Дисперсионный анализ

Материал из Википедии — свободной энциклопедии

Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях[1][2]. В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп. Разработан Р. Фишером для анализа результатов экспериментальных исследований. В литературе также встречается обозначение ANOVA (от англ. *ANalysis Of VAriance*)[3].

Содержание

Типы дисперсионного анализа

Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную. Зависимые переменные представлены значениями абсолютных шкал (шкала отношений). Независимые переменные являются номинативными (шкала наименований), то есть отражают групповую принадлежность, и могут иметь два или более значения (типа, градации или уровня). Примерами независимой переменной X_i с двумя значениями могут служить пол (женский: X_1, мужской: X_2) или тип экспериментальной группы (контрольная: X_1, экспериментальная: X_2). Градации, соответствующие независимым выборкам объектов, называются межгрупповыми, а градации, соответствующие зависимым выборкам, — внутригрупповыми.

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;

Математическая модель дисперсионного анализа

Математическая модель дисперсионного анализа представляет собой частный случай основной линейной модели. Пусть с помощью методов A_j $(1 \leq j \leq m)$ производится измерение нескольких параметров x_i $(1 \leq i \leq n)$, чьи точные значения — μ_i $(1 \leq i \leq n)$. В таком случае результаты измерений различных величин различными методами можно представить как:

$$x_{i,j} = \mu_i + a_{i,j} + e_{i,j},$$

где:

- $x_{i,j}$ — результат измерения i-го параметра по методу A_j;
- μ_i — точное значение i-го параметра;
- $a_{i,j}$ — систематическая ошибка измерения i-го параметра в группе по методу A_j;
- $e_{i,j}$ — случайная ошибка измерения i-го параметра по методу A_j.

Тогда дисперсии следующих случайных величин:
$x_{i,j}$
$x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*}$
$x_{i,*}$
$x_{*,j}$
(где:

$$x_{*,j} = \frac{1}{n} \sum_i x_{i,j},$$

$$x_{i,*} = \frac{1}{m} \sum_j x_{i,j},$$

$$x_{*,*} = \frac{1}{nm} \sum_{i,j} x_{i,j} \text{)}$$

выражаются как:

$$s^2 = \frac{1}{nm} \sum_i \sum_j (x_{i,j} - x_{*,*})^2$$

$$s_0^2 = \frac{1}{nm} \sum_i \sum_j (x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*})^2$$

$$s_1^2 = \frac{1}{n}\sum_i (x_{i,*} - x_{*,*})^2$$

$$s_2^2 = \frac{1}{m}\sum_j (x_{*,j} - x_{*,*})^2$$

и удовлетворяют тождеству:

$$s^2 = s_0^2 + s_1^2 + s_2^2$$

Процедура дисперсионного анализа состоит в определении соотношения систематической (межгрупповой) дисперсии к случайной (внутригрупповой) дисперсии в измеряемых данных. В качестве показателя изменчивости используется сумма квадратов отклонения значений параметра от среднего: SS (от англ. *Sum of Squares*). Можно показать, что общая сумма квадратов SS_{total} раскладывается на межгрупповую сумму квадратов SS_{bg} и внутригрупповую сумму квадратов SS_{wg}:

$$SS_{\text{total}} = SS_{\text{bg}} + SS_{\text{wg}}$$

Пусть точное значение каждого параметра есть его математическое ожидание, равное среднему генеральной совокупности $E(X) = M$. При отсутствии систематических ошибок групповое среднее и среднее генеральной совокупности тождественны: $M_j = M$. Тогда случайная ошибка измерения есть разница между результатом измерения $x_{i,j}$ и средним группы: $x_{i,j} - M_j$. Если же метод A_j оказывает систематическое воздействие, то систематическая ошибка при воздействии этого фактора есть разница между средним группы M_j и средним генеральной совокупности: $M_j - M$.

Тогда уравнение $x_{i,j} = \mu_i + a_{i,j} + e_{i,j}$ может быть представлено в следующем виде:

$x_{i,j} = M + (M_j - M) + (x_{i,j} - M_j)$, или

$x_{i,j} - M = (M_j - M) + (x_{i,j} - M_j)$.

Тогда

$$\sum_{i=1}^{n_j}(x_{i,j} - M)^2 = \sum_{i=1}^{n_j}(M_j - M)^2 + \sum_{i=1}^{n_j}(x_{i,j} - M_j)^2,$$

где

$$SS_{\text{total}} = \sum_{i=1}^{n_j}(x_{i,j} - M)^2$$

$$SS_{\text{bg}} = \sum_{i=1}^{n_j}(M_j - M)^2$$

$$SS_{\text{wg}} = \sum_{i=1}^{n_j} (x_{i,j} - M_j)^2$$

Следовательно

$$SS_{\text{total}} = SS_{\text{bg}} + SS_{\text{wg}}.$$

Аналогичным образом раскладываются степени свободы:

$df_{\text{total}} = df_{\text{bg}} + df_{\text{wg}}$, где

$$df_{\text{total}} = N - 1,$$

$$df_{\text{bg}} = J - 1,$$

$$df_{\text{wg}} = N - J,$$

и N есть объём полной выборки, а J — количество групп.

Тогда дисперсия каждой части, именуемая в модели дисперсионного анализа как «средний квадрат», или MS (от англ. *Mean Square*), есть отношение суммы квадратов к числу их степеней свободы:

$$MS_{\text{total}} = \frac{SS_{\text{total}}}{N - 1}$$

$$MS_{\text{bg}} = \frac{SS_{\text{bg}}}{J - 1}$$

$$MS_{\text{wg}} = \frac{SS_{\text{wg}}}{N - J},$$

Соотношение межгрупповой и внутригрупповой дисперсий имеет *F*-распределение (распределение Фишера) и определяется при помощи (*F*-критерия Фишера):

$$F_{df_{\text{bg}}, df_{\text{wg}}} = \frac{MS_{\text{bg}}}{MS_{\text{wg}}}.$$

Принципы и применение

Исходными положениями дисперсионного анализа являются

- нормальное распределение значений изучаемого признака в генеральной совокупности;
- равенство дисперсий в сравниваемых генеральных совокупностях;
- случайный и независимый характер выборки.

Нулевой гипотезой в дисперсионном анализе является утверждение о равенстве средних значений:

$$H_0\colon \quad \mu_1 = \mu_2 = \cdots = \mu_j.$$

При отклонении нулевой гипотезы принимается альтернативная гипотеза о том, что не все средние равны, то есть имеются, по крайней мере, две группы, отличающиеся средними значениями:

$$H_1\colon \quad \mu_1 \neq \mu_2 \neq \cdots \neq \mu_j.$$

При наличии трёх и более групп для определения различий между средними применяются *post-hoc* *t*-тесты или метод контрастов.

Однофакторный дисперсионный анализ

Простейшим случаем дисперсионного анализа является одномерный однофакторный анализ для двух или нескольких независимых групп, когда все группы объединены по одному признаку. В ходе анализа проверяется нулевая гипотеза о равенстве средних. При анализе двух групп дисперсионный анализ тождественен двухвыборочному *t*-критерию Стьюдента для независимых выборок, и величина *F*-статистики равна квадрату соответствующей *t*-статистики.

Для подтверждения положения о равенстве дисперсий обычно применяется критерий Ливена (*Levene's test*). В случае отвержения гипотезы о равенстве дисперсий основной анализ неприменим. Если дисперсии равны, то для оценки соотношения межгрупповой и внутригрупповой изменчивости применяется *F*-критерий Фишера:

$$F_{df_{\text{bg}},df_{\text{wg}}} = \frac{MS_{\text{bg}}}{MS_{\text{wg}}}.$$

Если *F*-статистика превышает критическое значение, то нулевая гипотеза не может быть принята (отвергается) и делается вывод о неравенстве средних. При анализе средних двух групп результаты могут быть интерпретированы непосредственно после применения критерия Фишера.

При наличии трёх и более групп требуется попарное сравнение средних для выявления статистически значимых отличий между ними. Априорный анализ включает метод контрастов, при котором межгрупповая сумма квадратов дробится на суммы квадратов отдельных контрастов:

$$SS_{\text{bg}} = SS_{\psi_1} + SS_{\psi_2} + \ldots + SS_{\psi_n},$$

где ψ есть контраст между средними двух групп, и затем при помощи критерия Фишера проверяется соотношение среднего квадрата для каждого контраста к внутригрупповому среднему квадрату:

$$F_{1,df_{\text{wg}}} = \frac{MS_{\psi_i}}{MS_{\text{wg}}}.$$

Апостериорный анализ включает *post-hoc* *t*-критерии по методам Бонферрони или Шеффе, а также сравнение разностей средних по методу Тьюки. Особенностью *post-hoc*-тестов является использование внутригруппового среднего квадрата MS_{wg} для оценки любых пар средних. Тесты по методам Бонферрони и Шеффе являются наиболее консервативными, так как они используют наименьшую критическую область при заданном уровне значимости α.

Помимо оценки средних дисперсионный анализ включает определение коэффициента детерминации R^2, показывающего, какую долю общей изменчивости объясняет данный фактор:

$$R^2 = \frac{SS_{\text{bg}}}{SS_{\text{total}}}.$$

Многофакторный дисперсионный анализ

- Многофакторный анализ позволяет проверить влияние нескольких факторов на зависимую переменную. Линейная модель многофакторной модели имеет вид:

$x_{i,j,k} = \mu_i + a_{i,j} + b_{i,k} + \ldots + (ab)_{i,j,k} + e_{i,j,k}$, где:

- - $x_{i,j,k}$ — результат измерения i-го параметра;
  - μ_i — среднее для i-го параметра;
  - $a_{i,j}$ — систематическая ошибка измерения i-го параметра в j группе по методу A;
  - $b_{i,k}$ — систематическая ошибка измерения i-го параметра в k группе по методу B;
  - $(ab)_{i,j,k}$ — систематическая ошибка измерения i-го параметра в j,k группе в силу комбинации методов A и B;
  - $e_{i,j,k}$ — случайная ошибка измерения i-го параметра.

В отличие от однофакторной модели, где имеется одна межгрупповая сумма квадратов, модель многофакторного анализа включает суммы квадратов для каждого фактора в отдельности и суммы квадратов всех взаимодействий между ними. Так, в двухфакторной модели межгрупповая сумма квадратов раскладывается на сумму квадратов фактора A, сумму квадратов фактора B и сумму квадратов взаимодействия факторов A и B:

$$SS_{\text{total}} = SS_A + SS_B + SS_{AB} + SS_{\text{wg}}.$$

Соответственно трёхфакторная модель включает сумму квадратов фактора A, сумму квадратов фактора B, сумму квадратов фактора C и суммы квадратов взаимодействий факторов A и B, B и C, A и C, а также взаимодействия всех трёх факторов A, B, C:

$$SS_{\text{total}} = SS_A + SS_B + SS_C + SS_{AB} + SS_{BC} + SS_{AC} + SS_{ABC} + SS_{\text{wg}}.$$

Степени свободы раскладываются аналогичным образом:

$df_{\text{total}} = df_A + df_B + df_{AB} + df_{\text{wg}}$, где

$$df_{\text{total}} = N - 1,$$

$$df_A = J - 1,$$

$$df_B = K - 1,$$

$$df_{AB} = (J - 1)(K - 1),$$

$$df_{\text{wg}} = N - JK,$$

и N есть объём полной выборки, J — количество уровней (групп) фактора A, а K — количество уровней (групп) фактора B.

В ходе анализа проверяются несколько нулевых гипотез:

- гипотеза о равенстве средних под влиянием фактора A: $H_0\colon \mu_{1,*} = \mu_{2,*} = \cdots = \mu_{j,*}$;
- гипотеза о равенстве средних под влиянием фактора B: $H_0\colon \mu_{*,1} = \mu_{*,2} = \cdots = \mu_{*,k}$;
- гипотеза об отсутствии взаимодействия факторов A и B: $H_0\colon (ab)_{j,k} = 0$ для всех j и k.

Каждая гипотеза проверяется с помощью критерия Фишера:

$$F_{df_A, df_{\text{wg}}} = \frac{MS_A}{MS_{\text{wg}}};$$

$$F_{df_B, df_{\text{wg}}} = \frac{MS_B}{MS_{\text{wg}}};$$

$$F_{df_{AB}, df_{\text{wg}}} = \frac{MS_{AB}}{MS_{\text{wg}}}.$$

При отвержении нулевой гипотезы о влиянии отдельного фактора принимается утверждение, что присутствует главный эффект фактора A (B, и т. д.). При отвержении нулевой гипотезы о взаимодействии факторов принимается утверждение о том, что влияние фактора A проявляется по-разному на разных уровнях фактора B. Обычно в таком случае результаты общего анализа признаются не имеющими силы, и влияние фактора A проверяется отдельно на каждом уровне фактора B с помощью однофакторного дисперсионного анализа или *t*-критерия.

Примечания

1. Дисперсионный анализ (http://www.statsoft.ru/home/textbook/modules/stanman.html#basic).
2. *Дисперсионный анализ* — статья из Большой советской энциклопедии. Большев, Л. Н..
3. А. Д. Наследов. Математические методы психологического исследования. СПб, 2008. ISBN 5-9268-0275-X

Литература

- *Шеффе Г.* Дисперсионный анализ, пер. с англ. — М., 1963.
- *Смирнов Н. В., Дунин-Барковский И. В.* Курс теории вероятностей и математической статистики для технических приложений. — 2. — М., 1965.