Лекція №2.

ТЕМА 2 ОСНОВИ ДИСПЕРСІЙНОГО АНАЛІЗУ

У практиці обробки результатів спостережень часто постає питання про те, наскільки істотно впливає зміна деякого фактора або групи факторів на вимірювану величину. Наприклад, можна досліджувати спільний вплив декількох економічних факторів, що не піддаються кількісному виміру на досліджуваний економічний показник, можна оцінити вплив властивостей сировини на показники якості продукції, кількості внесених добрив на врожайність і т.п.

У попередній главі була розглянута перевірка гіпотези про рівність математичних очікувань двох сукупностей при невідомих і однакових дисперсіях. Однак на практиці часто виникає необхідність узагальнення задачі, тобто перевірки при заданому рівні значимості за вибірковим середнім нульової гіпотези про рівність математичних очікувань p ($p > 2$) сукупностей, у яких також дисперсії невідомі й однакові. Для ефективного розв'язання таких задач застосовується новий підхід, що заснований на порівнянні дисперсій і тому його названо *дисперсійним аналізом* (розроблений англійським математиком–статистом Р.А. Фішером). Коротко суть цього аналізу зводиться до розчленовування загальної дисперсії ознаки на компоненти, обумовлені впливом конкретних факторів, і перевірці гіпотез про значимість їх впливу. Під *рівнем фактору* розуміють деяку його міру або стан. Моделі дисперсійного аналізу залежно від числа факторів класифікують як *однофакторні, двофакторні* й т.п. За метою дослідження виділяють такі моделі: *детермінована* – тут рівні всіх факторів заздалегідь фіксовані й перевіряють саме їхній вплив, *випадкова* – тут рівні кожного фактору отримані як випадкова вибірка з генеральної сукупності рівнів фактору, і *змішана* – тут рівні одних факторів заздалегідь фіксовані, а рівні інших – випадкова вибірка.

2.1 Однофакторний дисперсійний аналіз

В основі однофакторного дисперсійного аналізу лежить така теоретико – імовірнісна модель:

$$x_{ij} = \bar{x} + \alpha_j + \varepsilon_{ij}, \qquad (2.1)$$

де x_{ij} – значення ознаки X, отримане при i – му випробуванні на j – му рівні фактора; $\bar{x}$ – загальна середня величина ознаки X; α_j – ефект фактора на j – му рівні; ε_{ij} – випадковий компонент, що впливає на значення ознаки X в i – му спостереженні на j –му рівні фактору. Приймається припущення, що ε_{ij} взаємно незалежні й мають нормальний закон розподілу $N(0;\sigma^2)$.

2.2 Однакове число випробувань на різних рівнях

Нехай на кількісно нормально розподілену ознаку (випадкову величину) *X* впливає фактор *F*, що має *p* постійних рівнів. Будемо припускати, що число спостережень (випробувань) на кожному рівні однакове й дорівнює *q*.

Нехай спостерігалося $n = pq$ значень x_{ij} ознаки *X*, де, i – номер випробування $(i = \overline{1,q})$, j – номер рівня фактора $(j = \overline{1,p})$. Результати спостережень наведені в табл. 2.1.

Таблиця 2.1

Номер випробування	Рівні фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
...	…	…	…	
q	x_{q1}	x_{q2}	…	x_{qp}
Групова середня	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_p$

Позначимо через $\bar{x}_j$ – групове середнє *j-ї* групи.

Необхідно перевірити, чи суттєво впливає деякий якісний фактор *F*, що має *p* рівнів F_1, F_2 **,…,** F_p**,** на досліджувану ознаку *X.*

Якщо вважати, що елементи стовпців табл. 3.1, позначених $F_j (j = \overline{1,p})$, без елементів груповий середньої, є чисельними значеннями випадкових величин $X_1, X_2, \cdots, X_p$, що мають нормальний закон розподілу з математичними очікуваннями відповідно $m_1, m_2, \cdots, m_p$ й однаковими дисперсіями, то дана задача зводиться до перевірки нульової гіпотези $H_0 : m_1 = m_2 = \cdots = m_p$, здійснюваної в дисперсійному аналізі

$$\bar{x}_j = \frac{1}{q}\sum_{i=1}^{q} x_{ij} \qquad (2.2)$$

а через $\bar{x}$ *– загальне середнє*

$$\bar{x} = \frac{1}{pq}\sum_{j=1}^{p}\sum_{i=1}^{q} x_{ij} = \frac{1}{p}\sum_{j=1}^{p}\bar{x}_j \qquad (2.3)$$

тому що $\sum_{i=1}^{q} x_{ij} = q\bar{x}_j$

Розглянемо суму квадратів відхилень спостережуваних значень x_{ij} від загальної середньої $\bar{x}$:

$$\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}\right)^2=\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j+\bar{x}_j-\bar{x}\right)^2=\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)^2+$$
$$+2\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)\left(\bar{x}_j-\bar{x}\right)+\sum_{j=1}^{p}\sum_{i=1}^{q}\left(\bar{x}_j-\bar{x}\right)^2 \qquad (2.4)$$

причому

$$\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)\left(\bar{x}_j-\bar{x}\right)=\sum_{j=1}^{p}\left(\bar{x}_j-\bar{x}\right)\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)=0$$

оскільки

$$\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)=\sum_{i=1}^{q}x_{ij}-q\bar{x}_j=\sum_{i=1}^{q}x_{ij}-q\frac{1}{q}\sum_{i=1}^{q}x_{ij}=0$$

Тому, взявши до уваги, що

$$\sum_{j=1}^{p}\sum_{i=1}^{q}\left(\bar{x}_j-\bar{x}\right)^2=q\sum_{j=1}^{p}\left(\bar{x}_j-\bar{x}\right)^2$$

ми можемо основну тотожність (15.4) записати в такому вигляді:

$$\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}\right)^2=q\sum_{j=1}^{p}\left(\bar{x}_j-\bar{x}\right)^2+\sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)^2$$

або в скороченому виді

$$S_{общ}=S_{факт}+S_{ост} \qquad (2.5)$$

В (2.5) $S_{общ}$ – *загальна сума* квадратів відхилень спостережуваних значень від загальної середньої, $S_{факт}$ – *факторна сума* квадратів відхилень групових середніх від загальної середньої, котра характеризує розсіювання між групами, $S_{ост}$ – *залишкова сума* квадратів відхилень спостережуваних значень групи від своєї групової середньої, котра характеризує розсіювання усередині групи.

Практично залишкову суму знаходять за рівнянням

$$S_{ост}=S_{общ}-S_{факт} \qquad (2.6)$$

яке випливає з (3.5).

Одержимо більше зручні формули для розрахунків $S_{общ}$ і $S_{факт}$:

$$S_{îáù} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij}^2 - 2x_{ij}\bar{x} + \bar{x}^2\right) = \sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij}^2 -$$

$$- 2\bar{x}\sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij} + \sum_{j=1}^{p}\sum_{i=1}^{q}\bar{x}^2 = |(15.3)| = \sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij}^2 - 2\frac{1}{pq}\left(\sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij}\right)^2 + \qquad (2.7)$$

$$+ \frac{1}{pq}\left(\sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij}\right)^2 = \sum_{j=1}^{p}P_j - \frac{1}{n}\left(\sum_{j=1}^{p}R_j\right)^2;$$

$$S_{ôàêò} = q\sum_{j=1}^{p}\left(\bar{x}_j - \bar{x}\right)^2 = q\sum_{j=1}^{p}\left(\bar{x}_j^2 - 2\bar{x}_j\bar{x} + \bar{x}^2\right) = q\sum_{j=1}^{p}\bar{x}_j^2 - 2q\bar{x}\sum_{j=1}^{p}\bar{x}_j +$$

$$+ qp\bar{x}^2 = |(15.2);(15,3)| = \frac{1}{q}\sum_{j=1}^{p}\left(\sum_{i=1}^{q}x_{ij}\right)^2 - 2q\frac{1}{p}\sum_{j=1}^{p}\bar{x}_j\sum_{j=1}^{p}\bar{x}_j + \qquad (2.8)$$

$$+ \frac{q}{p}\left(\sum_{j=1}^{p}\bar{x}_j\right)^2 = \frac{1}{q}\sum_{j=1}^{p}\left(\sum_{i=1}^{q}x_{ij}\right)^2 - \frac{q}{p}\left(\sum_{j=1}^{p}\bar{x}_j\right)^2 = \frac{1}{q}\sum_{j=1}^{p}\left(\sum_{i=1}^{q}x_{ij}\right)^2 -$$

$$- \frac{1}{pq}\left(\sum_{j=1}^{p}\sum_{i=1}^{q}x_{ij}\right)^2 = \frac{1}{q}\sum_{j=1}^{p}R_j^2 - \frac{1}{pq}\left(\sum_{j=1}^{p}R_j\right)^2,$$

де $P_j = \sum_{i=1}^{q}x_{ij}^2$ – сума квадратів значень ознаки на рівні F_j;

$R_j = \sum_{i=1}^{q}x_{ij}$ – сума значень ознаки на рівні F_j

Зауваження 1. Обчислення за формулами (3.7) і (3.8) можна спростити, якщо виконати заміну змінних $y_{ij} = x_{ij} - C$, де C приблизно дорівнює загальній середній.

Підставляючи в ці формули $x_{ij} = y_{ij} + C$ та $R_j = \sum_{i=1}^{q}x_{ij} =$

$\sum_{i=1}^{q}(y_{ij} + C) = \sum_{i=1}^{q}y_{ij} + qC = T_j + qC$,

де $T_j = \sum_{i=1}^{q}y_{ij}$, з урахуванням позначення $Q_j = \sum_{i=1}^{q}y_{ij}^2$, одержимо

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(y_{ij}+C\right)^2 - \frac{1}{pq}\left(\sum_{j=1}^{p}\left(T_j+qC\right)\right)^2 = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(y_{ij}^2+2y_{ij}C+C^2\right) -$$

$$-\frac{1}{pq}\left(\sum_{j=1}^{p}T_j+pqC\right)^2 = \sum_{j=1}^{p}\left(Q_j+2CT_j+qC^2\right) - \frac{1}{pq}\left(\left(\sum_{j=1}^{p}T_j\right)^2+2pqC\sum_{j=1}^{p}T_j+p^2q^2C^2\right) =$$

$$= \sum_{j=1}^{p}Q_j + 2C\sum_{j=1}^{p}T_j + pqC^2 - \frac{1}{pq}\left(\sum_{j=1}^{p}T_j\right)^2 - 2C\sum_{j=1}^{p}T_j - pqC^2 = \sum_{j=1}^{p}Q_j - \frac{1}{n}\left(\sum_{j=1}^{p}T_j\right)^2 \tag{2.9}$$

$$S_{факт} = \frac{1}{q}\sum_{j=1}^{p}T_j^2 - \frac{1}{pq}\left(\sum_{j=1}^{p}T_j\right)^2 \tag{2.10}$$

Зауваження 2. Якщо спостережувані значення x_{ij} – десяткові дроби з k знаками після коми, то доцільно перейти до чисел $y_{ij} = 10^k x_{ij} - C$, де C – приблизно середнє значення чисел $10^k x_{ij}$. Хоча при цьому факторна й залишкова дисперсії збільшуються в 10^{2k} разів, їхнє відношення не зміниться.

У розкладанні (3.5) закладена основна ідея дисперсійного аналізу. Однак у ньому аналізуються не самі суми квадратів відхилень, а так названі середні квадрати, що є незміщеними оцінками відповідних дисперсій, які виходять діленням сум квадратів відхилень на відповідне число ступенів волі.

Нагадаємо, що *число ступенів волі* визначається як загальне число ступенів спостережень мінус число об'єднуючих їхніх рівнянь. Тому число ступенів волі загальної дисперсії дорівнює $pq–1 = n–1$, тому що один ступінь волі втрачається при визначенні середньої. Аналогічне число ступенів волі факторної дисперсії дорівнює $p–1$, тому що групові середні варіюють навколо однієї загальної середньої. Нарешті, число ступенів волі залишкової дисперсії дорівнює n-p, тому що використовується p співвідношень при обчисленні p групових середніх $\bar{x}_j$;.

Використовуючи отримані значення сум квадратів і чисел ступенів волі, можна обчислити незміщені оцінки трьох дисперсій:

$$s_{общ}^2 = \frac{S_{общ}}{n-1}; \qquad s_{факт}^2 = \frac{S_{факт}}{p-1}; \qquad s_{ост}^2 = \frac{S_{ост}}{n-p} \tag{2.11}$$

Зауваження 3. Число ступенів волі залишкової дисперсії $pq–p$ дорівнює різниці між числами ступенів волі загальної й факторної дисперсій. Дійсно,

$$pq–1–(p–1) = pq–p = n–p.$$

Перевірка гіпотези H_0 про рівність групових математичних очікувань ґрунтується на порівнянні дисперсій $S^2_{факт}$ і $S^2_{ост}$. Виявляється, якщо гіпотеза H_0 вірна, то вірна й гіпотеза про рівність факторної й залишкової дисперсій, що перевіряється за критерієм Фішера – Снедекора

$$F = S^2_{факт} / S^2_{ост} \tag{2.12}$$

який має ступені волі $k_1 = p-1$ і $k_2 = n-p$.

Якщо нульова гіпотеза про рівність групових середніх помилкова, то помилково й гіпотезу про рівність $S^2_{факт}$ і $S^2_{ост}$. Справедливі також і зворотні твердження.

Подамо схему перевірки нульової гіпотези у випадку правобічної критичної області:

$$\left.\begin{array}{c} p;\quad q;\quad n = pq \\ k_1 = p-1;\ k_2 = n-p \\ \alpha \end{array}\right\} \to H_0 : m_1 = m_2 = \ldots = m_p \to$$

$$\to F_{набл} = s^2_{факт} / s^2_{ост} \to H_1 : (\text{правост.крит.обл.}) \to \tag{2.13}$$

$$\to F_{кр}(\alpha; k_1; k_2) \longrightarrow x^{пр}_{кр} \to \begin{cases} F_{набл} < x^{пр}_{кр} \to & \textit{Гіпотеза приймається} \\ F_{набл} > x^{пр}_{кр} \to & \textit{Гіпотеза відхиляється} \end{cases}$$

Зауваження 4. Якщо $S^2_{факт} < S^2_{ост}$, то звідси виникає справедливість нульової гіпотези H_0 і, виходить, немає необхідності використовувати критерій F.

Зауваження 5. Якщо немає впевненості в справедливості припущення про рівність дисперсій розглянутих p сукупностей, то це припущення варто перевірити попередньо, наприклад за критерієм Кочрена.

Приклад. Зроблено по 4 випробування на кожному із чотирьох рівнів. Результати випробувань наведені в табл. 3.2. Методом дисперсійного аналізу при рівні значимості α = 0,05 перевірити нульову гіпотезу про рівність групових середніх. Передбачається, що вибірки взяті з нормальних сукупностей.

Знайдемо за формулами

$$\bar{x}_j = \frac{1}{4}\sum_{i=1}^{4} x_{ij},\ \ \bar{\sigma}^2_j = \frac{1}{4}\sum_{i=1}^{4}\left(x_{ij} - \bar{x}_j\right)^2,\ \ j = \overline{1,4}$$

групові вибіркові середні й дисперсії, результати обчислень яких занесемо в табл. 2.2.

Таблиця 2.2

Номер випробування	Рівні фактора *Fj*

i	F_1	F_2	F_3	F_4
1	140	150	148	150
2	144	149	149	155
3	142	152	146	154
4	145	150	147	152
Групове середнє, $\bar{x}_j$	142,75	150,25	147,5	152,75
Вибіркова дисперсія, $\bar{\sigma}_j^2$	3,6875	1,1875	1,25	3,6875

Перш ніж проводити дисперсійний аналіз, переконаємося в тому, що нульова гіпотеза

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$

про рівність групових дисперсій для вибірок, взятих з нормальних генеральних сукупностей, не суперечить результатам спостережень. Оскільки вибірки містять однаковий об'єм q = 4, то при перевірці нульової гіпотези доцільно скористатися критерієм Кочрена. Для цього:

1. Знайдемо виправлені вибіркові дисперсії:

$$S_1^2 = \frac{\bar{\sigma}_1^2 q}{q-1} = \frac{3{,}6875 \cdot 4}{3} = 4{,}917; \qquad S_2^2 = 1{,}583; \quad S_3^2 = 1{,}667; \quad S_4^2 = 4{,}917.$$

2. Знайдемо спостережуване значення критерію Кочрена:

$$G_{набл} = S_{\max} / \left(S_1^2 + S_2^2 + S_3^2 + S_4^2\right) = 4{,}917 / 13{,}084 \approx 0{,}376.$$

3. Знайдемо з таблиці (додаток 8), за рівнем значимості α= 0,05, число ступенів волі k = 4 –1 = 3 і число вибірок l = *4* критичну точку G(0,05; 3; 4) = $g_{кр}$ = 0,6841.

4. Оскільки $G_{набл} < g_{кр}$, то гіпотезу H_0 приймаємо.

Гіпотеза, що перевіряється в дисперсійному аналізі, має вигляд:

H_0: $m_1=m_2=m_3=m_4$,

де m_i – математичне очікування *i–оїї* генеральної сукупності.

Для спрощення розрахунку віднімемо C=148 з кожного спостережуваного значення: $y_{ij} = x_{ij} - 148$. Складемо розрахункову табл. 2.3.

Таблиця 2.3

Номер випробування	Рівні фактора Fj								Підсумковий стовпець
i	F_1		F_2		F_3		F_4		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	y_{i4}	y_{i4}^2	
1	–8	64	*2*	4	0	0	*2*	4	–

2	–4	16	1	1	1	1	7	49	
3	–6	36	4	16	–2	4	6	36	
4	–3	9	2	4	–1	1	4	16	
$Q_j = \sum_{i=1}^{4} y_{ij}^2$	–	125	–	25	–	6	–	105	$\sum_{i=1}^{4} Q_j = 261$
$T_j = \sum_{i=1}^{4} y_{ij}$	–21	–	9	–	–2	–	19	–	$\sum_{i=1}^{4} T_j = 5$
T_j^2	441	–	81	–	4	–	361	–	$\sum_{i=1}^{4} T_j^2 = 887$

Користуючись табл. 15.3 і з огляду на те, що число рівнів фактору $p = 4$, число випробувань на кожному рівні $q = 4$, знайдемо за формулами (3.9) і (3.10) загальну й факторну суми квадратів відхилень:

$S_{общ} = 261 – 25/16 = 259{,}4375;$

$S_{факт} = 887/4 – 25/16 = 220{,}1875.$

Знайдемо залишкову суму квадратів відхилень:

$S_{ост} = S_{общ} - S_{факт} = 39,25.$

Факторну й залишкову дисперсії знайдемо за формулами (3.11):

$S_{факт}^2 = 220{,}1875/3 \approx 73{,}4;$

$S_{ост}^2 = 39{,}25/(16-4) \approx 3{,}27.$

Перевірку гіпотези H_0 про рівність групових математичних очікувань проведемо за схемою (15.13):

$$\left.\begin{array}{l} p = 4;\ q = 4; \\ n = 16;\ k_1 = 3 \\ k_2 = 12;\ \alpha = 0{,}05 \end{array}\right\} \rightarrow H_0 : m_1 = m_2 = m_3 = m_4 \rightarrow F_{набл} = s_{факт}^2 / s_{ост}^2 = 22{,}45 \rightarrow$$

$$\rightarrow H_1 : (правост.крит.обл.) \rightarrow F_{кр}(0{,}05;3;12) \rightarrow x_{кр}^{пр} = 3{,}49 \rightarrow F_{набл} > x_{кр}^{пр} \rightarrow$$

$\rightarrow$ *Гіпотеза відхиляється*

Таким чином, групові середні істотно розрізняються. Якщо потрібно зрівняти середні попарно, то варто скористатися критерієм Стьюдента.

2.3 Неоднакове число випробувань на різних рівнях

Нехай число випробувань на різних рівнях різне, а саме: зроблено q_1 випробувань на рівні F_1, q_2 випробувань – на рівні F_2, ..., q_p випробувань – на рівні F_p . У цьому випадку загальну суму квадратів відхилень також знаходять

за формулою (15.7), де $P_1 = \sum_{i=1}^{q_1} x_{i1}^2$, $P_2 = \sum_{i=1}^{q_2} x_{i2}^2$, …, $P_p = \sum_{i=1}^{q_p} x_{ip}^2$ – суми квадратів значень, що спостерігалися, ознаки відповідно на *рівнях* F_1, F_2, …, F_p; $R_1 = \sum_{i=1}^{q_1} x_{i1}^2$, $R_2 = \sum_{i=1}^{q_2} x_{i2}^2$, … , $R_p = \sum_{i=1}^{q_p} x_{ip}^2$ –суми значень, що спостерігалися, ознаки відповідно на *рівнях* F_1, F_2, …, F_p; $n=q_1+q_2+...+q_p$ – об'єм вибірки, а факторну суму квадратів відхилень знаходять за формулою

$$S_{факт} = \sum_{j=1}^{p}\left(R_j^2 / q_j\right) - \left(\sum_{j=1}^{p} R_j\right)^2 / n \qquad (2.14)$$

Якщо для спрощення обчислень уводиться заміна змінних $y_{ij} = x_{ij} - C$, де C приблизно дорівнює загальній середній, то формула для загальної суми квадратів відхилень має вигляд (15.9),

де

$$Q_1 = \sum_{i=1}^{q_1} y_{i1}^2, \quad Q_2 = \sum_{i=1}^{q_2} y_{i2}^2 \quad , ..., \quad Q_p = \sum_{i=1}^{q_p} y_{ip}^2,$$

$$T_1 = \sum_{i=1}^{q_1} y_{i1}, \quad T_2 = \sum_{i=1}^{q_2} y_{i2} \quad , ..., \quad T_p = \sum_{i=1}^{q_p} y_{ip},$$

а формула для факторної суми відхилень має вигляд

$$S_{факт} = \sum_{j=1}^{p}\left(T_j^2 / q_j\right) - \left(\sum_{j=1}^{p} T_j\right)^2 / n \qquad (2.15)$$

Інші обчислення здійснюють, як і у випадку однакового числа випробувань за формулами (3.6) і (3.11).

Приклад. Проведено 22 випробування, з яких 7 на першому рівні фактору, 6 – на другому, 5 – на третьому й 4 – на четвертому. Результати випробувань наведені в табл. 3.4. Методом дисперсійного аналізу при рівні значимості α = 0,05 перевірити нульову гіпотезу про рівність групових середніх. Передбачається, що вибірки взяті з нормальних сукупностей з однаковими дисперсіями.

Таблиця 2.4

Номер випробування	Рівні фактора F_j.			
i	F_1	F_2	F_3	F_4

1	1,3	1,4	1,44	1,27
2	1,27	1,3	1,4	1,05
3	1,21	1,28	1,28	1,24
4	1,09	1,27	1,28	1,22
5	1,03	1,24	1,06	
6	1,01	1,08	—	
7	1,09	–	–	–

Для спрощення розрахунку помножимо кожне спостережуване значення x_{ij} на 100 і віднімемо $3 = 122$ (зауваження 2). Складемо розрахункову таблицю 2.5.

Таблиця 2.5

Номер випробування	Рівні фактора F_j.								Підсумковий стовпець
i	F_1		F_2		F_3		F_4		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	y_{i4}	y_{i4}^2	
1	8	64	18	324	22	484	5	25	
2	5	25	8	64	18	324	–17	289	
3	– 1	1	6	36	6	36	2	4	
4	–13	169	5	25	6	36	0	0	
5	–19	361	2	4	–16	256	–	–	
6	–21	441	– 14	196	–	–	–	–	
7	–13	169							
$Q_j = \sum_{i=1}^{4} y_{ij}^2$	–	1230	–	649	–	1136	–	318	$\sum_{i=1}^{4} Q_j = 3333$
$T_j = \sum_{i=1}^{4} y_{ij}$	–54	–	25	–	36	–	–10	–	$\sum_{i=1}^{4} T_j = -3$
T_j^2	2916	–	625	–	1296	–	100	–	–

Використовуючи табл. 3.5, знайдемо загальну й факторну суми квадратів відхилень:

$$S_{общ} = \sum_{j=1}^{4} Q_j - \left(\sum_{j=1}^{4} T_j \right)^2 / n = 3333 - 9/22 \approx 3332{,}6;$$

$$S_{факт} = \sum_{j=1}^{4} \left(T_j^2 / q_j \right) - \left(\sum_{j=1}^{4} T_j \right)^2 / n = 416{,}571 + 104{,}167 +$$

$$+ 259{,}2 + 25 - 9/22 = 804{,}938 - 9/22 \approx 804{,}529.$$

Знайдемо залишкову суму квадратів відхилень:

$$S_{ост} = S_{общ} - S_{факт} = 2528{,}071 .$$

Факторну й залишкову дисперсії знайдемо за формулами (3.11):

$S^2_{факт} = S_{факт}/(p–1) = 804{,}529/3 \approx 268{,}176;$

$S^2_{ост} = S_{ост}/(n–p) = \approx 2528{,}071/18 \approx 140{,}448.$

Порівняємо $S^2_{факт}$ й $S^2_{ост}$ за критерієм Фішера–Снедекора, використовуючи схему (3.13)

$$\left.\begin{array}{l} p=4;\ n=22; \\ k_1=3;\ k_2=18 \\ \alpha=0{,}05 \end{array}\right\} \to H_0: m_1=m_2=m_3=m_4 \to F_{набл}=s^2_{факт}/s^2_{ост}=$$

$$=268{,}176/140{,}448\approx 1{,}9094 \to H_1:(правост.\ крит.\ обл.)\to$$

$$\to F_{кр}(0{,}05;3;18)\longrightarrow x^{пр}_{кр}=3{,}16 \to F_{набл}<x^{пр}_{кр}\to$$

$$\to Гіпотеза\ приймається$$

2.4 Поняття про двофакторний дисперсійний аналіз

В основі двофакторного дисперсійного аналізу лежить наступна теоретико – імовірнісна модель:

$$x_{ijk} = \bar{x} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad (2.16)$$

де x_{ijk} – значення ознаки *X* в *k-му* спостереженні на *i-му* рівні фактора *A* і на *j-му* рівні фактора *B;* $\bar{x}$ – загальна середня величина ознаки *X;* α_i – ефект впливу фактору *A* на *i-му* рівні; β_j – ефект впливу фактору *B* на *j-му* рівні; γ_{ij} – ефект спільного впливу факторів (позначимо *A*×*B);* ε_{ijk} – незалежні нормально розподілені випадкові компоненти $N(0;\sigma^2)$, що являють собою відхилення ознаки від відповідних середніх.

Нехай необхідно виявити вплив двох факторів *(A* і *B)* та їхню взаємодію на деяку ознаку *X*. Спостереження проводяться при фіксованих рівнях факторів *A* і *B*. Оскільки для кожного поєднання факторів спостереження повторюється *n* разів, то відповідно одержимо *n* значень ознаки. Дані спостережень зручно подати у вигляді таблиці 3.6, у якій значення ознаки *X* позначене через x_{ijk}, де $i=\overline{1,q}$ *(q* – число спостережень фактора *A);* $j=\overline{1,p}$ *(p* – число спостережень фактора *B);* $k=\overline{1,n}$ *(k* – порядковий номер спостереження для кожного поєднання рівнів).

Таблиця 3.6

Рівень фактора A	Рівень фактора B				Сума P_i
	B_1	B_2	…	B_p	
A_1	x_{111}, x_{112} $x_{113}, \ldots, x_{11n}$	x_{121}, x_{122} $x_{123}, \ldots, x_{12n}$	…	x_{1p1}, x_{1p2} $x_{1p3}, \ldots, x_{1pn}$	$\sum_{j=1}^{p}\sum_{k=1}^{n} x_{1jk}$
A_2	x_{211}, x_{212} $x_{213}, \ldots, x_{21n}$	x_{221}, x_{222} $x_{223}, \ldots, x_{22n}$	…	x_{2p1}, x_{2p2} $x_{2p3}, \ldots, x_{2pn}$	$\sum_{j=1}^{p}\sum_{k=1}^{n} x_{2jk}$
…	…	…	…	…	…
A_q	x_{q11}, x_{q12} $x_{q13}, \ldots, x_{q1n}$	x_{q21}, x_{q22} $x_{q23}, \ldots, x_{q2n}$	…	x_{qp1}, x_{qp2} $x_{qp3}, \ldots, x_{qpn}$	$\sum_{j=1}^{p}\sum_{k=1}^{n} x_{qjk}$
Сума R_j	$\sum_{i=1}^{q}\sum_{k=1}^{n} x_{i1k}$	$\sum_{i=1}^{q}\sum_{k=1}^{n} x_{i2k}$	…	$\sum_{i=1}^{q}\sum_{k=1}^{n} x_{ipk}$	$\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} x_{ijk}$

За даними таблиці одержимо такі середні: загальна середня

$$\bar{x} = \frac{1}{qpn}\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} x_{ijk} \tag{2.17}$$

середні за рядками

$$\overline{P}_i = \frac{1}{pn}P_i = \frac{1}{pn}\sum_{j=1}^{p}\sum_{k=1}^{n} x_{ijk},\ i=\overline{1,q} \tag{2.18}$$

середні за стовпчиками

$$\overline{R}_j = \frac{1}{qn}R_j = \frac{1}{qn}\sum_{i=1}^{q}\sum_{k=1}^{n} x_{ijk},\ \ j=\overline{1,p} \tag{2.19}$$

середні для кожного окремого блоку таблиці

$$\overline{Q}_{ij} = \frac{1}{n}Q_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk} \tag{2.20}$$

Суми квадратів відхилень від загальної середньої $\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{ijk}-\bar{x}\right)^2$ розкладемо на складові. Для цього представимо $x_{ijk}-\bar{x}$ в еквівалентній формі

$$x_{ijk}-\bar{x} = \left(\overline{R}_j-\bar{x}\right)+\left(\overline{Q}_{ij}-\overline{R}_j-\overline{P}_i+\bar{x}\right)+\left(x_{ijk}-\overline{Q}_{ij}\right)$$

Оскільки всі перехресні добутки, при піднесенні правої частини останнього вираження у квадрат, дорівнюють нулю, то в підсумку одержимо

$$\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{ijk}-\bar{x}\right)^2=\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\bar{R}_j-\bar{x}\right)^2++\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\bar{P}_i-\bar{x}\right)^2+$$

$$+\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\bar{Q}_{ij}-\bar{R}_j-\bar{P}_i+\bar{x}\right)^2+\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{ijk}-\bar{Q}_{ij}\right)^2$$

або в скороченому вигляді

$$S=S_1+S_2+S_3+S_4. \qquad (2.21)$$

У (15.21) чотири складові: сума квадратів, пов'язана із впливом фактора *B,* фактора *A,* їхньої взаємодії, і складова, що характеризує залишкову суму квадратів (суму квадратів усередині кожного блоку таблиці).

Загальне число ступенів волі, мабуть, дорівнює *N–1,* де *N* – загальне число спостережень (*N = qpn)*. Число ступенів волі між стовпцями дорівнює *p* – 1, між рядками *q* –1, для взаємодії (*p* – 1)(*q* – 1), усередині осередків *pq*(*n* – 1) = *N* – *pq*. Для перевірки знайдемо суму *p* –1 + *q* –1 + +(*p*–1)(*q* – 1)+ *N* – *pq* = *N* –1.

Схема дисперсійного аналізу, заснована на отриманих вище даних, представлена в табл. 2.7.

Таблиця 2.7

Фактор	Сума квадратів	Число ступенів волі	Оцінка дисперсії
B	$S_1=qn\sum_{j=1}^{p}\left(\bar{R}_j-\bar{x}\right)^2$	$p–1=k_1$	$s_1^2=\frac{S_1}{p-1}$
A	$S_2=pn\sum_{i=1}^{q}\left(\bar{P}_i-\bar{x}\right)^2$	$q–1=k_2$	$s_2^2=\frac{S_2}{q-1}$
$A\times B$	$S_3=n\sum_{i=1}^{q}\sum_{j=1}^{p}\left(\bar{Q}_{ij}-\bar{R}_j-\bar{P}_i+\bar{x}\right)^2$	$(p–1)\times(q–1)=k_3$	$s_3^2=\frac{S_3}{(p-1)(q-1)}$
Залишковий	$S_4=\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{ijk}-\bar{Q}_{ij}\right)^2$	$N–pq=k_4$	$s_4^2=\frac{S_4}{N-pq}$
Сума	$S=\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{ijk}-\bar{x}\right)^2$	$N–1$	$s^2=\frac{S}{N-1}$

Оскільки при проведенні аналізу інтерес являє вплив кожного фактору порізно й вплив їхньої взаємодії, то знаходимо відповідно три значення:

$$F_B=s_1^2/s_4^2;\quad F_A=s_2^2/s_4^2;\quad F_{AB}=s_3^2/s_4^2 \qquad (2.22)$$

Необхідні для аналізу суми квадратів відхилень (15.21) можемо одержати також за такими формулами:

$$S_1 = \frac{1}{nq}\sum_{j=1}^{p} R_j^2 - \frac{G^2}{N} \tag{2.23}$$

$$S_2 = \frac{1}{np}\sum_{i=1}^{q} P_i^2 - \frac{G^2}{N} \tag{2.24}$$

$$S_3 = \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{p} Q_{ij}^2 - \frac{1}{nq}\sum_{j=1}^{p} R_j^2 - \frac{1}{np}\sum_{i=1}^{q} P_i^2 + \frac{G^2}{N} \tag{2.25}$$

$$S_4 = \sum_{i=1}^{q}\sum_{j=1}^{p}\left(\sum_{k=1}^{n} x_{ijk}^2 - \frac{Q_{ij}^2}{n}\right) \tag{2.26}$$

де

$$G = \sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n} x_{ijk},\ \ Q_{ij} = \sum_{k=1}^{n} x_{ijk},\ \ F_{ij} = \sum_{k=1}^{n} x_{ijk}^2 - \frac{Q_{ij}^2}{n} \tag{2.27}$$

При рівні значимості α визначаємо критичні точки:

$$F_{кр}(\alpha;k_1;k_4) = x_{кр}^B;\ F_{кр}(\alpha;k_2;k_4) = x_{кр}^A;\ F_{кр}(\alpha;k_3;k_4) = x_{кр}^{AB}$$

Якщо:

$F_B > x_{кр}^B$, то нульова гіпотеза про відсутність впливу фактору *B* відхиляється;

$F_A > x_{кр}^A$, то нульова гіпотеза про відсутність впливу фактору *A* відхиляється;

$F_{AB} > x_{кр}^{AB}$, то нульова гіпотеза про відсутність спільного впливу факторів *A* і *B* відхиляється.

Приклад. При рівні значимості $\alpha = 0{,}05$ перевірити, чи існує вплив факторів *A* і *B,* а також їх спільного впливу на ознаку *X*, для результатів випробувань, наведених у табл. 3.8.

Таблиця 2.8

Рівень фактора *A*	Рівень фактора *B*		
	B₁	***B₂***	***B₃***
A₁	10; 8; 7; 10	8; 12; 14; 12	15; 8; 10; 10
A₂	12; 8; 8; 7	12; 13; 11; 14	13; 15; 12; 10

Використовуючи табл. 2.8 і формули (3.23) ÷ (3.27), складемо табл. 2.9 і обчислимо суми квадратів відхилень $S_i (i = \overline{1,4})$

Таблиця 2.9

B / ***A***	***B₁***		***B₂***		***B₃***		$\sum_{j=1}^{3} Q_{ij}^2$	P_i	P_i^2
	x_{i1k}	x_{i1k}^2	x_{i2k}	x_{i2k}^2	x_{i3k}	x_{i3k}^2			

B / A	B_1		B_2		B_3		$\sum_{j=1}^{3} Q_{ij}^2$	P_i	P_i^2
	x_{i1k}	x_{i1k}^2	x_{i2k}	x_{i2k}^2	x_{i3k}	x_{i3k}^2			
A_1	10	100	8	64	15	225	–	–	–
	8	64	12	144	8	64			
	7	49	14	196	10	100			
	10	100	12	144	10	100			
Q_{1j}	35	–	46	–	43	–	–	124	15376
Q_{1j}^2	1225	–	2116	–	1849	–	5190	–	–
$\sum_{k=1}^{4} x_{1jk}^2$	–	313	–	548	–	489	–	–	–
F_{1j}	6,75		19		26,75		$\sum_{j=1}^{3} F_{1j} = 52{,}5$		
A_2	12	144	12	144	13	169	–	–	–
	8	64	13	169	15	225			
	8	64	11	121	12	144			
	7	49	14	196	10	100			
Q_{2j}	35	–	50	–	50	–	–	135	18225
Q_{2j}^2	1225	–	2500	–	2500	-	6225	-	-
$\sum_{k=1}^{4} x_{2jk}^2$	–	321	–	630	–	638	–	–	–
F_{2j}	14,75		5		13		$\sum_{j=1}^{3} F_{2j} = 32{,}75$		
R_j	70	–	96	–	93	–	–	G=259	$\sum_{i=1}^{2} P_i^2 = 33601$
R_j^2	4900	-	9216	-	8649	-	$\sum_{j=1}^{3} R_i^2 = 22765$		

З табл. 2.9 маємо:

$$S_1 = \frac{1}{8}\sum_{j=1}^{3} R_j^2 - \frac{G^2}{24} = \frac{1}{8}\cdot 22765 - \frac{529}{24} \approx 50{,}58;$$

$$S_2 = \frac{1}{12}\sum_{i=1}^{2} P_i^2 - \frac{G^2}{24} = \frac{1}{12}\cdot 33601 - \frac{529}{24} \approx 5{,}04;$$

$$S_3 = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{3} Q_{ij}^2 - \frac{1}{8}\sum_{j=1}^{3} R_j^2 - \frac{1}{12}\sum_{i=1}^{2} P_i^2 + \frac{G^2}{24} =$$

$$= \frac{1}{4}(5190 + 6225) - 2845{,}625 - 2800{,}083 + 2795{,}0417 \approx 3{,}083;$$

$$S_4 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left(\sum_{k=1}^{4} x_{ijk}^2 - Q_{ij}^2/4\right) = 52{,}5 + 32{,}75 = 85{,}25$$

Дані дисперсійного аналізу наведемо в табл. 2.10.

Таблиця 2.10

Джерело варіації		Число ступенів волі	Оцінка дисперсії
Фактор *B*	50,28	2	25,29
Фактор *A*	5,04	1	5,04
A×*B*	3,083	2	1,54
Залишкова варіація	85,25	18	4,74
$\sum$	143,95	23	–

Визначимо спостережувані значення критерію:

$F_{набл}^{B}$= 25,29/4,74 ≈ 5,34; $F_{набл}^{A}$=5,04/4,74 ≈ 1,06;

$F_{набл}^{AB}$= 1,54/4,74≈0,32

і порівняємо їх з відповідними критичними значеннями:

$F_{кр}(0{,}05;\ 2;\ 18) = x_{кр}^{B} = 3{,}55;\ F_{кр}(0{,}05;\ l;\ 18) = x_{кр}^{A} = 4{,}41;$

$F_{кр}(0{,}05;\ 2;\ 18) = x_{кр}^{AB} = 3{,}55.$

Оскільки $F_{набл}^{B} > x_{кр}^{B}$, то нульова гіпотеза про відсутність впливу фактору *B* відхиляється й варто зробити висновок про значимості впливу цього фактору. Оскільки $F_{набл}^{A} > x_{кр}^{A}$ і $F_{набл}^{AB} > x_{кр}^{AB}$, то немає підстави для відхилення відповідних нульових гіпотез.

2.5 Питання для самоперевірки теми 2

1. У чому суть дисперсійного аналізу?
2. Наведіть класифікацію моделей дисперсійного аналізу за числом

факторів і за метою дослідження.

3. Запишіть теоретико – ймовірнісну модель для однофакторного дисперсійного аналізу.
4. Що називається груповою і загальною середньою? Назвіть формули для їхнього обчислення.
5. Вивести основну тотожність для суми квадратів відхилень спостережуваних значень x_{ij} від загальної середньої $\bar{x}$.
6. Вивести формули для розрахунку $S_{общ}$ й $S_{факт}$ через P_j і R_j.
7. Вивести формули для розрахунку $S_{общ}$ й $S_{факт}$ через Q_j. і T_j.
8. Наведіть формули для обчислення незміщених оцінок трьох дисперсій:. $S^2_{общ}$, $S^2_{факт}$, $S^2_{ост}$
9. Наведіть схему для перевірки нульової гіпотези про рівність групових середніх.
10. Наведіть основні й спрощені формули для розрахунку $S_{общ}$ й $S_{факт}$ при неоднаковому числі випробувань на різних рівнях.
11. Запишіть теоретико – ймовірнісну модель для двофакторного дисперсійного аналізу.
12. Наведіть формули для обчислення середніх при двофакторному дисперсійному аналізі.
13. Дайте характеристику чотирьох складового розкладання суми квадратів відхилень від загальної середньої.
14. Наведіть формули для розрахунку S_1,, S_2, S_3 і S_4.