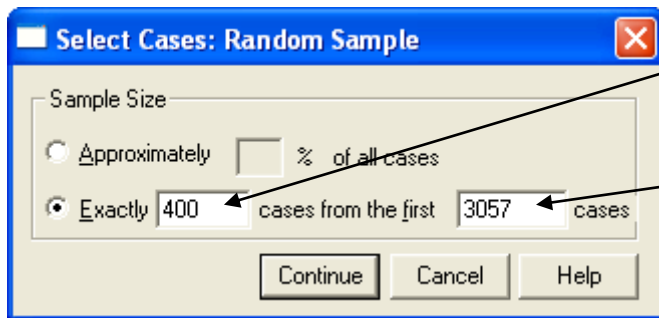
- **Approximately** (Приблизно). Аналітик вказує процент анкет, які потрібно відібрати зі всього масиву. SPSS створить випадкову вибірку з обсягом, що приблизно відповідає зазначеному відсотковій спостережень. Наприклад, можна відібрати 50% анкет зі всього масиву (див. рис. 3.6 а).

- **Exactly** (Точно): Аналітик указує точну кількість анкет у випадковій вибірці. Крім того, тут треба задати кількість спостережень, з яких буде витягнута вибірка. Звісно, що це число не може бути меншим ніж кількість анкет у вибірці та не повинне перевищувати загальної кількості анкет у файлі даних. При аналізі масивів соціологічної інформації зазвичай для створення вибірки з масиву анкет вказують кількість анкет у масиві. Так, наприклад, щоб з масиву, який містить 3057 анкет, випадково вилучити 400 анкет, необхідно задати такі параметри, які вказані на рисунку 3.6 б.



Вкажіть який процент анкет потрібно випадково відібрати зі всього масиву

Рис. 3.6 а. Відбір 50% анкет зі всього масиву



1. Вкажіть кількість анкет у вибірці

2. Вкажіть кількість анкет , з яких буде вилучатися випадкова вибірка

Рис.3.6 б. Вилучення 400 анкет з масиву, що містить 3057 анкет

Рис. 3.6. Два способи створення випадкової вибірки з масиву анкет

Необхідно мати на увазі, що кожного разу вилучається нова випадкова вибірка з наявного масиву анкет, тобто повторне застосування даної операції дасть результати відмінні від попередніх.

Література до теми

Бююль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005. С. 104-121.

Наследов, А.Д. *SPSS: Компьютерный анализ данных в психологии и социальных науках*. СПб.: Питер, 2005. С. 74-76.

Наследов, А. *IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных*. – СПб.: Питер, 2013. С. 61-64.

Додаткова література

Горбачик, А.П., Сальникова, С.А. *Анализ данных социологических исследований средствами SPSS*. Луцьк: «Вежа», 2008.

Питання для самоконтролю

1. Для чого призначені фільтри?
2. Коли у соціолога виникає потреба відібрати з масиву даних певні спостереження?
3. Які існують способи відбору спостережень в SPSS?
4. Соціологічна інформація та соціальна інформація. В чому різниця?
5. Чому англomовному сегменті немає визначення терміну «sociological information»?

6. Що таке масив даних? Чи тотожні поняття «масив даних» та «масив анкет»?
7. Чим відрізняється обробка даних від аналізу даних?
8. У чому полягають методологічні принципи аналізу соціологічних даних?

Тема 4. Модифікація даних в SPSS: створення нових змінних

Для проведення аналізу часто буває необхідно виконати перетворення даних, тобто на основі зібраних даних створити нові змінні та / або змінити їхнє кодування. Подібні перетворення називаються модифікацією даних. В SPSS існує багато можливостей для модифікації даних. До найважливіших з них відносяться: перекодування значень; підрахунок частоти появи певних значень; обчислення нових змінних при виконанні певної умови; обчислення нових змінних шляхом використання різних арифметичних виразів (математичних формул); автоматичне перекодування; агрегування даних тощо.

Найчастіше бувають потрібні такі можливості:

✓ Створення нової змінної на основі однієї змінної – перекодування значень (здійснюється за допомогою команди ***Transform (Перетворити) → Recode (Перекодувати)***);

✓ Створення нової змінної на основі кількох змінних, тобто обчислення нових змінних відповідно до визначених умов, зокрема створення різноманітних індексів (здійснюється за допомогою команди ***Transform (Перетворити) → Compute (Розрахувати)***);

✓ Підрахунок зустрічальності значень у спостереженнях (команда ***Transform (Перетворити) → Count Values within Cases (Підрахувати значення в спостереженнях)***).

4.1. Створення нової змінної на основі однієї ознаки

Перекодування даних використовують в тих випадках, коли потрібно модифікувати шкалу. Наприклад, для подальшого аналізу необхідно об'єднати кілька альтернатив відповіді в одну або інтерпретація результатів буде більш наочною, якщо змінити кодування альтернатив відповіді.

Перекодування даних часто використовують в тих випадках, коли для подальшого аналізу необхідно об'єднати кілька альтернатив відповіді в одну.

Наприклад, в анкету закладено питання в «Якому місті Ви проживаєте?». Соціолог хоче перевірити гіпотезу про відмінності електоральних переваг жителів заходу і сходу України. Для цього можна створити нову змінну, що матиме дві альтернативи:

1. Захід.
2. Схід.

Інший приклад. В анкеті (масив st09.sav) ознаки 134-145 змінюються від 7 до 1, а з змістовної точки зору їх краще кодувати числами від - 1 до 1.

Щоб здійснити перекодування в SPSS необхідно скористатися командою ***Transform (Перетворити) → Recode (Перекодувати)***, як показано на рис. 4.1. При цьому треба мати на увазі, що зберігати перекодовані значення можна у тій самій змінній або перенести їх в іншу змінну. Якщо ми проведемо перекодування в існуючу змінну, всі її старі

значення будуть стерті. Тому рекомендується завжди перекодовувати в нові змінні, щоб не втратити наявну інформацію.

Для цього виберіть в підміню пункт *Into Different Variables ... (У інші змінні)*. Відкриється діалогове вікно *Recode into Different Variables (Перекодувати в інші змінні)*.

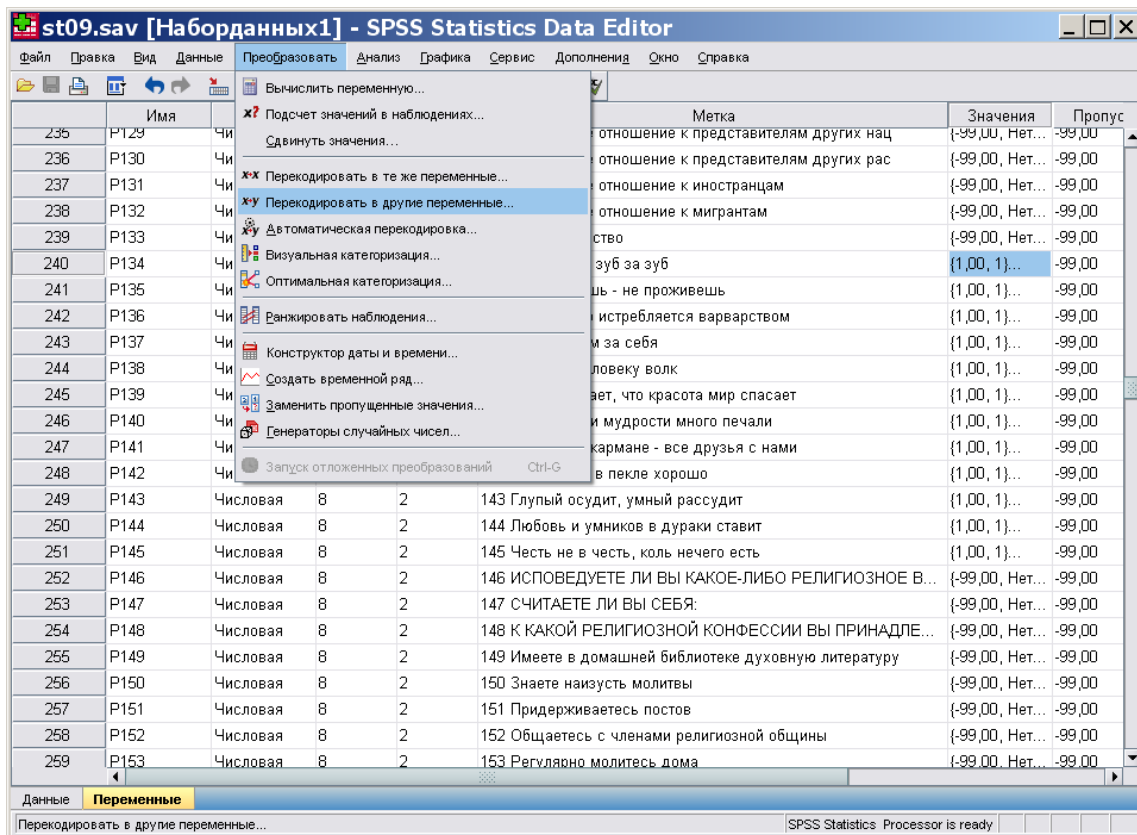


Рис. 4.1. Результаты виконання команди *Transform (Перетворити)* → *Recode (Перекодувати)*

У діалоговому вікні *Recode into Different Variables (Перекодувати в інші змінні)* задайте змінну, яка буде перекодована, ім'я та мітку нової змінної (рис. 4.2), а потім натисніть клавішу «*Старі та нові змінні*», щоб перейти до завдання відповідності старих і нових значень змінної (рис.4.3).

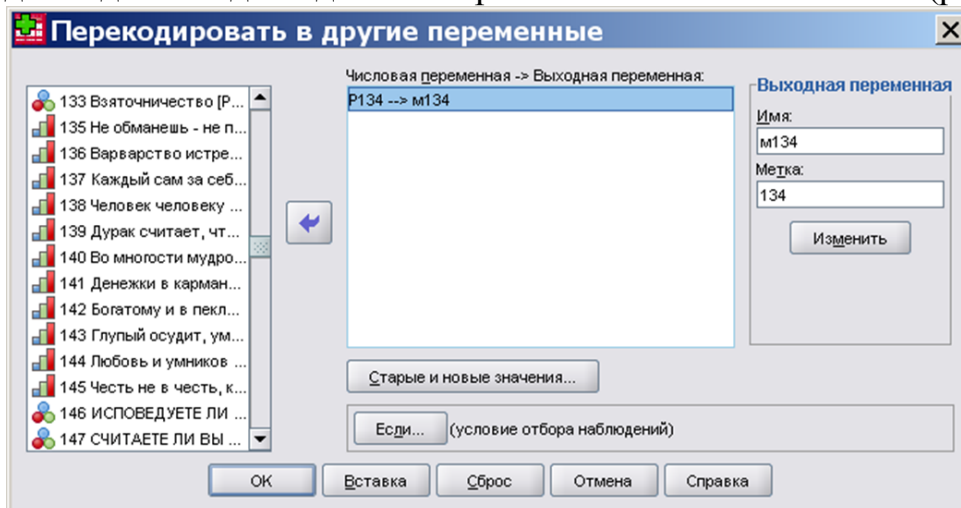


Рис. 4.2. Діалогове вікно *Recode into Different Variables* (Перекодувати в інші змінні)

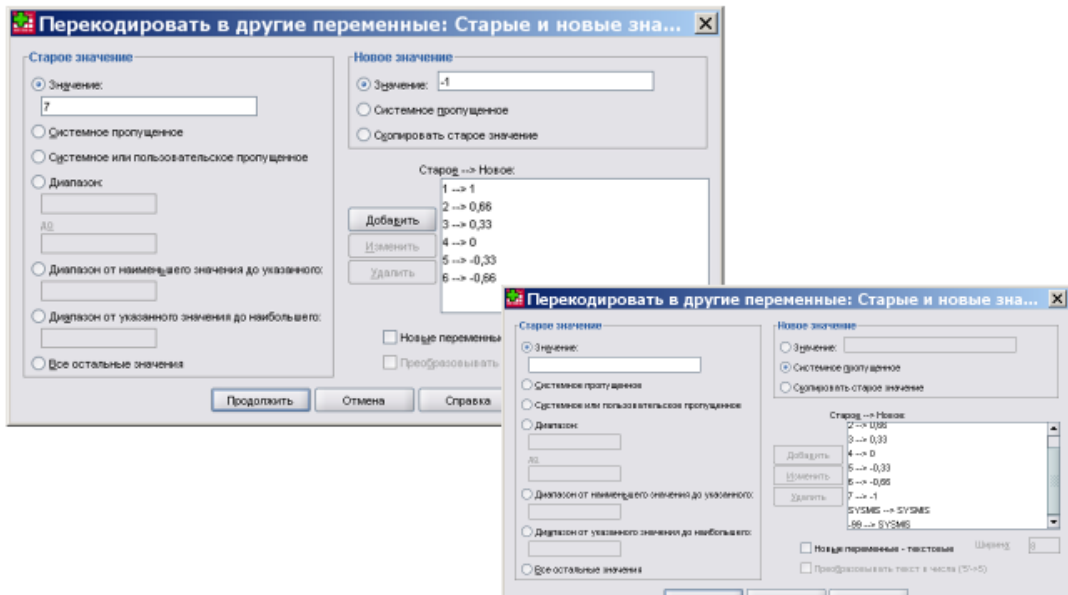


Рис. 4.3. Приклад встановлення відповідності старих та нових значень

Як переконатися у правильності перекодування? Розрахуйте одновимірні розподіли за старою та новою змінними, порівняйте їх та переконайтеся, що вони відрізняються лише значеннями змінної (див. рис. 4.4).

134

		Частота	Процент	Валидний процент	Кумп'ятивний и процент
Валидные	-1,00	499	16,3	17,4	17,4
	,66	639	20,9	22,3	39,7
	,33	515	16,8	18,0	57,7
	,00	664	21,7	23,2	80,9
	,33	216	7,1	7,5	88,4
	,66	174	5,7	6,1	94,5
	1,00	157	5,1	5,5	100,0
	Итого	2864	93,7	100,0	
Пропущенные	Системные пропущенные	194	6,3		
Итого		3058	100,0		

134 Око за око, зуб за зуб

		Частота	Процент	Валидний процент	Кумп'ятивний и процент
Валидные	1,00	157	5,1	5,5	5,5
	2,00	174	5,7	6,1	11,6
	3,00	216	7,1	7,5	19,1
	4,00	664	21,7	23,2	42,3
	5,00	515	16,8	18,0	60,3
	6,00	639	20,9	22,3	82,6
	7,00	499	16,3	17,4	100,0
	Итого	2864	93,7	100,0	
Пропущенные	-99,00	194	6,3		
Итого		3058	100,0		

Рис. 4.4. Результат перекодування змінної «ступінь згоди з твердженням, що треба слідувати принципу око за око, зуб за зуб» (Стара змінна вимірювалася значеннями від 1 до 7, нова змінна – від -1 до +1)

4.2. Створення нової змінної на основі кількох ознак

Наприклад, потрібно визначити, чим відрізняються студенти, які отримують технічну або економічну освіту від гуманітаріїв за своїми ціннісними орієнтаціями (. В цьому випадку можна досліджувати двовимірні розподіли змінної p_{253} (профіль навчання) і змінних p_{53-p70} , що характеризують ступінь важливості для респондентів певних термінальних цінностей. Крім того, для виявлення відмінностей ціннісних орієнтацій, наприклад, прагнення до самореалізації можна використовувати змінну p_{253} як групуючу і порівняти результати U-тесту по Манну і Уїтні для змінної P_{68} (цінність розвитку, реалізації своїх здібностей) при значеннях $p_{253} = 3$ (технічний профіль навчання), $p_{253} = 4$ (економічний профіль навчання) і $p_{253} = 1$ (гуманітарний профіль навчання). Якщо ж потрібно порівняти студентів-технарів і студентів-економістів зі студентками-гуманітаріями, виникає проблема: в цьому випадку з'являються дві групуючі змінні. У подібних ситуаціях допомагає створення допоміжної змінної. Цій змінній можна привласнити значення 1, коли спостереження відповідає студенту технічного профілю навчання, 2 - студенту-економісту, 3 - студентці гуманітарної спеціальності. Потім допоміжна змінна може використовуватись як групуюча при подальшому аналізі.

Щоб побудувати таку змінну, виберіть у меню команду **Transform (Перетворити) → Compute ... (Обчислити)** для відкриття відповідного діалогового вікна (рис. 4.5).

Потім задайте вихідну змінну, наприклад, **gruppe**, а в поле чисельного вираження введіть значення **1** (рис. 4.5). Натисніть кнопку **If ...**, щоб задати логічну умову.

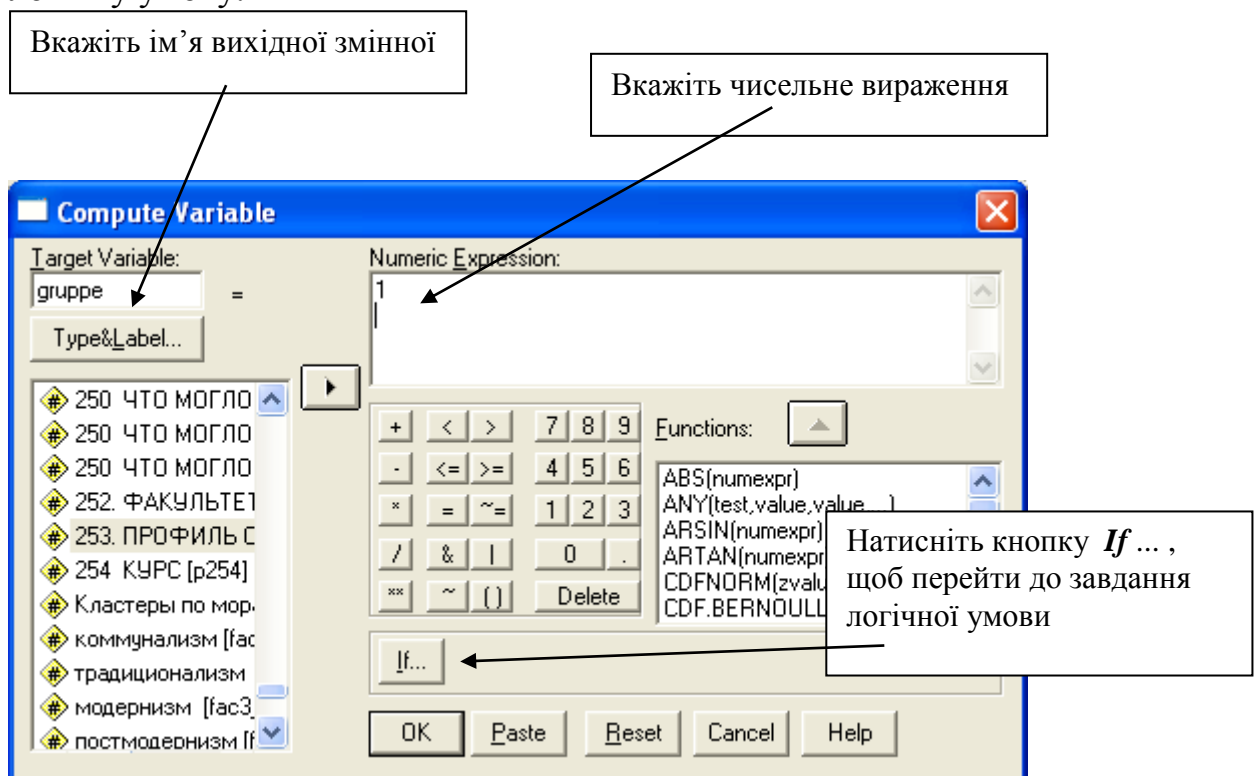


Рис. 4.5. Діалогове вікно **Compute ... (Обчислити)**.

У діалоговому вікні *If ...* вкажіть умову $p253 = 3 \text{ and } p204 = 1$ (див рис. 4.6).

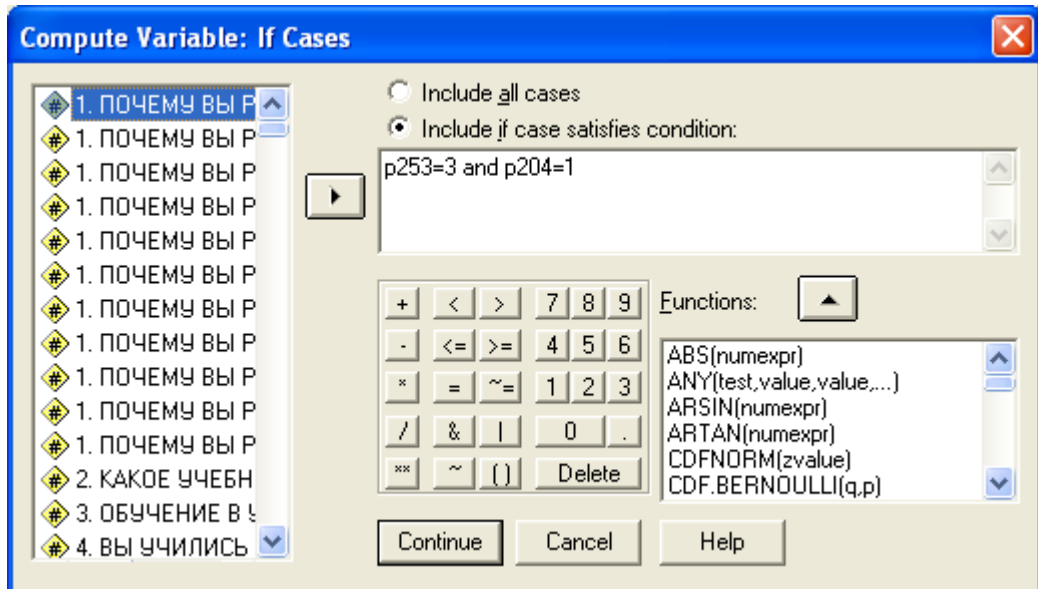


Рис. 4.6. Завдання логічної умови відбору юнаків технічного профілю навчання ($p253 = 3 \text{ and } p204 = 1$)

Закрийте діалоги кнопками *Continue* і *OK*. В результаті буде створена нова змінна, яка поки ще має лише одне значення – 1.

Повторіть процес для створення другого значення. Знову задайте вихідну змінну *gruppe*, але чисельне вираження 2. У діалозі *If ...* сформулюйте умову $p253 = 4 \text{ and } p204 = 1$ (див рис. 4.7).

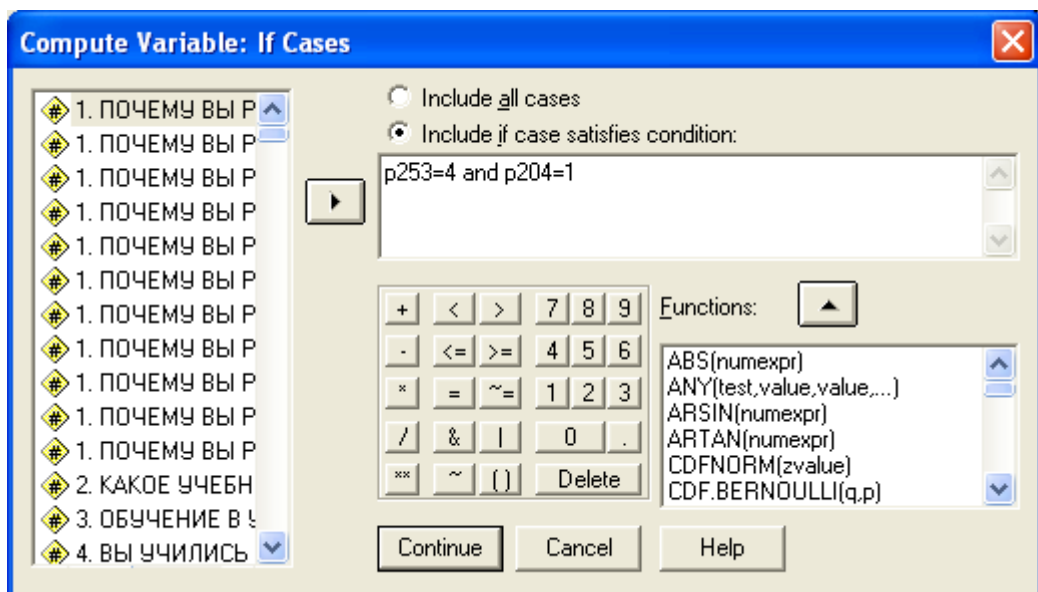


Рис. 4.7. Завдання логічної умови відбору юнаків економічного профілю навчання ($p253 = 4 \text{ and } p204 = 1$)

На запитання пакету *Change existing variables?* (рис. 4.8), яке з'являється після закриття діалогів, відповідайте ствердно (**OK**).

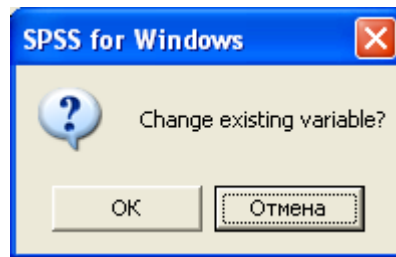


Рис. 4.8. Підтвердження внесених змін у створювану перемінну

Закрийте діалоги кнопками *Continue* і *OK*.

Повторіть процес; знову задайте вихідну змінну *gruppe*, але чисельне вираження 3. У діалозі *If...* сформулюйте умову $p253 = 1$ and $p204 = 2$.

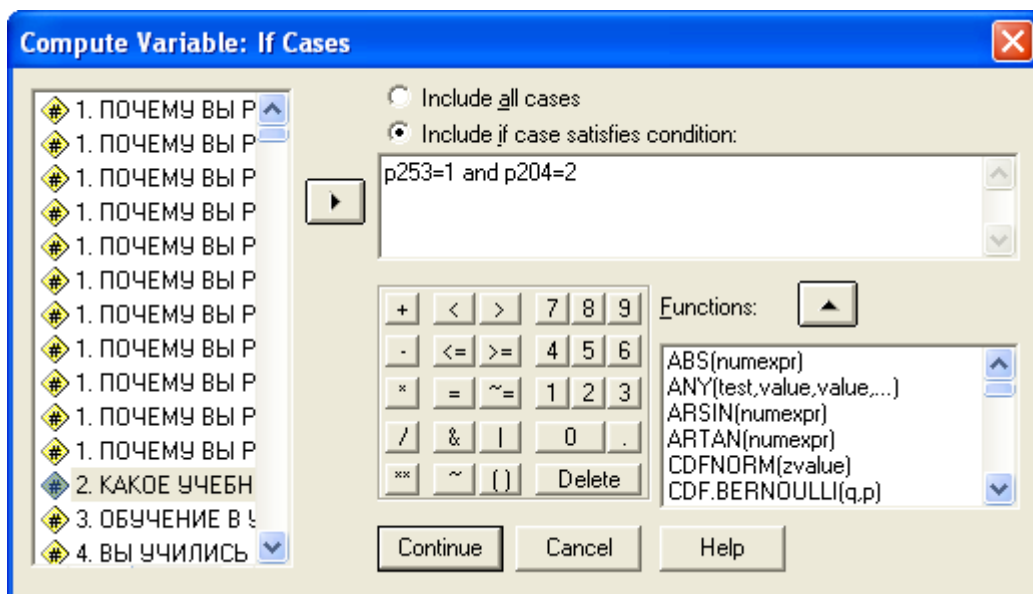


Рис. 4.9. Завдання логічної умови відбору дівчат гуманітарного профілю навчання ($p253 = 1$ and $p204 = 2$)

На запитання *Change existing variables ?*, яке з'являється після закриття діалогів, знову відповідайте ствердно (**OK**).

Після цього у редакторі даних з'явиться нова змінна *gruppe*, яка в спостереженнях, відповідних сформульованим умовам, має значення 1, 2 або 3.

Щоб з цією змінною було зручно працювати при подальшому аналізі даних, рекомендується відразу словесно описати числові значення змінної, застосовуючи правила роботи з *Редактором Даних* (див. рис. 4.10).

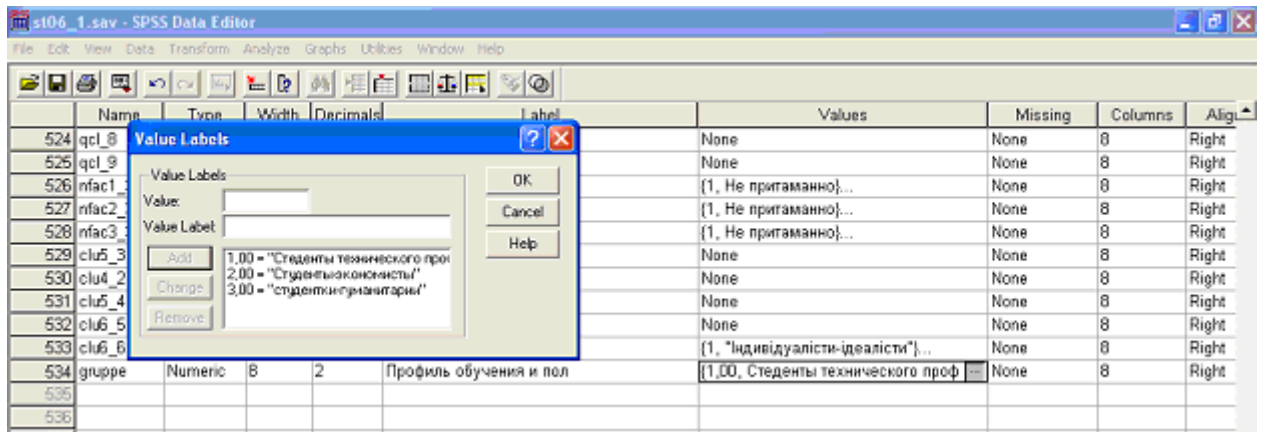


Рис. 4.10. Введення словесного опису кількісних значень нової змінної

4.4. Підрахунок зустрічальності значень у спостереженнях в SPSS

Діалогове вікно *Count Values within Cases...* (*Підрахувати зустрічальність значень у спостереженнях*) дає можливість створити змінну, в якій міститься результат підрахунку того, скільки разів певне значення зустрічається в наборі змінних. Наприклад, в анкеті може міститися список журналів з прапорцями так / ні, щоб відзначити читають респонденти цей журнал чи ні. Можна створити нову змінну, в якій буде підраховано кількість відповідей так. Така змінна буде показувати загальне число журналів, які він читає кожний респондент.

Можна підрахувати кількість телепередач, яким опитувані надають перевагу. Наприклад, зробимо це на основі масиву st09.sav. В анкеті є відповідне питання:

Яким телепередачам Ви надаєте перевагу?

1. Теленовинам
2. Публіцистичним програмам, політичним ток-шоу
3. Художнім фільмам
4. Телесеріалам
5. Розважальним програмам, ігровим шоу, вікторинам тощо
6. Музичним передачам, концертам
7. Спортивним програмам
8. Науково-популярним, освітнім програмам
9. Передачам про здоров'я, кулінарію тощо
10. Бізнес-інформації, рекламі
11. Телевізор не дивлюсь

Можна побачити, що шкала вимірювання ознаки – номінальна із сумісними альтернативами. Якщо респонденти обрали певну альтернативу відповіді (тобто певний вид телепередач) – це буде закодовано 1, якщо ні – 0. Для створення перемінної, що містить кількість телепередач, які дивляться опитані, треба виконати команду *Transform (Перетворити) Count... (Підрахувати)*. У результаті відкриється діалогове вікно *Count Occurrences of Values within Cases (Підрахувати кількість значень у спостереженнях)*, в

якому необхідно задати ім'я нової перемінної, мітку та значення, що будуть підраховуватись (див. рис. 4.11).

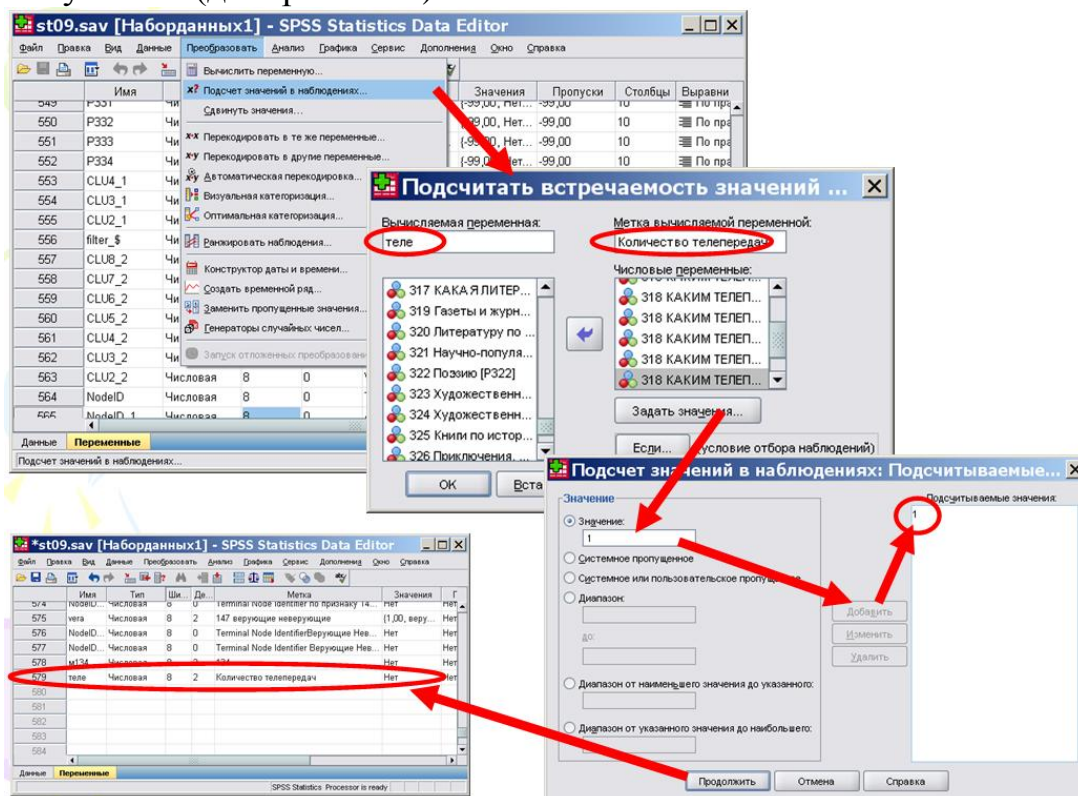


Рис. 4.11. Порядок дій при створенні змінної, що містить кількість телепередач, яким надають перевагу опитані

Після виконання цих дій у масив даних буде додано нову змінну, яку можна використовувати для виконання будь-яких розрахунків, наприклад, побудувати одновимірний розподіл та подивитися скільки телепередач в середньому дивляться українські студенти сьогодні (див. рис. 4.12 та рис. 4.13). При виконанні розрахунків та інтерпретації результатів варто мати на увазі, що створена перемінна є метричною.

Кількість телепередач

	Частота	Процент	Валидный процент	Кумулятивны и процент
Валидные	,00	76	2,5	2,5
1,00	911	29,8	29,8	32,3
2,00	516	16,9	16,9	49,1
3,00	666	21,8	21,8	70,9
4,00	454	14,8	14,8	85,8
5,00	237	7,8	7,8	93,5
6,00	116	3,8	3,8	97,3
7,00	33	1,1	1,1	98,4
8,00	22	,7	,7	99,1
9,00	15	,5	,5	99,6
10,00	11	,4	,4	100,0
11,00	1	,0	,0	100,0
Итого	3058	100,0	100,0	

Рис. 4.12. Одновимірний розподіл за створеною ознакою

Описательные			Статистика	Стд. ошибка
Количество телепередач	Среднее		2,7145	,03148
	95% доверительный интервал для среднего	Нижняя граница	2,6528	
		Верхняя граница	2,7762	
	5% усеченное среднее		2,5927	
	Медиана		3,0000	
	Дисперсия		3,030	
	Стд. отклонение		1,74060	
	Минимум		,00	
	Максимум		11,00	
	Размах		11,00	
	Межквартильный размах		3,00	
	Асимметрия		,994	,044
	Экссесс		1,310	,089

Рис. 4.13. Описові статистики за створеною ознакою

Література до теми

Бююль А., Цёфель П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005. С. 122-142.

Горбачик А.П., Сальникова С.А. *Аналіз даних соціологічних досліджень засобами SPSS*. Луцьк: «Вежа», 2008. С. 105-122.

Наследов А.Д. *SPSS: Компьютерный анализ данных в психологии и социальных науках*. СПб.: Питер, 2005. С. 63-73.

Питання для самоконтролю

1. В яких випадках виникає потреба у створенні нової змінної?
2. Чим відрізняється перекодування значень змінної від обчислення значень?
3. Коли виникає потреба у перекодуванні значень змінної?
4. Які аналітичні завдання вирішують завдяки створенню нових змінних?
5. Коли виникає потреба створювати нові змінні на основі кількох змінних?
6. В яких випадках корисно застосовувати підрахунок зустрічальності значень?

Тема 5. Кореляційний аналіз та двовимірні розподіли

5.1. Кореляційний аналіз та кореляційна залежність

Кореляційний аналіз – сукупність методів виявлення кореляційної залежності між кількома ознаками. Мета кореляційного аналізу – виявлення зв'язку між випадковими величинами, що дозволяє забезпечити отримання інформації про одну змінну за допомогою іншої змінної (або кількох інших змінних).

Кореляційна залежність - це залежність, де вплив окремих змінних виявляється тільки як ймовірність появи різних значень іншої змінної. Особливістю кореляційної залежності, на відміну від функціональної, є те, що одному значенню X відповідають кілька значень Y . Нагадаємо, що якщо значенню однієї величини (X) відповідає певне значення іншої (Y), то вважають, що між цими величинами має місце функціональна залежність (рис. 5а).

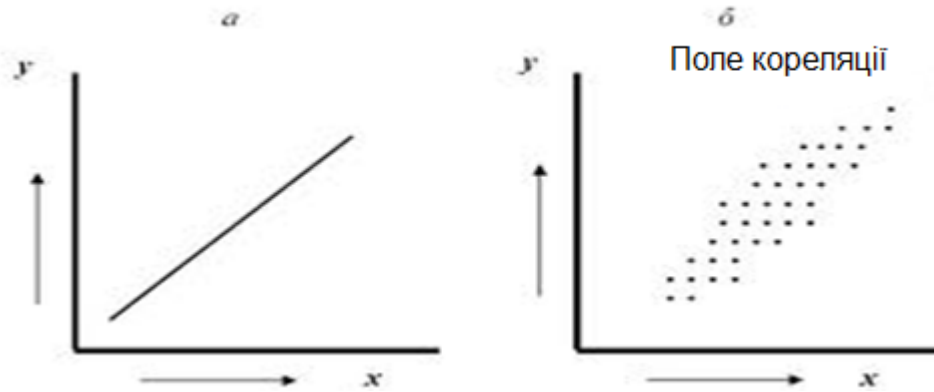


Рис 5.1а. Функціональна залежність

Рис 5.1б. Кореляційна залежність

Рис. 5.1. Функціональна та кореляційна залежності

Кореляційні залежності проявляються у найрізноманітніших формах, які класифікують за багатьма критеріями:

- Лінійна / нелінійна. При цьому лінійна залежність розуміється, як зв'язок між досліджуваними величинами, при якому рівномірним змінам однієї величини відповідають рівномірні зміни іншої; нелінійна – зв'язок між величинами, при якому рівномірним змінам однієї величини відповідають нерівномірні зміни іншої, причому ця нерівномірність має певний закономірний характер;

- Пряма / зворотна. Пряма залежність – зі збільшенням значення незалежної змінної (X), збільшуються значення залежної змінної (Y), зворотна залежність – зі збільшенням значення незалежної змінної (X), зменшуються значення залежної (Y);

- Сильна / слабка (встановлюється за абсолютним значенням коефіцієнту кореляції);

- Статистично значуща (з певною долею вірогідності можна розповсюджувати результати на генеральну сукупність) / статистично незначуща (результати не можна розповсюджувати на генеральну сукупність). *Примітка:* значущість залежить від обсягу вибірки, і на вибірках великого обсягу (в декілька тисяч об'єктів) навіть порівняно невеликі коефіцієнти кореляції (менш ніж 0,1) будуть статистично значущими.

В залежності від кількості досліджуваних ознак аналіз кореляційних залежностей поділяють на:

- ✓ Кореляційний аналіз двох ознак (крос табуляція, аналіз взаємозв'язку двох ознак, аналіз двовимірних розподілів).
- ✓ Кореляційний аналіз трьох ознак.
- ✓ Кореляційний аналіз чотирьох та більшої кількості ознак - багатовимірний кореляційний аналіз.

Кореляційний аналіз трьох ознак насамперед застосовують з метою усунення можливих помилок, пов'язаних з виявленням хибних кореляцій, широко відомим прикладом яких є кореляція сорту губної помади з політичними переконаннями жінки. Такі помилки зазвичай виникають у разі наявності якогось неврахованого в аналізі фактору (чи факторів), що впливає на кожну з досліджуваних змінних і, тим самим, породжує «кореляцію» між ними. У наведеному прикладі такими факторами є суспільне становище і рівень добробуту жінки. Кореляційний аналіз трьох ознак проводиться шляхом введення контрольної змінної, що дозволяє перевірити чи дійсно існує кореляційна залежність між аналізованими змінними, чи не є виявлена кореляція хибною або опосередкованою.



Багатовимірний кореляційний аналіз передбачає аналіз кореляцій чотирьох і більше змінних. Проте останнім часом він вкрай рідко застосовується соціологами, оскільки з розвитком доступного програмного забезпечення з'явилася можливість активно використовувати факторний аналіз, який вирішує більш широке коло завдань, ніж кореляційний аналіз багатьох змінних.

5.2. Аналіз двовимірних розподілів

У соціологічних дослідженнях найчастіше виникає потреба аналізувати зв'язки між двома ознаками, тобто проводити двовимірний кореляційний аналіз.

Найпоширенішими практичними прийомами кореляційного аналізу двох ознак є такі: 1) побудова кореляційної таблиці та її змістовна інтерпретація; 2) обчислення коефіцієнтів кореляції та інтерпретація їхніх значень.

Двовимірна кореляційна таблиця (кростабуляція, таблиця зв'язаності) – це матриця, у якій на перетині i -го рядка й j -го стовпця знаходиться n_{ij} – частота сумісної появи відповідних значень двох ознак x_i та y_j (див. табл. 5.1 та табл. 5.2). Крім того в такій таблиці поряд з частотами наводять відсотки по рядку та/або стовпцю.

Таблиця 5.1

Загальний вигляд кореляційної таблиці

X	Y				Результат за рядками N(x)
	y_1	y_2	...	y_j	
x_1	n_{11}	n_{12}	...	n_{1j}	$N(x_1)$
x_2	n_{21}	n_{22}	...	n_{2j}	$N(x_2)$
x_i	n_{i1}	n_{i2}	...	n_{ij}	$N(x_i)$
Результат за стовпцями N(y)	$N(y_1)$	$N(y_2)$...	$N(y_i)$	N

Побудова кореляційної таблиці (тобто двовимірного розподілу або таблиці крос-табуляції) в SPSS здійснюється за допомогою команди *Descriptive statistics (Описові статистики)* → *Crosstabs (Таблиці спряженості)*, у результаті виконання якої виводиться діалогове вікно *Crosstabs* (див. рис. 5.2), у якому необхідно вказати досліджувані ознаки у полях *Row (Строка)* та *Column (Стовпець)*.

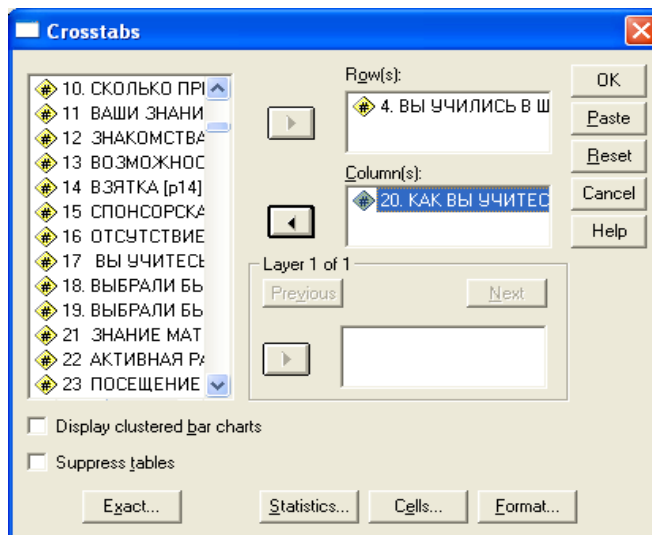


Рис. 5.2. Побудова двовимірного розподілу в пакеті SPSS: Діалогове вікно *Crosstabs*

У результаті отримаємо двовимірний розподіл (двовимірну кореляційну таблицю):

		20. КАК ВЫ УЧИТЕСЬ					
		Только на 5	На 4 и 5	На 4	На 4 и 3	На 3	Total
4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	Только на 'отлично'	Count 132	222	53	25	7	439
	% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	30,1%	50,6%	12,1%	5,7%	1,6%	100,0%
	% within 20. КАК ВЫ УЧИТЕСЬ	40,9%	18,6%	6,9%	4,2%	4,7%	14,5%
На 'хорошо' и 'отлично'	Count	167	816	425	197	43	1648
	% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	10,1%	49,5%	25,8%	12,0%	2,6%	100,0%
	% within 20. КАК ВЫ УЧИТЕСЬ	51,7%	68,4%	55,3%	33,3%	28,9%	54,5%
В основном на 'хорошо'	Count	17	117	235	222	53	644
	% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	2,6%	18,2%	36,5%	34,5%	8,2%	100,0%
	% within 20. КАК ВЫ УЧИТЕСЬ	5,3%	9,8%	30,6%	37,5%	35,6%	21,3%
На 'хорошо' и 'удовлетворительно'	Count	5	35	52	138	36	266
	% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	1,9%	13,2%	19,5%	51,9%	13,5%	100,0%
	% within 20. КАК ВЫ УЧИТЕСЬ	1,5%	2,9%	6,8%	23,3%	24,2%	8,8%
В основном на 'удовлетворительно'	Count	2	3	3	10	10	28
	% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?	7,1%	10,7%	10,7%	35,7%	35,7%	100,0%
	% within 20. КАК ВЫ УЧИТЕСЬ	0,6%	0,3%	0,4%	1,7%	6,7%	0,9%
Total		Count 323	1193	768	592	149	3025
% within 4. ВЫ УЧИЛИСЬ В ШКОЛЕ...?		10,7%	39,4%	25,4%	19,6%	4,9%	100,0%
% within 20. КАК ВЫ УЧИТЕСЬ		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Рис. 5.3. Вигляд двовимірного розподілу у пакеті SPSS

Для розрахунку коефіцієнтів кореляції у діалоговому вікні *Crosstabs* треба натиснути кнопку *Statistics*, щоб перейти до відповідного діалогового вікна (див. рис. 5.3).

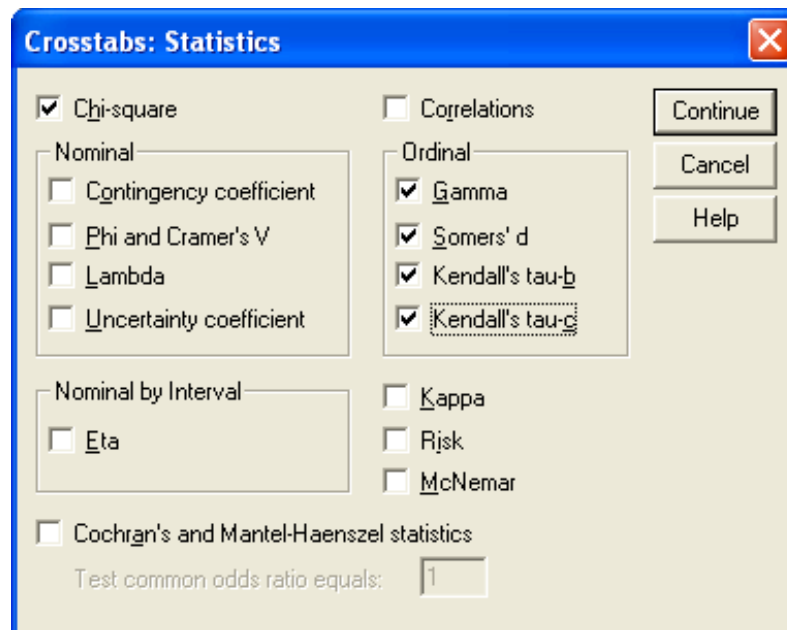


Рис. 5.4. Розрахунок коефіцієнтів кореляції двовимірного розподілу: Діалогове вікно *Statistics*

У результаті SPSS розрахує всі вказані коефіцієнти кореляції (див. рис. 5.5).

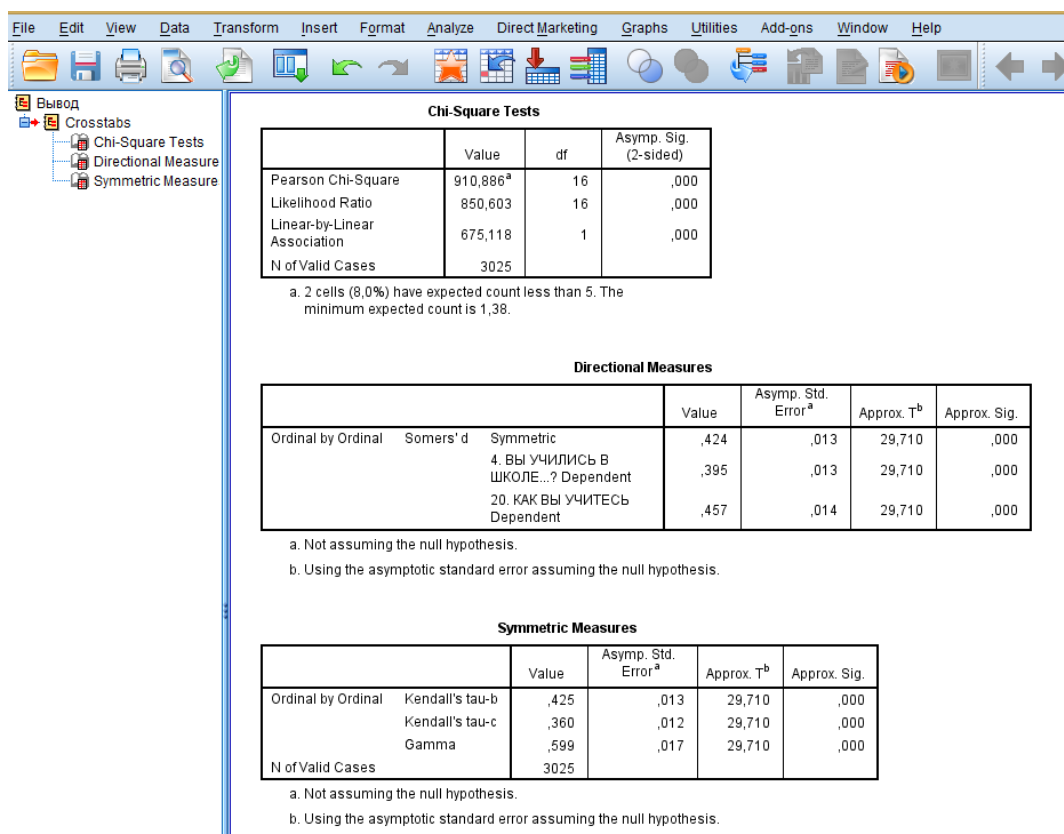


Рис. 5.5. Вигляд розрахованих коефіцієнтів кореляції

Існує велика кількість різноманітних коефіцієнтів кореляції, варіативність яких зумовлена застосуванням різних принципів (інакше кажучи різних моделей поняття «зв'язок»).

Найвідоміші принципи побудови коефіцієнтів кореляції:

Принцип спільної появи подій (коефіцієнти Чупрова, Крамера та ін.);
Принцип ПЗП - пропорційного зменшення помилки (коефіцієнти лямбда або тау Гудмана-Крускала);

Принцип ковариації (коефіцієнт парної кореляції Пірсона, коефіцієнти рангової кореляції Спірмена і Кенделла).

Обмеження у застосуванні кожного коефіцієнту кореляції головним чином пов'язані з рівнем вимірювання досліджуваних ознак.

Розглянемо статистики, які розраховує SPSS, та обговоримо в яких випадках їх можна застосовувати.

Статистика хі-квадрат (Chi-square) - критерій, який використовується для перевірки статистичної значущості спостережуваних зв'язків у таблицях двовимірних розподілів. Він *зазначає наявність чи відсутність зв'язку між двома ознаками*; розраховується на основі зіставлення теоретичного розподілу (коли зв'язок відсутній) з емпіричним, представленим у таблиці, що аналізується.

Хі-квадрат приймає значення від 0 до $+\infty$, проте аналітика цікавить лише його статистична значущість, яка інтерпретується як ймовірність наявності зв'язку між досліджуваними ознаками.

Найбільш відомим та поширеним є хі-квадрат Пірсона. Проте можливості його застосування обмежуються певними умовами.

Умови та обмеження застосування критерію хі-квадрат Пірсона:

- Показники, повинні бути виміряні за номінальною або за порядковою шкалою.
- Групи, що зіставляються, повинні бути незалежними, тобто критерій хі-квадрат не повинен застосовуватися з метою порівняння спостережень "до-після". У цих випадках проводиться **тест Мак-Немара** (для порівняння двох пов'язаних сукупностей) або розраховується **Q-критерій Кохрена** (в разі порівняння трьох і більше груп).
- При аналізі чотирьохпільних таблиць **очікувані значення в кожній з клітинок** повинні бути не менше 10. У тому випадку, якщо хоча б в одній клітинці очікуване явище приймає значення від 5 до 9, критерій хі-квадрат повинен розраховуватися з **поправкою Йейтса**. Якщо хоча б в одній клітинці очікуване явище менше 5, то для аналізу повинен використовуватися **точний критерій Фішера**.
- У разі аналізу багатопільних таблиць очікуване число спостережень не повинно приймати значення менше 5 більше, ніж в 20% клітинок.

SPSS розраховує декілька варіантів критерія Хі-квадрат:

- Pearson Chi-Square (Хі-квадрат за Пірсоном).
- Likelihood Ratio (Відношення правдоподібності) - При великих вибірках значення близькі до *хі-квадрат Пірсона*. При малих вибірках значення зазвичай трохи менше, а тому вважається деякими авторами за кращий.

- **Linear-by-Linear Association (Залежність лінійно-лінійна)** - Якщо обидві змінні в таблиці є кількісними, то при позначці елемента χ^2 -квадрат розраховується критерій лінійно-лінійного зв'язку.

Коефіцієнти кореляції свідчать про силу (щільність) зв'язку між ознаками. Вибір коефіцієнта залежить від рівня виміру. Рівень вимірювання для двовимірної таблиці визначається мінімальним рівнем вимірювання однієї зі змінних. Наприклад, якщо двовимірний розподіл побудований з номінальної та порядкової змінних, можна використовувати тільки номінальні коефіцієнти кореляції.

Correlations. Коли ми задаємо обчислення *Correlations*, SPSS розраховує коефіцієнт кореляції Пірсона або Спірмена.

Лінійний коефіцієнт кореляції Пірсона r (коефіцієнт парної кореляції) використовується для метричних шкал розподілених нормально. Він характеризує силу лінійного зв'язку між змінними. Може приймати значення в інтервалі від -1 до +1. Якщо r набув близького до «0» значення, це підстава стверджувати про відсутність лінійного зв'язку між досліджуваними змінними. Однак у цьому випадку можлива нелінійна взаємозалежність, що потребує додаткової перевірки та застосування інших коефіцієнтів. Коефіцієнт кореляції Пірсона не описує криву залежності і не підходить для опису складних, нелінійних залежностей.

Коефіцієнт рангової кореляції Спірмена ρ можна застосовувати не тільки до метричних, а й до порядкових змінних. Для розрахунку ρ Спірмена не потрібно ніяких припущень про нормальність розподілу ознак у генеральній сукупності. Може приймати значення в інтервалі від -1 до +1. Інтерпретація аналогічна інтерпретації коефіцієнта кореляції Пірсона r .

Коефіцієнти для номінальних даних (Nominal)

Номінальні змінні тільки розбивають досліджувану сукупність на класи, які неможливо впорядкувати (наприклад, чоловіки і жінки). Для них можна розрахувати наступні коефіцієнти кореляції.

Коефіцієнт контингенції (Contingency coefficient), або коефіцієнт спряженості Пірсона. Міра зв'язку, заснована на χ^2 -квадрат. Приймає значення в інтервалі $[0;1)$, причому 0 означає відсутність зв'язку між змінними рядка і стовпця, а значення, близьке до 1, - високий ступінь зв'язку між цими змінними. Чим більше значення, тим сильніше зв'язок. Максимально можливе значення залежить від числа рядків і стовпців в таблиці. Але є недолік: максимальне значення 1 не досягає (повний зв'язок).

Φ і V Крамера (Phi and Cramer's V). Коефіцієнт ϕ (ϕ) – це міра зв'язку, що обчислюється поділом статистики χ^2 -квадрат на обсяг вибірки і взяттям кореня квадратного з результату. Його можна використовувати тільки для таблиць 2×2 , в інших випадках він може перевищити значення 1. V Крамера - це міра зв'язку, також заснована на статистиці χ^2 -квадрат. Приймає значення в інтервалі $[0;1]$, але дорівнює 1 тільки у випадку квадратної таблиці (кількість строк дорівнює кількості стовпців).

Таким чином, на критерії Хі-квадрат засновані наступні коефіцієнти кореляції: коефіцієнт спряженості, Φ і V Крамера. Якщо Хі-квадрат не надійний, їх застосовувати не можна.

Лямбда (Lambda). Розраховуються симетричні і асиметричні коефіцієнти лямбда та асиметричні коефіцієнти тау Гудмена-Краскала.

Лямбда (λ) та тау (τ) Гудмена-Краскала - міри зв'язку, які відображають відносно зниження помилки, коли значення незалежної змінної використовуються для передбачення значень залежної змінної (такі коефіцієнти також називаються коефіцієнтами пропорційної редукції помилки). Значення 1 означає, що незалежна змінна точно прогнозує значення залежної. Значення 0 означає, що незалежна змінна абсолютно марна для передбачення залежної. Приймають значення в діапазоні $[0;1]$, де «0» означає відсутність зв'язку, чим ближче до 1, тим сильніший зв'язок. Лямбда має недолік: якщо всі модальні частоти незалежної змінної знаходяться в одному стовпці або рядку таблиці, то $\lambda = 0$.

Коефіцієнт невизначеності (Uncertainty coefficient). оцінює пропорційну редукцію невизначеності, що вимірюється за допомогою ентропії. Чим ближче значення до 1, тим більшою мірою значення незалежної змінної X усуває невизначеність того, яке значення прийме Y . Принцип дії коефіцієнту схожий на принцип дії лямбда або тау Гудмана-Краскала, з тією лише різницею, що замість помилок прогнозу тут використовується поняття невизначеності, мірою якої є ентропія. Обчислюються як симетрична, так і несиметрична версії коефіцієнта невизначеності.

Коефіцієнти для порядкових даних (Ordinal)

Порядкові змінні – це змінні, значення яких упорядковані певним чином. Для них можливо розрахувати наступні коефіцієнти.

Коефіцієнт гамма Гудмана-Краскала (Gamma). Симетрична міра зв'язку між двома порядковими змінними, значення якої змінюються між -1 і 1. Значення, близькі по абсолютній величині до 1, вказують на сильний зв'язок змінних; значення вище 0 говорять про прямий зв'язок, менше 0 – про зворотній. Значення, близькі до 0, говорять про слабкий зв'язок або його відсутність. Для таблиць спряженості двох змінних обчислюється гамма нульового порядку. Якщо ж таблиця спряженості включає більше двох змінних, для кожної підтаблиці обчислюється умовна гамма.

Коефіцієнт кореляції d Сомерса (Somers' d). Міра зв'язку між двома порядковими змінними, змінюється між - і 1. Значення, близькі по абсолютній величині до 1, вказують на сильний зв'язок між двома змінними, а значення, близькі до 0, - на слабкий зв'язок або його відсутність. Це асиметричне розширення міри гамма, що відрізняється тільки включенням числа пар, які не мають збігів (зв'язків) з незалежною змінною. Обчислюється також симетрична версія цієї статистики. *d Сомерса є асиметричним розширенням гамма Гудмена і Краскала* (є спрямованою мірою). Кожен із

чинників може при цьому по черзі розглядатися в якості залежного, а інший - незалежного.

Коефіцієнт кореляції тау-бі Кендала (Kendall's tau-b). Непараметрична міра кореляції для порядкових або рангових змінних, яка враховує можливі збіги значень (зв'язку). Знак коефіцієнта вказує напрямок зв'язку, а його модуль - силу зв'язку, причому, чим він більший, тим зв'язок сильніше. Значення змінюються в діапазоні між -1 і +1, однак значення -1 і +1 можна отримати тільки для квадратних таблиць. Цей коефіцієнт краще використовувати для квадратних таблиць, тому що тільки в квадратних таблицях його величина може досягати 1 або -1. Для прямокутних таблиць краще використовувати тау-сі.

Коефіцієнт кореляції тау-сі Кендала (Kendall's tau-c). Непараметрична міра зв'язку для порядкових змінних, що ігнорує можливі збіги значень (зв'язку). Знак коефіцієнта вказує напрямок зв'язку, а його модуль - силу зв'язку, причому, чим він більший, тим зв'язок сильніше. Значення змінюються в діапазоні між -1 і +1, однак значення -1 і +1 можна отримати тільки для квадратних таблиць.

Коефіцієнт для таблиці, де залежна змінна номінальна, а незалежна метрична (Nominal by Interval)

У ситуації, коли одна з змінних категоріальна, а інша - кількісна, виберіть статистику Ета. Значення категоріальної змінної повинні бути закодовані числами.

Ета (Eta) - міра зв'язку між змінними рядка і стовпця, значення якої змінюються від 0 (відсутність зв'язку) до 1 (сильний зв'язок). Індикатор Ета підходить для залежної змінної, вимірної за інтервальною шкалою (наприклад, дохід) і незалежної змінної з обмеженим числом категорій (наприклад, групи за віком). Обчислюються два значення Ета: одне розглядає змінну рядку як інтервальну змінну, а інше - змінну стовпця як інтервальну змінну.

Інші коефіцієнти

Каппа (Каппа). Каппа Коена вимірює згоду думок двох експертів, які оцінюють одні і ті ж об'єкти. Значення 1 вказує на повну згоду. Значення 0 вказує на те, що згода - не більше ніж випадковість. Каппа ґрунтується на квадратній таблиці, в якій значення рядків і стовпців виміряні за однією і тією же шкалою. Будь-якій клітинці, яка має спостережені значення для однієї змінної, але не має для іншої, присвоюється частота, рівна 0. Каппа не розраховується, якщо тип зберігання даних (текстовий або числовий) не однаковий для обох змінних. Для текстових змінних, обидві змінні повинні мати однакову задану довжину.

Ризик (Risk). Міра сили зв'язку для таблиць 2 x 2 міра сили зв'язку між присутністю фактору і виникненням події. Якщо довірчий інтервал для цієї статистики включає 1, припущення про те, що фактор пов'язаний з подією,

буде невірним. Якщо наявність фактору зустрічається рідко, то в якості оцінки відносного ризику можна використовувати відношення шансів.

МакНемара (McNemar). Непараметричний критерій для двох пов'язаних дихотомічних змінних. Перевіряє зміни у відгуках за допомогою розподілу хі-квадрат. Корисний для виявлення змін в відгуках, обумовлених експериментальним втручанням в плани до-і-після. Для великих квадратних таблиць видаються результати критерію симетричності Мак-Немара - Боукера.

Статистики Кокрена і Мантеля-Хенцеля (Cochran's and Mantel-Haenszel statistics). Вони можуть використовуватися для перевірки умовної незалежності дихотомічної факторної змінної і дихотомічної змінної відгуку при заданих коваріаційних структурах, що задаються однією або більшою кількістю змінних шару (керуючих змінних). Зауважимо, що в той час як інші статистики обчислюються пошарово, статистики Кокрена і Мантеля-Хенцеля обчислюються відразу для всіх верств.

Як ми бачимо, всі коефіцієнти кореляції, призначені для вимірювання сили зв'язку між ознаками, мають значення, що варіюють у межах від 0 до 1 або від -1 до +1. Отже, їхні абсолютні значення знаходяться у межах від 0 до 1. Інтерпретація цих значень здійснюється наступним чином (див. табл. 5.2).

Таблиця 5.2

Відповідність значення коефіцієнту кореляції та тісноти зв'язку ознак

Абсолютне значення коефіцієнту кореляції – кількісна характеристика	Тіснота зв'язку - якісна характеристика
1,00	Зв'язок функціональний
0,90-0,99	Дуже сильний
0,70-0,89	Сильний
0,50-0,69	Значний
0,30-0,49	Помірний
0,10-0,29	Слабкий
0,00-0,19	Зв'язок відсутній

Нижче наведена зведена таблиця, де коефіцієнти класифікуються в залежності від їхньої придатності для аналізу ознак певних рівнів вимірювання та принципів побудови.

Додатково коефіцієнти кореляції можна поділити на дві групи: спрямовані та симетричні. Спрямовані міри дозволяють аналізувати силу зв'язку з двох боків: 1) вплив X на Y (при цьому "X" розглядається як незалежна змінна, "Y" - залежна змінна; 2) вплив Y на X ("X" - залежна змінна; "Y" - незалежна змінна).

Симетричні міри характеризують силу взаємозв'язку.

До *спрямованих мір (Directional Measures)* відносяться: d Сомерса, λ (лямбда), τ (тау) Гудмана-Краскела, коефіцієнт невизначеності.

До *симетричних мір (Symmetric Measures)* відносяться: r (ер) Пірсона, ρ (ро) Спірмена, γ (гама), τ (тау) Кендалла — τ_b и τ_c , ρ (ро) Спірмена, ϕ (фі), спряженості Пірсона, V Крамера.

Таблиця 5.3

Коефіцієнти кореляції

Принцип Рівень вимірювання	Коваріація	Порівняння порядку рангів	Хі-квадрат	Прогноз або пропорційне зниження помилки	Ентропія
Метричний	r (ер) Пірсона, ρ (ро) Спірмена	γ (гама), τ (тау) Кендалла — τ_b и τ_c , d Сомерса, ρ (ро) Спірмена	ϕ (фі), спряженості Пірсона, V Крамера	λ (лямбда), τ (тау) Гудмана- Краскела	Коефіцієнт невизначе- ності
Порядковий		γ (гама), τ (тау) Кендалла — τ_b и τ_c , d Сомерса, ρ (ро) Спірмена	ϕ (фі), спряженості Пірсона, V Крамера	λ (лямбда), τ (тау) Гудмана- Краскела	Коефіцієнт невизначе- ності
Номінальний			ϕ (фі), спряженості Пірсона, V Крамера	λ (лямбда), τ (тау) Гудмана- Краскела	Коефіцієнт невизначе- ності

5.3. Візуалізація двовимірних розподілів

Вибір методу візуалізації двовимірного розподілу зумовлюється шкалами вимірювання аналізованих ознак.

Найпоширеніші способи візуалізації двовимірних розподілів:

- ✓ Для метричних шкал – діаграми розсіювання;
- ✓ Для порядкових шкал – діаграми розсіювання, інтервальні середні значення, стовпчасті діаграми;
- ✓ Для номінальних шкал – стовпчасті діаграми.

Діаграма розсіювання (точкова діаграма, англ. Scatter plot) – це двовимірний графік для зображення спільного розподілу двох кількісних змінних. Такі діаграми є найвідомішим засобом візуального представлення

кореляційного поля. Такі діаграми є найкращим методом візуалізації кореляції ознак, виміряних метричними шкалами.

Кожен об'єкт з вибірки представляється у вигляді точки, координатами якої служать відповідні йому значення двох змінних. По вигляду діаграми розсіювання можна судити про напрямок (прямий чи зворотній) та форму зв'язку між двома представленими на діаграмі змінними. По осі X відкладається одне значення, по осі Y – інше. Таким чином в евклідовому просторі візуалізується весь масив даних.

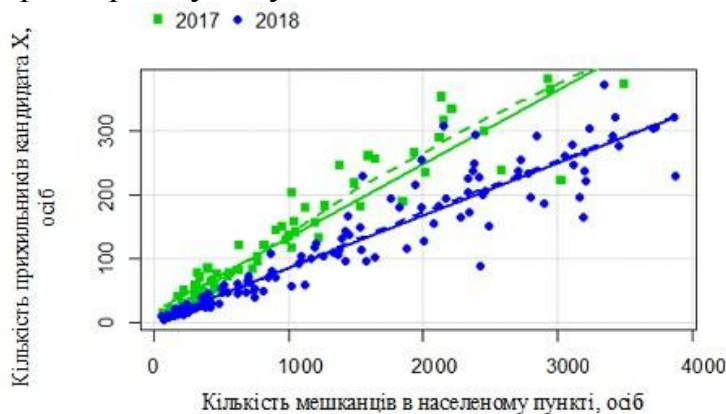


Рис 5.6а.

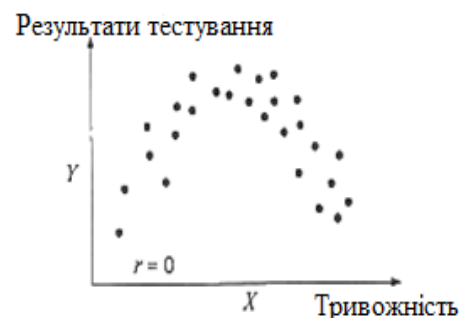


Рис. 5.6.б

Рис. 5.6. Приклади діаграм розсіювання, що демонструють (а) лінійний та (б) нелінійний зв'язки

Діаграми розсіювання дають можливість обрати коефіцієнт кореляції, що буде застосовуватися для подальшого аналізу. Так, якщо діаграма розсіювання демонструє наявність нелінійного зв'язку, дослідник розуміє, що слід відмовитися від використання коефіцієнта кореляції Пірсона, який призначений для відстеження лише лінійних кореляцій, та застосувати, наприклад, кореляційне відношення, яке дозволяє вивчати нелінійні зв'язки.

Для візуалізації залежностей ознак, виміряних **порядковими** шкалами можна застосовувати: 1) діаграми розсіювання, проте вони не дуже наочні (див рис. 5.7); 2) інтервальні середні значення, які корисні для візуалізації загальних тенденцій (рис. 5.8); 3) стовпчасті діаграми (рис. 5.9).

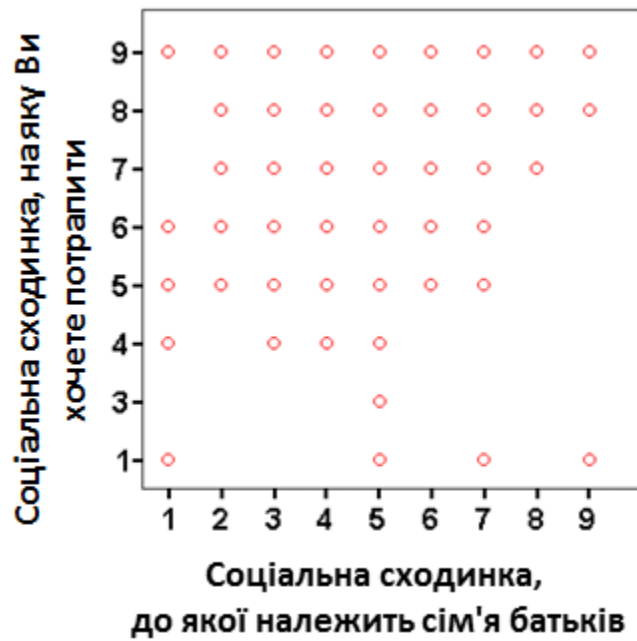


Рис. 5.7. Приклад діаграми розсіювання для порядкових шкал (побудовано в SPSS)

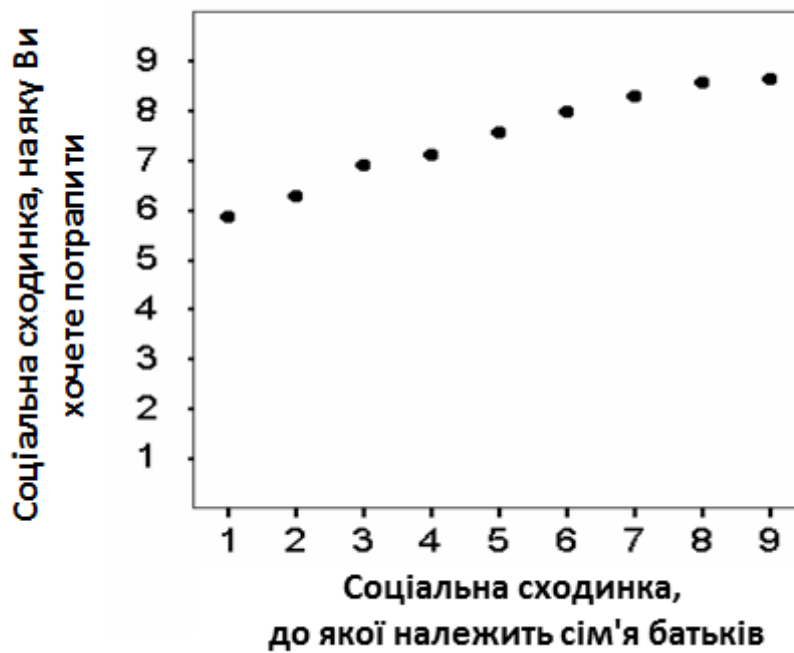


Рис. 5.8. Приклад діаграми для візуалізації середніх значень для кожного інтервалу (побудовано в SPSS)

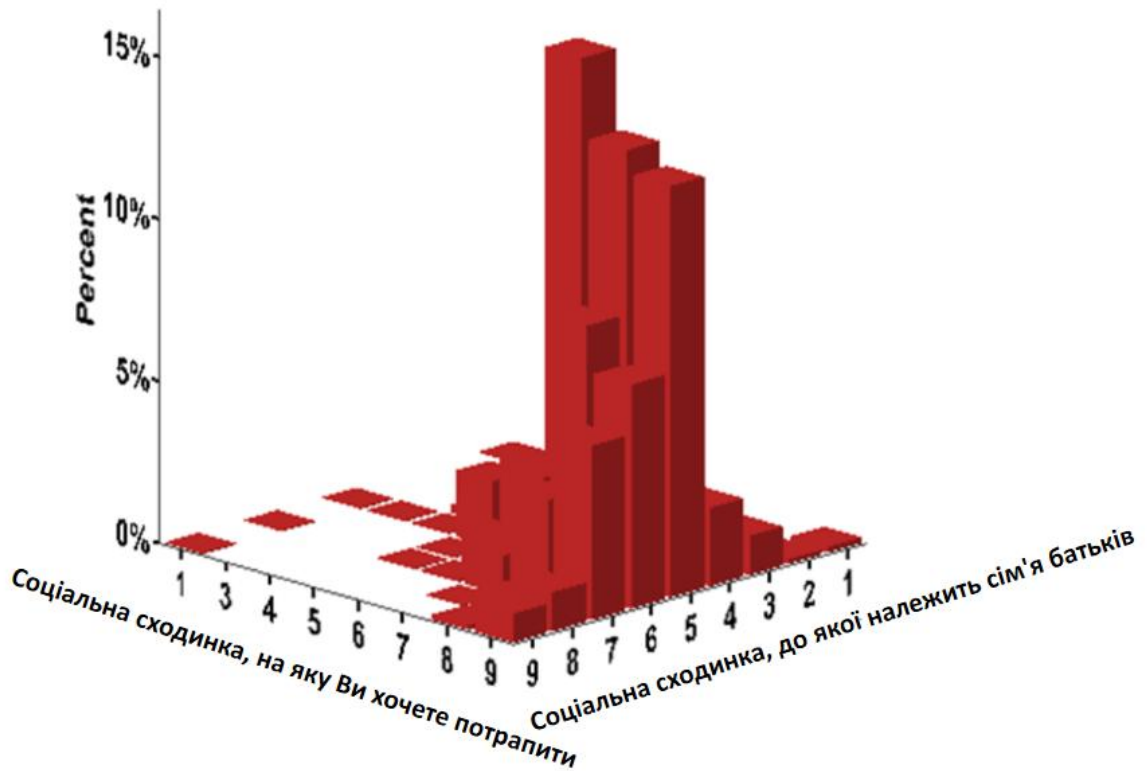


Рис. 5.9. Приклад стовпчастої діаграми для візуалізації порядкових шкал (побудовано в SPSS)

Найпоширенішим способом візуалізації кореляції ознак, виміряних номінальними шкалами, є стовпчасті діаграми (рис. 5.10).

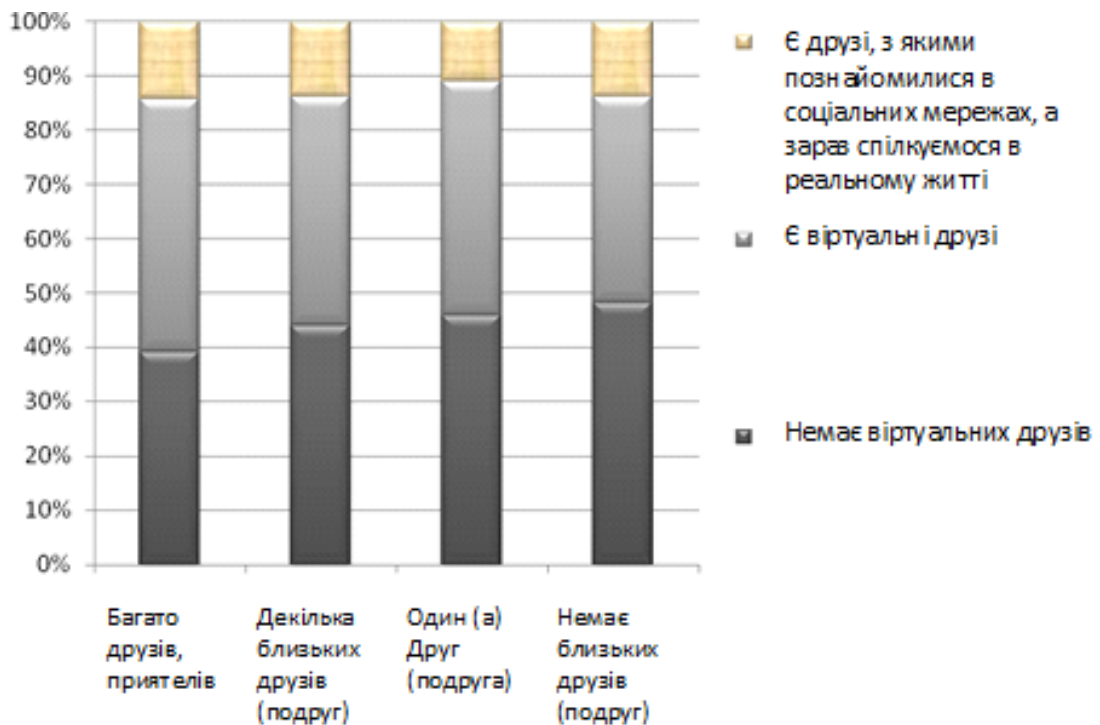


Рис. 5.10. Приклад стовпчастої діаграми для візуалізації номінальних шкал (побудовано в Microsoft Excel)

Література до теми

Бюль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005. С. 180-206.

Горбачик, А.П., Сальникова, С.А. *Аналіз даних соціологічних досліджень засобами SPSS*. – Луцьк: «Вежа», 2008. С. 55-78.

Паніна, Н.В. *Технологія соціологічного дослідження: Курс лекцій / 2-е видання, доповнене*. К., 2007. С.241-269.

Паніотто, В.І., Максименко, В.С., Харченко, Н.М. *Статистичний аналіз соціологічних даних*. К.: КМ Академія, 2004. С. 65-143.

Додаткова література

Daniel Stockemer. *Quantitative Methods for the Social Sciences. A Practical Introduction with Examples in SPSS and Stata*. Springer International Publishing AG 2019. P. 125-132.

Андреева, М.М., Волков, Р. В. Корреляционный анализ в социологических исследованиях. *Вестник Казанского технологического университета*. 2013. № 7 (16). С. 271-274. URL: <http://cyberleninka.ru/article/n/korrelyatsionnyy-analiz-v-sotsiologicheskikh-issledovaniyah>.

Бойко, А.Ф., Блинова, Т.А. Из практики корреляционного анализа в социологических исследованиях. *Международный научно-исследовательский журнал*. 2015. № 8-5 (39). С. 45-47. URL: <http://cyberleninka.ru/article/n/iz-praktiki-korrelyatsionnogo-analiza-v-sotsiologicheskikh-issledovaniyah>.

Кислова О.Н. Визуализация социологических данных как альтернатива традиционным методам дескриптивного анализа. *Методологія, теорія та практика соціологічного аналізу сучасного суспільства. Збірник наукових праць*. Харків: видавничий центр Харківського національного університету імені В.Н. Каразіна, 2008. С. 142-152.

Крыштановский, А.О. *Анализ социологических данных с помощью пакета SPSS*. М.: ГУ ВШЭ, 2007. С. 39-81.

Толстова, Ю.Н. *Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками*. М.: Научный мир, 2003. С. 164-319. URL: Доступно на: <http://www.ecsocman.edu.ru/db/msg/65788>

Питання для самоконтролю

1. Що таке «кореляційний аналіз»?
2. Що таке «двовимірний розподіл»?
3. Як будується двовимірний розподіл?
4. Статистика хі-квадрат: визначення, умови застосування, інтерпретація та види
5. Які коефіцієнти кореляції застосовуються для номінальних даних?

6. Які коефіцієнти кореляції застосовуються для порядкових даних?
7. Які коефіцієнти кореляції застосовуються для метричних даних?
8. На що треба звертати увагу в першу чергу: на силу зв'язку або на її статистичну значущість?
9. Яким чином можна візуалізувати двовимірні розподіли?

Тема 6. Статистичні висновки: статистичне оцінювання та перевірка гіпотез

6.1. Статистичний виведення і статистичні висновки

Зазвичай соціолог досліджує обмежену частину генеральної сукупності (вибірку), по вивченню якої він робить висновки про генеральну сукупність. При цьому, щоб отримати результати, які відповідають дійсності, необхідно застосовувати відповідні аналітичні методи, розроблені в межах математичної статистики та теорії ймовірності.

Процес отримання висновків про якусь сукупність на основі вивчення випадкових вибірок називається *статистичним виведенням*. Сукупність методів статистичного виводу називають *аналітичною (індуктивною) статистикою*.

Застосування методів статистичного виведення є обов'язковим, коли висновки, зроблені на основі вибірових даних, переносяться на всю сукупність, наприклад, при вибіровому дослідженні електоральних настроїв виборців певної країни результати дослідження розповсюджуються на електорат всієї країни. При цьому соціолог повинен мати на увазі, що процедури статистичного виведення можна застосовувати лише до випадкових (імовірнісних) вибірок, які витягнуті з генеральної сукупності за допомогою методів випадкового відбору.

У результаті застосування статистичного виведення дослідник отримує певні статистичні висновки, які зазвичай мають практичне значення.

Статистичний висновок – це певне твердження про параметри досліджуваної генеральної сукупності, зроблені на основі вивчення вибірки. Статистичні висновки завжди мають ймовірнісний характер та підрозділяються на такі різновиди: висновки, зроблені завдяки статистичному оцінюванню (точковому або інтервальному), й висновки, отримані у результаті перевірки статистичних гіпотез (див. рис. 6.1).



Рис. 6.1. Різновиди статистичних висновків

Необхідно зазначити, що математичним підґрунтям процедур статистичного виведення є **закон великих чисел**, загальний зміст якого полягає у тому, що спільна дія великої кількості випадкових факторів призводить до результату, що майже не залежить від випадку, оскільки при великій кількості повторень піддається дуже незначним коливанням.

Описові вибірові характеристики прийнято називати **статистиками**, а відповідні їм характеристики генеральної сукупності – **параметрами**.

Параметри генеральної сукупності – числові характеристики, що описують генеральну сукупність (наприклад, генеральні середні значення, міри варіації, коефіцієнти кореляції тощо).

Статистики – ті самі характеристики, але розраховані для вибірки (вибірові середні, міри варіації, коефіцієнти кореляції тощо).

Статистика є оцінкою параметра

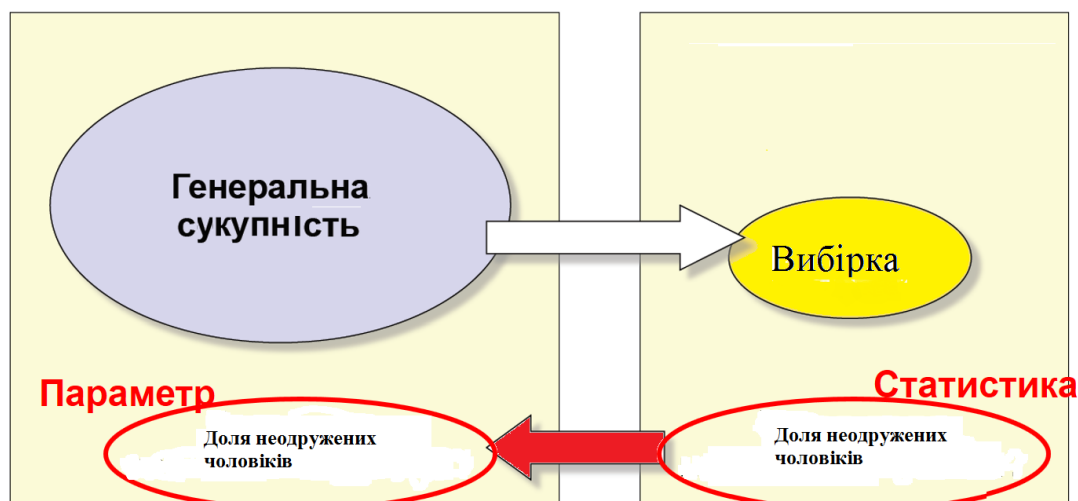


Рис. 6.2. Висновки про параметри генеральної сукупності роблять на основі дослідження статистик вибіркової сукупності

Статистична методологія аналізу даних виходить з того, що вибірові характеристики (статистики) є випадковими величинами, які обчислюються на основі дослідження вибірок, які випадково вилучаються з генеральної сукупності. Параметри генеральної сукупності є не випадковими, але невідомими величинами.

Статистики обчислюються на основі інформації, що міститься у вибірових даних. Вони зовсім не обов'язково будуть тотожними відповідним параметрам генеральної сукупності. Будь-яка вибірка зазвичай дає значення статистик, які відрізняються від дійсних значень параметрів сукупності. У зв'язку із цим виникає потреба визначити, наскільки *точно* (тобто з якою *ймовірністю*) кожна статистика відбиває відповідні параметри генеральної сукупності. Таким чином, при вибіровому дослідженні соціолог повинен урахувувати, що *достовірність висновків про закономірності соціальних явищ, залежить від якості вибірки*, оскільки вибірова сукупність

служує моделлю генеральної сукупності в тому сенсі, що в неї з прийнятною точністю відтворюються статистичні розподіли досліджуваних ознак.

Невідомі параметри генеральної сукупності (популярність політичного лідера, рейтинг телевізійного каналу, ціннісні орієнтації сучасної молоді тощо), які оцінюють за результатами вибіркового дослідження, неможливо визначити точно; мова може йти тільки про приблизну оцінку. В цьому контексті виникає потреба звернутись до поняття «ймовірність».

Ймовірність – числова характеристика можливості того, що випадкова подія відбудеться в умовах, які можуть бути відтворені необмежену кількість раз.

Оскільки вибірка включає тільки частину одиниць генеральної сукупності, в вибіркових дослідженнях завжди існує ризик допустити помилку. Незважаючи на те, що зі збільшенням кількості одиниць спостереження такий ризик зменшується, він все ж має місце у будь-якому вибірковому дослідженні. Висновки, зроблені за результати вибіркового дослідження, завжди мають ймовірнісну природу. Саме тому результати соціологічних досліджень треба подавати сумісно з оцінкою ймовірності помилки оприлюднених висновків (див. приклади на рис. 6.3 - 6.4).

ЯК УКРАЇНЦІ ДОТРИМУЮТЬСЯ УМОВ КАРАНТИНУ

Київський міжнародний інститут соціології (КМІС) з 21 до 24 березня 2020 року провів опитування на онлайн платформі Inpoll (<https://inpoll.org/>). Всього було опитано 402 респондента, що мешкають у всіх регіонах України (на підконтрольній території) за квотною вибіркою, що репрезентативна для міського населення України віком від 18 років, яке користується Інтернетом (приблизно 80% міського населення).

Статистична похибка вибірки (з імовірністю 0.95 і без врахування дизайн-ефекту) не перевищує 5%.

Результати дослідження показують, що майже 90% населення (точніше 88%) турбує поширення коронавірусу в Україні, з них 35% дуже сильно турбує. Ті, кого поширення вірусу не турбує, складають лише 6%.

Рис. 6.3. Презентація результатів дослідження КМІС «Як українці дотримуються умов карантину» (URL: <https://kiis.com.ua/?lang=ukr&cat=reports&id=925&page=1&t=7>)

У соціологічних дослідженнях зазвичай прийнято створювати вибірки таким чином, щоб за будь-якою ознакою результати вибіркового дослідження відрізнялись від параметрів генеральної сукупності не більше, ніж на 5%. Таким чином, соціолог робить висновки з ймовірністю не меншою, ніж 0,95. Тут мається на увазі, що ймовірність помилки не повинна перевищувати 0,05, а ймовірність правильних висновків повинна бути не менше, ніж 0,95. Проте дослідники завжди прагнуть зробити ймовірність помилки якомога меншою (див. рис.6.4).

ДИНАМІКА РЕЙТИНГУ ПІДТРИМКИ ПОЛІТИЧНИХ ЛІДЕРІВ І ПАРТІЙ ПОРІВНЯНО З ПРЕЗИДЕНТСЬКИМИ І ПАРЛАМЕНТСЬКИМИ ВИБОРАМИ 2019 РОКУ

Прес-реліз підготував Антон Грушецький, заступник директора КМІС

З 8 по 18 лютого 2020 року Київський міжнародний інститут соціології (КМІС) провів власне всеукраїнське опитування громадської думки. Методом особистого інтерв'ю опитано 2038 респондентів, що мешкають у 110 населених пунктах усіх регіонів України (крім АР Крим) за 4-х ступеневою стохастичною вибіркою, що репрезентативна для населення України віком від 18 років.

У Луганській і Донецькій областях опитування проводилося тільки на території, що контролюється українською владою.

Статистична похибка вибірки (з імовірністю 0,95 і за дизайн-ефекту 1,5) не перевищує: 3,3% для показників близьких до 50%, 2,8% — для показників близьких до 25%, 2,0% — для показників близьких до 10%, 1,4% — для показників близьких до 5%.

Рейтинг кандидатів на посаду Президента України в першому турі

100% у стовпчику	Рейтинг 1, % серед усіх респондентів	Рейтинг 2, % серед тих, хто визначився з кандидатом	Офіційні результати першого туру виборів, 2019	Різниця
Зеленський Володимир Олександрович	24,8	44,2	30,24	+14,0
Бойко Юрій Анатолійович	7,3	13,1	11,67	+1,4
Порошенко Петро Олексійович	6,5	11,6	15,95	-4,3
Тимошенко Юлія Володимирівна	3,9	6,9	13,4	-6,5
Смешко Ігор Петрович	2,8	4,9	6,04	-1,1
Ляшко Олег Валерійович	2,2	4,0	5,48	-1,5
Гриценко Анатолій Степанович	1,8	3,3	6,91	-3,6
Кошулинський Руслан Володимирович	1,4	2,5	1,62	+0,9
Вілкул Олександр Юрійович	1,1	1,9	4,15	-2,2
Безсмертний Роман Петрович	0,6	1,1	0,14	+0,9
Мороз Олександр Олександрович	0,4	0,8	0,06	+0,7
Кармазін Юрій Анатолійович	0,4	0,7	0,08	+0,6
Тимошенко Юрій Володимирович	0,3	0,6	0,62	+0,0
Інші загалом	1,8	3,3	2,2	+1,0
ЗАКРЕСЛИВ БИ ВСІХ КАНДИДАТІВ У БЮЛЕТЕНІ / ЗІПСУВАВ БЮЛЕТЕНЬ	4,3	1,2	1,18	---
ВАЖКО СКАЗАТИ / НЕ ВИЗНАЧИЛИСЯ	24,6	---	---	---
ВІДМОВА ВІДПОВІДАТИ	1,5	---	---	---
НЕ БРАВ БИ УЧАСТІ У ГОЛОСУВАННІ	14,2	---	---	---

Рис. 6.4. Презентація результатів дослідження КМІС «Динаміка рейтингу підтримки політичних лідерів і партій порівняно з президентськими і парламентськими виборами 2019 року» (URL: <https://kiis.com.ua/?lang=ukr&cat=reports&id=918&page=1&t=1>)

6.2. Статистичне оцінювання параметрів генеральних сукупностей

Розрізняють *точкове* та *інтервальне* оцінювання параметрів генеральної сукупності.

Точкове оцінювання передбачає отримання приблизного значення досліджуваного параметра у вигляді одного числа, розрахованого на основі вибірових даних. Приклади точкових оцінок:

✓ Середній прибуток респондентів з вибірки розглядається як оцінка середнього прибутку осіб, що становлять генеральну сукупність.

✓ Відсотки тих опитаних респондентів, що виказують намір голосувати за певних політичних лідерів, розглядаються як приблизна оцінка електоральних намірів досліджуваної сукупності.

✓ Вибірковий відсоток студентів, які відповіли, що вступили до ВНЗ для того, щоб подовжити час безтурботного існування, розглядається як оцінка цієї характеристики у генеральній сукупності.

✓ Середній час перегляду певного телеканалу, отриманий у наслідку вибірового опитування розглядається як оцінка середнього часу перегляду цього телеканалу особами, що складають генеральну сукупність.

Оскільки точкові оцінки заздалегідь не є точними, то частіше використовують інтервальне оцінювання.

Інтервальне оцінювання – спосіб отримання оцінки для невідомого значення параметра генеральної сукупності за допомогою інтервалу його припустимих значень і визначення ймовірності того, що в цьому інтервалі перебуває істинне значення. Таким чином, замість того щоб користуватися вибіровим середнім ми можемо говорити про інтервал $a < \mu < b$, у якому з тією чи іншою ймовірністю міститься середнє генеральної сукупності.

Довірчим інтервалом називають припустиме відхилення спостережуваних значень від справжніх. Розмір цього припущення визначається дослідником з урахуванням вимог до точності інформації.

Довірчий інтервал будується за даними вибірового дослідження для оцінювання параметра генеральної сукупності.

Довірчий інтервал показує, в якому діапазоні розташуються результати вибірових спостережень. Передбачається, що значення досліджуваного параметра із заданою довірчою ймовірністю P перебуває в цьому інтервалі. Якщо довірчий інтервал звузити, то ймовірність попадання в нього досліджуваного параметра (наприклад, середнього значення) зменшиться. Відповідно при розширенні інтервалу ймовірність стає більше.

Довірчий інтервал – це інтервал, що покриває оцінюваний параметр генеральної сукупності з заданою довірчою ймовірністю.

Довірча ймовірність являє собою надійність виміру та показує ймовірність припустимої помилки. Її прийнято позначати $P = (1 - \alpha)$, де α – це ймовірність помилки. Довірча ймовірність задається дослідником, який відповідно до вимог щодо точності інформації вибирає P зі стандартних значень 0.95, 0.99, 0.999. Стандартні значення ймовірності помилки α (тобто

ймовірності того, що значення параметра перебуває поза межами даного інтервалу) становить, відповідно, 0.05, 0.01 чи 0.001.

6.3. Довірчий інтервал для середнього

Побудуємо довірчий інтервал для середнього генеральної сукупності, що відповідає нормальному закону розподілу. Припустимо, у нас є проста випадкова вибірка з цієї генеральної сукупності. Обсяг вибірки дорівнює n . Потрібно побудувати довірчий інтервал, який з заданою довірчою ймовірністю P буде містити середнє генеральної сукупності:

$$\bar{x} - E < \mu < \bar{x} + E \quad (6.1)$$

де \bar{x} - середнє арифметичне значення змінної, обчислене за вибіркою;
 μ - середнє генеральної сукупності;
 E – точність інтервальної оцінки.

Візуально довірчий інтервал для середнього можна представити наступним чином (див. рис. 6.5):



Рис. 6.5. Довірчий інтервал для середнього

За вибіркою ми можемо обчислити вибіркоче середнє \bar{x} , тобто точкову оцінку для середнього генеральної сукупності. Отже знаходження довірчого інтервалу зводиться до обчислення точності інтервальної оцінки E , знання якої дозволить обчислити межі довірчого інтервалу. При цьому ми будемо ґрунтуватися на відомих властивостях закону нормального розподілу.

Існує два випадки, які необхідно розглянути:

- 1) вибірка велика ($n \geq 30$);
- 2) вибірка мала ($n \leq 30$).

Розрахунок довірчого інтервалу для генерального середнього на основі даних великих вибірок ($n \geq 30$)

У випадку великої вибірки, обсяг якої $n \geq 30$, середнє генеральної сукупності, що має нормальний закон розподіл, з довірчою ймовірністю $P = 1 - \alpha$ знаходиться в інтервалі:

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (6.2)$$

де \bar{x} - середнє арифметичне значення змінної, обчислене за вибіркою;

n – обсяг вибірки;

$z_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;

σ – стандартне відхилення.

Точність інтервальної оцінки E знаходиться за формулою:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \text{ де}$$

n – обсяг вибірки;

$z_{\alpha/2}$ – довірчий коефіцієнт для довірчої ймовірності $P = 1 - \alpha$;

σ – стандартне відхилення в генеральній сукупності.

Послідовність дій для знаходження довірчого інтервалу

1. За вибіркою обчислити вибіркове середнє; визначити значення стандартного відхилення (якщо для генеральної сукупності σ невідомо, для великих вибірок можна застосовувати його вибіркову оцінку s).

2. По таблиці нормального закону знайти z -значення для обраної дослідником довірчої ймовірності P (див рис. 6.6).

Значення довірчої ймовірності задає сам дослідник залежно від того, яку ступінь точності розрахунків вимагає дослідження.

3. Обчислити точність інтервальної оцінки за формулою:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

4. Підставити отримані значення в формулу для довірчого інтервалу:

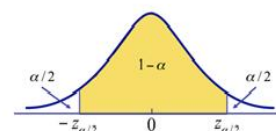
$$\bar{x} - E < \mu < \bar{x} + E$$

5. Написати відповідь.

Як по таблиці нормального закону знайти z -значення для обраної дослідником довірчої ймовірності P ?

Стандартний нормальний розподіл

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890



Довірча ймовірність дорівнює $P = 0,95$
 Їй відповідає $\alpha = 0,05$, $\alpha/2 = 0,025$.
 Площа = $0,95 + 0,025 = 0,975$

Довірча ймовірність	Площа	Z-значення
0,95 чи 95%	0,9750	1,96

Рис. 6.6. Знаходження z-значення по таблиці нормального закону для певної довірчої ймовірності P

Таблиця 6.1

Z-значення для найчастіше використовуваних довірчих ймовірностей

Z-значення	Довірча ймовірність
Z=1,96	P=95%
Z=2	P=95,5%
Z=2,56	P=99%
Z=3	P=99,7%

Приклад розрахунку довірчого інтервалу середнього ($n \geq 30$)

Припустимо, що ректор університету хоче дізнатися, який середній вік студентів, які навчаються на денному відділенні. З попередніх досліджень відомо, що стандартне відхилення дорівнює 2 роки. За допомогою випадкового відбору відібрано 50 студентів та обчислено вибіркоче середнє, яке дорівнює 20,3 р. Завдання: побудувати 95%-ий довірчий інтервал для середнього генеральної сукупності.

Вирішимо це завдання, послуговуючись раніше запропонованою послідовністю дій.

1. Вибіркове середнє $\bar{x} = 20,3$, стандартне відхилення відомо з попередніх досліджень: $\sigma = 2$.

2. Знайдемо по таблиці 6.1 довірчий коефіцієнт, який визначається у відповідності до заданої довірчої ймовірності. В умові завдання $P = 95\%$. Отже, згідно з таблицею z-значення $z_{\alpha/2} = 1,96$.

3. Обчислимо точність інтервальної оцінки:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1,96 \frac{2}{\sqrt{50}} = 0,55$$

4. Підставимо отримані значення в формулу для довірчого інтервалу (6.1):

$$20,3 - 0,55 < \mu < 20,3 + 0,55$$

5. Відповідь. Середній вік студентів університету з ймовірністю 0,95 знаходиться в інтервалі між 19,75 р. та 20,85 р., іншими словами:

$$19,75 < \mu < 20,85$$

Розрахунок довірчого інтервалу для генерального середнього на основі даних малих вибірок ($n \leq 30$)

Довірчий інтервал для середнього генеральної сукупності, що відповідає закону нормального розподілу, з довірчою ймовірністю $P = 1 - \alpha$ знаходиться в інтервалі:

$$\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}$$

(6.3),

де \bar{x} - середнє арифметичне значення змінної, обчислене за вибіркою;

n – обсяг вибірки;

$t_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;

s – вибіркове стандартне відхилення.

При побудові довірчого інтервалу замість нормального розподілу використовують t-розподіл Стюдента. Розподіл Стюдента схожий на стандартний нормальний розподіл, він має дзвоноподібну форму, симетричний відносно середнього, крива не стикається з віссю X.

Розподіл Стюдента відрізняється від стандартного нормального розподілу тим, що дисперсія t-розподілу більше 1, цей розподіл являє собою сімейство кривих, що розрізняються кількістю ступенів свободи. Зі збільшенням обсягу вибірки t-розподіл наближається до нормального.

Кількість ступенів свободи (degrees of freedom) - це кількість значень, які можуть вільно змінюватися після того, як за вибіркою було обчислено значення статистики, тобто кількість ступенів свободи t-розподілу при побудові довірчого інтервалу на одиницю менше обсягу вибірки: $df = n - 1$

Точність інтервальної оцінки E знаходиться за формулою:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}, \text{ де}$$

n – обсяг вибірки;

$t_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;

s – вибіркове стандартне відхилення.

Послідовність дій для знаходження довірчого інтервалу:

1. За вибіркою обчислити вибіркове середнє \bar{x} та визначити вибіркє стандартне відхилення s .

2. За таблицею розподілу Ст'юдента знайти t -значення для довірчої ймовірності $P=1 - \alpha$ та кількості ступенів свободи $df = n - 1$ (див. рис. 6.7).

Для знаходження t -значень використовують таблиці розподілу Ст'юдента.

3. Обчислити точність інтервальної оцінки за формулою:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

4. Підставити отримані значення в формулу для довірчого інтервалу:

$$\bar{x} - E < \mu < \bar{x} + E$$

5. Написати відповідь.

Як знайти t -значення по таблиці розподілу Ст'юдента при заданій довірчій ймовірності P та певній кількості ступенів свободи?

Розподіл Ст'юдента

df	Одностороння область				
	0,005	0,01	0,025	0,05	0,1
	Двостороння область				
	0,01	0,02	0,05	0,1	0,2
1	63,656	31,821	1,706	6,314	3,078
2	9,925	6,965	4,303	2,920	1,886
3	5,841	4,541	3,182	2,353	1,638
4	4,604	3,747	2,776	2,132	1,533
5	4,032	3,365	2,571	2,015	1,476
6	3,707	3,143	2,447	1,943	1,440
7	3,499	2,998	2,365	1,895	1,415
8	3,355	2,896	2,306	1,860	1,397
9	3,250	2,821	2,262	1,833	1,383
10	3,169	2,764	2,238	1,812	1,372
11	3,106	2,718	2,219	1,796	1,363
12	3,055	2,681	2,201	1,782	1,356
13	3,012	2,650	2,186	1,771	1,350
14	2,977	2,624	2,172	1,761	1,345
15	2,947	2,602	2,159	1,753	1,341
16	2,921	2,583	2,148	1,746	1,337
17	2,898	2,567	2,138	1,740	1,333
18	2,878	2,552	2,129	1,734	1,330
19	2,861	2,539	2,121	1,729	1,328
20	2,845	2,528	2,114	1,725	1,325
24	2,831	2,518	2,107	1,721	1,323

Рис. 6.7. Знаходження t -значення по таблиці розподілу Ст'юдента при заданій довірчій ймовірності P та певній кількості ступенів свободи

Приклад розрахунку довірчого інтервалу для середнього на основі даних малих вибірок ($n \leq 30$)

Припустимо, що ми бажаємо визначити, як б'ється серце студентів на іспиті. У 20 студентів, які здавали державний іспит, серце билось в

середньому зі швидкістю 96 ударів в хвилину. Стандартне відхилення вибірки дорівнює 5 ударів у хвилину. Завдання: знайти 95% довірчий інтервал для генерального середнього.

Вирішимо це завдання.

1. Нам відомо, що вибіркоче середнє дорівнює 96, стандартне відхилення 5, тобто $\bar{x} = 96, s = 5$.

2. Обсяг вибірки $n = 20$, отже кількість ступенів свободи $df = 20 - 1 = 19$. Довірча ймовірність 95% чи 0,95, тобто $\alpha = 0,05$. Знаходимо t -значення по таблиці розподілу Стьюдента, користуючись значеннями для двосторонньої області, та виявляємо, що $t_{\alpha/2} = 2,093$ (див. рис. 6.7).

3. Обчислюємо точність інтервальної оцінки:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}} = 2,093 \frac{5}{\sqrt{20}} = 2,34$$

4. Підставимо отримані значення в формулу для розрахунку довірчого інтервалу:

$$96 - 2,34 < \mu < 96 + 2,34.$$

5. Пишемо відповідь. Середнє число ударів серця у студентів, які здавали державний іспит, з довірчою ймовірністю 95% знаходиться в межах:

$$93,66 < \mu < 98,34 \text{ (ударів в хвилину).}$$

6.4. Довірчий інтервал для частки (відсотку)

Розглянемо, як побудувати довірчий інтервал для частки (відсотку) ознаки. Пам'ятаємо, що частка (відносна частота) є результатом ділення відсотку на 100, відповідно, частка помножена на 100 – це відсоток.

Отже, припустимо, що у нас є випадкова вибірка обсягу n з генеральної сукупності обсягу N та нам відомо q – частку респондентів, які дали певну відповідь на запитання анкети (див. рис. 6.8).

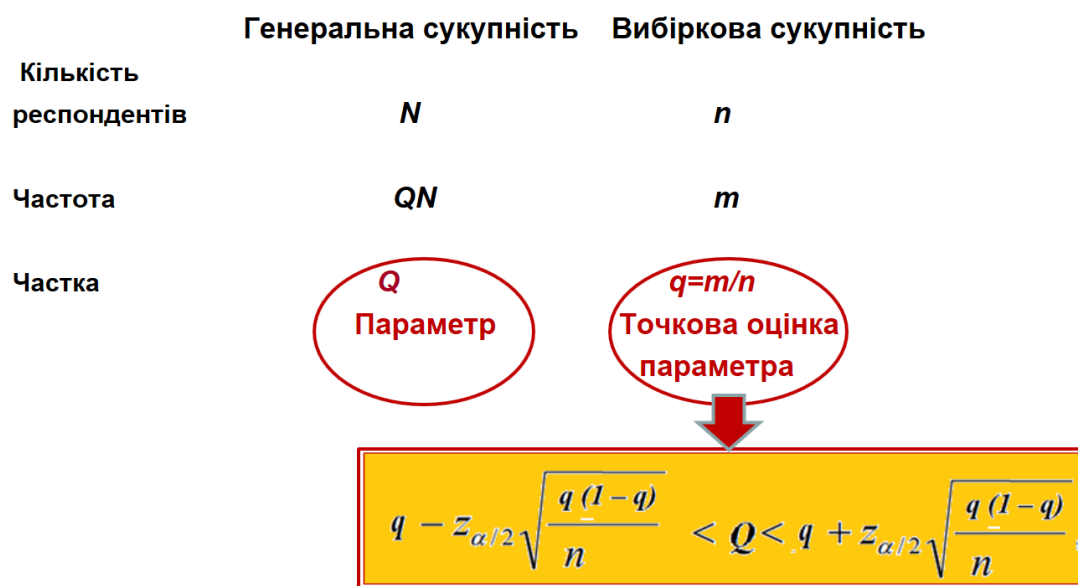


Рис. 6.8. Довірчий інтервал для частки як інтервальна оцінка параметра генеральної сукупності

Ми будемо шукати інтервальну оцінку параметру генеральної сукупності в наступному вигляді:

$$q - E < Q < q + E, \text{ де}$$

Q – частка досліджуваної ознаки у генеральній сукупності;

q – частка досліджуваної ознаки у вибірці;

E – точність інтервальної оцінки.

З теорії ймовірностей відомо, що для послідовності експериментів, значення яких змінюється за принципом *так/ні*, застосовують біноміальний розподіл – дискретний ймовірнісний розподіл, що характеризує кількість успіхів кожен з яких відбувається з ймовірністю p . Формули, засновані на біноміальному розподілі, точні, але дуже громіздкі. У зв'язку з цим соціологи зазвичай застосовують нормальний розподіл, який достатньо добре апроксимує біноміальний, якщо вибірки не занадто малі та виконуються наступні умови: $nq > 5$ та $n(q-1) > 5$. У цьому випадку точність інтервальної оцінки E розраховують за формулою:

$$E = z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}}, \text{ де}$$

q – частка досліджуваної ознаки у вибірці;

$z_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;

n – обсяг вибірки.

Отже, формула для розрахунку довірчого інтервалу для частки (відсотку) приймає вигляд:

$$q - z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}} < Q < q + z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}}, \quad (6.4),$$

де Q – частка досліджуваної ознаки у генеральній сукупності;

q – частка досліджуваної ознаки у вибірці;

$z_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;

n – обсяг вибірки.

Послідовність дій для знаходження довірчого інтервалу для долі (проценту):

1. За вибіркою обчислити частку ознаки.
2. Перевірити умови застосування нормального розподілу: $n > 30$, $nq > 5$ та $n(q-1) > 5$.
3. Знайти z -значення для заданої довірчої ймовірності $P = 1 - \alpha$.
4. Обчислити точність інтервальної оцінки за формулою:

$$E = z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}}$$

5. Підставити отримані значення в формулу для довірчого інтервалу:

$$q - E < Q < q + E$$

6. Написати відповідь.

Приклад. Підтримка для мера. В результаті проведеного опитування 829 жителів міста з'ясувалося, що 417 опитаних (51,5%) планують підтримати на майбутніх виборах кандидатуру чинного мера. Кореспондент місцевого ЗМІ оприлюднив цю інформацію та зробив висновок, що більше половини жителів міста підтримують кандидатуру діючого мера.

Чи правильний висновок зробив кореспондент? Можна чи ні на підставі наявних даних стверджувати, що більше половини жителів міста підтримують перевибори діючого мера на наступний термін?

Щоб відповісти на ці запитання необхідно розрахувати довірчий інтервал, що ми й зробимо.

1. За результатами дослідження 51,5% опитаних планують підтримати на майбутніх виборах кандидатуру чинного мера. Отже, частка ознаки у вибірці складає 0,515.

2. Перевіримо умови застосування нормального розподілу: $n = 829 > 30$, $nq = 829 \cdot 0,515 > 5$ та $n(1-q) = 829 \cdot (1-0,515) > 5$.

3. Виберемо рівень довірчої ймовірності 95%, для нього, як відомо, z -значення дорівнює 1,96.

4. Обчислимо точність інтервальної оцінки:

$$E = z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}} = 1,96 \sqrt{\frac{0,515 \cdot 0,485}{829}} = 0,034$$

5. Підставимо отримані значення в формулу для довірчого інтервалу:

$$0,515 - 0,034 < Q < 0,515 + 0,034$$

6. Напишемо відповідь:

$$0,481 < Q < 0,549 \text{ чи } 48,1\% < Q < 54,9\%.$$

Отриманий результат можна інтерпретувати в такий спосіб. Частка ознаки генеральної сукупності з ймовірністю 95% знаходиться в межах між 48,1% і 54,9% голосів. Отже, несправедливо по результатам опитування стверджувати, що більше половини виборців будуть голосувати за вибори діючого мера на наступний термін.

6.5. Розрахунок довірчих інтервалів в SPSS

Довірчий інтервал для середнього значення змінної в SPSS можна обчислити, виконавши команду *Analyze (Аналіз) → Descriptive Statistics*

(Дескриптивні статистики) → *Explore...* (Досліджувати), як показано на рис. 6.9.

Найпростіший спосіб розрахувати довірчий інтервал в SPSS полягає у використанні можливостей розвідувального аналізу.

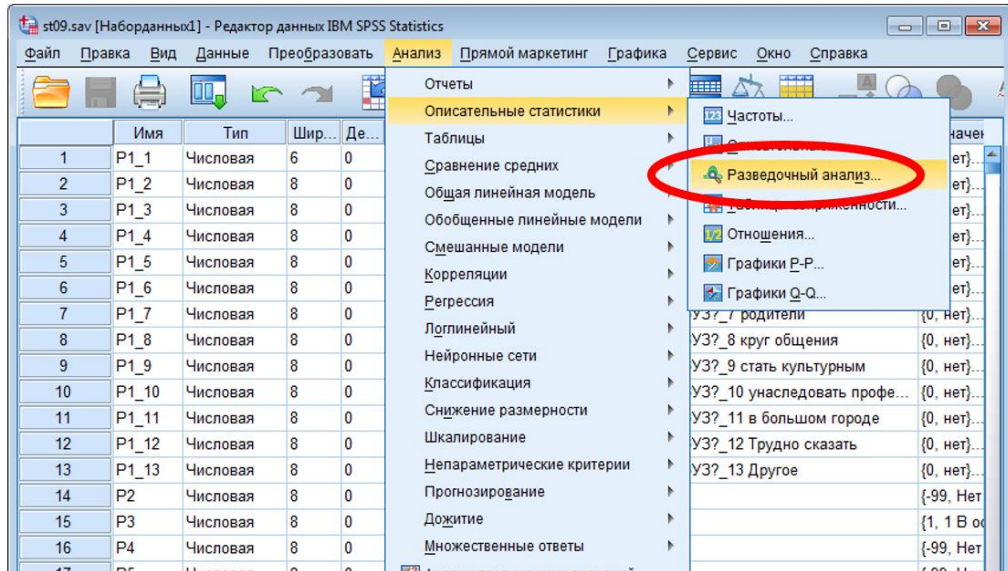


Рис. 6.9. Виклик процедури розрахунку довірчих інтервалів в SPSS

Відкриється діалогове вікно *Explore (Досліджувати)*, в якому потрібно задати змінні, що будуть досліджуватися, та рівень довірчої ймовірності (рис. 6.10).

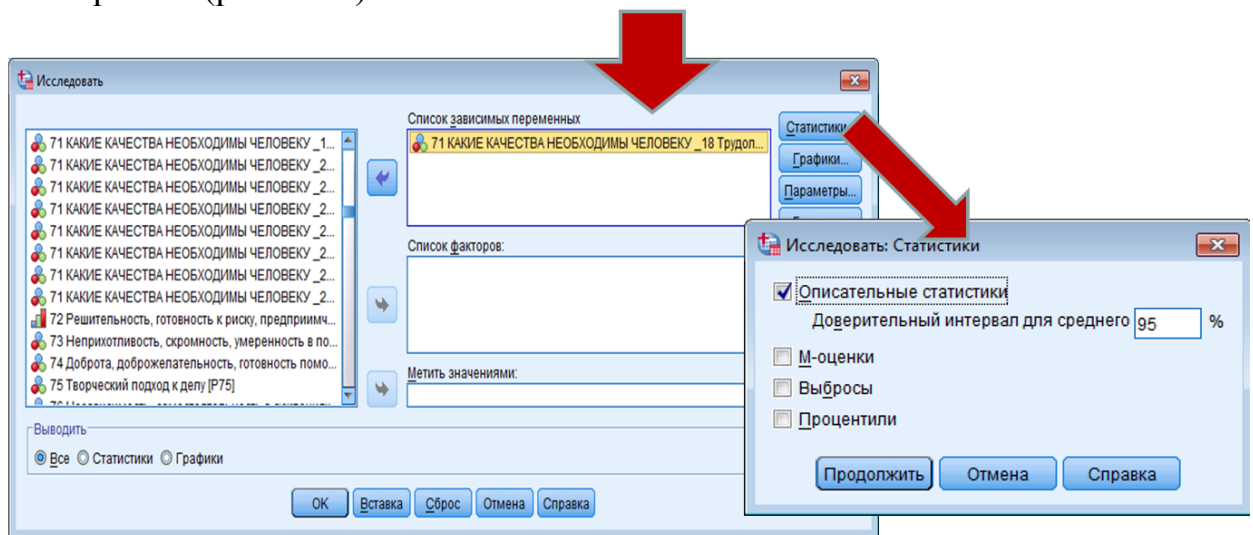


Рис. 6.10. Розрахунок довірчих інтервалів в SPSS

У результаті отримаємо таке:

Описательные

		Статистика	Стд. ошибка	
71 КАКИЕ КАЧЕСТВА НЕОБХОДИМЫ ЧЕЛОВЕКУ_18 Трудолюбие	Среднее	,3600	,00868	
	95% доверительный интервал для среднего	Нижняя граница Верхняя граница	,3430 ,3771	
	5% усеченное среднее		,3445	
	Медиана		,0000	
	Дисперсия		,230	
	Стд. отклонение		,48009	
	Минимум		,00	
	Максимум		1,00	
	Размах		1,00	
	Межквартильный размах		1,00	
	Асимметрия		,583	,044
	Экссесс		-1,661	,089

Рис. 6.11. Результати виконання процедури Explore у пакеті SPSS

Отже, довірчий інтервал складає (0,343; 0,5771), тобто відсоток студентів, які вважають, що працьовитість є необхідною якістю людини, в генеральній сукупності міститься між 34,3% та 37,7%.

На що треба звернути увагу!

1. SPSS розраховує довірчі інтервали тільки для середнього. Якщо виникає потреба розрахувати довірчий інтервал для відсотку, то необхідно номінальну шкалу перетворити у фіктивні змінні, які приймають значення 0-ні та 1-так. У цьому випадку середнє значення буде інтерпретуватися як частка.

2. Якщо розрахунки довірчого інтервалу вручну та в SPSS у вас не точно співпадають, перевірте виконання умов. Для малих вибірок застосовуємо тест Стьюдента, для великих - нормальний розподіл. Для частки майже на увазі, що наведена в лекції формула - це апроксимація. Тут для повної точності треба застосовувати біноміальний розподіл.

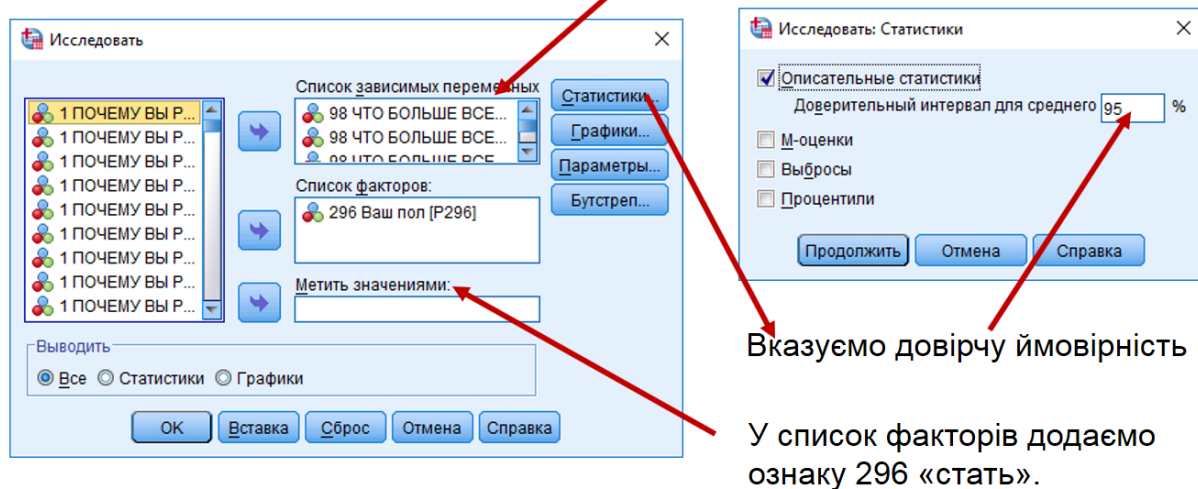
Побудова довірчих інтервалів для окремих груп респондентів

Для цього виконаємо команду *Аналіз* → *Описові статистики* → *Розвідницький аналіз*. У список залежних змінних додаємо ознаку «чинники досягнення життєвого успіху». У список факторів - стать. Після чого натискаємо кнопку «*Статистики*» та вказуємо рівень довірчої ймовірності (рис. 6.12).

Побудова довірчих інтервалів для окремих груп респондентів

Виконаємо команду **Аналіз** → **Описові статистики** → **Розвідницький аналіз**.

У список залежних змінних додаємо ознаку 98 «чинники досягнення життєвого успіху».



Вказуємо довірчу ймовірність

У список факторів додаємо ознаку 296 «стать».

Рис. 6.12. Побудова довірчих інтервалів для окремих груп респондентів

У результаті отримуємо довірчі інтервали для чоловіків і жінок за чинниками досягнення життєвого успіху (рис. 6.13).

		Описательные		Статистика	Стд. ошибка
		296 Ваш пол			
98 ЧТО БОЛЬШЕ ВСЕГО СПОСОБСТВУЕТ ЖИЗНЕННОМУ УСПЕХУ? _1 Влиятельные друзья	1 Мужской	Среднее		,5360	,01590
		95% доверительный интервал для среднего	Нижняя граница	,5048	
			Верхняя граница	,5672	
			5% усеченное среднее	,5400	
			Медиана	1,0000	
			Дисперсия	,249	
			Стд. отклонение	,49895	
			Минимум	,00	
			Максимум	1,00	
			Размах	1,00	
			Межквартильный размах	1,00	
			Асимметрия	-,145	,078
			Эксцесс	-1,983	,156
	2 Женский	Среднее			,4538
95% доверительный интервал для среднего			Нижняя граница	,4320	
			Верхняя граница	,4756	
			5% усеченное среднее	,4487	
			Медиана	,0000	
			Дисперсия	,248	
			Стд. отклонение	,49799	
		Минимум	,00		

Рис. 6.13. Результат побудови довірчих інтервалів для окремих груп респондентів

Цікавою є можливість обчислювати та одночасно візуалізувати довірчі інтервали в таблицях, що настроюються. Ця можливість є у версіях SPSS, починаючи з 24. У старіших версіях такого зробити не можна.

Для візуалізації довірчих інтервалів виконаємо наступні дії (рис. 6.14 та 6.15):

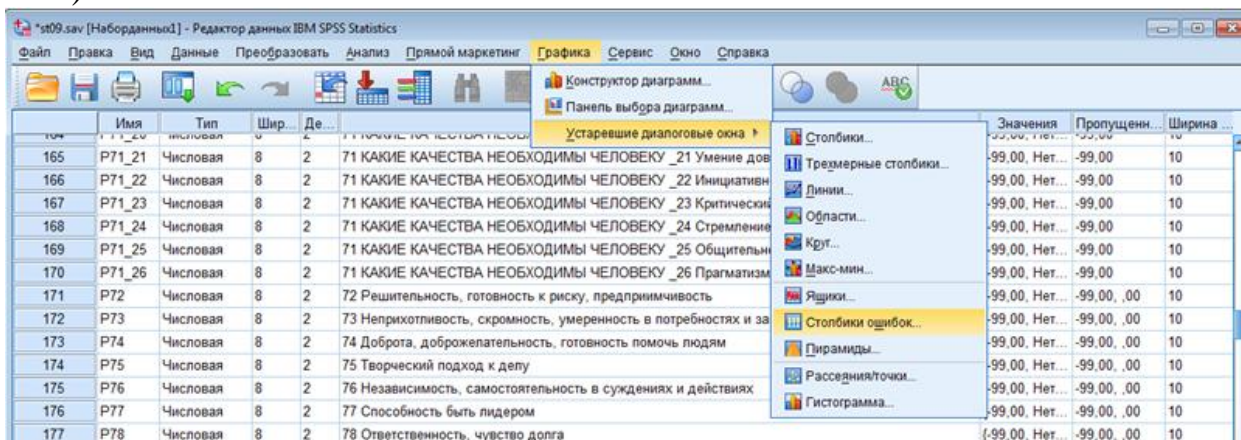


Рис. 6.14. Створення простих діаграм стовпчиків помилок підсумків по групах спостережень

Після виконання команди *Графіка* → *Столбчатые ошибки* задаємо змінні: кількісну, для середнього якої розраховується та візуалізується довірчий інтервал, та якісну, що слугує для поділу на групи (див. рис. 6.15).

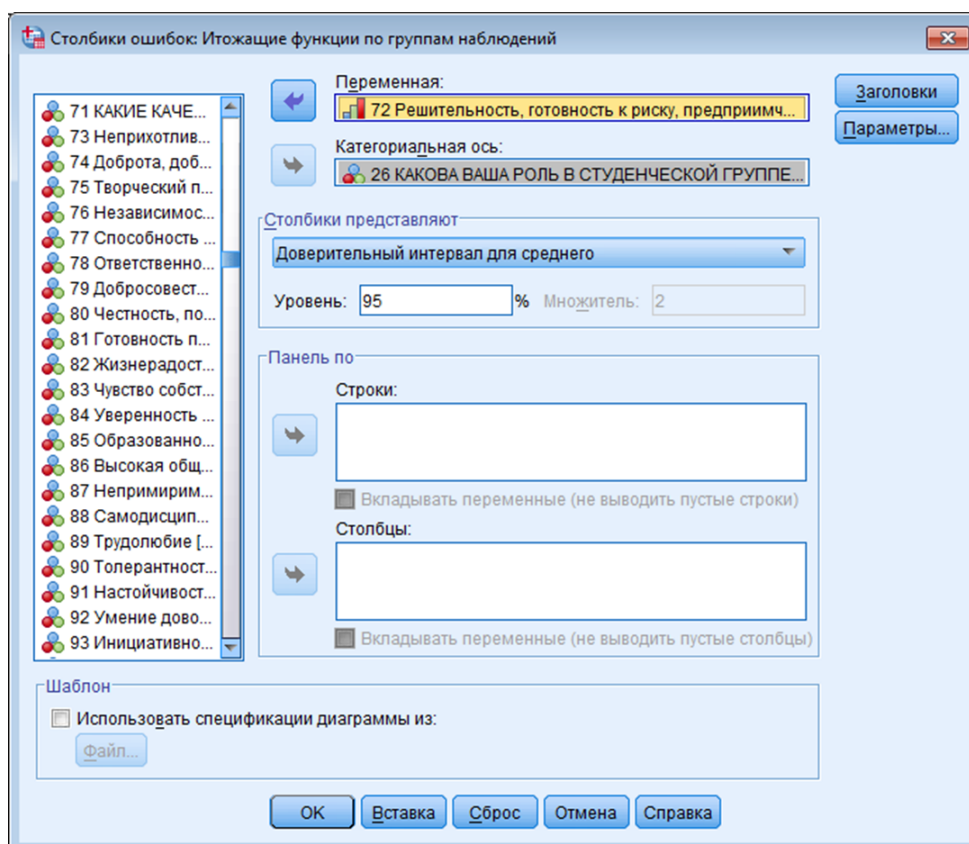


Рис. 6.15. Задаємо змінні та довірчу ймовірність

У результаті створюється діаграма, що підсумовує розподіл однієї кількісної змінної за категоріями іншої змінної. Змінна, по якій респонденти поділяються на групи, переміщується у в поле «*категоріальна вісь*». Ця змінна може бути як числовою, та й текстовою. Стовпчики помилок для змінної створюються для кожної категорії змінної категоріальної осі.

Потім треба вибрати одну альтернативу зі списку «*Стовпчики представляють*», щоб задати характеристику, яка буде представлена стовпчиками помилок. Ми обираємо «*Довірчий інтервал для середнього*».

У полі «*Рівень*» задаємо бажаний рівень довіри – довірчу ймовірність. Тиснемо **OK** та отримуємо результат:

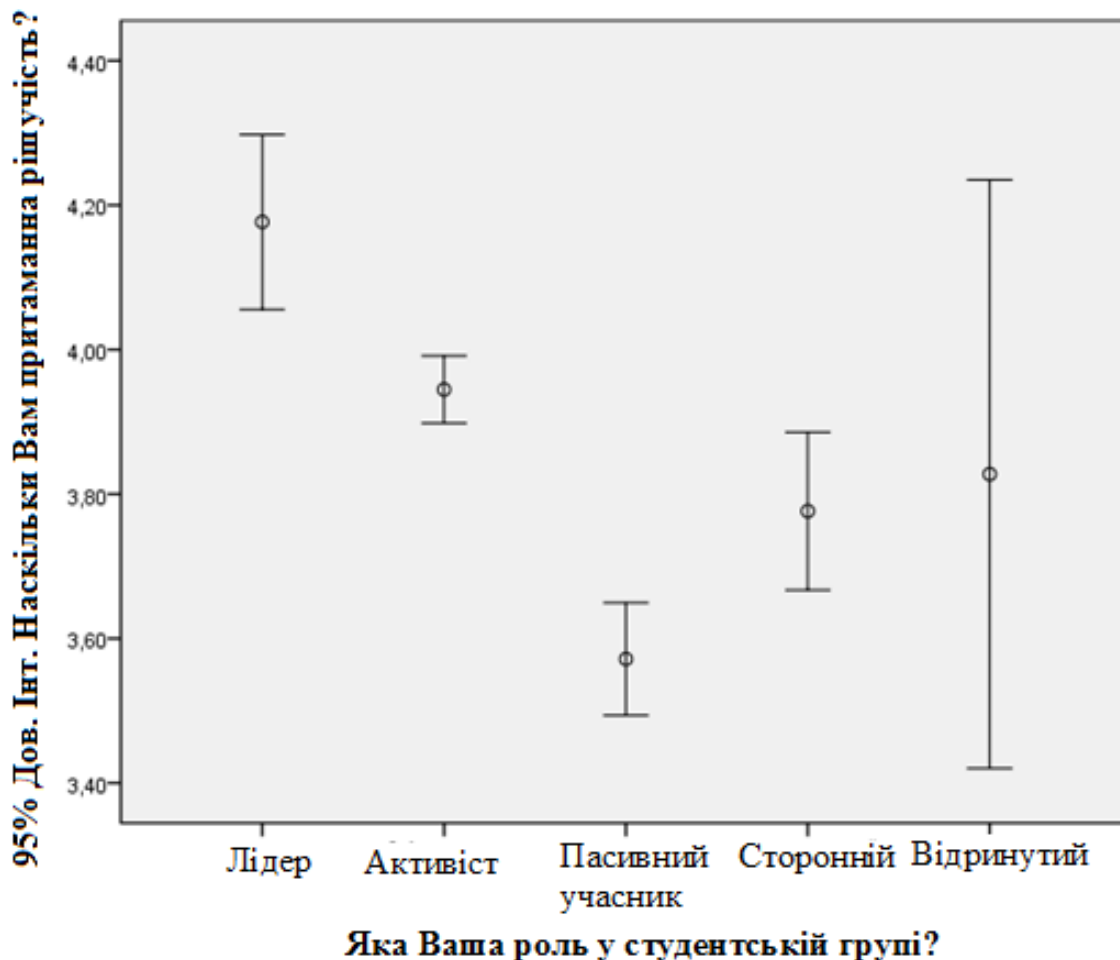


Рис. 6.16. Результати розрахунку та візуалізації довірчих інтервалів по групах

6.6. Приклади побудови довірчих інтервалів в SPSS та вручну

1. Побудова довірчого інтервал для частки (в SPSS і вручну)

Обираємо будь-яку дихотомічну ознаку з масиву, у якій 1 – «так» (альтернатива обрана), 0 – «ні» (альтернатива не обрана). В анкеті ця ознака вимірюється за допомогою шкали номінальної з сумсними альтернативами.

Наприклад, беремо масив st09.sav, питання «Частіше за все я згадую про свою країну, коли стикаюся з...» (вкажіть не більше 3 варіантів відповіді):

1. Іноземцями
2. Державними установами або їх представниками (правоохоронними органами, медичними або освітніми закладами, збройними силами, прикордонною/митною, податковою службою і т. ін.)
3. Перемогами/досягненнями своїх співвітчизників
4. Поразками/невдачами своїх співвітчизників
5. Державними символами
6. Людьми, які розмовляють іншою мовою
7. Діяльністю держави з утвердження державної мови, певних цінностей, поглядів, релігії (конфесії)
8. Згадуваннями про історію мого народу
9. Діяльністю політичних діячів
10. Творами українського мистецтва (діятература, музика, живопис, театр і т.ін.)

Розрахуємо в SPSS довірчий інтервал для частки тих, хто згадує про свою країну, коли стикається з творами українського мистецтва (ознака P202_10). В (Аналіз → Описові статистики → Розвідницький аналіз). Отримуємо:

Таблиця 6.2

Зведення обробки спостережень

	Спостереження					
	Валідні		Пропущені		Всього	
	N	Відсоток	N	Відсоток	N	Відсоток
202 НАЙЧАСТІШЕ Я ЗГАДУЮ ПРО СВОЮ КРАЇНУ, КОЛИ СТИКАЮСЯ З ТВОРАМИ УКРАЇНСЬКОГО МИСТЕЦТВА	3058	100,0%	0	,0%	3058	100,0%

Таблиця 6.3

Описові статистики

		Статистика	Стд. похибка
202	Середнє	0,3676	0,00872
НАЙЧАСТІШЕ Я ЗГАДУЮ ПРО СВОЮ КРАЇНУ, КОЛИ СТИКАЮСЯ З ТВОРАМИ УКРАЇНСЬКОГО МИСТЕЦТВА	95% довірчий інтервал	Нижня границя 0,3505	
	для середнього	Верхня границя 0,3847	
	5% усічене середнє	0,3528	
	Медіана	0,0000	
	Дисперсія	0,233	
	Стд. відхилення	0,48222	
	Мінімум	0,00	
	Максимум	1,00	
	Розмах	1,00	
	Міжквартильний розмах	1,00	
	Асиметрія	0,550	0,044
	Ексцес	-1,699	0,089

Користуючись формулою розрахунку довірчого інтервалу для частки, перевіряємо результати розрахунків в SPSS.

$$q - z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}} < Q < q + z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}} \quad (6.4)$$

де Q – частка досліджуваної ознаки у генеральній сукупності;
 q – частка досліджуваної ознаки у вибірці;
 $z_{\alpha/2}$ – довірчий коефіцієнт, який відповідає довірчій ймовірності $P = 1 - \alpha$;
 n – обсяг вибірки.

Довірчий коефіцієнт

- $Z=1$
- $Z=1,65$
- **$Z=1,96$**
- $Z=2$
- **$Z=2,56$**
- $Z=3$

Довірча ймовірність

- $P=0,683$ или 68,3%
- $P=0,90$ или 90%
- $P=0,95$ или 95%**
- $P=0,955$ или 95,5%
- $P=0,99$ или 99%**
- $P=0,997$ или 99,7%

У вибірковій сукупності частка респондентів, які найчастіше згадують про свою країну, коли стикаються з творами українського мистецтва, дорівнює приблизно 0,37 (точніше 0,3676, як показано у таблиці 6.3). У генеральній же сукупності з довірчою ймовірністю 0,95 частка знаходиться у довірчому інтервалі між 0,35 та 0,38, тобто у відсотковому вираженні: від 35% до 38%.

Розраховуємо довірчий інтервал вручну:

$$E = z_{\alpha/2} \sqrt{\frac{q(1-q)}{n}}$$

$$E = 1,96 * \sqrt{\frac{0,37 * 0,63}{3058}} = 0,017 = 0,02.$$

$$0,37 - 0,02 < Q < 0,37 + 0,02$$

$$0,35 < Q < 0,39$$

Розраховані вручну значення довірчого інтервалу збігаються з розрахунками в SPSS.

2. Побудова довірчого інтервалу для середнього (в SPSS і вручну)

Обираємо порядкову ознаку з 5 та більше градаціями (її можна вважати псевдометричною). Наприклад, беремо масив st09.sav, питання «Наскільки цінним особисто для вас є цікава, творча робота?», варіанти відповіді:

1. Зовсім не цінно
2. Не дуже цінно
3. Важко відповісти однозначно
4. Цінно
5. Дуже цінно

Розраховуємо в SPSS довірчий інтервал для середнього (ознака 53) (Аналіз → Описові статистики → Розвідницький аналіз). Отримуємо:

Таблиця 6.4

Зведення обробки спостережень

	Спостереження					
	Валідні		Пропущені		Всього	
	N	Відсоток	N	Відсоток	N	Відсоток
53 НАСКІЛЬКИ ЦІНИМ ОСОБИСТО ДЛЯ ВАС Є ЦІКАВА, ТВОРЧА РОБОТА?	3015	98,6%	43	1,4%	3058	100,0%

Таблиця 6.5

Описові статистики

			Статистика	Стд. похибка
53 НАСКІЛЬКИ ЦІНИМ ОСОБИСТО ДЛЯ ВАС Є ЦІКАВА, ТВОРЧА РОБОТА?	Середнє		4,2322	,01534
	95% довірчий інтервал	Нижня границя	4,2021	
	для середнього	Верхня границя	4,2623	
	5% усічене середнє		4,3136	
	Медіана		4,0000	
	Дисперсія		,710	
	Стд. відхилення		,84252	
	Мінімум		1,00	
	Максимум		5,00	
	Розмах		4,00	
	Міжквартильний розмах		1,00	
	Асиметрія		-1,096	,045
	Екссес		1,092	,089

Користуючись формулою розрахунку довірчого інтервалу для середнього, перевіряємо результати розрахунків в SPSS.

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

де σ - стандартне відхилення, $z_{\alpha/2}$ - довірчий коефіцієнт

Довірчий коефіцієнт

- Z=1
- Z=1,65
- **Z=1,96**
- Z=2
- **Z=2,56**
- Z=3

Довірча ймовірність

- P=0,683 или 68,3%
- P=0,90 или 90%
- P=0,95 или 95%**
- P=0,955 или 95,5%
- P=0,99 или 99%**
- P=0,997 или 99,7%

У вибірковій сукупності середнє значення дорівнює 4,23 (цікава, творча робота є цінною). У генеральній же сукупності з довірчою ймовірністю 0,95 середнє знаходиться в інтервалі 2,20-2,26.

$$E = Z \frac{\sigma}{\sqrt{n}}$$

$$E = 1,96 * \frac{0,84}{\sqrt{3015}} = 0,03$$

$$2,23-0,03 < p < 2,23+0,03$$

$$2,20 < p < 2,26$$

Розраховані вручну значення довірчого інтервалу збігаються з розрахунками в SPSS.

3. Як побудувати в SPSS довірчий для відсотку?

Обираємо будь-яку номінальну ознаку з масиву. Наприклад, беремо масив st09.sav, питання «*Поведінка людини у групі пов'язана з виконанням певної ролі. Якою є ваша роль у студентській групі? (виберіть один варіант відповіді)*»:

1. Лідер
2. Активіст
3. Пасивний учасник
4. «Сторонній», незалежний спостерігач
5. «Відринута» людина

Наприклад, нам цікаво дізнатися, скільки активістів у генеральній сукупності. Точкову оцінку даного параметру знайти дуже просто, треба розрахували одновимірний розподіл за ознакою «Роль у студентській групі» та отримати відповідне значення – 55,4%. Аналізована ознака виміряна номінальною шкалою, отже, щоб розрахувати середнє значення та відповідний довірчий інтервал, треба перекодувати змінну, що нас цікавить, тобто створити фіктивну змінну, що має дві альтернативи: 1 – активіст та 0 –

не активіст. Для цього здійснюємо операцію створення нової змінної (команда *Transform (Перетворити)* → *Recode (Перекодувати)*). У результаті отримуємо нову змінну «Активіст» з варіантами 1 – так та 0 – ні. Ця змінна є псевдометричною, для неї можна розраховувати середнє значення, яке має сенс та інтерпретується як частка.

Розраховуємо довірчий інтервал для нової змінної.

Таблиця 6.6

Зведення обробки спостережень

	Спостереження					
	Валідні		Пропущені		Всього	
	N	Відсоток	N	Відсоток	N	Відсоток
Роль у студентській групі	3012	98,5%	46	1,5%	3058	100,0%

Таблиця 6.7

Описові статистики

	Статистика	Стд. похибка
Активіст Середнє	0,5544	0,00906
95% довірчий інтервал для середнього	Нижня границя Верхня границя	0,5367 0,5722
5% усічене середнє	0,5605	
Медіана	1,0000	
Дисперсія	0,247	
Стд. відхилення	0,49711	
Мінімум	0,00	
Максимум	1,00	
Розмах	1,00	
Міжквартильний розмах	1,00	
Асиметрія	-0,219	0,045
Ексцес	-1,953	0,089

У вибірковій сукупності відсоток активістів дорівнює 55,4%. У генеральній же сукупності з довірчою ймовірністю 0,95 відсоток активістів знаходиться в інтервалі між 54% та 57%.

6.7. Статистична перевірка гіпотез

Гіпотеза – це наукове припущення у вигляді висловлювання, істинність або хибність якого невідомі, але можуть бути перевірені емпірично. Гіпотеза

висувається для пояснення якого-небудь явища й потребує емпіричної перевірки. Гіпотезою може бути, наприклад, припущення про взаємозв'язок між незалежною й залежною змінними, що висувається для пояснення яких-небудь явищ і потребує перевірки або припущення про структуру і характер об'єкта, що підлягає вивченню.

У соціологічному дослідженні гіпотези конкретизують мету дослідження, являють собою основний методологічний інструмент, що організує процес дослідження та зумовлює його внутрішню логіку.

Припущення — головний елемент будь-якої гіпотези. Припущення є відповіддю на поставлене запитання про сутність, причину, зв'язки спостережуваного явища. Припущення містить те знання, до якого доходять унаслідок узагальнення фактів. Припущення – це серцевина гіпотези, навколо якої відбувається вся пізнавальна і практична діяльність. Припущення у гіпотезі, з одного боку, є підсумком попереднього пізнання, тобто тим, до чого доходять унаслідок спостереження й узагальнення фактів; з другого боку, воно є відправною крапкою подальшого вивчення явища, визначення напрямку, яким має відбуватися все дослідження. Гіпотеза дає змогу не тільки пояснити наявні факти, а й виявити нові, на які не була б звернута увага, коли не була б висунута ця гіпотеза.

Жодне соціологічне дослідження не може обійтися без висунення гіпотез. Взагалі можна сказати, що головна його мета - це спростування або підтвердження будь-якого припущення дослідника про соціальну дійсність на основі зібраних емпіричних даних. Логіка класичного соціологічного дослідження може бути описана наступним чином: 1) висувається гіпотеза; 2) збираються дані щодо досліджуваного феномену; 3) на основі статистичного аналізу зібраних даних робиться висновок – спростовується чи підтверджується висунута гіпотеза.

Проте ланцюжок гіпотеза-дані-висновок містить в собі низку питань, вирішення яких не завжди є легким завданням, особливо для дослідників, що не мають достатнього досвіду. Найчастіше утруднення полягають у необхідності переведення висунутої гіпотези на мову математичної статистики, іншими словами, виникає питання щодо переформулювання змістовної гіпотези у вигляді статистичної гіпотези, яка здатна поєднати емпіричні дані з дослідницькими припущеннями та обґрунтовано відповісти на питання про те, наскільки ці припущення (гіпотези) відповідають дійсності. Отже, щоб висвітлити це питання насамперед розглянемо, що таке статистична гіпотеза та які є види статистичних гіпотез, а також, у чому полягає перевірка статистичних гіпотез.

Статистичною гіпотезою називається будь-яке припущення щодо виду або параметрів невідомого закону розподілу. У конкретній ситуації статистичну гіпотезу формулюють як припущення на певному рівні статистичної значущості про властивості генеральної сукупності за оцінками вибірки.

Статистичні гіпотези призначені для перевірки спостережуваних величин або подій. Наприклад, середній дохід населення країни за останні 5

років збільшився; студентська молодь Харкова та Львова однаково оцінює важливість цінності самореалізації тощо. Перевірка таких гіпотез здійснюється шляхом зіставлення з результатами спостережень. Проте, як відомо, результати спостережень залежать від випадку. Тому статистичні гіпотези мають ймовірнісний характер, отже, формулюються з урахуванням рівня статистичної значущості.

Формулювання статистичних гіпотез передбачає виділення нульової (H_0) та альтернативної (H_1) гіпотез, поєднання яких і є статистичною гіпотезою. Наприклад, $H_0: \alpha = \beta$ (розбіжностей немає – нульова гіпотеза) і $H_1: \alpha \neq \beta$ (розбіжності є – альтернативна гіпотеза).

Статистична перевірка гіпотез - процедура прийняття рішення, чи слід на основі даних вибіркового дослідження прийняти певне припущення стосовно характеристик (параметрів) генеральної сукупності.

Перевірка гіпотези зводиться до прийняття нульової гіпотези чи її відхилення на користь альтернативної. При цьому нульова гіпотеза (яка постулює відсутність розбіжностей, кореляції тощо) вважається справедливою доти, доки не будуть знайдені факти, що їй суперечать. Якщо не знайдено протиріч, на основі яких нульова гіпотеза повинна бути відхилена, то вона відкрита для подальшої перевірки.

Основні типи статистичних гіпотез:

✓ *Гіпотези щодо невідомого значення параметра розподілу (метод перевірки – побудова довірчих інтервалів).*

✓ *Гіпотези про взаємозв'язок ознак (метод перевірки – кореляційний аналіз).*

✓ *Гіпотези про розбіжності (метод перевірки – тести статистичної значущості розбіжностей).*

Крім того часто виникає потреба у перевірці гіпотез щодо виду розподілу. Така перевірка найчастіше здійснюється задля перевірки того, наскільки емпіричний розподіл досліджуваної ознаки відповідає нормальному розподілу. Необхідність такої перевірки зумовлена тим, що багато статистичних методів можна застосовувати тільки при умові нормальності. Отже, перевірка на нормальність розподілу дозволяє соціологу-аналітику обрати адекватний статистичний метод, який надалі буде застосований для аналізу соціологічних даних та отримання змістовних висновків.

Розглянемо приклади формулювання нульової та альтернативної гіпотез:

Приклад 1. Гіпотеза щодо невідомого значення параметра розподілу.

H_0 : Відсоток мешканців України, кого турбує поширення коронавірусу, знаходиться у межах інтервалу (95%; 85%).

H_1 : Відсоток мешканців України, кого турбує поширення коронавірусу, знаходиться поза межами інтервалу (95%; 85%).

Приклад 2. Гіпотеза про взаємозв'язок ознак.

H_0 : Зв'язок між ознаками «якість життя» та «цінність самореалізації» відсутній ($r=0$).

H_1 : Існує зв'язок між ознаками «якість життя» та «цінність самореалізації» ($r \neq 0$).

Приклад 3. Гіпотези про розбіжності.

H_0 : Середня ефективність аналізованого виду реклами дорівнює нулю.

H_1 : Середня ефективність аналізованої реклами відрізняється від нуля (це свідчить про ефективність реклами)

Приклад 4. Гіпотеза щодо виду розподілу.

H_0 : Розподіл ознаки «матеріальне положення» відповідає закону нормального розподілу.

H_1 : Розподіл ознаки «матеріальне положення» не відповідає закону нормального розподілу.

Нульова (H_0) та альтернативна (H_1) гіпотези приймаються на основі одного й того самого правила, яке називається критерієм нульової гіпотези. Однак умови їх прийняття відрізняються. Альтернативна гіпотеза приймається у випадку, коли нульова не підтверджується та має бути відхилена. Прийняття альтернативної гіпотези відбувається за умови зафіксованого значення ймовірності помилки I роду α , яке обирається із значень 0,1; 0,05; 0,01 та називається рівнем значущості. Нульова гіпотеза приймається у тих випадках, коли її неможливо відхилити.

Таблиця 6.8

Можливі рішення при перевірці гіпотез

У ГЕНЕРАЛЬНІЙ СУКУПНОСТІ	РІШЕННЯ (на основі вибірки):	
	Прийняти H_0	Відхилити H_0
H_0 вірна	<i>Правильне прийняття H_0</i> ($1 - \alpha$) – рівень довіри	<i>Помилка I роду (α)</i> - це ймовірність відкинути вірну гіпотезу (називають також рівень значущості)
H_0 невірна (тобто вірна H_1)	<i>Помилка II роду (β)</i> - це ймовірність прийняття помилкової гіпотези.	<i>Правильне відхилення H_0</i> ($1 - \beta$) – потужність критерію

Прийняття чи відхилення гіпотези здійснюється на основі вибірових даних, що зумовлює існування ймовірності помилки. Ймовірність відхилення гіпотези H_0 , якщо вона вірна, називається *помилкою першого роду* або рівнем значущості і позначається α . Ймовірність прийняття вірної гіпотези називають *рівнем довіри* ($1 - \alpha$). Ймовірність прийняття гіпотези H_0 , якщо вона не вірна, називається *помилкою другого роду* і позначається β .

Ймовірність відхилення невірної основної гіпотези позначається ($1 - \beta$) і називається *потужністю критерію*.

Гіпотезу соціологічного дослідження формулюють у вигляді змістовного твердження, яке може охоплювати велике коло проблем, наприклад: «Відбувається постмодернізація ціннісної свідомості студентства». Перевірку такої гіпотези неможливо здійснити застосовуючи один метод чи сформулювати у вигляді однієї статистичної гіпотези. Для її перевірки формулюють (також у вигляді змістовних тверджень) низку різноманітних робочих гіпотез, кожна з яких згодом необхідно переформулювати у вигляді статистичних гіпотез, перевірка яких дозволить досліднику робити висновки, ґрунтуючись на емпіричному масиві даних, наприклад, є робоча гіпотеза: підвищення рівня якості життя призводить до зростання цінності самореалізації. Вона передбачає необхідність перевірки статистичної гіпотези про взаємозв'язок ознак «якість життя» та «цінність самореалізації», отже, зумовлює потребу змістовного аналізу двовимірного розподіл цих ознак та вивчення кількісних значень коефіцієнтів кореляції. Інший приклад. Робоча гіпотеза – цінність самореалізації зростає протягом 2001-2009 рр. Для її перевірки можна застосовувати аналіз розбіжностей ґрунтуючись на результатах опитувань 2001 та 2009 років, перевірити статистичну гіпотезу щодо статистичної значущості розбіжностей.

Розглянемо ще кілька прикладів формулювання гіпотез.

Приклад №1. Фірма розробила два різних препарати, що дозволяють боротися з важким перебігом захворювання COVID-19 (назвемо їх препарати X і Y) і хоче дізнатися різниться чи ні вплив даних ліків на хворих. З 50 чоловік з важким перебігом захворювання випадково вибираються 20 і випадково ці 20 діляться на дві групи по 10 чоловік. Перша група протягом тижня приймає препарат X, друга - препарат Y. Потім у всіх хворих фіксується вміст кисню в крові. Висунута змістовна гіпотеза: препарати X і Y по-різному впливають на вміст кисню в крові хворого COVID-19.

Приклад №2. Дослідник хоче дізнатися, як впливає тривалість практичних занять на успішність студентів. Припустимо, він обрав такий шлях: з 200 студентів випадково вибрав 50 осіб і протягом місяця спостерігав за їх успішністю. Далі він збільшив тривалість практичних занять на 10 хвилин і протягом наступного місяця дивився на успішність все тих же 50 студентів. Потім він порівняв результати кожного студента до і після збільшення тривалості практичних занять. Висунута змістовна гіпотеза: тривалість практичного заняття впливає на успішність студента.

Приклад №3. З 200 студенток випадково було обрано 80 осіб, і ці 80 осіб розділили на дві групи по 40 у кожній. Одній групі задавали питання без установки: «Скільки ви готові заплатити за натуральний шампунь?», а другій групі задавали питання з установкою: «Скільки ви готові заплатити за натуральний шампунь, якщо відомо, що люди, які користуються натуральними шампунями, менше на 10-15% страждають від ламкості волосся?» Дослідник припускав, що позитивна інформація про продукт, що міститься у другому питанні, вплине на респондента, і люди, що

відповідають на питання з установкою, будуть готові заплатити за шампунь більше, ніж ті, яким було запропоновано питання без установки. Висунута змістовна гіпотеза: постановка питання впливає на відповідь респондента.

Перед нами три приклади, кожен з яких демонструє формулювання змістовної гіпотези. Тепер перетворимо наші змістовні гіпотези в статистичні: сформулюємо нульову і альтернативну гіпотези. Крім того розглянемо у чому полягає завдання дослідника та висновки, які можна отримати на основі перевірки статистичних гіпотез у кожному з наших трьох прикладів.

Нагадуємо зміст позначень:

\bar{x} - середнє арифметичне значення змінної, обчислене за вибіркою;

μ - середнє арифметичне значення змінної у генеральній сукупності.

Таблиця 6.9

Приклади формулювання гіпотез

***	Приклад №1	Приклад №2	Приклад №3
Змістовна гіпотеза	Препарати X і Y по-різному впливають на вміст кисню в крові у хворих COVID-19 з важким перебігом захворювання	Тривалість практичного заняття впливає на успішність студентів	Постановка питання впливає на відповідь респондента
Завдання дослідника	Порахувати середній вміст кисню в крові в першій і в другій групах хворих після тижневого використання досліджуваних препаратів, відповідно \bar{x}_1 і \bar{x}_2	1. Порахувати середній бал студента до збільшення тривалості практичних занять 2. Порахувати середній бал студента після збільшення тривалості практичних занять 3. Порахувати для кожного студента різницю між середніми до і після 4. Знайти середнє арифметичне різниць для всіх студентів, що позначається Dх	Порахувати, скільки в середньому готові заплатити за натуральний шампунь респонденти, які відповідають на питання з установкою і на питання без установки, відповідно \bar{x}_1 і \bar{x}_2
Нульова гіпотеза	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_D = 0$	$H_0 : \mu_1 - \mu_2 = 0$

Таблиця 6.9 (продовження)

	Приклад №1	Приклад №2	Приклад №3
Сенс нульової гіпотези	μ_1 і μ_2 - середні генеральних сукупностей, з яких взято вибірки з середніми \bar{x}_1 і \bar{x}_2 . Нульова гіпотеза говорить про те, що вплив обох ліків на вміст кисню в крові в середньому незначний, і якщо навіть вибіркові середні нерівні, то це пояснюється лише похибкою вибірки або іншими незалежними від нас причинами	μ_D - середнє різниць для студентів в генеральній сукупності. Нульова гіпотеза говорить про те, що насправді немає різниці між середнім балом студента до і після збільшення тривалості практичного заняття, і якщо навіть вибіркове середнє різниць відмінно від нуля, то це пояснюється лише похибкою вибірки або іншими незалежними від нас причинами	Оскільки H_0 збігається з H_0 в прикладі №1, то пояснення можна знайти в першій колонці (див. приклад 1)
Альтернативна гіпотеза	$H_1: \mu_1 - \mu_2 \neq 0$	$H_1: \mu_D \neq 0$	$H_1: \mu_1 - \mu_2 \neq 0$
Висновок щодо змістовної гіпотези	Якщо ми приймаємо нульову гіпотезу - препарати надають однаковий вплив (різниці між середніми немає), то ми відкидаємо змістовну гіпотезу, в іншому випадку - ми приймаємо змістовну гіпотезу	Якщо ми приймаємо нульову гіпотезу - тривалість практичного заняття не впливає на успішність, то ми відкидаємо змістовну гіпотезу і навпаки	Якщо ми приймаємо нульову гіпотезу - питання не впливає на вибір респондента, то ми відкидаємо змістовну гіпотезу і навпаки

Для того щоб перевірити гіпотезу необхідно здійснити наступне:

1. Перетворити змістовну гіпотезу у статистичну: сформулювати нульову (H_0) і альтернативну (H_1) гіпотези.
2. Вибрати метод та критерій перевірки відповідно до змісту гіпотез і наявних статистичних даних.
3. Вибрати рівень значущості, який контролює допустиму ймовірність помилки першого роду.
4. Порахувати значення критерію і порівняти його з критичним.
5. Відкинути або прийняти нульову гіпотезу.
6. Змістовно проінтерпретувати отримані результати.

Література до теми

Горбачик, А.П., Сальникова, С.А. *Аналіз даних соціологічних досліджень засобами SPSS. Навчально-методичний посібник*. Луцьк: «Вежа», 2008. С. 129-145.

Иванов, О.В. *Статистика: Учебный курс для социологов и менеджеров. Часть 2. Доверительные интервалы. Проверка гипотез. Методы и их применение*. М. 2005. С. 3-25.

Крыштановский, А.О. *Анализ социологических данных с помощью пакета SPSS*. М.: ГУ ВШЭ, 2007. С. 37-38.

Паніотто, В.І., Максименко, В.С., Харченко, Н.М. *Статистичний аналіз соціологічних даних*. К.: «КМ Академія», 2004. С. 177-217.

Толстова, Ю.Н. *Математико-статистические модели в социологии (математическая статистика для социологов) : учеб. Пособие*. М.: ГУ-ВШЭ, 2008. С. 71-95.

Додаткова література

Бююль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005.

Донченко, В. С., Сидоров, М. В., Шарапов, М. М. *Теорія ймовірностей та математична статистика. Навчальний посібник*. К.: Академія, 2009.

Руденко В. М. *Математична статистика. Навч. посіб.* К.: Центр учбової літератури, 2012.

Толстова, Ю. Н. *Анализ социологических данных (Методология, дескриптивная статистика, изучение связей между номинальными признаками)*. М.: Научный мир, 2003.

Толстова, Ю. Н. Проверка статистических гипотез. *Социологический словарь*. М., 2014. URL: <http://ponjatija.ru/node/9029>.

Питання для самоконтролю

1. Яке спостереження має назву вибіркового?
2. Чому при вибіркового спостереженні присутні помилки?
3. Чим відрізняється параметр від статистики?
4. Що таке точкова оцінка параметрів генеральної сукупності?
5. Що таке інтервальна оцінка параметрів генеральної сукупності?
6. Які вибірки вважають малими?
7. У результаті проведеного опитування 800 жителів міста з'ясувалося, що 54% опитаних планують підтримати на майбутніх виборах кандидатуру чинного президента. При цьому відомо, що статистична похибка вибірки становить 5%. Чи можна на цій підставі прогнозувати, що діючий президент буде переобраний на наступний термін?
8. Після проведення рекламної кампанії було проведено опитування для виявлення ефективності реклами. В результаті виявилось, що рекламу продукту А запам'ятали 55% опитаних, а рекламу продукту Б - 51%. Помилка вибірки складає 2%. Чи можна стверджувати, що реклама

продукту Б більше запам'яталася споживачам, ніж реклама продукту А?

Тема 7. Аналіз розбіжностей

7.1. Аналіз розбіжностей та статистична значущість розбіжностей

Аналіз розбіжностей застосовується для порівняння результатів дослідження двох (або більшої кількості) груп респондентів з метою визначення ступеня реальної відмінності в їхніх перевагах чи поведінці, наприклад, ціннісних орієнтаціях, електоральних настроях, споживчій поведінці тощо. Цей метод є головним інструментом аналізу результатів кількісних порівняльних досліджень. Також він застосовується для аналізу динаміки показників, які вимірюються двічі (чи більше) на одній і тій же вибірці через певний проміжок часу.

Метою аналізу розбіжностей є виявлення та подальше дослідження специфічних характеристик заданих груп респондентів, які статистично значущо відрізняються між собою. Статистична значущість розбіжностей розраховується з метою оцінки відмінності відсотків, середніх значень, дисперсій, коефіцієнтів кореляції тощо. Проте у соціологічних дослідженнях найчастіше виникає потреба порівняння відсотків та середніх, на чому ми й зосередимо увагу.

У практиці соціологічних досліджень досить часто зустрічаються ситуації, коли в ході попереднього аналізу (на підставі досвіду дослідника або статистичного аналізу масиву емпіричних даних, отриманих в результаті вибіркового опитування) висувається гіпотеза щодо поділу всієї вибіркової сукупності на певні групи, які значно розрізняються за своїми характеристиками. Порівняння одномірних розподілів (для якісних ознак) або середніх значень (для кількісних ознак) може показувати, що респонденти з виділених груп розрізняються. Проте виявлені відмінності стосуються лише вибіркової сукупності. Для того, щоб з упевненістю констатувати наявність розходжень у генеральній сукупності необхідно переконатися, що знайдені відмінності не зумовлені випадковими факторами, тобто необхідно розрахувати їхню статистичну значущість. Отже, **аналіз розходжень передбачає дві процедури: 1) виявлення розбіжностей; 2) оцінку їхньої статистичної значущості.**

Виявлення розбіжностей полягає у зіставленні відповідей досліджуваних груп респондентів на одне і теж запитання. Кількісними показниками, що дозволяють здійснити таке порівняння, є відсотки або середні значення (в залежності від типу шкали досліджуваної ознаки).

Статистична значущість розбіжностей показує, чи дійсно знайдені розходження існують у генеральній сукупності. Ймовірність істинності зроблених висновків не повинна бути менша за 0,95, тобто ймовірність помилки (p – рівень значущості) має бути менше 0,05 ($p \leq 0,05$).

Схема аналізу розбіжностей:

✓ *Висуваємо гіпотезу про відмінності двох або більше груп. Наприклад, сімейне благополуччя є більшою цінністю для жінок, ніж для чоловіків.*

✓ *Виявляємо розбіжності, тобто розраховуємо відповідні статистики одномірних розподілів за аналізованою ознакою (наприклад, середні значення цінності сімейного благополуччя в кожній групі, тобто серед жінок і серед чоловіків).*

✓ *Проводимо статистичну оцінку значущості виявлених розбіжностей. Наприклад, розраховуємо статистичну значущість розбіжностей відсотків або середніх (залежно від типу шкали аналізованої ознаки).*

✓ *Змістовно інтерпретуємо виявлені відмінності (якщо вони є статистично значущими). Якщо розбіжності виявилися статистично незначущими, то інтерпретувати їх не має сенсу.*

Зверніть увагу! Перед інтерпретацією розбіжностей **обов'язково** слід перевірити їхню статистичну значущість. Не можна описувати та інтерпретувати розбіжності не перевіривши їх статистичну значущість, оскільки це може призвести до помилки першого роду (див. тему «Перевірка статистичних гіпотез»).

Статистична значущість розбіжностей – кількісний показник вірогідності того, що знайдені на основі вибіркового дослідження відмінності правильно віддзеркалюють розбіжності у генеральній сукупності.

Розрахунок статистичної значущості розбіжностей – трудомістка процедура, яку зазвичай виконують застосовуючи відповідне програмне забезпечення, наприклад, SPSS.

Для розрахунку статистичної значущості розбіжностей існує велика кількість тестів, серед яких аналітик повинен обрати адекватний досліджуванім даним. Вибір методу перевірки статистичної значущості розходжень визначається **типом шкали аналізованої ознаки, обсягами вибірок, формою її розподілу** (нормальністю або ненормальністю), **залежністю/незалежністю вибірок** та **кількістю вибірок** (одна, дві чи більше).

Тип шкали визначає, що саме соціолог буде порівнювати: відсотки чи середні значення.

Обсяги вибірок визначає вибір критерію: t-критерій чи z-критерій. Якщо вибірки достатньо великі, то застосовують z-статистики, якщо малі, то застосовують t-критерій (як було показано у попередньому розділі).

Форма розподілу аналізованої ознаки, тобто відповідність чи невідповідність нормальному розподілу обумовлює застосування **параметричних чи непараметричних методів (тестів, критеріїв)**. Параметричні методи ґрунтуються на відомому вигляді розподілу генеральної сукупності (як правило, нормальному), використовує параметри цієї сукупності (середні, дисперсії тощо). Критерій розбіжності називають непараметричним, якщо він не базується на припущенні щодо виду розподілу генеральної сукупності та не використовує параметри цієї сукупності.

Методи перевірки форми розподілу:

- візуальна: *Analyze (Аналіз) → Descriptive Statistics (Дескриптивні статистики) → Frequencies... (Частоти) → Charts*

(Діаграми) → Histograms (Гістограми) → With normal curve (З кривою нормального розподілу). Інтерпретація: візуальне співвідношення форми гістограми з кривою нормального розподілу.

- тест Колмогорова-Смірнова для перевірки статистичної гіпотези щодо форми розподілу: *Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → 1-Sample KS (З однієї вибірки)*.

H_0 : Емпіричний розподіл не відрізняється від нормального.

H_1 : Емпіричний розподіл відрізняється від нормального.

Інтерпретація: якщо *Asymp. Sig. (2-tailed) $\leq 0,05$* , то гіпотеза H_0 відхиляється, розподіл не є нормальним. Для таких змінних варто застосовувати непараметричні тести. Якщо *Asymp. Sig. (2-tailed) $> 0,05$* , то гіпотеза H_0 не може бути відхилена, розподіл є нормальним. Для таких змінних застосовують параметричні тести.

- тест Шапіро-Уїлка для перевірки форми розподілу: *Analyze (Аналіз) → Descriptive Statistics (Дескриптивні статистики) → Explore... (Досліджувати) → Plots (Графіки) → Normality plots with tests (Графіки та критерії для перевірки нормальності)*.

H_0 : Емпіричний розподіл не відрізняється від нормального.

H_1 : Емпіричний розподіл відрізняється від нормального.

Інтерпретація: якщо *Asymp. Sig. (2-tailed) $\leq 0,05$* , то гіпотеза H_0 відхиляється, розподіл не є нормальним. Для таких змінних варто застосовувати непараметричні тести. Якщо *Asymp. Sig. (2-tailed) $> 0,05$* , то гіпотеза H_0 не може бути відхилена, розподіл є нормальним. Для таких змінних застосовують параметричні тести.

Залежність або незалежність досліджуваних вибірок впливає на вибір методу перевірки статистичної значущості розбіжностей, оскільки у статистиці для залежних та незалежних вибірок розроблені окремі методи. Це зумовлює необхідність розглянути сутність даних понять. Вибірки називаються залежними (пов'язаними або парними), якщо можна встановити гомоморфізм, тобто відповідність, коли одному випадку з вибірки X відповідає один і тільки один випадок з вибірки Y і навпаки. Приклади залежних вибірок: пари близнюків; чоловіки й дружини; два виміри якої-небудь ознаки до й після експериментального впливу. Залежні вибірки характеризуються кореляцією думок респондентів, які утворюють пару. У випадку відсутності взаємозв'язку між вибірками вони вважаються незалежними. Саме незалежні вибірки найчастіше досліджуються соціологами.

Кількість вибірок також є чинником вибору методу перевірки статистичної значущості розбіжностей, оскільки для аналізу двох та більшої кількості вибірок існують різні методи. Коли виникає необхідність порівняти одночасно більше за дві вибірки, то застосовують дисперсійний аналіз чи його непараметричні аналоги: ранговий однофакторний аналіз Краскела-Уолліса, ранговий критерій Фрідмана тощо.

7.2. Т-тести

Критерій розбіжності називають *параметричним*, якщо він ґрунтується на наявному вигляді розподілу генеральної сукупності (як правило, нормальному) або використовує параметри цієї сукупності (середні, дисперсії тощо). Найчастіше використовуються t-тести (у випадку двох вибірок) та однофакторний дисперсійний аналіз (у випадку більше двох вибірок).

Таблиця 7.1

Параметричні тести для перевірки значущості розбіжностей

	Незалежні вибірки	Залежні вибірки
Дві вибірки	<p>t-тест для незалежних вибірок (тест Стьюдента)</p> <p>Виклик процедури: <i>Analyze</i> → <i>Compare Means</i> → <i>Independent Samples T Test</i>.</p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>	<p>t-тест для залежних вибірок</p> <p>Виклик процедури: <i>Analyze</i> → <i>Compare Means</i> → <i>Paired-Samples T Test</i>.</p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>
Більше двох вибірок	<p>Однофакторний дисперсійний аналіз</p> <p>Виклик процедури: <i>Analyze</i> → <i>Compare Means</i> → <i>One-Way ANOVA</i>).</p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>	<p>Однофакторний дисперсійний аналіз з повторними вимірами</p> <p>Виклик процедури: <i>Analyze (Аналіз)</i> → <i>General Linear Model (Загальна лінійна модель)</i> → <i>Repeated Measures... (Повторні виміри)</i>.</p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>

Т-тести для двох незалежних вибірок

Т-тести призначені для встановлення розбіжностей між середніми значеннями досліджуваного показника у двох групах респондентів. Досліджуваний показник – незалежна змінна, яка має бути кількісною (метричною чи інтервальною). Незалежна змінна – категоріальна, що поділяє масив на дві групи.

Нульова (H_0) та альтернативна (H_1) гіпотези t тесту незалежних вибірок можуть бути виражені у такий спосіб:

$H_0: \mu_1 = \mu_2$ (середні значення у двох групах рівні).

$H_1: \mu_1 \neq \mu_2$ (середні значення у двох групах не рівні, відрізняються між собою).

Розглянемо застосування t -тесту для двох незалежних вибірок на прикладі перевірки гіпотези, що дівчата гірше ставляться до такого явища, як аборт, ніж юнаки. Тобто оцінимо статистичну значущість розбіжності у ставленні до абортів серед юнаків та дівчат (масив st06.sav, стать – ознака p204, ставлення до абортів – ознака p98).

Для розрахунку статистичної значущості розбіжностей необхідно виконати команду: *Analyze (Аналіз) → Compare Means (Порівняння середніх) → Independent-Samples T Test (t-тесту для незалежних вибірок*. У результаті на екрані з'явиться діалогове вікно *Independent-Samples T Test (t-тест для незалежних вибірок)*, в якому треба задати: змінну чи змінні, що підлягають аналізу; змінну, за якою виділяються групи, що будуть порівнюватися та аналізуватися; довірчу ймовірність, відповідно до якої будуть розраховуватися довірчі інтервали (рис. 7.1).

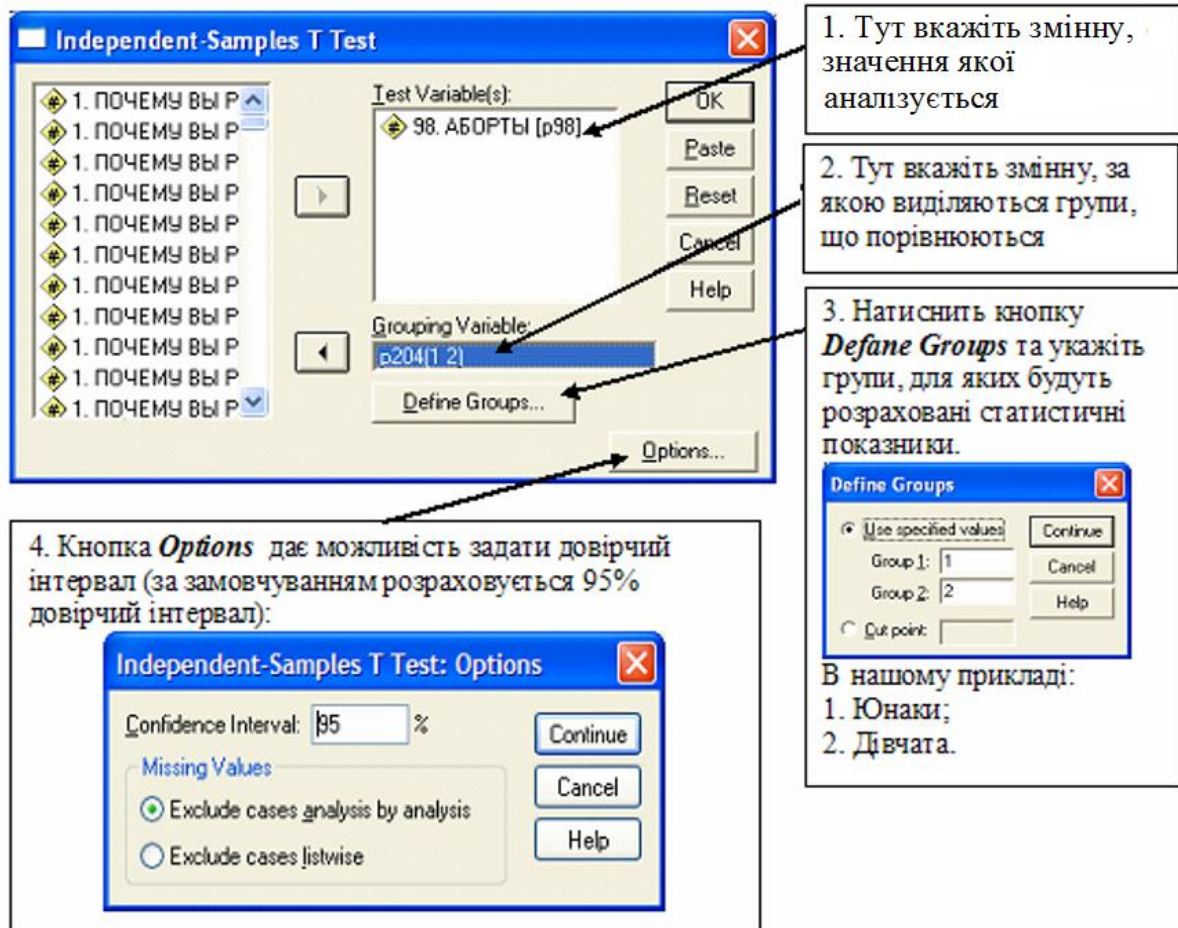


Рис. 7.1. Діалогове вікно Independent-Samples T Test (t-тест для незалежних вибірок)

У результаті виконання процедури розрахунку t-тесту для незалежних вибірок SPSS виведе дві таблиці:

- 1) **Group Statistics (Статистики груп)**, де представлено кількість спостережень, середні значення, стандартні відхилення й стандартні помилки середніх в обох групах (табл. 7.2);
- 2) **Independent Samples Test (Т-тест для незалежних вибірок)**, що містить кількісні показники, які дозволяють встановити статистичну значущість розбіжності між середніми значеннями (табл. 7.3).

Таблиця 7.2

Group Statistics (Статистики груп за ознакою «98. Ставлення до абортів»)

204. Стать	N (кількість спостережень)	Mean (середнє значення)	Std. Deviation (стандартне відхилення)	Std. Error Mean (стандартна помилка середніх)
Чол.	1305	2,66	1,288466	0,035667
Жін.	1661	2,60	1,172269	0,028764

При інтерпретації результатів таблиці 7.3 необхідно мати на увазі, що t-тест Стьюдента розраховується за різними формулами в залежності від того відрізняються чи ні дисперсії досліджуваної змінної у аналізованих групах.

Якщо передбачаються рівні дисперсії, при розрахунку t-тесту використовуються об'єднані відхилення; коли дисперсії не є рівними, обчислення ґрунтується на необ'єднаних відхиленнях та враховує поправку на ступінь свободи.

Однорідність дисперсій перевіряється за допомогою тесту Левена, результати якого SPSS виводить поряд з результатами тесту Стьюдента.

Таблиця 7.3

Independent Samples Test (Т-тест для незалежних вибірок за ознакою «98. Ставлення до абортів»)

	Levene's Test for Equality of Variances (тест Левена на рівність дисперсій)		t-test for Equality of Means (тест Стьюдента на рівність середніх)						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed (дисперсії рівні)	18,937	0,000	1,00	2964,00	0,317	0,045	0,045	-0,044	0,134
Equal variances not assumed (дисперсії не рівні)			0,99	2665,91	0,323	0,045	0,046	-0,045	0,135

Якщо тест Левена показує, що дисперсії не відрізняються (тобто значення Sig більше за 0,05), то потрібно дивитися на значення тесту Стьюдента Sig. (2-tailed) у рядку «*Equal variances assumed*» (*Рівність дисперсій виявлена*).

Якщо тест Левена показує, що дисперсії відрізняються (Sig менш ніж 0,05), то необхідно дивитися на значення Sig. (2-tailed) у рядку «*Equal variances not assumed*» (*Рівність дисперсій не виявлена*).

В нашому прикладі тест Левена виявив, що дисперсії рівні, отже, статистичну значущість тесту Стьюдента на рівність середніх дивимось у другому рядку: Sig. (2-tailed) = 0,323 > 0,05, що свідчить про статистичну незначущість розбіжностей у ставленні до абортів серед юнаків та дівчат.

Т-тести для двох залежних (пов'язаних) вибірок

Дослідження залежних (парних, пов'язаних) вибірок не є поширеним завданням при аналізі результатів масових соціологічних опитувань.

Тест парних вибірок зазвичай використовується для тестування наступного:

- ✓ статистичної різниці між двома часовими точками; статистичної різниці між двома умовами;
- ✓ статистичної різниці між двома вимірюваннями, наприклад, зробленими до та після рекламного впливу;
- ✓ статистичної різниці між узгодженою парою.

Нульова (H_0) та альтернативна (H_1) гіпотези t -тесту для двох залежних вибірок можуть бути виражені у такий спосіб:

$H_0 : \mu_1 = \mu_2$ (середні значення у двох групах рівні).

$H_1 : \mu_1 \neq \mu_2$ (середні значення у двох групах не рівні).

Т-тести для залежних вибірок інколи використовують замість аналізу кореляційних таблиць.

У якості прикладу розглянемо, яким чином можна перевірити гіпотезу: студентська молодь цінує матеріальний добробут вище, ніж цікаву, творчу роботу. Емпіричною базою слугуватиме масив st09.sav, ознаки «53. Цінність цікавої, творчої роботи» та «54. Цінність матеріального добробуту».

Для розрахунку t -тесту для двох (залежних, парних) вибірок у SPSS треба виконати команду *Analyze (Аналіз) → Compare Means (Порівняння середніх) → Paired-Samples T Test (t-тест для залежних вибірок)* (див. рис. 7.2), після чого у діалоговому вікні *Paired-Samples T Test* необхідно вказати аналізовані змінні та задати довірчу ймовірність (рис. 7.3).

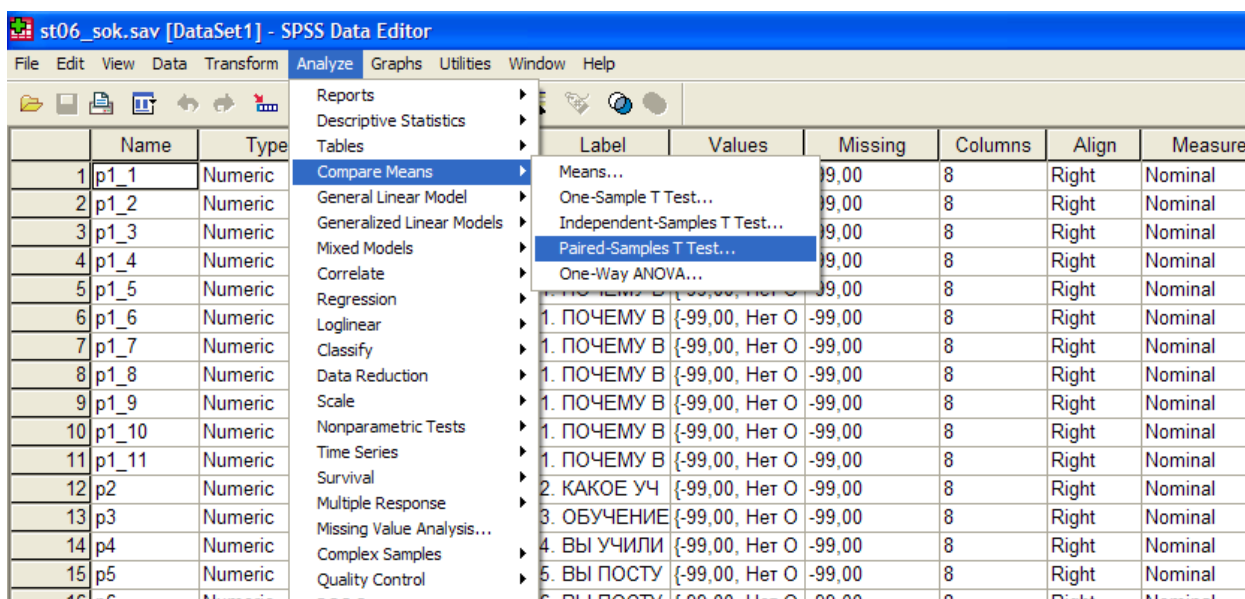


Рис. 7.2. Розрахунок Paired-Samples T Test (t-тест для залежних вибірок)

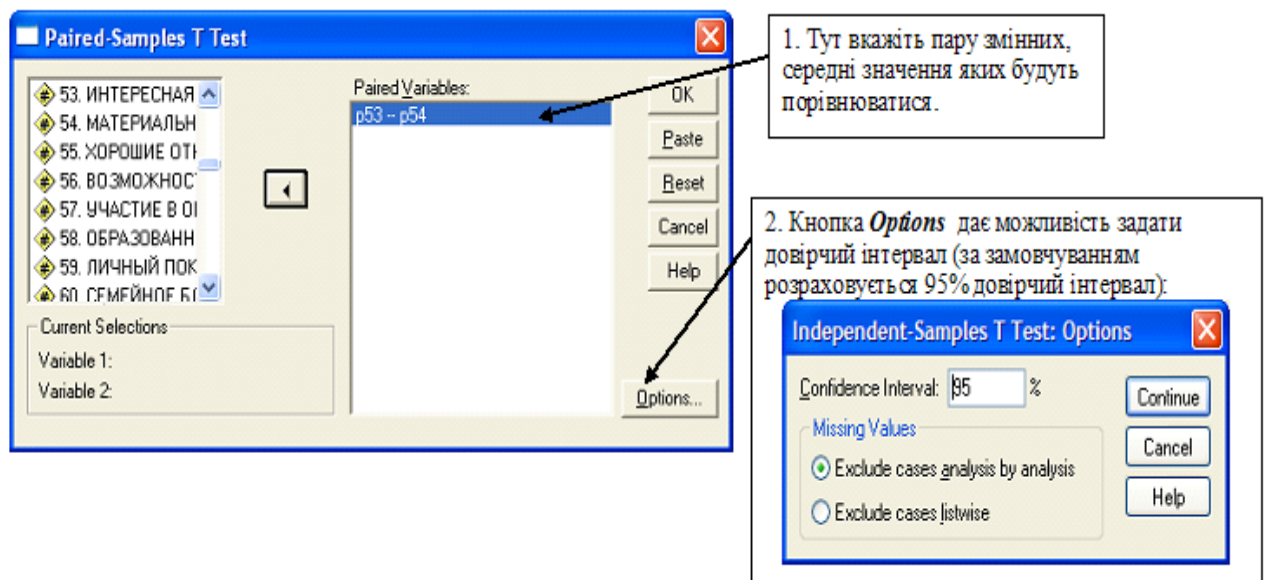


Рис. 7.3. Діалогове вікно *Paired-Samples T Test* (*t*-тест для залежних вибірок)

У результаті виконання процедури розрахунку *t*-тесту для залежних вибірок SPSS презентує три таблиці, перші дві з яких містять допоміжні результати, остання – головну інформацію:

- ***Paired Samples Statistics* (Статистики для парних вибірок)**, у якій наведено середні значення досліджуваних ознак (Mean), кількість валідних спостережень (N), стандартне відхилення (Std. Deviation) та стандартна похибка вибіркового середнього (Std. Error Mean) (див. табл. 7.4);
- ***Paired Samples Correlations* (Кореляції для парних вибірок)**, що містить коефіцієнт кореляції між змінними (Correlation) та його статистичну значущість (Sig) (див. табл. 7.5);
- ***Paired Samples Test* (Тест для парних вибірок)**, у якій представлено: різницю середніх значень (Mean), відповідне стандартне відхилення (Std. Deviation), стандартна похибка (Std. Error Mean), довірчий інтервал для заданої довірчої ймовірності та головний результат – кількісний показник статистичної значущої розбіжності у стопці з назвою «Sig. (2-tailed)», значення якого дають можливість зробити висновок щодо наявності чи відсутності розбіжностей у відповідях респондентів на два різні запитання (див. табл. 7.6)

Таблиця 7.4

Paired Samples Statistics (Статистики для парних вибірок)

		Mean	N	Std. Deviation	Std. Error Mean
Pair	53. Цікава, творча робота	3,285	2947	,72642	,01338
	54. Матеріальний добробут	3,514	2947	,61386	,01131

Таблиця 7.5

Paired Samples Correlations (Кореляції для парних вибірок)

		N	Correlation	Sig.
Pair	53. Цікава, творча робота 54. Матеріальний добробут	2947	,097	,000

Таблиця 7.6

Paired Samples Test (Тест для парних вибірок)

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
53. Цікава, творча робота 54. Матеріальний добробут	-0,229	0,905	0.0167	-0.262	-0,197	-13,765	2946	,000

У цьому прикладі розбіжності виявилися статистично значущими, проте незначними (середні значення 3,285 проти 3,514), що також підтверджує низьке значення коефіцієнта кореляції (0,097).

Т-тести для однієї вибірки

У соціологічних дослідженнях одновибіркові t-тести зазвичай використовуються, щоб з'ясувати таке:

✓ Чи існує статистично значуща різниця між вибірковим середнім та відомим (або гіпотетичним) значенням середнього генеральної сукупності?

✓ Чи існує статистично значуща різниця між вибірковим середнім певної групи респондентів та відомим значенням середнього даної групи у генеральній сукупності?

✓ Чи репрезентативна вибірка? (Тут порівнюють вибіркові середні з даними, наприклад, Державної служби статистики України).

Одновибіркові t-тести можуть порівнювати лише одне середнє значення вибірки з заданою константою, яка розглядається як еталон чи бажане значення. Ці тести не призначені порівнювати вибіркові середні між кількома групами.

Нульова гіпотеза (H_0) та альтернативна гіпотеза (H_1) одного зразка T-тесту може бути виражена як:

$H_0: \mu = x$ (вибіркове середнє дорівнює пропонованому еталонному значенню).

$H_1: \mu \neq x$ (вибіркове середнє не дорівнює пропонованому еталону)

де μ - еталонна константа, запропонована для середнього генеральної сукупності, а x - середнє значення, розраховане за вибіркою.

T-тести для однієї вибірки в SPSS розраховують за допомогою команди *Analyze (Аналіз) → Compare Means (Порівняння середніх) → One-Sample T Test... (t-тест для однієї вибірки)*, у результаті виконання якої відкриється діалогове вікно *One-Sample T Test*, де необхідно задати:

- ✓ Змінну, середнє значення якої буде порівнюватися з деяким еталонном;
- ✓ Еталонне значення;
- ✓ Довірчий інтервал (див. рис. 7.4).

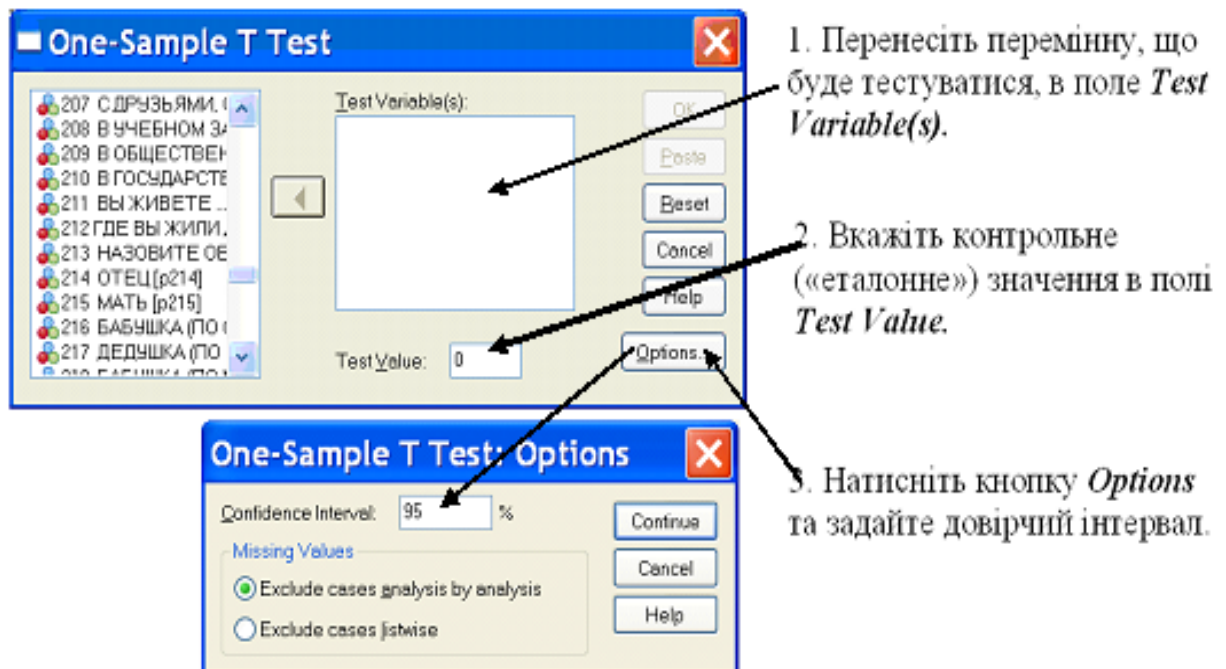


Рис. 7.4. Діалогове вікно One-Sample T Test (t-тест для однієї вибірки)

7.3. Дисперсійний аналіз

Дисперсійний аналіз (англ. ANalysis Of VAriance – ANOVA) – метод статистичного аналізу, призначений для дослідження впливу однієї або кількох якісних (номінальних або порядкових) змінних на залежну кількісну змінну (метричну або інтервальну). Незалежні змінні інтерпретуються як фактори, що за гіпотезою дослідника зумовлюють коливання залежної змінної. Ці фактори відбивають групову приналежність і можуть мати більше двох градацій.

Головна мета дисперсійного аналізу – дослідити статистичну значущість розбіжностей між середніми значеннями залежної кількісної змінної по групах фактору. Досягається це шляхом розкладання загальної дисперсії залежної змінної на складові: міжгрупову і внутрішньогрупову дисперсії. Аналіз цих компонентів дисперсії дає можливість дослідити частку впливу кожного фактору на залежну змінну. Дисперсійний аналіз застосовується для порівняння статистичної значущості розбіжностей середніх в двох та більше групах. Якщо за допомогою дисперсійного аналізу порівнювати дві групи, його результати збігаються з результатами звичайного t-критерію.

Дисперсійний аналіз поділяють на однофакторний та багатфакторний (див. рис. 7.5). У випадку однофакторного дисперсійного аналізу вивчається наявність чи відсутність впливу на результуючий показник одного якісного фактору. Відповідно, вирішуються два завдання: 1) загальна оцінка статистичної значущості розбіжностей між груповими середніми; 2) змістовна інтерпретація тих розбіжностей, які виявилися статистично значущими.

Статистична гіпотеза ANOVA може бути сформульована наступним чином:

$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ (середні значення у всіх аналізованих групах рівні);

H_1 : принаймні одне середнє значення μ_i відрізняється від інших, де k – кількість досліджуваних груп, μ_i – середнє значення аналізованої ознаки у групі під номером i ($i = 1, 2, \dots, k$).

Багатфакторний дисперсійний аналіз призначений для вивчення впливу кількох незалежних змінних (факторів) на залежну змінну. Відмінною особливістю багатфакторного дисперсійного аналізу є можливість оцінити не лише вплив кожної незалежної змінної окремо, але також їх взаємодію - залежність впливу одних факторів від значень інших факторів. Так, наприклад, в результаті двофакторного дисперсійного аналізу отримують вплив першого та другого факторів, а крім того їх взаємовплив. Можливість виявити ефекти взаємодії між факторами та, відповідно, перевіряти більш складні гіпотези є вагомою перевагою багатфакторного дисперсійного аналізу.

Види дисперсійного аналізу

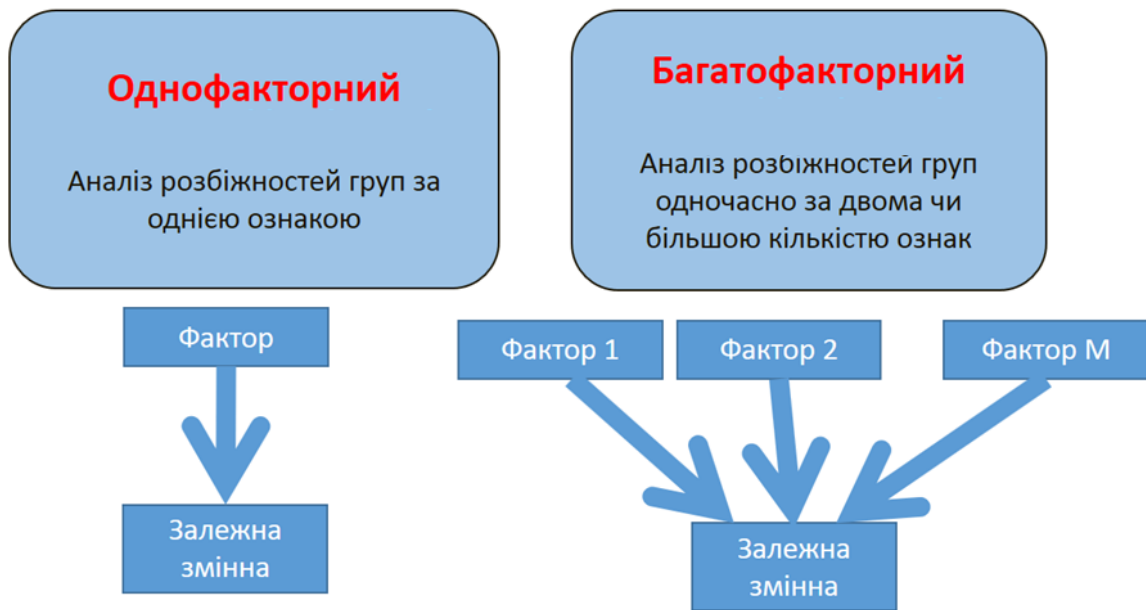


Рис. 7.5. Види дисперсійного аналізу

Дисперсійний аналіз відноситься до параметричних методів, що зумовлює необхідність контролювати відповідність розподілу досліджуваної (залежної) змінної нормальному закону. У випадках, коли розподіл не є нормальним, замість F-критерію (критерію Фішера) застосовують його непараметричні альтернативи, зокрема, критерій Крускала-Уолліса або критерій Фрідмана.

Приклад застосування однофакторного дисперсійного аналізу

Завдання. Перевірити гіпотезу щодо наявності розходжень у ціннісних орієнтаціях студентів в залежності від профілю навчання. Емпірична база – результати суцільного опитування студентів 5 курсу ХНУ імені В.Н.Каразіна (дослідження 2013р., файл 5kurs.sav).

Розглянемо, які дії потрібні для вирішення цього завдання.

Крок 1. Чи дозволяють наявні дані застосовувати середні значення?

Розглянемо які дані є для аналізу. В інструментарій закладено блок запитань стосовно ціннісних пріоритетів. Респондентів запитували «Наскільки цінними особисто для Вас є ... ?» (нижче були перелічені 18 цінностей, які респонденти мали оцінити за наступною шкалою: 5 – дуже цінно; 4 – скоріше цінно; 3 – частково цінно, частково ні; 2 – скоріше не цінно; 1 – зовсім не цінно; 0 – важко відповісти. Можна побачити, що альтернатива відповіді 0 (важко відповісти) спотворює шкалу, яка без неї мала би бути порядковою. Отже, якщо ми хочемо застосувати середні значення, необхідно перетворити шкали шляхом перекодування 0 як пропущеного значення.

Крок 2. Розрахунок та візуалізація середніх значень. Розрахуємо середні значення у групах в залежності від профілю навчання (рис. 7.6).

Розрахуйте середні значення. Незалежна змінна – профіль навчання, залежні змінні – ціннісні орієнтації.

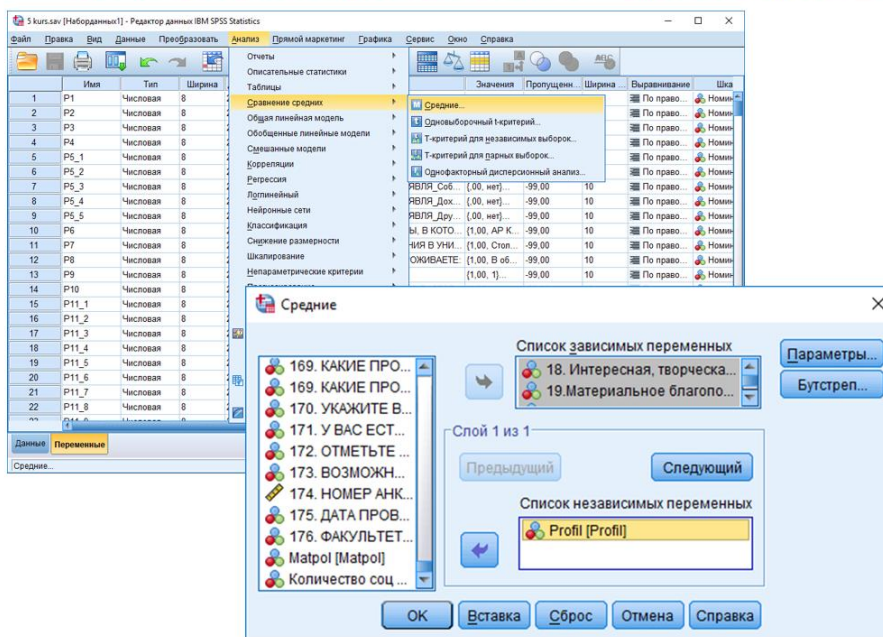


Рис. 7.6. Розрахунок середніх значень у групах в залежності від профілю навчання

У результаті отримаємо таблицю, яка є недостатньо наочною, отже, візуалізуємо результати, наприклад, у вигляді пелюсткової діаграми (рис. 7.7).

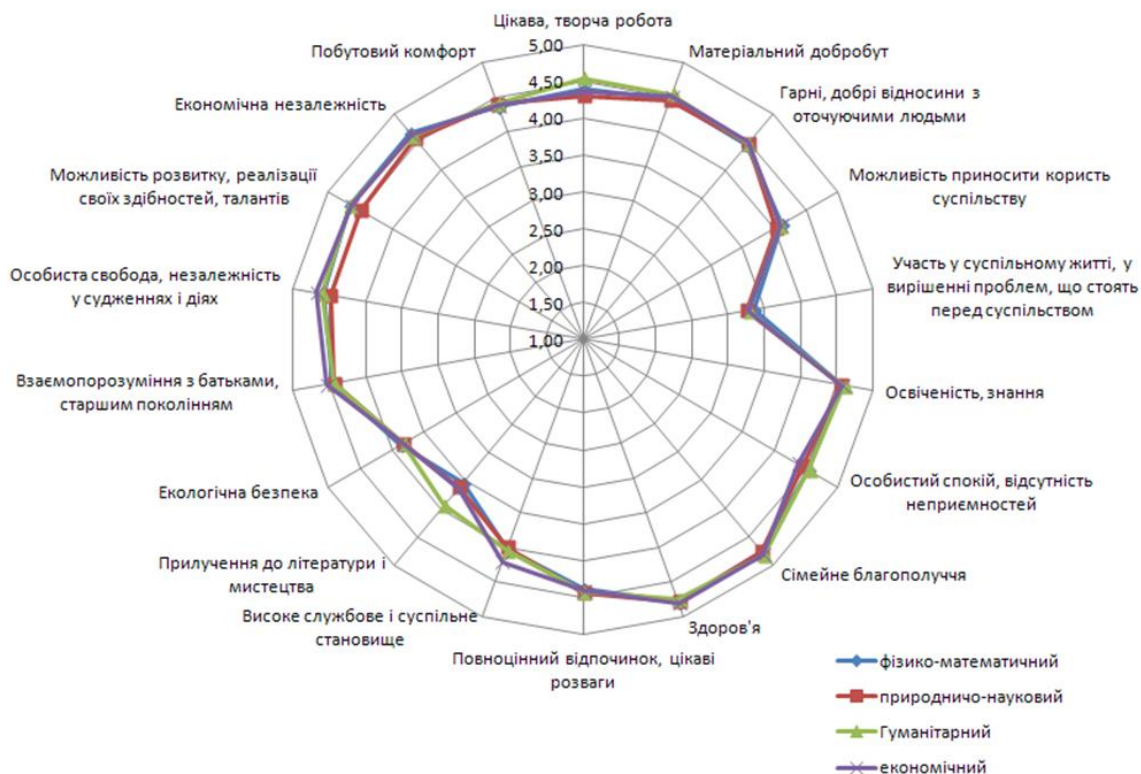


Рис. 7.7. Ціннісні орієнтації в залежності від профілю навчання (візуалізація за допомогою Microsoft Excel) (середні значення; інтервал від 1 – «зовсім не цінно» до 5 – «дуже цінно»)

Діаграма демонструє, що ціннісні орієнтації не дуже сильно розрізняються в аналізованих групах. Невеликі відмінності можна побачити лише в оцінці таких цінностей як залучення до літератури та мистецтва; цікава, творча робота; економічна незалежність; особиста свобода, незалежність у судженнях та діях; самореалізація; високе службове та суспільне положення. Отже, виникає питання, чи можна знайдені невеликі відмінності вважати підтвердженням нашої гіпотези. Для відповіді необхідно перевірити статистичну значущість розбіжностей.

Крок 3. Дисперсійний аналіз. Розрахунок статистичної значущості розбіжностей у кількох групах може бути здійснений за допомогою однофакторного дисперсійного аналізу, для проведення якого в SPSS застосовують команду *Analyze* → *Compare Means* → *One-Way ANOVA* (рис. 7.8), у результаті виконання якої з'явиться відповідне діалогове вікно, в якому у полі **Factor** задається незалежна (групуюча) змінна, а у полі **Dependent List** – залежна, тобто та, для якої визначаються оцінювані середні значення (у нашому прикладі – цінності).

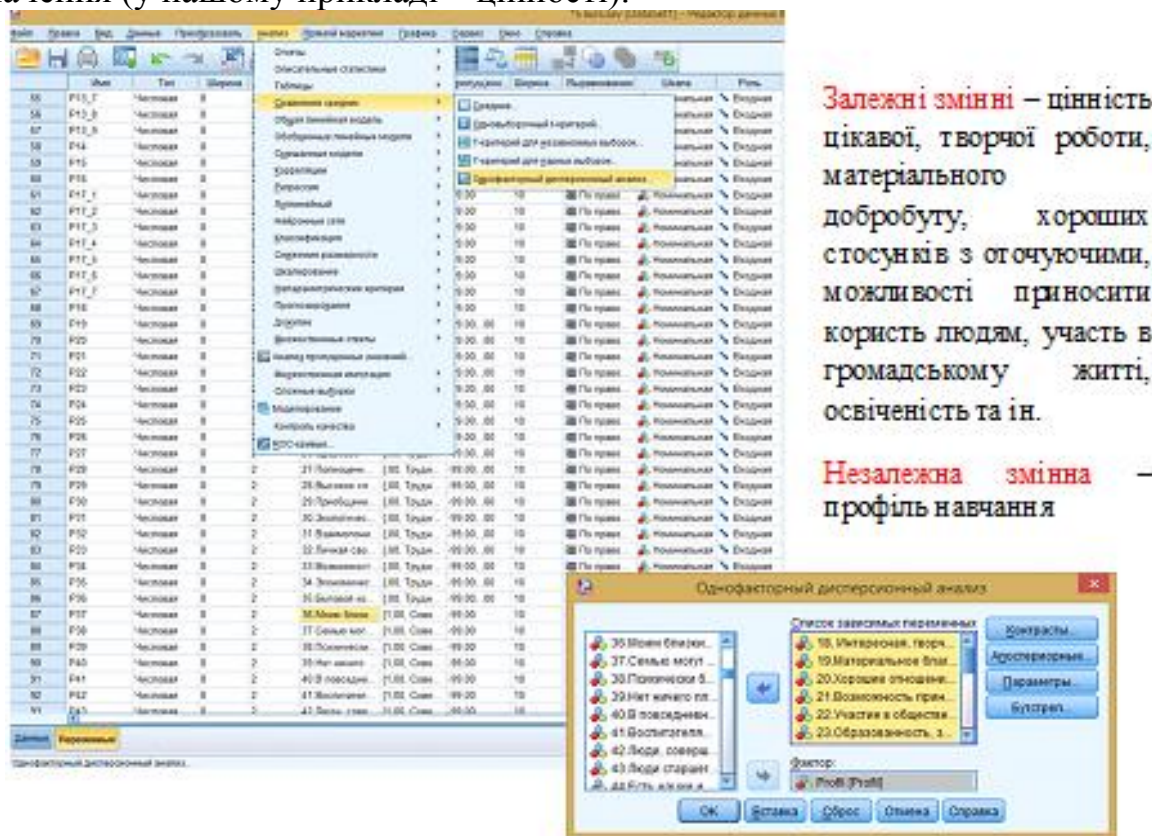


Рис. 7.8. Виклик процедури дисперсійного аналізу

Головним результатом процедури дисперсійного аналізу є ймовірність помилки p , що відповідає тестовому значенню **F-критерію**, яке виводиться в правій колонці під заголовком "**Sig.**" ("**Значущість**"). Її величина свідчить про статистичну значущість розбіжностей, зокрема, у нашому прикладі, значущість розбіжностей ціннісних орієнтирів у групах по профілю навчання.

Таблиця 7.7

**Результати виконання однофакторного дисперсійного аналізу
(залежна змінна – цінність цікавої, творчої роботи; незалежна змінна –
профіль навчання)**

18. Цікава, творча робота

	Сума квадратів	ст.св.	Середній квадрат	F	Знч.
Між групами	11,246	3	3,749	4,917	,002
Всередині груп	769,287	1009	,762		
Разом	780,533	1012			

Зверніть увагу, що дисперсійний аналіз тільки показує значущі чи ні розбіжності. Для інтерпретації самих розбіжностей необхідно розрахувати середні значення в кожній групі.

Враховуючи, що допустима ймовірність помилки в соціологічних дослідженнях дорівнює 0,05 (тобто 5%), інтерпретація значущості F-статистики здійснюється наступним чином: якщо $Sig \leq 0,05$, то виявлені розбіжності є статистично значущими на 5% рівні та соціолог має право їх змістовно інтерпретувати.

Для зручності змістовної інтерпретації ми пропонуємо звести до однієї таблиці результати розрахунку середніх значень та F-статистики. Така таблиця дає можливість відкинути всі статистично незначущі розходження та зосередити увагу на статистично значущих.

Таблиця 7.8

**Середні значення ціннісних орієнтацій в залежності від профілю
навчання та статистична значущість розбіжностей за F-статистикою
(інтервал від 1 – «зовсім не цінно» до 5 – «дуже цінно»)**

Наскільки цінними особисто для Вас є ... ?	Профіль навчання				Статистична значущість розбіжностей за F-статистикою
	Фізико-математичний	Природничо-науковий	Гуманітарний	Економічний	
Цікава, творча робота	4,39	4,3	4,55	4,36	0,001
Матеріальний добробут	4,46	4,46	4,54	4,51	0,402
Гарні, добрі відносини з оточуючими людьми	4,46	4,47	4,47	4,49	0,977
Можливість приносити користь суспільству	4,12	4,02	4,09	4,1	0,669
Участь у суспільному житті, у вирішенні проблем, що стоять перед суспільством	3,36	3,25	3,28	3,27	0,747
Освіченість, знання	4,55	4,56	4,6	4,58	0,765

Таблиця 7.8 (продовження)

	Профіль навчання				Статистична значущість розбіжностей за F статистикою
	Фізико-математичний	Природничо-науковий	Гуманітарний	Економічний	
Особистий спокій, відсутність неприємностей	4,52	4,45	4,54	4,37	0,081
Сімейне благополуччя	4,77	4,76	4,82	4,81	0,544
Здоров'я	4,8	4,79	4,76	4,81	0,606
Повноцінний відпочинок, цікаві розваги	4,38	4,44	4,42	4,41	0,884
Високе службове і суспільне становище	4,04	4	4,04	4,21	0,084
Прилучення до літератури і мистецтва	3,55	3,62	3,94	3,64	0,000
Екологічна безпека	3,87	3,82	3,82	3,83	0,964
Взаємопорозуміння з батьками, старшим поколінням	4,46	4,43	4,45	4,53	0,543
Особиста свобода, незалежність у судженнях і діях	4,58	4,49	4,6	4,67	0,021
Можливість розвитку, реалізації своїх здібностей, талантів	4,66	4,51	4,66	4,63	0,009
Економічна незалежність	4,66	4,56	4,61	4,62	0,389
Побутовий комфорт	4,36	4,4	4,41	4,38	0,904

Крок 4. Інтерпретація результатів.

Ціннісні пріоритети п'ятикурсників мало варіюють в залежності від профілю навчання, більшість розходжень виявилися статистично незначущими (див. табл.). Невеликі статистично значущі розходження виявились за оцінками п'ятикурсниками важливості для них лише наступних чотирьох цінностей:

- *Прилучення до літератури та мистецтва*, що найвище оцінюється гуманітаріями (середнє значення дорівнює 3.94) та найнижче оцінюються студентами природничо-наукового профілю навчання (3.55).

- *Цікава, творча робота* – найбільш високо оцінюється гуманітаріями (4.55) та найнижче оцінюються студентами природничо-наукового профілю навчання (4.3).

- *Особиста свобода, незалежність у судженнях і діях*, що однаково високо цінують економісти (4.67) та гуманітарії (4.66), а студенти природничо-наукового профілю навчання оцінюють найнижче з аналізованих груп (4.49).

- *Можливість розвитку, реалізації своїх здібностей, талантів* студенти фізико-математичного та гуманітарного профілю навчання оцінюють однаково високо (4.66), а представники природничо-наукового профілю навчання оцінюють трохи менше (4.51).

Висновок. В цілому гіпотеза не підтвердилась, більшість цінностей мають однаково важливість для всіх опитаних. Незначні статистично значущі розбіжності є лише в оцінці важливості (1) прилучення до літератури та мистецтва, (2) цікавої, творчої роботи, (3) особистої свободи, незалежності у судженнях і діях та (4) самореалізації.

Багатофакторний дисперсійний аналіз – дисперсійний аналіз із двома чи більше незалежними змінними (факторами). Він призначений для вивчення залежності досліджуваної змінної відразу від багатьох ознак, а також виявляє ефекти взаємодії факторів.

Зверніть увагу, що багатофакторний та багатовимірний дисперсійні аналізи розрізняються! Фактори – незалежні змінні. Коли незалежних змінних більше, ніж одна, аналіз називають багатофакторним.

Багатомірним називають дисперсійний аналіз, коли є кілька залежних змінних.

Найпростішим різновидом багатофакторного дисперсійного аналізу є двофакторний дисперсійний аналіз, який ми розглянемо на прикладі перевірки гіпотези, що статусні домагання студентів зумовлені соціальним положенням батьків та матеріальним станом сім'ї.

Багатофакторний дисперсійний аналіз дозволяє дослідити залежність середніх значень однієї змінної (у нашому прикладі – середніх значень статусних домагань груп респондентів) від кількох факторів, наприклад, матеріального стану сім'ї та соціального положення батьків. Отже, у нашому прикладі:

Залежна змінна – статусні домагання.

Фактори (незалежні змінні): 1) матеріальний стан сім'ї; 2) соціальне положення батьків.

Проведення багатофакторного дисперсійного аналізу в SPSS має певні особливості, які ми зараз розглянемо.

Насамперед звернемо увагу, що інструментом вимірювання статусних домагань було питання №249, котре в анкеті виглядало наступним чином:

НАШЕ СУСПІЛЬСТВО МОЖНА УЯВИТИ У ВИГЛЯДІ СХОДИНОК, НА ЯКИХ ЗНАХОДЯТЬСЯ ЛЮДИ, ЩО НАЛЕЖАТЬ ДО РІЗНИХ СОЦІАЛЬНИХ ПРОШАРОК. **ЯКУ СОЦІАЛЬНУ СХОДИНКУ ЗАЙМАЄ СІМ'Я ВАШИХ БАТЬКІВ І НА ЯКУ СХОДИНКУ ХОТІЛИ Б ПОТРАПИТИ ВИ?** (зробіть позначку як праворуч, так і ліворуч).

248. Соціальна сходинка, до якої належить сім'я Ваших батьків	249. Соціальна сходинка, на яку хотіли б потрапити Ви
<p>Найвищий прошарок</p> <p>9. _____</p> <p>8. _____</p> <p>7. _____</p> <p>6. _____</p> <p>Середній прошарок 5. _____</p> <p>4. _____</p> <p>3. _____</p> <p>2. _____</p> <p>Найнижчий 1. _____ прошарок</p>	<p>9. _____</p> <p>8. _____</p> <p>7. _____</p> <p>6. _____</p> <p>5. _____</p> <p>4. _____</p> <p>3. _____</p> <p>2. _____</p> <p>1. _____</p>
10. Важко відповісти	10. Важко відповісти

Рис. 7.9. Вигляд питання «Яку соціальну сходинку займає сім'я Ваших батьків, і на яку сходинку хотіли б потрапити Ви?» в анкеті

До початку обробки даних треба вказати, що 10 – це пропущене значення. Якщо цього не зробити, середні значення будуть обчислені неправильно.

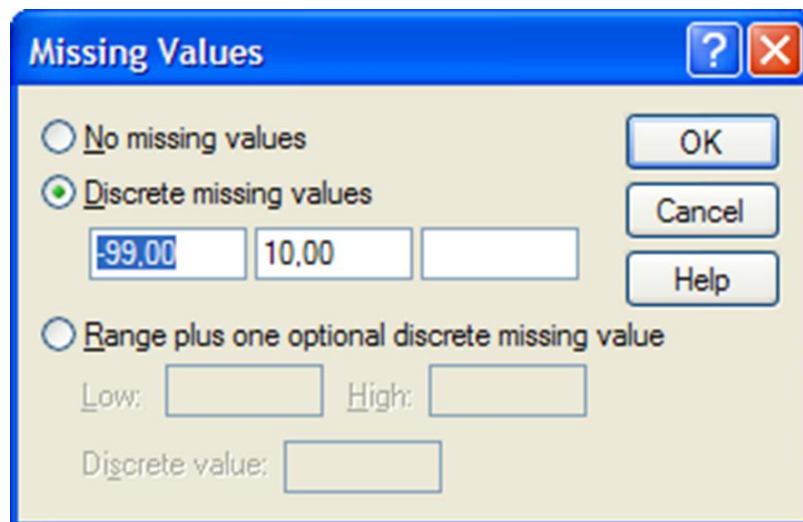


Рис. 7.10. Діалогове вікно Missing Values (Пропущені значення)

Тепер можна викликати процедуру багатофакторного дисперсійного аналізу: *Analyze (Аналіз) → General Linear Model (Загальна лінійна модель) → Univariate... (Одновимірна)*. Відкриється діалогове вікно *Univariate (Одновимірна)*, у якому вказуємо залежну та незалежні змінні:

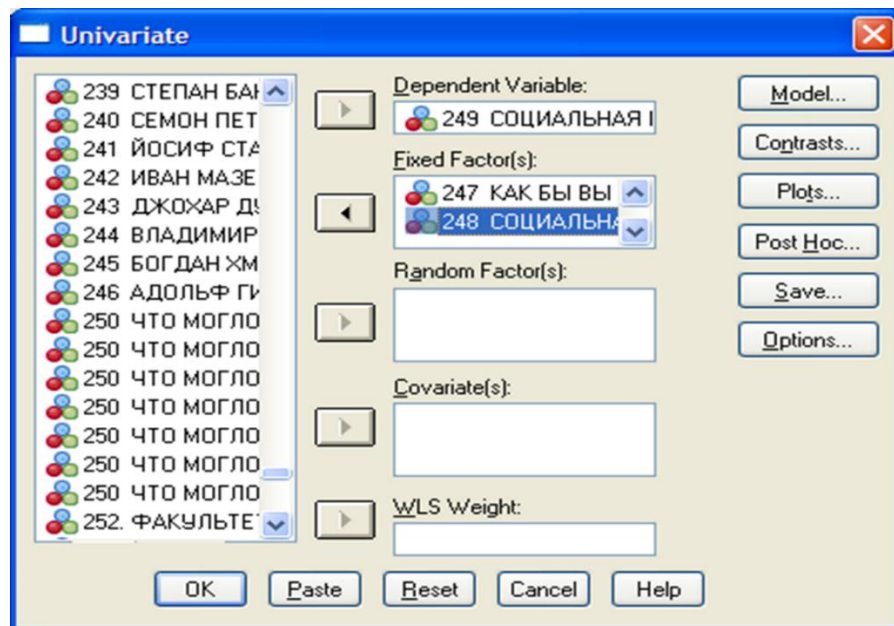


Рис. 7.11. Діалогове вікно багатofакторного дисперсійного аналізу

У результаті отримаємо таке:

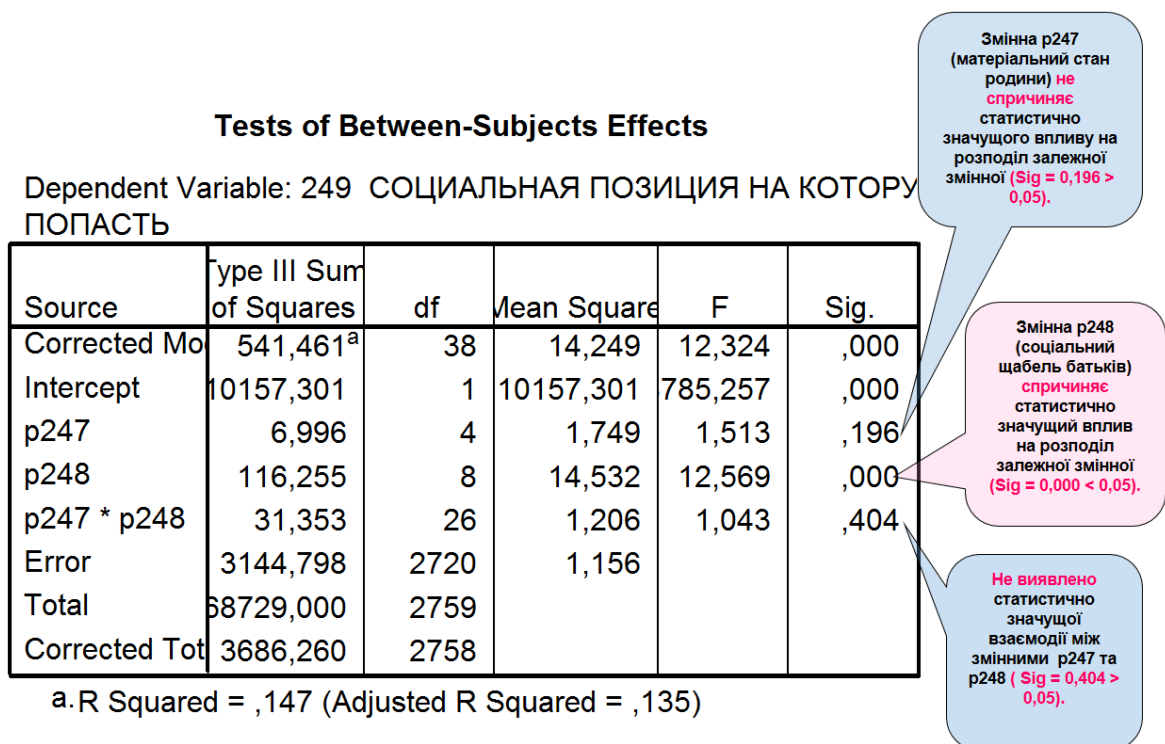


Рис. 7.12. Результати виконання процедури багатofакторного дисперсійного аналізу

Ці результати дають можливість відповісти на три питання:

1. Чи існує статистично значуща відмінність статусних домагань 9-ти груп студентів, чії батьки займають 9 різних шаблеь уявних соціальних сходів?

Відповідь ґрунтується на аналізі *Sig*, яка перевищує 0,05: змінна «Матеріальне становище сім'ї» не спричиняє статистично значущого впливу на розподіл змінної «Соціальна сходинка, на яку хотіли б потрапити».

2. Чи існує статистично значуща відмінність статусних домагань між 5-тьма групами з різним матеріальним станом?

Змінна «Соціальна сходинка батьків» спричиняє статистично значущий вплив на розподіл змінної «Соціальна сходинка, на яку хотіли б потрапити» ($Sig = 0,000 < 0,05$).

3. Чи існує статистично значуща взаємодія змінних «Матеріальний стан сім'ї» та «Соціальне положення батьків»?

Не виявлено статистично значущої взаємодії між змінними «Матеріальне становище сім'ї» та «Соціальна сходинка, на яку хотіли б потрапити» ($Sig > 0,05$).

7.4. Непараметричні тести

Непараметричні тести – це методи статистичного висновку, які не ґрунтуються на інформації щодо параметрів генеральної сукупності. Вони, на відміну від параметричних, можуть застосовуватися для кількісних змінних, розподіл яких не підпорядковується нормальному закону. Ці методи аналізують не самі виміряні значення, а їхні ранги, отже, вони підходять для вивчення порядкових змінних.

Таблиця 7.9

Непараметричні тести для перевірки значущості розбіжностей

	Незалежні вибірки	Залежні вибірки
Дві вибірки	<p>Виклик процедури: <i>Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Legacy Dialogs... (Застарілі діалогові вікна) → 2 Independent Samples... (2 незалежні вибірки)</i></p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо $Sig. (2-tailed) \leq 0,05$, то розбіжності вважаються значущими на 5%-му рівні. Якщо $Sig. (2-tailed) > 0,05$, то розбіжності вважаються не значущими на такому ж рівні.</p>	<p>Виклик процедури: <i>Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Legacy Dialogs... (Застарілі діалогові вікна) → 2 Related Samples... (2 залежні вибірки)</i></p> <p>Статистична гіпотеза: H_0: розбіжностей немає. H_1: розбіжності є.</p> <p>Інтерпретація розрахунків: Якщо $Sig. (2-tailed) \leq 0,05$, то розбіжності вважаються значущими на 5%-му рівні. Якщо $Sig. (2-tailed) > 0,05$, то розбіжності вважаються не значущими на такому ж рівні.</p>

Таблиця 7.9 (продовження)

	Незалежні вибірки	Залежні вибірки
Більше двох вибірок	<p>Виклик процедури: <i>Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Legacy Dialogs... (Застарілі діалогові вікна) → K Independent Samples... (k незалежних вибірок)</i></p> <p>Статистична гіпотеза: <i>H₀: розбіжностей немає.</i> <i>H₁: розбіжності є.</i></p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>	<p>Виклик процедури: <i>Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Legacy Dialogs... (Застарілі діалогові вікна) → K Related Samples... (k залежних вибірок)</i></p> <p>Статистична гіпотеза: <i>H₀: розбіжностей немає.</i> <i>H₁: розбіжності є.</i></p> <p>Інтерпретація розрахунків: Якщо Sig. (2-tailed) ≤ 0,05, то розбіжності вважаються значущими на 5%-му рівні. Якщо Sig. (2-tailed) > 0,05, то розбіжності вважаються не значущими на такому ж рівні.</p>

Порівнюючи параметричні та непараметричні критерії, слід зазначити, що перші мають більшу потужність, вони здатні з більшою вірогідністю відкидати нульову гіпотезу, якщо вона невірна. Отже, у випадках, коли аналізовані змінні розподілені нормально, бажано віддавати перевагу параметричним методам. Проте практика свідчить, що переважна більшість даних, одержуваних у результаті соціологічних опитувань, не підкоряються нормальному закону розподілу, тому виникає потреба у застосуванні непараметричних критеріїв.

Переваги непараметричних методів:

- *Непараметричне тестування не потребує ніяких припущень щодо характеру розподілу генеральної сукупності, з якої взято досліджувану вибірку.*
- *Методи непараметричного тестування можуть застосовуватися навіть тоді, коли вибірка дуже мала.*

Недолік непараметричних методів:

- *Непараметричні тести мають нижчу потужність, ніж параметричні. Це означає, що при застосуванні непараметричних методів ймовірність відхилення нульової гіпотези, коли вірна альтернативна, є нижчою.*

SPSS дозволяє розрахувати чотири непараметричних тести (див. рис. 7.13):

- Критерій Манна-Уїтні.
- Критерій екстремальних реакцій Мозеса.
- Двовибірковий критерій Колмогорова-Смірнова.
- Критерій серій Вальда-Вольфовица.

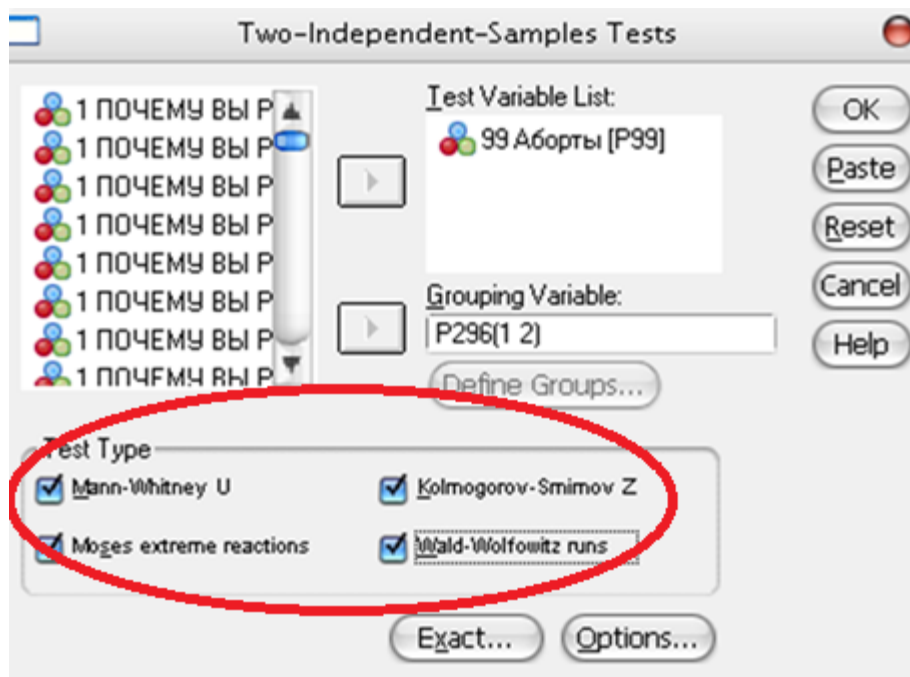


Рис. 7.13. Непараметричні тести для незалежних вибірок в SPSS

U-тест Манна - Уїтні (англ. *Mann – Whitney U-test*) – найвідоміший статистичний непараметричний критерій порівняння середніх двох незалежних вибірок, заснований на порівнянні рангів.

Існує два шляхи розрахунку U-критерію Манна-Уїтні в SPSS:

1) *Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Independent Samples (Незалежні вибірки) → Settings (Налаштування) → поставити галочку в блоці Customize Tests (Налаштувати критерії) поряд з Mann-Whitney U (2 Samples) (U-критерій Манна-Уїтні) (2 вибірки)* (див. рис. 7.14).

2) *Analyze (Аналіз) → Nonparametric Tests (Непараметричні тести) → Legacy Dialogs (Застарілі діалогові вікна) → Two Independent Samples (Для двох незалежних вибірок) поставити галочку в блоці Test Type (Тип тесту) поряд з Mann-Whitney U (U-критерій Манна-Уїтні)* (див. рис. 7.15).

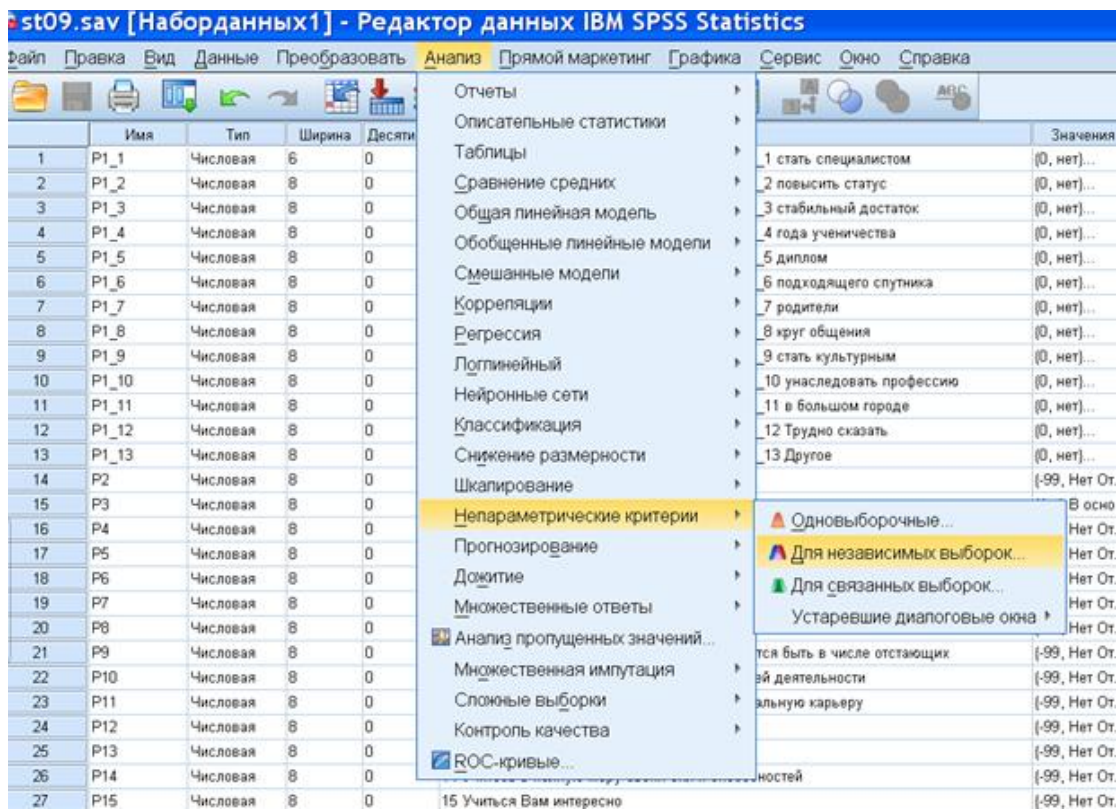


Рис. 7.14. Перший спосіб розрахунку U-критерію Манна-Уїтні в SPSS

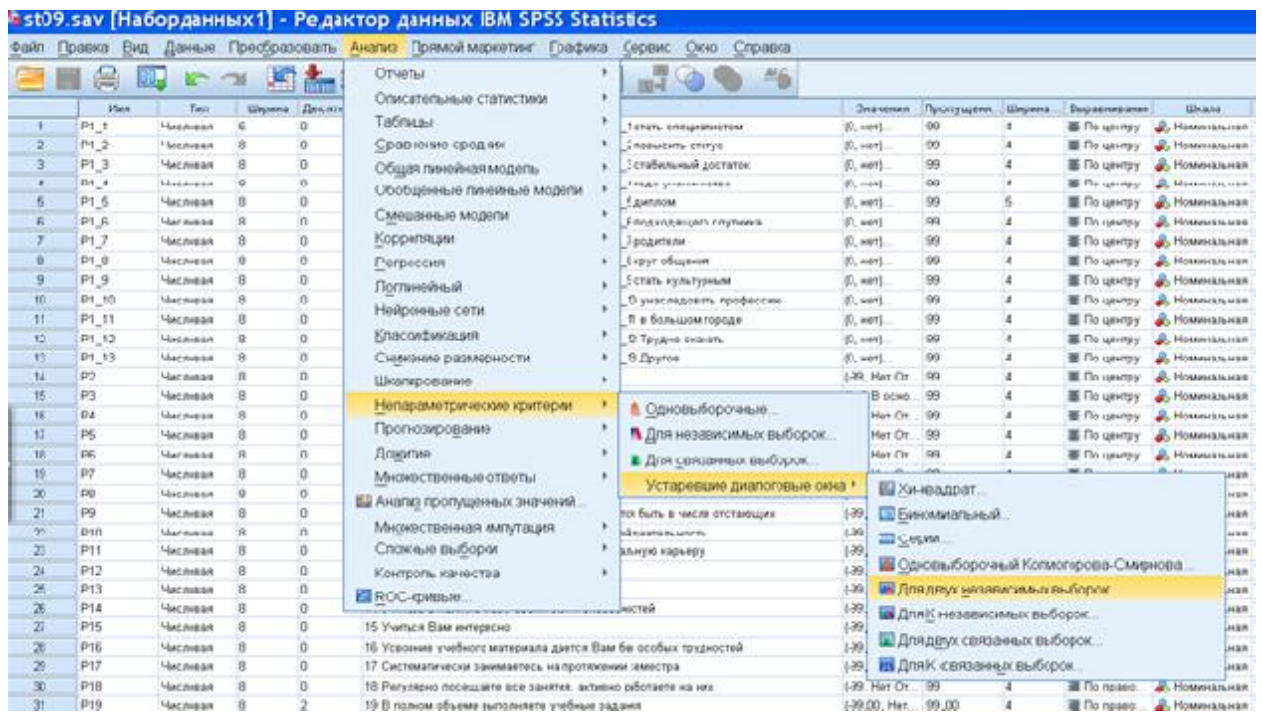


Рис. 7.15. Другий спосіб розрахунку U-критерію Манна-Уїтні в SPSS

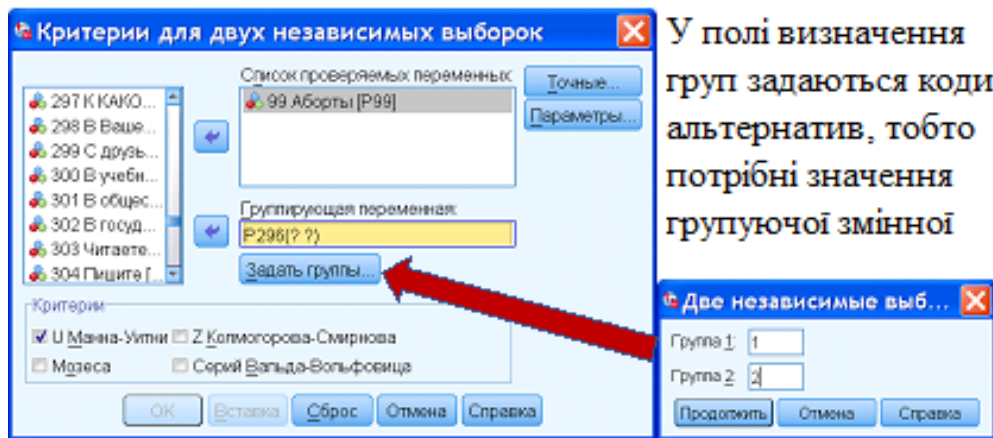


Рис. 7.16. Визначення груп, що будуть порівнюватися з застосуванням U-критерію Манна-Уїтні

У результаті розрахунків SPSS виведе наступні показники:

NPar Tests

[DataSet1] F:\Массивы\st09.sav

➔ Mann-Whitney Test

Ranks				
	296 Ваш пол	N	Mean Rank	Sum of Ranks
99 Аборты	1 Мужской	978	1553,75	1519567,50
	2 Женский	1997	1455,80	2907232,50
	Total	2975		

Test Statistics ^a	
	99 Аборты
Mann-Whitney U	912229,50
Wilcoxon W	2907232,5
Z	-3,124
Asymp. Sig. (2-tailed)	,002

a. Grouping Variable: 296 Ваш пол

Рис. 7.17. Результати розрахунку U-критерію Манна-Уїтні в SPSS

Критерій екстремальних реакцій Мозеса передбачає вплив експериментальної змінної на деякі об'єкти в одному напрямі, а на інші – у протилежному. Критерій виявляє екстремальні у порівнянні з контрольною групою реакції. Він зосереджується навколо розмаху контрольної групи, та показує силу впливу на цей розмах екстремальних значень з експериментальної групи, коли експериментальна та контрольна групи об'єднано. Контрольна група задається значенням для групи 1 у діалоговому вікні «Дві незалежні вибірки: задати групи». Спостереження з обох груп об'єднуються та ранжуються. Розмах контрольної групи обчислюється як

У полі визначення груп задаються коди альтернатив, тобто потрібні значення групуючої змінної

У першій таблиці виводиться кількість спостережень, усереднені ранги та рангова сума для двох вибірок (причому більшим значенням привласнюються нижчі рангові місця)

У другій таблиці представлено: тестова величина U, що виявлена за допомогою тесту Манна й Уїтні; W-тест Уїлкоксона; точне значення ймовірності помилки p, яке рекомендується використовувати, коли кількість спостережень є меншою 30; тестова величина z, виявлену тестом Колмогорова-Смірнова, а також ймовірність помилки p стосовно неї, яку застосовують за умови кількості спостережень, більшої 30

різниця рангів найбільшого та найменшого значень у контрольній групі, збільшена на 1. Оскільки випадкові винятки можуть легко викривити величину розмаху, 5% спостережень з кожного кінця розподілу контрольної групи автоматично відсікаються.

Таблиця 7.10

Результати розрахунку критерію Мозеса (1)

Частоти

296 Ваша стаття	N
99 Аборти 1 Чоловіча (Контрольне)	978
2 Жіноча (Експериментальне)	1997
Всього	2975

Таблиця 7.11

Результати розрахунку критерію Мозеса (2)

Статистики критерію^{a,b}

	Value Значення	99 Аборти
Спостережень в контрольованій групі		2207
розмах	Знч. (1-стор.)	,000
Розмах усіченої контрольної групи	Знч. (1-стор.)	2207 ,000
Викиди усічені з кожного кінця		48

a. Критерій Мозеса

b. Групуєча змінна: 296 Ваша стаття

Критерій Z Колмогорова-Смірнова та критерій серій Вальда-Вольфовица носять більш загальний характер та виявляють розбіжності між розподілами у їхньому розташуванні та формі. Критерій серій об'єднує та ранжує спостереження з обох груп. Якщо обидві вибірки узяті з однієї генеральної сукупності, обидві групи мають бути випадково розподіленими за проранжованими даними.

Таблиця 7.12

Результати розрахунку критерію Вальда-Вольфовица

Статистики критерію^{b, c}

	Кількість серій	Z	Асимпт. значущість (одностор.)
99 Аборти	Можливий мінімум	6 ^a	-54,347 ,000
	Можливий максимум	1957 ^a	26,718 1,000

a. Міжгрупових зв'язків 5, спостережень в них 2975.

b. Критерій Вальда-Вольфовица

c. Групуєча змінна: 296 Ваша стаття

Двовибірковий тест Колмогорова-Смірнова призначений для перевірки гіпотези про збіг розподілу в двох вибірках. Він ґрунтується на максимумі модуля різниці між емпіричними функціями розподілу для обох вибірок. Якщо ця різниця є значущо великою, розподіли вважаються різноманітними.

Статистика критерію - абсолютна величина (модуль) різниці емпіричних функцій розподілу у вказаних вибірках:

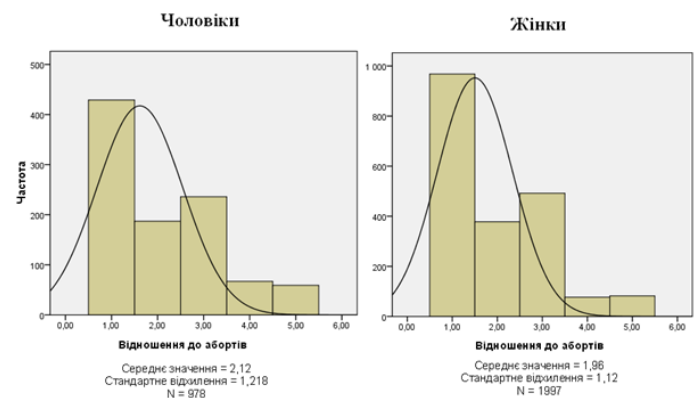
$$K_S = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \max_x |F_1(x) - F_2(x)|$$

де N_1 и N_2 - обсяги вибірок.

Двовибірковий тест Колмогорова-Смірнова

Частоти

296 Ваша стать		N
99 Аборти	1 Чоловіча	978
	2 Жіноча	1997
	Загалом	2975



Статистики критерію^a

		99 Аборти
Різниця екстремумів	Модуль	,049
	Позитивні	,049
	Негативні	,000
Статистика Z Колмогорова-Смірнова		1,261
Асимптот. знч. (двостороння)		0,83

a. Групуєча змінна: 296 Ваша стать

Рис. 7.18. Результати розрахунку двовибіркового критерію Колмогорова-Смірнова

7.5. Аналіз розбіжностей відсотків (часток)

Перевірка статистичної значущості розбіжностей відсотків (часток) є розповсюдженим завданням, необхідність вирішення якого достатньо часто виникає у процесі аналізу результатів кількісних соціологічних досліджень. При цьому головним питанням є таке: чи не зумовлені виявлені розбіжності випадковими факторами, чи дійсно розрізняються частки досліджуваної ознаки в аналізованих групах у генеральній сукупності? Пошук відповіді зводиться до перевірки статистичної гіпотези про рівність часток деякої ознаки у двох вибірках. При аналізі результатів соціологічних досліджень

зазвичай використовують методи, призначені для незалежних вибірок досить великого обсягу, у яких розподіл наближається до нормального. На практиці це означає, що виконуються вимоги: $n_1 v_1^B > 5$, $n_2 v_2^B > 5$, $n_1(1-v_1^B) > 5$, $n_2(1-v_2^B) > 5$ (пояснення позначень див. у табл. 7.13).

Таблиця 7.13

Основні позначення

	Вибірка №1	Вибірка №2
Обсяг	n_1	n_2
Частка ознаки у вибірковій сукупності	v_1^B	v_2^B
Відсоток ознаки у вибірковій сукупності	$n_1 v_1^B$	$n_2 v_2^B$
Обсяг генеральної сукупності	N	N
Частка ознаки у генеральній сукупності (невідомо)	v_1^G	v_2^G
Відсоток ознаки у генеральній сукупності (невідомо)	$N v_1^G$	$N v_2^G$

Насамперед сформулюємо статистичну гіпотезу:

$H_0: v_1^G - v_2^G = 0$, різниця часток досліджуваної ознаки в аналізованих групах у генеральній сукупності дорівнює нулю.

$H_1: v_1^G \neq v_2^G$, різниця часток досліджуваної ознаки в аналізованих групах у генеральній сукупності не дорівнює нулю. (пояснення позначень див. у табл. 7.13).

Статистикою для перевірки цієї гіпотези слугуватиме z^2 :

$$z = \frac{|v_1^B - v_2^B|}{\sigma_z},$$

$$\text{де } \sigma_z = \sqrt{v^B(1-v^B) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

$$v^B = \frac{v_1^B n_1 + v_2^B n_2}{n_1 + n_2}$$

Відомо, що на 5%-му рівні значущості критичні точки статистики z дорівнюють $\pm 1,96$. Область, що знаходиться між ними ($-1,95 < z < 1,96$), є областю прийняття гіпотези, поза цими критичними точками ($z \leq -1,96$ або $z \geq 1,96$) – критичною областю. Якщо $-1,95 < z < 1,96$, то гіпотеза H_0

² Див.: Паніотто, В.І., Максименко, В.С., Харченко, Н.М. Статистичний аналіз соціологічних даних. К.: «КМ Академія», 2004. С.202.

приймається. У випадку, коли $z \leq 1,96$ або $z \geq 1,96$, гіпотеза відкидається на 5%-му рівні значущості.

Безумовно, розрахунок статистичної значущості розходжень часток та відсотків – дуже трудомістка процедура, яку наразі навряд чи хтось з дослідників-практиків виконує вручну. Нажаль у пакеті SPSS ця процедура не реалізована. Проте існує велика кількість онлайн-калькуляторів, у яких легко виконати потрібні розрахунки.

SPSS пропонує опосередкований підхід до вирішення завдання перевірки статистичної значущості розходжень відсотків, який відомий як **аналіз залишків**.

Аналіз залишків є специфічним різновидом аналізу таблиць двовимірних розподілів, що ґрунтується на стандартизації, дуже схожій на статистичну стандартизацію за допомогою z-значень. Якщо z-значення дозволяють стандартизувати звичайні розподіли, щоб отримати можливість їхнього порівняння, то стандартизовані залишки нормалізують дані при тестуванні гіпотез χ^2 -квадрат, що сприяє інтерпретації таблиць спряженості (двовимірних розподілів). Використання стандартизованих залишків у контексті аналізу результатів соціологічних досліджень поширилася з першої половини 1990-х років, даний метод дає корисні результати при мікроаналізі двовимірних таблиць спряженості. Недоліком цього методу є обмеження: очікувані частоти повинні бути більше, ніж 5.

Залишок (англ. *residuals*) – це різниця між спостережуваним і очікуваним значеннями досліджуваної ознаки. Очікуване значення являє собою кількість спостережень в осередку таблиці двовимірного розподілу за умови незалежності змінних у рядках та стовпцях. Позитивні (негативні) значення залишку вказує на те, що в осередку є більше (менше) спостережень, ніж в разі, якби двовимірний розподіл був рівномірним.

Стандартизований (нормований) залишок (англ. *standardized residuals*) являють собою залишок, поділений на оцінку його стандартного відхилення:

$$R = \frac{F_o - F_e}{\sqrt{F_e}}$$

де F_o – спостережуване значення (*observed count*), тобто це кількість спостережень в осередку таблиці двовимірного розподілу за умови незалежності змінних у рядку і стовпці;

F_e – очікуване значення (*expected count*), тобто це кількість спостережень в осередку таблиці двовимірного розподілу за умови незалежності змінних у рядку і стовпці.

Стандартизовані залишки мають середнє 0 і стандартне відхилення 1.

Виявлення розходжень частот та відсотків в пакеті SPSS із застосуванням критерію χ^2 -квадрат здійснюється за допомогою команди: **Descriptive statistics (Описові статистики) → Crosstabs (Таблиці спряженості)**). Розглянемо на прикладі, яким чином здійснюється аналіз у даному випадку. Припустимо, що необхідно перевірити гіпотезу, що сімейне

благополуччя вище цінується жінками, ніж чоловіками (файл st09.sav). Для цього можна побудувати звичайний двовимірний розподіл (див. табл. 7.14) та порівняти розподіли відповідей (ряди у двовимірній частотній таблиці) між певними групами респондентів (стовпці в цій таблиці).

Таблиця 7.14

Таблиця двовимірного розподілу ознак «Сімейне благополуччя» та «Стать» (масив st09.sav)

			296 Ваша стать		Разом
			Чоловіча	Жіноча	
60. Цінність сімейного благополуччя	Зовсім не цінно	Частота	7	2	9
		% в 296 Ваша стать	,7%	,1%	,3%
	Не дуже цінно	Частота	13	14	27
		% в 296 Ваша стать	1,3%	,7%	,9%
	Важко сказати	Частота	49	44	93
		% в 296 Ваша стать	5,0%	2,2%	3,1%
	Цінно	Частота	226	258	484
		% в 296 Ваша стать	23,2%	13,0%	16,4%
	Дуже цінно	Частота	679	1664	2343
		% в 296 Ваша стать	69,7%	84,0%	79,3%
Разом	Частота	974	1982	2956	
	% в 296 Ваша стать	100,0%	100,0%	100,0%	

Проте тест хі-квадрат, що розраховується для цієї таблиці, не може показати статистичну значущість розбіжностей відсотків в кожному рядку (див. табл. 7.15).

Таблиця 7.15

Chi-Square Tests (Критерій хі-квадрат)

	Value Значення	df Ступені свободи	Asymp. Sig. (2-sided) Асимпт. значущість (2-стор.)
Pearson Chi-Square Хі-квадрат Пірсона	85,508 ^a	4	,000
Likelihood Ratio Відношення правдоподібності	81,891	4	,000
Linear-by-Linear Association Лінійно-лінійний зв'язок	75,922	1	,000
N of Valid Cases Кількість валідних спостережень	2956		

a 1 cells (10,0%) have expected count less than 5. The minimum expected count is 2,97.

(а. В 1 (10,0%) клітинці таблиці очікувана частота менше 5. Мінімальна очікувана частота дорівнює 2,97).

Для того, щоб з'ясувати, чи є статистично значущими розбіжності між двома (або більше) досліджуваними групами необхідно здійснити аналіз стандартизованих залишків, для чого треба застосувати команду *Descriptive statistics (дескриптивні статистики)* → *Crosstabs (Таблиці спряженості)*. З'явиться діалогове вікно *Crosstabs*. Перенесіть ознаку «60 Сімейне благополуччя» у список змінних рядків, ознаку «296 Ваша стаття» – у список змінних стовпців, і в діалоговому вікні, що відкриває кнопкою *Cells...*, крім виводу спостережуваних частот (прапорець *Observed* у групі *Counts*), задайте також вивід очікуваних частот (прапорець *Expected*) та нормованих залишків (прапорець *Standardized* у групі *Residuals*), як показано на рисунку 7.19.

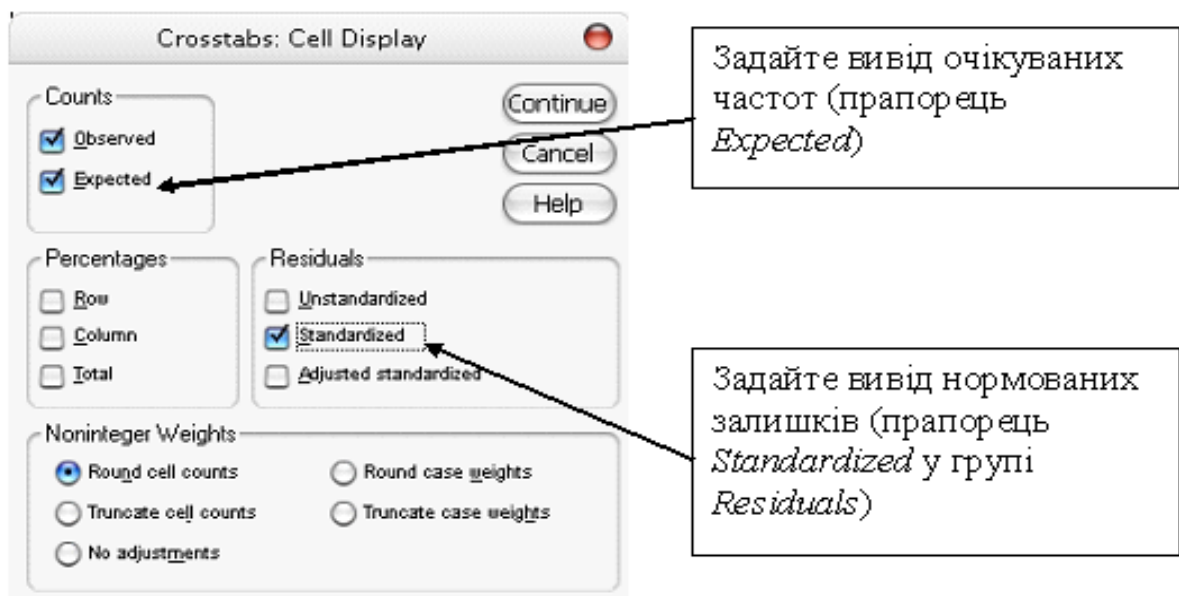


Рис. 7.19. Розрахунок очікуваних частот та стандартизованих залишків

Результати розрахунку очікуваних частот та стандартизованих (нормованих) залишків будуть внесені у таблицю двовимірного розподілу, як показано у табл. 7.16.

Таблиця 7.16

Таблиця двовимірного розподілу ознак «Цінність сімейного благополуччя» та «Стать» з очікуваними частотами та стандартизованими залишками (Масив st09.sav)

60. Цінність сімейного благополуччя		296. Ваша стать		Разом
		Чоловіча	Жіноча	
Зовсім не цінно	Count (Частота)	7	2	9
	Expected Count (Очікувана частота)	3,0	6,0	9,0
	% в 296 Ваша стать	,7%	,1%	,3%
	Std. Residual (Стандартизований залишок)	2,3	-1,6	
Не дуже цінно	Count (Частота)	13	14	27
	Expected Count (Очікувана частота)	8,9	18,1	27,0
	% в 296 Ваша стать	1,3%	,7%	,9%
	Std. Residual (Стандартизований залишок)	1,4	-1,0	
Важко сказати	Count (Частота)	49	44	93
	Expected Count (Очікувана частота)	30,6	62,4	93,0
	% в 296 Ваш пол	5,0%	2,2%	3,1%
	Std. Residual (Стандартизований залишок)	3,3	-2,3	
Цінно	Count (Частота)	226	258	484
	Expected Count (Очікувана частота)	159,5	324,5	484,0
	% в 296 Ваш пол	23,2%	13,0%	16,4%
	Std. Residual (Стандартизований залишок)	5,3	-3,7	
Дуже цінно	Count (Частота)	679	1664	2343
	Expected Count (Очікувана частота)	772,0	1571,0	2343,0
	% в 296 Ваша стать	69,7%	84,0%	79,3%
	Std. Residual (Стандартизований залишок)	-3,3	2,3	
Разом	Count (Частота)	974	1982	2956
	Expected Count (Очікувана частота)	974,0	1982,0	2956,0
	% в 296 Ваш пол	100,0%	100,0%	100,0%

Що означають стандартизовані залишки? Величини залишків дозволяють судити про те, наскільки сильно спостережувані значення відрізняються від очікуваних, або які значення найбільше відхиляються від нульової гіпотези (якщо вона вірна, залишки повинні дорівнювати нулю).

Відсоток загального числа спостережень в рядку і наведені стандартизовані залишки використовуються для виявлення клітин таблиці, які найбільше відхиляються від нормального розподілу, що є підставою для виявлення статистичних залежностей. Значні стандартизовані залишки

дозволяють оцінити схильність чи відсутність схильності аналізованої групи респондентів думати або діяти певним чином.

Загальне правило для інтерпретації стандартизованих залишків:

Якщо значення стандартизованого залишку менше -2, то спостережувана частота у клітинці таблиці менше очікуваної частоти.

Якщо значення стандартизованого залишку більше +2 то спостережувана частота у клітинці таблиці більша за очікувану частоту.

Якщо значення стандартизованого залишку менше -2 або більше +2 , то існує значуща розбіжність між спостережуваною й очікуваною частотами на 5% рівні (тобто з ймовірністю 95%).

Інші граничні значення інтерпретуються відповідно до наступної таблиці, але при цьому треба пам'ятати, що ці правила можуть застосовуватися, тільки в тому випадку, якщо очікувані частоти не менше 5.

Таблиця 7.17

Відповідність значень стандартизованого залишку та рівня значущості

Абсолютні значення стандартизованого залишку	Рівень значущості	Інтерпретація
≥ 2	$\leq 0,05$	Розбіжності статистично значущі на 5% рівні
$\geq 2,6$	$\leq 0,01$	Розбіжності статистично значущі на 1% рівні
$\geq 3,3$	$\leq 0,001$	Розбіжності статистично значущі на 0,1% рівні

Література до теми

Бююль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2002. С. 221-255, 323-340.

Крыштановский, А.О. *Анализ социологических данных с помощью пакета SPSS*. М.: Изд.дом ГУ ВШЭ, 2007. С.82-114.

Наследов А. *SPSS: компьютерный анализ в психологии и социальных науках*. Питер, 2005. С. 137-236.

Паніотто, В.І., Максименко, В.С., Харченко, Н.М. *Статистичний аналіз соціологічних даних*. – К.: «КМ Академія», 2004. С.195-217.

Толстова, Ю.Н. *Математико-статистические модели в социологии (математическая статистика для социологов) : учеб. Пособие*. М.: ГУ-ВШЭ, 2008. С. 113-120.

Додаткова література

Daniel Stockemer. *Quantitative Methods for the Social Sciences. A Practical Introduction with Examples in SPSS and Stata*. Springer International Publishing AG 2019. P. 101-124.

Венгер, Г.С. Порівняльний аналіз особливостей якості сімейного життя в традиційних і дистантних сім'ях. *Збірник наукових праць К-ПНУ імені Івана Огієнка, Інституту психології імені Г.С.Костюка НАПН України*.

Головаха, Є., Горбачик, А., Паніна, Н. *Україна та Європа: результати міжнародного порівняльного соціологічного дослідження*. Київ : Інститут соціології НАН України, 2006. URL: https://i-soc.com.ua/assets/files/library/european_social_survey_ua.pdf

Руденко, В.М. *Математична статистика*. К.: Центр учбової літератури, 2012.

Савельєв, Ю. Б. Аналіз порівняльний у соціології. *Велика українська енциклопедія*. 2019. URL: <https://vue.gov.ua/>

Толстова, Ю. Н., Куликова, А. А., Рыжова, А. В., Юдин, Б. Г. *Математическая статистика для социологов: задачник : учеб. Пособие для вузов* М.: Изд. дом Гос. ун-та — Высшей школы экономики, 2010.

Толстова, Ю.Н. *Анализ социологических данных (Методология, дескриптивная статистика, изучение связей между номинальными признаками)*. М.: Научный мир, 2003. URL: <http://www.ecsocman.edu.ru/images/pubs/2003/05/12/0000086337/>.

Чуприна, О. О., Назарова, О. Ю. Аналіз розбіжностей в оцінках бідності та соціальної нерівності в Україні // *Вісник Харківського національного університету. Серія «Економічна»*. 2014. № 1096. С. 118–124.

Питання для самоконтролю

1. Що таке «статистичний аналіз розбіжностей»?
2. Коли використовується аналіз розбіжностей?
3. Які є процедури аналізу розходжень?
4. Які є процедури для перевірки статистичної значущості розбіжностей?
5. Чим визначається вибір методу перевірки статистичної значущості розходжень?
6. Які методи перевірки статистичної значущості розходжень ви можете назвати?
7. Чим відрізняються параметричні та непараметричні тести?
8. Які параметричні тести Ви знаєте? В яких випадках вони використовуються?
9. Що таке «дисперсійний аналіз»? Коли він застосовується?
10. Які непараметричні тести Ви знаєте? В яких випадках вони використовуються?
11. Навіщо звертатися до розрахунку статистичної значущості під час застосування аналізу розбіжностей відсотків?

12. Які існують методи перевірки значущості розбіжностей відсотків (часток) в пакеті SPSS?

Тема 8. Факторний аналіз

8.1. Сутність та формальна модель факторного аналізу

Факторний аналіз – низка методів багатовимірного статистичного аналізу, призначених для дослідження кореляційних зв'язків, що має метою виявлення латентних факторів, які детермінують значення спостережуваних ознак. При цьому під *латентним фактором* у контексті факторного аналізу називають неспостережувану (латентну) змінну, що зумовлює кореляції між наявними (вимірними) змінними. Кожний фактор поєднує змінні, які тісно корелюють між собою, та інтерпретується як чинник взаємозв'язку цих змінних.

Латентний (від лат. *Latent*) – прихований. *Латентна змінна* – це змінна, значення якої в ході спостереження не доступні для безпосереднього вимірювання, а можуть бути оцінені відповідно до висунутої гіпотези за допомогою значень наявних змінних, які були виміряні.

Фактор (від лат. *Factor*) – чинник, рушійна сила будь-якого процесу, що визначає характер чи окремі риси процесу. У факторному аналізі **фактор** розуміється як сконструйована математичними засобами латентна змінна, яка тісно корелює з певною групою наявних (вимірних) змінних.

Факторний аналіз дозволяє звести велику кількість змінних, які закладені в інструментарій соціологічного дослідження, до значно меншої кількості факторів, які інтерпретуються з урахуванням як сутності формальної моделі факторного аналізу, так і специфіки феномену, що вивчається.

Застосування факторного аналізу обмежується певними вимогами до вихідних даних. Насамперед, змінні мають бути кількісними, оскільки факторний аналіз ґрунтується на обчисленні коефіцієнтів кореляції Пірсона. Крім того модель факторного аналізу передбачає, що спостереження мають бути незалежними та повинна виконуватись умова нормальності їх багатовимірного розподілу.

Головна ідея факторного аналізу полягає в тому, що кореляційні зв'язки між великою кількістю змінних визначаються існуванням меншого числа гіпотетичних (неспостережуваних) змінних (факторів). Отже, соціолог може перейти від аналізу численних емпіричних показників до вивчення глибинних, прихованих факторів, які часто взагалі не підлягають вимірюванню.

Загальною моделлю факторного аналізу є наступна лінійна модель:

$$X_i = \sum_{j=1}^k a_{ij} f_j + U_i, \quad i = 1, \dots, n, \quad (1)$$

де X_i – значення i -ої змінної, що представлено у вигляді лінійної комбінації k загальних факторів;

f_j – загальні фактори;

U_i – характерні фактори, що є унікальними (специфічними) для кожної ознаки X_i ;

a_{ij} – факторні навантаження, які є регресійними коефіцієнтами, що показують «внесок» кожного з k факторів у сконструйовану змінну (фактор).

Передбачається, що $k < n$ (k – кількість латентних факторів, n – кількість змінних, що підлягають факторному аналізу).

Фактори f_j конструюють таким чином, щоб найкраще (з мінімальною похибкою) представити X_i . Їх прийнято розділяти на загальні (f_j) і характерні (U_i). Фактори f_j називають загальними, оскільки у формальній моделі всі змінні X_i представлені як їхні функції, тобто загальні фактори впливають на всі досліджувані змінні. Відмінність характерних факторів від загальних полягає в тому, що кожний характерний фактор є унікальним для кожної змінної X_i та впливає лише на неї (має ненульове значення лише для однієї ознаки спостереження).

Засадою класичного факторного аналізу є твердження, що X_i стандартизовані (середнє значення кожної ознаки дорівнює нулю, тобто $\bar{X}_i = 0$, а дисперсія – одиниці, $\sigma_i = 1$), фактори f_j є незалежними та не пов'язані зі специфічними факторами U_i (хоча існують моделі, що спираються на інші припущення). Це означає, що фактори f_j також є стандартизованими.

Отже, факторні навантаження a_{ij} збігаються з коефіцієнтами кореляції між загальними факторами та змінними X_i . Дисперсія X_i є сумою квадратів факторних навантажень і дисперсії специфічного фактору:

$$S_{x_i}^2 = H_i^2 + S_{u_i}^2, \text{ де } H_i^2 = \sum_k a_{ik}^2,$$

де H_i^2 – спільність, $S_{u_i}^2$ – специфічність. Спільність є частиною дисперсії змінних, що пояснюється факторами; специфічність – частина дисперсії, що не пояснюється факторами.

Відповідно до постановки завдання, факторний аналіз має визначати фактори, коли сумарна спільність максимальна, а специфічність – мінімальна.

Звичайні припущення, що дозволяють додати моделі факторного аналізу статистичний зміст, полягають у наступному: фактори є випадковими величинами, розподілені за нормальним законом та задані в стандартній формі; характерні фактори незалежні між собою і відносно загальних факторів. Так, з'являється можливість визначення факторних навантажень вихідних ознак, використовуючи різного роду статистичні процедури. Послугуючись значенням факторних навантажень, можна здійснити змістовну інтерпретацію та дати назву виділеним факторам.

Рівняння (1) свідчить, що кожна змінна може бути представлена у вигляді суми внесків кожного із загальних факторів. З іншого боку, кожний з k факторів може бути виражений у вигляді лінійної комбінації спостережуваних змінних:

$$F_j = W_{j,1} * V_1 + W_{j,2} * V_2 + \dots + W_{j,p} * V_n, \quad (2)$$

де $W_{j,i}$ – навантаження j -го фактора на i -ю змінну;
 n - кількість змінних.

Значення факторних навантажень можуть бути розраховані для кожного окремого спостереження, що дає можливість застосовувати фактори у якості нових змінних в подальшому аналізі даних.

Факторне навантаження – це коефіцієнт кореляції i -тої змінної та j -того фактору, що презентує якою мірою ця i -та змінна пов'язана з j -тим фактором. Факторні навантаження приймають значення в діапазоні від -1 до 1 .

Матриця факторних навантажень є головним результатом виконання всіх розрахунків факторного аналізу. Саме ця матриця дає можливість соціологу виявити латентні фактори, що зумовлюють досліджуване явище, проінтерпретувати їх та зробити змістовні висновки щодо сутності досліджуваного феномену.

Головні цілі застосування факторного аналізу:

- 1) зменшення кількості змінних (редукція даних, зниження розмірності простору ознак);
- 2) опосередковане, непряме оцінювання ознак, що є предметом соціологічного аналізу, у випадку неможливості або незручності їхнього прямого виміру;
- 3) визначення структури взаємозв'язків між змінними (експлораторний факторний аналіз);
- 4) генерування нових ідей на основі знайденої структури взаємозв'язків (експлораторний факторний аналіз);
- 5) перевірка (підтвердження чи спростування) теоретичних моделей факторного типу шляхом оцінки відповідності емпіричних даних теорії, яка перевіряється (конфірматорний факторний аналіз);
- 6) структурування та компактна візуалізація емпіричних даних.

У практиці соціологічних досліджень найчастіше застосовують експлораторний факторний аналіз, який можна застосовувати навіть на ранніх стадіях дослідження складних феноменів, коли ще не повністю склалися теоретичні уявлення щодо їхньої специфіки. Проте несправедливо оминати увагою можливість конфірматорного підходу, який може принести плідні результати в контексті перевірки існуючих теорій.

Експлораторний (розвідницький) факторний аналіз здійснюється у дослідженнях прихованої факторної структури без певного припущення про кількість факторів та їх навантаження; він має метою виявити щось цікаве, непрогнозоване дослідником.

Конфірматорний (підтверджувальний) факторний аналіз припускає наявність чітко сформульованої факторної моделі досліджуваного явища,

він призначений для перевірки гіпотез про кількість факторів та їх навантаження. Факторна модель, що пов'язує спостережувані та латентні змінні, конструюється на засадах знань предметної області, а гіпотези про структуру моделі ґрунтуються на аналізі природи досліджуваних факторів з урахуванням теоретичних здобутків та емпіричних даних.

Методи побудови факторів є методами факторизації, тобто методами розкладання на прості множники. До них відносять такі методи: метод головних компонент, головних факторів, канонічну факторизацію Рао, факторизацію образів, альфа-факторизацію, а також незважену й зважену факторизацію за методом найменших квадратів. Універсального пакета програм, у якому були б реалізовані всі методи факторизації, не існує. Пакет ОСА використовує метод головних компонент, пакет SPSS – метод головних компонент, незважений МНК (метод найменших квадратів), узагальнений (зважений) МНК, метод максимальної правдоподібності, метод факторизації головної осі, метод альфа-факторизації та аналіз образів. Найпопулярнішим у соціологічній практиці є метод головних компонент, який націлений на пояснення якомога більшої частини загальної дисперсії результативної ознаки користуючись мінімальною кількістю змінних.

Етапи проведення факторного аналізу:

1. На першому етапі необхідно переконатися, що дані, які підлягатимуть факторизації, не тільки є кількісними (метричними чи інтервальними), а й відповідають формальній моделі факторного аналізу, тобто дають можливість замінити групи значно пов'язаних змінних відповідними факторами. Якщо між досліджуваними змінними немає кореляцій, то факторний аналіз проводити безглуздо.

Для перевірки адекватності емпіричних даних моделі факторного аналізу існує кілька статистик, зокрема в SPSS можна застосувати критерії адекватності вибірки Кайзера-Мейера-Олкина (КМО) та сферичності Бартлетта. Ці критерії дозволяють переконатися, що досліджувані

Критерій Кайзера-Мейера-Олкина (Kaiser-Meyer-Olkin Measure of Sampling Adequacy, КМО) досліджує приватні коефіцієнти кореляції та дозволяє переконатися, що наявні змінні пов'язані між собою.

Критерій КМО перевіряє статистичну гіпотезу про те, що приватні кореляції між змінними відсутні ($H_0: r=0$, кореляції відсутні; $H_1: r \neq 0$, кореляції є). Якщо значення КМО-статистики не перевищує 0,5, нульова гіпотеза приймається та робиться висновок про відсутність кореляцій у генеральній сукупності, тобто про неадекватність емпіричних даних факторній моделі. Високі значення критерію КМО свідчать про адекватність даних моделі факторного аналізу. Прийнято характеризувати придатність даних наступним чином:

- значення критерію КМО, більші ніж 0,9 – безумовна адекватність;
- більші ніж 0,8 – висока адекватність;
- більші ніж 0,7 – прийнятна адекватність;
- більші ніж 0,6 – задовільна адекватність;
- більші ніж 0,5 – низька адекватність;

- менші ніж 0,5 – факторний аналіз не слід застосовувати.

Bartlett's Test of Sphericity (критерій сферичності Бартлетта) перевіряє гіпотезу, що кореляційна матриця сукупності - це одинична матриця, в якій всі діагональні елементи рівні 1, а всі інші рівні 0. Перевірка заснована на перетворенні детермінанта кореляційної матриці в статистику хі-квадрат. При великому значенні статистики нульову гіпотезу відхиляють. Якщо ж нульову гіпотезу не відхиляють, то доцільність виконання факторного аналізу викликає сумніви. Значення р-рівня, що є меншими 0,05, вказують на прийнятність (адекватність) даних для проведення факторного аналізу.

2. На другому етапі обирається метод факторизації (частіш за все метод головних компонентів) та вирішується проблема кількості факторів. Найбільш поширеним способом визначення мінімальної кількості факторів для адекватного пояснення спостережуваних кореляцій між первинними змінними є застосування критерію Кайзера, який заснований на припущенні, що слід вилучати та інтерпретувати всі фактори, власні значення яких перевищують одиницю. Власне значення являє собою дисперсію, яка обумовлена дією одного фактору. Коли фактор не виділяє дисперсію, еквівалентну, принаймні, дисперсії однієї змінної, він не може зацікавити дослідника, який шукає латентні змінні, що значніше інформативніші, ніж окремі ознаки, які були виміряні.

3. Третій етап – обертання факторної структури, тобто поворот факторних осей таким чином, щоб досягти найкращої інтерпретації знайдених факторів. Існують різні методи обертання факторних осей – варимакс, кватримакс, еквимакс, промакс тощо. На практиці найчастіше використовується метод варимакс, який дає ортогональне рішення з простою структурою та істотно полегшує інтерпретацію факторів.

4. Четвертий етап – виконання всіх розрахунків з урахуванням потреб подальшого аналізу.

5. П'ятий етап полягає у змістовній інтерпретації матриці факторних навантажень, що завершується створенням конструктів, які відповідають кожному з побудованих факторів.

8.2. Порядок виконання процедури факторного аналізу на прикладі пошуку чинників моральних преференцій

Сьогодні факторний аналіз здійснюється за допомогою сучасного програмного забезпечення. Саме тому ми розглянули сутність факторного аналізу дуже стисло, не акцентуючи уваги на математичному формалізмі. Всі необхідні розрахунки можна виконати в будь-якому пакеті обробки статистичних даних. Проте проінтерпретувати отримані результати необхідно самому соціологу, ніяка комп'ютерна програма не допоможе. Саме тому ми пропонуємо розглянути застосування методу факторного аналізу на основі розрахунків SPSS, звертаючи особливу увагу на інтерпретацію

статистичних показників, які необхідні для отримання дослідницьких висновків.

В якості прикладу візьмемо завдання пошуку латентних факторів, що зумовлюють існування певних моральних феноменів у середовищі сучасного українського студентства. Для цього використаємо масив st06.sav, зокрема, відповіді респондентів на питання про ставлення й ступінь згоди з дванадцятьма парними твердженнями, що охоплюють основні сфери людського життя (в анкеті вони представлені афоризмами, біблейськими максимами та прислів'ями).

Для вимірювання ступеню згоди з твердженнями, що репрезентують певні моральні принципи окремого респондента, в анкеті було використано спеціальний блок запитань (див. табл. 8.1, ознаки 131-142). Шкали мають 7 градацій: цифрами 7, 6 та 5 закодована ступінь прихильності до змісту певного принципу соціальної взаємодії (7 – максимальна прихильність), 4 – нейтральна позиція, 3, 2, 1 – ступені прихильності до протилежного змісту того ж принципу (максимальна прихильність закодована цифрою 1).

Таблиця 8.1

Блок питань анкети для вимірювання ступеня згоди з висловленнями, що репрезентують певні моральні принципи

<u>З ЯКИМ ІЗ НАВЕДЕНИХ У КОЖНІЙ ПАРІ ТВЕРДЖЕНЬ ВИ ЗГОДНІ Й ЯКОЮ МІРОЮ?</u>		
131. Око за око, зуб за зуб	7_6_5_ 4 _3_2_1	Хто вдарить тебе в праву щоку, оберни до нього й ліву
132. Не обдуриш - не проживеш	7_6_5_ 4 _3_2_1	Краще бути бідняком, ніж жити з гріхом
133. Варварство знищується варварством	7_6_5_ 4 _3_2_1	Любить ворогів Ваших
134. Кожен сам за себе	7_6_5_ 4 _3_2_1	Один за всіх, всі за одного
135. Людина людині вовк	7_6_5_ 4 _3_2_1	Людина людині друг, товариш і брат
136. Дурень вважає, що краса мир спасає	7_6_5_ 4 _3_2_1	Краса врятує світ
137. При многості мудрості, множиться й клопіт	7_6_5_ 4 _3_2_1	Краще більше знати, ніж багато мати
138. Усі друзі хороші, коли в людини є гроші	7_6_5_ 4 _3_2_1	Май не 100 рублів, а май 100 друзів
139. Багатому й у пеклі рай	7_6_5_ 4 _3_2_1	Верблюдові легше пройти через голчине вушко, ніж багатому в Боже Царство ввійти
140. Дурний осудить, розумний розсудить	7_6_5_ 4 _3_2_1	Не судить, щоб й Вас не судили
141. Любов і розумників на дурні обертає	7_6_5_ 4 _3_2_1	Лише закоханий має право на звання людини

Таким чином, *вихідними даними для факторного аналізу є матриця розміру $n \times t = 3057 \times 12$ ($n = 3057$ – це кількість анкет у масиві даних, $t = 12$*

– кількість тверджень, ступінь згоди з якими дає можливість дослідження моральних орієнтацій респондентів). Гіпотеза, що має бути перевірена, полягає у припущенні про існування латентної структури моральних орієнтацій студентської молоді, що не може бути виявлена безпосередньо, але вона побічно впливає на соціальну поведінку українських студентів. Метою застосування факторного аналізу в даному випадку є пошук латентних факторів, що детермінують моральні орієнтації українського студентства. Тут треба підкреслити, що ми застосовуємо експлораторний факторний аналіз, оскільки не знаємо ні кількість факторів, ні їхній зміст.

Розглянемо процедуру виконання факторного аналізу.

1. Перевірка адекватності даних передбачає аналіз рівня вимірювання досліджуваних ознак та використання критеріїв Кайзера-Мейєра-Олкіна (КМО) та сферичності Бартлетта для перевірки адекватності вибірки.

Шкали, які застосовувались для вимірювання моральних орієнтацій, є інтервальними, що відповідає моделі факторного аналізу.

Для перевірки можливості застосування факторного аналізу до наявних даних використовують критерії адекватності вибірки Кайзера-Мейєра-Олкіна (КМО) та сферичності Бартлетта.

Ми рекомендуємо завжди починати факторний аналіз з перевірки можливості його застосування. Для цього в SPSS слід виконати такі дії.

Оберіть у меню *Analyze (Аналіз) → Data Reduction (Редукація обсягу даних) → Factor... (Факторний аналіз)*. Відкриється діалогове вікно *Factor Analysis (Факторний аналіз)* (див. рис. 8.1), в якому слід зазначити змінні, до яких планується застосувати процедуру факторизації, та натиснути кнопку *Descriptives*. У результаті на екрані з'явиться діалогове вікно *Descriptives (Дескриптивні статистики)*, в якому необхідно встановити прапорець навпроти пункту *KMO and Bartlett's Test of Sphericity* (див. рис. 8.2). Саме ці тести дають змогу виявити придатність вихідних емпіричних даних факторному аналізу.

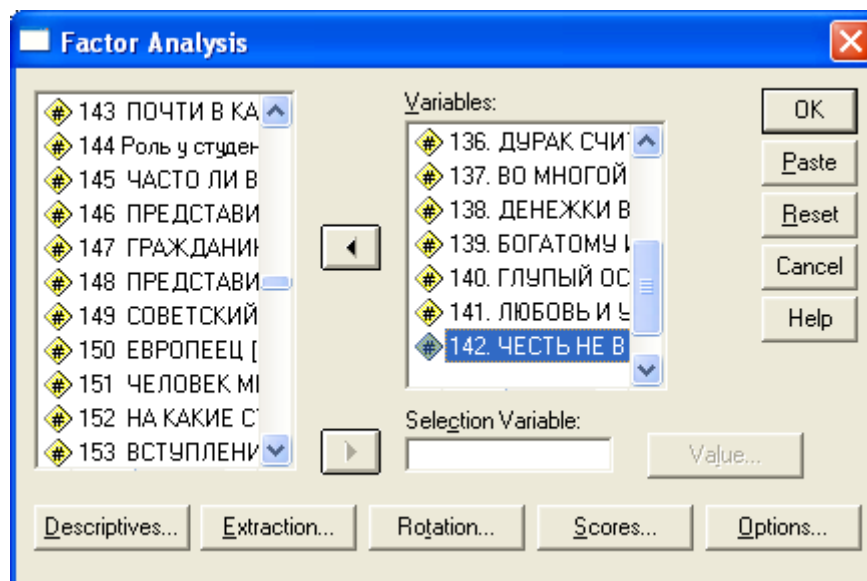


Рис. 8.1. Діалогове вікно *Factor Analysis (Факторний аналіз)*

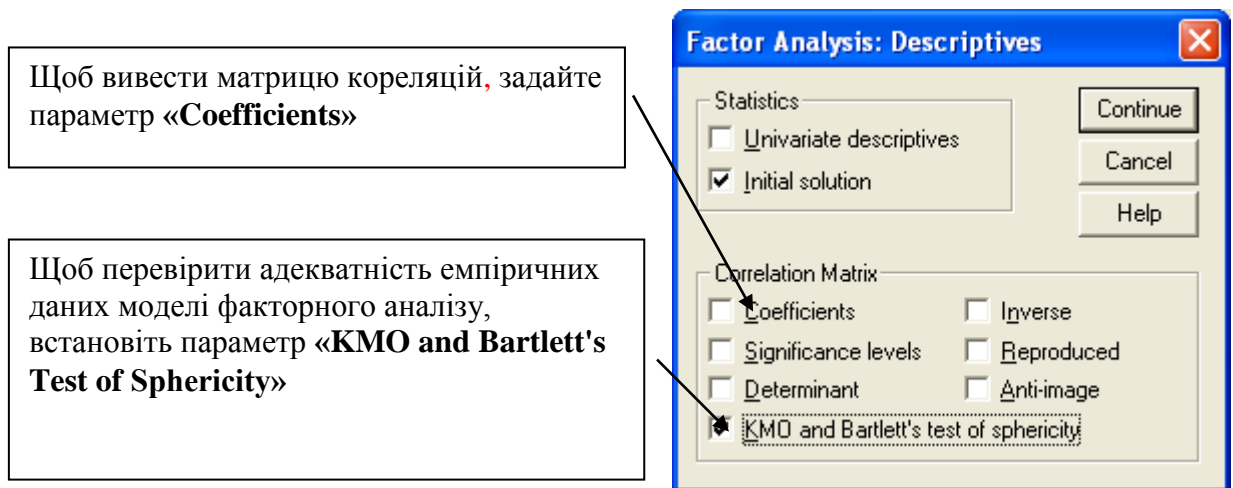


Рис. 8.2. Діалогове вікно *Factor Analysis: Descriptives* (Факторний аналіз: Дескриптивні статистики)

У нашому прикладі результати тестів КМО та Бартлетта засвідчили високу адекватність даних факторній моделі (див. табл. 8.2), що дозволяє проводити аналіз далі, інтерпретувати результати та робити змістовні висновки.

Таблиця 8.2
Загальний вигляд таблиці, що представляє результати тестів КМО і сферичності Бартлетта (KMO and Bartlett's Test)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,844
Bartlett's Test of Sphericity	Approx. Chi-Square	5126,788
	df	66
	Sig.	,000

Вихідним елементом для подальших розрахунків факторного аналізу є кореляційна матриця. Для її створення SPSS спочатку стандартизує значення змінних (z-перетворення); потім за допомогою стандартизованих значень розраховуються кореляційні коефіцієнти Пірсона між досліджуваними змінними. Для створеної кореляційної матриці визначаються власні значення та відповідні їм власні вектори, для визначення яких використовуються оцінні значення діагональних елементів матриці (відносні дисперсії простих факторів). Цю операцію SPSS виконує без вказівок соціолога-аналітика, оскільки вона є обов'язковою для всіх видів факторного аналізу. Без неї неможливі подальші кроки факторизації досліджуваних змінних.

Матрицю кореляцій можна вивести як один з результатів факторного аналізу (див. табл. 8.3), проте вона зазвичай не застосовується при змістовній

інтерпретації. Щоб розрахувати матрицю кореляцій та вивести її у файл результатів необхідно у діалоговому вікні *Descriptives (Дескриптивні статистики)* встановити прапорець навпроти пункту *Coefficients* (див. рис. 8.2).

Таблиця 8.3

Матриця кореляцій

	131	132	133	134	135	136	137	138	139	140	141	142
131	1	0,353	0,422	0,18	0,191	0,132	0,215	0,103	0,212	0,169	0,187	0,179
132	0,353	1	0,308	0,253	0,299	0,172	0,344	0,188	0,307	0,183	0,142	0,3
133	0,422	0,308	1	0,223	0,297	0,177	0,227	0,133	0,257	0,203	0,168	0,175
134	0,18	0,253	0,223	1	0,488	0,21	0,252	0,324	0,219	0,119	0,144	0,225
135	0,191	0,299	0,297	0,488	1	0,296	0,287	0,372	0,28	0,206	0,201	0,211
136	0,132	0,172	0,177	0,21	0,296	1	0,249	0,204	0,112	0,154	0,231	0,167
137	0,215	0,344	0,227	0,252	0,287	0,249	1	0,228	0,26	0,174	0,236	0,239
138	0,103	0,188	0,133	0,324	0,372	0,204	0,228	1	0,248	0,177	0,135	0,22
139	0,212	0,307	0,257	0,219	0,28	0,112	0,26	0,248	1	0,226	0,136	0,197
140	0,169	0,183	0,203	0,119	0,206	0,154	0,174	0,177	0,226	1	0,176	0,191
141	0,187	0,142	0,168	0,144	0,201	0,231	0,236	0,135	0,136	0,176	1	0,255
142	0,179	0,3	0,175	0,225	0,211	0,167	0,239	0,22	0,197	0,191	0,255	1

2. Вирішення проблеми кількості факторів. В ході проведення факторного аналізу соціологу необхідно вирішити скільки факторів вилучати з досліджуваних даних. Відзначимо, що в процесі послідовного виділення факторів методом факторного аналізу кожний вилучений фактор містить у собі меншу мінливість, ніж попередній. Рішення про те, коли варто зупинити процедуру вилучення факторів, є довільним, однак існують певні рекомендації. Вони дозволяють раціонально обрати кількість факторів, які потім будуть проінтерпретовані соціологом.

Найбільш поширеним способом є застосування *критерію Кайзера (критерій власних значень)*, який базується на припущенні, що слід вилучати (а потім й інтерпретувати) всі фактори, власні значення яких перевищують 1. Власне значення є дисперсією, що зумовлена дією одного фактору. Цей показник ще називають інформативністю фактору. Коли фактор не вилучає дисперсію, еквівалентну дисперсії принаймні однієї змінної, він не може зацікавити дослідника.

Щоб зрозуміти, що означає відсоток дисперсії кожного фактору, слід мати на увазі наступні міркування. В нашому прикладі беруть участь 12 змінних, кожна з яких має дисперсію рівну 1. Це означає, що найбільша мінливість, яка потенційно може бути вилучена, дорівнює 12 разів по 1. Відсоток дисперсії, що пояснюється однією змінною, розраховується таким чином: $100/12=8,333$. Кожен з виділених факторів повинен пояснювати відсоток дисперсії, більший цього значення.

Інформативність фактору – дисперсія, що пояснюється цим фактором. Інформативність фактору може бути виражена власним значенням або у відсотках.

Слід звернути увагу на ще один важливий показник – кумулятивну (накоплену, сумарну) дисперсію вилучених факторів. Цей показник можна інтерпретувати, як частину досліджуваного феномену (в нашому прикладі – моральних орієнтацій студентської молоді), що описується всіма факторами. У нашому прикладі цей показник є дуже низьким (47,642%). У соціальних дослідженнях рекомендується виділяти таку кількість факторів, що пояснює принаймні 60% дисперсії. Але в окремих випадках допускається перехід вимірних ознак до факторів (які описують лише половину досліджуваного соціального явища), якщо це зумовлено необхідністю наявного представлення отриманих результатів. За наявності приходиться «сплачувати» втрату частини інформації, що міститься в емпіричних даних.

Кумулятивна (сумарна) дисперсія всіх вилучених факторів – частка варіації, що пояснена цими факторами. Бажано, щоб цей показник перевищував 60%.

У таблиці 8.4 «*Total Variance Explained*» (Сумарна дисперсія, що пояснюється факторним аналізом) можна побачити, що три фактори мають власні значення, які перевищують одиницю. Отже, за критерієм Кайзера слід обирати три фактори для аналізу. Перший фактор пояснює 29,168% сумарної дисперсії, другий фактор – 9,890% і третій фактор – 8,584%.

Таблиця 8.4

**Сумарна дисперсія, що пояснюється факторним аналізом
(Total Variance Explained)**

Фактори	Первинні власні значення (Initial Eigenvalues)			Сума квадратів після обертання (Rotation Sums of Squared Loadings)		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,500	29,168	29,168	2,023	16,857	16,857
2	1,187	9,890	39,058	1,971	16,424	33,281
3	1,030	8,584	47,642	1,723	14,361	47,642
4	,928	7,729	55,371			
5	,873	7,272	62,643			
6	,793	6,607	69,250			
7	,735	6,125	75,375			
8	,680	5,670	81,045			
9	,660	5,497	86,542			
10	,597	4,973	91,515			
11	,549	4,579	96,094			
12	,469	3,906	100,000			

Критерій кам'янистого осипу або *критерій Кетела* є графічним методом, який також може застосовуватися для вирішення завдання щодо вилучення кількості факторів, які слід виділяти та інтерпретувати у факторному аналізі. Діаграма «кам'янистого осипу» відображає власні значення, які представлені в таблиці 8.4, у вигляді простого графіка (див. рис. 8.3).

У SPSS діаграма «кам'янистого осипу» має назву «Scree Plot», яка складається з двох частин: англійського слова «scree», що означає «щебінь» і слова «plot», що в англійському – «графічне зображення». Така діаграма використовується для того, щоб найменш значущі фактори – «щебінь» – можна було відокремити від найбільш значущих. Ці значущі фактори на графіку утворюють свого роду схил, тобто ту частину лінії, яка характеризується крутим підйомом. У діаграмі такий крутий підйом спостерігається в області перших трьох факторів. Якщо подивитися на графік, можна помітити, що схил, тобто область значущих факторів, спостерігається вище третього фактору, а нижче цього фактору розташувався «щебінь», тобто область незначущих факторів.

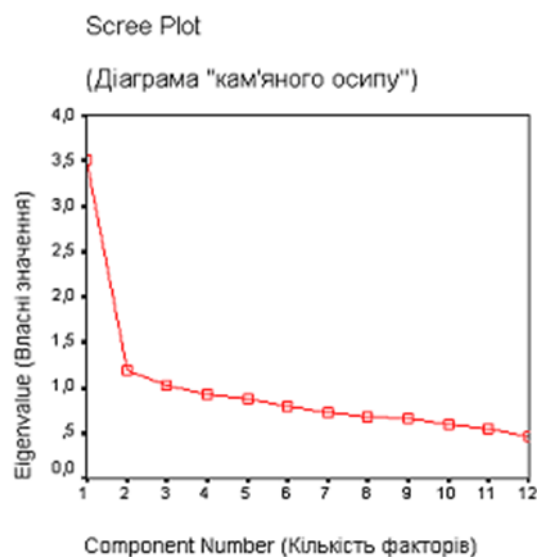


Рис. 8.3. Діаграма «кам'янистого осипу», що показує кількість факторів, які підлягають інтерпретації

У конфірматорних дослідження можна застосовувати *критерій визначення кількості факторів*, що заснований на наявності *попередньої інформації*. Цей критерій передбачає, що у дослідника вже є інформація про те, скільки факторів необхідно виділяти. Такою інформацією можуть виступати розроблена теорія факторного типу та попередні емпіричні дослідження. Але в цьому випадку все одне рекомендується звертати увагу на статистичні показники, що представлені у таблиці 8.3, оскільки апріорні уявлення соціолога можуть не повною мірою відповідати дійсності.

Для установки параметрів вилучення факторів у SPSS необхідно у діалоговому вікні *Factor Analysis (Факторний аналіз)* (див. рис. 8.1) натиснути на кнопку *Extraction*, після чого на екрані з'явиться вікно *Factor*

Analysis: Extraction (Факторний аналіз: Вилучення факторів) (див. рис. 8.4).

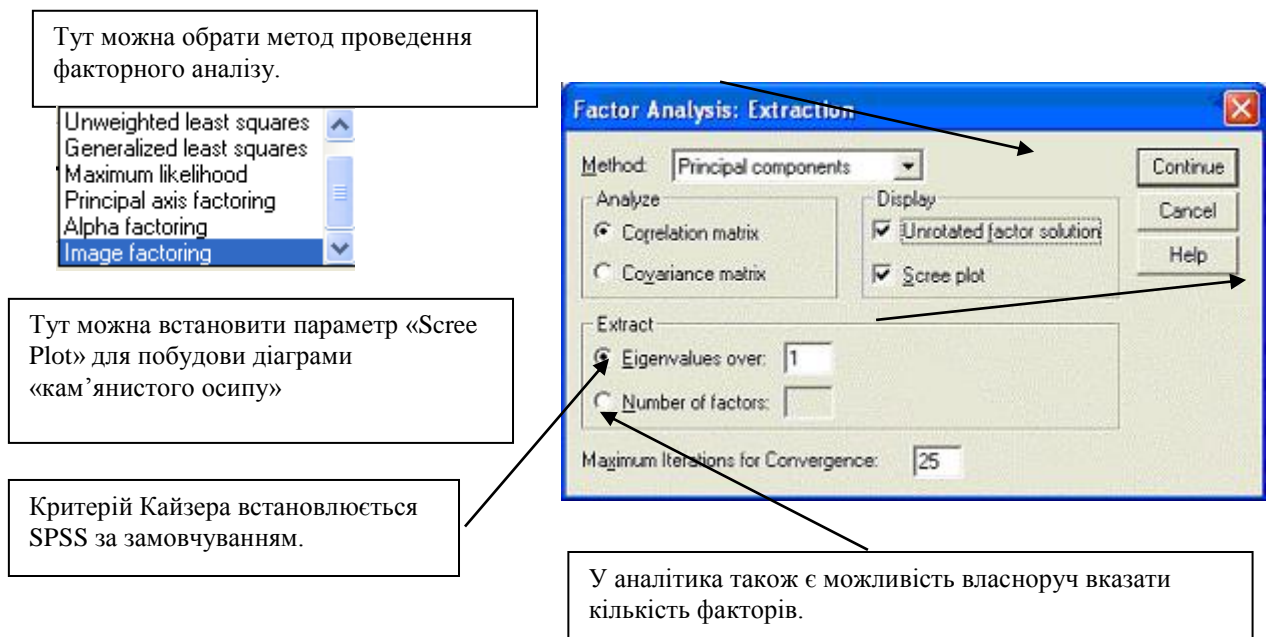


Рис. 8.4. Діалогове вікно *Factor Analysis: Extraction* (Факторний аналіз: Вилучення факторів)

3. Операції обертання факторних осей дозволяють так розташувати фактори, щоб кожний з них містив лише невелику кількість змінних, що мають високі навантаження. Ця операція дає можливість повернути факторні осі таким чином, щоб досягти найкращої інтерпретації вилучених факторів за рахунок виокремлення факторів, що відзначені високими факторними навантаженнями для деяких змінних і низькими – для всіх інших.

Для того, щоб реалізувати обертання факторних осей в SPSS треба натиснути кнопку **Rotation... (Обертання)**, яка дозволяє обрати метод обертання (див. рис. 8.5). Рекомендується активувати метод варімакс та залишати активованим параметр виводу поверненої матриці факторів. Крім того, можна організувати вивід факторних навантажень у графічному вигляді, у якому перші три фактори будуть представлені в тривимірному просторі. Якщо є лише два фактори, то буде побудовано графічне зображення розташування досліджуваних змінних у двовимірному просторі.

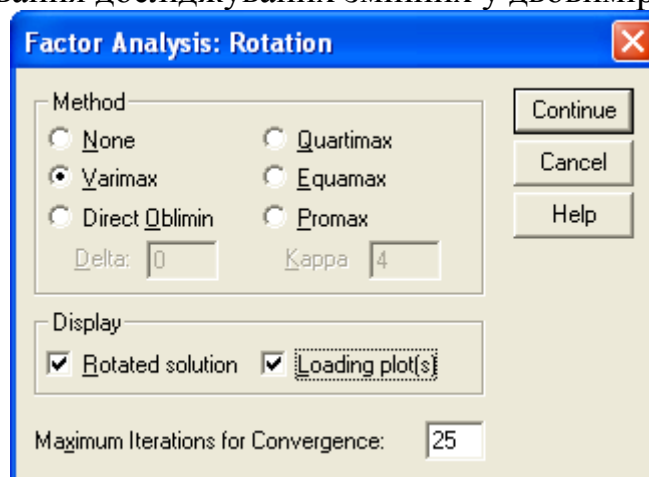


Рис. 8.5. Діалогове вікно *Factor Analysis: Rotation* (Факторний аналіз: Обертання)

Розглянемо методи обертання факторних осей, які зазвичай поділяють на ортогональні та косокутні. Ортогональним обертанням називають таке обертання факторних осей, що зберігає прямокутну систему координат. У результаті ортогонального обертання вилучають фактори, які не корелюють між собою. До ортогональних методів обертання факторних осей відносять варімакс, квартімакс і еквімакс.

Косокутне обертання не зберігає прямокутну систему координат. У результаті одержують фактори, що корелюють між собою. Інколи такі методи обертання бувають корисними, але найчастіше соціологи застосовують метод варімакс, який забезпечує зменшення кількості змінних, що пов'язані з кожним фактором. Він максимізує дисперсію факторних навантажень за кожним фактором і призводить до обертання факторних осей у центр скупчення точок, що зображують у факторному просторі досліджувані ознаки. У результаті такого обертання інтерпретація факторів, як правило, значно полегшується.

Метод квартімакс має тенденцію до виділення генерального фактору, що спрощує інтерпретацію за рахунок зменшення кількості факторів, пов'язаних з кожною змінною. Метод еквімакс дає проміжний ефект. Метод облімін реалізує косокутне обертання результатів варімакс-обертання. При цьому фактори розташовуються в просторі вихідних змінних не цілком перпендикулярно один до одного з обліком їх взаємної кореляції. У випадку незначного ступеню кореляції факторів, результати облімін-обертання майже не відрізняються від методу варімакс.

4. Розрахунок показників, необхідних для подальшого аналізу.

За допомогою кнопки **Options... (Опції)** Ви зможете організувати вивід у файл результатів коефіцієнтів, відсортованих за розміром. Для цього треба активувати опцію *Sorted by size* (Сортування за розміром). Коли невеликі значення факторних навантажень не виводяться на екран, це значно полегшує візуальне сприйняття результатів. Для цього слід активувати опцію *Suppress absolute values less than ...* та вказати, які саме значення не потрібно виводити (зазвичай не виводять значення, менші 0,4). Кнопка **Options... (Опції)** також призначена для обробки пропущених значень (тобто НеВідповідей). Тут забезпечуються такі можливості: вилучити з аналізу анкети, що містять НеВідповіді; замінити пропущені значення середніми значеннями відповідних змінних (див. рис. 8.6).

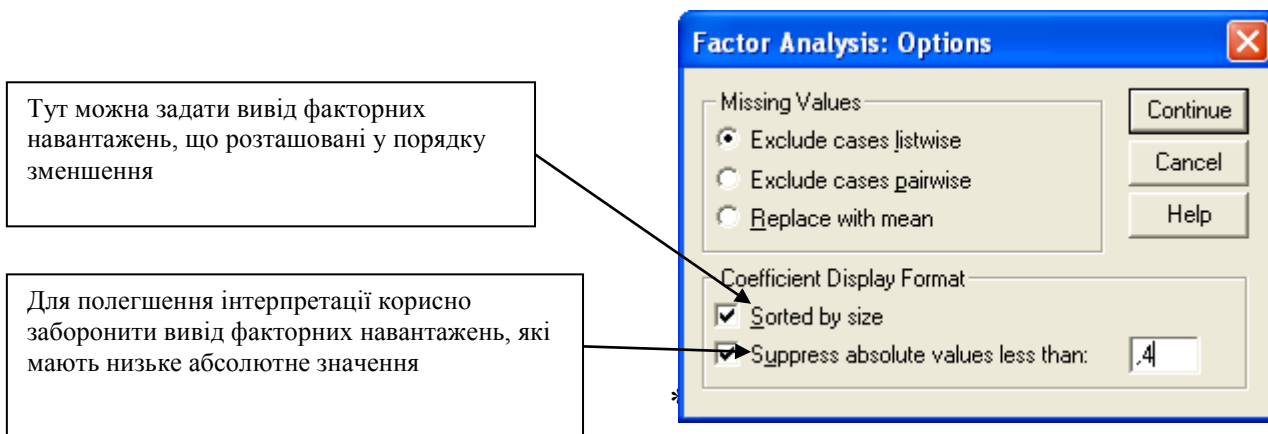


Рис. 8.6. Діалогове вікно *Factor Analysis: Options* (Факторний аналіз: Опції)

Щоб знайти значення факторів і зберегти їх у вигляді додаткових змінних застосуйте кнопку *Scores...* (Значення) і відзначте *Save as variables* (Зберегти як змінні). Автоматично SPSS здійснює регресійний метод створення нової змінної (див. рис. 8.7).

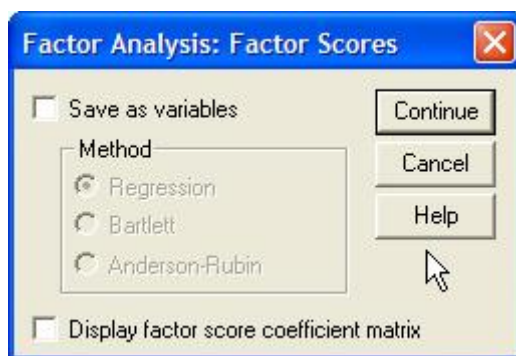


Рис. 8.7. Діалогове вікно *Factor Analysis: Factor Scores* (Факторний аналіз: Значення факторів)

Якщо ця опція буде встановлена, слід виконати команду збереження масиву даних, оскільки в нього будуть додані нові змінні-фактори. Назви цих змінних SPSS генерує таким чином: *fac1_1*, *fac2_1* і *fac3_1*. Ці змінні містять значення факторів, що обчислені для кожної анкети з вихідного масиву. Якщо Ви переглянете файл даних після проведення факторного аналізу та збереження масиву з новими змінними, то зможете побачити нормалізовані значення факторів, які, знаходяться у межах від -4 до +4.

5. Інтерпретація факторів здійснюється на основі аналізу матриці факторних навантажень (див. табл. 8.5).

Таблиця 8.5

Матриця факторних навантажень після обертання факторних осей

Моральні принципи	Фактори		
	1	2	3
131. «Око за око, зуб за зуб» - «Хто вдарить тебе в праву щоку, оберни до нього й ліву»	,764		
133. «Варварство викорінюється варварством» - «Любіть ворогів Ваших»	,715		
132. «Не обдуриш - не проживеш» - «Краще бути бідняком, ніж жити з гріхом»	,633		
139. «Багатому й у пеклі рай» - «Верблюдові легше пройти крізь голчине вушко, ніж багатому в Боже Царство ввійти»	,476		
134. «Кожен сам за себе» - «Один за всіх, всі за одного»		,737	
135. «Людина людині вовк» - «Людина людині друг, товариш і брат»		,725	
138. «Усі друзі хороші, коли в людини є гроші» - «Май не 100 рублів, а май 100 друзів»		,700	
141. «Любов і розумника на дурня обертає» - «Лише закоханий має право на звання людини»			,792
136. «Дурень вважає, що краса мир спасає» - «Краса врятує світ»			,574
142. «Честь не до честі, коли їсти нічого» - «Краще око втратити, ніж добре ім'я»			,530
137. «При многості мудрості, множиться й клопіт» - «Краще більше знати, ніж багато мати»			,415
140. «Дурний осудить, розумний розсудить» - «Не судить, щоб й Вас не судили»			,413

Матриця факторних навантажень містить коефіцієнти кореляції вимірних змінних з факторами. Інтерпретація кожного фактору проводиться за наступною схемою. Насамперед, навантаження, що належать до аналізованого фактору, розташовуються у порядку зменшення абсолютних значень. Після цього здійснюється відбір тих ознак, які мають максимальні абсолютні значення факторних навантажень. Останній крок є ключовим та найскладнішим - відібрана група ознак змістовно аналізується, виявляється загальна властивість, що їх поєднує, та вишукується назва, яка найкраще відображає сутність знайденої властивості.

Ми вилучили три фактори, кожний з яких є біполярним континуумом певної властивості, що включає всі ступені її прояву – від максимальної виразності до мінімальної.

Перший фактор, що одержав назву «толерантність», є континуумом ступенів прояву толерантності-ксенофобії в моральній сфері. Він увібрав у себе наступні ознаки:

- Око за око, зуб за зуб - Хто вдарить тебе в праву щоку, оберни до нього й ліву (факторне навантаження = 0,764);
- Варварство знищується варварством - Любить ворогів Ваших (0,715);
- Не обдуриш - не проживеш-Краще бути бідняком, ніж жити з гріхом (0,633);
- Багатому й у пеклі рай - Верблюдові легше пройти через голчине вушко, ніж багатому в Боже Царство ввійти (0,476).

Другий фактор містить ознаки, які є характеристиками індивідуалізму чи колективізму, тому його можна назвати колективізм. У тих випадках, коли значення змінної «колективізм» дорівнює нулю, це свідчить про абсолютний прояв індивідуалізму. Тобто фактор «колективізм» є континуумом прояву колективістських – індивідуалістичних орієнтацій моральних принципів сучасного українського студентства. Він утворений такими ознаками:

- Кожен сам за себе-Один за всіх, всі за одного (0,737);
- Людина людині вовк-Людина людині друг, товариш і брат (0,725);
- Усі друзі хороші, коли в людини є гроші-Май не 100 рублів, а май 100 друзів (0,700).

До третього фактору увійшли знаки, які можна розглядати як прояви духовності. Він також є континуумом, що представляє полярність «духовність- прагматичний матеріалізм»:

- Любов і розумників на дурні обертає-Лише закоханий має право на звання людини (0,792);
- Дурень вважає, що краса мир спасає-Краса врятує світ (0,574);
- Честь не до честі, коли їсти нічого-Краще око втратити, ніж добре ім'я (0,530);
- При многості мудрості, множиться й клопіт - Краще більше знати, ніж багато мати (0,415);
- Дурний осудить, розумний розсудить-Не судить, щоб й Вас не судили (0,413).

Якщо назви першого та другого факторів, на наш погляд, навряд чи можуть визвати дискусії, то назва третього фактору потребує пояснення смислу. Ми схильні у поняття «духовність» вкладати той смисл, що є протилежним крайньому прояву матеріалістичності, тобто прагматизму. Але це не той прагматизм, який призводить до конструктивних рішень, а, якщо можна так сказати, «примітивний, тваринний прагматизм». У цьому контексті має сенс згадати трактування феномену «духовність», що часто ототожнюється з прилученням до загальнолюдських цінностей, спрямованістю на задоволення безкорисливих духовних потреб (у знаннях, спілкуванні, естетичному задоволенні).

Література до теми

Бююль, А., Цёфель, П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2002. С. 368-383.

Наследов, А.Д. *SPSS: Компьютерный анализ данных в психологии и социальных науках*. СПб.: Питер, 2005. С. 280-297.

Крыштановский, А.О. *Анализ социологических данных с помощью пакета SPSS*. – 2-е изд. – М.: Изд. дом ГУ ВШЭ, 2007. – С. 191-205.

Паніотто, В.І., Максименко, В.С., Марченко, Н.М. *Статистичний аналіз соціологічних даних*. – К.: Вид.дім “КМ Академія”, 2004. С. 218-231.

Толстова, Ю.Н. *Измерение в социологии. Курс лекций*. М.: ИНФРА-М, 1998. С. 92-103.

Додаткова література

Бессокирная, Г.П. Факторный анализ: традиции использования и новые возможности. *Социология: методология, методы, математические модели*. 2000. № 12. С. 142-153.

Бірон, Б. В. Опитувальник внутрішньої мотивації. конфірмаційний факторний аналіз української версії. *Теоретичні і прикладні проблеми психології*. 2017. № 3. С. 52-65. URL: http://nbuv.gov.ua/UJRN/Tipprr_2017_3_7/

Дегтярев, Г.П. Факторный анализ в социологическом исследовании: вопросы интерпретации. *Комплексный подход к анализу данных в социологии*. М.: ИС АН СССР, 1989. С. 168 - 184.

Кислова, О. М., Ніколаєвська, А. М. Досвід застосування технології інтелектуального аналізу даних (ІАД) при вивченні моральних феноменів. *Вісник Одеського національного університету*. Том. 12. Випуск 6. Серія «Соціологія і політичні науки». 2007. С. 621-628.

Кислова, О. Н., Кузина, И. И. Методы многомерного анализа в исследовании детерминант моральной дифференциации украинских студентов. *Вісник Харківського національного університету імені В.Н. Каразіна “Соціологічні дослідження сучасного суспільства: методологія, теорія, методи”*. № 800. 2008. С. 92-99.

Судін, Д. Ю. Факторний аналіз неметричних даних: евристичний потенціал категоріального аналізу головних компонент. *Вісник ХНУ імені В. Н. Каразіна. Серія «Соціологічні дослідження сучасного суспільства: методологія, теорія, методи»*. 2012. № 999. С. 84-90. URL: <https://periodicals.karazin.ua/ssms/article/view/13924>.

Харман, Г. *Современный факторный анализ*. М.: Статистика, 1972.

Питання для самоконтролю

1. Які дані застосовують у факторному аналізі?
2. Як перевірити адекватність емпіричних даних моделі факторного аналізу?
3. У чому полягає гіпотеза факторного аналізу?

4. Що таке експлораторний факторний аналіз?
5. Що таке конфірмаційний факторний аналіз?
6. Які методи факторизації Ви знаєте?
7. Що таке факторне навантаження?
8. Як визначають кількість факторів, що підлягають інтерпретації?
9. Що таке «інформативність фактору»?
10. Що показує кумулятивна дисперсія вилучених факторів?
11. Навіщо потрібно застосовувати обертання факторних осей?
12. Що є головним результатом розрахунків процедури факторного аналізу?
13. Для вирішення яких завдань соціологи застосовують факторний аналіз?

Тема 9. Кластерний аналіз

9.1. Сутність кластерного аналізу

Кластерний аналіз – узагальнена назва великого набору алгоритмів автоматичного розпізнавання образів, які використовуються для групування об'єктів в однорідні кластери (групи, типи). Кластерний аналіз є описовою процедурою, він не дозволяє зробити статистичні висновки, але надає можливість здійснити пошук сталих сполучень властивостей соціальних об'єктів (або явищ), розглянутих відповідно до описових гіпотез у кількох вимірах одночасно та вивчити багатовимірну структуру досліджуваної сукупності. Основні етапи кластерного аналізу: 1) відбір ознак, що за гіпотезою дослідника є чинниками диференціації досліджуваних об'єктів по кластерах; 2) вибір комп'ютерної програми обробки даних, в якій буде реалізовано кластерний аналіз (SPSS, STATISTICA або ін.); 3) вибір способу визначення відстані між об'єктами; 4) застосування одного з алгоритмів кластерного аналізу; 5) перевірка надійності отриманих результатів; 6) змістовна інтерпретація отриманої кластерної структури.

У соціології кластерний аналіз активно застосовують як засіб емпіричної типологізації, що останнім часом набуває все більшого значення, оскільки створює передумови побудови типологій, заснованих на емпіричних даних, а не на ідеальних моделях. Кластерний аналіз почав відігравати помітну роль в соціології, зокрема в кількісних соціологічних дослідженнях, починаючи з 70-х років ХХ сторіччя, що було зумовлено розвитком інформаційних технологій та поширенням доступного програмного забезпечення, без якого реалізувати кластерний аналіз неможливо. Значне зростання популярності кластерного аналізу в контексті аналізу кількісних даних відмічалось багатьма соціологами, так, напр., американський соціолог Кеннет Д. Бейлі в книзі «Типології та таксономії: вступ до класифікаційних методів», виданій в 1994 р., навіть прогнозував повне заміщення традиційних для соціології методів типологізації на кластерний аналіз. Повного заміщення

не сталося, прийшло розуміння, що кластерний аналіз не «панацея на всі випадки», а лише дуже корисний різновид типологічного аналізу.

Наразі кластерний аналіз застосовують у двох випадках: 1) з метою емпіричної типологізації, тобто для побудови типології, що заснована на емпіричних даних, а не на ідеальній моделі; 2) для перевірки гіпотези щодо відповідності дійсності теоретично сконструйованої типології.

Основою коректної (з соціологічної точки зору) реалізації кластерного аналізу є, по-перше, теоретично обґрунтований відбір ознак, що за гіпотезою дослідника є чинниками диференціації досліджуваних об'єктів по кластерах. По-друге, грамотна інтерпретація отриманих результатів, яка передбачає виявлення критерію типологізації (в кількісних дослідженнях цьому сприяє попереднє проведення факторного аналізу або багатовимірного шкалювання); опис специфічних характеристик кластерів на основі аналізу не тільки кластерних центроїдів, а й допоміжних ознак, які не брали участі в кластеризації, але сприяють поясненню досліджуваного явища (наприклад, ознак соціально-демографічного блоку).

Останнім часом кластерний аналіз все частіше застосовується соціологами в контексті аналізу якісних (насамперед текстових) даних, метод кластерного аналізу став повноцінним інструментом аналізу соціальних мереж. Таким чином, кластерний аналіз завойовує нові сфери застосування в соціології, історія розвитку даного методу не закінчена.

Особливість кластерного аналізу: На відміну від факторного, кластерний аналіз спирається на розрахунок «відстані» між досліджуваними об'єктами (респондентами), а не кореляційних зв'язків. Гомогенні респонденти (за певними відповідями) об'єднуються в одну групу, гетерогенні – у різні.

Алгоритм процедури кластерного аналізу:

- 1) Визначення сукупності змінних для групування (виокремлення змінних, що є значущими для сегментації респондентів).
- 2) Визначення методу кластеризації.
- 3) Вибір мір відстані (для порівняння об'єктів між собою).
- 4) Прийняття рішення щодо кількості кластерів.
- 5) Інтерпретація виокремлених кластерів.
- 6) Оцінка достовірності й надійності кластеризації.

Ще одним етапом процедури кластерного аналізу, який не завжди потрібно застосовувати, є нормалізація даних (використовується для того, щоб зробити шкали співставними за розмірністю, звести до єдиних інтервалів за допомогою z-нормалізації). Нормалізація даних потрібна в тих випадках, коли в ознаки, за якими проводиться кластеризація виміряні різними шкалами, наприклад, 3-х та 7-ми значними.

Ключові терміни

Кластер (cluster) - скупчення, згусток. Кластером називають групу схожих об'єктів.

Кластерні центроїди представляють середні значення в кластері по кожній змінній, які були основою кластеризації. Кластерні центроїди дозволяють описати «середнього представника» кожного кластеру.

Статистики, пов'язані з кластерним аналізом

План агломерації, об'єднання (*agglomeration schedule*). Дає інформацію про об'єкти (події, випадки), що мають об'єднатися у процесі ієрархічної кластеризації.

Кластерний центроїд (*cluster centroid*). Середнє значення змінних для усіх випадків чи об'єктів у конкретному кластері.

Кластерні центри (*cluster centers*). Вихідні початкові точки в неієрархічній кластеризації. Кластери утворюють навколо цих центрів, або зерен кластеризації.

Приналежність кластеру (*cluster membership*). Вказує кластер, якому належить кожен випадок або об'єкт.

Деревовидна діаграма (*dendrogram*). Її також називають деревовидний граф – графічний засіб для показу результатів кластеризації. Вертикальні лінії представляють об'єднувані кластери. Положення вертикальної лінії на шкалі відстані (горизонтальна вісь) показує відстані, за якими об'єднували кластери. Деревовидну діаграму читають зліва направо.

Найбільш поширені методи кластерного аналізу

Ієрархічні методи

Неієрархічні методи (тобто методи k-середніх)

Ієрархічні методи кластеризації

- Агломеративні (AGNES): кластери групуються, об'єднуються (від множинності до єдності).

- Divisive Analysis (DIANA): кластери роз'єднуються (з цілого на частини).

Ці методи є наглядними: існує графік дендрограма, що показує кроки процесу об'єднання/роз'єднання кластерів.

Неієрархічні методи (k-середніх) використовують для великих масивів даних. Ці методи мають певні недоліки:

- Необхідно знати певну кількість кластерів;
- Необхідно задавати кластерні центри;
- Картина кластеризації дуже мінлива, неоднозначна, ненадійна (алгоритм занадто чутливий до «викидів», які можуть робити середнє хибним. Можливим рішенням цієї проблеми є використання модифікації алгоритму - алгоритм k-медіани);
- Результати застосування методу залежать від того, яким чином розташовані анкети в масиві (в залежності від того, яка анкета взята першою, змінюватимуться обсяги кластерів).

9.2. Реалізація ієрархічного кластерного аналізу в пакеті SPSS

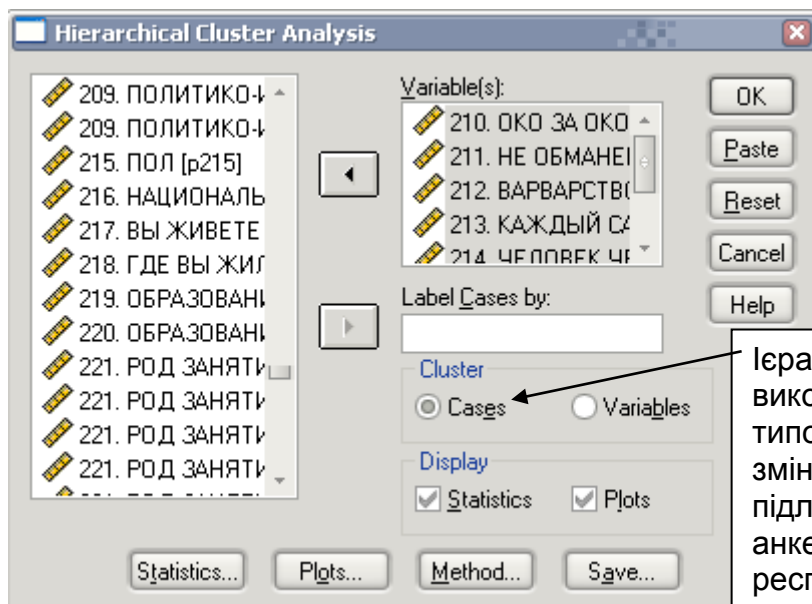
Масив st01.sav

Ознаки 210-214, що визначають міру згоди з перерахованими моральними принципами

Метод Варда

Метрика - квадрат евклідової відстані

Для виклику процедури ієрархічного кластерного аналізу необхідно виконати команду *Analyze* → *Classify* → *Hierarchical Cluster*

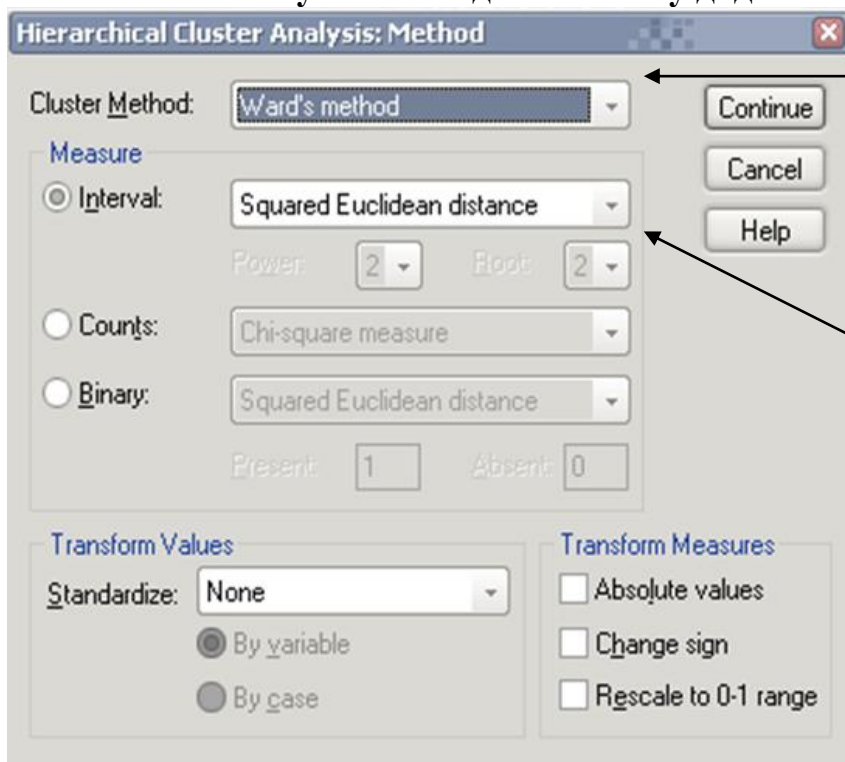


Далі необхідно вибрати метод утворення кластерів, а також метод розрахунку дистанційної міри і міри подібності. У пакеті SPSS передбачені наступні методи об'єднання в кластери:

- **Зв'язок між групами** (Between - groups linkage) Дистанція між кластерами дорівнює середньому значенню дистанцій між усіма можливими парами спостережень.
- **Зв'язок усередині груп** (Within - groups linkage). Розраховується на підставі усіх можливих пар спостережень, що належать обом кластерам, причому враховуються також і пари спостережень, що утворюються усередині кластерів.
- **Відстань між найближчими сусідами** - найближчими об'єктами кластерів (Nearest neighbour). Тут відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах.
- **Відстань між найдалшими сусідами** (Furthest neighbour). Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто "найбільш видаленими сусідами").

- **Відстань між центрами кластерів** (Centroid clustering) або **центроїдний метод**. Недоліком цього методу є те, що центр об'єднаного кластеру обчислюється як середнє центрів об'єднаних кластерів, без урахування їх обсягу.
- **Метод медіан** - той же центроїдний метод, але центр об'єднаного кластеру обчислюється як середнє усіх об'єктів (Median clustering).
- **Метод Варда**. Відстанню між кластерами є приріст суми квадратів відстаней об'єктів до центрів кластерів, що розраховується в результаті їх об'єднання. Тут використовуються методи дисперсійного аналізу. Найбільш ефективний в роботі з соціологічними даними.

Натисніть кнопку Methods для виклику додаткового вікна:

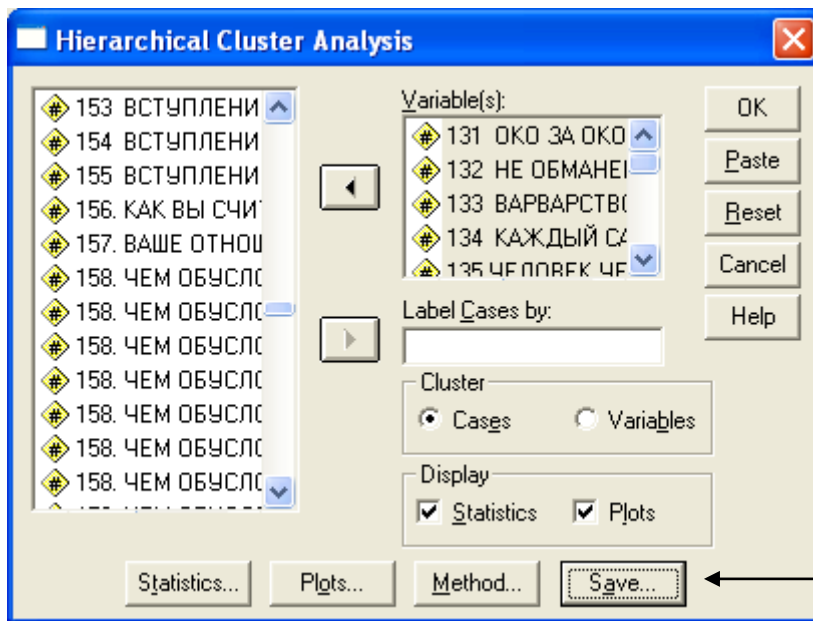


Тут необхідно обрати метод об'єднання в кластери. Як правило використовується метод Варда

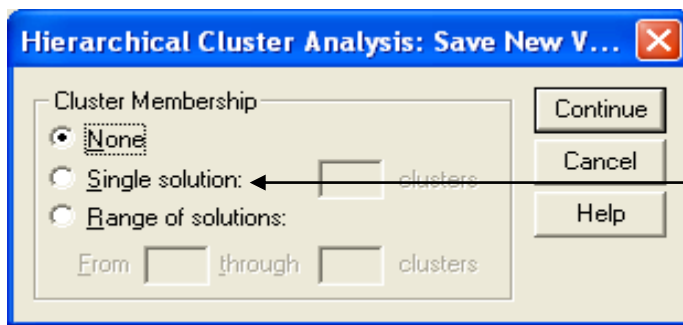
Тут необхідно обрати метод об'єднання в кластери. Як правило використовується метод Варда

За результатами обчислень визначаємо кількість кластерів

Визначившись з певною кількістю кластерів, знову замовляємо процедуру кластерного аналізу з метою збереження кластерних змінних. Це дає можливість провести аналіз кожного кластеру окремо.



Викликається при необхідності утворення змінної, що показує номер кластеру для кожної анкети

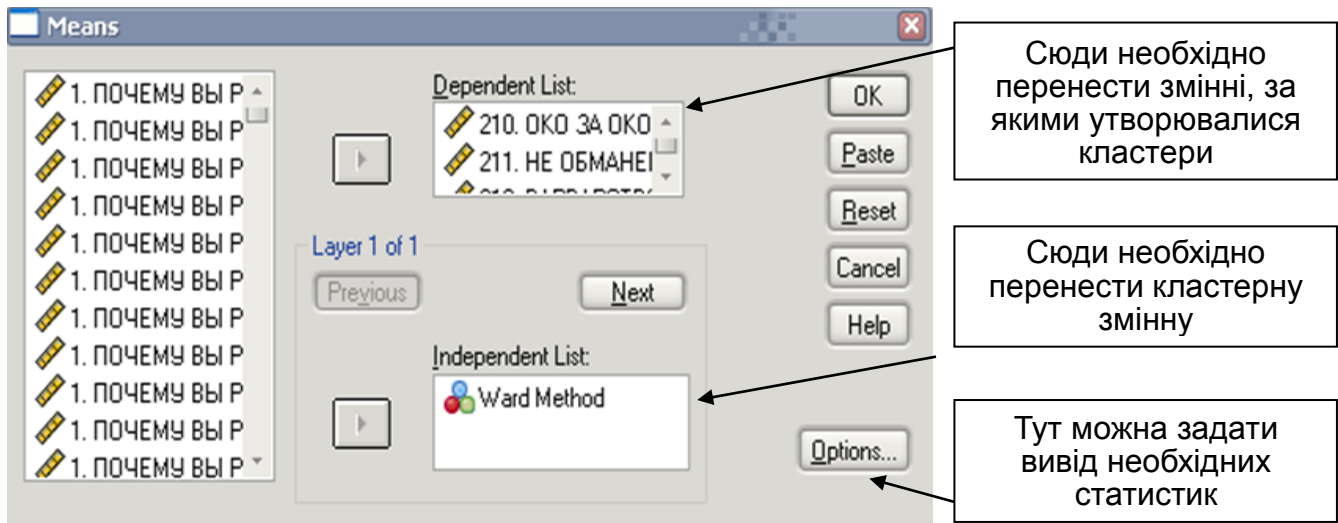
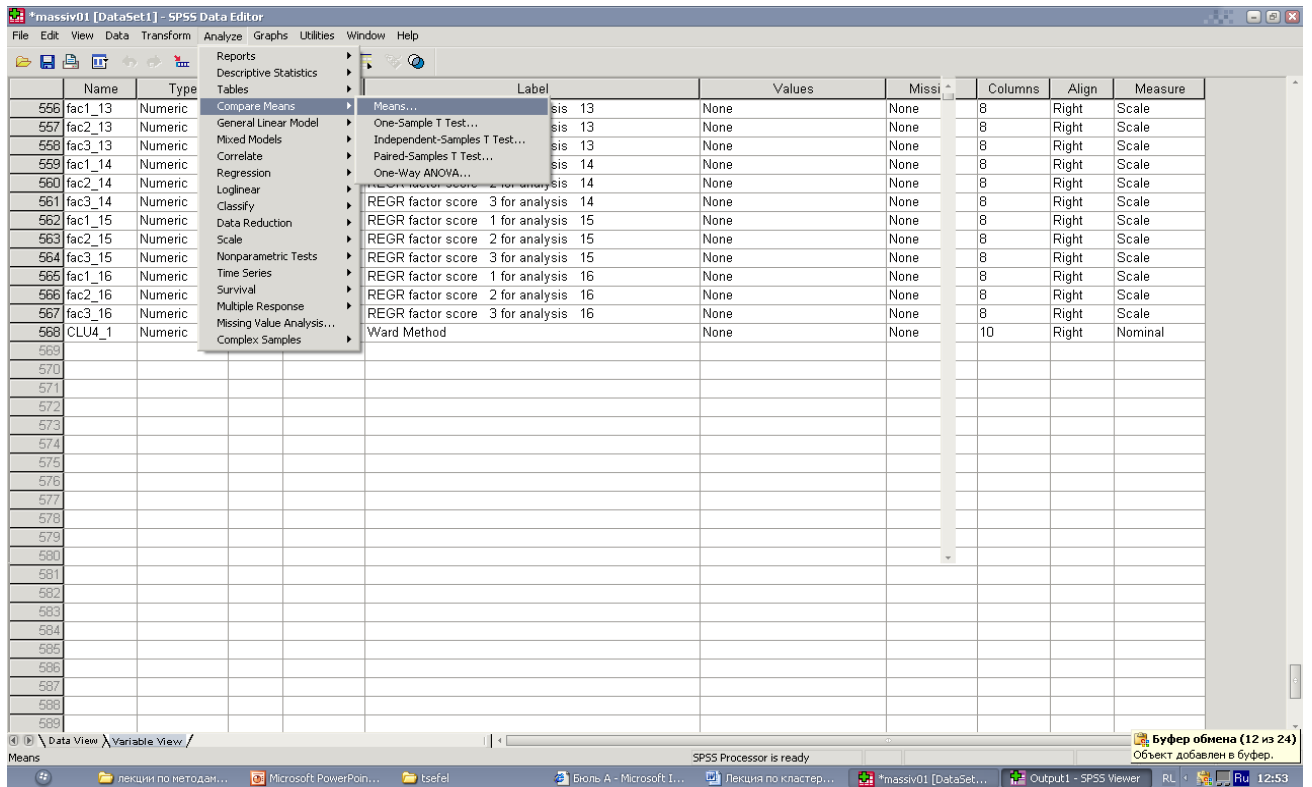


Встановлюємо збереження одного рішення (Single solutions), тобто усі кластери будуть збережені в одній змінній

Збережена кластерна змінна знаходиться в кінці списку Variable View

	Name	Type	Width	Decimals	Label	
556	fac1_13	Numeric	11	5	REGR factor score 1 for analysis 13	None
557	fac2_13	Numeric	11	5	REGR factor score 2 for analysis 13	None
558	fac3_13	Numeric	11	5	REGR factor score 3 for analysis 13	None
559	fac1_14	Numeric	11	5	REGR factor score 1 for analysis 14	None
560	fac2_14	Numeric	11	5	REGR factor score 2 for analysis 14	None
561	fac3_14	Numeric	11	5	REGR factor score 3 for analysis 14	None
562	fac1_15	Numeric	11	5	REGR factor score 1 for analysis 15	None
563	fac2_15	Numeric	11	5	REGR factor score 2 for analysis 15	None
564	fac3_15	Numeric	11	5	REGR factor score 3 for analysis 15	None
565	fac1_16	Numeric	11	5	REGR factor score 1 for analysis 16	None
566	fac2_16	Numeric	11	5	REGR factor score 2 for analysis 16	None
567	fac3_16	Numeric	11	5	REGR factor score 3 for analysis 16	None
568	CLU4_1	Numeric	8	0	Ward Method	None
569						
570						
571						
572						
573						

Для наступного аналізу необхідно побудувати таблицю кластерних центрів, для чого виконуємо команду: *Analyze/ Compare means(Порівняння середніх)/ Means*



Таблиця кластерних центрів. Під час інтерпретації необхідно пам'ятати, на підставі яких шкал проводилася кластеризація, оскільки інакше буде незрозуміло, що означають середні значення, представлені в таблиці.

Report

Ward Method		210. ОКО ЗА ОКО	211. НЕ ОБМАНЕ ШЬ - НЕ ПРОЖИВ ЕШЬ	212. ВАРВАРС ТВО ИСКОРЕ НЯЕТСЯ ВАРВАРС ТВОМ	213. КАЖДЫЙ САМ ЗА СЕБЯ	214. ЧЕЛОВЕК ЧЕЛОВЕКУ ВОЛК
1	Mean	4.9936	4.0552	4.2490	2.2863	2.4840
	N	779	779	779	779	779
2	Mean	6.0595	5.5833	5.7083	5.8006	4.8690
	N	336	336	336	336	336
3	Mean	4.4738	3.9323	3.7969	5.1231	3.8585
	N	325	325	325	325	325
4	Mean	1.7652	1.9697	1.4621	2.3333	1.7500
	N	132	132	132	132	132
Total	Mean	4.8429	4.1813	4.2335	3.6279	3.2163
	N	1572	1572	1572	1572	1572

Інтерпретація результатів ієрархічного кластерного аналізу:

- 1) кількість спостережень, що підлягають кластеризації, зазначена в таблиці

Case Processing Summary

Case Processing Summary ^{a,b}					
Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
1572	93,6	108	6,4	1680	100
a Squared Euclidean Distance used					
b Ward Linkage					

- 2) Головне питання кластерного аналізу – питання про кількість кластерів. Тут відсутні чіткі правила, що дозволяють швидко прийняти рішення, але можна керуватися наступними.

- У ієрархічній кластеризації в якості критерію можна використовувати відстані, за якими об'єднують кластери.
- У неієрархічній кластеризації зображують графік залежності відношення сумарної внутрішньогрупової дисперсії до міжгрупової дисперсії від кількості кластерів. Точка, в якій спостерігається вигин або різкий поворот, вказує на прийнятну кількість кластерів.

- Відносні розміри кластерів мають бути досить виразними. (Безглуздо створювати кластер з одним випадком).

Визначення кількості кластерів за таблицею Agglomeration Schedule

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1653	1680	0,00	0	0	681
2	1652	1679	0,00	0	0	16
1566	1	2	10788,55	1559	1556	1568
1567	4	6	11506,12	1554	1563	1568
1568	1	4	12801,96	1566	1567	1570
1569	3	5	14515,80	1565	1564	1571
1570	1	11	17120,91	1568	1562	1571
1571	1	3	23446,60	1570	1569	0

Кількість кластерів = $N - S$, де N - кількість анкет, що підлягали кластеризації, S - крок, після якого спостерігається перший скачок коефіцієнта. Кількість кластерів = $1572 - 1568 = 4$.

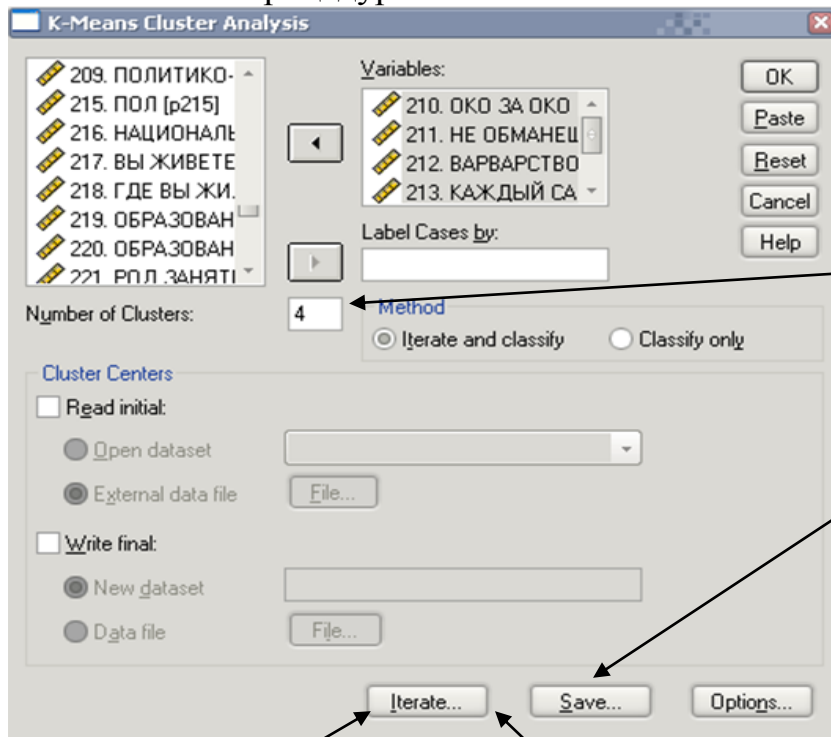
Інтерпретація кластерів – опис кластерних центроїдів

N кластера	Обсяг кластера	210. Око за око	211. Не обдуриш - не проживеш	212. Варварство викорінюється варварством	213. Кожен сам за себе	214. Людина людині - вовк
1	779	4,99	4,06	4,25	2,29	2,48
2	336	6,06	5,58	5,71	5,80	4,87
3	325	4,47	3,93	3,80	5,12	3,86
4	132	1,77	1,97	1,46	2,33	1,75

9.3. Реалізація неієрархічного кластерного аналізу в пакеті SPSS

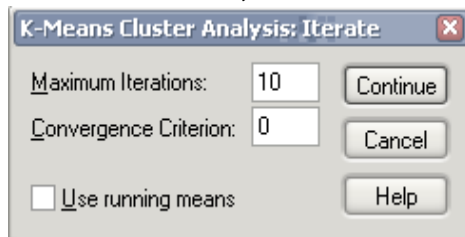
Для виклику процедури неієрархічного кластерного аналізу виберіть у головному меню послідовність пунктів *Analyze* → *Classify* → *K-Means Cluster*

Головне вікно процедури



Задаємо певну кількість кластерів. У нашому випадку, спираючись на результати ієрархічного кластерного аналізу, ми можемо виділити 4 кластери.

Заходимо в це меню з метою створення нової змінної, що покаже, до якого кластеру відноситься кожна анкета. Встановлюємо прапорець



У даному (під)меню можна задати кількість ітерацій (обертань). Автоматично стоїть 10 ітерацій. Встановіть, наприклад, 100

Таблиця Iteration History не підлягає змістовній інтерпретації. Вона вказує на кінець кластерного аналізу. У нижньому рядку таблиці мають бути нулі. Якщо в нижньому рядку стоять інші числа, необхідно повторити обчислення, задавши більшу кількість ітерацій.

Iteration History

Iteration	Change in Cluster Centers			
	1	2	3	4
1	3.662	4.561	4.324	4.548
2	.306	.701	.408	.766
3	.141	.304	.575	.435
4	.045	.220	.276	.229
5	.048	.130	.200	.150
6	.022	.107	.122	.055
7	.017	.060	.080	.055
8	.006	.059	.076	.037
9	.017	.041	.062	.036
10	.007	.038	.034	.000
11	.004	.032	.029	.000
12	.000	.000	.000	.000

- a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 12. The minimum distance between initial centers is 9.055.

Основна таблиця (Final Cluster Centers) підлягає змістовній інтерпретації. У ній знаходяться середні значення за усіма ознаками в кожному кластері, на основі яких здійснюється інтерпретація. Під час інтерпретації необхідно пам'ятати, на підставі яких шкал проводилася кластеризація, оскільки інакше буде незрозуміло, що означають середні значення в таблиці

Final Cluster Centers

	Cluster			
	1	2	3	4
210. ОКО ЗА ОКО	5.46	6.07	4.45	3.04
211. НЕ ОБМАНЕШЬ - НЕ ПРОЖИВЕШЬ	4.42	5.52	4.09	2.49
212. ВАРВАРСТВО ИСКОРЕНЯЕТСЯ ВАРВАРСТВОМ	4.86	5.77	3.66	2.32
213. КАЖДЫЙ САМ ЗА СЕБЯ	2.18	5.84	5.00	1.88
214. ЧЕЛОВЕК ЧЕЛОВЕКУ ВОЛК	2.65	5.11	3.42	1.86

Таблиця **Number of Cases in each Cluster** вказує на кількість респондентів у кожному кластері

Number of Cases in each Cluster

Cluster	1	512.000
	2	339.000
	3	401.000
	4	320.000
Valid		1572.000
Missing		110.000

Оцінка надійності й достовірності

Маючи декілька висновків, зроблених в процесі кластерного аналізу, не слід приймати певне рішення про кластеризацію, не виконавши оцінку надійності й достовірності цього рішення. Формальні процедури оцінки надійності й достовірності рішень кластеризації досить складні і не завжди виправдані. Проте наступні процедури забезпечать адекватну перевірку якості кластерного аналізу.

- Виконуйте кластерний аналіз на підставі одних і тих же даних, але з використанням різних способів вимірювання відстані. Порівняйте результати, отримані на основі різних вимірювань відстані, щоб визначити, наскільки співпадають отримані результати.
- Використовуйте різні методи кластерного аналізу і порівняйте отримані результати.
- Розбийте дані на дві рівні частини випадковим чином. Виконайте кластерний аналіз окремо для кожної половини. Порівняйте кластерні центроїди двох підвбірок.
- Випадковим чином видаліть деякі змінні. Виконайте кластерний аналіз за скороченим набором змінних. Порівняйте результати з отриманими на основі повного набору змінних.
- У неієрархічній кластеризації рішення може залежати від порядку випадків в наборі даних. Виконайте аналіз кілька разів, змінюючи порядок випадків, до отримання стабільного рішення.

Заключні зауваження:

Для реалізації засад кластерного аналізу в соціологічних дослідженнях слід послуговуватися доволі об'ємним алгоритмом процедури реалізації першого.

За допомогою розвитку сучасних технологій, у соціологічній практиці найчастіше використовують ієрархічні методи кластеризації.

Слід пам'ятати, що ключовим аспектом кластерного аналізу є оцінка достовірності та надійності виокремлення кластерів, тому використання декількох методів та їх комбінація є актуальними дослідницькими завданнями.

9.4. Приклад застосування кластерного аналізу в соціологічному дослідженні

Застосування кластерного аналізу в дослідженні ролі особистісного фактора в досягненні життєвого успіху

Масив st13.sav. Дослідження лідерів студентського самоврядування проведено кафедрою прикладної соціології під керівництвом В.С. Бакірова у 2013 р.

Виконання:

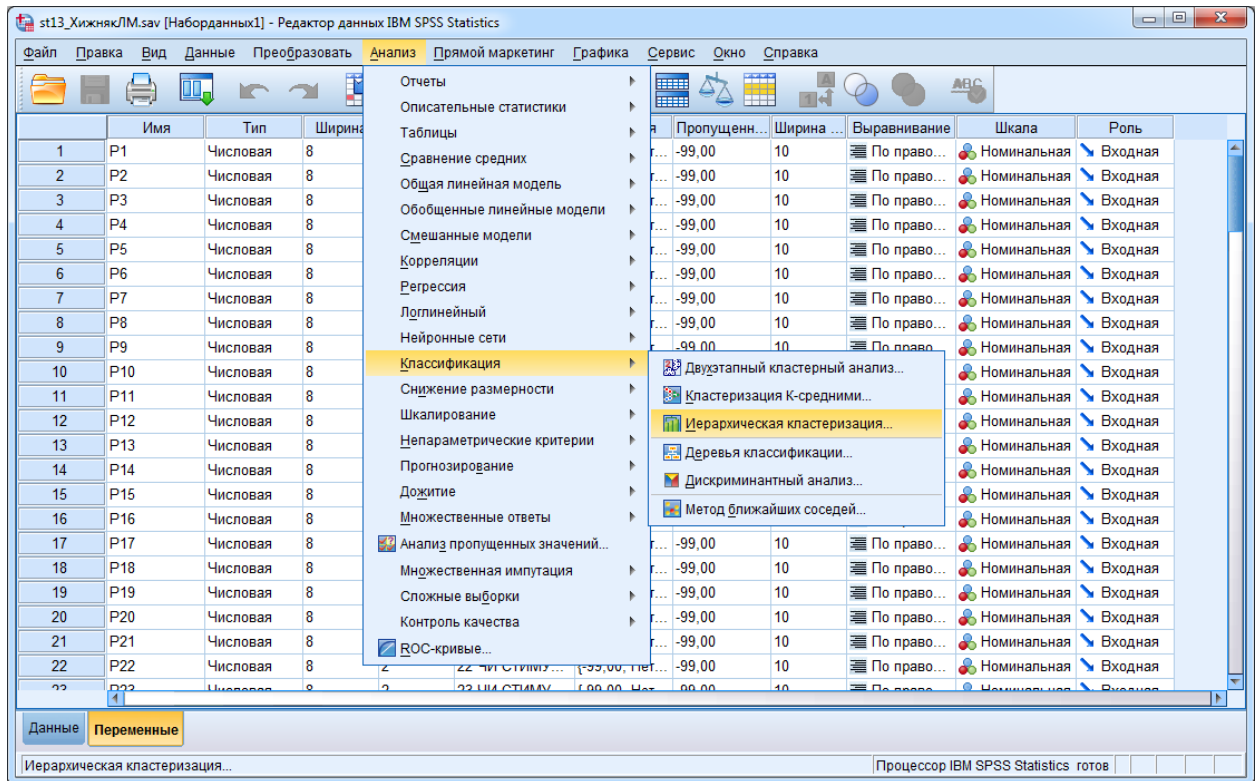


Рис. 9.1. Виклик процедури кластерного аналізу

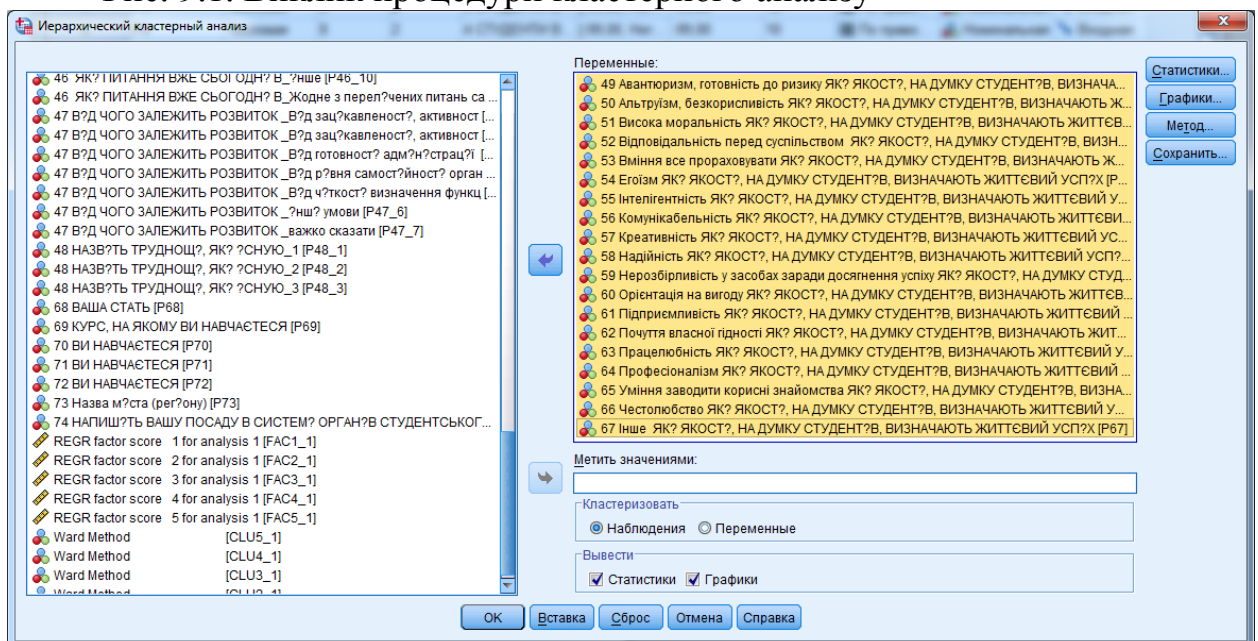


Рис. 9.2. Задаємо змінні для проведення кластеризації

Проаналізуємо роль особистісного фактору в досягненні життєвого успіху. Для цього розглянемо експертні оцінки лідерів студентського самоврядування ЗВО України, ґрунтуючись на їх відповідях на блок запитань анкети щодо важливості певних особистісних якостей для досягнення життєвого успіху. Так, респондентів запитували «Які якості, на думку сучасних студентів, визначають життєвий успіх людини?» (нижче були перелічені різноманітні якості, важливість яких лідери студентського самоврядування мали оцінити за п'ятибальною шкалою: 5 – «визначає повною мірою»; 4 – «скоріше визначає»; 3 – «може визначає, може не визначає»; 2 – «скоріше не визначає»; 1 – «не визначає»). Порівняння ступеня важливості цих якостей проводилось шляхом побудови ієрархії на основі порівняння середніх значень, розрахованих по масиву опитаних (див. табл. 9.1).

Таблиця 9.1

Ієрархія особистісних якостей, важливих для життєвого успіху, в оцінках лідерів студентського самоврядування України

№	Особистісні якості (середні значення)*	Коефіцієнт варіації (%)** $V = \frac{\sigma}{\bar{x}} \cdot 100\%$
1	Креативність 4,58	16
2	Комунікабельність 4,53	16
3	Професіоналізм 4,49	18
4	Працелюбність 4,4	20
5	Уміння заводити корисні знайомства 4,35	18
6	Надійність 4,25	19
7	Інтелігентність 4,15	25
8	Вміння все прораховувати 4,09	23
9	Почуття власної гідності 4,06	20
10	Честолюбство 3,7	23
11	Відповідальність перед суспільством 3,58	26
12	Висока моральність 3,58	29
13	Авантюризм, готовність до ризику 3,48	31
14	Орієнтація на вигоду 3,32	28
15	Підприємливість 3,02	31
16	Нерозбірливість у засобах заради досягнення... 2,83	36
17	Егоїзм 2,73	44
18	Альтруїзм, безкорисливість 2,72	38

N=490, 2013 р.

* Ієрархію побудовано на основі середніх значень відповідей респондентів щодо ступеня важливості перелічених в анкеті особистісних якостей для досягнення життєвого успіху. Інтервал: від 5 – «визначає повною мірою» до 1 – «не визначає».

**Коефіцієнт варіації - це відношення середньоквадратичного відхилення до середнього арифметичного, представлене у відсотках

Можна побачити, що перші три місця ієрархії посідають такі якості, як креативність, комунікабельність, професіоналізм. Егоїзм та альтруїзм, що вважаються найбільш важливими в моделях «економічної людини» та «соціологічної людини», займають останні місця – 17 та 18, відповідно. Проте це не є свідомством емпіричного спростування теоретичного підґрунтя цих моделей, оскільки погляди наших респондентів на обумовленість життєвого успіху певними особистісними якостями дуже варіативні, досліджувана сукупність не є статистично однорідною, зокрема за такими ознаками як егоїзм (коефіцієнт варіації дорівнює 44%), альтруїзм, безкорисливість (38%) та нерозбірливість у засобах заради досягнення успіху (36%).

Зі статистики відомо, що коефіцієнт варіації дає можливість судити про ступінь однорідності досліджуваної сукупності, що є передумовою коректного застосування мір центральної тенденції в аналізі емпіричної інформації: якщо за якоюсь ознакою $V \leq 33\%$, то за нею сукупність досить однорідна; якщо V перевищує 33%, то сукупність недостатньо однорідна і середні значення не можуть коректно відобразити реальні закономірності. У таких випадках досліджувану сукупність необхідно розділити, щоб мати змогу робити статистичні висновки.

Для вирішення такої проблеми зазвичай застосовують диференціацію сукупності за найбільш важливою ознакою. Проте ми апріорно не готові назвати таку одну ознаку, оскільки вважаємо, що у наших реаліях «економічна людина» співіснує з «соціологічною людиною», при цьому егоїзм вважається найбільш важливою якістю в моделі «економічної людини», а альтруїзм – в моделі «соціологічної людини». Крім того не можна не мати на увазі, що ці моделі ймовірно не охоплюють всю багатогранність досліджуваного феномену.

У зв'язку зі сказаним, звернемось до емпіричних даних та перевіримо гіпотезу про існування кількох типів сприйняття життєвого успіху у середовищі сучасного студентства. Для вирішення цього завдання ми застосуємо кластерний аналіз, який, як відомо, є поширеним методом побудови емпіричних типологій, особливо у тих випадках, коли є потреба урахувати багатомірність досліджуваного феномену.

Кластерний аналіз показав, що спостерігається диференціація думок студентів про те, якими особистісними якостями обумовлюється досягнення життєвого успіху: приблизно третина експертів (33 %) акцентують увагу на якостях, що характерні для моделі «економічної людини» і дві третини (67 %) – на якостях, що властиві моделі «соціологічної людини» (рис. 9.3). Однак у межах соціологічної та економічної моделі досягнення життєвого успіху було виявлено значну диференціацію думок респондентів відносно якостей, що сприяють досягненню життєвого успіху. Це свідчить про необхідність

виділення більшої кількості кластерів, у зв'язку з чим ми скористалися чотирикластерною моделлю.

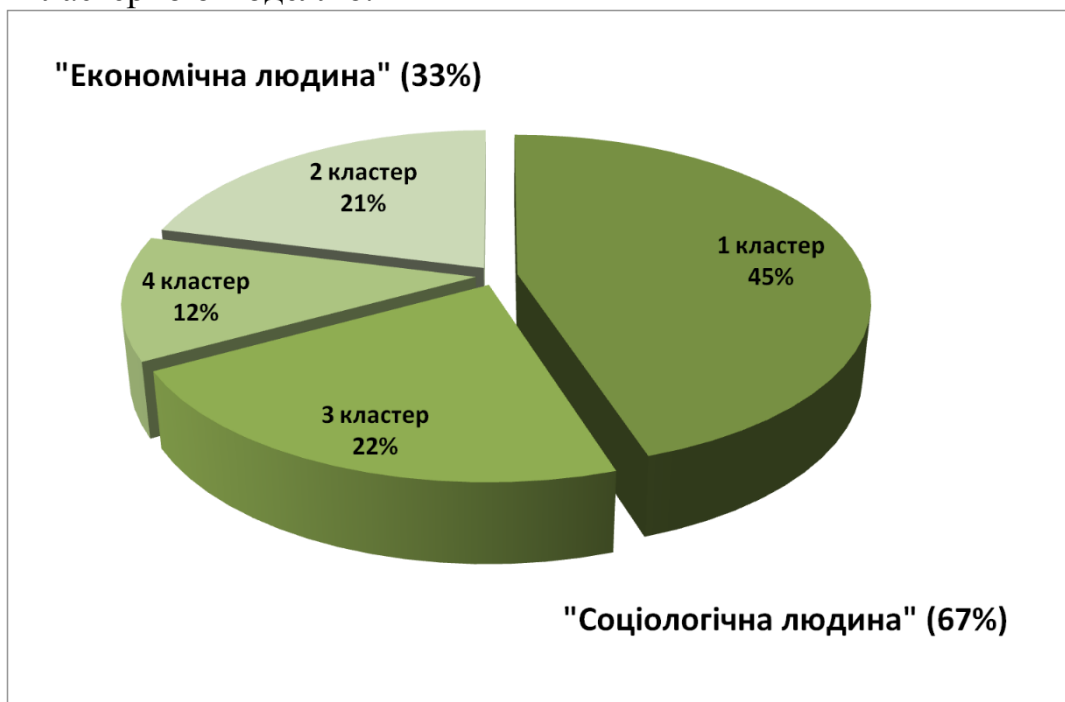


Рис. 9.3. Співвідношення прихильників моделей «соціологічна людина» та «економічна людина»

Відмінності прихильників моделей соціологічної та економічної людини обумовлюються багатовимірною конфігурацією сприйняття респондентами значущості певних якостей, що визначають досягнення життєвого успіху (рис. 9.4). Зазначимо, що на цьому рисунку чітко видно, що розбіжності обумовлюються перш за все різним сприйняттям егоїзму як особистісної якості, що визначає досягнення життєвого успіху. Серед прихильників моделі соціологічної людини це значення дорівнює 2,2 (скоріше не визначає) проти 3,8 (скоріше визначає) у середовищі прихильників моделі економічної людини.

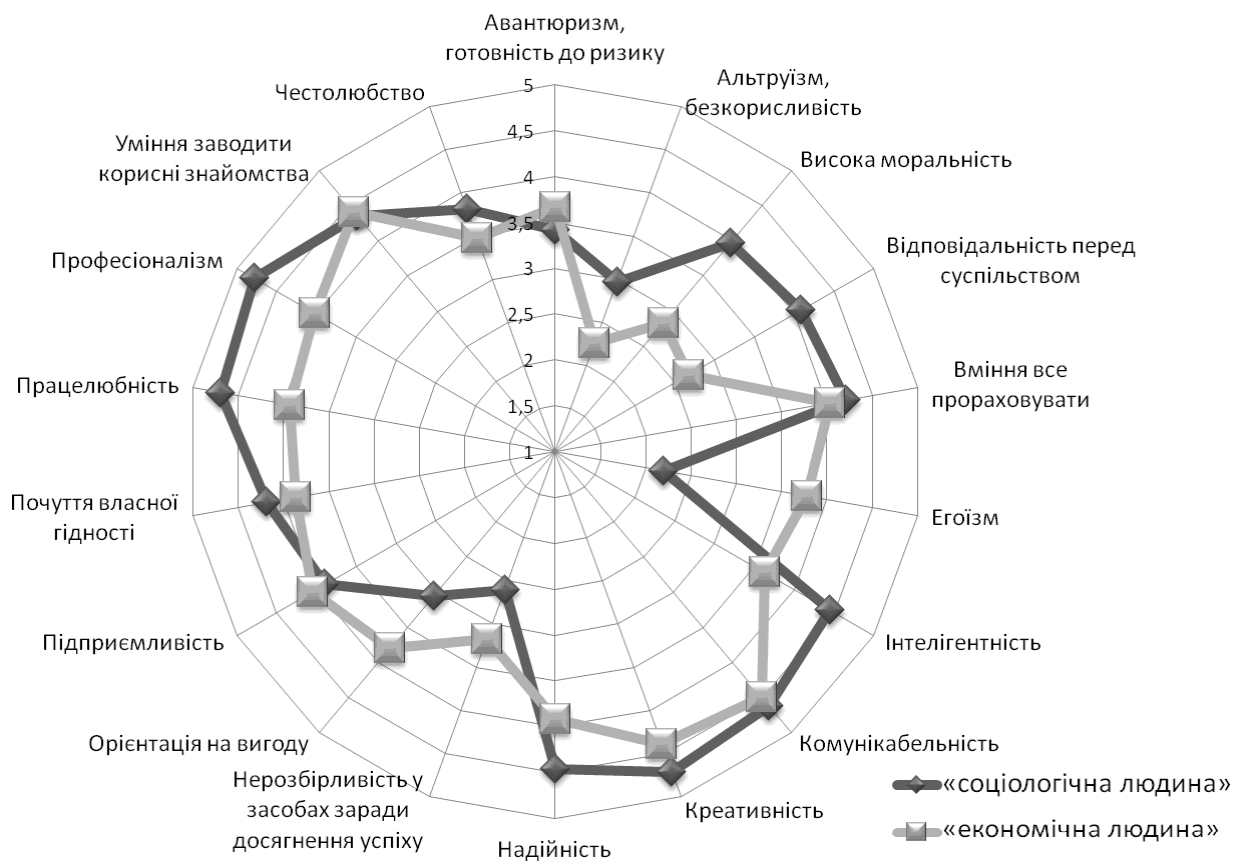


Рис. 9.4. Важливість особистісних якостей для життєвого успіху в оцінках прихильників моделей «соціологічна людина» та «економічна людина» (середні значення коливаються від 5 – «визначає повною мірою» до 1 – «не визначає»)

Аналіз результатів дає можливість побачити значні розходження й за ознаками відповідальність перед суспільством (4,1 проти 2,7), висока моральність (4 проти 2,8), інтелігентність (4,5 проти 3,6).

Окремої уваги заслуговують й відрізнєння в ієрархіях особистісних якостей, важливих для життєвого успіху, за оцінками прихильників моделей «соціологічна людина» та «економічна людина». Перші фокусуються на професіоналізмі, другі – на комунікабельності (рис. 9.5).

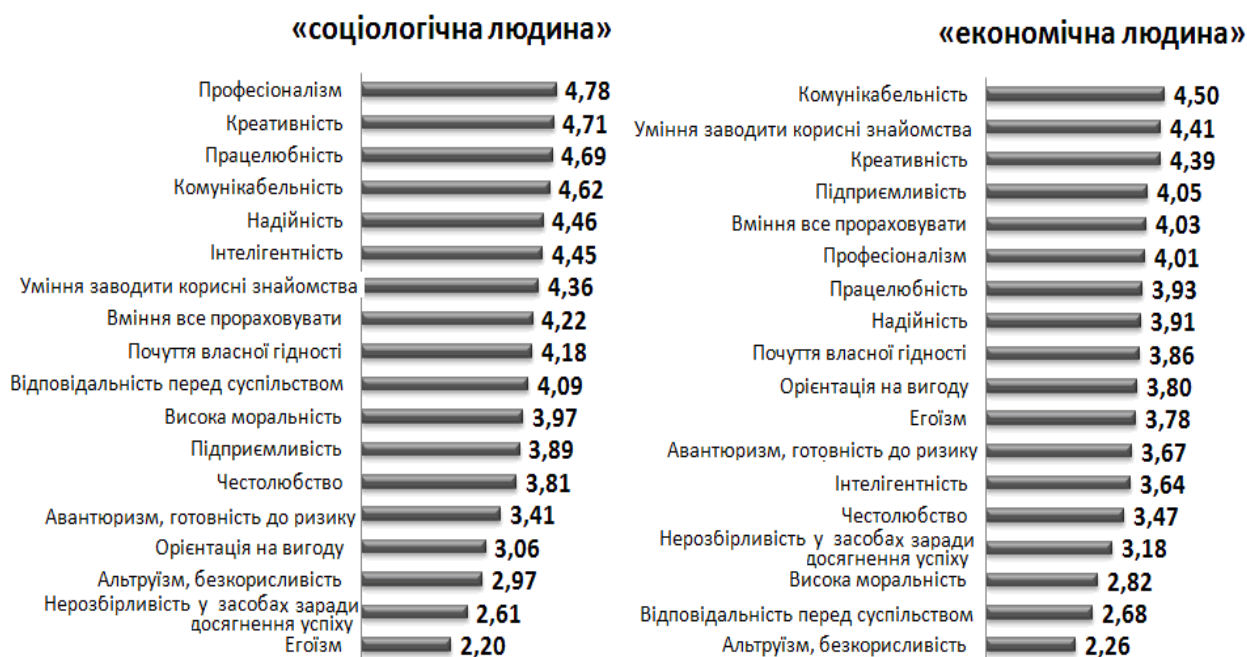


Рис. 9.5. Ієрархія особистісних якостей, важливих для життєвого успіху, за оцінками прихильників моделей «соціологічна людина» та «економічна людина» (середні значення коливаються від 5 – «визначає повною мірою» до 1 – «не визначає»)

Проте кластери «соціологічна людина» та «економічна людина» не повною мірою підтверджують теоретичні основи цих моделей. Чому? Ми вважаємо, що це пов'язано з існуванням підтипів у межах цієї загальної класифікації. Це явно демонструє чотирікластерна модель (рис. 9.6).

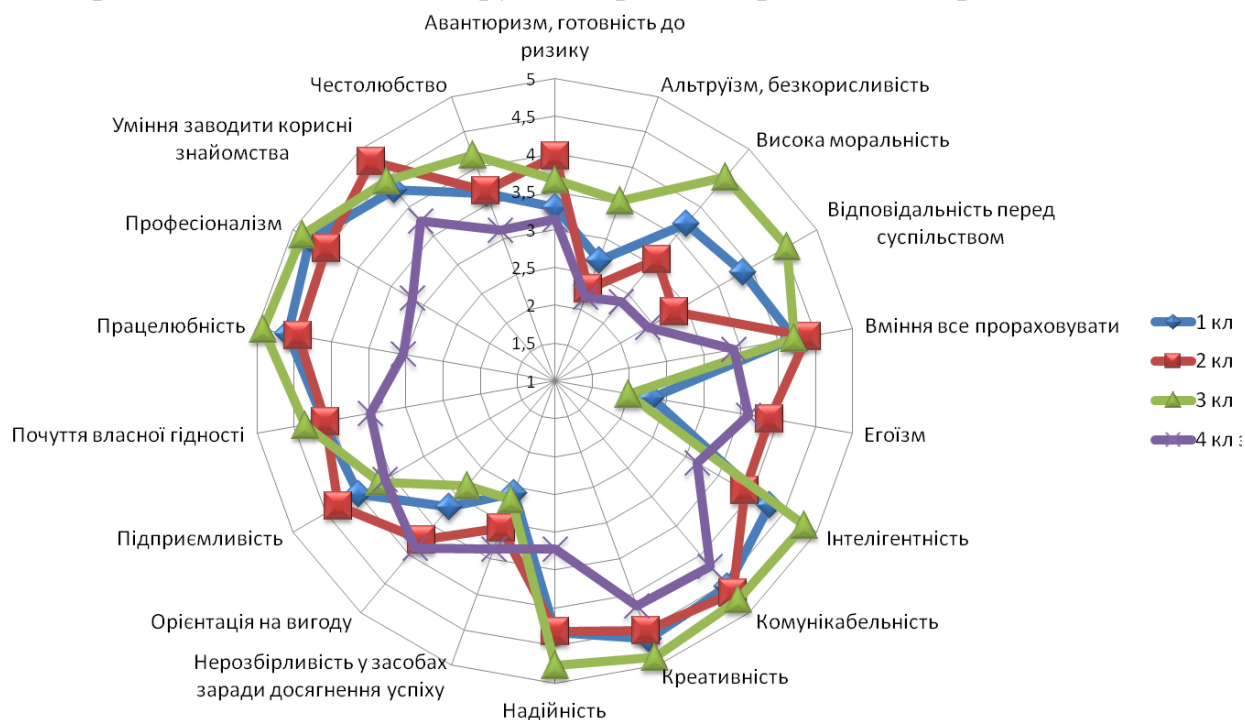


Рис. 9.6. Важливість особистісних якостей для життєвого успіху в оцінках представників чотирьох кластерів (середні значення коливаються від 5 – «визначає повною мірою» до 1 – «не визначає»)

Перш за все зазначимо, що 1 та 3 кластери – представляють різновиди «соціологічної людини», 2 та 4 - «економічної людини». Розглянемо їх більш детально. Перш за все звернемо увагу на різновиди «економічної людини» (див. рис. 9.7).

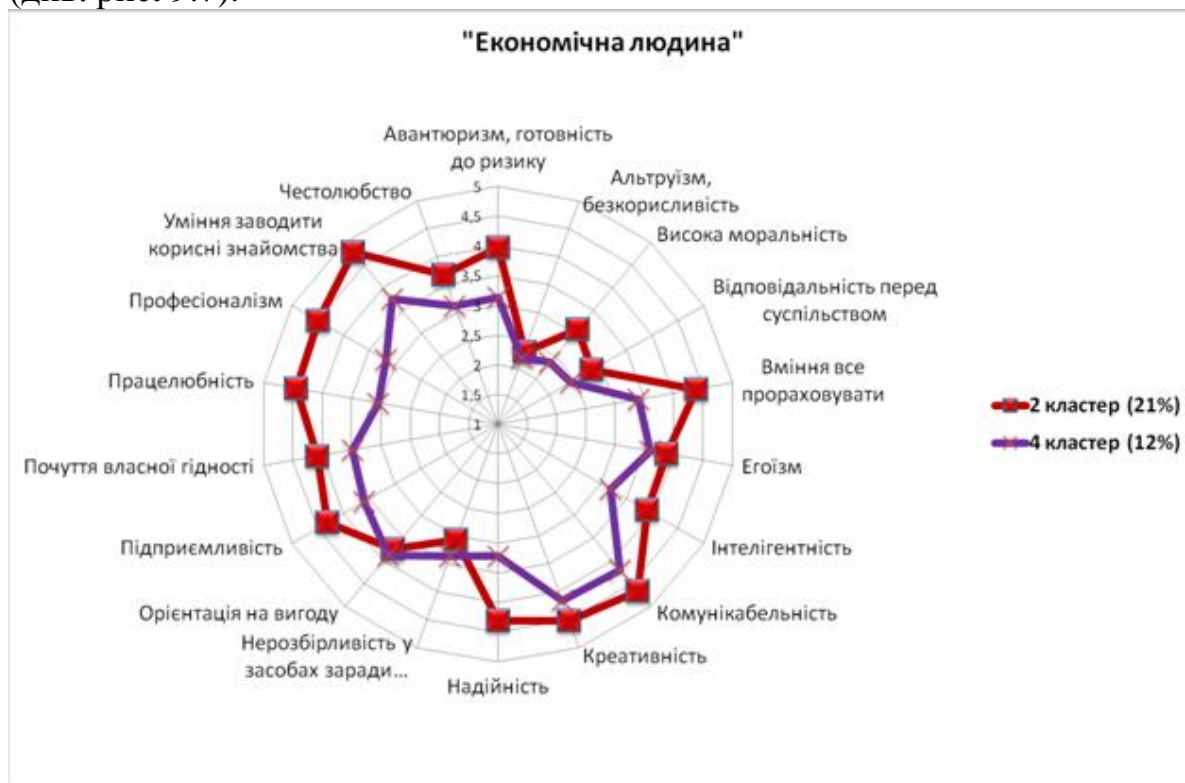


Рис. 9.7. Підтипи «людини економічної»

Розглянемо *другий кластер* (21%), що представляє традиційну модель «людини економічної» - раціонального, розумного егоїста, який прагне отримати максимальну вигоду, використовуючи всі можливі (в тому числі і легітимні) кошти. Його представники орієнтовані на комунікбельність, креативність, професіоналізм, працьовитість, надійність і, одночасно вміння заводити потрібні знайомства і все прорахувати. Вони вважають, що егоїзм - корисна якість при прагненні досягти життєвого успіху, а альтруїзм - ні.

Четвертий кластер - найменший (12%) і відрізняється найбільшою неоднорідністю думок своїх представників. Найбільш яскраво вираженими особливостями цієї групи є низькі показники в оцінці значущості альтруїзму, моральності, відповідальності перед суспільством, інтелігентності, професіоналізму, працьовитості та ін. (див. рис. 9.7). Ще однією особливістю представників даного кластеру є висока оцінка орієнтації на вигоду і нерозбірливість у засобах досягнення успіху. При цьому ми бачимо низькі оцінки значущості розважливості, надійності і вміння заводити корисні знайомства. Найбільше представники четвертого кластеру цінують креативність та комунікбельність, що є також центральною тенденцією по всьому масиву опитаних. Представники даного кластеру є втіленням «людини економічної», але аж ніяк не розумної, оскільки досягнення

життєвого успіху вони пов'язують насамперед з егоїзмом, орієнтацією на вигоду, комунікабельністю і креативністю, які вважаються ними повноцінними заміниками професіоналізму, працьовитості, інтелігентності та ін. Саме представники цього кластера - бич нашої епохи. Вони - це ті, у кого маса необгрунтованих амбіцій, мало знань і умінь, але є стійке прагнення до досягнення успіху, хто здатний «йти по кістках», використовуючи всі засоби для досягнення особистої вигоди. Збільшення питомої ваги таких людей в органах влади, здатне перекрити будь-які спроби більшості досягти життєвого успіху шляхом підвищення професіоналізму, рівня знань, відповідальності перед суспільством. Саме через них сьогодні альтруїзм розглядається як нова форма ризику, своєрідний авантюризм.

Модель «соціологічної людини» представлена 1 і 3 кластерами (див. рис. 9.8). Їх відмінності обумовлені диференціацією думок щодо важливості таких якостей як орієнтація на вигоду (3,2 в першому кластері, 2,8 - у другому), підприємливість (3,9 і 3,2, відповідно), егоїзм (2,3 і 1,9), а також альтруїзм (2,7 в першому кластері і 3,5 - у другому), висока моральність (3,7 і 4,5, відповідно), відповідальність перед суспільством (3,9 і 4,5), честолюбство (3,6 і 4,1), інтелігентність (4,3 і 4,8). Представники цих кластерів дуже схожі між собою по багатовимірній конфігурації їхніх поглядів на якості, що сприяють досягненню життєвого успіху, але 3 кластер - це більш явно виражений підтип соціологічної людини, значно орієнтованої на відповідальність перед суспільством, високу моральність, альтруїзм і при цьому інтелігентного, комунікабельного, креативного, працьовитого, що володіє високим професіоналізмом, здоровим честолюбством і почуттям власної гідності.

"Соціологічна людина"



Рис. 9.8. Підтипи «людини соціологічної»

Підводячи підсумок, відзначимо, що емпіричні дані чітко показують наявність значної диференціації студентства за сприйняттям факторів життєвого успіху. Певні характеристики стратегій «людини економічної» домінують у свідомості студентів, проте вони занадто переплетені з діаметрально протилежними (в рамках обраної теоретичної концепції) характеристиками «людини соціологічної», що свідчить про доцільність більш детального дослідження даного феномена.

Література до теми

Обов'язкова література

Бююль А., Цёфель П. *SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей*. СПб.: ООО «ДиаСофтЮП», 2005. С. 384-408.

Крыштановский А.О. *Анализ социологических данных с помощью пакета SPSS*. М.: ГУ ВШЭ, 2007. С. 205-216.

Наследов А.Д. *SPSS: Компьютерный анализ данных в психологии и социальных науках*. СПб.: Питер, 2005. С. 315-330.

Паніотто В.І., Максименко В.С., Марченко Н.М. *Статистичний аналіз соціологічних даних*. – К.: Вид.дім “КМ Академія”, 2004. С. 232-241.

Додаткова література

Пациорковский В.В., Пациорковская В.В. *SPSS для социологов. Учебное пособие*. ИСЭПН РАН. М., 2005. 433 с.

IBM SPSS Statistics Base 20. *Фирменное руководство пользователя по работе с модулем "Statistics Base" в IBM SPSS Statistics 20.0*. 2011. 363 с.

Наследов А. *IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных*. СПб.: Питер, 2013. С. 273-288.

Питання для самоконтролю

1. В чому полягає сутність кластерного аналізу? З якою метою він застосовується?
2. Які вам відомі алгоритми кластерного аналізу?
3. В чому полягає різниця між ієрархічними та неієрархічними методами кластерного аналізу?
4. Які чинники виникнення помилок кластеризації?
5. Як оцінити надійність й достовірність результатів кластерного аналізу?

Тема 10. Регресійний аналіз та його застосування у соціологічних дослідженнях

10.1. Сутність регресійного аналізу

Регресійний аналіз — розділ математичної статистики, присвячений методам аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу не з'ясовує чи істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії.

Регресійний аналіз використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія.

Математико-статистичний апарат, що дозволяє встановити вид кореляційної залежності **називається регресійним аналізом**, а функція, яка описує цю залежність, називається **рівнянням регресії**.

Кореляційний аналіз - статистичний метод, що дозволяє визначити, чи існує залежність між змінними і наскільки вона сильна.

Регресійний аналіз - статистичний метод, який використовується для опису характеру зв'язку між змінними (позитивна чи негативна, лінійна або нелінійна залежність) і прогнозування значень залежної змінної при зміні значень незалежної змінної.

Порівняння коефіцієнтів кореляції і регресії:

Коефіцієнт кореляції

- Приймає значення в діапазоні від -1 до +1
- Безрозмірна величина
- Показує силу зв'язку між ознаками
- Знак коефіцієнта говорить про напрямок зв'язку

Коефіцієнт регресії

- Може приймати будь-які значення
- Прив'язаний до одиниць виміру обох ознак
- Показує структуру зв'язку між ознаками
- Знак коефіцієнта говорить про напрямок зв'язку

Змінна X , на підставі якої робляться прогнози, називається незалежною (факторна ознака). Змінна Y , значення якої визначаються значеннями X , називається залежною (результуюча ознака).

Завдання, які вирішуються за допомогою регресійного аналізу:

- Побудова моделей, що пояснюють механізм впливу факторних ознак на результат;
- Статистичний прогноз - обчислення значення результативної змінної при різних значеннях факторів;

- Графічне прогнозування поведінки однієї змінної в залежності від зміни іншої.

Етапи регресійного аналізу:

- Виявлення наявності взаємозв'язку між ознаками і її форми.
- Побудова регресійної моделі: лінія регресії і рівняння регресії.
- Прогноз (цей етап виконується не завжди).

Залежно від типу розв'язуваної задачі вибирається вид регресійного аналізу:

- Лінійний регресійний аналіз (для метричних ознак):
 - проста лінійна регресія (одна залежна змінна, одна незалежна змінна);
 - множинна лінійна регресія (одна залежна змінна, дві чи більше незалежних змінних);
- Нелінійний регресійний аналіз (для метричних ознак);
- Бінарна логістична регресія (залежна змінна - категоріальна);
- Мультиномінальна логістична регресія (залежна змінна - категоріальна);
- Порядкова регресія;
- Пробіт-аналіз та ін.

Незалежні змінні можуть бути *категоріальними, дихотомічними* або *фіктивними*.

Фіктивна змінна - дихотомічна змінна, введення якої дозволяє:

- розподілити досліджувану сукупність на два ряди, щоб не будувати дві окремі моделі регресії для кожного з них, а одразу порівняти ці дві групи між собою;
- включити до побудови регресійних моделей змінні, виміряні недихотомічними номінальними шкалами.

Види регресії в залежності від типів шкал ознак:

Вимірювання залежної змінної	Вимірювання незалежної(их) змінної(их)		
	Усі X метричні або інтервальні	Усі X порядкові	Деякі інтервальні або метричні, деякі – порядкові або номінальні
Інтервальна або метрична	Лінійна регресія	Лінійна регресія з використанням фіктивних змінних	Лінійна регресія з використанням фіктивних змінних

Порядкова	Множинна логістична регресія	Порядкова регресія	Множинна логістична регресія з використанням фіктивних змінних
Номінальна недихотомічна	Множинна логістична регресія	Множинна логістична регресія з використанням фіктивних змінних	Множинна логістична регресія з використанням фіктивних змінних
Номінальна дихотомічна	Бінарна логістична регресія	Бінарна логістична регресія з використанням фіктивних змінних	Бінарна логістична регресія з використанням фіктивних змінних

Оскільки найбільш простою формою залежності в математиці є пряма, то в кореляційному і регресійному аналізі найбільш популярні лінійні моделі.

Приклади:

Необхідно встановити, як зумовлює утворення респондента його вибір того чи іншого кандидата на виборах - проста лінійна регресія, тому одна незалежна змінна (освіта).

Необхідно встановити як впливає освіту, матеріальне становище і вік респондента на вибір того чи іншого кандидата на виборах - множинна лінійна регресія, так як кілька незалежних змінних (освіта, матеріальне становище, вік).

Типи змінних, що беруть участь в лінійному регресійному аналізі:

- залежна змінна (інтервальна або порядкова)
- незалежні змінні (інтервальна, порядкова, дихотомічна)

10.2. Приклад розрахунку множинної лінійної регресії в пакеті SPSS

Розглянемо приклад розрахунку множинної лінійної регресії в пакеті SPSS. Опитування присвячено вивченню життєвого успіху чоловіків і жінок Білорусії 1997 р. В опитуванні взяло участь 300 осіб.

Завдання: перевірити як фінансове становище респондентів, наявність роботи та оцінка власного впливу на хід свого життя позначаються на загальній задоволеності життям.

Змінні, які взяли участь в аналізі:

- залежна змінна - «Задоволеність життям» - інтервальна

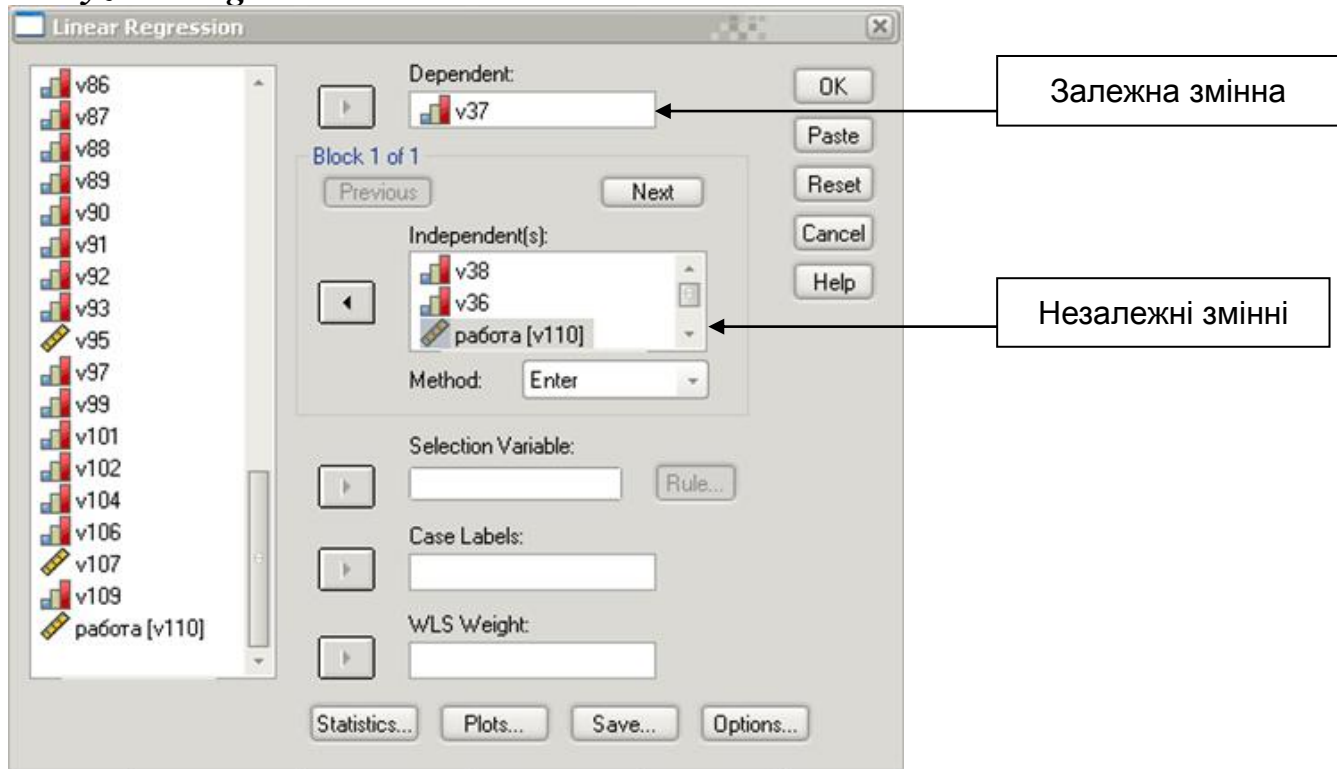
V37 Наскільки Ви задоволені своїм життям?

Не задоволений (а) 1 2 3 4 5 6 7 8 9 10 Задоволений (а)

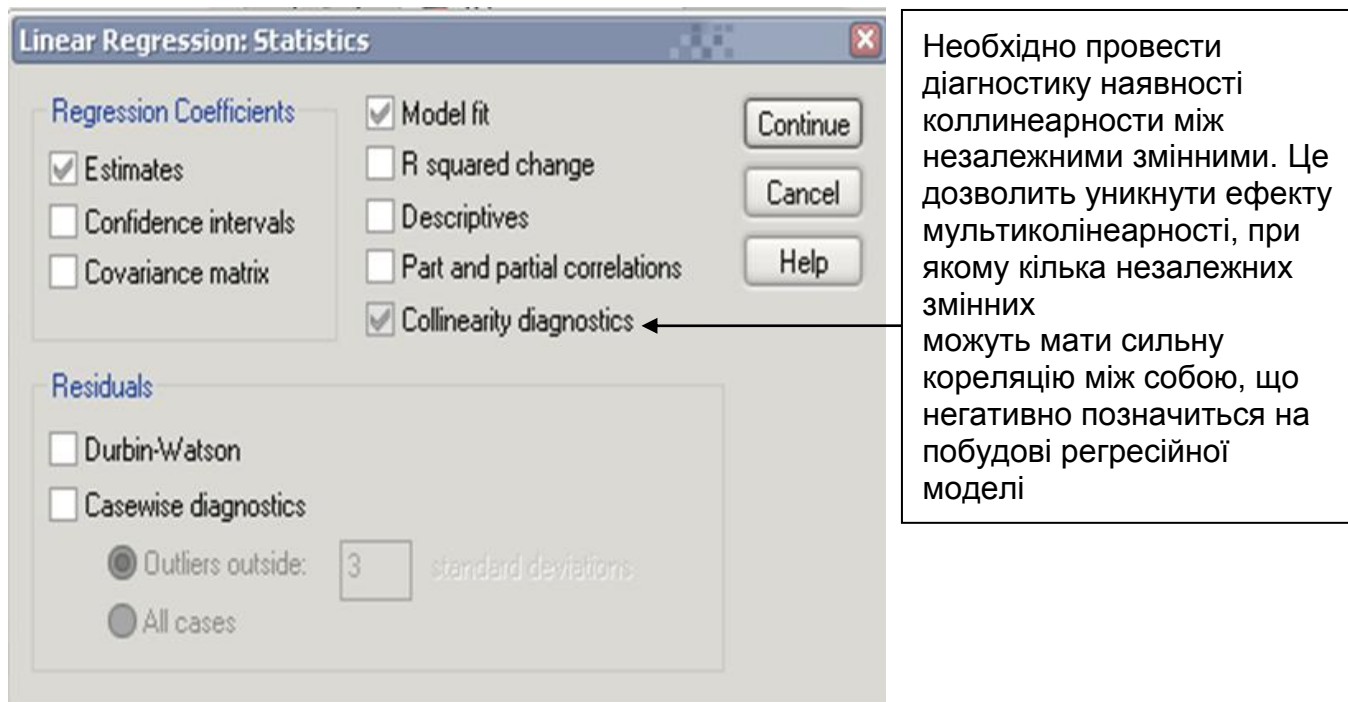
- незалежні змінні - v36 «задоволеність фінансовим становищем сім'ї» (інтервальна), v38 «ступінь впливу на хід свого життя» (інтервальна), «наявність роботи» - нова змінна, отримана шляхом перекодування v99 (чи працюєте Ви в теперішній час), де альтернатива «1» відповідає наявності роботи у респондента, а альтернатива «0» - її відсутності.

Процедура лінійного регресійного аналізу:

Analyze → *Regression* → *Linear...*



Вікно Statistics:



Інтерпретація результатів:

Перша таблиця Variables Entered / Removed (b) вказує на метод регресійного аналізу і на змінні, які брали участь в ньому, і не підлягає інтерпретації.

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	робота ^a v38, v36	.	Enter

a. All requested variables entered.

b. Dependent Variable: v37

У таблиці Model Summary представлений коефіцієнт множинної кореляції R між залежною і незалежними змінними, який вимірюється від 0 до 1 і показує наскільки добре обрані незалежні змінні здатні визначати поведінку залежної. У нашому випадку $R = 0,554$ Коефіцієнт детермінації R^2 показує яку частку дисперсії залежної змінної пояснюють незалежні змінні. $R^2 = 0,307$ - це означає, що даній регресійної моделлю описано тільки 30,7% випадків, що є нормальним показником для соціальних наук.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.554 ^a	.307	.300	1.979

a. Predictors: (Constant), робота, v38, v36

b. Dependent Variable: v37

У таблиці ANOVA (b) представлені результати дисперсійного аналізу, які вказують на значимість розглянутих вище коефіцієнтів множинної кореляції і детермінації. У нашому випадку $\text{Sig.} = .000$, що говорить про значне відміну коефіцієнтів від 0.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	513.225	3	171.075	43.668	.000 ^a
	Residual	1159.612	296	3.918		
	Total	1672.837	299			

a. Predictors: (Constant), работа, v38, v36

b. Dependent Variable: v37

У таблиці Coefficients (a) містяться основні практичні результати, на підставі яких можна побудувати змістовну модель.

Показник VIF вказує на стан мультиколінеарності, тобто на сильну кореляцію незалежних змінних між собою (що унеможливорює подальшу інтерпретацію регресійної моделі). Величина даного показника не повинна перевищувати число 10. У нашому випадку даній ситуації не спостерігається.

Стовпець Beta містить стандартизовані коефіцієнти регресії, а стовпець Sig. вказує на рівень значущості коефіцієнтів.

З таблиці видно, що незначущим є коефіцієнт, що відноситься до змінної «Наявність роботи» (так як Sig.> 0,05). З цього можна зробити висновок, що наявність у респондентів роботи не впливає на задоволеність життям.

Решта коефіцієнти є значущими. Отже, задоволеність фінансовим становищем (V36) має сильний прямий вплив на загальну задоволеність життям, тобто чим вище задоволеність фінансовим становищем сім'ї, тим вище задоволеність життям.

Слабкіший вплив на задоволеність життям має оцінка можливості впливати на хід свого життя (V38), тобто чим вище респонденти оцінюють вплив на своє життя, тим сильніше їх задоволеність своїм життям.

Нестандартизовані коефіцієнти регресії (стовпець B) і константа (в нашому випадку вона вище 1) використовуються для побудови рівняння регресії, і показують, яка зміна значення залежної змінної відбувається при одиничній зміні відповідної незалежної змінної.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.310	.312		4.196	.000		
	v38	.157	.044	.177	3.578	.000	.962	1.040
	v36	.525	.053	.485	9.829	.000	.960	1.042
	робота	.231	.230	.049	1.004	.316	.987	1.013

a. Dependent Variable: v37

Рівняння регресії будується на основі таблиці Coefficients:
”Задоволеність життям” = $1.31 + 0,157 * V38 + 0,525 * V36 + 0,231 * \text{робота}$, де
V36 - ”фінансовий стан”, V38 - ”самооцінка впливу на хід власного життя”.

10.3. Приклад застосування регресійного аналізу: визначення факторів довіри до політичних інститутів в Україні

Регресійний аналіз може використовуватися як засіб збагачення соціологічних теорій. Теорії середнього рівня покликані узагальнювати і структурувати емпіричні дані в межах окремих областей соціологічного знання. Ми висунули гіпотезу щодо наявності ендогенних та екзогенних довіри до політичних інститутів. Ці групи факторів виокремлені на базі культурологічних та інституційних підходів: у культурологічних теоріях робиться наголос на загальних звичаях, цінностях та нормах соціальної групи як факторах довіри (що інкорпуються протягом соціалізації) – екзогенні фактори, у інституційних теоріях акцентується увага на довірі як продукті раціонального та осмисленого вибору, заснованого на індивідуальній та колективній оцінці ефективності політичних інститутів – ендогенні фактори.

Завдяки емпіричним соціологічним дослідженням, які іноді називають третім, нижнім, рівнем соціологічних теорій, які виводять теоретичні закономірності з конкретних даних, ми мали змогу перевірити цю гіпотезу. Так, для виявлення факторів довіри до політичних інститутів були використані масиви **міжнародних досліджень «Євразійський барометр»** («Умови життя, спосіб життя і здоров'я населення країн колишнього СРСР», 2000-2004, n=2400; «Здоров'я в перехідний період: тенденції у здоров'ї населення та політика в сфері охорони здоров'я в країнах СНД», 2009-2013, n=2000) та **міжнародних досліджень цінностей** (третя (1994-1998, n=2811), п'ята (2005-2009, n=1000) і шоста (2010-2014, n=1500) хвилі дослідження світових цінностей; третя (1999-2001, n=1195) хвиля дослідження європейських цінностей).

У якості метод аналізу даних була використана лінійна регресія методом «покрокового виключення», який дозволяє відстежувати ступінь значущості впливу змінних і, за підсумками покрокового відбору, відсікати предиктори, вплив яких не значущий (в якості рівня значущості застосовується значення 0,05).

Залежна змінна - агрегований показник «Довіра до політичних інститутів». Він побудований на підставі 5 ознак, які вимірюють рівень довіри на національному рівні - інститутам президентства, уряду, парламентаризму, політичним партіям; на регіональному рівні - інститутам місцевих органів влади. Агрегований показник може набувати значень від -1 до +1, де «-1» характеризує повну недовіру представленій сукупності політичних інститутів, а «+1» - повну довіру, відповідно значення «0» характеризує нейтральне ставлення.

Незалежні змінні (предиктори) - набір потенційних (теоретично обґрунтованих) ендогенних і екзогенних факторів довіри до політичних інститутів. Ендогенні - внутрішні щодо політичній системі фактори довіри до політичних інститутів. Екзогенні - зовнішні щодо політичній системі фактори довіри до політичних інститутів.

Потенційні **ендогенні фактори** довіри до політичних інститутів:

- 1) рівень політичної активності респондентів;
- 2) наявність політичних свобод;
- 3) оцінка політичних режимів;
- 4) краща система правління (комуністична система, управління країною армією, монархія, сильний лідер при розпуск парламенту і скасування вільних виборів, демократія)
- 5) задоволеність функціонуванням основних державних механізмів (розвитком демократії в країні, розвитком економіки, системою освіти, системою охорони здоров'я, тим, як уряд справляється зі своїми обов'язками, тим, як вирішуються питання регіонального рівня місцевими органами влади, системою соціального захисту в країні);
- 6) оцінка діяльності органів влади країни в динаміці (оцінка за п'ятибальною шкалою ситуації за радянських часів, сьогодні і через 10 років);
- 7) членство в політичних партіях (дихотомічна змінна);
- 8) ступінь важливості політики в житті, інтерес до політики, частота відстеження політичних подій в ЗМІ.

Потенційні **екзогенні фактори** довіри до політичних інститутів:

- 1) соціально-демографічні характеристики опитаних (вік, освіта, матеріальне становище в динаміці);
- 2) ступінь задоволеності матеріальним становищем на сьогоднішній день;
- 3) психологічний стан респондентів, який визначається відповідно до відповідей на питання про загальної задоволеності життям, віднесенням себе за «шкалою щастя» до тієї чи іншої позиції (від «1» - «Зовсім нещасний» до «10» - «Абсолютно щасливий»);
- 4) релігійна ідентичність (дихотомічна змінна);
- 5) ступінь генералізованої довіри респондентів (питання звучить наступним чином: «Якою мірою Ви згодні з твердженням, що більшості людей можна довіряти?»);
- 6) активність членства в громадських організаціях (дихотомічна змінна);
- 7) приналежність до релігійної конфесії (дихотомічна змінна);
- 8) стан економіки України в динаміці;
- 9) потреба в кардинальній зміні економічної системи.

Опис процедури. З набору потенційних ендогенних і екзогенних факторів довіри до політичних інститутів покроково виділяється спочатку ті, які емпірично демонструють найбільшу кореляцію з залежною змінною «Довіра до політичних інститутів». Далі з числа предикторів, що залишилися (потенційних факторів довіри до політичних інститутів), вибираються той, який має другий за значенням коефіцієнт кореляції (у регресійній моделі - стандартизований коефіцієнт β) і т.п. Процес триває до тих пір, поки не

будуть відібрані всі предиктори, які істотно впливають на залежну змінну. За допомогою цієї процедури ми отримуємо спочатку перелік ендогенних, потім екзогенних факторів, які надають статистично значущий вплив на змінну «Довіра до політичних інститутів» в Україні і будуть надалі задіяні при побудові підсумкової моделі.

За результатами регресійного аналізу для України актуальними виявилися такі фактори. **Ендогенні:** ступінь задоволеності тим, як уряд виконує свої обов'язки, оцінка існуючої політичної системи, а також ступінь задоволеності розвитком демократії в країні, актуальна оцінка діяльності органів влади, оцінка рядовими громадянами своєї здатності впливати на уряд країни, оцінка альтернативних існуючої політичних систем і ступінь важливості політики в життя людини. **Екзогенні фактори:** прогноз щодо стану економіки країни і матеріального благополуччя своєї родини, оцінка попереднього стану економіки, оцінка поточного рівня власного матеріального становища, ступінь генералізованої довіри, вік і рівень освіти респондента.

10.4. Проблеми, пов'язані з можливістю застосування регресійного аналізу

1. Класичний регресійний аналіз призначений для *метричних шкал*.
2. *Мультиколінеарність*. Основний недолік регресійної моделі в разі мультиколінеарності - нестійкі значення коефіцієнтів моделі.
3. *Гетероскедастичність*. При побудові регресійного рівняння треба враховувати одне з обмежень цього методу - вимога гомоскедастичності, або друга умова Гауса-Маркова. Суть цього обмеження проста: розкид точок навколо лінії регресії повинен бути досить рівномірний по всій протяжності лінії ікс.

Способи вирішення вказаних проблем:

1. *Використання фіктивних змінних* (Прочитати: Малхотра Н.К. Маркетинговые исследования. Практическое руководство. – М.:Изд.дом «Вильямс», 2002. – С. 672 (Регрессия с использованием фиктивных переменных)
Аргунова К.Д. Качественный регрессионный анализ в социологии. Методическое пособие. М., 1990.
Марчук А.А. Учет качественных параметров в регрессионном анализе. <http://www.rusvs.ru/data/files/articles/dummyvar.pdf>).

2. *Мультиколінеарність* - існування лінійної залежності або сильної кореляції між незалежними змінними, які беруть участь у побудові регресійного рівняння.

Мультиколінеарність є проблемою лише в разі множинної регресії!

Основні наслідки мультиколінеарності:

1. Знижується точність оцінок параметрів, яка проявляється в зростанні помилок деяких оцінок, в значному збільшенні дисперсії оцінок параметрів;
2. Оцінки деяких параметрів стають статистично незначущими;

3. Оцінки деяких параметрів стають чутливими до обсягів сукупності спостережень.

Існує кілька ознак, за якими може бути встановлена наявність мультиколінеарності:

1) Коефіцієнт детермінації R^2 досить високий, але деякі з коефіцієнтів регресії статистично незначущі, тобто вони мають низькі t-статистики.

2) Парна кореляція між малозначущими пояснюючими змінними досить висока. Однак цей показник буде надійним лише в разі двох пояснюючих змінних. При більшій їх кількості доцільнішим є використання приватних коефіцієнтів кореляції.

3) Високі приватні коефіцієнти кореляції. Приватні коефіцієнти кореляції визначають силу лінійної залежності між двома змінними, без урахування впливу на них інших змінних, в той час як парні коефіцієнти кореляції цей вплив враховують. Тому при вивченні багатовимірних зв'язків в ряді випадків парні коефіцієнти кореляції можуть давати абсолютно невірні уявлення про характер зв'язку між двома змінними. Наприклад, між двома предикторами X_1 і X_2 може бути високий позитивний парний коефіцієнт кореляції не тому, що один з них стимулює зміну іншого, а тому, що вони обидва змінюються в одному напрямку під впливом інших змінних, як врахованих в моделі, так і, можливо, неврахованих. Тому необхідно вимірювати дійсну силу лінійного зв'язку між двома предикторами, очищену від впливу на розглянуту пару предикторів інших факторів. Коефіцієнт кореляції між двома змінними, очищений від впливу інших змінних, - приватний (рос. частный) коефіцієнт кореляції.

3. У статистиці, послідовність випадкових величин називається *гетероскедастичною*, якщо випадкові величини мають різну дисперсію. Термін означає «різна дисперсія» і походить від грецького слова «гетеро» («інший») і «skedasis» («дисперсії»). На відміну, послідовність випадкових величин називається гомоскедастичною, якщо вона має постійну дисперсію.

Що робити у випадках гетероскедастичності?

- 1) Використовувати зважений метод найменших квадратів - LS (w).
- 2) Перевизначити змінні.
- 3) Обчислення стандартних помилок робити з поправкою на гетероскедастичність (метод Уайта) - LS (h).

Література до теми

Обов'язкова література

Дрейлер Н., Смит Г. *Прикладной регрессионный анализ* / Нормана Р. Дрейпер, Гарри Смит. – М.: Диалектика, 2007. – 912 с.

Паніотто В.І., Максименко В.С., Марченко Н.М. *Статистичний аналіз соціологічних даних*. – К.: Вид.дім “КМ Академія”, 2004. С. 144-169.

Додаткова література

Кузіна, Ірина. (2015). Довіра до політичних інститутів в Україні: концептуалізація феномену та фактори формування [дис.канд.соціол.наук], <http://dissertations.karazin.ua/sociology/resources/2b73e117fd1597d804af05a2a81a527c.pdf>.

Кузьминова Т.В. *Моделирование динамики безработицы (построение регрессионных и нейросетевых моделей динамики безработицы в России с учетом временных лагов)* // Социология: 4М, 2003. – №16.

Пациорковский В.В., Пациорковская В.В. SPSS для социологов. Учебное пособие. ИСЭПН РАН. – М., 2005. – С. 275-292.

Ротмистров А. Н., Толстова Ю. Н. Проблемы построения нелинейных регрессионных моделей в социологии: номинальные шкалы, синергетические эффекты, поиск эффективной системы предикторов. *Математическое моделирование социальных процессов*. 2014. № 16. С. 159-178.

Черчилль Г.А. Маркетинговые исследования. – СПб: Питер, 2001. С.642-651.

Питання для самоконтролю

1. В чому полягає сутність регресійного аналізу? З якою метою він застосовується?
2. Назвіть етапи регресійного аналізу.
3. Які ви можете назвати проблеми, пов'язані з можливістю застосування регресійного аналізу?
4. Які існують види регресії в залежності від типів шкал ознак?