2. Двофакторний варіансний аналіз

Нехай дані про деяку мінливу величину поділяються на m груп за ознакою A і n груп за ознакою B. Одержимо mn класифікаційних підгруп. Припустимо, що для кожної підгрупи проводиться лише одне спостереження.

Позначимо через x_{ij} - спостереження в $i-$й групі за ознакою A, та в $j-$й групі за ознакою B. Тоді всі mn спостережень можна записати в наступній таблиці

$$\begin{array}{l} 1 \ldots\ldots\ldots\ldots j \ldots\ldots\ldots n \\ x_{11} \ldots\ldots\ldots\ldots x_{1j} \ldots\ldots\ldots x_{1n} \\ x_{i1} \ldots\ldots\ldots\ldots x_{ij} \ldots\ldots\ldots x_{in} \\ x_{m1} \ldots\ldots\ldots\ldots x_{mj} \ldots\ldots\ldots x_{mn} \end{array} \qquad (*)$$

Позначимо через $x_{i\bullet} -$ середнє $i-$ої групи за ознакою A ($i-$ рядка)

$$x_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad (i = \overline{1,m}),$$

через $x_{\bullet j}$ - середнє $j-$ої групи за ознакою B

$$x_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \quad (j = \overline{1,n}),$$

через $x_{\bullet\bullet} -$ загальне середнє всіх спостережень

$$x_{\bullet\bullet} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}.$$

Повна мінливість всіх спостережень виражається девіацією

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - x_{\bullet\bullet})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\left(x_{ij} - x_{i\bullet} - x_{\bullet j} + x_{\bullet\bullet}\right) + \left(x_{i\bullet} - x_{\bullet\bullet}\right) + \left(x_{\bullet j} - x_{\bullet\bullet}\right)\right]^2 = \qquad (3)$$

яку запишемо у вигляді

$$= n\sum_{i=1}^{m}(x_i - x_{\bullet\bullet})^2 + m\sum_{j=1}^{n}(x_{\bullet j} - x_{\bullet\bullet})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - x_{i\bullet} - x_{\bullet j} + x_{\bullet\bullet})^2$$

(мінливість між групами ознаки *A* або девіація , мінливість між групами ознаки *B* та залишкова мінливість). Тут три суми подвійних добутків дорівнюють нулю (за властивістю середнього арифметичного).

Таким чином повна мінливість розкладається на мінливість між групою *A*, мінливість між групою *B* та залишкову.

Кожна з цих дивіацій має своє число $d.f.$ (ступенів вільності)

$$mn - 1 \qquad m - 1 \qquad n - 1 \qquad mn - (m + n - 1) = (m - 1)(n - 1)$$

Сума чисел ступенів вільності справа = числу ступенів вільності зліва. Якщо тотожність (3) поділити на $(mn-1)$, то одержимо, що повна варіанса є опуклою лінійною комбінацією варіанс між групами ознак A, між групами ознаки *B* та залишковою варіансою.

Варіанси справа позначимо через $S_A^2, \quad S_B^2, \quad S_r^2$.
Тоді три варіанси:

$$S_A^2 = \frac{1}{m-1} n \sum_{i=1}^{m} (x_{i\bullet} - x_{\bullet\bullet})^2$$

$$S_B^2 = \frac{1}{n-1} \cdot m \sum_{j=1}^{n} (x_{\bullet j} - x_{\bullet\bullet})^2$$

$$S_r^2 = \frac{1}{(m-1) \quad (n-1)} \sum_{i=1}^{m} \quad \sum_{j=1}^{n} (x_{ij} - x_{i\bullet} - x_{\bullet j} + x_{\bullet\bullet})^2$$

$$\frac{1}{mn-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - x_{\bullet\bullet})^2 = \frac{m-1}{mn-1} \quad n \cdot \frac{1}{m-1} \sum_{i=1}^{m} (x_i - x_{\bullet\bullet})^2 + \frac{n-1}{mn-1} m \frac{1}{n-1} \sum_{j=1}^{n} (x_{\bullet j} - x_{\bullet\bullet})^2$$

$$+ \frac{(m-1)\,(n-1)}{mn-1} \cdot \frac{1}{m-1} \cdot \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - x_{i\bullet} - x_{j\bullet} - x_{\bullet\bullet})^2$$
.

Якщо припустити, що спостереження однорідні і взяті з нормальної генеральної сукупності, то варіанси S_A^2, S_B^2, S_r^2 є незалежними оцінками дисперсії генеральної сукупності (тобто справа кожна з варіанс незалежна від двох інших)
Звідси слідує для перевірки $H:$ однорідності можна вибрати статистику Фішера.

Формулюємо паралельно дві гіпотези:

$H_A:$ коли вплив груп ознаки A не істотний $(=0)$

Два доведення гіпотези вибираємо статистику

$$(1^*) \quad F_A = \frac{S_A^2}{S_r^2}$$

тобто гіпотезу $H_A:$ доводимо незалежно від впливу груп ознак B, а гіпотезу H_B - незалежно від впливу груп ознаки A.

$H_B:$ коли вплив груп ознаки B не істотний $(=0)$

Для доведення цієї гіпотези вибираємо статистику

$$(2^*) \quad F_B = \frac{S_B^2}{S_r^2}$$
.

Якщо гіпотеза про однорідність даних табл. $(*)$ з генеральної популяції вірна , то статистики $(1^*), \quad (2^*)$ мають розподіл Фішера відповідно з $d_\bullet f_\bullet = (m-1, (m-1)(n-1))$ та $d_\bullet f_\bullet = (n-1, (m-1)(n-1))$. Це дозволяє визначити критичні значення для обидвох гіпотез.
Обчислення при двофакторному варіансному аналізі оформляємо у вигляді таблиці при одному спостереженні в кожній підгрупі

Мінливість	Девіація	$d_\bullet f_\bullet$	Варіанса

між групами A	$n\sum_{i=1}^{m}(x_{i\bullet}-x_{\bullet\bullet})^2$	$m-1$	S_A^2
між групами B	$m\sum_{j=1}^{n}(x_{\bullet j}-x_{\bullet\bullet})^2$	$n-1$	S_B^2
Залишкова	$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-x_{i\bullet}-x_{j\bullet}+x_{\bullet\bullet})^2$	$(m-1)(n-1)$	S_r^2
Повна	$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-x_{\bullet\bullet})^2$	$mn-1$	-

Останній рядок є сумою трьох попередніх, а це служить контролем правильності обчислень

Приклад .Затрати матеріалу на виготовлення деякого виробу трьома різними технологіями (A) на 4-х різних заводах (B) були такі:

B A	1	2	3	4
1	25	20	30	25
2	30	40	40	50
3	23	18	20	27

При рівні значущості $\alpha = 0{,}10$ перевірити гіпотезу про те, що рівень затрат матеріалу на виріб не впливає ані на вибір технології ані на вибір заводу.

$H_A:$ вплив технології $A = 0$

$H_B:$ вплив заводу $B = 0$

Для перевірки H проводимо варіансний аналіз

$m = 3$ $n = 4$

$$x_{1\bullet} = \frac{1}{4}\sum_{i=1}^{4} x_i$$

$x_{1\bullet} = 25$ $x_{\bullet 1} = 26$

$x_{2\bullet} = 40$ $x_{\bullet 2} = 26$

$x_{3\bullet} = 22$ $x_{\bullet 3} = 30$

$x_{\bullet 4} = 34$

$x_{\bullet\bullet} = 29$

Результати

Мінливість	Девіація	$d.f.$	Варіанса
між технологіями A	744	2	$S_A^2 = 372$
між заводами B	132	3	$S_B^2 = 44$
Залишкова	164	6	$S_r^2 = 27{,}33$
Повна	1040	11	-

$$F_{Аетп} = \frac{S_A^2}{S_r^2} = \frac{372}{27{,}33} = 13{,}61 \qquad F_{Аеем} > F_{Акк}$$

$H_A -$ відкидаємо. Тип технології істотно впливає на рівень затрат матеріалу при виготовлені.

$$\left.\begin{array}{l} \alpha = 0{,}10 \\ d.f. = (2,\ 6) \end{array}\right| \qquad F_{Акк} = 4{,}76 \qquad F_{Веем} < F_{Акк}$$

$H_B -$ приймаємо.

Заводи не впливають істотно на рівень затрат матеріалу.