Дисперсионный анализ

Дисперсионный анализ

Дисперсионный анализ (от латинского Dispersio – рассеивание / на английском Analysis Of Variance - ANOVA) применяется для исследования влияния одной или нескольких качественных переменных (факторов-качественные, количественные, случайные) на одну зависимую количественную переменную (отклик).

В дисперсионном анализе используется свойство аддитивности дисперсии независимых факторов.

Р.А.Фишер в 1938 году впервые определил дисперсионный анализ как «отделение дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам»

Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость изучаемой случайной величины. Для этого проводят *разложение суммарной дисперсии на составляющие*, обусловленные независимыми факторами.

Проверка значимости оценок дисперсий проводиться по *F-критерию Фишера:*

применяют для сравнения двух независимых нормально распределенных выборочных совокупностей.

Выборочные дисперсии S_1^2, S_2^2 различаются значимо, если частное S_1^2 / S_2^2 превышает табличный $F_{кр}$ *критерий Фишера* для принятой доверительной вероятности *p* и чисел степеней свободы $f_1=n_1-1$, $f_2=n_2-1$.

Допущения в дисперсионном анализе:

1.случайные ошибки наблюдений имеют нормальное распределение;

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x_i < +\infty)$$

где μ и σ^2 - математическое ожидание и генеральная дисперсия случайной величины X.

2.факторы влияют только на изменение средних значений, а дисперсия наблюдений остается постоянной; эксперименты равноточны.

Факторы, рассматриваемые в дисперсионном анализе, бывают двух родов:

1) со случайными уровнями (выбор уровней производится из бесконечной совокупности возможных уровней - <u>модель со случайными уровнями факторов</u>);

2) с фиксированными (все уровни фиксированы – <u>модель с фиксированными уровнями факторов);</u>

3) модель смешанного типа (часть факторов рассматривается на фиксированных уровнях, а уровни остальных выбираются случайным образом).

Однофакторный дисперсионный анализ

Задачей однофакторного дисперсионного анализа является изучение влияния одного фактора *A* (количественного или качественного), который принимает *k* различных значений (уровней факторов), на рассматриваемый признак (отклик).

На *i-м уровне производиться* n_i наблюдений, результаты которых представлены:

$$
\begin{array}{lll}
y_{11} & y_{21} & y_{к_1}, \\
y_{12} & y_{22} & y_{к_2}, \\
\vdots & \vdots & \vdots \\
y_{1n_1} & y_{2n_2} & y_{kn_k}.
\end{array}
$$

$$y_{ij} = \mu + d_i + \varepsilon_{ij}$$

где μ - суммарный эффект во всех опытах;

d_i - эффект фактора А на i-м уровне (i = 1,2,…,k);

ε_{ij} - ошибка измерения на i-м уровне.

Предположим, что наблюдения на фиксированном уровне фактора нормально распределены относительно среднего значения $\mu + d_i$ с обшей дисперсией σ^2.

Общее число опытов равно N :

$$N=n_1+n_2+\ldots+n_k.$$

Проверяется нулевая гипотеза равенства средних значений на различных уровнях фактора A :

$$m_1=m_2=\ldots=m_k=m.$$

Расчеты при равном числе опытов на каждом уровне фактора *A* : $n_1=n_2=\ldots=n_k=n$.

Таблица Исходные данные для однофакторного дисперсионного анализа с равным числом повторений опытов

Номер наблюдения	Уровни фактора A			
	a_1	a_2	...	a_k
1	y_{11}	y_{21}		y_{k1}
2	y_{12}	y_{22}		y_{k2}
.	.	.		.
.	.	.		.
.	.	.		.
n	y_{1n}	y_{2n}		y_{kn}
Итоги	$A_1=\sum_{j=1}^{n} y_{1j}$	$A_2=\sum_{j=1}^{n} y_{2j}$		$A_k=\sum_{j=1}^{n} y_{kj}$

Обозначим среднее значение наблюдений на *i*-том уровне:

$$\overline{y_i} = \frac{\sum_{j=1}^{n} y_{ij}}{n} = \frac{A_i}{n}$$

общее среднее значение для всей выборки из N наблюдений:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{y_i}\,.$$

Общую выборочную дисперсию разложим на составляющие, которые характеризовали бы вклад фактора *A* и фактора случайности.

$$s^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{y})^2}{N-1} = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}\right)^2}{N}\right]$$

Определим выборочную дисперсию на каждом уровне:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n} y_{ij}\right)^2}{n}\right]$$

$$i = 1, 2, \ldots, k.$$

Критерий Кохрана (проверка однородности системы) применяют для сравнения *k* независимых нормально распределенных выборочных совокупностей равных объемов n_i=*const* с дисперсиями S_i^2. Выборочные дисперсии различаются значимо, если *критерий Кохрана* *G* превышает табличный $G_{кр}$ для принятой доверительной вероятности *p* и числа степеней свободы *f=k-1.*

$$G = \frac{S_{\max}^2}{\sum_{i=1}^{k} S_i^2}$$

Если между выборочными дисперсиями нет значимых различий, для оценки генеральной дисперсии σ^2, характеризующей фактор случайности, используют выборочную дисперсию $S_{ош}^2$: (f=k(n-1)=N-k).

$$s_{ош}^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right]$$

Приближенную оценку для дисперсии фактора *A* можно получить следующим образом:

$$\sigma_A^2 \approx s^2 - s_{ош}^2$$

Более точную оценку для σ_A^2 можно получить, рассматривая отклонения средних $\overline{y}_i$ на отдельных уровнях от общего среднего всей выборки $\overline{y}$

$$\frac{1}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 \approx \sigma_A^2 + \frac{\sigma^2}{n} \approx \sigma_A^2 + \frac{s_{ош}^2}{n}$$

$$\sigma_A^2 \approx \frac{1}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 - \frac{s_{\text{ош}}^2}{n}$$

Введем следующее обозначение, дисперсия фактора A (проверка нулевой гипотезы по критерию Фишера):

$$s_A^2 = \frac{n}{k-1}\sum_{i=1}^{k}(y_i - \overline{y})^2 \approx n\sigma_A^2 + s_{\text{ош}}^2$$

Влияние фактора является значимым, если:

$$\frac{s_A^2}{s_{\text{ош}}^2} > F_{1-p}(f_1,\ f_2),\ f_1 = k-1,\ f_2 = k(n-1) = N-k.$$

Алгоритм:

1) итоги по столбцам

$$A_i = \sum_{j=1}^{n} y_{ji};$$

2) сумму квадратов всех наблюдений

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}^2;$$

3) сумму квадратов итогов по столбцам , деленную на число наблюдений в столбце

$$SS_2 = \frac{1}{n} \sum_{i=1}^{k} A_i^2$$

4) квадрат общего итога, деленный на число всех наблюдений (корректирующий член)

$$SS_3 = \frac{1}{N} \left(\sum_{i=1}^{k} A_i \right)^2$$

5) сумма квадратов для столбца

$$SS_A = SS_2 - SS_3;$$

6) $SS_{общ}$ - общая сумма квадратов, равная разнице между суммой квадратов всех наблюдений и корректирующим членом

$$SS_{общ} = SS_1 - SS_3;$$

7) $SS_{ост}$ - остаточная сумма квадратов для оценки ошибки эксперимента

$$SS_{ост} = SS_1 - SS_2;$$

8) дисперсия s_A^2

$$s_A^2 = \frac{SS_A}{k-1};$$

9) дисперсия $s_{ош}^2$

$$s_{ош}^2 = \frac{SS_{ост}}{k(n-1)}$$

Результаты расчета представляются в виде таблицы дисперсионного анализа.

Таблица 6. **Однофакторный дисперсионный анализ (с равным числом повторений опытов)**

Иточник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
A	$k-1$	$SS_A = SS_2 - SS_3$	$s_A^2 = \frac{SS_A}{k-1}$	$n\sigma_A^2 + \sigma_{ош}^2$
Остаток	$k(n-1)$	$SS_{ост} = SS_1 - SS_2$	$s_{ош}^2 = \frac{SS_{ост}}{k(n-1)}$	$\sigma_{ош}^2$
Общая сумма	$kn-1$	$SS_{общ} = SS_1 - SS_3$	$\frac{SS_{общ}}{kn-1}$	

Если неравенство, $$\frac{s_A^2}{s_{ош}^2} > F_{1-p}(f_1,\ f_2),\ f_1 = k-1,\ f_2 = k(n-1) = N-k.$$

справедливо, то различие между s_A^2 и $s_{ош}^2$ значимо, следовательно значимо влияние фактора *A*.

$$\sigma_A^2 \approx \frac{s_A^2 - s_{ош}^2}{n}$$

Нулевая гипотеза отвергается и различие между средними считается значимым.
Для выявления различности средних применяют критерии Стьюдента, Фишера или ранговый критерий Дункана.

Если выборочные дисперсии различаются в пределах случайного разброса, то следующим шагом является сравнение выборочных средних.

Выборочные средние различаются значимо, если *t-критерий Стьюдента* превышает табличный $t_{p,f}$ для принятой доверительной вероятности *p* и числа степеней свободы объединенной выборки $f=n_1+n_2-2$.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{S_{1,2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

где $S_{1,2}^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ – средневзвешенная дисперсия.

Пример: применение однофакторного дисперсионного анализа для выяснения влияния вида галоидного алкила (фактор А) на процесс полимеризации.

Номер наблюдения	Уровни фактора *A*				
	a_1	a_2	a_3	a_4	a_5
1	79,80	87,30	42,45	76,0	70,70
2	86,30	69,60	64,3	83,5	64,65
3	86,50	81,75	78,9	72,80	38,50
4	92,30	77,95	61,00	89,00	77,00
5	76,50	83,65	31,30	76,50	91,50
6	87,05	64,80	72,85	87,45	68,00
7	82,50	67,30	58,65	74,50	38,05
8	90,00	75,45	52,50	93,15	79,95
Итоги	A1=680,95	A2=607,8	A3-461,95	A4=652,9	A5=528,35

Двухфакторный дисперсионный анализ

Изучается влияние на процесс одновременно двух факторов А и В. Фактор А исследуется , на уровнях a_1, a_2,..., a_k .Фактор В – на уровнях b_1,b_2,...,b_m .

Таблица 7. **Данные для двухфакторного дисперсионного анализа с повторениями**

B	A						Итоги
	a_1	a_2	...	a_i	...	a_k	
b_1	$y_{111}, y_{112}, \ldots, y_{11n}$	$y_{211}, y_{212}, \ldots, y_{21n}$	...	$y_{i11}, y_{i12}, \ldots, y_{i1n}$	...	$y_{k11}, y_{k12}, \ldots, y_{k1n}$	B_1
b_2	$y_{121}, y_{122}, \ldots, y_{12n}$	$y_{221}, y_{222}, \ldots, y_{22n}$	...	$y_{i21}, y_{i22}, \ldots, y_{i2n}$	...	$y_{k21}, y_{k22}, \ldots, y_{k2n}$	B_2
b_j	$y_{1j1}, y_{1j2}, \ldots, y_{1jn}$	$y_{2j1}, y_{2j2}, \ldots, y_{2jn}$	...	$y_{ij1}, y_{ij2}, \ldots, y_{ijn}$	...	$y_{kj1}, y_{kj2}, \ldots, y_{kjn}$	Bj
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
b_m	$y_{1m1}, y_{1m2}, \ldots, y_{1mn}$	$y_{2m1}, y_{2m2}, \ldots, y_{2mn}$	...	$y_{im1}, y_{im2}, \ldots, y_{imn}$	...	$y_{km1}, y_{km2}, \ldots, y_{kmn}$	Bm
Итоги	A_1	A_2	...	A_i	...	A_k	

Общее число наблюдений равно *N=nkm*

$$y_{ijq} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijq}$$

μ -Общее среднее;

α_i -эффект фактора А на *i*-м уровне, *i*=1,2,…, *K*;

β_j -эффект фактора В на *j*-м уровне, *j*=1,2,…*m*;

$\alpha_i\beta_j$ -эффект взаимодействия факторов, представляем собой отклонение среднего по наблюдениям в (ij)-й серии от суммы первых- трех членов в модели

ε_{ijq} -учитывает вариацию внутри серии наблюдений (ошибка воспроизводимости)

Если предположить, что между факторами нет взаимодействия, то можно использовать линейную модель: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Таблица 8. Данные для двухфакторного дисперсионного анализа без повторений

B	A				Итоги
	a_1	a_2	...	a_k	
b_1	y_{11}	y_{21}	...	y_{k1}	B_1
b_2	y_{12}	y_{22}	...	y_{k2}	B_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
b_m	y_{1m}	y_{2m}	...	y_{km}	B_m
Итоги	A_1	A_2	.	A_k	

Линейная модель:

Через $\overline{y}_i$ и $\overline{y}_j$ обозначим соответственно средние значения по строкам и столбцам: $\overline{y}_i = \frac{A_i}{m}$,

$$\overline{y}_j = \frac{B_j}{k} ,$$

А $\overline{\overline{y}}$ - среднее всех результатов $\overline{\overline{y}} = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij} .$

Рассеяние средних по столбцам $\overline{y}_1, \overline{y}_2, \ldots \overline{y}_k$ относительно общего среднего $\overline{\overline{y}}$ не зависит от фактора B, т.к. все уровни фактора B усреднены.

Это рассеяние связано с влиянием фактора А и случайного фактора. Так как дисперсия среднего в *m* раз меньше дисперсии единичного измерения, имеем:

$$\frac{1}{k-1} \sum_{i=1}^{k} (\overline{y}_i - \overline{\overline{y}})^2 \approx \sigma^2_A + \frac{\sigma^2}{m} .$$

Линейная модель:

В свою очередь, рассеяние в средних по строкам на зависит от фактора А и связано с влиянием фактора В:

$$\frac{1}{m-1}\sum_{j=1}^{m}(\overline{y}_j' - \overline{\overline{y}})^2 \approx \sigma_B^2 + \frac{\sigma^2}{k}. \qquad (а)$$

Эти равенства позволяют оценить влияние факторов А и В, если известна оценка дисперсии.

Для оценки фактора случайности при отсутствии параллельных наблюдений, найдем дисперсию наблюдений по i-му столбцу:

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2. \qquad (б)$$

Эта дисперсия обусловлена влиянием фактора B и фактора случайности

$$s_i^2 \approx \sigma_B^2 + \sigma^2.$$

Вычитая (б) из (а), получим

$$\sigma^2 - \frac{\sigma^2}{k} \approx \frac{1}{k(m-1)} \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \overline{y}_i)^2 - \frac{1}{m-1} \sum_{j=1}^{m} (\overline{y}'_j - \overline{\overline{y}})^2.$$

Отсюда

$$\sigma^2 \cong \frac{1}{(k-1)(m-1)} \left[\sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \overline{y}_i)^2 - k \sum_{j=1}^{m} (\overline{y}'_j - \overline{\overline{y}})^2 \right].$$

Обозначим полученную оценку для дисперсии σ^2 через $S_{ош}^2$

Введем следующие обозначения:

$$s_A^2 = \frac{m}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{\overline{y}})^2 \approx m\sigma_A^2 + s_{\text{ош}}^2 ,$$

$$s_B^2 = \frac{k}{m-1}\sum_{j=1}^{m}(\overline{y}'_j - \overline{\overline{y}})^2 \approx k\sigma_B^2 + s_{\text{ош}}^2 .$$

Величины s_A^2 и s_B^2 можно считать выборочными дисперсиями с (к-1) и (m-1) степенями свободы соответственно. Проверяют нулевые гипотезы о незначимости влияния факторов А и В по критерию Фишера.

$$F = \frac{s_A^2}{s_{\text{ош}}^2} < F_{1-p}(f_1,\ f_2), \quad f_1 = k-1; \quad f_2 = (k-1)(m-1),$$

Нулевая гипотеза значима, α_i=0.

Если

$$F = \frac{s_A^2}{s_{ош}^2} > F_{1-p}(f_1,\ f_2),$$

нулеваая гипотеза отвергается и влияние фактора А считается значимым.
Аналогично, если

$$F = \frac{s_B^2}{s_{ош}^2} < F_{1-p}(f_1,\ f_2),$$

$$f_1 = m - 1; \quad f_2 = (k-1)(m-1),$$

Гипотеза принимается, βj =0. При справедливости неравенства:

$$s_B^2 = \frac{k}{m-1}\sum_{j=1}^{m}(\bar{y}'_j - \bar{\bar{y}})^2 \approx k\sigma_B^2 + s_{ош}^2 .$$

Влияние фактора В считается значимым.

При проведении дисперсионного анализа в условиях линейной модели, используют следующий алгоритм расчета:

Находят :

1) Итоги по столбцам

$$A_i = \sum_{j=1}^{m} y_{ij}, \quad i = 1, 2, \ldots, k;$$

2) Итоги по строкам

$$B_j = \sum_{i=1}^{k} y_{ij}, \quad j = 1, 2, \ldots, m;$$

3) Сумму квадратов всех наблюдений

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij}^2;$$

4) Сумму квадратов итогов по столбцам , деленную на число наблюдений в столбце

$$SS_2 = \frac{1}{m} \sum_{i=1}^{k} A_i^2;$$

5) Сумму квадратов итогов по строкам., деленную на число наблюдений в строке

$$SS_3 = \frac{1}{k} \sum_{j=1}^{m} B_j^2 ;$$

6) Квадрат общего итога, деленный на число всех наблюдений (корректирующий член)

$$SS_4 = \frac{1}{mk} \left(\sum_{i=1}^{k} A_i \right)^2 = \frac{1}{mk} \left(\sum_{j=1}^{m} B_j \right)^2 ;$$

7) Сумму квадратов для столбца: $SS_A = SS2 - SS4$;

8) Сумму квадратов для строки: $SS_B = SS3 - SS4$;

9) Общую сумму квадратов, равную разнице между суммой квадратов всех наблюдений и корректирующим членом: $SS_{общ} = SS1 - SS4$;

10) Остаточную сумму квадратов:

$$SS_{ост} = SS_{общ} - SS_A - SS_B = SS1 - SS2 - SS3 + SS4;$$

11) Дисперсию s_A^2: $s_A^2 = SS_A/(k-1)$

12) Дисперсию s_B^2: $s_B^2 = SS_B/(m-1)$

13) Дисперсию $s_{ош}^2$: $s_{ош}^2 = \dfrac{SS_{ост}}{(k-1)(m-1)}$

Таблица 9. **Двухфакторный дисперсионный анализ (без повторения опытов)**

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
A	$k-1$	$SS_A = SS_2 - SS_4$	$s_A^2 = \dfrac{SS_A}{k-1}$	$m\sigma_A^2 + \sigma_{\text{ош}}^2$
B	$m-1$	$SS_B = SS_3 - SS_4$	$s_B^2 = \dfrac{SS_B}{m-1}$	$k\sigma_B^2 + \sigma_{\text{ош}}^2$
Остаток	$(k-1)(m-1)$	$SS_{\text{ост}} = SS_1 - SS_2 - SS_3 + SS_4$	$s_{\text{ош}}^2 = \dfrac{SS_{\text{ост}}}{(k-1)(m-1)}$	$\sigma_{\text{ош}}^2$
Общая сумма	$km-1$	$SS_{\text{общ}} = SS_1 - SS_4$		

Рассмотрим **модель со взаимодействием** факторов А и В. Пусть при каждом сочетании уровней факторов А и В проводится n параллельный опытов. Имеется целая серия наблюдений $y_{ij1}, y_{ij2},\ldots,y_{ijn}$. Выборочная дисперсия результатов в каждой ячейке, где (n-1)-степень свободы:

$$s_{ij}^2 = \frac{1}{n-1}\sum_{u=1}^{n}(y_{iju} - \bar{y}_{ij})^2$$

Если выборочные дисперсии по всем ячейкам однородны, их можно усреднить и использовать полученную средневзвешенную дисперсию в качестве оценки для дисперсии воспроизводимости σ^2 :

$$s_{\text{ош}}^2 = \frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m}s_{ij}^2$$

Число степеней свободы равно *mk(n-1)*

Более удобная формула для вычисления дисперсии воспроизводимости

$$s^2_{\text{ош}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{m} \sum_{u=1}^{n} y^2_{iju} - \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} y^2_{ij}}{n}}{mk(n-1)},$$

где y_{ij} - сумма наблюдений в ij – й ячейке.

При проведении дисперсионного анализа в условиях модели с учетом взаимодействия факторов А и В, удобно использовать следующий алгоритм расчета:

1) суммы наблюдений в каждой ячейке

$$y_{ij} = \sum_{u=1}^{n} y_{iju},$$

$$j = 1, 2, \ldots, m; \quad i = 1, 2, \ldots, k;$$

2) квадрат сумм наблюдений в каждой ячейке

$$y_{ij}^2 = \left(\sum_{u=1}^{n} y_{iju} \right)^2;$$

3) итоги по столбцам

$$A_i = \sum_{j=1}^{m} \sum_{u=1}^{n} y_{iju};$$

4) итоги по строкам

$$B_j = \sum_{i=1}^{k} \sum_{u=1}^{n} y_{iju};$$

5) сумму всех наблюдений (общий итог)

$$\sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{u=1}^{n} y_{iju} = \sum_{i=1}^{k} A_i = \sum_{j=1}^{m} B_j;$$

6) сумму квадратов всех наблюдений

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{u=1}^{n} y_{iju}^2;$$

7) Сумму квадратов итогов по столбцам , деленную на число наблюдений в столбце

$$SS_2 = \frac{1}{mn} \sum_{i=1}^{k} A_i^2;$$

8) Сумму квадратов итогов по строкам., деленную на число наблюдений в строке

$$SS_3 = \frac{1}{kn} \sum_{j=1}^{m} B_j^2;$$

9) Квадрат общего итога, деленный на число всех наблюдений (корректирующий член)

$$SS_4 = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju}\right)^2}{N} = \frac{1}{mkn}\left(\sum_{i=1}^{k} A_i\right)^2 = \frac{1}{mkn}\left(\sum_{j=1}^{m} B_j\right)^2;$$

10) Сумму квадратов для столбца $SS_A = SS_2 - SS_4;$

11)Сумму квадратов для строки $SS_B = SS_3 - SS_4;$

12) Сумму квадратов для дисперсии воспроизводимости

$$SS_{\text{ош}} = SS_1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}^2}{n};$$

13) Общую сумму квадратов, равную разнице между суммой квадратов всех наблюдений и корректирующим членом $SS_{\text{общ}} = SS_1 - SS_4;$

14) Остаточную сумму квадратов отклонений для эффекта взаимодействия АВ

$$SS_{AB} = SS_{\text{общ}} - SS_A - SS_B - SS_{\text{ош}};$$

15) дисперсию s^2_A

$$s^2_A = SS_A/(k-1);$$

16) дисперсию s^2_B

$$s^2_B = SS_B/(m-1);$$

17) дисперсию s^2_{AB}

$$s^2_{AB} = \frac{SS_{AB}}{(k-1)(m-1)};$$

18) дисперсию воспроизводимости

$$s^2_{\text{ош}} = \frac{SS_{\text{ош}}}{mk(n-1)}.$$

Таблица 10. **Двухфакторный дисперсионный анализ для модели со случайными уровнями (с повторными опытами)**

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
A	$k-1$	$SS_A = SS_2 - SS_4$	$s_A^2 = \dfrac{SS_A}{k-1}$	$nm\sigma_A^2 + n\sigma_{AB}^2 + \sigma_{\text{ош}}^2$
B	$m-1$	$SS_B = SS_3 - SS_4$	$s_B^2 = \dfrac{SS_B}{m-1}$	$nk\sigma_B^2 + n\sigma_{AB}^2 + \sigma_{\text{ош}}^2$
AB	$(k-1)(m-1)$	$SS_{AB} = SS_1 - SS_2 - SS_3 + SS_4$	$s_{AB}^2 = \dfrac{SS_{AB}}{(k-1)(m-1)}$	$n\sigma_{AB}^2 + \sigma_{\text{ош}}^2$
Остаток (ошибка)	$mk(n-1)$	$SS_{\text{ош}} = SS_1 - \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}^2}{n}$	$s_{\text{ош}}^2 = \dfrac{SS_{\text{ош}}}{mk(n-1)}$	$\sigma_{\text{ош}}^2$
Общая сумма	$mkn-1$	$SS_{\text{общ}} = SS_1 - SS_4$		

Для оценки значимости фактора А необходимо составить дисперсионное отношение вида

$$F = s_A^2 / s_{AB}^2.$$

Влияние фактора *A* признается значимым, если

$$s_A^2 / s_{AB}^2 > F_{1-p}(f_1, f_2),$$

$$f_1 = k - 1; \quad f_2 = (k-1)(m-1).$$

влияние фактора *B* считается значимым, если

$$s_B^2 / s_{AB}^2 > F_{1-p}(f_1, f_2),$$

$$f_1 = m - 1, f_2 = (k-1)(m-1).$$

Пример:

B	A			
	a_1	a_2	a_3	a_4
b1	13,2	4,7	53,4	13,6
	13,9	5,8	48,3	13,2
b2	18,9	19,8	14,0	9,5
	21,0	17,9	13,2	8,6
b3	7,3	38,2	5,1	54,4
	8,5	37,7	5,9	55,2
b4	20,0	60,1	19,6	58,2
	20,8	60,9	18,5	59,7

Многофакторный дисперсионный анализ

Латинские и гипер-греко-латинские квадраты.

Полным факторным экспериментом(ПФЭ) называют эксперимент, в котором встречаются все возможные сочетания уровни изучаемых факторов. Дробным факторным экспериментом(ДФЭ) – эксперимент, в котором пропущены некоторые сочетания уровней.

Рассмотрим трехфакторный дисперсионный анализ при одинаковом числе уровней n для каждого фактора. Полный перебор сочетаний уровней факторов потребует N опытов $N=n^3$

Число опытов можно значительно сократить, используя ДФЭ по схеме латинского квадрата, введенного впервые Фишером. Латинский квадрат - n:n – это квадратная матрица, составленная из n элементов(чисел или букв) таким образом, что каждый элемент повторяется в каждой строке и в каждом столбце только один раз.

A	*B*	*C*
B	*C*	*A*
C	*A*	*B*

Стандартным или каноническим латиннским квадратами называются такие квадраты, у которых первая строка и первый столбец построены в алфавитном порядке или в порядке натурального ряда.

Таблица 11. 2×2 латинский квадрат

A	B	
	b_1	b_2
a_1	c_1	c_2
a_2	c_2	c_1

Таблица 12. План эксперимента $n = 2$; $N = 4$

Номер	A	B	C	y
1	a_1	b_1	c_1	y_1
2	a_1	b_2	c_2	y_2
3	a_2	b_1	c_2	y_3
4	a_2	b_2	c_1	y_4

Результат наблюдения, полученный при ПФЭ

$$y_{ijq} = \mu + \alpha_i + \beta_j + \gamma_q + \alpha_i\beta_j + \alpha_i\gamma_q + \beta_j\gamma_q + \alpha_i\beta_j\gamma_q + \varepsilon_{ijq}.$$

При применении латинского квадрата предполагают, что результаты взаимодействия незначимы и применяют линейную модель

$$y_{ijq} = \mu + \alpha_i + \beta_j + \gamma_q + \varepsilon_{ijq}$$

Алгоритм расчета: Для этого определяют

1) итоги по строкам Ai, столбцам Bj, и латинским буквам Cq.

Например, для латинского квадрата 3:3 итоги по строкам

$$A_1 = y_1 + y_2 + y_3, \quad A_2 = y_4 + y_5 + y_6, \quad A_3 = y_7 + y_8 + y_9;$$

Итоги по столбцам

$$B_1 = y_1 + y_4 + y_7, \quad B_2 = y_2 + y_5 + y_8, \quad B_3 = y_3 + y_6 + y_9;$$

Итоги по латинским буквам

$$C_1 = y_1 + y_6 + y_8,$$
$$C_2 = y_2 + y_4 + y_9,$$
$$C_3 = y_3 + y_5 + y_7;$$

Таблица 14. Латинский квадрат 3×3

A	B			Итоги
	b_1	b_2	b_3	
a_1	c_1 y_1	c_2 y_2	c_3 y_3	A_1
a_2	c_2 y_4	c_3 y_5	c_1 y_6	A_2
a_3	c_3 y_7	c_1 y_8	c_2 y_9	A_3
Итоги	B_1	B_2	B_3	

2) сумму квадратов всех наблюдений

$$SS_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij}^2;$$

3) Сумму квадратов итогов по строкам, деленную на число наблюдений в строке

$$SS_2 = \frac{1}{n} \sum_{i=1}^{n} A_i^2;$$

4) Сумму квадратов итогов по столбцам, деленную на число наблюдений в столбце

$$SS_3 = \frac{1}{n} \sum_{j=1}^{n} B_j^2;$$

5)Сумму квадратов итогов по латинским буквам , деленную на число наблюдений, соответствующих каждой букве

$$SS_4 = \frac{1}{n}\sum_{q=1}^{n} C_q^2 ;$$

6) Квадрат общего итога, деленный на число всех наблюдений(корректирующий член)

$$SS_5 = \frac{1}{n^2}\left(\sum_{i=1}^{n} A_i\right)^2 = \frac{1}{n^2}\left(\sum_{j=1}^{n} B_j\right)^2 = \frac{1}{n^2}\left(\sum_{q=1}^{n} C_q\right)^2 ;$$

7) сумму квадратов для строки

$$SS_A = SS_2 - SS_5 ;$$

8) сумму квадратов для столбца

$$SS_B = SS_3 - SS_5 ;$$

9) сумму квадратов для латинской буквы

$$SS_C = SS_4 - SS_5 ;$$

10) Общую сумму квадратов, равную разнице между суммой квадратов всех наблюдений и корректирующим членом

$$SS_{общ} = SS_1 - SS_5 ;$$

11) остаточную сумму квадратов

$$SS_{ост} = SS_{общ} - SS_A - SS_B - SS_C = SS_1 - SS_5 - SS_2 + SS_5 - SS_3 + $$
$$+ SS_5 - SS_4 + SS_5 = SS_1 - SS_2 - SS_3 - SS_4 + 2SS_5.$$

12) дисперсию s^2_A

$$s^2_A = SS_A/(n-1);$$

13) дисперсию s^2_B

$$s^2_B = SS_B/(n-1);$$

14) дисперсию s^2_C

$$s^2_C = SS_C/(n-1);$$

15) дисперсию $s^2_{ош}$

$$s^2_{ош} = \frac{SS_{ост}}{(n-1)(n-2)}.$$

Значимость линейных эффектов проверяют по критерию Фишера.

$$s^2_A / s^2_{ош} < F_{1-p}(f_1, f_2),$$

$$s^2_B / s^2_{ош} < F_{1-p}(f_1, f_2),$$

$$s^2_C / s^2_{ош} < F_{1-p}(f_1, f_2),$$

Таблица 15. **Дисперсионный анализ латинского квадрата (без повторных опытов)**

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
A	$n-1$	$SS_A = SS_2 - SS_5$	$s_A^2 = \dfrac{SS_A}{n-1}$	$n\sigma_A^2 + \sigma_{\text{ош}}^2$
B	$n-1$	$SS_B = SS_3 - SS_5$	$s_B^2 = \dfrac{SS_B}{n-1}$	$n\sigma_B^2 + \sigma_{\text{ош}}^2$
C	$n-1$	$SS_C = SS_4 - SS_5$	$s_C^2 = \dfrac{SS_C}{n-1}$	$n\sigma_C^2 + \sigma_{\text{ош}}^2$
Остаток (ошибка)	$(n-1)(n-2)$	$SS_{\text{ост}} = SS_1 - SS_2 - SS_3 - SS_4 + 2SS_5$	$s_{\text{ош}}^2 = \dfrac{SS_{\text{ост}}}{(n-1)(n-2)}$	$\sigma_{\text{ош}}^2$
Общая сумма	$n^2 - 1$	$SS_{\text{общ}} = SS_1 - SS_5$		