Дисперсионный анализ

Дисперсионный анализ

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящих от различных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния. Дисперсионный анализ находит применение в различных областях науки и техники.

Известно, что многие признаки и свойства живых организмов находятся под влиянием различных факторов: наследственности, условий среды, внутренних факторов организма, искусственного отбора. Степень и направленность воздействия различных факторов неодинаковы, поэтому важно определить долю влияния отдельных факторов на изменчивость признака. Для решения подобной задачи используют метод дисперсионного анализа, разработанный Р.Фишером.

Сущность дисперсионного анализа состоит в установлении роли отдельных факторов в изменчивости признака.

В зависимости от количества изучаемых факторов различают однофакторный и многофакторный дисперсионный анализ. Рассмотрим подробнее метод однофакторного дисперсионного анализа.

Однофакторный дисперсионный анализ

Предположим, что имеется K выборок с объемами

$n_1, n_2, \ldots, n_k$ $N=n_1+n_2+\ldots+n_k$, и наблюдения можно представить в виде $x_{ij} = a_j + \varepsilon_{ij}$

где i - номер наблюдения в выборке; j - номер выборки;

a_j - групповые математические ожидания;

ε_{ij} - случайные ошибки с $M(\varepsilon_{ij}) = 0$, о которых предполагается, что они независимы и одинаково расположены.

Подобная ситуация возникает, когда существует некий фактор, принимающий различных значений (называемых уровнями), и каждая группа объектов, чьи признаки мы примеряем, подвергается воздействию определенного уровня этого фактора. Методы математической статистики, изучающие воздействие одного фактора на объекты и их признаки, называют в совокупности однофакторным анализом.

Предполагается, что ошибки нормально распределены: $\varepsilon_{ij} \in N(0, \sigma^2)$

Тогда можно изучать влияние фактора, вычисляя дисперсии некоторых величин. Совокупность этих методов называют однофакторным дисперсионным анализом.

Основной гипотезой, нуждающейся в проверке, является гипотеза о равенстве групповых средних $H_0: a_1 = a_2 = \ldots = a_k$. Иными словами, проверяют гипотезу о том, что фактор вообще не влияет на наблюдения. В случае нормальных ошибок ее можно проверить, вычислив две разные оценки дисперсии.

Рассмотрим группу экспериментальных животных, подвергнутых ультрафиолетовому облучению. В процессе эксперимента измерялась температура тела животных. Результаты измерений были занесены в таблицу:

Температура тела животных

№ испытания	Уровень фактора А (мощность ультрафиолетового облучения)		
	A1	A2	A3
1	37,4	37,8	38,0
2	37,3	37,9	37,9
3	37,0	37,5	38,4
4	36,6	37,4	38,3
$\overline{x}_j$	37,15	37,65	38,15

Физический фактор А (ультрафиолетовое излучение) имеет $m = 3$ постоянных уровней (3 различных мощности облучения). На всех уровнях распределения случайной величины X (температуры тела животного) предполагается нормальным, а дисперсии одинаковыми, хотя и неизвестными.

В данном эксперименте число проведенных наблюдений при действии каждого из уровней фактора одинаково.

Все значения величины X, наблюдаемые при каждом фиксированном уровне фактора A_j, составляют группу, и в последней строке таблицы представлены соответствующие выборочные групповые средние, вычисленные по формуле

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Здесь n – число испытаний, j – номер столбца, i - номер строки, в которой расположено данное значение случайной величины. Общая средняя арифметическая всех nm наблюдений находится как

$$\overline{x} = \frac{1}{m}\sum_{j=1}^{m} \overline{x}_j$$

Факторная сумма

Факторная сумма квадратов отклонений групповых средних от общей средней $\bar{x}$, которая характеризует рассеивание «между группами» (т.е. рассеивание за счет исследуемого фактора):

$$S_{\text{факт}} = n\sum_{j=1}^{m}(\bar{x}_j - \bar{x})^2$$

Остаточная сумма

Остаточная сумма квадратов отклонений наблюдаемых значений группы от своей групповой средней $\bar{x}_j$, которая характеризует рассеивание «внутри групп» (за счет случайных причин):

$$S_{\text{ост}} = \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2 + \sum_{i=1}^{n}\left(x_{i2} - \bar{x}_2\right)^2 + \ldots + \sum_{i=1}^{n}\left(x_{il} - \bar{x}_l\right)^2$$

Общая сумма

Общая сумма квадратов отклонений наблюдаемых значений от общей средней $\bar{x}$:

$$S_{\text{общ}} = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(x_{ij} - \bar{x} \right)^2$$

Можно доказать следующее равенство:

$$S_{\text{общ}} = S_{\text{факт}} + S_{\text{ост}}$$

С помощью $S_{\text{общ}}$, $S_{\text{факт}}$, $S_{\text{ост}}$ производится оценка общей, факторной и остаточной дисперсий:

$$S^2_{\text{общ}} = \frac{1}{mn-1} S_{\text{общ}}$$

$$S^2_{\text{факт}} = \frac{1}{m-1} S_{\text{факт}}$$

$$S^2_{\text{ост}} = \frac{1}{m(n-1)} S_{\text{ост}}$$

В основе однофакторного дисперсионного анализа лежит тесная связь между различием в групповых средних $\bar{x}_j$ и соотношением между двумя видами дисперсий – факторной, которая характеризует влияние фактора А на величину Х, и остаточной, которая характеризует влияние случайных причин. Сравнивая факторную дисперсию с остаточной по величине их отношения судят, насколько сильно проявляется влияние фактора.

Показатель критерия Фишера

Для сравнения двух дисперсий используют **показатель критерия Фишера**

$$F_{\text{эксп}} = S^2_{\text{факт}} / S^2_{\text{ост}}$$

При этом при заданном уровне значимости проверяют нулевую гипотезу о равенстве факторной и остаточной дисперсии (изучаемый фактор не вызывает изменчивости признака) при конкурирующей гипотезе об их неравенстве (изучаемый фактор вызывает изменчивость признака).

По таблице критических значений распределения Фишера-Снедекора при уровне значимости, равном половине заданного уровня α, находят критическое значение $F_{kp}(\alpha/2; k_1; k_2)$. Здесь $k_1 = m - 1$; $k_2 = n(m-1)$

Если $F_{\text{эксп}} < F_{kp}$, нулевую гипотезу считают согласующейся с результатами наблюдений. Если $F_{\text{эксп}} > F_{kp}$, то эту гипотезу отвергают в пользу конкурирующей.

Замечание. Если окажется, что $S^2_{факт} < S^2_{ост}$, следует сделать вывод об отсутствии влияния фактора А на Х.

Если проверка покажет значимость различий между $S^2_{факт}$ и $S^2_{ост}$,следует сделать вывод о существенном влиянии фактора А на Х.

Пример

Имеются данные о настриге шерсти овец в зависимости от их живой массы (табл. 2).

Требуется определить достоверность разницы в настриге шерсти овец в зависимости от их живой массы с уровнем вероятности суждения 0,05.

Для расчета показателей вариации настриг шерсти овец возведем в квадрат (табл. 3).

Таблица 2
Настриг шерсти овец, кг

Живая масса овец, кг	Овцы								Сумма	Численность групп	Средний квадрат суммы
	1	2	3	4	5	6	7	8			
	x_{ij}								$\sum_{j=1}^{n_i} x_{ij}$	n_i	$\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i}$
До 52,4	5,7	6,3	5,9	6,5	6,1	–	–	–	30,5	5	186,05
52,5–54,6	5,7	6,3	6,9	6,8	6,4	6,1	6,3	6,7	51,2	8	327,68
Свыше 54,6	7,2	7,1	6,9	6,3	7,0	6,8	7,1	–	48,4	7	334,65
Итого	×	×	×	×	×	×	×	×	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$ = 130,1	$\sum_{i=1}^{k} n_i$ = N = = 20	$\sum_{i=1}^{k}\left(\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i}\right)$ = = 848,38

Таблица 3

Квадрат настрига шерсти овец

Живая масса овец, кг	Овцы								Сумма квадратов
	1	2	3	4	5	6	7	8	
	x_{ij}^2								$\sum_{j-1}^{n_i} x_{ij}^2$
До 52,4	32,46	39,69	34,81	42,25	37,21	–	–	–	186,45
52,5–54,6	32,49	39,69	47,61	46,24	40,96	37,21	39,69	44,89	328,78
Свыше 54,6	51,84	50,41	47,61	39,69	49,00	46,24	50,41	–	335,2
Итого	×	×	×	×	×	×	×	×	$\sum_{i-1}^{k}\sum_{j-1}^{n_i} x_{ij}^2 =$ $= 850,43$

Показатели вариации будут равны:

общая вариация:

$$w_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2}{N} = 850{,}43 - \frac{130{,}1^2}{20} = 4{,}1295$$

групповая вариация:

$$w_{гр} = \sum_{i=1}^{k}\left(\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i}\right) - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2}{N} = 848{,}38 - \frac{130{,}1^2}{20} = 2{,}0809$$

Остаточная вариация:

$$w_{ост} = w_0 - w_{гр} = 4{,}1295 - 2{,}0809 = 2{,}0486$$

Рассчитаем число степеней свободы вариации:

общей: $v_0 = N - 1 = 20 - 1 = 19$

групповой: $v_{гр} = k - 1 = 3 - 1 = 2$

остаточной вариации: $v_{ост} = N - k = 20 - 3 = 17$

Отсюда дисперсии будут равны:

общая: $S_0^2 = \frac{w_0}{v_0} = \frac{4{,}1295}{19} = 0{,}2173$

групповая: $S_{гр}^2 = \frac{w_{гр}}{v_{гр}} = \frac{2{,}0809}{2} = 1{,}0405$

остаточная: $S_{ост}^2 = \frac{w_{ост}}{v_{ост}} = \frac{2{,}0486}{17} = 0{,}1205$

Фактическое значение *F*-критерия для групповой и остаточной дисперсий:

$$F_{факт} = \frac{S_{гр}^2}{S_{ост}^2} = \frac{2{,}0809}{0{,}1205} = 8{,}63$$

Табличное значение *F*-критерия при уровне значимости 0,05, 2 степенях свободы для групповой дисперсии и 17 степенях свободы для остаточной дисперсии равно 3,59 (таблица «Значение *F*-критерия Фишера при уровне значимости 0,05».

Результаты дисперсионного анализа представлены в табл.4.

Однофакторный дисперсионный анализ

Источники вариации	Вариация (сумма квадратов отклонений)	Степень свободы вариации	Дисперсия	Отношение дисперсий	
				фактическое	табличное
Групповая	4,1295	19	0,2173	8,63	3,59
Остаточная	2,0809	2	1,0405	1	x
Общая	2,0486	17	0,1205	x	x

Данные таблицы показывают, что фактическое отношение дисперсий больше табличного, следовательно, разница в среднем настриге шерсти по группам овец с различной живой массой достоверна. Живая масса овец оказывает влияние на их настриг шерсти.