Дисперсионный анализ

Дисперсионный анализ (Р. Фишер, 1920 г.) – группа методов математической статистики для анализа результатов наблюдений, зависящих от **нескольких** одновременно действующих факторов.

Идея дисперсионного анализа заключается в **разбиении** общей дисперсии изучаемой случайной величины на независимые составляющие. Каждая из них характеризует влияние того или иного фактора или их взаимодействие, а их **сравнение** позволяет оценить **зна́чимость влияния** факторов на исследуемую величину.

Предположения дисперсионного анализа:

1. Исследуемые факторы стохастически **независимы**. С точки зрения способов отбора информации это означает независимость выборочных результатов наблюдения (отдельных выборок или слоев – они не преобразуются друг в друга с помощью какого-либо алгоритма).

2. Исследуемые факторы, каждый по отдельности, подчиняются **нормальным** законам распределения.

3. **Дисперсии** σ_i^2 исследуемых факторов **однородны** (априори приблизительно одного порядка).

Идею дисперсионного анализа о разбиении дисперсии изучим на примере однофакторного эксперимента по установлению связи выходного фактора системы ($\boldsymbol{\eta}$) с одним входным фактором ($\boldsymbol{\xi}$).

Входной фактор $\boldsymbol{\xi}$ задается своими k уровнями, значения которых в дисперсионном анализе не существенны, важны лишь их номера:
$j = 1, 2,\ldots, k$.

В однофакторном эксперименте при каждом j-ом уровне входного фактора проводится серия замеров выходного фактора. Каждый такой замер имеет номер:
$i = 1, 2,\ldots,$

Тогда результат единичного *i*-го замера выходного фактора ***η*** при ***j*-м уровне** входного фактора (в *j*-й серии наблюдений, группе, слое) можно представить в виде:

$$Y_{ji} = b_j + \varepsilon_{ji},$$

где b_j - математическое ожидание фактора η при ***j*-м уровне** исследуемого входного фактора;

ε_{ji} - погрешность наблюдения, независимые стохастические компоненты наблюдений, распределенные по единому нормальному закону с нулевым математическим ожиданием и дисперсией σ^2.

Допустим, что все предположения дисперсионного анализа выполнены:

- исследуемый (единственный входной) фактор **независим**;

- исследуемый фактор подчиняется **нормальному закону распределения**;

- единственная дисперсия входного фактора **«однородна»**.

Гипотеза: **выходной фактор зависит от входного**, т.е. математические ожидания b_j различаются значимо, тогда b_j можно рассматривать как функцию от номера j уровня входного фактора:

$$b_j = \mu + T_j,$$

где μ – математическое ожидание фактора η при **всех** уровнях исследуемого входного фактора,

T_j – добавок к μ от **влияния** исследуемого входного фактора.

Таким образом, **дисперсионная модель** ***однофакторного дисперсионного*** анализа имеет вид:

$$y_{ji} = \mu + T_j + \varepsilon_{ji}.$$

Однако ни μ , ни b_j известными быть не могут, вместо них можно использовать их оценки $\overline{\overline{y}}$ и $\overline{y}_j$: $y_{ji} = \overline{y}_j + \delta_{ji}$

где δ_{ji} - независимые стохастические компоненты наблюдений, тоже распределенные по единому нормальному закону с нулевым математическим ожиданием и дисперсией σ^2.

Рассмотрим дисперсионную сумму квадратов отклонений в выражении несмещенной оценки общей дисперсии всего эксперимента:

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{\bar{y}})^2, \text{ где } N = \sum_{j=1}^{k} N_j.$$

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{\bar{y}})^2 = \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j + \bar{y}_j - \bar{\bar{y}})^2 =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{y}_j - \bar{\bar{y}})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)(\bar{y}_j - \bar{\bar{y}}) =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{k} N_j(\bar{y}_j - \bar{\bar{y}})^2 + 2\sum_{j=1}^{k}(\bar{y}_j - \bar{\bar{y}})\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j).$$

Но $2\sum_{j=1}^{k}(\bar{y}_j - \bar{\bar{y}})\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j) = 0$, так как $\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j) = 0$ по определению $\bar{y}_j$.

Первое слагаемое $\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2$ дает оценку рассеяния **внутри** серий

наблюдений (отклонения единичных замеров от средней **внутри** серии), т.е. отражает влияние всех **неучтенных** факторов. Поэтому выражение:

$$s_0^2=\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2$$

называется **остаточной (внутренней) дисперсией**.

Второе слагаемое $\sum_{j=1}^{k} N_j(\bar{y}_j - \bar{\bar{y}})^2$ дает оценку рассеяния **между**

сериями наблюдений (отклонения средних по сериям от общего среднего), т.е. отражает **влияние** изменения входного **фактора.**

Поэтому выражение: $s_A^2 = \frac{1}{k-1}\sum_{j=1}^{k} N_j(\bar{y}_j - \bar{\bar{y}})^2$

называется **межгрупповой дисперсией**.

Основное уравнение дисперсионного анализа:

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{\bar{y}})^2=\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2+\sum_{j=1}^{k}N_j(\bar{y}_j-\bar{\bar{y}})^2$$

или $$(N-1)\cdot s^2=(N-k)\cdot s_0^2+(k-1)\cdot s_A^2$$

Если в последнем уравнении: $s_A^2=s_0^2$, то $s^2=s_A^2=s_0^2$.

Отсюда: если все выборочные данные подчиняются **одному** и тому же нормальному закону распределения (с общими математическим ожиданием и дисперсией), то различие между s_A^2 и s_0^2 должно быть **незначимым**.

Для подтверждения выдвинутой гипотезы о зависимости выходного фактора от единственного входного необходимо

значимое превосходство межгрупповой дисперсии s_A^2 над

остаточной s_0^2.

Критерий Р. Фишера

Гипотеза: все выборочные данные по всем слоям подчиняются **одному** и тому же нормальному закону распределения (с общими математическим ожиданием и дисперсией), т.е. различие между s_A^2 и s_0^2 должно быть **незна́чимо**.

Из 13-й строки таблицы ***выборочных функций*** используется закон распределения $\frac{s_A^2}{s_0^2}$ Фишера: $F_{1-a}(f_1,f_2)$ при вероятности 1 – a и двух числах степеней свободы: f_1 для **большей** дисперсии и f_2 для **меньшей**.

Три возможных исхода *критерия Р. Фишера*:

– если межгрупповая дисперсия **ЗНАЧИМО БОЛЬШЕ** остаточной:

$$\frac{s_A^2}{s_0^2} > F_{1-\alpha}(k-1, N-k),$$

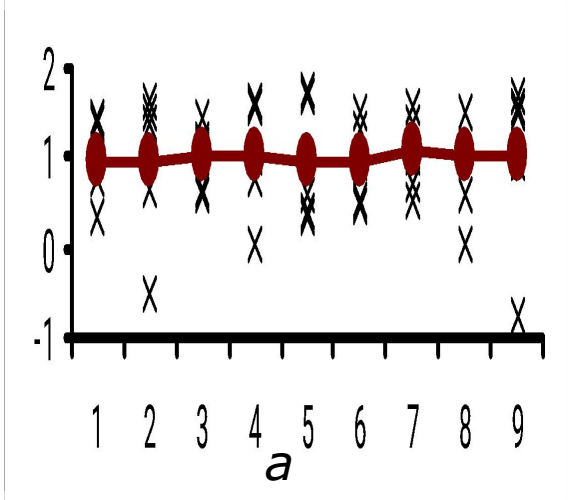
то **влияние фактора существенно** и его необходимо учитывать;

– если остаточная дисперсия **ЗНАЧИМО БОЛЬШЕ** межгрупповой:

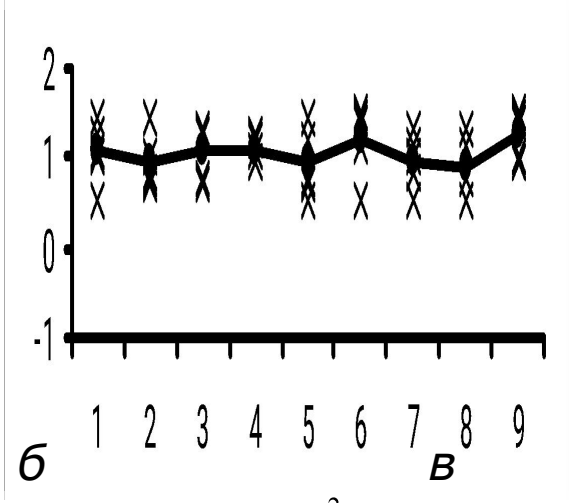
$$\frac{s_0^2}{s_A^2} > F_{1-\alpha}(N-k, k-1),$$

то **влияние фактора несущественно** и им можно пренебречь;

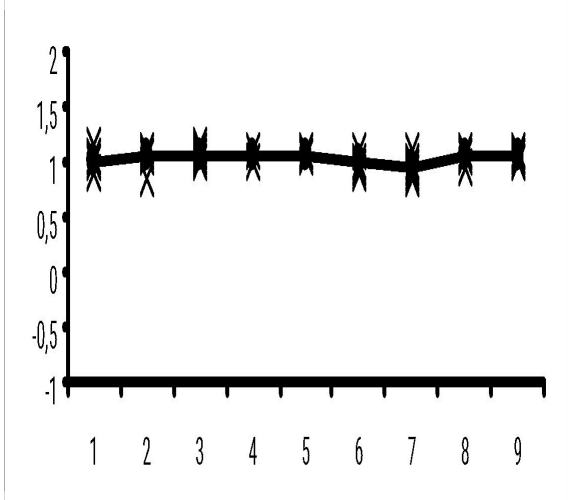
– в противном случае влияние исследуемого фактора сравнимо с погрешностью эксперимента или влиянием неучтенных факторов, поэтому конкретный вывод невозможен.



а



б



в

а) бóльшая дисперсия – остаточная: $\frac{s_0^2}{s_A^2} = 8,07 > F_{1-\alpha}(N-k, k-1) = 5,15$ –

влияние **неучтенных** факторов значительно, они "забивают" возможную зависимость от исследуемого входного фактора, признать которую нельзя.

б) бóльшая дисперсия – межгрупповая, но отношение дисперсий не достигает критического значения: $\frac{s_A^2}{s_0^2} = 1,21 < F_{1-\alpha}(k-1, N-k) = 3,04$ –

уверенный вывод о влиянии или невлиянии исследуемого входного фактора сделать нельзя.

в) межгрупповая дисперсия **значимо** больше остаточной:

$\frac{s_A^2}{s_0^2} = 9,02 > F_{1-\alpha}(k-1, N-k) = 3,04$ – влияние исследуемого входного фактора существенно.

Алгоритм дисперсионного анализа

1. Проверка независимости (или некоррелированности) исследуемых факторов методами корреляционного анализа. Обеспечение некоррелированности.

2. Проверка нормального распределения исследуемых факторов по критерию согласия Пирсона. При необходимости пересмотр факторов.

3. Проверка однородности дисперсий по критерию Фишера. При необходимости замена факторов.

4. Разбиение общей дисперсии в соответствии с задачей исследований.

5. Вычисление необходимых межгрупповых и остаточных дисперсий и проверка гипотез о значимости их различия с помощью критерия Фишера.

(6). Анализ отклонений средних от общего среднего (проверка гипотезы о равенстве математических ожиданий) с помощью критерия

знаков для *k* величин: $\frac{\bar{y}_i - \bar{\bar{y}}}{s_0}\sqrt{N_i}$, а при больших N_i и *k* еще и проверка

нормального распределения k величин (4-я или 5-я строка табл. 10 § 5.4): $\frac{\bar{y}_i - b}{\sigma}\sqrt{N_i}$ или $\frac{\bar{y}_i - b}{s}\sqrt{N_i}$.

(7). Если гипотеза о равенстве математических ожиданий отвергнута, то можно определить доверительные интервалы для них с помощью распределения Стьюдента с $N - k$ степенями свободы для функции

$$\frac{\bar{y}_i - b_i}{s_0}\sqrt{N_i}.$$