# Лекція Тема 8. АНАЛІЗ ЗАКОНОМІРНОСТЕЙ ДЛЯ ЯКІСНИХ ДАНИХ

## 8.1. Особливості аналізу якісних даних. Сутність критерію «хі-квадрат»

Статистичні методи, які розглядалися у попередніх розділах даного підручника, призначені для аналізу кількісних ознак. Однак, багато ознак неможливо виміряти числом. Наприклад, можна бути чоловіком або жінкою, студентом або ні. Можна бути лікарем, юристом, геологом та ін. Тут ми маємо справу з якісними (атрибутивними) ознаками. Ці ознаки не пов'язані між собою жодними арифметичними співвідношеннями, впорядкувати їх не можна (будь-які арифметичні дії з якісними ознаками виконувати не можна, навіть якщо вони проранжовані). Єдиний спосіб описання якісних ознак полягає в тому, щоб підрахувати кількість об'єктів, що мають одне і те ж значення. Крім того, можна підрахувати, яка частка загальної кількості об'єктів має те чи інше значення.

Якісні (нечислові) змінні традиційно мають широке застосування в соціологічних опитуваннях, у ході яких збирають інформацію шляхом опитувань, проведення інтерв’ю тощо.

Останнім часом такі змінні набувають усе більшого розповсюдження і в економічних дослідженнях. Причиною тому, перш за все, є прагнення врахувати людський фактор у бізнесі. У всьому світі проводяться бізнес-опитування керівників підприємств, менеджерів щодо ситуації як на самому підприємстві, так і в навколишньому ринковому середовищі. Це можуть бути опитування, що здійснюються з метою порівняння нинішнього і попередніх (наприклад, минулих років) рівнів основних фінансово-виробничих показників, або їх можливих змін у майбутньому. Можуть проводитися також опитування щодо змін умов для здійснення бізнесу в регіоні, які безпосередньо впливають на обсяг продаж продукції підприємства. При цьому варіанти відповідей в анкетах зазвичай бувають такими: «ситуація не змінилася», «відбулося покращення ситуації», «відбулося погіршення ситуації», «складно відповісти».

Такий підхід до формулювання варіантів відповідей є особливо актуальним у сучасних умовах, коли суттєвий вплив на отримання інформації має так звана «комерційна таємниця». Так, під час

здійснення опитування менеджерів щодо обсягів розрахунків з певними постачальниками або покупцями, респонденти зазвичай не мають бажання озвучувати конкретні цифрові дані, а віддають перевагу відповідям, які лише окреслюють ті чи інші зміни ситуації: покращилася, погіршилася тощо.

Здійснюються бізнес-опитування й серед населення. Так, опитування щодо вибору переважного продукту із списку відомих дозволяє приймати управлінські рішення щодо випуску нових для підприємства видів продукції. Опитування населення щодо переваг подорожування (в межах країни, до країн Європи, Азії тощо) є підставою для прийняття управлінських рішень у фірмах, що займаються наданням туристичних послуг.

Крім того, постійний моніторинг змін, що відбуваються останнім часом на регіональних або національному ринках, призвів до необхідності розробки певних методик здійснення рейтингових оцінок різних суб’єктів господарювання: банківських установ, фірм, навчальних закладів тощо. Для цього бізнес-структурам присвоюються певні ранги. Очевидно, що ранги, в основу яких може бути покладена низка показників, не можна вважати звичайними кількісними змінними, оскільки сказати, наприклад, що другий ранг є величною в два рази більшою (кращою), ніж перший ранг, не можна.

У деяких випадках разом із новими анкетами респонденти отримують результати їх попередніх опитувань. Така практика є особливо корисною в наступних випадках:

по-перше, для того, щоб з’ясувати чи змінилися погляди опитуваних щодо поставлених проблем;

по-друге, для підвищення ступеня відповідальності (уникнення формального підходу) до оцінки досліджуваної проблеми;

по-третє, для більш оперативного проведення дослідження, оскільки в сучасних умовах інформація зазвичай дуже швидко

«старіє»;

по-четверте, для порівняння змін варіантів відповідей останнього опитування з попередніми, що відбулися внаслідок тих чи інших об’єктивних причин.

Оскільки якісні дані не мають числового вираження, то основними питаннями, які виникають при аналізі таких даних, є:

1. підрахунок частот або часток, тобто визначення кількості одиниць в кожній групі або визначення їх кількості у відсотках (частках) до загальної кількості одиниць досліджуваної сукупності;
2. знаходження способу (методу) оцінити рівень точності, з якою частоти, які визначені за даними вибіркової сукупності, відповідають частотам у всій генеральній сукупності;
3. перевірка необхідної гіпотези (чи то рівності відсотків, чи то визначення наявності або відсутності взаємозв’язку між досліджуваними ознаками) за допомогою тих чи інших статистичних критеріїв.

Для аналізу якісних даних, представлених не числами, а категоріями, використовують систему статистичних критеріїв, серед яких особливе місце займає критерій «хі-квадрат» (  2 ).

Критерій «хі-квадрат» був запропонований Карлом Пірсоном ще в 1900 році. За допомогою даного критерію здійснюється оцінка значущості відмінностей між фактичною кількістю випадків або якісних характеристик (отриманих у результаті вибіркового спостереження), що потрапили в кожну категорію, і теоретичною їх кількістю, яку можна очікувати в отриманих групах.

Іншими словами, ***критерій (тест) «хі-квадрат»*** заснований на частотах, які являють собою кількість одиниць вибіркової сукупності, що потрапили в кожну з категорій якісної ознаки. Завданням критерію є встановлення суттєвості (або не суттєвості) відмінностей між фактичними частотами і очікуваними, опорними частотами (у випадку справедливості нульової гіпотези) [16, с. 879].

Розраховується ***критерій «хі-квадрат»*** за наступною формулою:



 2 

(*m*  *m*)2 *m*

, (8.1)

де *m* – фактичні частоти;

*m* – очікувані (опорні) частоти.

Необхідно звернути увагу на те, що під час розрахунку критерію підсумовування здійснюється по всіх категоріях або комбінаціях категорій.

Методика визначення очікуваних (опорних) частот залежить від того, яка саме гіпотеза в даному випадку перевіряється.

Наприклад, потрібно порівняти думку покупців про продукцію двох меблевих фірм і оцінити, чи дійсно існують відмінності між оцінками, які надаються продукції першої фірми, та оцінками, які відносяться до продукції другої фірми. Самі оцінки меблів повинні бути представлені у вигляді якісних характеристик, тобто повинні

виражатися позитивними і негативними епітетами: комфортна, красива, міцна чи незручна, груба, ненадійна та ін.

Для того, щоб зробити висновок про те, що продукція першої фірми більше подобається покупцям, ніж продукція другої фірми, необхідно довести, що позитивні оцінки зустрічаються статистично значно частіше в першій групі відгуків (про продукцію першої фірми), ніж у другій групі відгуків (про продукцію другої фірми). Здійснити порівняння частот тих чи інших якісних оцінок дозволяє виконати саме критерій «хі-квадрат».

Цей критерій цілком підходить і для порівняння того, наскільки отримана частота появи об'єктів збігається з очікуваною частотою їх появи. Таке порівняння необхідно здійснювати, наприклад, для того, щоб відповісти на питання про те, наскільки виправданою є якість прогнозів, зокрема, прогнозів погодних явищ: дощу, снігу, граду, урагану тощо. Але якщо групу очікуваних подій (прогнозів) прийняти в якості результатів спостережень у якихось уже реалізованих умовах, припустимо, в іншій країні, то це дозволить порівнювати частоти природних явищ двох країн, а не оцінювати точність прогнозів – у цьому полягає гнучкість використання критерію «хі- квадрат».

## 8.2. Використання критерію «хі-квадрат» щодо подібності розподілів

Критерій «хі-квадрат» щодо подібності розподілів (рівності відсотків) використовують для того, щоб відповісти на питання щодо того, чи з однаковою частотою зустрічаються значення ознаки в емпіричному і теоретичному розподілах або в двох і більше емпіричних розподілах (порівняння розподілів двох якісних ознак). Тобто, за допомогою критерію оцінюють, наскільки подібними є розподіли емпіричних і теоретичних частот. Якщо відмінності між ними незначні, то можна вважати, що відхилення емпіричних частот від теоретичних обумовлені лише випадковістю. Якщо ж, навпаки, ці розподіли виявляться достатньо різними, то можна вважати, що відмінності між ними значущі.

Таким чином, критерій «хі-квадрат» використовують для перевірки гіпотези про те, що не існують відмінності між фактичними відсотками, які побудовані за даними, взятими із деякої генеральної сукупності, і очікуваними (опорними) величинами.

Нагадаємо, що сутність і види статистичних гіпотез детально розглядаються у Розділі 4 підручника.

У даному випадку ***гіпотези*** формулюються так:

* нульова гіпотеза ( *H* 0 ): значення відсотків у досліджуваній сукупності відповідають (рівні) набору відомих, заданих опорних величин;
* альтернативна (дослідницька) гіпотеза ( *H*1 ): значення відсотків у досліджуваній сукупності не рівні заданому набору опорних величин, щонайменше для однієї з категорій існує відмінність.

Критерій «хі-квадрат» можна застосовувати лише у випадку виконання двох умов:

* 1. Отримані дані є випадковою вибіркою з досліджуваної генеральної сукупності.
  2. Загальна кількість спостережень (обсяг вибірки) повинна бути не меншою 50, а частоти в кожній групі повинні бути не меншими п’яти.

У формулі (7.1), яка використовується для розрахунку критерію

«хі-квадрат», фігурують так звані «очікувані (опорні) частоти». Ми вже відмічали, що спосіб визначення цих частот залежить від сутності гіпотези, яка перевіряється. У даному випадку ***очікуване значення частоти*** для кожної категорії розраховується так:

*m*  *z*  *n* , (8.2)

де *z* – задані, опорні значення *часток* в генеральної сукупності;

*n* – обсяг (кількість) вибірки.

Після розрахунку критерію його значення порівнюють із критичним (табличним) значенням. При цьому зазвичай перевірка здійснюється з 95%-м рівнем довірчої ймовірності, тобто рівень надійності (значущості) складає 5 % (100 – 95).

Оскільки під час визначення критерію «хі-квадрат» здійснюється підсумовування співвідношень (формула 7.1), то величина критерію залежить від кількості доданків: за інших рівних умов чим більшою є їх кількість, тим більшим буде значення самого критерію, оскільки кожен із доданків вносить свій вклад у загальну суму. Із цього випливає, що для кожної кількості незалежних змінних критерій «хі-квадрат» буде мати власний розподіл.

Слід визначитися з тим, що таке кількість незалежних змінних. Вона завжди на одиницю менша, ніж кількість категорій номінальної ознаки. Чому це так? Тому що коли ми маємо певну вибіркову

сукупність, для якої підраховано суму частот, то одну із частот ми можемо завжди визначити шляхом віднімання від загальної їх кількості решти інших частот. Це інакше називають кількістю ступенів свободи.

Отже, в даному випадку ***кількість ступенів свободи***

визначається наступним чином:

*k*  *l* 1, (8.3)

де *l* – кількість категорій (варіантів) досліджуваної ознаки.

Якщо значення «хі-квадрат» більше критичного значення з таблиці розподілу «хі-квадрат» для відповідної кількості ступенів свободи, то це є доказом того, що отримані частоти суттєво відрізняються від очікуваних частот, виходячи з опорних значень відсотків. У такому випадку слід відхилити нульову гіпотезу і прийняти досліджувану (альтернативну) гіпотезу, зробивши висновок про те, що отримані вибіркові відсотки значно (суттєво) відрізняються від заданих опорних значень.

Якщо значення критерію «хі-квадрат» менше критичного значення з таблиці розподілу «хі-квадрат», то отримані частоти не значно відрізняються від частот, які можна було б очікувати виходячи з опорних значень відсотків. У такому випадку слід прийняти нульову гіпотезу як допустиму і зробити висновок, що отримані вибіркові відсотки не мають значних відмінностей від заданих опорних значень.

Розглянемо розрахунок та перевірку критерію «хі-квадрат» на такому прикладі.

У відділі чоловічого взуття магазину протягом тижня проводилося наступне спостереження. Для кожного потенційного клієнта, що заходив до відділу магазину, фіксувався бажаний ним тип

«літнього» взуття. У табл. 8.1 наведені дані за минулий тиждень, а також відповідні відсотки в цей же час у минулому році.

Необхідно:

1. оцінити зміну попиту на окремі види взуття шляхом порівняння відсотків;
2. розглядаючи відсотки за минулий рік як точні, визначити і оцінити критерій «хі-квадрат»;
3. надати адміністрації магазину поради щодо необхідності прийняття управлінських рішень.

*Таблиця 8.1*

## Результати опитування клієнтів за останній тиждень і дані за минулий рік

|  |  |  |
| --- | --- | --- |
| Тип взуття | Кількість клієнтів | |
| частоти за останній тиждень, одиниць | дані за минулий рік,  % |
| Кросівки | 124 | 16,2 |
| Черевики типу «босоніжки» | 256 | 26,4 |
| Шкіряні черевики | 302 | 38,5 |
| Черевики із замінників шкіри | 138 | 18,9 |
| Усього | 820 | 100,0 |

Для відповіді на перше запитання потрібно визначити частки за останній тиждень (результати наведено в табл. 8.2).

*Таблиця 8.2*

## Частки за останній тиждень та минулий рік

|  |  |  |
| --- | --- | --- |
| Тип взуття | Кількість клієнтів, у % до загальної кількості | |
| дані за останній тиждень | дані за минулий рік |
| Кросівки | 15,1 | 16,2 |
| Черевики типу «босоніжки» | 31,6 | 26,4 |
| Шкіряні черевики | 36,5 | 38,5 |
| Черевики із замінників шкіри | 16,8 | 18,9 |
| Усього | 100,0 | 100,0 |

Як видно із таблиці 7.2, відбулися певні зміни в структурі попиту на взуття: зросла частка тих, хто надає перевагу черевикам типу «босоніжки», у той же час зменшилася частка тих, хто бажає купувати шкіряні черевики, інші зміни були незначними. Але чи є суттєвими всі ці зміни, для того щоб пропонувати адміністрації змінювати структуру закупівель? На це питання краще відповісти шляхом перевірки тесту «хі-квадрат».

Для цього спочатку необхідно знайти очікувані частоти (формула 7.2). Результати розрахунків наведено в табл. 7.3.

Слід звернути увагу на те, що сума очікуваних частот повинна дорівнювати сумі фактичних частот (у даному випадку – 820).

Маючи очікувані частоти можемо визначити критерій «хі- квадрат» за формулою 8.1:

 2  (124 132,8)2

132,8

 (256  216,5)2

216,5

 (302  315,7)2

315,7

 (138 155,0)2

155,0

 10,2

*Таблиця 8.3*

## Розрахунок очікуваних частот за минулий рік

|  |  |
| --- | --- |
| Тип взуття | Очікувані частоти |
| Кросівки | 820 \* 0,151 = 132,8 |
| Черевики типу «босоніжки» | 820 \* 0,316 = 216,5 |
| Шкіряні черевики | 820 \* 0,365 = 315,7 |
| Черевики із замінників шкіри | 820 \* 0, 168 = 155,0 |
| Усього | 820,0 |

Для того, щоб порівняти отримане значення критерію «хі- квадрат» із табличним, необхідно знайти кількість ступенів свободи за формулою 8.3:

*k*  4 1  3

Використовуючи таблиці розподілу «хі-квадрат» (які можна легко знайти в Інтернеті) визначаємо, що для 3-х ступенів свободи і рівня значущості 5 % його значення складає 7,8.

Оскільки фактичне значення статистики "хі-квадрат" значно більше табличного, то нульова гіпотеза відхиляється (отримані частки істотно відрізняються від опорних значень відсотків). Тому можна пропонувати адміністрації магазину внести зміни в структуру продажів.

## 8.3. Використання статистичних критеріїв

**для оцінки тісноти зв'язку між якісними змінними**

При наявності двовимірного набору якісних даних (коли є інформація по 2-х і більше змінних для кожного об’єкта) можна перевірити наявність зв'язку між змінними. Вважають, що дві якісні змінні є **незалежними**, якщо знання значення однієї з них не допомагає передбачити значення іншої. Іншими словами, ймовірності однієї змінної такі ж, як умовні ймовірності при заданих опорних значеннях іншої змінної.

Наприклад, уявімо, що в генеральній сукупності відсоток браку при фрезеруванні деталей у бригаді механічного цеху складає 2 %. Однак, коли працює бригада Іванова, відсоток браку становить 7 %. У цьому випадку знання значення однієї змінної (прізвище конкретного

бригадира) допомагає спрогнозувати значення іншої змінної (обсяг браку), оскільки 2 % і 7 % суттєво відрізняються між собою. Поява браку є більш імовірною під час роботи бригади Іванова і менш імовірною, коли працює будь-яка інша бригада. Отже, ці дві змінні не є незалежними.

Вимірювання тісноти зв'язку між якісними змінними засновано на побудові так званих частотних таблиць, які інакше називають таблицями спряженості.

Таблиця спряженості дещо відрізняється від звичайної таблиці розподілу, яка містить інформацію про категорії (значення) ознаки і частоти (або частки) їх виникнення. Таблиця спряженості – це таблиця спільного розподілу частот для двох категоріальних змінних, вона містить вибіркові оцінки ймовірнісних розподілів двох випадкових величин.

Слід відмітити, що таблиця спряженості є найбільш універсальним засобом вивчення статистичних зв'язків, оскільки в ній можуть бути представлені змінні з будь-яким способом вимірювання. Сам термін «таблиця спряженості» зобов'язаний своїм походженням тій обставині, що на основі аналізу подібних таблиць можна робити висновки щодо пов'язаності (спільної появи) значень одних категоріальних змінних разом з деякими значеннями інших змінних. Зв'язок між номінальними змінними, власне кажучи, і виражається у вигляді подібних сполучень.

При побудові таблиці спряженості рядкам відповідають категорії однієї змінної, а стовпчикам – іншої змінної. При цьому бажано, щоб групи явищ, які виділяються як причинний фактор, розміщувалися у лівій частині таблиці (рядки); а явища, які трактуються як наслідок дії фактору, – у верхній частині таблиці (стовпчики).

Зовні таблиця спряженості має такий вигляд (табл. 8.4).

*Таблиця 8.4*

## Таблиця спряженості

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Градації за першою ознакою | Градації за другою ознакою | | | | ∑ |
| 1 | 2 | … | m |
| 1 | *a*11 | *a*12 | *…* | *a*1*m* |  |
| 2 | *a*21 | *a*22 |  | *a*2*m* |  |
| … | **…** | **…** | *…* | *…* |  |
| n | *an*1 | *an*2 | *…* | *anm* |  |
| ∑ |  |  |  |  |  |

Останнім часом у економічній літературі все частіше використовується розширене розуміння таблиці спряженості. Пропонується в якості її елементів наводити не лише частоти, але й інші показники: наприклад, у клітинках статево-вікової таблиці розподілу населення можна наводити середній рівень доходів людей, які мають дану стать і вік, що суттєво розширює можливості проведення аналізу.

За таблицями спряженості зазвичай обчислюються коефіцієнти взаємної спряженості, засновані на тестовій статистиці «хі-квадрат». У тому випадку, коли не можна виконати умови застосування статистики хі-квадрат, рекомендується користуватися так званими теоретико-інформаційними (експертними) способами вимірювання зв'язків.

Кореляцію між порядковими змінними можна вимірювати за допомогою коефіцієнтів рангової кореляції, широко використову- ваними серед яких є коефіцієнти рангової кореляції Спірмена (  ) і Кендела ( ).

Ці коефіцієнти можуть бути використані для визначення щільності зв’язку як між кількісними, так і між якісними ознаками за умови, якщо їх значення будуть певним чином упорядковані або ранжовані.

***Коефіцієнт кореляції рангів (коефіцієнт Спірмена)*** розраховується за такою формулою (у випадку відсутності пов’язних рангів):

  1

6 *d* 2

*n*(*n*2 1) ,

*i*

(8.4)

де *i* - квадрат різниці рангів;

*d*

2

*n* – кількість рангів.

Коефіцієнт Спірмена може приймати будь-які значення в інтервалі [-1; 1]. Зв'язок між ознаками можна визнати статистично істотним, якщо значення коефіцієнту перевищує 0,5.

Розглянемо методику розрахунку коефіцієнта рангової кореляції Спірмена на наступному прикладі. Серед покупців молочних продуктів і фахівців з їх якості було здійснено опитування щодо оцінки рівня якості продукції шести підприємств виробників вершкового масла.

Результати опитування наведено в табл. 8.5.

*Таблиця 8.5*

## До розрахунку коефіцієнта рангової кореляції Спірмена

|  |  |  |  |
| --- | --- | --- | --- |
| Підприємства- виробники | Ранг продукції підприємства, який виставлено | | Квадрат різниці рангів *d* 2 |
| покупцями | фахівцями |
| А | 2 | 3 | 1 |
| Б | 3 | 2 | 1 |
| В | 1 | 1 | 0 |
| Г | 5 | 6 | 1 |
| Д | 6 | 4 | 4 |
| Ж | 4 | 5 | 1 |
| Усього | - | - | 8 |

Визначаємо Критерій Спірмена:

  1

6 8

6(62 1)

 0,8

Результати розрахунків свідчать про те, що між оцінками, які виставили фахівці та покупці продукції підприємств, є достатньо тісний зв'язок.

***Коефіцієнта рангової кореляції Кендала*** розраховується за наступною формулою:

де *n* – кількість рангів;

  2*S* ,

*n*(*n* 1)

(8.5)

*S* – сума балів, яка знаходиться як різниця двох складових:

*S*  *S* 1  *S*2 ,

де *S*1 – кількість рангів, які перевищують номер рангу, що відповідає

результативній ознаці;

*S* 2 – кількість рангів, які менші даного рангу результативної

ознаки.

Аналогічно коефіцієнту Спірмена, зв'язок між ознаками можна визнати істотним, якщо значення коефіцієнту рангової кореляції Кендала перевищує 0,5.

Визначимо коефіцієнт рангової кореляції Кендала за даними попереднього прикладу.

Для цього спочатку упорядкуємо підприємства за рангами, які їм виставили покупці (табл. 8.6).

*Таблиця 8.6*

## До розрахунку коефіцієнта рангової кореляції Кендала

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Підприємства- виробники | Ранг продукції підприємства, який виставлено | | *S*1 | *S* 2 |
| покупцями | фахівцями |
| В | 1 | 1 | 5 | 0 |
| А | 2 | 3 | 3 | 1 |
| Б | 3 | 2 | 3 | 0 |
| Ж | 4 | 5 | 1 | 1 |
| Г | 5 | 6 | 0 | 1 |
| Д | 6 | 4 | 0 | 0 |
| Усього | - | - | 12 | 3 |

Визначаємо суму різниці рангів (при цьому *S* розраховується за рангами, які виставлені фахівцями, оскільки ми впорядкували підприємства за рангами покупців):

*S*  12  3  9

Розраховуємо коефіцієнт Кендала:

  2  9 6(6 1)

 0,6

Як бачимо, коефіцієнт Кендала дав більш сувору оцінку тісноти зв’язку між рівнями якості продукції підприємств, які їм виставили покупці та фахівці (коефіцієнт Спірмена склав 0,8).

Для визначення щільності зв’язку між довільною кількістю ранжованих ознак використовується множинний коефіцієнт рангової кореляції – ***коефіцієнт конкордації*** (*W*), який визначається за формулою:

*W*

де *m* – кількість факторів;

 12*S* ,

*m*2  (*n*3  *n*)

(8.6)

*n* – кількість спостережень;

*S* – відхилення суми квадратів рангів від середньої квадратів рангів.

За наявності пов’язаних рангів формула коефіцієнта конкордації набуває наступного вигляду:

*W*  12*S* ,

*m*

*m*2  (*n*3  *n*)  *m*  *T*

*j*

*j* 1

(8.7)

де *T j*

 1 / 12

*m*



*j*  1

( *t* 3

 *t j* );

*t* – кількість пов’язаних рангів за окремими ознаками.

*j*

Коефіцієнт конкордації може набувати будь-яких значень в інтервалі [-1; 1].

Для оцінювання тісноти зв’язку між номінальними ознаками (які неможливо певним чином ранжувати) використовують ***критерій «хі- квадрат»***. Він дає можливість відповісти на питання, чи є дві змінні незалежними, чи ні, виходячи з таблиці спостережуваних частот, складеної для двовимірного набору якісних даних (таблиці спряженості).

Іншими словами, даний критерій перевіряє нульову гіпотезу щодо того, існує чи ні зв'язок між ознаками, які складають таблицю спряженості.

При цьому нульова і альтернативна ***гіпотези*** формулюються наступним чином:

* нульова гіпотеза ( *H* 0 ): дві досліджувані змінні не залежать одна від одної;
* альтернативна (дослідницька) гіпотеза ( *H*1 ): дві змінні пов'язані; іншими словами, вони не є незалежними одна від іншої.

***Очікувані частоти*** в даному випадку визначають так: для кожної комбінації категорій частоту однієї категорії множать на частоту іншої категорії і отриманий результат ділять на загальну кількість спостережень (обсяг вибірки). Це можна записати у вигляді формули:

*Частота категорії*  *Частота категорії* 



*однієї*

*змінної*

  

*другої*



*змінної*

*m*     

*п*

де *n* – кількість спостережень, тобто загальний осяг вибірки.

(7.8)

Чому саме так обчислюють очікувані частоти? Згадаймо, що відповідно до теорії ймовірності для двох незалежних подій імовірність того, що обидві вони настануть, дорівнює добутку ймовірностей цих подій.

При цьому повинні виконуватися наступні умови:

1. Набір даних повинен бути випадковою вибіркою, що отримана з досліджуваної генеральної сукупності.
2. У кожній клітинці таблиці спряженості очікувана частота повинна бути не меншою п'яти.

***Кількість ступенів свободи*** в даному випадку визначають як добуток частот кожної з категорій за мінусом одиниці:

*k*  (*l*1 1)  (*l*2 1) , (8.9)

де *l*1 *i l*2 – відповідно частота першої та другої категорій.

Оскільки використання ***критерію «хі-квадрат»*** для перевірки незалежності передбачає сумування для кожної з двох категорій, то формула 7.1 набуває вигляду:

 2  (*m*  *m*)2 *m*



(8.10)

Якщо отримане значення критерію «хі-квадрат» перевищує критичне його значення з таблиці розподілу «хі-квадрат», то це означає, що спостережувані частоти значно відрізняються від тих, що мали б місце, якби досліджувані змінні були незалежними. У такому випадку слід відхилити нульову гіпотезу про незалежність змінних і прийняти досліджувану (альтернативну) гіпотезу, зробивши висновок, що між змінними є істотний зв'язок.

Якщо значення критерію «хі-квадрат» менше табличного (критичного) значення, то це є доказом того, що спостережувані дані не значно відрізняються від тих, яких можна було б очікувати, якби змінні в генеральної сукупності були незалежними. У цьому випадку слід прийняти нульову гіпотезу про незалежність змінних (як прийнятну можливість) і зробити висновок, що між змінними відсутній суттєвий зв'язок [16, с. 885].

Розглянемо розрахунок критерію «хі-квадрат» та перевірку гіпотези щодо незалежності змінних на наступному прикладі.

Фірма розглядає питання щодо розширення своєї діяльності в регіоні. Для цього здійснено опитування працівників фірми. Під час опитування працівникам було запропоноване таке запитання: "Як ви вважаєте, умови для бізнесу у вашому регіоні стали кращими, погіршилися або залишилися незмінними?" Усього було опитано 470 працівників. Результати опитування наведено в таблиці 8.5.

Необхідно з’ясувати:

1. чи залежать відповіді від категорії працівників, тобто, чи є залежність між категорією опитуваних і варіантом відповіді, яку вони надавали?
2. чи є підґрунтя для розширення бізнесу в регіоні?

*Таблиця 8.5*

## Результати опитування щодо перспектив розширення бізнесу

|  |  |  |
| --- | --- | --- |
| Варіанти відповіді | Кількість опитаних, осіб | |
| менеджери | інші службовці |
| Умови покращилися | 23 | 92 |
| Умови залишилися незмінними | 37 | 168 |
| Умови погіршилися | 11 | 81 |
| Не знаю | 15 | 43 |
| Усього | 86 | 384 |

Для відповіді на поставлені запитання необхідно визначити критерій «хі-квадрат» і здійснити перевірку гіпотези щодо незалежності змінних.

На першому етапі визначаємо очікувані частоти за формулою

8.8. Результати розрахунків наведено в табл. 7.6.

*Таблиця 8.6*

## Розрахунок очікуваних частот

|  |  |  |  |
| --- | --- | --- | --- |
| Варианти відповіді | Очікувані частоти | | Разом |
| менеджери | інші службовці |
| Умови покращилися | 86\*115/470=21,0 | 384\*115/470=94,0 | 115 |
| Умови залишилися незмінними | 86\*205/470=37,5 | 384\*205/470=167,5 | 205 |
| Умови погіршилися | 86\*92/470=16,8 | 384\*92/470=75,2 | 92 |
| Не знаю | 86\*58/470=10,6 | 384\*58/470=47,4 | 58 |
| Усього | 85,9 | 384,1 | 470 |

Необхідно звернути увагу на те, що сума очікуваних частот повинна дорівнювати сумі фактичних частот (у даному випадку – 470).

Для порівняння критерію «хі-квадрат» з його критичним (табличним) значенням необхідно визначити кількість ступенів свободи.

Для цього скористаємося формулою 8.9:

*k*  (4 1)  (2 1)  3

Отже, кількість ступенів свободи складає 3, оскільки у нас чотири категорії відповідей і дві категорії працівників.

Визначимо критерій «хі-квадрат» за формулою 7.10:

 2  (23  21)2

21

 (37  37,5)2

37,5

 (11  16,8)2

16,8

 (15 10,6)2

10,6

 (92  94)2 

94

 (168  167,5)2

167,5

 (81  75,2)2

75,2

 (43  47,4)2

47,4

 0,19 

0,01 

2,00

 1,83 

 0,04  0,001  0,45  4,52

За таблицями розподілу «хі-квадрат» для 3-х ступенів свободи і рівня значущості 5 % знаходимо критичне значення, яке дорівнює 7,815.

Отримане значення «xі-квaдpaт» (4,52) виявилося меншим, ніж його критичне значення (7,815), тому можна зробити висновок, що між двома якісними змінними (варіант відповіді та категорія працівника) відсутній значущий зв'язок.

Оскільки обидві категорії працівників відмітили або покращення умов для бізнесу, або їх незмінний рівень, то можна розглядати питання щодо розширення бізнесу в регіоні.

Як ми відмічали раніше, однією з головних умов коректного застосування критерію «хі-квадрат» є те, що в кожній клітинці таблиці спряженості очікувана частота повинна бути не меншою п'яти, інакше у результаті розрахунку критерію можна отримати хибні висновки.

Крім того, значення критерію «хі-квадрат», що отримують у результаті розрахунків, ґрунтуються на частотах, які є цілими числами. У той же час табличні значення для його розподілу складені для безперервної шкали. Тому англійський статистик Френк Йейтс (1902-1994) запропонував поправку до даного критерію, яку зараз називають «поправкою Йейтса».

Сутність поправки полягає у відніманні числа 0,5 з різниці між фактичною і очікуваною кількістю спостережень. Тобто, формула ***критерію «хі-квадрат» з «поправкою Йейтса»*** набуває такого вигляду:

 2 

( *m*  *m*  0,5)2



*m*

(8.11)

Дану формулу доцільно застосовувати в тому випадку, коли кількість частот в окремих клітинках таблиці спряженості менша п’яти.

Слід відмітити, що «поправка Йейтса» викликає багато суперечок між фахівцями, оскільки розрахований таким чином критерій «хі-квадрат» може давати дещо занижені результати. А це означає, що збільшується помилка ймовірності не знайти розходжень там, де вони є.

Іншим виходом із ситуації є визначення коефіцієнтів спряженості, які засновані на критерії «хі-квадрат». Серед них найбільш поширеними є коефіцієнти взаємної спряженості К. Пірсона (*С*) та А. Чупрова (*К*). Вважається, що ці критерії дають більш точні результати для невеликої кількості частот.

Формула ***критерію Пірсона*** має вигляд:

 2

 2  *n*

де *n* – обсяг вибірки.

*C* 

, (8.12)

Даний критерій не враховує кількість категорій для кожної з ознак. Більш точним є ***коефіцієнт спряженості Чупрова***, який розраховується за формулою:

 2

(*r* 1)(*c* 1)

де *n* – обсяг вибірки

*K* 

, (8.13)

*r* – кількість рядків у таблиці спряженості,

*c* – кількість стовпчиків у таблиці спряженості.

Модифікацією цього коефіцієнта є ***коефіцієнт взаємної спряженості Г. Крамера***, в знаменнику формули розрахунку якого береться мінімальне з двох значень: кількість рядків за мінусом одиниці, або кількість стовпчиків за мінусом одиниці:

 2  *n*

min{*r* 1, *c* 1}

*K* 

(8.14)

Усі коефіцієнти спряженості можуть набувати значень від 0 до

1. Якщо значення коефіцієнту наближається до 0, то це свідчить про відсутність зв’язку між досліджуваними ознаками, а якщо до 1, то між ними існує тісний взаємозв’язок.

## Питання для самоконтролю

1. У чому полягає відмінність між якісними та кількісними ознаками?
2. Що таке номінальна ознака?
3. У чому сутність критерію «хі-квадрат» і як він розраховується?
4. Що розуміють під опорними частотами?
5. Як визначаються опорні частоти під час перевірки гіпотези на подібність розподілів?
6. Що таке нульова і альтернативна гіпотези?
7. Які умови повинні виконуватися для того, щоб застосування критерію «хі-квадрат» було коректним?
8. Що таке кількість ступенів свободи?
9. Як розраховуються опорні частоти під час перевірки гіпотези на незалежність змінних?
10. У чому відмінність між таблицею спряженості та звичайної таблиці?
11. Як розраховується коефіцієнт кореляції рангів Спірмена?
12. Як розраховується коефіцієнт кореляції рангів Кендела?
13. При яких значеннях коефіцієнтів Спірмена та Кендала зв'язок між ознаками можна визнати істотним?
14. Як розраховується коефіцієнт конкордації?
15. Яких значень може набувати коефіцієнт конкордації?
16. Як визначаються опорні частоти під час перевірки гіпотези на незалежність змінних?
17. Що таке «поправка Йейтса»? Коли вона необхідна?
18. Як розраховується критерій Пірсона?
19. Які недоліки має критерій Пірсона?
20. Як розраховується критерій Чупрова?
21. Як розраховується критерій Крамера?

Яких значень можуть набувати критерії спряженості?