# Лекція Тема 7. ДИСПЕРСІЙНИЙ АНАЛІЗ У ДОСЛІДЖЕННІ БІЗНЕС-ПРОЦЕСІВ

## 7.1. Види дисперсій. Правило декомпозиції дисперсій

Основою будь-якого статистичного дослідження просторових даних про соціально-економічні явища є формування якісно однорідної сукупності. Тому серед статистичних методів особливе місце займає дисперсійний аналіз, одним з основних завдань якого є вимірювання однорідності сукупності.

При цьому якщо досліджувана сукупність розбита на групи за певною ознакою, то можна визначити дисперсію як у цілому по сукупності, так і в кожній групі, крім того можна визначити міжгрупову дисперсію.

Зведеною характеристикою розсіювання (варіації) значень кількісної ознаки в генеральній сукупності навколо свого середнього значення є ***загальна (генеральна) дисперсія***, яка розраховується за формулою:

 2 

(*X*  *X* )2

*n*

, (7.1)

де *X* – значення ознаки в кожної одиниці загальної сукупності;

*X* – загальна середня;

*n* – кількість одиниць сукупності.

Розглядаючи кожну групу як самостійну сукупність можна визначити ***дисперсію в групі***, яка характеризує розсіювання значень кількісної ознаки навколо групової середньої величини. Розрахунок дисперсії в групі здійснюється за формулою, аналогічною загальній дисперсії:

 2  (*Xіj*  *X j* )

2

*j* , (7.2)

*n*

*j*

де *X ij* – значення ознаки в кожної одиниці сукупності в *j* -й групі;

*X j* – середня величина ознаки в *j* -й групі;

*n j* – кількість одиниць в *j* -й групі.

Внутрішньогрупова дисперсія характеризує випадкову варіацію ознаки, яка відбувається під впливом інших факторів (крім групувального), і не залежить від ознаки, що покладена в основу групування.

Узагальнюючою мірою внутрішньогрупової варіації є ***середня з групових дисперсій***, яка розраховується за такою формулою:

 2 *f*

 2 

*j*

 *f j*

*j*

, (7.3)

де *f j* – частота або частка *j* -ї групи.

Мірою варіації групових середніх навколо загальної середньої є ***міжгрупова дисперсія***. Вона характеризує відмінності між варіантами досліджуваної ознаки, які виникають під дією лише однієї ознаки – фактора, який покладено в основу групування. Тобто, міжгрупова дисперсія показує, на скільки відрізняються між собою групові середні. Вона розраховується за формулою:

 2 

( *X*  *X* )2 *f*

 *f j*

*j j*

, (7.4)

Міжгрупова дисперсія може дорівнювати нулю, коли середні величини в кожній групі однакові (рівні) і вона є тим більшою, чим більше відрізняються середні між собою.

Загальна дисперсія складається з двох частин. Перша характеризує внутрішньогрупову, друга – міжгрупову варіацію. Тому загальна дисперсія дорівнює сумі середньої з внутрішньогрупових і міжгрупової дисперсій:

 2  ~~~~ 2   2

(7.5)

Взаємозв’язок дисперсій називають правилом декомпозиції (розкладання) дисперсій. На практиці цей взаємозв’язок досить часто використовується для визначення однієї з дисперсій на основі значень двох інших дисперсій.

Відношення міжгрупової дисперсії до загальної характеризує частку варіації результативної ознаки, яка обумовлена варіацією (впливом) групувальної ознаки. Це відношення називається ***емпіричним коефіцієнтом детермінації*:**

 2   2

 2

(7.6)

Для оцінки тісноти зв’язку між групувальною і результативною ознаками розраховується ***емпіричне кореляційне відношення***:

  (7.7)

 2

Емпіричне кореляційне відношення може набувати значень від 0 до 1. Чим більше його значення наближається до 1, тим більш тісним є зв'язок між групувальною і результативною ознаками.

Розглянемо розрахунок емпіричного кореляційного відношення на наступному прикладі.

24 студенти групи розподілені за кількістю пропущених занять із статистики на 2 групи: 1-а – студенти, які пропустили менше, ніж 10 % загальної кількості годин, 2-а – 10 % і більше. До 1-ї групи потрапили 18 студентів, середній бал рейтингу яких склав 78, причому бали (рейтинг) окремих студентів групи відхиляються від середнього рейтингу в середньому на 11 балів. У 2-ій групі середній бал склав 46, а коливання навколо середнього рейтингу – 16.

Потрібно визначити, чи впливає наявність або відсутність пропусків занять на варіацію балів рейтингу студентів групи, тобто, чи існує залежність між пропусками занять студентів та рівнем їх успішності зі статистики.

Для зручності складемо таблицю, в яку занесемо вихідні дані (табл. 7.1).

*Таблиця 6.1*

## Вихідні дані щодо розподілу студентів

|  |  |  |  |
| --- | --- | --- | --- |
| Групи студентів за кількістю пропусків занять, % | Кількість студентів, чол. ( *fі* ) | Середній бал ( *X і* ) | Варіація балів (*і* ) |
| До 10 | 18 | 78 | 11 |
| 10 і більше | 6 | 46 | 16 |
| Усього | 24 | - | - |

Для того, щоб відповісти на поставлене питання необхідно визначити емпіричне кореляційне відношення. Для цього нам необхідно мати міжгрупову та загальну дисперсії.

Визначимо середній бал рейтингу студентів групи в цілому на основі групових середніх за формулою середньої арифметичної зваженої:

*X*   *Хi fi*

 *fі*

Проміжні розрахунки наведено в табл. 6.2.

*X*  1680  70 *балів*

24

Визначаємо міжгрупову дисперсію за формулою 6.4 (проміжні розрахунки наведено в табл. 6.2).

 2  4608  192

24

*Таблиця 7.2*

## До розрахунку емпіричного кореляційного відношення

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Групи студентів за  кількістю про- пусків занять, % | ( *fі* ) | ( *X і* ) | *X і fi* | ( *X*  *X* )2 *f*  *і* *i* |  2  *і* |  2 *f*  *і i* |
| До 10 | 18 | 78 | 11 | 1152 | 121 | 2178 |
| 10 і більше | 6 | 46 | 16 | 3456 | 256 | 1536 |
| Усього | 24 | - | - | 4608 | - | 3714 |

Оскільки немає даних про рейтинг кожного студента, то не можна визначити загальну дисперсію за вихідною формулою 6.1, тому її можна визначити згідно правила декомпозиції дисперсій (формула 6.5). Для цього спочатку визначимо середню з внутрішньогрупових дисперсій за формулою 7.3 (проміжні розрахунки наведено в табл. 6.2):

~~~~ 2  3714  154,75

24

Загальна дисперсія складає:

 2  154,75 192  346,75

Визначаємо емпіричний коефіцієнт детермінації за формулою

7.6:

 2 

192

346,75

 0,554

*або*

55,4 %

Отже, 55,4 % загальної варіації балів рейтингу студентів пояснюється варіацією пропусків занять.

Емпіричне кореляційне відношення складає:

   0,744

0,554

Оскільки значення емпіричного кореляційного відношення наближається до 1, то між кількістю пропусків занять із статистики та рівнем успішності студенів (балами їх рейтингу) з дисципліни існує достатньо тісний зв'язок. Причому, за результатами аналітичного групування, яке міститься в табл. 6.1, можна побачити, що зв'язок цей обернений.

Отримані результати даного прикладу свідчать про те, що зв'язок між досліджуваними ознаками є, але якщо значення емпіричного кореляційного відношення складає близько 50 %, то дати відповідь відносно існування зв’язку складно. Такі питання (і не лише) допомагає вирішувати дисперсійний аналіз.

## 7.2. Сутність дисперсійного аналізу

В основі дисперсійного аналізу лежить правило декомпозиції дисперсій. Однак відмінність полягає в тому, що в дисперсійному аналізі здійснюється не простото розчленування загальної варіації досліджуваної ознаки на складові, які обумовлені впливом певних факторів, а проводиться перевірка гіпотези щодо суттєвості їх впливу (або значущості відмінностей між груповими середніми).

Дисперсійний аналіз широко застосовується в багатьох сферах діяльності: при дослідженні рівня життя населення, в медицині, хімії, фізиці тощо.

Зокрема, дисперсійний аналіз використовується для перевірки наявності відмінностей доходів чи витрат різних груп населення: наприклад він дозволяє встановити, чи є суттєвими регіональні відмінності в середніх доходах, чи ні. У цьому випадку загальна дисперсія варіації доходів населення аналізується як у зв'язку з відмінностями в самих регіонах, так і у зв'язку з відмінностями середніх значень доходів між регіонами. Очевидно, що чим більша частка загальної дисперсії припадає на відмінності між середніми значеннями доходів в окремих регіонах, тим вище ймовірність того, що ці групи (регіони) дійсно різні; разом з тим, чим більшою є частка дисперсії, яка припадає на внутрішньогрупові відмінності, тим вище ймовірність того, що відмінності між середніми значеннями доходів у різних груп мають суто випадковий характер.

Слід відмітити, що в економічних дослідженнях застосування дисперсійного аналізу є не таким поширеним, при тому, що важливість його використання в певних випадках є навіть більш значущою.

Розглянемо це на прикладі. З метою підвищення якості йогурту, що виготовляє підприємство, планується використовувати одну з чотирьох харчових добавок. Підприємство двадцять разів запускало виробництво йогурту з кожною добавкою і двадцять разів без добавки. Результат представляє собою набір із п’яти рядів даних

виробництва продукції (з кожною добавкою і без добавок). При цьому ми маємо один фактор – харчову добавку, який має декілька значень. Унаслідок того, що будь-який виробничий процес пов'язаний із багатьма факторами виробництва, то складно пояснити чим обумовлена зміна якості продукції: тією чи іншою добавкою, або суто випадковими обставинами, пов’язаними лише з мінливістю виробничого процесу. Саме для цього проводиться однофакторний дисперсійний аналіз, який дає можливість встановити, чи є суттєвими відмінності між харчовими добавками та їх впливом на рівень якості йогурту. Якщо ці відмінності суттєві, то можна далі більш детально вивчати доцільність здійснення витрат для впровадження у виробництво певної харчової добавки. У протилежному випадку можна зробити висновок, що суттєвих відмінностей між добавками немає, тому їх впровадження у виробничий процес скоріш за все є недоцільним.

Отже, в основі дисперсійного аналізу знаходиться припущення про те, що результати експерименту можна представити у вигляді певної суми складових. Наприклад, якщо досліджується вплив одного фактора, то модель, яка відображає результат експерименту, має вигляд [20, с. 207]:

*xij*

 *X* *і*  *ij* ,

(7.8)

де *xij* – значення ознаки, що отримані на *i* -му рівні фактора;

*X* – загальна середня;

*і* – ефект фактора на *і* -му рівні;

*ij* – випадковий компонент, який викликаний впливом інших

факторів.

Однофакторний дисперсійний аналіз є найбільш поширеним в економічних дослідженнях.

У тому випадку, коли визначається вплив двох факторів (А і В) та їх взаємодія, то модель, яка відображає результат експерименту, має вигляд:

*xijk*

 *X*  *i*   *j*   *ij*  *ijk* ,

(7.9)

де *xij* – результат експерименту в *k* -му спостереженні на *i* -му рівні

фактора А і *j* -му рівні фактора В;

*X* – загальна середня;

*і* – ефект фактора А;

 *j* – ефект фактора В;

 *ij* – ефект, який викликаний спільним впливом обох факторів;

*ij* – випадковий компонент, який викликаний впливом інших

(крім А і В) факторів.

Для проведення дисперсійного аналізу вся досліджувана сукупність розбивається на однорідні групи, які відрізняються рівнями факторів. Показником, що характеризує рівень однорідності, може бути коефіцієнт варіації.

Якщо групувальною ознакою є якісна, то кількість груп та їх найменування визначаються самою ознакою. Якщо ж групувальною ознакою є кількісна, то виділення однорідних груп здійснюється поступово. На першому етапі визначається середня величина показника, за яким сукупність поділяється на найменшу необхідну кількість груп. На другому етапі в кожній групі визначаються основні статистичні характеристики (середня величина, дисперсія, коефіцієнт варіації). Якщо отримані групи є дуже неоднорідними, то сукупність поділяють на більшу кількість груп і т. д.

## 7.3. Однофакторний дисперсійний аналіз

Як ми відмічали раніше, за допомогою однофакторного дисперсійного аналізу досліджується вплив одного фактора, тобто перевіряється значущість відмінностей між собою значень середніх величин декількох вибіркових сукупностей.

В однофакторному дисперсійному аналізі набір даних складається з *k* незалежних одновимірних вибірок із *k* генеральних сукупностей (або з однієї генеральної сукупності), елементи яких вимірюються в однакових одиницях.

Дані для однофакторного дисперсійного аналізу зручно подавати в такій формі (табл. 6.3).

У рядку «Усього» наводиться загальна кількість одиниць сукупності (загальний обсяг вибірки) *n* , загальна середня (по

сукупності в цілому) *X* і сума квадратів відхилень від середньої

величини в цілому по сукупності *S*2 .

Таблицю можна також транспонувати, зробивши рядки графами.

Для перевірки гіпотези щодо суттєвості відмінностей між середніми величинами у вибірках, що сформовані за факторною ознакою, використовується *F-критерій Фішера*.

В однофакторному дисперсійному аналізі статистичні гіпотези формулюються так:

* нульова гіпотеза ( *H* 0 ): усі середні рівні між собою;
* альтернативна (дослідницька) гіпотеза ( *H*1 ): не всі середні рівні між собою, принаймні існує хоча б одна пара вибіркових сукупностей, середні величини в яких відрізняються між собою.

*Таблиця 7.3*

## Форма подання даних для однофакторного дисперсійного аналізу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер вибірки | Значення досліджуваної ознаки | Обсяг вибірки | Середня | Сума квадратів відхилень від середніх |
| 1 | *X*11 ; *X* 21 ;…; *X i*1 ;…; *X n*1 | *n*1 | *X* 1 | *S*2  1 |
| 2 | *X* 12 ; *X* 22 ;…; *X i* 2 ;…; *X n* 2 | *n*2 | *X* 2 | *S*2  2 |
| … | … | … | … | … |
| *j* | *X* 1 *j* ; *X* 2 *j* ;…; *X ij* ;…; *X nj* | *n j* | *X j* | *S*2  *j* |
| … | … | … | … | … |
| *k* | *X*1*k* ; *X* 2 *k* ;…; *X ik* ;…; *X nk* | *nk* | *X k* | *S*2  *k* |
| Усього | | *n* | *X* | *S*2 |

Для коректного використання однофакторного дисперсійного аналізу повинні виконуватися такі *вимоги*:

1. Кількісні дані повинні бути неперервними, дискретні дані є менш бажаними.
2. Вибіркові сукупності повинні бути незалежними.
3. Незалежними повинні бути спостереження в кожній із вибірок.
4. Вибіркові сукупності повинні мати нормальний розподіл.
5. Дисперсії генеральних сукупностей, із яких здійснювалися

вибірки, повинні бути однаковими ( 2 = 2 =…= 2 ). Це означає, що всі

1 2 *k*

вибірки здійснюються або з однієї генеральної сукупності, або середні в генеральних сукупностях, з яких здійснюються вибірки, рівні між собою, що дозволяє використовувати для перевірки гіпотез стандартні статистичні таблиці.

***F-критерій Фішера*** для однофакторного дисперсійного аналізу визначається як співвідношення незміщених оцінок міжгрупової та середньої з внутрішньогрупових дисперсій:

 2

*F*  ~~~~ 2 (7.10)

Тобто, *F*-критерій показує, у скільки разів середні величини кожної з вибірок відрізняються між собою, тобто наскільки більше вони відрізняються між собою, ніж можна було цього очікувати, якби їх розбіжності були суто випадковими. Очевидно, що його значення можуть бути або додатними, або дорівнювати нулю. Дорівнювати нулю *F*-критерій може в тому випадку, коли міжгрупова варіація ознаки відсутня.

Для того, щоб перевірити істотність розходження дисперсій необхідно отримане значення *F*-критерію порівняти з табличним. Таблиці розподілу *F*-критерію для 5 % рівня значущості у випадку однофакторного дисперсійного аналізу містяться в *Додатку B*, також їх легко можна знайти в Інтернеті.

Слід звернути увагу на те, що у формулі *F*-критерію Фішера зіставляються не просто дисперсії, а їх незміщені оцінки. Відмінність полягає в тому, що при визначенні незміщених оцінок дисперсій суми квадратів відхилень діляться не на кількість одиниць сукупності, а на кількість ступенів свободи.

Поняття ***кількості ступенів свободи*** вів у вжиток англійський статистик Рональд Фішер (1890-1962), і воно означає кількість незалежних змінних у сумі.

Кількість ступенів свободи визначається по-різному для кожної з дисперсій. І тому критерій Фішера має два значення ступенів свободи, які відповідають двом порівнюваним дисперсіям: міжгрупової і середньої з внутрішньогрупових.

Для міжгрупової дисперсії кількість ступенів свободи складає *k –1* (*k* – кількість груп, на які розбита вся сукупність, тобто кількість одновимірних незалежних вибірок). Отже, ***незміщена оцінка***

***міжгрупової дисперсії*** (6.4) має вигляд:

 2 

*j*

( *X j*

 *X* )2 *f*

(7.11)

*k* 1

***Незміщена оцінка внутрішньогрупової дисперсії*** (7.2) розраховується за формулою:

 2  (*Xіj*  *X j* )

2

*j* 1

*n*

*j*

(7.12)

Для середньої із внутрішньогрупових дисперсій кількість ступенів свободи визначається як *n – k* . А отже і ***незміщена оцінка***

***середньої із внутрішньогрупових дисперсій*** (7.3) розраховується за формулою:

*j*

*j*

 2 

 2 ( *f*

1)

(6.13)

*n*  *k*

Незміщена оцінка середньої з внутрішньогрупових дисперсій може бути визначена і в інший спосіб:

 2 

( *X j*

 *X* )2

(7.14)

*n*  *k*

У чисельнику даної формули фігурує загальна сума квадратів відхилень варіантів (значень ознаки) від середніх величин в кожній групі.

Розрахунок оцінок дисперсій зручно розміщувати в спеціальній таблиці, яку називають таблицею дисперсійного аналізу [20, с. 210]. Для однофакторного дисперсійного аналізу вона має такий вигляд (табл. 7.4).

*Таблиця 7.4*

## Форма таблиці однофакторного дисперсійного аналізу

|  |  |  |  |
| --- | --- | --- | --- |
| Характеристика варіації | Сума квадратів відхилень | Кількість ступенів свободи | Оцінка дисперсії |
| Міжгрупова (систематична) | *S*2  (*X*  *X* )2 *n*  1 *j j* | *k* 1 | 2   2  *S*1  *k* 1 |
| Внутрішньогрупова (залишкова) | *S*2  (*X*  *X* )2  2 *ij j* | *n*  *k* | 2  ~~~~ 2  *S*2  *n*  *k* |
| Усього | *S*2  *S*  *S*  (*X*  *X* )2  1 2 *ij* | *n* 1 | - |

Дані табл. 6.4 дозволяють визначити співвідношення дисперсій, тобто *F*-критерій Фішера за формулою 7.10.

Після розрахунку *F*-критерію Фішера його значення порівнюється з критичним (табличним) значенням. Якщо фактичне значення *F*-критерію менше, ніж табличне, то слід прийняти нульову гіпотезу як прийнятну, і це означає, що розходження між

вибірковими середніми можна пояснити лише випадковістю, тобто результат не є статистично значущим.

Якщо фактичне значення *F*-критерію перевищує його табличне значення, то слід відхилити нульову гіпотезу і прийняти альтернативну, а це означає, що групові вибіркові середні суттєво відрізняються між собою, причому ці відмінності не можна пояснити лише випадковістю, тобто результат є статистично значущим.

Зазвичай для перевірки гіпотези використовують *F*-критерій із 5% рівнем значущості. При цьому *F*-розподіл Фішера, як ми вже відмічали вище, залежить від кількості ступенів свободи *v1 = k –1* і *v2* = *n – k*.

Розглянемо реалізацію однофакторного дисперсійного аналізу на прикладі.

Є такі дані про рівень годинної заробітної плати трьох категорій робітників промислового підприємства: жінки та чоловіки із стажем роботи і молодь без стажу роботи (табл. 7.5). Необхідно визначити, чи впливає категорія, до якої відносяться робітники, на рівень їх заробітної плати.

*Таблиця 7.5*

## Годинна заробітна плата робітників підприємства

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Жінки | | | Чоловіки | | | Молодь | | |
|  | годин- |  |  | годин- |  |  | годин- |  |
| номер  робіт- | на за-  робітна | ( *Х*  *X* )2 | номер  робітн | на за-  робітна | ( *Х*  *X* )2 | номер  робітн | на за-  робітна | ( *Х*  *X* )2 |
| ника | плата, |  | ика | плата, |  | ика | плата, |  |
|  | грн. |  |  | грн. |  |  | грн. |  |
| 1 | 48,30 | 15,68 | 11 | 47,28 | 1,49 | 22 | 46,97 | 7,02 |
| 2 | 42,56 | 4,75 | 12 | 49,36 | 10,89 | 23 | 42,16 | 5,66 |
| 3 | 44,61 | 0,02 | 13 | 45,12 | 0,88 | 24 | 45,12 | 0,34 |
| 4 | 43,12 | 2,62 | 14 | 43,96 | 4,41 | 25 | 40,78 | 14,14 |
| 5 | 41,32 | 11,70 | 15 | 47,54 | 2,19 | 26 | 42,14 | 5,76 |
| 6 | 42,20 | 6,45 | 16 | 44,80 | 1,59 | 27 | 43,19 | 1,82 |
| 7 | 43,22 | 2,31 | 17 | 42,31 | 14,06 | 28 | 45,87 | 1,77 |
| 8 | 44,64 | 0,01 | 18 | 44,20 | 3,46 | 29 | 44,56 | 0,00 |
| 9 | 46,32 | 2,50 | 19 | 45,12 | 0,88 | 30 | 48,11 | 12,74 |
| 10 | 47,12 | 5,66 | 20 | 48,44 | 5,66 |  |  |  |
|  |  |  | 21 | 48,51 | 6,00 |  |  |  |
| Усього | 443,41 | 51,70 | Усього | 506,64 | 51,52 | Усього | 398,90 | 49,26 |

Визначимо середньогодинну заробітну плату робітників кожної категорії та по підприємству в цілому за формулою середньої арифметичної простої на основі даних, наведених у табл. 7.5:

* перша група (жінки):

*X*  443,41  44,34 *грн*.

1 10

;

* друга група (чоловіки):
* третя група (молодь):
* підприємство в цілому:

*X*  506,64  46,06 *грн*.

2 11

;

*X*  398,90  44,32 *грн*.

;

3 9

*X*  443,41  506,64  398,90  44,97 *грн*.

30

За даними табл. 6.5 визначимо оцінки внутрішньогрупових дисперсій згідно формули 6.12:

* перша група (жінки):

 2  51,70  5,7444 ;

1 10 1

* друга група (чоловіки):

 2  51,52  5,1520 ;

2 111

* третя група (молодь):

 2  49,26  6,1575 .

3 9 1

Оцінку середньої із внутрішньогрупових дисперсій визначимо за формулою 6.13:

~~~~ 2  5,7444  9  5,1520 10  6,1575 8  5,6474

.

30  3

Визначимо також оцінку середньої з внутрішньогрупових дисперсій за іншою формулою – 6.14:

~~~~ 2  152,48  5,6474 .

327

На основі попередніх розрахунків середніх у кожній групі та в цілому по сукупності визначимо оцінку міжгрупової дисперсії за формулою 6.11:

 2  (44,34  44,97)2 10  (46,06  44,97)2 11 (44,32  44,97)2 9 

3 1

10,4203

Для зручності наведемо всі результати розрахунків у таблиці дисперсійного аналізу (табл. 6.6).

На основі даних табл. 7.6 визначаємо *F*-критерій Фішера згідно формули 7.10:

*F*  10,4203  1,845

5,6474

*Таблиця 7.6*

## Таблиця однофакторного дисперсійного аналізу

|  |  |  |  |
| --- | --- | --- | --- |
| Характеристика варіації | Сума квадратів відхилень | Кількість  ступенів свободи | Оцінка дисперсії |
| Міжгрупова (систематична) | *S* 2  20,8406  1 | 3-1=2 |  2 10,4203 |
| Внутрішньогру- пова (залишкова) | *S* 2  52,48  2 | 30-3=27 | ~~~~ 2  5,6474 |
| Усього | *S* 2  20,8406 52,48  73,3206 | 30-1=29 | - |

Нагадаємо, що *F*-критерій Фішера показує на скільки вибіркові середні відрізняються між собою (чисельник формули) з урахуванням рівня варіації у вибірках (знаменник).

У даному випадку міжгрупова варіація заробітної плати (яка обумовлена тим, до якої категорії належать робітники) у 1,845 разів більша, ніж можна було очікувати виходячи лише з варіації заробітної плати окремих робітників. Але чи можна стверджувати, що ці відмінності є значущими? Для відповіді на це питання необхідно порівняти отримане значення *F*-критерію з критичним (табличним) його значенням.

Критичне значення *F*-критерію при рівні значущості 5% і кількості ступенів свободи (3-1=2) і (30-3=27) складає 3,35 (згідно таблиць розподілу, наведених у *Додатку B*).

Оскільки фактичне значення *F*-критерію Фішера (1,845) менше, ніж критичне (3,35), то результат перевірки є незначущим. Тобто, слід прийняти нульову гіпотезу щодо рівності середніх. А це означає, що рівень середньої заробітної плати робітників різних категорій не суттєво відрізняється між собою, тобто практично не існує залежності між тим, до якої з категорій відноситься робітник, і тим, яку заробітну плату він отримує.

## 7.4. Багатофакторний дисперсійний аналіз

Набір аналізованих даних може бути не лише одновимірним, про що йшла мова у попередніх параграфах, а й представляти собою

більш складну структуру. У цьому випадку дисперсійний аналіз може дати відповіді на більш складні питання.

При цьому дані можуть бути представлені як набір вибірок із однієї або декількох генеральних сукупностей, які виражені в однакових одиницях (так само як у випадку однофакторного дисперсійного аналізу). Можливий також інший спосіб представлення даних: у вигляді багатовимірного набору даних, у якому обов’язково одна змінна є кількісною (основна вимірювана величина), а всі інші змінні є якісними, і ці якісні змінні в сукупності визначають, яким чином кількісні спостереження групуються у вибірки.

Багатофакторний дисперсійний аналіз дає можливість оцінити ефект кожного фактора, тобто вплив кожного фактора на результативний показник, а також ефект взаємодії факторів (показати як фактори пов’язані між собою). При чому чим більше факторів, тим більше видів взаємодії між ними. Допустимо, при трьохфакторному дисперсійному аналізі крім чистого впливу кожного фактора на результативний показник визначається взаємодія між кожними двома факторами (першим і другим, другим і третім, першим і третім), а також між трьома факторами разом.

Таким чином, багатофакторний дисперсійний аналіз застосовується тоді, коли в одному дисперсійному аналізі необхідно одночасно досліджувати не лише вплив факторів, а також можливу коваріацію змінних. При цьому змінні не є незалежними одна від іншої, а навпаки, корелюють між собою.

Для коректного застосовування багатофакторного дисперсійного аналізу повинні виконуватися такі ж вимоги, як і для однофакторного.

Розглянемо більш детально двохфакторний дисперсійний аналіз. У цьому випадку вибірки утворюють таблицю, в якій одному фактору відповідають рядки, а іншому – стовпчики, тому виникає три основних питання: "Чи впливає перший фактор на появу будь-яких відмінностей результативної ознаки?" "Чи впливає другий фактор на появу будь-яких відмінностей?" "Чи залежить вплив першого фактора від другого, або обидва фактори діють незалежно один від одного?" Перші два питання мають відношення до головних ефектів кожного з факторів, а третє питання має відношення до взаємодії факторів [16,

с. 828].

Фактори А і В вважаються **взаємодіючими**, якщо ефект фактора А залежить від рівня фактора В.

Інформацію для двохфакторного дисперсійного аналізу зручно наводити у вигляді такої таблиці (табл. 7.7) [20, с. 216].

В табл. 6.7: *p* – кількість рівнів фактора А*, q* – кількість рівнів фактора В, *k* – кількість вибірок.

За даними табл. 6.7 визначаються такі середні:

  *X ijk*

* загальна середня:

;

*X* 

*pqn*

* середні за рядками:

середні за стовпчиками:

*X j* 

*Xi* 

 *X ijk*

*pn* ;

 *X ijk*

*qn* ;

* середні для кожного варіанта взаємодії факторів (для кожної

 *X ijk*

окремої клітинки таблиці):

*X ij*  *n* .

*Таблиця 7.7*

## Форма подання даних для двохфакторного дисперсійного аналізу

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Рівень фактора *В* | Рівень фактора *А* | | | | Сума |
| *А1* | *А2* | *…* | *Аp* |
| *В1* | *X111; X112;…;*  *X11n* | *X211; X212;…;*  *X21n* | *…* | *Xp11; Xp12;…;*  *Xp1n* |  *Xi*1*k* |
| *В2* | *X121; X122;…;*  *X12n* | *X221; X222;…;*  *X22n* | *…* | *Xp21; Xp22;…;*  *Xp2n* |  *Xi* 2*k* |
| *…* | *…* | *…* | *…* | *…* | *…* |
| *Вq* | *X1q1; X1q2;…;*  *X1qn* | *X2q1; X2q2;…;*  *X2qn* | *…* | *Xpq1; Xpq2;…;*  *Xpqn* |  *Xiqk* |
| Сума |  *X*1 *jk* |  *X*2 *jk* | *…* |  *X pjk* |  *Xijk* |

Суми квадратів відхилень від загальної середньої розкладається на складові. Схема двохфакторного дисперсійного аналізу наведена в табл. 7.8 [20, с. 217].

Оскільки двохфакторний дисперсійний аналіз дозволяє визначити вплив кожного фактора окремо, а також їх взаємну дію, то відповідно визначають три значення *F*-критерію Фішера:

 2

*F*  4

* для фактора А



*p* 1; *N*  *pq* );

*А* 2 (кількість ступенів свободи складає

1

 2

*F*  3

* для фактора В



*q* 1; *N*  *pq* );

*В* 2 (кількість ступенів свободи складає

1

 2

* для взаємодії факторів А і В *FАВ*  2 (кількість ступенів



свободи складає

2

1

( *p* 1)(*q* 1); *N*  *pq* ).

*Таблиця 7.8*

## Форма таблиці двохфакторного дисперсійного аналізу

|  |  |  |  |
| --- | --- | --- | --- |
| Джерело варіації | Сума квадратів відхилень | Кількість ступенів свободи | Оцінка дисперсії |
| Фактор А | *S*2  *nq*(*X*  *X* )2  4 *i* | *p* 1 | 2   2  *S*4  4 *p* 1 |
| Фактор В | *S*3  *np*(*X j*  *X* )  2 2 | *q* 1 | 2   2  *S*3  3 *q* 1 |
| Взаємодія | *S*2  *n*(*X*  *X*  *X*  *X* )2 | ( *p* 1)(*q* 1) | 2   2  *S*2 |
| факторів А і В | 2 *ij i j* | 2 ( *p* 1)(*q* 1) |
| Залишкова | *S*1  (*Xijk*  *X* )  2 2 | *N*  *pq* | 2   2  *S*1  1 *N*  *pq* |
| варіація | *ij* |
| Усього | *S*2  (*X*  *X* )2  0 *ijk* | *N* 1 | - |

Це означає, що в двохфакторному дисперсійному аналізі використовуються три різних критерії:

* 1. для перевірки гіпотези про відсутність ефекту фактора А;
  2. для перевірки гіпотези про відсутність ефекту фактора В;
  3. для перевірки гіпотези про відсутність ефекту взаємодії факторів А і В.

Після визначення *F*-критеріїв Фішера перевірка гіпотез здійснюється за таким самим алгоритмом, як і в однофакторному дисперсійному аналізі.

Двохфакторний дисперсійний аналіз може мати два різновиди:

1. без повторень;
2. з повтореннями.

У першому випадку кожному рівню факторів відповідає тільки одна вибірка даних, у другому – певним рівням факторів може відповідати більше однієї вибірки даних.

Унаслідок складності розрахунків, особливо при великій кількості рівнів кожного фактора, для двохфакторного аналізу слід застосовувати або *Excel*, або спеціалізоване програмне забезпечення.

Для здійснення розрахунків в *Excel* необхідно активізувати

«Пакет аналізу» і далі здійснити наступні кроки:

1. Обрати опцію «Двохфакторний дисперсійни аналіз без повторень (або з повтореннями)».
2. У діалоговому вікні з’являється адреси ознак, їх необхідно вказати (із попередньо складеного в *Excel* масиву вихідних даних).
3. Вказати кількість рядків для вибірки (враховуючи кількість ступенів свободи).
4. Встановити рівень надійності коефіцієнтів регресії (за замовчуванням – 95 %).
5. Командою ОК вивести результати на новий робочий аркуш.

## Питання для самоконтролю

* 1. Як визначається та що характеризує загальна дисперсія?
  2. Як визначається та що характеризує внутрішньогрупова дисперсія?
  3. Як визначається та що характеризує міжгрупова дисперсія?
  4. Як визначається та що характеризує емпіричне кореляційне відношення?
  5. Які завдання вирішує однофакторний дисперсійний аналіз?
  6. Які завдання вирішує двохфакторний дисперсійний аналіз?
  7. Що таке нульова і альтернативна гіпотези?
  8. Як здійснюється перевірка гіпотез в однофакторному дисперсійному аналізі?
  9. Скільки нульових гіпотез перевіряється у двохфакторному дисперсійному аналізі?
  10. Які фактори вважаються взаємодіючими?
  11. Що таке «кількість ступенів свободи»?
  12. Як визначається *F*-критерій Фішера і що він характеризує?
  13. Які умови повинні виконуватися для того, щоб застосування *F*-критерію Фішера було коректним?
  14. Чим відрізняється двохфакторний дисперсійний аналіз «без повторень» від його аналізу «з повтореннями»?