Дисперсионный анализ.

Общая идея дисперсионного анализа.

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящий от разных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния.

Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение этих дисперсий позволяет оценить существенность влияния факторов на исследуемую величину.

Если исследовать влияние одного фактора на исследуемую величину, то речь идет об *однофакторном анализе*. Если изучается влияние двух факторов, то речь идет о *двухфакторной* анализе.

Рассмотрим вариант параметрического дисперсионного анализа. Если данные не имеют нормального распределения, то целесообразно использовать непараметрический критерий Краскела-Уоллиса.

Однофакторный дисперсионный анализ.

Предположим, что совокупности случайных величин имеют нормальное распределение и равные дисперсии. Пусть имеется m таких совокупностей, из которых произведены выборки объемом $n_1,n_2,\ldots,n_i,\ldots n_m$. Обозначим выборку из i-ой совокупности $(x_{i1},x_{i2},\ldots,x_{in})$. Тогда все выборки можно записать в виде следующей таблице, которая называется *матрицей наблюдений*.

Таблица 1.

Количество элементов совокупности (n) / Количество совокупностей (m)	1	2	…	j	…	n
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n1}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n2}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ini}
…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mnm}

Средние значения этих выборок обозначим через β_1, $\beta_2,\ldots,\beta_i$, $\ldots,\beta_m$. Теперь проверим гипотезу о равенстве этих средних (или о том, что фактор не влияет на исследуемую величину). Нулевую гипотезу запишем так:

H_0: $\beta_1=\beta_2=\ldots=\beta_i=\ldots=\beta_m$;

альтернативную - в виде

H_0: $\beta_1\neq\beta_2\neq\ldots\neq\beta_i\neq\ldots\neq\beta_m$.

Гипотеза H_0 проверяется сравнением внутригрупповых и межгрупповых дисперсий по F-критерию. Если расхождение между ними значительно, то нулевая гипотеза

принимается. В противном случае гипотеза о равенстве средних отвергается и делается заключение о том, что различие в средних обусловлено не только случайностями выборок, но и действием исследуемого фактора.

Рассмотрим структуру и межгрупповой и внутригрупповой дисперсии и способ их вычисления. Вернемся к таблице 1. Найдем средние арифметические членов каждой совокупности. Для первой совокупности обозначим среднюю арифметическую через $\bar{x}_{1*}$, для i- й – через $\bar{x}_{i*}$. Тогда имеем

$$\bar{x}_{1*} = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}; \ldots \bar{x}_{ij} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}; \ldots \bar{x}_{mj} = \frac{\sum_{j=1}^{n_m} x_{mj}}{n_m}.$$

Общую среднюю арифметическую всех m совокупностей обозначим через

$$\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}, \ldots или \ldots \bar{x} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_{i*}.$$

Найдем сумму квадратов отклонений x_{ij} от $\bar{x}$, т.е. $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$. Представим ее в виде

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{\bar{x}}_{i*} + \bar{x}_{i*} - \bar{x})^2 =$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*} - \bar{x})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})(\bar{x}_{i*} - \bar{x}), \quad (1)$$

причем

$$S = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})(\bar{x}_{i*} - \bar{x}) = \sum_{ji=1}^{n}(x_{ij} - \bar{x}_{i*})\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x}).$$

Но $\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*}) = 0$, так как это есть сумма отклонений переменных одной совокупности от средней арифметической этой же совокупности, т.е. S=0. Второй член суммы (1) запишем в виде

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*} - \bar{x})^2 = n\sum(\bar{x}_{i*} - \bar{x})^2.$$

Тогда основное тождество (1) можно представить следующим образом:

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2}_{Q} = \underbrace{n\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x})^2}_{Q_1} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2}_{Q_2},$$

или $Q = Q_1 + Q_2$.

Слагаемое Q_1 является суммой квадратов разностей между средними отдельных совокупностей и общей средней всей совокупности наблюдений; эта сумма называется *суммой квадратов отклонений между группами* и характеризует систематическое расхождение между совокупностями наблюдений. Величину Q1 называют иногда *рассеиванием по факторам* (т.е. за счет исследуемого фактора).

Слагаемое Q_2 представляет собой сумму квадратов разностей между отдельными наблюдениями и средней соответствующей совокупности; эта сумма *называется суммой квадратов отклонений внутри группы*. Она характеризует *остаточное рассеивание* случайной погрешности совокупностей.

Наконец Q называется *общей* или *полной суммой* квадратов отклонений отдельных наблюдений от общей средней $\bar{x}$.

Оценим дисперсии s_1^2, s_2^2, s^2:

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x})^2 = \frac{Q_1}{m-1},$$

$$s_2^2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2 = \frac{Q_2}{m(n-1)},$$

$$s^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2.$$

Затем произведем оценку различия между дисперсиями по F-критерию:

$$F = \frac{Q_1 / (m-1)}{Q_2 / m(n-1)}. \qquad (2)$$

Этот критерий подчиняется F-распределению с $k_1=(m-1)$ и $k_2=m(n-1)$ степенями свободы. Выбирая уровень значимости α, найдем по **таблице** (она прилагается к учебнику в приложении и называется «F-распределение для уровня значимости α=0,01» и другая таблица для α=0,05) соответствующий предел так, чтобы

$$P(F>Fa)=\alpha$$

Сравнивая между собой межгрупповую и остаточную дисперсии, по величине их отношения судят, насколько сильно проявляется влияние факторов (в этом сравнении и заключается основная идея дисперсионного анализа).

Однофакторный дисперсионный анализ удобно представить в виде таблице.

Таблица 2.

Компоненты дисперсии	Сумма квадратов	Число степеней свободы ,k	Средний квадрат	Оценка дисперсий
Межгрупповая	$\sum_i (\bar{x}_{i*} - \bar{x})^2$	m-1	$\frac{1}{m-1}\sum_i (\bar{x}_{i*} - \bar{x})^2$	s_1^2
Внутригрупповая	$\sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	m(n-1)	$\frac{1}{m(n-1)}\sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	s_2^2
Полная (общая)	$\sum_{ij} (x_{ij} - \bar{x})^2$	mn-1	$\frac{1}{mn-1}\sum_{ij} (x_{ij} - \bar{x})^2$	s^2

Многофакторный дисперсионный анализ.

Если исследуют действие двух, трех и т.д. факторов, то структура дисперсионного анализа та же, что и при однофакторном анализе, усложняются лишь вычисления. Рассмотрим задачу оценки действия двух одновременно действующих факторов. Предположим, что имеем несколько однотипных приборов и несколько видов материалов. Требуется выяснить, значимо ли влияние различных приборов и качество материала в партиях на качество результата исследования. Пусть фактор А – влияние приборов; фактор В – влияние качества материалов в партиях. Через x_{ij} обозначим результат исследования. Для простоты рассмотрим сначала случай, когда для каждого прибора и каждой партии материала проводится лишь одно исследование (наблюдение). Тогда матрицу наблюдений можно записать в виде таблицы.

Таблица 3.

Партия материалов (j) / Приборы (i)	B_1	B_2	…	B_j	…	B_v	$\bar{x}_{i*}$
A1	x_{11}	x_{12}	…	x_{1j}	…	x_{1v}	$\bar{x}_{1*}$
A2	x_{21}	x_{22}	…	x_{2j}	…	x_{2v}	$\bar{x}_{2*}$
…	…	…	…	…	…	…	…
Ai	x_{i1}	x_{i2}	…	x_{ij}	…	x_{iv}	$\bar{x}_{i*}$
…	…	…	…	…	…	…	…
Ar	x_{r1}	x_{r2}	…	x_{rj}	…	x_{rv}	$\bar{x}_{r*}$
$\bar{x}_{*j}$	$\bar{x}_{*1}$	$\bar{x}_{*2}$	…	$\bar{x}_{*j}$	…	$\bar{x}_{*v}$	$\bar{x}$

Пусть имеется $r(A_1,A_2,\ldots,A_i,\ldots,A_r)$ приборов. В матрице наблюдений им соответствуют r строк, которые назовем уровнями фактора А. Имеем $v(B_1,B_2,\ldots,B_j,\ldots,B_v)$ партия материалов. В матрице им соответствуют v столбцов, которые назовем уровнями фактора В.

Пересечением i-го и j-го уровней образует ij-ячейку, в которую записываются наблюдения, полученные при одновременном исследовании факторов А и В на i-ом и j-ом уровнях соответственно.

По каждому столбцу и строке вычислим среднее значение, а также общее среднее. Следующие величины вычисляем

$$\bar{x}_{i*}=\frac{1}{v}\sum_{j=1}^{v}x_{ij},\ldots\ldots\ldots\bar{x}_{*j}=\frac{1}{r}\sum_{i=1}^{r}x_{ij},\ldots\ldots\ldots\bar{x}=\frac{1}{rv}\sum_{i=1}^{r}\sum_{j=1}^{v}x_{ij}.$$

Основное тождество однофакторного анализа в данном случае принимает вид:

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2=\textit{прибавляем..и..отнимаем..}\bar{x}_{i*},\bar{x}_{*j},x=$$

$$=v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2+r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2+\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2=Q_1+Q_2+Q_3. \quad (1)$$

Слагаемое Q_1 представляет собой сумму квадратов разностей между средними по строкам и общим средним и характеризует изменение признака по фактору А. Слагаемое Q_2 представляет собой сумму квадратов разностей между средними по столбцам и общим средним и характеризует изменение признака по фактору В. Слагаемое Q_3 называется *остаточной* суммой квадратов и характеризует влияние неучтенных факторов. Сумма Q

называется *общей или полной* суммой квадратов отклонений отдельных наблюдений от общей средней.

Произведем оценку дисперсий:

$$s^2 = \frac{1}{rv-1}\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2 = \frac{Q}{rv-1},$$

$$s_1^2 = \frac{1}{r-1}v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2 = \frac{Q_1}{r-1},$$

$$s_2^2 = \frac{1}{v-1}r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2 = \frac{Q_2}{v-1},$$

$$s_3^2 = \frac{1}{(r-1)(v-1)}\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2 = \frac{Q_3}{(r-1)(v-1)}.$$

В двухфакторном анализе для выяснения значимости влияния факторов А и В на исследуемый признак сравнивают дисперсии по факторам с остаточной дисперсией, т.е. оценивают s_1^2 / s_3^2 и s_2^2 / s_3^2.

Известно, что если случайная величина распределена нормально, то отношение выборочных дисперсий имеет F-распределение. Поскольку нормальный закон распределения случайной величины является предпосылкой дисперсионного анализа, то имеем

$$F_A = \frac{Q_1/(r-1)}{Q_3/(r-1)(v-1)} = \frac{s_1^2}{s_3^2},$$

$$F_B = \frac{Q_2/(v-1)}{Q_3/(r-1)(v-1)} = \frac{s_2^2}{s_3^2}.$$

Полученные значения F_A и F_B сравнивают с табличными значениями при выбранном уровне значимости α. При $F_A<F_\alpha$ и $F_B<F_\alpha$ нулевая гипотеза о равенстве средних не отвергается, т.е. влияние факторов А и В на исследуемый признак незначимо.

Двухфакторный дисперсионный анализ удобно представить в виде таблицы.

Таблица 4.

Компонента дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсий
Между средними по строкам	$Q_1 = v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2$	r-1	s_1^2
Между средними по столбцам	$Q_2 = r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2$	v-1	s_2^2
Остаточная	$Q_3 = \sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2$	(r-1)(v-1)	s_3^2
Полная (общая)	$Q = \sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2$	rv-1	s^2