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 [redacted]

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:

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

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

Модель однофакторного дисперсійного аналізу шукаємо у вигляді:

$$y_{ik}=\mu+\alpha_i+e_{ik}, \quad i=\overline{1,I},\ k=\overline{1,N_i}\ \left(N_i \ge 1\right), \tag{1}$$

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 $e_{ik}, i=\overline{1,I},\ k=\overline{1,N_i}\ \left(N_i \ge 1\right)$ 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 $\left\{y_{ik}, i=\overline{1,I},\ k=\overline{1,N_i}\ \left(N_i \ge 1\right)\right\}$ 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

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

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

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

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

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:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0. \qquad (4)$$

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

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

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

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

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

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

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

$F_\gamma(\upsilon_1, \upsilon_2)$ – 100γ 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

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

де $\hat{\alpha} = (\hat{\mu} \quad \hat{\alpha}_1 \quad \hat{\alpha}_2 \quad \cdots \quad \hat{\alpha}_I)^T, \quad \overline{y}_{i\bullet} = \frac{1}{N_i}\sum\limits_{k=1}^{N_i} y_{ik},\ i = \overline{1,I}.$

З останніх I рівнянь системи (6) випливає, що

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

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

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

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

а оцінка відносного впливу відповідно

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

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

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

Тобто

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

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

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

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

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

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

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

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

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

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

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

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

Таблиця однофакторного дисперсійного аналізу

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

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

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

Скорочено останній результат можна записати таким чином:

$$S = S_e + S_A, \tag{12}$$

де

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

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* $\overline{y}_{i\bullet}$, плюс S_A, *сума квадратів відхилень середніх за градаціями* $\overline{y}_{i\bullet}$ *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**	$\gamma_{\max}$
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_e	$N-I$	$\overline{S}_e = \dfrac{S_e}{N-I}$		
	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

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

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:

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

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

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

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 $N_i = N_0, \quad i = \overline{1, I}$ $\left(\sum_{i=1}^{I} c_i = 0, N_0 \ge 1 \right)$. Нерівність, яка визначає цей довірчий інтервал, задається таким чином:

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

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:

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

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

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

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