Дисперсійний аналіз
(ANalysis Of VAriance, ANOVA)

Дисперсійний аналіз – це один з розділів аналізу даних, який займається побудовою математичних моделей істотних зв'язків між залежними кількісними змінними та незалежними якісними змінними.

Основи дисперсійного аналізу заклав у першій половині XX ст. відомий англійський математик, статистик Рональд Ейлмер Фішер (Ronald Aylmer Fisher).

Задачі дисперсійного аналізу виникають в різних галузях, а саме: у науці, бізнесі, економіці, фінансах, медицині, біології, промисловості, сільському господарстві, соціології тощо.

Приклад 1. Залежна змінна η – врожайність зернової культури, незалежна змінна ζ – сорт зернової культури, причому загальна кількість сортів зернової культури дорівнює I.

Приклад 2. Залежна змінна η – врожайність зернової культури, незалежна змінна ζ_1 – сорт зернової культури, причому загальна кількість сортів зернової культури дорівнює I_1,
незалежна змінна ζ_2 – вид добрива, загальна їх кількість I_2.

Приклад 3. Залежна змінна η – результати голосування за певну кандидатуру на виборах,
незалежна змінна ζ – передвиборча технологія, яка була використана під час передвиборчої компанії.

Приклад 4. Залежна змінна η – кількісний показник якості виплавленого металу,
незалежна змінна ζ – технологія, яка використана при його плавці.

Приклад 5. Залежна змінна η – прибуток від продажу товару, незалежна змінна ζ – маркетингова стратегія при його реалізації.

Приклад 6. Залежна змінна η_1 – кількісний показник ризику зараження віріоном (коронавірусом) SARS-CoV-2 людини,
незалежна змінна ζ_1 – стать,
незалежна змінна ζ_2 – група крові (0(I), A(II), B(III), AB(IV)),
незалежна змінна ζ_3 – резус-фактор крові (Rh^+, Rh^-),
незалежна змінна ζ_4 – раса пацієнта.

Якщо у цьому прикладі не одна скалярна залежна кількісна змінна η_1, а до неї ще додати декілька скалярних залежних кількісних змінних, наприклад:
η_2 – важкість перебігу COVID-19 (тривалість необхідного лікування),
η_3 – рівень смертності від хвороби COVID-19,
то це вже буде задача **багатовимірного дисперсійного аналізу** (Multivariate ANalysis Of VAriance, MANOVA).

Далі увага буде зосереджена на задачах ANOVA.

Постановка задачі дисперсійного аналізу. Нехай
η - залежна кількісна скалярна змінна,
$\zeta_1, \zeta_2, \ldots, \zeta_q$ - незалежні якісні скалярні змінні.

Потрібно по спостереженням над η та $\zeta_1, \zeta_2, \ldots, \zeta_q$
та апріорній інформації про невизначеності
побудувати математичну модель залежності η від $\zeta_1, \zeta_2, \ldots, \zeta_q$.

Незалежні якісні змінні $\zeta_1, \zeta_2, \ldots, \zeta_q$ ще називають *факторами*, а для їх нотації можуть використовуватися відповідні великі літери латинської абетки, тобто A, B, C, … .

Структура дисперсійного аналізу:

- Однофакторний дисперсійний аналіз
- Двофакторний дисперсійний аналіз
- Багатофакторний дисперсійний аналіз

Однофакторний дисперсійний аналіз

Постановка задачі однофакторного дисперсійного аналізу. Нехай
η - залежна кількісна скалярна змінна,
ζ - незалежна якісна скалярна змінна, яка набуває своїх значень з I градацій.

Необхідно за спостереженнями над залежною змінною η при активних різних градаціях незалежної змінної ζ побудувати математичну модель залежності змінної η від змінної ζ.

Фон: Приклад 1: η – врожайність зернової культури,
ζ – сорт зернової культури.

Припустимо, що при активній i-й градації змінної ζ доступні:

$$\eta: \quad y_{ik},\ i=\overline{1,I},\ k=\overline{1,N_i}\ \ (N_i \geq 1).$$

Загальна кількість спостережень $N=\sum_{i=1}^{I} N_i$.

Модель однофакторного дисперсійного аналізу шукаємо у вигляді:

$$y_{ik}=\mu+\alpha_i+e_{ik},\quad i=\overline{1,I},\ k=\overline{1,N_i}\ \ (N_i \geq 1), \qquad (1)$$

де
y_{ik} – k-те спостереження над η при активній i-й градації змінної ζ,
μ – середнє в деякому розумінні всіх таких спостережень,
α_i – *кількісний вираз відносного впливу i-ї градації змінної ζ* на η відносно μ,
e_{ik} – похибка k-го спостереження над η при активній i-й градації змінної ζ.

А *кількісний вираз абсолютного впливу i-ї градації змінної ζ* на η :

$$a_i=\mu+\alpha_i,\quad i=\overline{1,I}.$$

Припустимо, що похибки $e_{ik}, i=\overline{1,I},\ k=\overline{1,N_i}\ \left(N_i \geq 1\right)$ моделі (1) є:

- $e_{ik} \sim \mathcal{N}\left(0,\sigma^2\right), \sigma^2 > 0, \quad \forall i,k,$
- $\left\{e_{ik}\right\}$ - незалежні.

Таким чином, потрібно для моделі (1) за спостереженнями $\left\{y_{ik}, i=\overline{1,I},\ k=\overline{1,N_i}\ \left(N_i \geq 1\right)\right\}$ знайти оцінки невідомих параметрів $\mu,\alpha_1,\alpha_2,\ldots,\alpha_I$. Оцінки параметрів будемо шукати методом найменших квадратів.

Представимо систему рівнянь (1) у матричному вигляді:

$$y = X\alpha + e, \tag{2}$$

де

$$y, e \in \mathbb{R}^N, X \in M_{N,I+1}(\mathbb{R}), \alpha \in \mathbb{R}^{I+1}, N = \sum_{i=1}^{I} N_i,$$

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1N_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2N_2} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IN_I} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \\ 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \end{pmatrix}, \quad \alpha = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix}, \quad e = \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1N_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2N_2} \\ \vdots \\ e_{I1} \\ e_{I2} \\ \vdots \\ e_{IN_I} \end{pmatrix}.$$

Чи можемо безпосередньо скористатися методом найменших квадратів для визначення єдиної оцінки вектора невідомих параметрів α цієї моделі?

Ні, бо rank(X)=I.

Проте, враховуючи сутність μ, можна стверджувати, що додатково справедливо

$$\exists\{w_i\}_{i=1}^{I} : \forall i \;\; w_i > 0, \; \sum_{i=1}^{I} w_i = 1, \quad \sum_{i=1}^{I} w_i \alpha_i = 0. \tag{3}$$

Приклад. Нехай змінна η – врожайність зернової культури, змінна ζ – сорт зернової культури, причому загальна кількість сортів зернової культури дорівнює I, s_{ik} $(s_{ik} > 0)$ – площа k-го поля, яке засіяне i-м сортом зернової культури, $i = \overline{1,I}$, $k = \overline{1,N_i}$ $(N_i \geq 1)$. За μ візьмемо середню врожайність за усіма сортами зернової культури. Потрібно знайти вагові коефіцієнти $\{w_i\}_{i=1}^{I}$ для обмеження (3). (На с/р)

Перейдемо до **обчислення оцінки** $\hat{\alpha}$ **вектора невідомих параметрів** $\alpha = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_I)^T$ за допомогою МНК у моделі (2) при лінійних обмеженнях (3) за доступними спостереженнями y_{ik}, $i = \overline{1,I}$, $k = \overline{1,N_i}$ $(N_i \geq 1)$.

Крім цього представляє практичний інтерес з'ясування питання про відсутність відмінностей у впливах градацій ζ на залежну змінну η, тобто перевірка гіпотези

$$H_0 : a_1 = a_2 = \ldots = a_I, \qquad \gamma > 0,$$

яка еквівалентна *гіпотезі перевірки на значимість* (на значиме відхилення від нуля) *значень* $\{\alpha_i\}_{i=1}^{I}$, а саме:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0. \tag{4}$$

Перевірку гіпотези H_0 будемо здійснювати з рівнем значущості $\gamma > 0$.

Гіпотезу (4) можна також представити у такому вигляді:

$$H_0 : A\alpha = \theta, \tag{5}$$

де

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \quad A \in M_{I-1,I+1}(\mathbb{R}), \quad rank(A) = I - 1.$$

Зауважимо, що не врахована в останньому представленні (5) гіпотези H_0 рівність $\alpha_I = 0$ буде справедлива завжди, якщо взяти до уваги лінійне обмеження (3).

Таким чином, задача перевірки гіпотези (4) звелася до перевірки лінійної гіпотези (5) для лінійної регресійної моделі (2) з урахуванням лінійних обмежень (3).

Згадаємо відповідну теорему з регресійного аналізу про перевірку лінійної гіпотези для лінійної регресійної моделі

$$y = X\alpha + e, \quad \text{де } y, e \in \mathbb{R}^N, X \in M_{N,p}(\mathbb{R}), \alpha \in \mathbb{R}^p.$$

Припустимо, що

- вектор похибок: $\mathcal{N}\left(\theta_N, \sigma^2 E_N\right), \sigma^2 > 0,$
- $rank(X) = p.$

Позначимо $\mathcal{L} = \{\alpha : A\alpha = b, rank(A) = q\}, \quad A \in M_{q,p}(\mathbb{R}),$

$$Q(\alpha) = \|y - X\alpha\|^2, \qquad \hat{\alpha} = \arg\min_{\alpha} Q(\alpha) = \left(X^T X\right)^{-1} X^T y,$$

$$\hat{\alpha}_{\mathcal{L}} = \arg\min_{\alpha \in \mathcal{L}} Q(\alpha) = \hat{\alpha} - \left(X^T X\right)^{-1} A^T \left[A\left(X^T X\right)^{-1} A^T \right]^{-1} \left[A\hat{\alpha} - b\right].$$

Тоді область прийняття гіпотези $H_0 : A\alpha = b, \quad rank(A) = q\ (q < p),\ \gamma > 0$

має вигляд: $$F = \frac{\left[Q(\hat{\alpha}_{\mathcal{L}}) - Q(\hat{\alpha})\right] / q}{Q(\hat{\alpha}) / (N - p)} < F_{\gamma}(q, N - p).$$

У нашому випадку область прийняття гіпотези (5) набуває виду:

$$F = \frac{\left[Q(\hat{\alpha}_{\mathcal{L}}) - Q(\hat{\alpha})\right]/(I-1)}{Q(\hat{\alpha})/(N-I)} < F_{\gamma}(I-1, N-I),$$

де $Q(\alpha) = \|y - X\alpha\|^2$, $\hat{\alpha} = \arg\min_{\alpha} Q(\alpha)$,

$\mathcal{L} = \{\alpha : A\alpha = \theta,\ rank(A) = I-1\}$, $\hat{\alpha}_{\mathcal{L}} = \arg\min_{\alpha \in \mathcal{L}} Q(\alpha)$,

$F_{\gamma}(\upsilon_1, \upsilon_2) - 100\gamma$ відсоткова точка F-розподілу з параметрами υ_1 та υ_2.

Спочатку визначимо $Q(\hat{\alpha})$, а потім $Q(\hat{\alpha}_{\mathcal{L}}) - Q(\hat{\alpha})$. Очевидно, що оцінка $\hat{\alpha}$ є розв'язком такої системи нормальних рівнянь

$$X^T X\hat{\alpha} = X^T y,$$

до якої потрібно додати обмеження (3).

Останню систему можна переписати у вигляді

$$\begin{pmatrix} N & N_1 & N_2 & \cdots & N_{I-1} & N_I \\ N_1 & N_1 & 0 & \cdots & 0 & 0 \\ N_2 & 0 & N_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ N_{I-1} & 0 & 0 & \cdots & N_{I-1} & 0 \\ N_I & 0 & 0 & \cdots & 0 & N_I \end{pmatrix} \hat{\alpha} = \begin{pmatrix} \sum_{i=1}^{I}\sum_{k=1}^{N_i} y_{ik} \\ N_1\overline{y}_{1\cdot} \\ N_2\overline{y}_{2\cdot} \\ \vdots \\ N_{I-1}\overline{y}_{(I-1)\cdot} \\ N_I\overline{y}_{I\cdot} \end{pmatrix}, \qquad (6)$$

де $\hat{\alpha} = \begin{pmatrix} \hat{\mu} & \hat{\alpha}_1 & \hat{\alpha}_2 & \cdots & \hat{\alpha}_I \end{pmatrix}^T$, $\overline{y}_{i\cdot} = \frac{1}{N_i}\sum_{k=1}^{N_i} y_{ik}$, $i = \overline{1, I}$.

З останніх I рівнянь системи (6) випливає, що

$$N_i\hat{\mu} + N_i\hat{\alpha}_i = N_i\overline{y}_{i\cdot}, i = \overline{1, I},$$

$$\Rightarrow \quad \hat{\mu} + \hat{\alpha}_i = \overline{y}_{i\cdot}, \quad i = \overline{1, I}.$$

Тобто оцінка $\hat{a}_i$ абсолютного впливу $i-$ї градації змінної ζ на η має вигляд

$$\hat{a}_i = \overline{y}_{i\cdot}, \quad i = \overline{1,I},$$

а оцінка відносного впливу відповідно

$$\hat{\alpha}_i = \overline{y}_{i\cdot} - \hat{\mu}, \quad i = \overline{1,I}. \tag{7}$$

Оскільки перше рівняння системи є сумою всіх наступних, то замість нього використаємо обмеження (3):

$$\sum_{i=1}^{I} w_i \hat{\alpha}_i = 0, \quad \sum_{i=1}^{I} w_i = 1, \ w_i > 0, i = \overline{1,I}.$$

Врахування (7) дозволяє стверджувати, що

$$0 = \sum_{i=1}^{I} w_i \hat{\alpha}_i = \sum_{i=1}^{I} w_i \left(\overline{y}_{i\cdot} - \hat{\mu}\right) = \sum_{i=1}^{I} w_i \overline{y}_{i\cdot} - \hat{\mu} \sum_{i=1}^{I} w_i = \sum_{i=1}^{I} w_i \overline{y}_{i\cdot} - \hat{\mu}.$$

Тобто

$$\hat{\mu} = \sum_{i=1}^{I} w_i \overline{y}_{i\cdot}. \tag{8}$$

А це, у свою чергу, дозволяє переписати (7) таким чином:

$$\hat{\alpha}_i = \overline{y}_{i\cdot} - \sum_{i=1}^{I} w_i \overline{y}_{i\cdot}, \quad i = \overline{1,I}. \tag{9}$$

У підсумку оцінки МНК вектора α об'єкта (2)–(3) набули вигляду (8)-(9). А відповідне значення функціоналу якості буде дорівнювати

$$Q(\hat{\alpha}) = \left\| y - X\hat{\alpha} \right\|^2 = \sum_{i=1}^{I} \sum_{k=1}^{N_i} \left[y_{ik} - \left(\hat{\mu} + \hat{\alpha}_i\right) \right]^2 = \sum_{i=1}^{I} \sum_{k=1}^{N_i} \left[y_{ik} - \overline{y}_{i\cdot} \right]^2. \tag{10}$$

Знайдемо тепер значення $\left(Q(\hat{\alpha}_{\mathcal{L}}) - Q(\hat{\alpha})\right)$, але для цього потрібно знайти оцінку $\hat{\alpha}_{\mathcal{L}}$.

Урахування лінійних обмежень $\mathcal{L}$ у моделі (2) приводить її до представлення

$$\text{I} \quad y = \tilde{X}\mu + \tilde{e},$$

де $\tilde{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$, $\tilde{e}$ – відповідний вектор похибок математичної моделі.

Залишається тільки знайти оцінку $\hat{\mu}_{\mathcal{L}}$, яка є розв'язком системи нормальних рівнянь

$$\tilde{X}^T \tilde{X} \hat{\mu}_{\mathcal{L}} = \tilde{X}^T y,$$

який визначається таким чином:

$$\hat{\mu}_{\mathcal{L}} = \frac{1}{N} \sum_{i=1}^{I} \sum_{k=1}^{N_i} y_{ik} = \overline{y}.$$

Проте згідно наслідку до теореми про перевірку лінійної гіпотези для лінійної регресійної моделі справедливо

$$Q(\hat{\alpha}_{\mathcal{L}}) - Q(\hat{\alpha}) = \left\| X\hat{\alpha} - X\hat{\alpha}_{\mathcal{L}} \right\|^2 = \left\| X\hat{\alpha} - \tilde{X}\hat{\mu}_{\mathcal{L}} \right\|^2 =$$

$$= \sum_{i=1}^{I} \sum_{k=1}^{N_i} (\overline{y}_{i\cdot} - \overline{y})^2 = \sum_{i=1}^{I} N_i (\overline{y}_{i\cdot} - \overline{y})^2.$$

Тоді врахування (10) та останнього результату дозволяє записати область прийняття гіпотези H_0 у вигляді

$$F = \frac{\left[\sum_{i=1}^{I} N_i (\overline{y}_{i\cdot} - \overline{y})^2 \right] \Big/ (I-1)}{\left[\sum_{i=1}^{I} \sum_{k=1}^{N_i} (y_{ik} - \overline{y}_{i\cdot})^2 \right] \Big/ (N-I)} < F_{\gamma}(I-1, N-I). \qquad (11)$$

Таблиця однофакторного дисперсійного аналізу

Проаналізуємо *повну суму квадратів відхилень спостережень* y_{ik} *від загального середнього* $\overline{y}$. Дійсно,

$$\sum_{i=1}^{I} \sum_{k=1}^{N_i} (y_{ik} - \overline{y})^2 = \sum_{i=1}^{I} \sum_{k=1}^{N_i} \left[(y_{ik} - \overline{y}_{i\cdot}) + (\overline{y}_{i\cdot} - \overline{y}) \right]^2 =$$

$$= \sum_{i=1}^{I} \sum_{k=1}^{N_i} (y_{ik} - \overline{y}_{i\cdot})^2 + \sum_{i=1}^{I} N_i (\overline{y}_{i\cdot} - \overline{y})^2 \text{ (на с/р)}$$

Скорочено останній результат можна записати таким чином:

$$S = S_e + S_A, \qquad (12)$$

де

$$S = \sum_{i=1}^{I}\sum_{k=1}^{N_i}\left(y_{ik} - \overline{y}\right)^2, S_e = \sum_{i=1}^{I}\sum_{k=1}^{N_i}\left(y_{ik} - \overline{y}_{i\cdot}\right)^2, S_A = \sum_{i=1}^{I} N_i\left(\overline{y}_{i\cdot} - \overline{y}\right)^2.$$

Тобто отримали, що S, *повна сума квадратів відхилень спостережень* y_{ik} *від загального середнього* $\overline{y}$, дорівнює S_e, *сумі квадратів відхилень спостережень* y_{ik} *від середнього відповідної градації* $\overline{y}_{i\cdot}$, плюс S_A, *сума квадратів відхилень середніх за градаціями* $\overline{y}_{i\cdot}$ *від загального середнього* $\overline{y}$. Суму S_e називають ще *залишковою сумою квадратів.*

Результати однофакторного дисперсійного аналізу заносять у нижченаведену *таблицю однофакторного дисперсійного аналізу* (табл. 1.1).

Таблиця 1.1

Таблиця однофакторного дисперсійного аналізу

Джерело варіації	**Сума квадратів**	**Кількість ступенів свободи**	**Середня сума квадратів**	***F*-статистика**	γ_{max}
між градаціями	S_A	$I-1$	$\overline{S}_A = \dfrac{S_A}{I-1}$	$F = \dfrac{\overline{S}_A}{\overline{S}_e}$	γ_*
усередині градацій	$S_¶$	$N-I$	$\overline{S}_e = \dfrac{S_e}{N-I}$		
	S	$N-1$			

В останньому рядку табл. 1.1 підраховані суми за другим та третім стовпчиками, відповідно. Причому результат у другому стовпчику збігається з отриманим результатом (12).

Також у таблиці наведено:
значення статистики F згідно з (11), яка використовується для перевірки гіпотези (4),
та γ_* – значення максимального рівня значущості γ, при якому гіпотеза (4) буде справедлива.

Аналіз контрастів

Нехай у результаті перевірки гіпотези (4):

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0,$$

з деяким рівнем значущості $\gamma > 0$, виявилося, що вона не є справедливою. Тоді виникають запитання: з'ясувати, наскільки істотно абсолютні впливи $\{a_i\}_{i=1}^{I}$ градацій змінної ζ на η або їх середні за підмножинами відхиляються одне від одного; виявити підмножини градацій ζ однакового впливу на η зі статистичної точки зору, тобто *підмножини градацій ζ гомогенного впливу на η*. Інакше кажучи необхідно провести аналіз статистик такого виду:

$$a_i - a_j,\ a_5 - \frac{a_7 + a_8}{2},\ \frac{a_1 + a_2}{2} - \frac{a_3 + a_4 + a_5}{3},\ \ldots\ ,\text{ і т.п.}$$

Іншими словами, представляють інтерес лінійні комбінації абсолютних впливів $\{a_i\}_{i=1}^{I}$, у яких сума коефіцієнтів дорівнює нулю.

Означення. *Контрастом* змінної ζ щодо η називається статистика вигляду

$$\sum_{i=1}^{I} c_i a_i, \text{ у якої } \quad \sum_{i=1}^{I} c_i = 0,$$

де a_i – кількісний вираз абсолютного впливу i-ї градації змінної ζ на η, $i = \overline{1, I}$.

При розв'язанні вищепоставлених задач стане у нагоді проведення *перевірки на значимість контрасту*, а саме, перевірки гіпотези

$$H_0 : \sum_{i=1}^{I} c_i a_i = 0, \qquad (13)$$

з деяким рівнем значущості $\gamma > 0$ $\left(\sum_{i=1}^{I} c_i = 0\right)$.

Перевірка гіпотези (13) здійснюється у два кроки:

- будується довірчий інтервал для контрасту $\sum_{i=1}^{I} c_i a_i$ з рівнем довіри $(1-\gamma), \gamma > 0$;
- якщо нуль належить побудованому довірчому інтервалу для контрасту $\sum_{i=1}^{I} c_i a_i$, то гіпотезу (13) приймають, тобто вважають контраст незначущим із статистичної точки зору з обраним рівнем значущості $\gamma > 0$, інакше його вважають таким, що істотно відхиляється від нуля.

Розглянемо деякі підходи до побудови необхідних довірчих інтервалів для контрастів $\sum_{i=1}^{I} c_i a_i$ з рівнем довіри $(1-\gamma), \gamma > 0$, $\left(\sum_{i=1}^{I} c_i = 0\right)$.

Скористаємося раніше введеним позначенням:

$$\overline{S}_e = \frac{S_e}{N-I} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{N_i}\left(y_{ik} - \overline{y}_{i\cdot}\right)^2}{N-I}.$$

Спочатку звернемося до варіанта побудови довірчого інтервалу, який запропонував Генрі Шеффе (Henry Scheffe).

I. *Метод Шеффе.* Згідно із цим підходом довірчий інтервал для контрасту $\sum_{i=1}^{I} c_i a_i$ з рівнем довіри $(1-\gamma), \gamma > 0$ задається такою нерівністю:

$$\left| \sum_{i=1}^{I} c_i a_i - \sum_{i=1}^{I} c_i \bar{y}_{i\cdot} \right| \le \sqrt{\bar{S}_e \left(\sum_{i=1}^{I} \frac{c_i^2}{N_i} \right) (I-1) F_\gamma (I-1, N-I)} = \Delta_1, \qquad (14)$$

де $F_\gamma(\upsilon_1, \upsilon_2)$ – 100γ відсоткова точка F-розподілу з υ_1 та υ_2 ступенями свободи.

Або після розкриття модуля у нерівності (14), явний вигляд границь довірчого інтервалу буде мати представлення

$$\sum_{i=1}^{I} c_i \bar{y}_{i\cdot} - \Delta_1 \le \sum_{i=1}^{I} c_i a_i \le \sum_{i=1}^{I} c_i \bar{y}_{i\cdot} + \Delta_1. \qquad (15)$$

Інший, який Джон Вайлдер Тьюкі (John Wilder Tukey) запропонував підхід для аналізу контрастів, коли $N_i = N_0, \quad i = \overline{1,I}$.

II. *Метод Тьюкі.* Він розроблений для побудови довірчого інтервалу для контрасту $\sum_{i=1}^{I} c_i a_i$ з рівнем довіри $(1-\gamma), \gamma > 0$ у припущенні, що $N_i = N_0, \quad i = \overline{1,I} \quad \left(\sum_{i=1}^{I} c_i = 0, N_0 \ge 1 \right)$. Нерівність, яка визначає цей довірчий інтервал, задається таким чином:

$$\left| \sum_{i=1}^{I} c_i a_i - \sum_{i=1}^{I} c_i \bar{y}_{i\cdot} \right| \le \frac{1}{2} \sum_{i=1}^{I} |c_i| \sqrt{\frac{\bar{S}_e}{N_0}} q_\gamma (I, N-I) = \Delta_2, \qquad (16)$$

де $q_\gamma(\upsilon_1, \upsilon_2)$ – 100γ відсоткова точка стьюдентизованого розмаху з υ_1 та υ_2 ступенями свободи.

Після розкриття модуля в (16) маємо довірчий інтервал у вигляді:

$$\sum_{i=1}^{I} c_i \bar{y}_{i\cdot} - \Delta_2 \le \sum_{i=1}^{I} c_i a_i \le \sum_{i=1}^{I} c_i \bar{y}_{i\cdot} + \Delta_2. \qquad (17)$$

Зауваження. Цей розподіл є похідним від нормального розподілу та сконструйований таким чином. Нехай

- випадкові величини η_i – нормально розподілені з параметрами 0 та 1, $i=\overline{1,n_1}$,
- випадкова величина $\chi^2(n_2)$ має χ^2-розподіл з n_2 ступенями свободи,
- випадкові величини $\{\eta_i\}_{i=1}^{n_1}, \chi^2(n_2)$ – незалежні,

тоді випадкова величина

$$q_{n_1,n_2}=\frac{\max\limits_{i=1,n_1}\eta_i-\min\limits_{i=1,n_1}\eta_i}{\sqrt{\frac{\chi^2(n_2)}{n_2}}}$$

має *розподіл стьюдентизованого розмаху з n_1 та n_2 ступенями свободи.*

Ш. *Множинний t-метод.* Нехай M – кількість *a priori* обраних для аналізу контрастів. (Зазвичай контрасти вибирають для дослідження після експерименту.) Тоді наближена довірча область для множини цих вибраних контрастів $\left\{\sum\limits_{i=1}^{I}c_i^{(j)}a_i\right\}_{j=1}^{M}$ з рівнем довіри, не менше ніж $(1-\gamma)$, $\gamma>0$, задається системою довірчих інтервалів для кожного контрасту $\sum\limits_{i=1}^{I}c_i^{(j)}a_i$ з рівнем довіри $\left(1-\frac{\gamma}{M}\right)$, $\gamma>0$ і визначається таким чином:

$$\left|\sum_{i=1}^{I}c_i^{(j)}a_i-\sum_{i=1}^{I}c_i^{(j)}\overline{y}_{i\cdot}\right|\le\sqrt{\overline{S}_e\left(\sum_{i=1}^{I}\frac{\left(c_i^{(j)}\right)^2}{N_i}\right)}\,t_{\frac{\gamma}{2M}}(N-I)=\Delta_3^j,\ j=\overline{1,M}, \qquad (18)$$

де $t_\gamma(\nu)$ – 100γ відсоткова точка t-розподілу Стьюдента з ν ступенями свободи.

рівнем довіри $(1-\gamma), \gamma > 0$, в результаті чого отримали (для методів Шеффе та Тьюкі) відповідний довірчий інтервал виду $\left[\hat{k}-\Delta_i, \hat{k}+\Delta_i\right], \Delta_i > 0, i = \overline{1,2}$, де $\hat{k} = \sum_{i=1}^{I} c_i \overline{y}_{i\bullet}$.

Це дозволяє на другому кроці область прийняття гіпотези (13) для цих методів записати в такому вигляді:

$$\left(\hat{k}-\Delta_i\right)\left(\hat{k}+\Delta_i\right) \le 0 \Leftrightarrow \hat{k}^2 \le \Delta_i^2 \Leftrightarrow \left|\hat{k}\right| \le \Delta_i, i = 1,2,$$

тобто справедливість останньої нерівності означає, що відповідний контраст слід вважати таким, що незначимо відхиляється від нуля із

статистичної точки зору з рівнем значущості $\gamma > 0$, у протилежному випадку його треба вважати таким, що істотно відхиляється від нуля.

Остаточно область прийняття відповідної гіпотези матиме такий вигляд:

I. для методу Шеффе з урахуванням (15) одержуємо

$$\left|\sum_{i=1}^{I} c_i \overline{y}_{i\bullet}\right| \le \sqrt{\overline{S}_e \left(\sum_{i=1}^{I} \frac{c_i^2}{N_i}\right)(I-1) F_\gamma (I-1, N-I)} = \Delta_1,$$

II. для методу Тьюкі, якщо взяти до уваги (17), отримуємо

$$\left|\sum_{i=1}^{I} c_i \overline{y}_{i\bullet}\right| \le \frac{1}{2}\sum_{i=1}^{I} |c_i| \sqrt{\frac{\overline{S}_e}{N_0}}\, q_\gamma (I, N-I) = \Delta_2,$$

III. для множинного t-методу врахування (19) дозволяє записати

$$\left|\sum_{i=1}^{I} c_i^{(j)} \overline{y}_{i\bullet}\right| \le \sqrt{\overline{S}_e \left(\sum_{i=1}^{I} \frac{\left(c_i^{(j)}\right)^2}{N_i}\right)} t_{\frac{\gamma}{2M}} (N-I) = \Delta_3^j, j = \overline{1, M}.$$

η - залежна кількісна скалярна змінна,

а незалежні якісні скалярні змінні:

ζ_1 - фактор A, який набуває своїх значень з I_1 градацій,

ζ_2^{I} - фактор B, який набуває своїх значень з I_2 градацій.

Необхідно за спостереженнями над залежною змінною η при активних різних сполученнях градацій незалежних змінних ζ_1 та ζ_2 побудувати математичну модель залежності змінної η від змінних ζ_1 та ζ_2.

Фон: Приклад 2: η – врожайність зернової культури,

ζ_1 – сорт зернової культури, всього I_1 сортів,

ζ_2 – вид добрива, всього I_2 видів добрива.

Нехай при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2 доступно N_{ij} спостережень y_{ijk} над η, $i=\overline{1,I_1}$, $j=\overline{1,I_2}$, $k=\overline{1,N_{ij}}$ $\left(N_{ij}\geq 1\right)$. Тоді математичну *модель двофакторного дисперсійного аналізу* будемо шукати в такому вигляді:

$$y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk},\quad i=\overline{1,I_1},\ j=\overline{1,I_2},\ k=\overline{1,N_{ij}}\ \left(N_{ij}\geq 1\right), \quad (20)$$

де

y_{ijk} – k-те спостереження над η при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2,

μ – загальне середнє всіх спостережень у деякому розумінні,

α_i – *кількісний вираз відносного впливу i-ї градації змінної ζ_1 на η відносно* μ (або, **іншими словами**, *головний ефект i-го рівня фактора A*),

β_j – *кількісний вираз відносного впливу j-ї градації змінної ζ_2 на η відносно* μ (або, **іншими словами**, *головний ефект j-го рівня фактора B*),

γ_{ij} – *кількісний вираз відносного впливу взаємодії i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2 на η відносно* μ (або, **іншими словами**, *взаємодія i-го рівня фактора A та j-го рівня фактора B*),

e_{ijk} – похибка моделі k-го спостереження над η при активному сполученні i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2.

Причому всього спостережень доступно в кількості

$$N=\sum_{i=1}^{I_1}\sum_{j=1}^{I_2}N_{ij}.$$

У свою чергу, *кількісний вираз абсолютного впливу i-ї градації змінної ζ_1* **на η дорівнює**

$$a_i = \mu + \alpha_i, \quad i = \overline{1, I_1},$$

кількісний вираз абсолютного впливу j-ї градації змінної ζ_2 **на η має вигляд**

$$b_j = \mu + \beta_j, \quad j = \overline{1, I_2},$$

а *кількісний вираз абсолютного впливу взаємодії i-ї градації змінної ζ_1 та j-ї градації змінної ζ_2 на* **η визначається як**

$$c_{ij} = \mu + \gamma_{ij}, \quad i = \overline{1, I_1}, j = \overline{1, I_2}.$$

Припустимо, що похибки e_{ijk} моделі (20) є:

- $e_{ijk} \sim \mathcal{N}\left(0, \sigma^2\right), \sigma^2 > 0, \quad \forall i, j, k;$
- $\left\{e_{ijk}\right\}$ - незалежні.

Необхідно за доступними скалярними спостереженнями $\left\{y_{ijk},\ i = \overline{1, I_1},\ j = \overline{1, I_2},\ k = \overline{1, N_{ij}}\ \left(N_{ij} \geq 1\right)\right\}$ знайти оцінки невідомих параметрів:

$$\begin{aligned} &\mu, \\ &\alpha_1, \alpha_2, \ldots, \alpha_{I_1}, \\ &\beta_1, \beta_2, \ldots, \beta_{I_2}, \\ &\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1I_2}, \\ &\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2I_2}, \\ &\quad \ldots \\ &\gamma_{I_1 1}, \gamma_{I_1 2}, \ldots, \gamma_{I_1 I_2} \end{aligned} \tag{21}$$

математичної моделі (20). Шукаємо за допомогою МНК. Загальна кількість невідомих параметрів буде дорівнювати

$$1 + I_1 + I_2 + I_1 I_2 = \left(1 + I_1\right)\left(1 + I_2\right).$$

Далі схема розв'язання задачі двофакторного дисперсійного аналізу повністю аналогічна процедурі розв'язання задачі однофакторного дисперсійного аналізу.

Спочатку модель (20) переписується, як і раніше, а саме:

$$y = X\alpha + e,$$

причому

y – вектор-стовпчик з усіх спостережень y_{ijk},

α – вектор-стовпчик з усіх невідомих параметрів (21),

X – матриця відповідної розмірності, елементи кожного рядка якої всі дорівнюють нулю, окрім першого та трьох інших, що відповідають місцезнаходженню відповідних параметрів головних ефектів та попарної взаємодії у векторі α,

e – вектор-стовпчик з усіх похибок моделі

$\left\{e_{ijk},\ i = \overline{1, I_1},\ j = \overline{1, I_2},\ k = \overline{1, N_{ij}}\ \left(N_{ij} \geq 1\right)\right\}$.

У цьому випадку матриця X буде мати неповний ранг. Тому, щоб скористатися МНК при визначенні оцінок вектора α, необхідно врахувати додаткові лінійні обмеження, які справедливі для нього.

Дійсно, враховуючи зміст невідомих параметрів (21), можна стверджувати, що:

$$\exists \left\{v_i\right\}_{i=1}^{I_1}, \left\{w_j\right\}_{j=1}^{I_2} : \forall i\ \ v_i > 0, \forall j\ \ w_j > 0,$$

$$\begin{cases} \sum\limits_{i=1}^{I_1} v_i \alpha_i = 0, \\ \sum\limits_{j=1}^{I_2} w_j \beta_j = 0, \\ \sum\limits_{i=1}^{I_1} v_i \gamma_{ij} = 0, \quad j = \overline{1, I_2}, \\ \sum\limits_{j=1}^{I_2} w_j \gamma_{ij} = 0, \quad i = \overline{1, I_1}. \end{cases} \tag{22}$$

Визначення вагових коефіцієнтів $\{v_i\}_{i=1}^{I_1}$ та $\{w_j\}_{j=1}^{I_2}$ здійснюється відповідно до змісту конкретної постановки задачі.

Врахування лінійних обмежень (22) дозволяє однозначно визначити оцінку $\hat{\alpha}$ методом найменших квадратів у математичній моделі (20) за спостереженнями $\left\{y_{ijk}, i=\overline{1,I_1},\ j=\overline{1,I_2},\ k=\overline{1,N_{ij}}\ \left(N_{ij}\geq 1\right)\right\}$.

Окрім цього, цікавою є перевірка на значимість параметрів моделі двофакторного дисперсійного аналізу, і насамперед – перевірка з деяким рівнем значущості $\gamma > 0$ таких гіпотез:

$$H_0^A : \alpha_1 = \alpha_2 = \ldots = \alpha_{I_1} = 0, \tag{23}$$

$$H_0^B : \beta_1 = \beta_2 = \ldots = \beta_{I_2} = 0, \tag{24}$$

$$H_0^{AB} : \gamma_{ij} = 0, \quad i=\overline{1,I_1},\ j=\overline{1,I_2}. \tag{25}$$

Для розв'язання цих задач достатньо використати той самий математичний апарат, що й при розв'язанні відповідних задач в однофакторному дисперсійному аналізі. Проте формули стануть більш громіздкіми. Для їх спрощення наведемо розв'язок для випадку, коли справедливо

$$N_{ij} = N_0\ \left(N_0 \geq 1\right), \quad i=\overline{1,I_1},\ j=\overline{1,I_2}, \tag{26}$$

а відповідні вагові коефіцієнти $\{v_i\}_{i=1}^{I_1}, \{w_j\}_{j=1}^{I_2}$ усі однакові

$$v_i = \frac{1}{I_1}, \quad i=\overline{1,I_1}; \qquad w_j = \frac{1}{I_2}, \quad j=\overline{1,I_2}. \tag{27}$$

Тепер загальна кількість спостережень визначатиметься таким чином:

$$N = I_1 I_2 N_0.$$

У результаті використання методу найменших квадратів для визначення невідомих параметрів (21) у математичній моделі (20) за наявності лінійних обмежень (22) при справедливості припущень (26), (27) отримуємо такі їх оцінки:

$$\hat{\mu} = \bar{y},$$

$$\hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}, \quad i = \overline{1, I_1},$$

$$\hat{\beta}_j = \bar{y}_{\cdot j\cdot} - \bar{y}, \quad j = \overline{1, I_2},$$

$$\hat{\gamma}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}, \quad i = \overline{1, I_1}, j = \overline{1, I_2},$$

де

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} y_{ijk},$$

$$\bar{y}_{ij\cdot} = \frac{1}{N_0} \sum_{k=1}^{N_0} y_{ijk}, \quad i = \overline{1, I_1}, \; j = \overline{1, I_2},$$

$$\bar{y}_{i\cdot\cdot} = \frac{1}{I_2 N_0} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} y_{ijk}, \quad i = \overline{1, I_1},$$

$$\bar{y}_{\cdot j\cdot} = \frac{1}{I_1 N_0} \sum_{i=1}^{I_1} \sum_{k=1}^{N_0} y_{ijk}, \quad j = \overline{1, I_2}.$$

Таблиця двофакторного дисперсійного аналізу

Проведемо аналіз *повної суми квадратів відхилень спостережень* y_{ijk} *від загального середнього* $\bar{y}$. Дійсно,

$$\sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \left(y_{ijk} - \bar{y}\right)^2 = \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \Big[\left(y_{ijk} - \bar{y}_{ij\cdot}\right) +$$

$$+ \left(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}\right) + \left(\bar{y}_{i\cdot\cdot} - \bar{y}\right) + \left(\bar{y}_{\cdot j\cdot} - \bar{y}\right) \Big]^2 =$$

$$= \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{N_0} \left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2 +$$

$$+ N_0 \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \left(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}\right)^2 + \qquad \text{(на с/р)}$$

$$+ I_2 N_0 \sum_{i=1}^{I_1} \left(\bar{y}_{i\cdot\cdot} - \bar{y}\right)^2 + I_1 N_0 \sum_{j=1}^{I_2} \left(\bar{y}_{\cdot j\cdot} - \bar{y}\right)^2.$$

Останнє перетворення справедливо в силу того, що в усіх подвійних добутках квадратні дужки дорівнюють нулеві. Отриманий результат скорочено можна записати таким чином:

$$S = S_e + S_A + S_B + S_{AB}, \tag{28}$$

де

$$S = \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\sum_{k=1}^{N_0}\left(y_{ijk} - \overline{y}\right)^2, S_e = \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\sum_{k=1}^{N_0}\left(y_{ijk} - \overline{y}_{ij\cdot}\right)^2,$$

$$S_A = I_2 N_0 \sum_{i=1}^{I_1}\left(\overline{y}_{i\cdot\cdot} - \overline{y}\right)^2, S_B = I_1 N_0 \sum_{j=1}^{I_2}\left(\overline{y}_{\cdot j\cdot} - \overline{y}\right)^2,$$

$$S_{AB} = N_0 \sum_{i=1}^{I_1}\sum_{j=1}^{I_2}\left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}\right)^2.$$

Отже, для S, *повної суми квадратів відхилень спостережень* y_{ijk} *від загального середнього* $\overline{y}$, у випадку двофакторної моделі дисперсійного аналізу отримали розклад (28), аналогічний розкладу (12), отриманому для однофакторної моделі.

Результати двофакторного дисперсійного аналізу також заносять у відповідну *таблицю 2.1 двофакторного дисперсійного аналізу.*

Таблиця 2.1 двофакторного дисперсійного аналізу

Джерело варіації	**Сума квадратів**	**Кількість ступенів свободи**	**Середня сума квадратів**	***F*-статистика**	$\gamma_{\max}$
головні ефекти фактора *A*	S_A	$I_1 - 1$	$\overline{S}_A = \dfrac{S_A}{I_1 - 1}$	$F_A = \dfrac{\overline{S}_A}{\overline{S}_e}$	γ_A
головні ефекти фактора *B*	S_B	$I_2 - 1$	$\overline{S}_B = \dfrac{S_B}{I_2 - 1}$	$F_B = \dfrac{\overline{S}_B}{\overline{S}_e}$	γ_B
взаємодії факторів *A* та *B*	S_{AB}	$(I_1 - 1)(I_2 - 1)$	$\overline{S}_{AB} = \dfrac{S_{AB}}{(I_1 - 1)(I_2 - 1)}$	$F_{AB} = \dfrac{\overline{S}_{AB}}{\overline{S}_e}$	γ_{AB}
помилки	S_e	$I_1 I_2 (N_0 - 1)$	$\overline{S}_e = \dfrac{S_e}{I_1 I_2 (N_0 - 1)}$		
	S	$N - 1$			

В останньому рядку табл. 2.1, таблиці двофакторного дисперсійного аналізу, наведено суми по другому та третьому стовпчикам відповідно. Зауважимо, що результат у другому стовпчику збігається з уже отриманим результатом (28). Підраховані в таблиці значення $\gamma_A, \gamma_B, \gamma_{AB}$ – значення максимальних рівнів значущості γ, за яких гіпотези (23), (24), (25) будуть справедливі, а використання статистик F_A, F_B, F_{AB} дозволяє записати області прийняття гіпотез H_0^A, H_0^B, H_0^{AB}, відповідно:

для гіпотези H_0^A:

$$F_A < F_\gamma\left(I_1 - 1, I_1 I_2 (N_0 - 1)\right),$$

для гіпотези H_0^B:

$$F_B < F_\gamma\left(I_2 - 1, I_1 I_2 (N_0 - 1)\right),$$

для гіпотези H_0^{AB}:

$$F_{AB} < F_\gamma\left((I_1 - 1)(I_2 - 1), I_1 I_2 (N_0 - 1)\right),$$

де $F_\gamma(\upsilon_1, \upsilon_2)$ – 100γ відсоткова точка F-розподілу з параметрами υ_1 та υ_2.

Багатофакторний дисперсійний аналіз

...

Самостійна робота №6. З навчального посібника «Слабоспицький О.С. Дисперсійний аналіз даних, 2013» пропрацювати матеріал наведений у **Розділі 3.**: «Багатофакторний дисперсійний аналіз».

(пропустити 5 стор.)

[redacted]

3. БАГАТОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ

Після ознайомлення з проблемами однофакторного та двофакторного дисперсійного аналізу перейдемо до більш загальних випадків. У цьому розділі спочатку розглянемо задачу дисперсійного аналізу з трьома факторами, а потім проаналізуємо задачу багатофакторного дисперсійного аналізу. Розглянемо шляхи розв'язання подібних проблем та необхідний для цього математичний апарат.

3.1. Випадок наявності трьох факторів

Перш ніж звернутися до задачі дисперсійного аналізу в загальній постановці розглянемо випадок, коли на залежну кількісну змінну η можуть впливати три незалежні якісні скалярні змінні ζ_1, ζ_2 та ζ_3, які набувають своїх значень з I_1, I_2 та I_3 градацій, відповідно. До змінних ζ_1, ζ_2 та ζ_3 також будемо звертатися як до факторів A, B та C, відповідно.

Необхідно за спостереженнями над залежною змінною η, коли активні різні комбінації градацій незалежних змінних $\zeta_1, \zeta_2, \zeta_3$, сконструювати математичну модель залежності змінної η від змінних $\zeta_1, \zeta_2, \zeta_3$.

Математичну *модель трифакторного дисперсійного аналізу* будемо шукати у вигляді

$$y_{i_1 i_2 i_3 k} = \mu + \alpha_{i_1}^{(1)} + \alpha_{i_2}^{(2)} + \alpha_{i_3}^{(3)} + \\ + \beta_{i_1 i_2}^{(1,2)} + \beta_{i_1 i_3}^{(1,3)} + \beta_{i_2 i_3}^{(2,3)} + \gamma_{i_1 i_2 i_3}^{(1,2,3)} + e_{i_1 i_2 i_3 k}, \qquad (3.1)$$
$$i_1 = \overline{1, I_1}, i_2 = \overline{1, I_2}, i_3 = \overline{1, I_3}, k = \overline{1, N_{i_1 i_2 i_3}} \left(N_{i_1 i_2 i_3} \geq 1\right),$$

де $y_{i_1 i_2 i_3 k}$ – k-те спостереження над η при активному сполученні i_1-ї градації змінної ζ_1, i_2-ї градації змінної ζ_2 та i_3-ї градації змінної ζ_3, μ – загальне середнє всіх спостережень над η, $\alpha_{i_1}^{(1)}, \alpha_{i_2}^{(2)}, \alpha_{i_3}^{(3)}$ – *кількісний вираз відносного впливу* i_j*-ї градації змінної* ζ_j *на* η *відносно* μ ($j = \overline{1,3}$) (або, іншими словами, *головний ефект* i_1*-го рівня фактора* A, i_2*-го рівня фактора* B, i_3*-го рівня фактора* C, *відповідно*), $\beta_{i_1 i_2}^{(1,2)}, \beta_{i_1 i_3}^{(1,3)}, \beta_{i_2 i_3}^{(2,3)}$ – *кількісний вираз відносного впливу взаємодії* i_1*-ї градації змінної* ζ_1 *та* i_2*-ї градації змінної* ζ_2, i_1*-ї градації змінної* ζ_1 *та* i_3*-ї градації змінної* ζ_3, i_2*-ї градації змінної* ζ_2 *та* i_3*-ї градації змінної* ζ_3 *на* η *відносно* μ, *відповідно* (або, іншими словами, це відповідно попарні взаємодії для i_1*-го рівня фактора* A *та* i_2*-го рівня фактора* B, i_1*-го рівня фактора* A *та* i_3*-го рівня фактора* C, i_2*-го рівня фактора* B *та* i_3*-го рівня фактора* C), $\gamma_{i_1 i_2 i_3}^{(1,2,3)}$ – *кількісний вираз відносного впливу взаємодії* i_1*-ї градації змінної* ζ_1, i_2*-ї градації змінної* ζ_2 *та* i_3*-ї градації змінної* ζ_3 *на* η *відносно* μ (або, іншими словами, взаємодія трійки: i_1-ї градації змінної ζ_1, i_2-ї градації змінної ζ_2 та i_3-ї градації змінної ζ_3), $e_{i_1 i_2 i_3 k}$ – помилка моделі k-го спостереження над η при активному сполученні i_1-ї градації змінної ζ_1, i_2-ї градації змінної ζ_2 та i_3-ї градації змінної ζ_3.

Загальна кількість доступних спостережень у цьому випадку обчислюється таким чином:

$$N = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} N_{i_1 i_2 i_3}.$$

Для обчислення кількісного виразу деякого абсолютного впливу достатньо до відповідного кількісного виразу відносного впливу додати μ.

Припустимо, що в моделі (3.1) помилки $e_{i_1 i_2 i_3 k}$ є:

- нормально розподіленими з параметрами 0 та σ^2 $\left(\sigma^2 > 0\right)$;
- незалежними.

Необхідно за відомими спостереженнями $\left\{y_{i_1 i_2 i_3 k},\ i_1 = \overline{1, I_1}, i_2 = \overline{1, I_2}, i_3 = \overline{1, I_3}, k = \overline{1, N_{i_1 i_2 i_3}} \left(N_{i_1 i_2 i_3} \geq 1\right)\right\}$ знайти методом найменших квадратів оцінки невідомих параметрів

$$\mu,$$
$$\alpha_1^{(1)}, \alpha_2^{(1)}, \ldots, \alpha_{I_1}^{(1)},$$
$$\alpha_1^{(2)}, \alpha_2^{(2)}, \ldots, \alpha_{I_2}^{(2)},$$
$$\alpha_1^{(3)}, \alpha_2^{(3)}, \ldots, \alpha_{I_3}^{(3)},$$
$$\beta_{11}^{(1,2)}, \beta_{12}^{(1,2)}, \ldots, \beta_{1I_2}^{(1,2)},$$
$$\beta_{21}^{(1,2)}, \beta_{22}^{(1,2)}, \ldots, \beta_{2I_2}^{(1,2)},$$
$$\ldots$$
$$\beta_{I_1 1}^{(1,2)}, \beta_{I_1 2}^{(1,2)}, \ldots, \beta_{I_1 I_2}^{(1,2)},$$
$$\beta_{11}^{(1,3)}, \beta_{12}^{(1,3)}, \ldots, \beta_{1I_3}^{(1,3)},$$
$$\beta_{21}^{(1,3)}, \beta_{22}^{(1,3)}, \ldots, \beta_{2I_3}^{(1,3)},$$
$$\ldots$$
$$\beta_{I_1 1}^{(1,3)}, \beta_{I_1 2}^{(1,3)}, \ldots, \beta_{I_1 I_3}^{(1,3)},$$

$$\beta_{11}^{(2,3)},\beta_{12}^{(2,3)},\ldots,\beta_{1I_3}^{(2,3)},$$

$$\beta_{21}^{(2,3)},\beta_{22}^{(2,3)},\ldots,\beta_{2I_3}^{(2,3)},$$

$$\ldots$$

$$\beta_{I_21}^{(2,3)},\beta_{I_22}^{(2,3)},\ldots,\beta_{I_2I_3}^{(2,3)},$$

$$\gamma_{111},\gamma_{112},\ldots,\gamma_{11I_3},$$

$$\gamma_{121},\gamma_{122},\ldots,\gamma_{12I_3},$$

$$\ldots$$

$$\gamma_{1I_21},\gamma_{1I_22},\ldots,\gamma_{1I_2I_3}$$

$$\gamma_{211},\gamma_{212},\ldots,\gamma_{21I_3},$$

$$\gamma_{221},\gamma_{222},\ldots,\gamma_{22I_3},$$

$$\ldots$$

$$\gamma_{2I_21},\gamma_{2I_22},\ldots,\gamma_{2I_2I_3}$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\gamma_{I_111},\gamma_{I_112},\ldots,\gamma_{I_11I_3},$$

$$\gamma_{I_121},\gamma_{I_122},\ldots,\gamma_{I_12I_3},$$

$$\ldots$$

$$\gamma_{I_1I_21},\gamma_{I_1I_22},\ldots,\gamma_{I_1I_2I_3}$$

математичної моделі (3.1).

Загальна кількість невідомих параметрів буде дорівнювати

$$1+I_1+I_2+I_3+I_1I_2+I_1I_3+I_2I_3+I_1I_2I_3=(1+I_1)(1+I_2)(1+I_3).$$

Далі схема розв'язання задачі трифакторного дисперсійного аналізу після врахування відповідних лінійних обмежень на невідомі параметри повністю аналогічна процедурі розв'язання задачі однофакторного дисперсійного аналізу.

Використання попереднього апарата також дозволяє здійснити перевірку на значимість параметрів моделі трифакторного дисперсійного аналізу.

3.2. Побудова математичної моделі багатофакторного дисперсійного аналізу

Перейдемо до розгляду загального випадку, коли на залежну скалярну кількісну змінну η можуть впливати незалежні скалярні якісні змінні $\zeta_1, \zeta_2, \ldots, \zeta_q$, які набувають своїх значень з $I_1, I_2, \ldots, I_q$ градацій, відповідно. Як і раніше, до змінних $\{\zeta_1, \zeta_2, \zeta_3, \zeta_4, \ldots\}$ можна буде звертатися як до факторів $\{A, B, C, D, \ldots\}$, відповідно.

Наша мета – за спостереженнями над залежною змінною η, коли активні різні комбінації градацій незалежних змінних $\zeta_1, \zeta_2, \ldots, \zeta_q$, побудувати математичну модель залежності змінної η від змінних $\zeta_1, \zeta_2, \ldots, \zeta_q$. *Модель багатофакторного дисперсійного аналізу* будемо шукати у такому вигляді:

$$y_{i_1 i_2 i_3 \ldots i_q k} = \mu + \alpha_{i_1}^{(1)} + \alpha_{i_2}^{(2)} + \ldots + \alpha_{i_q}^{(q)} + \sum_{j_1 < j_2} \beta_{i_{j_1} i_{j_2}}^{(j_1, j_2)} + \sum_{j_1 < j_2 < j_3} \gamma_{i_{j_1} i_{j_2} i_{j_3}}^{(j_1, j_2, j_3)} + \ldots + \omega_{i_1 i_2 i_3 \ldots i_q}^{(1,2,3,\ldots,q)} + e_{i_1 i_2 i_3 \ldots i_q k}, \qquad (3.2)$$

$$i_1 = \overline{1, I_1}, i_2 = \overline{1, I_2}, \ldots, i_q = \overline{1, I_q}, k = \overline{1, N_{i_1 i_2 i_3 \ldots i_q}} \left(N_{i_1 i_2 i_3 \ldots i_q} \geq 1 \right),$$

де $y_{i_1 i_2 i_3 \ldots i_q k}$ – k-те спостереження над η при активному сполученні i_1-ї градації змінної ζ_1, i_2-ї градації змінної ζ_2, ..., i_q-ї градації змінної ζ_q, μ – загальне середнє всіх спостережень над η, $\alpha_{i_j}^{(j)}$ – *кількісний вираз відносного впливу* i_j*-ї градації змінної* ζ_j *на* η *відносно* μ ($j = \overline{1, q}$) (або, іншими словами, *головний ефект* i_j*-го рівня фактора* ζ_j ($j = \overline{1, q}$)),

$\beta_{i_{j_1} i_{j_2}}^{(j_1, j_2)}$ – *кількісний вираз відносного впливу взаємодії* i_{j_1} *-ї градації змінної* ζ_{j_1} *та* i_{j_2} *-ї градації змінної* ζ_{j_2} *на* η *відносно* μ (або, іншими словами, це попарні взаємодії для i_{j_1} *-го рівня фактора* ζ_{j_1} *та* i_{j_2} *-го рівня фактора* ζ_{j_2}), $\gamma_{i_{j_1} i_{j_2} i_{j_3}}^{(j_1, j_2, j_3)}$ – *кількісний вираз відносного впливу взаємодії* i_{j_1} *-ї градації змінної* ζ_{j_1}, i_{j_2} *-ї градації змінної* ζ_{j_2} та i_{j_3} *-ї градації змінної* ζ_{j_3} *на* η *відносно* μ (або, іншими словами, взаємодія трійки: i_{j_1} *-ї градації змінної* ζ_{j_1}, i_{j_2} *-ї градації змінної* ζ_{j_2} та i_{j_3} *-ї градації змінної* ζ_{j_3}), $\omega_{i_1 i_2 i_3 \ldots i_q}$ – *кількісний вираз відносного впливу взаємодії* i_1 *-ї градації змінної* ζ_1, i_2 *-ї градації змінної* ζ_2, … , i_q *-ї градації змінної* ζ_q *на* η *відносно* μ, $e_{i_1 i_2 i_3 \ldots i_q k}$ – помилка моделі k -го спостереження над η при активному сполученні i_1 -ї градації змінної ζ_1, i_2 -ї градації змінної ζ_2, … , i_q -ї градації змінної ζ_q.

Загальна кількість доступних спостережень у багатофакторному випадку дорівнює

$$N = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_q=1}^{I_q} N_{i_1 i_2 i_3 \ldots i_q} .$$

Як і раніше, для обчислення кількісного виразу деякого абсолютного впливу достатньо до відповідного кількісного виразу відносного впливу додати μ.

Щодо помилок $e_{i_1 i_2 i_3 \ldots i_q k}$ моделі (3.2) припускаємо, що вони є:

- нормально розподіленими з параметрами 0 та σ^2 $\left(\sigma^2 > 0\right)$;
- незалежними.

Необхідно за доступними спостереженнями

$$\left\{ y_{i_1 i_2 i_3 \ldots i_q k}, i_1 = \overline{1, I_1}, i_2 = \overline{1, I_2}, \ldots, i_q = \overline{1, I_q}, k = \overline{1, N_{i_1 i_2 i_3 \ldots i_q}} \left(N_{i_1 i_2 i_3 \ldots i_q} \geq 1 \right) \right\}$$

знайти методом найменших квадратів оцінки невідомих параметрів:

$$\mu, \left\{\alpha_{i_j}^{(j)}\right\}_{j=1}^{q}, \left\{\beta_{i_{j_1} i_{j_2}}^{(j_1, j_2)}\right\}_{\left(i_{j_1} i_{j_2}\right)}, \left\{\gamma_{i_{j_1} i_{j_2} i_{j_3}}^{(j_1, j_2, j_3)}\right\}_{\left(i_{j_1} i_{j_2} i_{j_3}\right)}, \ldots, \omega_{i_1 i_2 i_3 \ldots i_q}$$

математичної моделі (3.2).

Всього невідомих параметрів буде

$$\prod_{i=1}^{q}(1+I_i).$$

Далі порядок розв'язання задачі багатофакторного дисперсійного аналізу, після врахування відповідних лінійних обмежень на невідомі параметри, повністю аналогічний порядку розв'язання задачі однофакторного дисперсійного аналізу. Представляє також інтерес перевірка на значимість параметрів моделі багатофакторного дисперсійного аналізу, яка здійснюється вже розглянутим раніше шляхом.