Однофакторний та двофакторний дисперсійний аналіз

Підготував:

Студент групи СНм-51

Пімєнов А. В.

Структура моделі впливу одного та двох факторів

$$x_{ij} = \bar{x} + \alpha_j + \varepsilon_{ij}$$

де x_{ij} — значення ознаки X, одержане при i-му експерименті на j-му рівні фактора.

При цьому $M(\varepsilon_{ij}) = 0$ і ε_{ij}, як випадкові величини мають закон розподілу ймовірностей $N(0;\sigma^2)$ і між собою незалежні (K_{ij} =0).

$$x_{ij} = \bar{x} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

де x_{ijk} – значення ознаки X в i-му експерименті на j-му рівні впливу фактора A і на k-му рівні впливу фактора B;

$\bar{x}$ — загальна середня величина ознаки X;

α_i — ефект впливу фактора A на i-му рівні,

β_j — ефект впливу фактора B на j-му рівні;

γ_{ij} — ефект одночасного впливу факторів A і B;

ε_{ijk} — випадкова компонента.

Однофакторний дисперсійний аналіз

$$Y = \begin{bmatrix} y_{11} & \dots & y_{1p} \\ \dots & y_{ij} & \dots \\ y_{n1} & \dots & y_{np} \end{bmatrix}$$

Кожен стовпець (градацію фактора) треба розглядати як вибірку нормально розподілених випадкових величин $\xi_1, \xi_2, \dots, \xi_p$ з параметрами $M(\xi_j) = \mu_j$, $D(\xi_j) = \sigma^2$ для всіх j=1,…,p (дисперсії однакові).

Отже, для кожної градації фактора (стовпця таблиці даних) маємо фіксоване середнє значення, що є сталим у межах експерименту.

Гіпотезу для перевірки сформулюємо так:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p = \mu$$

Отже, дисперсія випадкової величини y_{ij}, зумовлена дією фактора на всіх рівнях, σ_R^2 = 0, і вся мінливість буде спричинена неврахованими факторами:

$$\sigma_T^2 = \sigma_D^2 \text{ або } D(y_{ij}) = \sigma_A^2 + \sigma_D^2$$

Схема обчислень для однофакторного дисперсійного аналізу

А. Обчислюємо генеральне середнє $\bar{y}$ і вибіркові середні $\bar{y}_i$

$$\bar{y} = \frac{1}{N}\sum_{i=i}^{n}\sum_{j=1}^{p} y_{ij}, N = np$$

для рівномірного однофакторного аналізу або

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{p}\sum_{i=i}^{n_j} y_{ij}, N = \sum_{j=1}^{p} n_j$$

для нерівномірного однофакторного аналізу.

Б. Знаходимо суми квадратів відхилень від відповідних середніх значень:

- сума, що характеризує мінливість, зумовлену досліджуваним фактором (факторна сума),

$$SS_R = n\sum_{j=1}^{p}(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2$$

- сума, що характеризує мінливість у межах кожної градації фактором (залишкова сума),

$$SS_D = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{i\,j} - \bar{y}_i)^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_i}(y_{i\,j} - \bar{y}_i)^2$$

- сума, що характеризує загальну мінливість (загальна або тотальна сума),

$$SS_T = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{i\,j} - \bar{y})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_i}(y_{i\,j} - \bar{y})^2$$

Справджується рівність

$$SS_T = SS_R + SS_D$$

В. Визначаємо оцінки дисперсій:

$$s_T^2 = \frac{SS_T}{N-1}; \quad s_R^2 = \frac{SS_R}{p-1}; \quad s_D^2 = \frac{SS_D}{N-p}$$

Г. Критерій Фішера для перевірки гіпотези H_0 має вигляд:

$$F = \frac{s_R^2}{s_D^2} = \frac{SS_R}{SS_D}\frac{(N-p)}{(p-1)}$$

Д. Якщо гіпотеза H_0 правильна, то F має розподіл Фішера зі ступенями вільності *df 1 = p – 1, df 2 = N – p.*
Для заданого рівня значущості α знаходимо критичні значення статистики F(α; df 1; df 2).

Обчислені значення записуємо у вигляді таблиці (табл. 1.1), (ANOVA).

Таблиця 2.1 – Результати однофакторного дисперсійного аналізу

Різновид дисперсії	Сума квадратів відхилень	Кількість ступенів вільності	Середній квадрат (оцінка дисперсії)	F-критерій
Факторна (між вибірками)	SS_R	p-1	MS_A	MS_A MS_D
Залишкова (у вибірці)	SS_D	N-p	MS_D	
Загальна	SS_T	N-1		

Схема обчислень для двофакторного дисперсійного аналізу

А. Знаходимо вибіркові середні (генеральне середнє $\bar{y}$, а також середнє в рядку y_i^r, стовпці y_j^c й клітинці $\bar{y}_{ij}$):

$$\bar{y} = \frac{1}{npq}\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n} y_{ijm};$$

$$y_i^r = \frac{1}{np}\sum_{j=1}^{p}\sum_{m=1}^{v} y_{ijm};$$

$$y_j^c = \frac{1}{nq}\sum_{i=1}^{q}\sum_{m=1}^{v} y_{ijm};$$

$$\bar{y}_{ij} = \frac{1}{n}\sum_{m=1}^{v} y_{ijm}.$$

Б. Обчислюємо суми квадратів відхилень від відповідних середніх:

- мінливість, зумовлену фактором *A*,

$$SS_A = np\sum_{i=1}^{q}(\bar{y}_i^r - \bar{y})^2$$

- мінливість, зумовлену фактором *B*,

$$SS_B = nq\sum_{j=1}^{p}(\bar{y}_j^c - \bar{y})^2$$

- мінливість, зумовлену взаємодією факторів A і B,

$$SS_{AB} = n\sum_{i=1}^{q}\sum_{j=1}^{p}(\bar{y}_{ij} - \bar{y}_i^r - \bar{y}_j^c + \bar{y})^2$$

- мінливість у межах кожної з клітинок

$$SS_D = \sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n}(y_{ijm} - \bar{y}_{ij})^2$$

- загальну мінливість спостережуваної ознаки (параметра)

$$SS_T = \sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n}(y_{ijm} - \bar{y})^2$$

Справджується рівність

$$SS_T = SS_A + SS_B + SS_{AB} + SS_D$$

B. Знаходимо оцінки дисперсій (середні квадратів відхилень)

$$MS_A = \frac{SS_A}{(q-1)}; MS_B = \frac{SS_B}{(p-1)};$$

$$MS_{AB} = \frac{SS_{AB}}{(q-1)(p-1)};$$

$$MS_D = \frac{SS_D}{\frac{pq}{(n-1)}}; MS_T = \frac{SS_T}{(N-1)}, N = npq.$$

Результати двофакторного дисперсійного аналізу записують у таблицю (табл. 1.2)

Різновид дисперсії	Сума квадратів відхилень	Кількість ступенів вільності	Середній квадрат (оцінка дисперсії)	F –критерій
Факторна для фактора A	SS_A	p-1	MS_A	MS_A/MS_D
Факторна для фактора B	SS_B	q-1	MS_B	MS_B/MS_D
Змішана для факторів AB	SS_{AB}	(p-1)(q-1)	MS_{AB}	MS_{AB}/MS_D
Залишкова	SS_D	pq(n-1)	MS_D	
Загальна	SS_T	npq-1		

Перевірка гіпотез двофакторного дисперсійного аналізу

Нехай $\mu_i^r, i=1,\dots,q$ – математичні сподівання рядків табл. 1.3, а $\mu_j^c, j=1,\dots,p$ – математичні сподівання стовпців.

Тоді $\alpha_i = \mu_i^r - \mu$ – ефект *i*-ї градації фактора A; $\beta_j = \mu_j^c - \mu$ – ефект *j*-ї градації фактора

$\gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j - \mu$ – ефект *j*-ї градації фактора B в умовах *i*-ї градації фактора A;

μ_{ij} – математичне сподівання у кожній з клітинок.

Рівні фактору	B_1	...	B_j	...	B_p
A_1	$y_{111},..,y_{11n}$	...	$y_{1j1},..,y_{1jn}$	...	$y_{1p1},..,y_{1pn}$
...	...	...	...	...	...
A_i	$y_{i11},..,y_{i1n}$	...	$y_{ij1},..,y_{ijn}$	...	$y_{ip1},..,y_{ipn}$
...	...	...	...	...	...
A_q	$y_{q11},..,y_{q1n}$	...	$y_{qj1},..,y_{qjn}$	...	$y_{qp1},..,y_{qpn}$

Сформулюємо гіпотези, які стверджують, що впливи факторів *A* і *B* на всіх рівнях однакові, а взаємовплив факторів нема:

$$H_0^A : \alpha_1 = \alpha_2 = \ldots = \alpha_q;$$

$$H_0^B : \beta_1 = \beta_2 = \ldots = \beta_p;$$

$H_0^{AB} : \gamma_{ij} = 0$ для всіх *i* та j.

Критерії для перевірки цих гіпотез мають такий вигляд:

$$F^A = \frac{MS_A}{MS_D} = \frac{SS_A}{SS_D} \frac{(N-qp)}{(q-1)}$$

$$F^B = \frac{MS_B}{MS_D} = \frac{SS_B}{SS_D} \frac{(N-qp)}{(p-1)}$$

$$F^{AB} = \frac{MS_{AB}}{MS_D} = \frac{SS_{AB}}{SS_D} \frac{(N-qp)}{(p-1)(q-1)}.$$

Якщо гіпотеза $H_0 = H_0^A H_0^B H_0^{AB}$ правильна (тобто одночасно виконуються всі три підгіпотези), т

$$\frac{MS_A}{MS_D}, \quad \frac{MS_B}{MS_D} \quad \text{і} \quad \frac{MS_{AB}}{MS_D}$$

$$F^A \geq F(\alpha; q-1; N-pq)$$

$$F^B \geq F(\alpha; p-1; N-pq)$$

$$F^{AB} \geq F(\alpha; (q-1)(p-1); N-pq)$$

Дякую за увагу!

Тернопіль 2010