Лекция 6

Факторный дисперсионный анализ. Выявление несущественных факторов. Обработка результатов эксперимента на основе регрессии

Автор Джомартова Ш.А.

Дисперсионный анализ

Постановка проблемы

Дисперсионный анализ является статистическим методом анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с целью выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс.

Методами дисперсионного анализа устанавливается наличие влияния заданного фактора на изучаемый процесс (на выходную переменную процесса) за счёт статистической обработки наблюдаемой совокупности выборочных данных.

Классификация методов дисперсионного анализа

По количеству анализируемых признаков

Однофакторный
(ANOVA Analysis of Variance)
(Анализ различий групп по одному признаку)

Многофакторный
(MANOVA)
(Анализ различий групп Одновременно по двум признакам и более)

Классификация методов дисперсионного анализа

По принципам анализа

Параметрический
(Для анализа нормально распределенных признаков в группах)

Непараметрический
(для анализа количественного признака независимо от вида его распределения в группах)

Классификация методов дисперсионного анализа

По анализируемым данным

Данные, полученные в несвязанных (независимых) выборках (в частности данные однократных наблюдении)

Данные, полученные в связанных (зависимых) выборках (в частности данные повторных наблюдений)

Однофакторный анализ

- Дисперсионный анализ, который рассматривает только одну переменную называется **однофакторным дисперсионным анализом** (One-Way ANOVA).



Эксперимент для выполнения однофакторного дисперсионного анадиза

- Введем обозначения:
- m - число вариантов фактора;
- n - число измерений при каждом варианте;
- a_{ij} - результат каждого измерения;
- $i = \overline{1,m}$ - номер варианта фактора;
- $j = \overline{1,n}$ -номер измерения.
- Будем исходить из следующей нулевой гипотезы:
  - наблюдения каждого варианта независимы;
  - наблюдения каждого варианта имеют нормальное распределение;
  - имеют одинаковую дисперсию σ^2:

Результаты эксперимента

№ варианта	Номер измерения						Средние значения
	1	2	…	j	…	n	
1	a_{11}	a_{12}	…	a_{1j}	…	a_{1n}	$\overline{a_1} = \frac{1}{n}\sum_{j=1}^{N} a_{1j}$
2	a_{21}	a_{22}	…	a_{2j}	…	a_{2n}	$\overline{a_2} = \frac{1}{n}\sum_{j=1}^{N} a_{2j}$
…	…	…	…	…	…	…	…
i	a_{i1}	a_{i2}	…	a_{ij}	…	a_{in}	$\overline{a_i} = \frac{1}{n}\sum_{j=1}^{N} a_{ij}$
…	…	…	…	…	…	…	…
m	a_{m1}	a_{m2}	…	a_{mj}	…	a_{mn}	$\overline{a_m} = \frac{1}{n}\sum_{j=1}^{N} a_{mj}$

Общее среднее арифметическое по всем $m \cdot n$ измерениям:

$$\bar{a} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}}{m \cdot n}$$

Сумма квадратов отклонений по всем $m \cdot n$ измерений, то есть по данным всего эксперимента:

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \bar{a})^2$$

Эту сумму квадратов отклонений можно разложить на два независимых слагаемых:

$$Q = n \sum_{i=1}^{m} (\bar{a}_i - \bar{a})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \bar{a})^2$$

Обозначим:

m m n

Вычисляем критерий Фишера

$$F = \frac{\frac{1}{m-1}Q_1}{\frac{1}{m(n-1)}Q_2}$$

- Если окажется $F > F_2$,то есть мы попали в область маловероятных значений *F*, то выдвинутая гипотеза *не подтверждается*. А это значит, что варианты фактора не однотипны. Но если $F \leq F_2$,то гипотеза об однородности вариантов фактора подтверждается, конечно, в рамках допустимого риска

Пример

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

Шаг 1. Гипотезы:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$$

$$H_1 : \text{не все средние равны}$$

Шаг 2. Критическая область

- Найдем критическое значение по таблице критических точек распределения Фишера.
- Уровень значимости $\alpha = 0{,}05$.
- Так как $k = 3$ и $N = 19$, то
- числитель $df = k - 1 = 3 - 1 = 2$
- знаменатель $df = N - k = 19 - 3 = 16$
- Критическое значение равно 3,633.
- Критическая область $F > 3{,}633$

Шаг 3. Вычисление статистики F

Шаг 3а. Подсчет средних

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

$\bar{x}_1 = 35{,}8$ $n_1 = 6$

$\bar{x}_2 = 44{,}6$ $n_2 = 7$

$\bar{x}_3 = 35{,}7$ $n_3 = 6$

$\bar{\bar{x}} = 39$

$N = n_1 + n_2 + n_3 = 19$

Шаг 3b. Расчет отклонений

$$SS_b = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2 =$$

$$= 6 \cdot (35{,}8 - 39)^2 + 7 \cdot (44{,}6 - 39)^2 + 6 \cdot (35{,}7 - 39)^2 =$$

$$= 344{,}1$$

$$SS_w = \sum (x - \bar{x}_i)^2 =$$

$$= (24 - 35{,}8)^2 + (27 - 35{,}8)^2 + \ldots + (48 - 35{,}8)^2 + (40 - 35{,}8)^2 +$$

$$+ (59 - 44{,}6)^2 + (35 - 44{,}6)^2 + \ldots + (54 - 44{,}6)^2 + (56 - 44{,}6)^2 +$$

$$+ (34 - 35{,}7)^2 + (29 - 35{,}7)^2 + \ldots + (40 - 35{,}7)^2 + (45 - 35{,}7)^2 =$$

$$= 1669{,}9$$

Шаг 3c. Расчет дисперсий

$$MS_B = \frac{SS_B}{k-1} = \frac{344{,}1}{2} = 172{,}06$$

$$MS_W = \frac{SS_W}{N-k} = \frac{1669{,}9}{16} = 104{,}37$$

Шаг 3d. Расчет статистики

$$F = \frac{MS_B}{MS_W} = \frac{172{,}06}{104{,}37} = 1{,}649$$

Шаг 4-5. Получение выводов, ответ

$$1{,}649 < 3{,}633$$

Полученное значение статистики не попало в критическую область. У нас нет оснований думать, что средние значения отличаются.

Ответ.

- Средний возраст рассматриваемых категорий персонала не различается значимо.

Пример

Заводы	Эксперименты и отклонения, см								
	1	2	3	…	26	27	…	49	50
№ 1	3	2	1	…	4	3	…	1	3
№ 2	2	0	4	…	3	2	…	2	3
№ 3	2	3	3	…	1	0	…	1	5

$$\overline{a_1} = \frac{3+2+1+\cdots+4+3+\cdots+1+3}{50} = 2{,}5 \text{ см},$$

$$\overline{a_2} = \frac{2+0+4+\cdots+3+2+\cdots+2+3}{50} = 2{,}8 \text{ см},$$

$$\overline{a_3} = \frac{2+3+4+\cdots+1+0+\cdots+1+5}{50} = 3{,}1 \text{ см}$$

$$\bar{a} = \frac{2{,}5+2{,}8+3{,}1}{3} = 2{,}8 \text{ см}.$$

$$\frac{1}{m-1}Q_1 = \frac{1}{m-1}n\sum_{i=1}^{n}(\bar{a}_i-\bar{a})^2 = \frac{50}{2}[(2{,}5-2{,}8)^2+(2{,}8-2{,}8)^2+(3{,}1-2{,}8)^2] = 4{,}5\text{см}^2$$

$$\frac{1}{m(n-1)}Q_2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(a_{ij}-\bar{a}_i\right)^2 = 2{,}2\text{см}^2$$

Расчет F -критерия: $F = \frac{1.5}{2.2} \approx 2.1$ $\qquad F_2 = 3.06$.

Выявление несущественных факторов

Несущественный фактор выявляется так.

Выполняются первый эксперимент из *N* наблюдений с учетом проверяемого фактора и второй эксперимент также из *N* наблюдений - без него. В обоих случаях фиксируются отклики *y*. Делается предположение, что обе выборки принадлежат одной генеральной совокупности, то есть, что проверяемый фактор - несущественный (это нулевая гипотеза). Дальнейшие рассуждения должны либо *не опровергнуть* эту гипотезу, либо посчитать ее *недостаточно обоснованной*.

Пример

Исследуется зависимость времени пребывания заявки в системе массового обслуживания от дисциплины выборки заявок из очереди: LIFO или FIFO. Проведены два эксперимента. Первый эксперимент из $N = 100$ наблюдений с дисциплиной FIFO и второй эксперимент также из $N = 100$ наблюдений с дисциплиной LIFO.

$$S_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{y} - y_i)^2}$$

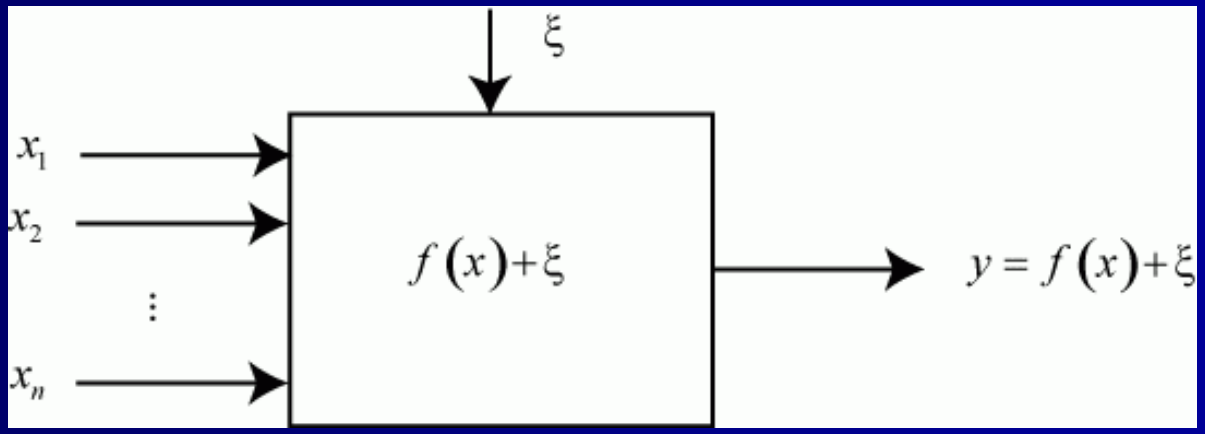
Выдвигается гипотеза о несущественности влияния на время пребывания заявки в системе массового обслуживания изменения дисциплины FIFO на LIFO.

Результаты измерений и вычислений:

$\bar{z} = 1{,}8\text{мин}, S_y = 2\ \text{мин}$;

Обработка результатов эксперимента на основе регрессии

- *Регрессией* понимают зависимость среднего значения какой-либо величины от некоторой другой (других) величины.
- *Регрессионный анализ* - это совокупность методов построения и исследования регрессионной зависимости между величинами (в нашем случае между факторами и откликом) по статистическим данным.



$$\bar{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \sum_{i=0}^{n} \beta_i x_i \, , x_0 = 1$$

для определения выражения *f(x)* надо:

- выбрать степень аппроксимирующего полинома - уравнения регрессии;
- определить коэффициенты регрессии.

Выбор уравнения регрессии

Планирование эксперимента при дисперсионном анализе

Дисперсионный анализ тесно связан с соответствующим планированием эксперимента. Удачно спланированный эксперимент, выявляя все необходимые эффекты, оказывается всегда либо более точным, либо менее трудоемким по сравнению с непродуманным экспериментом.

Если на результат эксперимента действуют одновременно несколько факторов, то наилучший эффект дает одновременный дисперсионный анализ всех этих факторов (многофакторный анализ).

Методы дисперсионного анализа позволяют исследовать и такой случай, когда некоторые сочетания уровней пропущены. Такой эксперимент называется дробным факторным экспериментом (ДФЭ). Планирование при ДФЭ приобретает особо важную роль, ибо пропущенные сочетания уровней не так-то просто нейтрализовать.

Планирование эксперимента при дисперсионном анализе

Такие способы планирования существуют и притом не единственные; согласно Фишеру их называют латинскими квадратами. Эти расположения приводятся в специальных справочниках; для примера приведен один вид такого квадрата:

	A_1	A_2	…	A_{k-1}	A_k
B_1	C_1	C_2	…	C_{k-1}	C_k
B_2	C_2	C_3	…	C_k	C_1
…	…	…	…	…	…
B_{k-1}	C_{k-1}	C_k	…	C_{k-3}	C_{k-2}
B_k	C_k	C_1	…	C_{k-2}	C_{k-1}

Планирование эксперимента при дисперсионном анализе

Схема расчетов для латинского квадрата очень похожа на обычный двухфакторный анализ:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{k} x_{ij}^2$$

Находим сумму квадратов по столбцам, деленную на число наблюдений в столбце:

$$Q_2 = \frac{1}{k}\sum_{i=1}^{k} X_i^2$$

Находим сумму квадратов итогов по строкам, деленную на число наблюдений в строке:

$$Q_3 = \frac{1}{k}\sum_{j=1}^{k} X_j'^2$$

Находим квадрат общего итога, деленный на число всех наблюдений:

$$Q_4 = \frac{1}{k^2}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{k^2}\left(\sum_{j=1}^{k} X_j'\right)^2$$

Находим сумму квадратов итогов по уровням фактора C, деленную на число уровней:

$$Q_5 = \frac{1}{k}\sum_{\nu=1}^{k} Y_\nu^2$$

Планирование эксперимента при дисперсионном анализе

Перейдем теперь к вычислению и оценке значимости дисперсий:

$$S_0^2 = \frac{Q_1 + 2Q_4 - Q_2 - Q_3 - Q_5}{(k-1)(k-2)}$$

$$S_A^2 = \frac{Q_2 - Q_4}{k-1},\ S_B^2 = \frac{Q_3 - Q_4}{k-1}$$

Если отличие будет значимым, то $\frac{S_A^2 - S_0^2}{k} \Rightarrow \sigma_A^2,\ \frac{S_B^2 - S_0^2}{k} \Rightarrow \sigma_B^2$

$$S_C^2 = \frac{Q_5 - Q_4}{k-1}$$

Если отличие будет значимым, то $\frac{S_C^2 - S_0^2}{k} \Rightarrow \sigma_C^2$

Спасибо за внимание