Основы системного анализа

Лекция 5 (9-я неделя)

3. Методы анализа экспериментальных данных

3.1. Дисперсионный анализ

3.2. Корреляционный анализ

3.3. Регрессионный анализ

3.1. Дисперсионный анализ

Методы *дисперсионного, корреляционного и регрессионного* анализов являются последовательными ступенями при исследовании связей между случайными величинами.

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

Корреляционный анализ позволяет оценить силу такой связи, а методами регрессионного анализа строится математическая модель и оценивается адекватность модели.

Дисперсионный анализ (ANOVA)– метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. Метод разработан Р. Фишером для анализа результатов экспериментальных исследований сложных систем.

Дисперсионный анализ является статистическим методом анализа результатов наблюдений случайной величины (*отклика*), зависящей от различных одновременно действующих независимых переменных (*факторов*), с целью оценки их влияния на исследуемую случайную величину.

Влияние различных факторов на изучаемую случайную величину (например, влияние технологического способа изготовления или режима нагрузки на долговечность технического изделия) приводит к изменению значений статистических характеристик этой величины: математического ожидания, дисперсии или моментов более высокого порядка.

С помощью дисперсионного анализа устанавливаются изменения дисперсии результатов эксперимента при изменении уровней изучаемого фактора. Если дисперсии будут отличаться значимо, то следует вывод о значимом влиянии фактора на среднее значение наблюдаемой случайной величины.

Дисперсионный анализ позволяет исследовать различие между группами данных, определять, носят ли эти расхождения случайный характер или вызваны конкретными обстоятельствами.

При выполнении эксперимента в разных условиях дисперсионный анализ поможет определить, насколько влияют внешние факторы на измерения, или отклонения носят случайный характер. Если на производстве для улучшения качества продукции изменяют режим процессов, то дисперсионный анализ позволяет оценить результаты воздействия данного фактора.

Классические методы дисперсионного анализа основываются на следующих предпосылках: распределение исходных случайных величин нормально; дисперсии экспериментальных данных одинаковы для всех условий эксперимента (т.е. для каждого уровня фактора).

Т.е. проведение дисперсионного анализа возможно при соблюдении следующих требований для каждого уровня фактора:

- наблюдения независимы и проводятся в одинаковых условиях;
- измеряемая случайная величина имеет нормальный закон распределения с постоянной для всех уровней фактора дисперсией σ^2.

Сущность метода состоит в разделении общей дисперсии на две части, одна из которых обусловлена случайной ошибкой (то есть внутригрупповой изменчивостью), а вторая связана с различием средних значений, и проверке гипотез о значимости влияния этих факторов на случайную величину.

Сравнивая компоненты дисперсии друг с другом с помощью *F*-критерия Фишера, можно определить, какая доля общей вариативности случайной величины обусловлена действием регулируемых факторов.

Основной целью сравнения компонентов дисперсии друг с другом посредством F-критерия Фишера является проверка нулевой гипотезы H0: «Средние величины (мат. ожидания) случайной величины на всех уровнях фактора одинаковы».

Принципиальный алгоритм дисперсионного анализа включает несколько основных этапов: ·

- установление основных источников варьирования результативного признака и определение объемов вариации (сумм квадратов отклонений) по источникам её образования;

- определение числа степеней свободы, соответствующих компонентам общей вариации;
- вычисления дисперсий как отношений соответствующих объемов вариации к их числу степеней свободы;
- анализ соотношения между дисперсиями;
- оценка достоверности разницы между средними и формулирование выводов.

Указанный алгоритм сохраняется как при простых моделях дисперсионного анализа, когда данные группируются по одному признаку, так и при сложных моделях, когда данные группируются по двумя и большим числом признаков. Однако с увеличением числа групповых признаков усложняется процесс разложение общей вариации по источникам ее образования.

В зависимости от типа и количества переменных различают:

- однофакторный и многофакторный дисперсионный анализ (одна или несколько независимых переменных);
- одномерный и многомерный дисперсионный анализ (одна или несколько зависимых переменных);
- дисперсионный анализ с повторными измерениями (для зависимых выборок);
- дисперсионный анализ с постоянными факторами, случайными факторами, и смешанные модели с факторами обоих типов.

Однофакторный дисперсионный анализ

Предположим, что анализируется влияние на случайную величину X фактора A, изучаемого в k группах (A_1, A_2,..., A_i,..., A_k). В каждой группе A_i проведены n наблюдений (x_{i1}, x_{i2},..., x_{in}), i = const случайной величины X (табл. 1). Следовательно, во всех k группах фактора A произведены $N = k \times n$ наблюдений.

Экспериментальные данные

Номер наблюдения	Группы 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}

Рассмотрим оценки различных дисперсий, возникающие при анализе таблицы результатов наблюдений.

Для оценки вариации данных, характеризующей изменение данных внутри одной группы A_i, вводится понятие оценки *групповой дисперсии*:

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2,\ i = \text{const}$$

где среднее значение отклика по i-й группе:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n} x_{ij},\ i = \text{const}.$$

Для оценки вариации данных, характеризующей рассеяние рассматриваемой случайной величины вне влияния фактора A, вводится понятие оценки *средней внутригрупповой дисперсии*:

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

Для оценки вариации данных, характеризующей влияние фактора, положенного в основу группы A_i, на общую выборку вводится понятие оценки *межгрупповой дисперсии*:

$$S_A^2 = \frac{1}{k - 1}\sum_{i=1}^{k} n_i\left(\bar{x}_i - \bar{x}_0\right)^2,$$

где общее среднее значение отклика по всем наблюдениям:

$$\overline{x}_0 = \frac{1}{k}\sum_{i=1}^{k}\overline{x}_i = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}\,.$$

Общая дисперсия выборки может быть представлена в виде суммы межгрупповой дисперсии (дисперсии вариантов) и средней внутригрупповой дисперсии (дисперсии ошибок):

$$\sigma = \sigma_A + \sigma_0\,,$$

а дисперсионное отношение имеет вид:

$$F = \frac{\sigma_A}{\sigma_0} = \frac{S_A^2}{S_0^2}\,.$$

Если фактор, положенный в основу группировки, не оказывает влияния на вариацию изучаемой случайной величины, то средняя внутригрупповая дисперсия будет отражать влияние прочих факторов, которые определяют вариацию внутри групп, и поэтому дисперсионное отношение будет близко к единице или отличаться от неё в силу присутствия случайных колебаний.

Проверка влияния фактора A на изменение математических ожиданий может быть сведена к сравнению межгрупповой и средней внутригрупповой дисперсии, оценками которых являются соответственно $S_A{}^2$ и $S_0{}^2$. Влияние фактора A признается значимым для математических ожиданий случайной величины, если значимо дисперсионное отношение F. Дисперсионное отношение признается значимым с доверительной вероятностью α, если

$$F = \frac{S_A^2}{S_0^2} > F_\alpha(d.f_1; d.f_2),$$

где $F_\alpha(d.f_1;\ d.f_2)$ – квантиль F распределения с $d.f_1$ и $d.f_2$ степенями свободы, определяемый таблично

Для проверки значимости результата (т.е. случайности или неслучайности отклонения математических ожиданий) учитывается число степеней свободы $d.f$. Оценка *групповой дисперсии* имеет число степеней свободы $d.f_i = (n - 1)$, оценка *межгрупповой дисперсии* имеет число степеней свободы $d.f_1 = (k - 1)$, а оценка *средней внутригрупповой дисперсии* имеет число степеней свободы $d.f_2 = k(n - 1) = (N - k)$.