Варіансний аналіз

Варіансний аналіз (*В.А.*) – це метод статистичного дослідження впливу різних факторів на яке не-будь явище, який базується на порівнянні варіанс. Він застосовується тоді, коли вибірки можна згрупувати. Основна задача варіансного аналізу полягає в тому, щоб дослідити мінливості, викликані різними факторами. При цьому повну мінливість розкладаємо на доданки за факторами.

При *В.А.* істотно те, що вибірки беруться із нормальної популяції і статистичні доведення проводяться з допомогою критерію Фішера. Якщо генеральна сукупність (популяція) не є нормальною, то оцінки середніх арифметичних не є незалежними від варіанси і відношення варіанс при аналізі стають залежними.

В.А. може бути : однофакторний,

двофакторний і т.д.

1. Однофактоний варіансний аналіз

Нехай дано m груп (класів, рівнів) незалежних спостережень над деякою одновимірною кількістю мінливою величиною.

$$m = 2,3,\ldots\ldots$$

Позначимо через x_{ij} - $j-$е спостережень в $i-$й групі, а через n_i - обсяг $i-$ї групи. Тоді всі m груп спостережень можна записати в такій таблиці:

$$\begin{array}{c}
x_{11}\ldots\ldots\ldots\ldots x_{1j}\ldots\ldots\ldots\ldots x_{1n_{1}} \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_{i1}\ldots\ldots\ldots\ldots x_{ij}\ldots\ldots\ldots\ldots x_{in_{i}} \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_{m1}\ldots\ldots\ldots\ldots x_{mj}\ldots\ldots\ldots\ldots x_{mn_m}
\end{array}$$

Позначимо через N обсяг всіх спостережень

$$N = n_1 + \ldots\ldots\ldots + n_i + \ldots\ldots + n_m$$

через

$x_{i\bullet}$ - середнє спостереження в $i-$й групі

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

через

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

$$x_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}$$

Повна мінливість всіх спостережень виражається з допомогою дивіації

$$(1) \qquad \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - x_{\bullet\bullet})^2 =$$

Запишемо її у вигляді

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

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

Подвоєний добуток $=0$ на основі доведеного твердження: сума відхилень елементів статистичного матеріалу від середнього арифметичного $=0$.
Ця тотожність з алгебри представляє розклад квадратичної форми з ліва на суму двох квадратичних форм справа .
Таким чином повна дивіація розкладається на суму двох дивіацій:
дивіація між групами

$$\sum n_i \ (x_{i\bullet} - x_{\bullet\bullet})^2$$

і дивіація у групах

$$\sum \sum (x_{ij} - x_{i\bullet})^2$$

Кожна з цих дивіацій має своє число ступенів вільності. Повна дивіація має $(N-1)$ ступенів вільності. Дивіація між групами має $(m-1)$ ступенів вільності. Дивіація у групах має $(N-m)$ ступенів вільності. Очевидно, що між цими ступенями вільності існує тотожність

$$N-1 = (m-1) + (N-m)$$

Поділимо тотожність (1) на $(N-1)$ одержимо нову тотожність (2)

$$\underbrace{\frac{1}{N-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - x_{\bullet\bullet})^2}_{\text{повна варіанса}} = \frac{m-1}{N-1} \underbrace{\frac{1}{m-1}\sum_{i=1}^{m} n_i (x_{i\bullet} - x_{\bullet\bullet})^2}_{\text{варіанса між групами } S_1^2} + \frac{N-m}{N-1}\cdot \underbrace{\frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - x_{\bullet\bullet})^2}_{\text{варіанса у групах } S_2^2} \qquad (2)$$

$$S^2 = \frac{m-1}{N-1} S_1^2 + \frac{N-m}{N-1} S_2^2$$

Коефіцієнти при варіансах справа додатні і сума їх =1.

Таким чином тотожність (2) вказує на те, що повна варіанса є опуклою лінійною комбінацією варіанси між групами та варіанси у групах. Позначимо варіанси через S_1^2 і S_2^2 .

Припустимо, що m груп спостережень однорідні, тобто взяті у тої самої генеральної сукупності. Тоді середнє кожної групи буде оцінкою сподівання генеральної сукупності.

Якщо припустити що генеральна сукупність нормально розподілена, то варіанса між групами S_1^2 та варіанса у групах S_2^2 є незалежними і незміщеними оцінками дисперсії нормальної генеральної сукупності

Означення. Оцінка називається незміщеною, якщо її сподівання дорівнює оцінюваному параметру, тобто $E(S_1^2) = E(S_2^2) = \sigma^2$)

Тому для перевірки гіпотези однорідності можна використати критерій Фішера, основану на статистиці

$$F = \frac{S_1^2}{S_2^2}$$

Якщо при заданому рівні значущості α і даних ступенях вільності $d_{\bullet}f_{\bullet} = (m-1, N-m)$ $F_{emn} > F_{kp}$, то гіпотезу однорідності відкидаємо. В протилежному випадку – гіпотеза не суперечить емпіричним даним.

Обчислення при однофакторному варіантному аналізу оформляємо у вигляді наступної таблиці.

Мінливість	Девіація	$d_{\bullet}f_{\bullet}$	Варіанса
між групами	$\sum_{i=1}^{m} n_i\,(x_{i0} - x_{\bullet\bullet})^2$	$m-1$	S_1^2
у групах	$\sum_{i-1}^{m} \sum_{j=1}^{n_i} (x_{ij} - x_{i\bullet})^2$	$N-m$	S_2^2
повна	$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - x_{\bullet\bullet})^2$	$N-1$	-

Відмітимо, що останній рядок для девіацій та ступенів вільності є сумою двох попередніх рядків, що служить контролем правильності обчислення найтрудомісткішим буває обчислення девіації в групі.

Приклад. Тривалість життя в годинах 4-х вибіркових електроламп була такою:

Дано	час (год)	тривалості	4-х	вибірок	електричних	ламп	
1-а вибірка:	1600	1610	1650	1680	1700	1720	1800	-
2-а вибірка:	1580	1640	1640	1700	1750	-	-	-
3-а вибірка:	1450	1550	1600	1620	1640	1660	1740	1820
4-а вибірка:	1510	1520	1530	1600	1600	1800	-	-

На основі цих даних здійснити варіансний аналіз і показати, що критерій Фішера не дозволяє відкинути гіпотезу однорідності

$$x_{1\bullet} = \frac{1600 + 1610 + 1650 + 1680 + 1700 + 1720 + 1800}{7}$$

$$x_{2\bullet} = \frac{1580 + 1640 + 1620 + 1700 + 1750}{5}$$

Середнє спостережень в 1-й виборці(групі) (це сума всіх ел-в вибірки(групі) (це сума всіх ел-в статистичного матеріалу поділена на обсяг статистичного матеріалу

Тут

$n_1 = 7$

$n_2 = 5$

$n_3 = 8$ $\quad m = 4$

$n_4 = 6$ $\quad N = 26$

$x_{1\bullet} = 1680$ $\quad x_{2\bullet} = 1662$ $\quad x_{3\bullet} = 1636{,}25$ $\quad x_{4\bullet} = 1568{,}33$

$x_{\bullet\bullet} = 1637{,}3$ - (загальне середнє спостережень)

Обчислення оформляємо у вигляді таблиці (варіансного аналізу)

Мінливість	Девіація	$d_\bullet f_\bullet$	Варіанса
між групами	44360	3	$S_1^2 = 14767$
у групах	151351	22	$S_2^2 = 6880$
повна	195711	25	-

$F_{емп} = \dfrac{14767}{6880} = 2{,}14$ (емпіричне відношення варіанс)

$$\left.\begin{array}{l} \alpha = 0{,}10 \\ d_\bullet f_\bullet = (3{,}22) \end{array}\right\} \quad F_{кр} = 3{,}05$$

$F_{емп} < F_{кр}$ Гіпотезу однорідності приймаємо.