Лекція 18. Двофакторний аналіз

Двохфакторний дисперсійний аналіз (англ. Two-way analysis of variance, або two-way ANOVA) дозволяє встановити одночасний вплив двох факторів, а також взаємодію між цими факторами. При наявності більше двох факторів говорять про багатофакторний дисперсійний аналіз (англ. Multifactor ANOVA).

Розглянемо одну з реалізацій двофакторного дисперсійного аналізу, яка найчастіше застосовується на практиці.

Для проведення аналізу необхідно мати результати спостережень над досліджуваною ознакою, які представляються у вигляді таблиці спостережень.

B \ A	A_1	A_2	…	$A_к$	$\bar{y}_{*j}$
B_1	y_{11}	y_{21}	…	y_{k1}	$\bar{y}_{*1}$
B_2	y_{12}	y_{22}	…	y_{k2}	$\bar{y}_{*2}$
…	…	…	…	…	…
B_n	y_{1n}	y_{2n}	…	y_{kn}	$\bar{y}_{*n}$
$\bar{y}_{i*}$	$\bar{y}_{1*}$	$\bar{y}_{2*}$	…	$\bar{y}_{k*}$	$\bar{y}$

де y_{ij} - значення ознаки; A - фактор з k рівнями: $A_1, A_2, ..., A_k$; B - фактор з n рівнями: $B_1, B_2, ..., B_n$.

$\bar{y}_{*j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij}$ - середнє значення по рядкам.

$\bar{y}_{i*} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ - середнє значення по стовпцям.

$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} \bar{y}_{*j} + \frac{1}{k}\sum_{i=1}^{k} \bar{y}_{\mathrm{i}*} = \frac{1}{k+n}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$ - загальне середнє значення.

Очевидно, що $\bar{y}_{*i}$ є оцінкою математичного сподівання, тобто

$$m_1^B = \bar{y}_{*1};\ m_2^B = \bar{y}_{*2};\ ...\ m_n^B = \bar{y}_{*\mathrm{n}}$$
$$m_1^A = \bar{y}_{1*};\ m_2^A = \bar{y}_{2*};\ ...\ m_k^A = \bar{y}_{\mathrm{k}*}$$

Необхідно перевірити гіпотезу про рівність математичних очікувань генеральних сукупностей, тобто для фактора *А*.

$H_0^A : m_1^A = m_2^A = ... = m_k^A;$

$H_1^A : m_1^A \neq m_2^A \neq ... \neq m_k^A$

Для фактора *B*

$H_0^B : m_1^B = m_2^B = ... = m_k^B;$

$H_1^B : m_1^B \neq m_2^B \neq ... \neq m_n^B$

Ці гіпотези ми перевіряємо за відомими оцінками Безпосередня перевірка цих гіпотез не раціональна. Р. Фішер запропонував замість цих гіпотез перевіряти однорідність оцінок дисперсій а саме: розкид значень $m_i^A\left(i=1,2,...,k\right)$ можна виміряти за допомогою факторної дисперсії, яку позначають D_A розкид значень $m_j^B\left(j=1,2,...,n\right)$ можна виміряти за допомогою факторної дисперсії, яку позначають D_B. Чим сильніше розрізняються між собою середні величини, тим більше факторна дисперсія. Крім впливу факторів на вихідну величину впливають випадкові фактори (помилки досвіду). Розкид значень за рахунок випадкових причин вимірюється дисперсією, яка називається залишкової і позначається $D_{залиш}$.

Отже нульові і протилежні гіпотези для факторів *A* і *B* можна записати так:

$H_0^A : m_1^A = m_2^A = ... = m_k^A \Leftrightarrow D_A = D_{залиш}$ - фактор *A* впливає не значимо;

$H_1^A : m_1^A \neq m_2^A \neq ... \neq m_k^A \Leftrightarrow D_A > D_{залиш}$ - фактор *A* впливає значимо;

$H_0^B : m_1^B = m_2^B = ... = m_n^B \Leftrightarrow D_B = D_{залиш}$ - фактор *B* впливає не значимо;

$H_1^B : m_1^B \neq m_2^B \neq ... \neq m_n^B \Leftrightarrow D_B > D_{залиш}$ - фактор *B* впливає значимо.

Перевіримо гіпотези: $\begin{matrix} H_0^A : D_A = D_{залиш}; & H_0^B : D_B = D_{залиш}; \\ H_1^A : D_A > D_{залиш}; & H_1^B : D_B > D_{залиш}. \end{matrix}$

Однорідність дисперсій перевіряється за критерієм Фішера:

$$F_{набл}^A = \frac{D_A}{D_{залиш}}; \quad F_{набл}^B = \frac{D_B}{D_{залиш}},$$

де

$$D_A = \frac{n \cdot \sum_{i=1}^{k} \left(\bar{y}_{i*} - \bar{y} \right)^2}{k-1}; \quad D_B = \frac{k \cdot \sum_{j=1}^{n} \left(\bar{y}_{*j} - \bar{y} \right)^2}{n-1}; \quad D_{залиш} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \left(y_{ij} - \bar{y}_{i*} - \bar{y}_{*j} + \bar{y} \right)^2}{(k-1)(n-1)}.$$

Критичне значення визначається за таблицею Фішера:

$$F^A_{крит} = F\left(\alpha;\ f_1 = k-1;\ f_2 = (k-1)(n-1)\right);$$

$$F^B_{крит} = F\left(\alpha;\ f_1 = n-1;\ f_2 = (k-1)(n-1)\right).$$

Якщо $F_{набл} \le F_{крит}$, дисперсії неоднорідні, а це значить, що фактор впливає значимо.

Критерій Фішера застосовується, якщо $D_{факт} > D_{залиш}$; якщо $D_{факт} < D_{залиш}$, то без перевірки робиться висновок про незначущість впливу факторів.

Приклад 18.1.

Мають такі показники: *Y* - продуктивність праці; *A* - концентрація виробництва; *B* - тип обладнання.

Таблиця спостережень:

A \ B	B_1	B_2	$\bar{y}_{i*}$
A_1	25	30	27,5
A_2	50	60	55
A_3	80	90	85
$\bar{y}_{*j}$	51,7	60	$\bar{y} = 55{,}8$

При рівні значущості α=0,05 перевірити чи є значимим вплив факторів *A* і *B* на вихідну величину *Y*, тобто перевірити гіпотези:

$H_0^A : m_1^A = m_2^A = \ldots = m_k^A \Leftrightarrow D_A = D_{залиш}$ - фактор *A* впливає не значимо;

$H_1^A : m_1^A \ne m_2^A \ne \ldots \ne m_k^A \Leftrightarrow D_A > D_{залиш}$ - фактор *A* впливає значимо;

$H_0^B : m_1^B = m_2^B = \ldots = m_n^B \Leftrightarrow D_B = D_{залиш}$ - фактор *B* впливає не значимо;

$H_1^B : m_1^B \ne m_2^B \ne \ldots \ne m_n^B \Leftrightarrow D_B > D_{залиш}$ - фактор *B* впливає значимо.

На першому етапі обчислимо оцінки математичних сподівань:

A - фактор з n рівнями: $A_1, A_2, A_3 \quad (n=3)$.

B - фактор з k рівнями: $B_1, B_2 \quad (k=2)$.

За факторами А

$\overline{y}_{*j} = \dfrac{1}{k}\sum_{i=1}^{k} y_{ij}$ - середнє значення по рядкам:

$$\overline{y}_{*1} = \frac{25+30}{2} = 27,5; \quad \overline{y}_{*2} = \frac{50+60}{2} = 55; \quad \overline{y}_{*3} = \frac{80+90}{2} = 85.$$

За фактором В

$\overline{y}_{i*} = \dfrac{1}{n}\sum_{j=1}^{n} y_{ij}$ - середнє значення по стовпцям.

$$\overline{y}_{1*} = \frac{25+50+80}{3} = 51,7; \quad \overline{y}_{2*} = \frac{30+60+90}{3} = 60.$$

$\overline{y} = \dfrac{1}{n}\sum_{j=1}^{n}\overline{y}_{*j} = \dfrac{1}{k}\sum_{i=1}^{k}\overline{y}_{i*} = \dfrac{1}{k+n}\sum_{i=1}^{k}\sum_{j=1}^{n}\overline{y}_{ij} = \dfrac{1}{k\cdot n}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$ - загальне середнє значення:

$$\overline{y} = \frac{1}{5}\cdot(51,7+60+27,5+55+85) = 55,84.$$

Очевидно, що $\overline{y}_{*i}$ є оцінкою математичного сподівання, тобто

$m_1^B = \overline{y}_{*1}; m_2^B = \overline{y}_{*2};$

$m_1^A = \overline{y}_{1*}; m_2^A = \overline{y}_{2*};\ m_3^A = \overline{y}_{3*}.$

Обчислимо оцінки дисперсії:

$$D_A = \frac{k\cdot\sum_{j=1}^{n}\left(\overline{y}_{*\text{j}} - \overline{y}\right)^2}{n-1} = \frac{2\cdot\left[(27,5-55,8)^2 + (55-55,8)^2 + (85-55,8)^2\right]}{3-1} = 1654,2;$$

$$D_B = \frac{n\cdot\sum_{i=1}^{k}\left(\overline{y}_{i*} - \overline{y}\right)^2}{k-1} = \frac{3\left[(51,7-55,8)^2 + (60,0-55,8)^2\right.}{2-1} = 103,35;$$

$$D_{залиш}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(y_{ij}-\overline{y}_{i*}-\overline{y}_{*\mathrm{j}}+\overline{y}\right)^2}{(k-1)(n-1)}=\frac{1}{(2-1)(3-1)}\Big[\left(25-51{,}7-27{,}5+55{,}8\right)^2+$$

$$+\left(30-60-27{,}5+55{,}8\right)^2+\left(50-51{,}7-55+55{,}8\right)^2+\left(60-60-55+55{,}8\right)^2+$$

$$+\left(80-51{,}7-85+55{,}8\right)^2+\left(90-60-85+55{,}8\right)^2\Big]=4{,}175\approx 4{,}2.$$

Однорідність дисперсій перевіряється за критерієм Фішера.

Для цього обчислимо F - статистику:

$$F^{A}_{набл}=\frac{D_A}{D_{залиш}}=\frac{1654{,}2}{4{,}2}=393{,}9;\;\; F^{B}_{набл}=\frac{D_B}{D_{залиш}}=\frac{103{,}35}{4{,}2}=24{,}6\,.$$

При рівні значущості α=0,05 за таблицею критичних точок розподілу Фішера визначимо критичне значення критерію

$$F^{A}_{крит}=F\left(0{,}05;\; f_1=n-1=3-1=2;\; f_2=(k-1)(n-1)=1\cdot 2=2\right)=19;$$

$$F^{B}_{крит}=F\left(0{,}05;\; f_1=k-1=2-1=1;\; f_2=(k-1)(n-1)=2\right)=18{,}51.$$

$\left(F^{A}{}_{набл}=393{,}9\right)>\left(F^{A}{}_{крит}=19\right)\Leftrightarrow$ фактор A впливає значимо;

$\left(F^{B}{}_{набл}=24{,}6\right)>\left(F^{B}{}_{крит}=18{,}51\right)\Leftrightarrow$ фактор B впливає значимо;

Висновок: Як видно з результатів, розрахункове значення величини F-статистики для фактора A (концентрація виробництва) F=393,6, а критична область утворюється правостороннім інтервалом (19,0; +∞). Оскільки F=393,9 попадає в критичну область, гіпотеза H_0^A не береться, тобто вважається, що в цьому експерименті концентрація виробництва мала вплив на продуктивність праці.

Розрахункове значення величини F-статистики для фактора B (тип обладнання) F=24,6, а критична область утворюється правостороннім інтервалом (18,51;+∞). Оскільки F=24,6 попадає у критичну область, гіпотезу H_0^B не приймається, тобто. вважається, що в даному експерименті тип обладнання також вплинув на продуктивність праці.

Отже, на продуктивність праці в цій групі підприємств значимо впливає і концентрація виробництва і тип обладнання