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.

Однофакторный дисперсионный анализ

Предположим, что анализируется влияние на случайную величину X фактора A, изучаемого на k уровнях (A_1, A_2, …, A_k). На каждом уровне A_i проведены n наблюдений (x_{i1}, x_{i2}, …, x_{in}) случайной величины X.

Расположим эксперимен-тальные данные в виде таблицы

Номер наблюдения	Уровни фактора A					
	A_1	A_2	…	A_i	…	A_k
1	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}
….	…	…	…	…	…	…
j	x_{1j}	x_{2j}	…	x_{ij}	…	x_{kj}
…	…	…	…	…	…	…
n	x_{1n}	X_{2n}	…	x_{in}	…	x_{kn}
Σ	X_1	X_2	…	X_i	…	X_n

Однофакторный дисперсионный анализ

Рассмотрим оценки различных дисперсий, возникающие при анализе таблицы результатов наблюдений. Для оценки дисперсии, характеризующей изменение данных на уровне Ai (по строкам таблицы), имеем:

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \overline{x}_i)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right].$$

Из предпосылок дисперсионного анализа следует, что должно иметь место равенство всех дисперсий. При выполнении этого условия находим оценку дисперсии, характеризующей рассеяние значений x_{ij} вне влияния фактора A, по формуле:

$$S_0^2 = \frac{1}{k}\sum_{i=1}^{k} S_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x}_i)^2 = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right]$$

Однофакторный дисперсионный анализ

Для упрощения вычислений приведем алгоритм их выполнения. Вычисляем последовательно суммы:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 \qquad Q_2 = \frac{1}{n}\sum_{i=1}^{k} X_i^2 \qquad Q_3 = \frac{1}{kn}\left(\sum_{i=1}^{k} X_i\right)^2$$

$$S_0^2 = \frac{Q_1 - Q_2}{k(n-1)} \qquad S_A^2 = \frac{Q_2 - Q_3}{k-1}$$

Сравниваем S_A^2 и S_0^2 устанавливаем наличие влияния фактора А.

Если $\dfrac{k(n-1)}{k-1}\dfrac{Q_2 - Q_3}{Q_1 - Q_2} > F_\alpha[k-1; k(n-1)]$, то влияние А – значимо.

Двухфакторный дисперсионный анализ

Рассмотренный ранее однофакторный дисперсионный анализ обладает информативностью, не большей, чем методы множественного сравнения средних. Информативность дисперсионного анализа возрастает при одновременном изучении влияния нескольких факторов.

Рассмотрим случай, когда анализируется влияние одновременно двух факторов А и В.

Двухфакторный дисперсионный анализ

Пусть результаты эксперимента представлены таблицей:

B	Уровни фактора A						Σ
	A_1	A_2	…	A_i	…	A_k	
B_1	x_{11}	x_{21}	...	x_{i1}	...	x_{k1}	X_1'
B_2	x_{12}	x_{22}	...	x_{i2}	...	x_{k2}	X_2'
….	...	...	...	...	...	...	…
B_j	x_{1j}	x_{2j}	...	x_{ij}	...	x_{kj}	X_j'
…	...	...	...	...	...	...	…
B_m	x_{1n}	X_{2n}	...	x_{in}	...	x_{kn}	X_m'
Σ	X_1	X_2	…	X_i	…	X_n	

Двухфакторный дисперсионный анализ

Дисперсионный анализ для двухфакторных таблиц проводится в следующей последовательности. Вычисляются суммы:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 \qquad Q_2 = \frac{1}{m}\sum_{i=1}^{k} X_i^2 \qquad Q_3 = \frac{1}{k}\sum_{j=1}^{m} X_j^2 \qquad Q_4 = \frac{1}{mk}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{mk}\left(\sum_{j=1}^{k} X_{j'}\right)^2$$

Далее находятся оценки дисперсий:

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(m-1)} \qquad S_A^2 = \frac{Q_2 - Q_4}{k-1} \qquad S_B^2 = \frac{Q_3 - Q_4}{m-1}$$

Если $\dfrac{S_A^2}{S_0^2} > F_\alpha(f_1, f_2)$, то влияние фактора А признается значимым.

Если $\dfrac{S_B^2}{S_0^2} > F_\alpha(f_1, f_2)$, то влияние фактора В признается значимым.

Двухфакторный дисперсионный анализ

Приведенный анализ предполагает независимость факторов A и B. Если они зависимы, то взаимодействие факторов C=AB также является фактором, которому соответствует своя дисперсия. Для того чтобы выделить такое взаимодействие, необходимы параллельные наблюдения в каждой клетке таблицы, т.е. при каждом сочетании факторов A и B на уровнях A_i и B_j соответственно необходимо не одно наблюдение, а серия наблюдений.

Для оценки влияния взаимодействия факторов AB вычисляем дополнительную сумму:

$$Q_5 = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{v=1}^{n} x_{ijv}^2$$

Далее анализ проводится, как и ранее, с той лишь разницей, что в клетках таблицы вместо отдельных значений используется их средние значения. Вычисляется оценка дисперсии и проверяется значимость взаимодействия факторов:

$$S_{AB}^2 = \frac{Q_5 - nQ_1}{mk(n-1)} \qquad \frac{nS_0^2}{S_{AB}^2} > F_\alpha(f_1, f_2) \qquad f_1 = (k-1)(m-1) \qquad f_2 = mk(n-1)$$

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 С, 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:

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

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

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

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

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

ВОПРОСЫ ?