Двухфакторный дисперсионный анализ

Рассмотрим случай влияния двух *факторов* на *отклик* X. В этом случае *дисперсионный анализ* основывается на результатах эксперимента, проводимого на различных *уровнях* каждого из факторов.

Будем предполагать, что взаимосвязь между факторами отсутствует*. Для простоты изложения ограничимся случаем, когда для каждой пары уровней рассматриваемых факторов проводится по одному наблюдению. Через l_A обозначим число уровней фактора A, а через l_B — число уровней фактора B. Тогда общее число наблюдений для всех возможных пар уровней факторов A и B равно $n = l_A l_B$.

Математическую модель двухфакторного дисперсионного анализа в этом случае можно представить в виде

$$X_{ij} = \mu_0 + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = \overline{1, l_A}, \quad j = \overline{1, l_B}, \tag{8.12}$$

где X_{ij} — отклик X на i-м уровне фактора A и j-м уровне фактора B; $\mu_0 = \mathbf{M}X$; α_i, β_j — неслучайные величины, характеризующие вклады в X_{ij}, обусловленные действием соответствующих факторов A и B; ε_{ij} — случайная величина, характеризующая вклад в X_{ij}, обусловленный действием неучтенных факторов.

Предположения, сделанные в **8.2** относительно случайных величин ε_{ij}, остаются в силе. При этом

$$\mathbf{M}X_{ij} = m_0 + \alpha_i + \beta_j$$

и $\alpha_1 + \ldots + \alpha_{l_A} = \beta_1 + \ldots + \beta_{l_B} = 0$, что и означает независимость факторов A и B.

Поскольку в модели (8.12) взаимодействие факторов отсутствует, *проверка гипотез* о влиянии факторов A и B на отклик X проводится отдельно для каждого фактора. Рассмотрим критерии для проверки гипотез о влиянии фактора A (фактора B) на отклик X. Введем обозначения

$$\overline{X}_{i\cdot} = \frac{1}{l_B}\sum_{j=1}^{l_B} X_{ij}, \qquad \overline{X}_{\cdot j} = \frac{1}{l_A}\sum_{i=1}^{l_A} X_{ij}, \qquad \overline{X} = \frac{1}{l_A l_B}\sum_{i=1}^{l_A}\sum_{j=1}^{l_B} X_{ij}.$$

Общая сумма квадратов отклонений X_{ij} от выборочного среднего $\overline{X}$ может быть представлена в виде

$$Q(\vec{X}_n) = \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X})^2 = l_B \sum_{i=1}^{l_A}(\overline{X}_{i\cdot} - \overline{X})^2 +$$

$$+ l_A \sum_{i=1}^{l_B}(\overline{X}_{\cdot j} - \overline{X})^2 + \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X})^2$$

(в этом можно убедиться с помощью рассуждений, аналогичных приведенным в **8.2**). Отсюда вытекает равенство

$$Q(\vec{X}_n) = Q_A(\vec{X}_n) + Q_B(\vec{X}_n) + Q_0(\vec{X}_n), \tag{8.13}$$

где слагаемое

$$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A}(\overline{X}_{i\cdot} - \overline{X})^2$$

обусловлено отличием *выборочных средних* $\overline{X}_{i\cdot}$ и $\overline{X}$, т.е. влиянием фактора A на отклик X; слагаемое

$$Q_B(\vec{X}_n) = l_A \sum_{i=1}^{l_B}(\overline{X}_{\cdot j} - \overline{X})^2$$

обусловлено отличием выборочных средних $\overline{X}_{\cdot j}$ и $\overline{X}$, т.е. влиянием фактора B на отклик X; слагаемое

$$Q_0(\vec{X}_n) = \sum_{i=1}^{l_A}\sum_{j=1}^{l_B}(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X})^2$$

учитывает влияние всех факторов, в том числе и неучтенных.

Проверка гипотез о влиянии факторов A и B на отклик X основана на сравнении *статистик* $Q_A(\vec{X}_n)$ и $Q_B(\vec{X}_n)$ с $Q_0(\vec{X}_n)$.

Проверим, например, гипотезу H_0 о том, что фактор A не влияет на отклик X, т.е. $\alpha_i = 0$, $i = \overline{1, l_A}$.

Если гипотеза H_0 верна, то при сделанных выше предположениях относительно ε_{ij}, $i = \overline{1, l_A}$, $j = \overline{1, l_B}$, статистики $Q_A(\vec{X}_n)/\sigma^2$ и $Q_0(\vec{X}_n)/\sigma^2$ независимы и имеют χ^2*-распределение* с числом степеней свободы $l_A - 1$ и $(l_A - 1)(l_B - 1)$ соответственно, а статистики

$$S_A^2(\vec{X}_n) = \frac{Q_A(\vec{X}_n)}{l_A - 1} \quad \text{и} \quad S_0^2(\vec{X}_n) = \frac{Q_0(\vec{X}_n)}{(l_A - 1)(l_B - 1)} \tag{8.14}$$

являются *несмещенными оценками* дисперсии σ^2 отклика* X. Отсюда следует (см. **Д.3.1**), что

$$F = \frac{S_A^2(\vec{X}_n)}{S_0^2(\vec{X}_n)} \sim F\big(l_A-1,\, (l_A-1)(l_B-1)\big). \qquad (8.15)$$

Гипотеза H_0 не противоречит результатам наблюдений, если выборочное значение $f_{\text{в}}$ статистики $S_A^2(\vec{X}_n)/S_0^2(\vec{X}_n)$ не превосходит $f_{\text{кр}} = f_{1-\alpha}\big(l_A-1,\, (l_A-1)(l_B-1)\big)$ для заданного *уровня значимости* α. В противном случае, т.е. если

$$f_{\text{в}} > f_{\text{кр}},$$

гипотезу H_0 отклоняют.

Если приходится отвергать гипотезу H_0, то может возникнуть необходимость в проверке одной из гипотез $H_0^{(i)}$, согласно которой влияние на отклик оказывает i-й уровень фактора A, т.е. проверяют гипотезу

$$H_0^i\colon\ \alpha_1 = \ldots = \alpha_{i-1} = \alpha_{i+1} = \ldots = \alpha_{l_A} = 0,\ \ \alpha_i \neq 0.$$

Пусть $i=1$, а

$$\overline{X}_{\cdot(2\ldots l_A)} = \frac{1}{(l_A-1)l_B}\sum_{i=2}^{l_A}\sum_{j=1}^{l_B} X_{ij}.$$

Тогда сумма квадратов $Q_A(\vec{X}_n)$ может быть представлена в виде

$$Q_A(\vec{X}_n) = Q'_A(\vec{X}_n) + Q''_A(\vec{X}_n), \qquad (8.16)$$

где

$$Q'_A(\vec{X}_n) = \frac{(l_A-1)l_B}{l_A}\big(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\big)^2,$$

$$Q''_A(\vec{X}_n) = l_B\sum_{i=2}^{l_A}\big(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\big)^2.$$

Действительно, учитывая равенства

$$\overline{X} = \frac{1}{l_A l_B}\sum_{i=1}^{l_A}\sum_{j=1}^{l_B} X_{ij} = \frac{1}{l_A l_B}\sum_{i=2}^{l_A}\sum_{j=1}^{l_B} X_{ij} + \frac{1}{l_A l_B}\sum_{j=1}^{l_B} X_{1j} =$$

$$= \frac{l_A-1}{l_A}\overline{X}_{\cdot(2\ldots l_A)} + \frac{1}{l_A}\sum_{j=1}^{L_B} X_{ij},$$

находим

$$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}\right)^2 =$$

$$= l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \frac{l_A - 1}{l_A} \overline{X}_{\cdot(2\ldots l_A)} - \frac{1}{l_A} \overline{X}_{1\cdot}\right)^2 =$$

$$= l_B \sum_{i=1}^{l_A} \left(\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) + \frac{1}{l_A}\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)\right)^2.$$

В полученной сумме преобразуем каждое слагаемое по формуле квадрата суммы. В результате находим

$$Q_A(\vec{X}_n) = l_B \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + \frac{l_B}{l_A^2} \sum_{i=1}^{l_A} \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)^2 +$$

$$+ 2\frac{l_B}{l_A} \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) =$$

$$= l_B \sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + l_B\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 +$$

$$+ \frac{l_B}{l_A}\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)^2 + 2\frac{l_B}{l_A} \sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right).$$

Так как в силу определения величин $\overline{X}_{i\cdot}$ и $\overline{X}_{\cdot(2\ldots l_A)}$

$$\sum_{i=2}^{l_A} \overline{X}_{i\cdot} = (l_A - 1)\overline{X}_{\cdot(2\ldots l_A)},$$

то

$$\sum_{i=2}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) = \sum_{i=2}^{l_A} \overline{X}_{i\cdot} - (l_A - 1)\overline{X}_{\cdot(2\ldots l_A)} = 0.$$

Поэтому

$$\sum_{i=1}^{l_A} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)\left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right) =$$

$$= -\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + \left(\overline{X}_{\cdot(2\ldots l_A)} - \overline{X}_{1\cdot}\right)\sum_{i=1}^{l_A}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right) = -\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2.$$

Собирая теперь все слагаемые, получаем

$$Q_A(\vec{X}_n) = l_B\sum_{i=2}^{l_A}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2 + l_B\left(1 - \frac{2}{l_A} + \frac{1}{l_A}\right)\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot(2\ldots l_A)}\right)^2,$$

что равносильно (8.16).

Для проверки гипотезы $H_0^{(i)}$ по результатам наблюдений используют статистику

$$F = \frac{S_A''(\vec{X}_n)^2}{S_0^2(\vec{X}_n)},$$

где

$$S_A''(\vec{X}_n)^2 = \frac{Q_A''(\vec{X}_n)}{l_A - 2}.$$

Эта статистика имеет *распределение Фишера* с числом степеней свободы $l_A - 2$ и $(l_A - 1)(l_B - 1)$, если гипотеза $H_0^{(i)}$ верна*.

Аналогично строятся критерии для проверки влияния фактора B на отклик X.

Порядок проведения двухфакторного анализа представим в виде таблицы (табл. 8.3).

Пример 8.5. В табл. 8.5 приведены опытные данные спектрографического исследования с целью проверки влияния различных фотопленок (фактор A) и электродов (фактор B) на величину X (отклик), характеризующую интенсивность света.

Таблица 8.5

Уровни фактора $B(j)$	Уровни фактора $A(i)$				
	1	2	3	4	5
1	4	18	26	38	44
2	3	19	25	35	43
3	6	18	24	28	39
4	7	13	21	31	38

В данном случае фактор A имеет $l_A = 5$ уровней, фактор B — $l_B = 4$ уровня, число опытов равно $n = l_A l_B = 20$.

Проверим на уровне значимости $\alpha = 0{,}01$ гипотезы:

H_0^A — отсутствие влияния фактора A на величину X;

H_0^B — отсутствие влияния фактора B на величину X.

Для этого рассчитаем

$$\overline{x} = \frac{1}{l_A l_B} \sum_{i=1}^{5} \sum_{j=1}^{4} x_{ij} = \frac{1}{20}(20 + 68 + 96 + 132 + 164) = 24.$$

Значения статистик $\overline{X}_{i\cdot}$ и $\overline{X}_{\cdot j}$, вычисленные по формулам

$$\overline{x}_{i\cdot} = \frac{1}{l_B} \sum_{j=1}^{4} x_{ij} \qquad \text{и} \qquad \overline{x}_{\cdot j} = \frac{1}{l_A} \sum_{i=1}^{5} x_{ij},$$

приведены соответственно в табл. 8.6 и 8.7.

Таблица 8.6

i	1	2	3	4	5
$\overline{X}_{i\cdot}$	20	68	96	132	164

Таблица 8.7

i	1	2	3	4
$\overline{X}_{\cdot j}$	26	25	23	22

Далее вычисляем

$$Q_A = l_B \sum_{i=1}^{l_A} (\overline{x}_{i\cdot} - \overline{x})^2 = 4(361 + 49 + 81 + 289) = 3120,$$

$$Q_B = 5 \sum_{j=1}^{4} (\overline{x}_{\cdot j} - \overline{x})^2 = 5(4 + 1 + 1 + 4) = 50,$$

$$Q = l_B \sum_{i=1}^{5} \sum_{j=1}^{4} (x_{ij} - \overline{x})^2 =$$

$$= 400 + 36 + 4 + 196 + 400 + 441 + 25 + 1 + 100 + 361 +$$

$$+ 324 + 36 + 16 + 225 + 289 + 121 + 9 + 49 + 196 = 3229.$$

Таблица 8.8

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средняя сумма квадратов	Статистика
Фактор A	$Q_A = 3120$	$l_A - 1 = 4$	$S_A^2 = \dfrac{Q_A}{l_A - 1} = 780$	$F_в = \dfrac{S_A^2}{S_0^2} = 158{,}54$
Фактор B	$Q_B = 50$	$l_B - 1 = 3$	$S_B^2 = \dfrac{Q_B}{l_B - 1} = 16{,}67$	$F_в = \dfrac{S_B^2}{S_0^2} = 3{,}39$
Ошибки	$Q_0 = 59$	$(l_A - 1) \times (l_B - 1) = 12$	$S_0^2 = \dfrac{Q_0}{(l_A - 1)(l_B - 1)} = 4{,}92$	
Сумма	$Q = 3229$	$l_A l_B - 1 = 19$		

Находим разность вычисленных величин:

$$Q_0 = Q - Q_A - Q_B = 59.$$

Полученные результаты сведем в таблицу (табл. 8.8). Поскольку

$$f_в = 158{,}7 > f_{кр} = f_{0{,}01}(4, 12) = 5{,}41,$$

то гипотезу H_0^A следует отвергнуть. Гипотезу H_0^B следует принять, так как

$$f_в = 3{,}39 < f_{кр} = f_{0{,}01}(3, 12) = 5{,}92.$$