ДИСПЕРСИОННЫЙ АНАЛИЗ

Основные задачи дисперсионного анализа

Дисперсионный анализ предназначен для проверки наличия зависимости нормально распределенной результативной случайной величины Y от нескольких факторов (факторных величин), а именно для выявления причинно-следственной связи между вариацией факторов и вариацией результативных признаков.

Суть дисперсионного анализа состоит в разложении дисперсии признака на составляющие, обусловленные влиянием конкретных факторов и проверке гипотез о значимости их влияния.

Классификация моделей дисперсионного анализа

Модели дисперсионного анализа классифицируются следующим образом:

1) в зависимости от числа факторов на однофакторные, двухфакторные и т.д.;

2) по природе факторов на детерминированные (M_1), случайные (M_2) и смешанные, в зависимости от того какими являются уровни факторов.

Параметрический однофакторный дисперсионный анализ

Постановка задачи

Пусть требуется проверить наличие влияния на результативный признак одного контролируемого фактора A, имеющего m уровней A_j, $j = 1,2,...m$. Наблюдаемые значения результативного признака Y на каждом из фиксированных уровней A_j обозначим y_{ij}, $i = \overline{1, n_j}$, где n_j - число объектов наблюдения.

Для изучения случайных величин $\xi_1, \xi_2, \dots, \xi_m$ рассматриваем априорные выборки $\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}$, где $\xi_{j,n_j}, j = 1..m$

Реализации априорных выборок представлены в таблице :

Таблица -Реализация априорных выборок

A_1 (для ξ_1)	A_2 (для ξ_2)	…	A_m (для ξ_m)
y_{11}	y_{12}	…	y_{1m}
y_{21}	y_{22}	…	y_{2m}
…	y_{23}	…	y_{3m}
$y_{n_1 1}$	…	…	y_{4m}
	$y_{n_2 2}$	…	…

Однофакторная модель дисперсионного анализа

Апостериорная модель однофакторного дисперсионного анализа:

$$y_{ij} = a + \alpha_j + z_{ij}, j = \overline{1,m}, i = \overline{1,n_j}$$

где a – некоторое общее среднее,

α_j – отклонение от среднего, вызванное влиянием фактора на j уровень,

z_{ij} – величина отклонения y_{ij} от $a + \alpha_j$

Априорная модель:

M_1 – уровни фактора А фиксированы

$$\xi_{ij} = a + \alpha_j + \varepsilon_{ij}, \text{где} \sum_{j=1}^{m} \alpha_j = 0 - \text{отклонение}$$

M_2 – уровни фактора А случайны

$$\xi_{ij} = a + \delta_j + \varepsilon_{ij}$$

Требования к δ:

$M\delta_j = 0$

$cov(\delta_j, \delta_s) = M\left((\delta_j - M\delta_j)(\delta_s - M\delta_s)\right) = M\delta_j \cdot \delta_s = 0, \forall\ j \neq s$

$D\delta_j = M\delta_j^2 = \sigma^2$(один для всех уровней)

$cov(\delta_s, \varepsilon_{ij}) = M\delta_s \cdot \varepsilon_{ij} = 0$

$D_{\xi_{ij}} = \sigma_\varepsilon^2$

$D_{\delta_j} = \sigma_\delta^2$

Формулировка гипотезы об отсутствии влияния фактора А на результативный признак

В зависимости от изучаемой модели относительно α_j предполагается:

- модель M_1 – α_j - фиксированные величины, такие что $\sum \alpha_j n_j = 0$

  H_0: $\alpha_j = 0 \ \ \forall j = \overline{1,m}$, то есть нет влияния фактора A на Y;

- модель M_2 – α_j - случайные величины, удовлетворяющие условиям - $M\alpha_j = 0$; $M\alpha_j \alpha_{j'} = 0$ $\forall j \neq j'$; $M\alpha_j \varepsilon_{ij} = 0 \quad \forall i, j$; $M\alpha_j^2 = \sigma^2$ - факторная дисперсия

  H_0: $\sigma_\alpha^2 = 0$, то есть нет влияния фактора A на Y.

Основные характеристики однофакторного дисперсионного анализа

$\bar{y}_{*j}(y_{j,n_j}) = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ - групповые средние (средние уровней A_j);

$\bar{y}_{**}(y_{1,n_1}, y_{2,n_2}, \ldots, y_{m,n_m}) = \bar{y}_{**} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} y_{ij} = \frac{1}{n}\sum_{j=1}^{m} n_j * \bar{y}_{*j}$ - общая средняя результативного признака, где $N = \sum_{j=1}^{m} n_j$.

Апостериорные суммы квадратов отклонений:

$Q_{\text{факт}}(y_{1,n_1}, y_{2,n_2}, \dots, y_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{y}_{*j} - \bar{y}_{**})^2 = \sum_{j=1}^{m} n_j * (\bar{y}_{*j} - \bar{y}_{**})^2$- факторная сумма квадратов отклонений;

$Q_{\text{ост}}(y_{1,n_1}, y_{2,n_2}, \dots, y_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{*j})^2$- остаточная сумма квадратов отклонений;

$Q_{\text{общ}}(y_{1,n_1}, y_{2,n_2}, \dots, y_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{**})^2$- общая сумма квадратов отклонений.

Априорные суммы квадратов отклонений:

$$Q_{\text{факт}}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{y}_{*j(\xi_{1,n_1}, \xi_{2,n_2}, \dots \xi_{m,n_m})} - \bar{y}_{**}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m})$$

$$\sum_{j=1}^{m} n_j * (\bar{y}_{*j}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}) - \bar{y}_{**}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}))^2$$

- факторная сумма квадратов отклонений;

$Q_{\text{ост}}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\xi_{ij} - \bar{y}_{*j}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}))^2$ - остаточная сумма квадратов отклонений;

$Q_{\text{общ}}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\xi_{ij} - \bar{y}_{**}(\xi_{1,n_1}, \xi_{2,n_2}, \dots, \xi_{m,n_m}))^2$ - общая сумма квадратов отклонений.

$$Q_{общ} = Q_{факт} + Q_{ост}$$

Несмещенные оценки общей, факторной и остаточной дисперсий

$$S^{\wedge}2_{\text{общ}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)} = \frac{Q_{\text{общ}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)}}{N-1};$$

$$S^{\wedge}2_{\text{факт}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)} = \frac{Q_{\text{факт}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)}}{m-1};$$

$$S^{\wedge}2_{\text{ост}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)} = \frac{Q_{\text{ост}\left(\xi_{1,n_1},\xi_{2,n_2},\cdots,\xi_{m,n_m}\right)}}{N-m}.$$

Проверка гипотезы об отсутствии влияния фактора А на результативный признак

$$H_0: \sigma^2_{факт} = \sigma^2_{ост} .$$

$$F(\xi_{1,n_1},\dots,\xi_{m,n_m}) = \frac{\frac{1}{m-1}Q_{факт}(\xi_{1,n_1},\dots,\xi_{m,n_m})}{\frac{1}{n-m}Q_{ост}(\xi_{1,n_1},\dots,\xi_{m,n_m})} = \frac{S^{\wedge 2}_{факт}(\xi_{1,n_1},\xi_{2,n_2},\dots\xi_{m,n_m})}{S^{\wedge 2}_{ост}(\xi_{1,n_1},\xi_{2,n_2},\dots\xi_{m,n_m})} \sim F(m-1;n-m)$$

Проверка гипотезы о равенстве двух средних выбранных уровней

Если влияние фактора доказано, то можно проверить гипотезы:

$H_0: a_j = a_{j'}$

$H_1: a_j \neq a_{j'}$

Для проверки нулевой гипотезы строится статистика:

$$F = \frac{\dfrac{(y_{*j} - y_{*j'})^2}{1}}{\dfrac{Q_{ост}}{N-m}} \cdot \frac{n_j n_{j'}}{n_j + n_{j'}},$$

распределенная по закону Фишера-Снедекора с $\nu_1 = 1$ и $\nu_2 = N - m$ степенями свободы.

Проверка гипотезы о значении уровня фактора

При проверке гипотезы $H_0: a = a_0$ используется:

- в случае модели M_1 статистика: $F = \dfrac{\dfrac{N(y_{**} - a_0)^2}{1}}{\dfrac{Q_{ост}}{N-m}}$, имеющая F – распределение с $\nu_1 = 1$ и $\nu_2 = N - m$ степенями свободы;

- в случае модели M_2 и $n_j = n$ статистика: $F = \dfrac{\dfrac{N(y_{**} - a_0)^2}{1}}{\dfrac{Q_{факт}}{m-1}}$, имеющая F – распределение с $\nu_1 = 1$ и $\nu_2 = m-1$ степенями свободы.

Точечная и интервальная оценка дисперсий

Несмещенную точечную оценку для факторной дисперсии можно уточнить:

$$\hat{S}^2_{\substack{факт \\ уточ}} = (\hat{S}^2_{факт} - \hat{S}^2_{ост}) \cdot \frac{N(m-1)}{N^2 - \sum n_j^2}.$$

Интервальная оценка для $D(\varepsilon_{ij}) = \sigma^2$ с надежностью γ имеет вид:

$$\frac{Q_{ост}}{x^2(\frac{1-\gamma}{2}, N-m)} \le \sigma^2 \le \frac{Q_{ост}}{x^2(\frac{1+\gamma}{2}, N-m)}.$$

НЕПАРАМЕТРИЧЕСКИЙ ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Критерий Краскела-Уоллиса проверяет однородность распределения k случайных величин при альтернативной гипотезе сдвига. Критерий Краскела–Уоллиса

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(n+1),$$

где $n = \sum_{i=1}^{k} n_i$, R_i – сумма рангов i-ой выборки, $i = \overline{1,\ldots,k}$, при справедливости нулевой гипотезы и $n_i \geq 5$ и $k \geq 4$ имеет приблизительно распределение «Хи-квадрат» с числом степеней свободы $\nu = k - 1$.

Медианный тест обладает меньшей мощностью и основан на подсчете числа наблюдений каждой выборки, которые попадают выше или ниже общей медианы выборок, и вычисляет затем значение статистики «Хи-квадрат» для таблицы сопряженности $2 \times k$, где k – число рассматриваемых случайных величин.

Двухфакторный дисперсионный анализ

Постановка задачи

Необходимо исследовать влияние двух факторов A и B на результативный нормально распределенный признак Y.

$A_i,\ i=\overline{1,m};\ B_j,\ j=1,l$ - уровни факторов.

При этом возможны два случая:

1. каждой паре уровней факторов A_i и B_j соответствует одно наблюдаемое значение результативного признака y_{ij}.
2. для каждой пары уровней A_i и B_j имеется $n(n>1)$ наблюдений y_{ijk}.

Модель двухфакторного дисперсионного анализа (случай I)

Пусть каждой паре уровней факторов A_i и B_j соответствует одно наблюдаемое значение результативного признака y_{ij}, то есть наблюденные значение можно представить в виде таблицы с двумя входами:

A_i \ B_j	B_1	B_2	…	B_l
A_1	y_{11}	y_{12}	…	y_{1l}
A_2	y_{21}	y_{22}	…	y_{2l}
…	…	…	…	…
A_m	y_{m1}	y_{m2}	…	y_{ml}

Апостериорная модель дисперсионного анализа будем рассматривать в виде:

$$y_{ij} = \mu + \alpha_i + \beta_j + z_{ij} i = \overline{1,m}, j = \overline{1,l}$$

Априорная модель:

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

где a – общая генеральная средняя;

ε_{ij} - независимые нормально распределенные остатки, с $M\varepsilon_{ij} = 0$ и $D\varepsilon_{ij} = \sigma^2$, $i = \overline{1,m}$; $j = \overline{1,l}$;

α_i, β_j - отклонения от a, обусловленные влиянием соответствующих уровней факторов A и B.

Если уровни факторов A_i и B_j фиксированные (модель M_1), то α_i и β_j есть неслучайные величины, удовлетворяющие очевидным условиям

$$\sum_{i=1}^{m} \alpha_i = 0; \ \sum_{j=1}^{l} \beta_j = 0.$$

Формулировка гипотез об отсутствии влияния факторов на результативный признак

Если уровни факторов A_i и B_j фиксированные **(модель** M_1**),** то α_i и β_j есть неслучайные величины, удовлетворяющие очевидным условиям

$$\sum_{i=1}^{m}\alpha_i = 0;\ \sum_{j=1}^{l}\beta_j = 0.$$

Отсутствие влияния уровней факторов на изменения результативного признака - нулевые гипотезы - формулируются в виде:

$$H_0:\ \alpha_i = 0,\ i = \overline{1,m};$$
$$H_0:\ \beta_j = 0,\ j = \overline{1,l}.$$

Если уровни факторов A_i и B_j случайные **(модель M_2)**, то α_i и β_j считают независимыми между собой и с ε_{ij} случайными величинами распределенными нормально с $M\alpha_j = M\beta_j = 0$ и $D\alpha_i = \sigma_\alpha^2;\ D\beta_j = \sigma_\beta^2$.

H_0: $\sigma_\alpha^2 = 0$;
H_0: $\sigma_\beta^2 = 0$.

Если уровни фактора A – случайные, а B – фиксированные (**смешанная модель**), то α_i независимые между собой и с ε_{ij} случайные величины с $M\alpha_j = 0,\ D\alpha_i = \sigma_\alpha^2;\ \beta_j$ - неслучайные величины, удовлетворяющие условию $\sum \beta_j = 0$.

H_0: $\sigma_\alpha^2 = 0$;
H_0: $\beta_j = 0,\ j = \overline{1,l}$.

Аналогично строиться смешанная модель, в которой фактор A имеет фиксированные уровни, а фактор B – случайные.

Разложение дисперсии

$$Q_{общ} = Q_A + Q_B + Q_{ост},$$

где $$Q_A = l\sum_{i=1}^{m}(y_{i*} - y_{**})^2;$$

$$Q_B = m\sum_{j=1}^{l}(y_{*j} - y_{**})^2;$$

$$Q_{ост} = \sum_{i=1}^{m}\sum_{j=1}^{l}(y_{ij} - y_{*j} - y_{i*} + y_{**})^2$$

Проверка гипотезы об отсутствии влияния факторов на результативный признак

Для проверки нулевой гипотезы об отсутствии влияния одного из факторов $D \in \{A; B\}$ рассматривается статистика:

$$F = \frac{\frac{Q_D}{n_D - 1}}{\frac{Q_{ост}}{N - n_D}}, \text{ где } n_D = \begin{cases} m, D = A \\ l, D = B \end{cases}$$

распределенная по закону Фишера-Снедекора с $\nu_1 = n_D - 1$ и $\nu_2 = N - n_D$ степенями свободы.

Модель двухфакторного дисперсионного анализа (случай II)

В общем случае, когда для каждой пары уровней A_i и B_j имеется $n(n>1)$ наблюдений, модель дисперсионного анализа представляется в виде:

$$y_{ijk} = a + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad i = \overline{1,m},\ j = \overline{1,l},\ k = \overline{1,n},$$

где y_{ijk} - $к$-ое наблюдение результативного признака для i-го уровня фактора A и j-го уровня фактора B;
a – общая генеральная средняя;
α_i, β_j - отклонения от a, обусловленные влиянием соответствующих уровней A_i и B_j;
$(\alpha\beta)_{ij}$ - отклонения от a, обусловленные совместным влиянием уровней факторов A и B;
$\varepsilon_{ijk} \in (0,\sigma)$ и независимы между собой.

Формулировка гипотез об отсутствии влияния факторов на результативный признак

Если уровни факторов A_i и B_j фиксированные (**модель** M_1), то отклонения α_i, β_j и $(\alpha\beta)_{ij}$ - неслучайные величины, удовлетворяющие условиям:

$$\sum_{i=1}^{m}\alpha_i = 0;\ \sum_{j=1}^{l}\beta_j = 0;\ \sum_{i=1}^{m}(\alpha\beta)_{ij} = 0;\ \sum_{j=1}^{l}(\alpha\beta)_{ij} = 0.$$

Нулевые гипотезы об отсутствии влияния:

фактора A – H_0: $\alpha_i = 0;\ i = \overline{1,m}$;

фактора B – H_0: $\beta_j = 0;\ j = \overline{1,l}$;

совместного влияния факторов A и B – H_0: $(\alpha\beta)_{ij} = 0; i = \overline{1,m};\ j = \overline{1,l}$.

В случае **модели** M_2 α_i, β_j и $(\alpha\beta)_{ij}$ есть независимые между собой и с ε_{ijk} случайные величины, распределенные нормально с нулевым математическим ожиданием и с дисперсиями σ_α^2, σ_β^2 и $\sigma_{\alpha\beta}^2$.

Нулевые гипотезы от отсутствии влияния:

фактора A – H_0: $\sigma_\alpha^2 = 0$;

фактора B – H_0: $\sigma_\beta^2 = 0$;

совместного влияния факторов A и B – H_0: $\sigma_{\alpha\beta}^2 = 0$.

Для смешанной модели, когда, к примеру, уровни фактора A случайные, а фактора B – фиксированные, отклонения α_i и $(\alpha\beta)_{ij}$ независимые между собой и с ε_{ijk} нормально распределены случайные величины с нулевыми математическими ожиданиями, с дисперсиями σ_α^2 и $\sigma_{\alpha\beta}^2$, при этом $\sum_{i=1}^{m}(\alpha\beta)_{ij} \neq 0$, а $\sum_{j=1}^{l}(\alpha\beta)_{ij} = 0;\ \sum_{j=1}^{l}\beta_j = 0$.

Нулевые гипотезы об отсутствии влияния факторов имеют вид:

фактора $A - H_0$: $\sigma_\alpha^2 = 0$;

фактора $B - H_0$: $\beta_j = 0;\ j = \overline{1,l}$;

совместного влияния факторов A и $B - H_0$: $\sigma_{\alpha\beta}^2 = 0$.`

Аналогично строится другая смешанная модель.

Разложение дисперсии

$$Q_{общ} = Q_A + Q_B + Q_{AB} + Q_{ост},$$

где

$$Q_{общ} = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(y_{ijk} - y_{***}\right)^2;$$

$$Q_A = l \cdot n \cdot \sum_{i=1}^{m}\left(y_{i**} - y_{***}\right)^2;$$

$$Q_B = m \cdot n \cdot \sum_{j=1}^{l}\left(y_{*j*} - y_{***}\right)^2;$$

$$Q_{AB} = n \cdot \sum_{i=1}^{m}\sum_{j=1}^{l}\left(y_{ij*} - y_{i**} - y_{*j*} + y_{***}\right)^2;$$

$$Q_{ост} = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(y_{ijk} - y_{ij*}\right)^2.$$

Проверка гипотезы об отсутствии влияния факторов на результативный признак

Вариации	Сумма квадратов	Число степеней свободы	Несмещенные оценки дисперсий	M_1	M_2	Смешанная модель	
						A – фиксир. B – случ.	A – случ. B – фиксир.
				$F_{набл.}$	$F_{набл.}$	$F_{набл.}$	$F_{набл.}$
A	Q_A	$m-1$	1. $Q_A/(m-1)$	1:4	1:3	1:4	1:3
B	Q_B	$l-1$	2. $Q_B/(l-1)$	2:4	2:3	2:3	2:4
AB	Q_{AB}	$(m-1)(l-1)$	3. $Q_{AB}/(m-1)(l-1)$	3:4	3:4	3:4	3:4
$Остат.$	$Q_{ост}$	$ml(n-1)$	4. $Q_{ост}/(ml(n-1))$				

НЕПАРАМЕТРИЧЕСКИЙ ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Для проверки однородности распределения $k > 2$ зависимых совокупностей следует использовать непараметрические альтернативы двухфакторного дисперсионного анализа, например, критерий Фридмана:

$$F = \frac{12}{kn(n+1)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_{ij} \right)^2 - 3k(n+1),$$

где R_{ij} – ранг i-го объекта по j-му признаку.

Критерий Фридмана при справедливости нулевой гипотезы аппроксимируется распределением «Хи-квадрат» с числом степеней свободы n-1.