ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

Статистическое моделирование и прогнозирование
Лекция 5. Дисперсионный анализ

Семёнов Михаил Евгеньевич
к. ф.-м. н., доцент ОЭФ ИЯТШ

Томский политехнический университет
15 апреля 2020 г.

Содержание

ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Качественные и количественные признаки

Качественным признаком называют признак отдельные варианты которого выражаются в виде понятий или наименований. Он может быть представлен в виде альтернативного или формального признака:

- Альтернативный признак – признак имеющий два противоположных значения.
- Формальный признак – признак, по сути относимый к качественному, но представленный числом (например, успеваемость студентов можно представить в виде формального признака).

Количественным признаком называют признак отдельные варианты которого имеют числовое выражение и отличаются по величине, то есть варьируются.

Дисперсионный анализ предназначен для оценки влияния одного или нескольких факторов (качественных величин) на количественную случайную величину.

В случае, когда рассматривается влияние только одного качественного признака, имеющего конечное число уровней градаций, дисперсионный анализ называется *однофакторным*.

Однако очень редко тот или иной процесс определяется только одним фактором. Напротив, обычно наблюдается одновременное влияние многих факторов. *Двухфакторный дисперсионный анализ* (англ. two-way analysis of variance, или two-way ANOVA) позволяет установить одновременное влияние двух факторов, а также взаимодействие между этими факторами.

Предположим, что одна и та же случайная величина X с одинаковой точностью измеряется при k различных значениях фактора. Если анализируемый фактор оказывает существенное влияние на X, то наблюдения на одном уровне будут значимо отличаться от наблюдений на других уровнях, и, следовательно, средние значения на разных уровнях будут различными. И наоборот, если фактор не оказывает влияние на рассматриваемую случайную величину, то средние значения X на различных уровнях будут статистически незначимо отличаться друг от друга.

Представим результаты наблюдений в виде таблицы:

i	n_i	x_{ij}
1	n_1	$x_{11}, x_{12}, \ldots, x_{1n_1}$
2	n_2	$x_{21}, x_{22}, \ldots, x_{2n_2}$
$\ldots$	$\ldots$	$\ldots$
k	n_k	$x_{k1}, x_{k2}, \ldots, x_{kn_k}$

где i – уровни фактора; $j = 1, 2, \ldots, n_i$ – номера наблюдений на i-ом уровне; n_i -– количество наблюдений на i-ом уровне; x_{ij} — наблюдаемые значения.

Условия проведения дисперсионного анализа

При проведении дисперсионного анализа предполагается выполнение следующих условий:

1. Результаты наблюдений x_{ij} – это независимые случайные величины, то есть $cov(x_{ij}, x_{lm}) = 0$, где $i \neq$l и/или $j \neq m$.
2. Совокупности наблюдаемых значений $\{x_{i1}, x_{i2}, ..., x_{in_i}\}$ на каждом уровне i нормально распределены: $N(a_i, \sigma_i)$, где a_i, σ_i — среднее и дисперсия i-го уровня.
3. Дисперсии распределений на всех уровнях $i = 1, 2, ..., k$ одинаковы: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = const$.

Гипотеза

Гипотеза: С учётом выдвинутых условий формулируется нулевая гипотеза о равенстве средних всех уровней $H_0 : a_1 = a_2 = ... = a_k$ при альтернативной, что хотя бы одно из указанных равенств нарушается $H_1 : \exists a_l \neq a_m$, где $l \neq m$.

Статистика: Рассмотрим следующие величины:

$$\bar{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij},\ \bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{x_i},\ n = \sum_{i=1}^{k} n_i, \tag{1}$$

где $\bar{x_i}$ – средние значения i-го уровня; $\bar{x}$ – общее среднее значение всех n величин.

$$Q_t = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2,\ Q_d = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2,\ Q_r = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2, \qquad (2)$$

где Q_t - сумма квадратов отклонений отдельных наблюдений x_{ij} от общего среднего $\bar{x}$; Q_d – сумма квадратов отклонений средних значений уровней $\bar{x}_i$ от общей средней $\bar{x}$, которая характеризует различия между средними значениями отдельных уровней и определяется влиянием рассматриваемого фактора; Q_r – сумма квадратов отклонений отдельных наблюдений x_{ij} от средних значений своего уровня $\bar{x}_i$, которая обусловлена наличием неучтённых факторов и называется остаточным рассеянием или суммой квадратов внутри групп.

Гипотеза

Можно доказать, что имеет место равенство $Q_t = Q_d + Q_r$, причём, левая часть равенства имеет $n-1$ степень свободы, первое слагаемое в правой части – $(k-1)$ степень свободы, а второе – $(n-k)$, и каждая сумма квадратов, делённая на соответствующее число степеней свободы, будет представлять несмещённую оценку дисперсии случайной величины X. При этом, величина $\frac{1}{n-1}Q_t$ в любом случае является несмещённой оценкой дисперсии X, а величины $\frac{1}{k-1}Q_d$ и $\frac{1}{n-k}Q_r$ – только в рамках гипотезы о равенстве средних значений уровней фактора, то есть при отсутствии влияния исследуемого фактора на случайную величину X.

Тогда при согласии с нулевой гипотезой $H_0 : a_1 = a_2 = ... = a_k$ статистика F_s будет иметь распределение Фишера с числами степеней свободы числителя $k-1$, и знаменателя $n-k$:

$$F_s = \frac{\frac{1}{k-1}Q_d}{\frac{1}{n-k}Q_r} = \frac{(n-k)\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2} \sim F_{\frac{k-1}{n-k}}. \tag{3}$$

Критерий: Гипотеза H_0 принимается, если $F_s < F_{\alpha,\frac{k-1}{n-k}}$; и отвергается в противном случае, где $F_{\alpha,\frac{k-1}{n-k}}$ – квантиль уровня α.

Вышеуказанный выбор критической области $F_s \geqslant F_{\alpha,\frac{k-1}{n-k}}$ определяется тем, что при выполнении альтернативной гипотезы H_1 статистика F_s неограниченно возрастает с ростом объёма выборки n.

Постановка заданий

1. В ходе исследования были получены значения количественного признака для трёх различных уровней качественного признака (фактора).
   D = (4,0; 4,5; 4,3; 5,6; 4,9; 5,4; 3,8; 3,7; 4,0),
   B = (4,5; 4,9; 5,0; 5,7; 5,5; 5,6; 4,7; 4,5; 4,7),
   S = (5,4; 4,9; 5,6; 5,8; 6,1; 6,3; 5,5; 5,0; 5,0).
   Используя методику однофакторного дисперсионного анализа, определите: значимо ли влияние изменения качественного признака на величину признака количественного.
   Сохранить скрипт в файле (*.r).

2 Установите пакет HSAUR2. Используя методику двухфакторного дисперсионного анализа, установите для экспериментальных данных *weightgain*: одновременное влияние двух факторов, а также взаимодействие между этими факторами.
Сохраните скрипт в файле (*.r). Сохраните экспериментальные данные *weightgain* в файле (*.csv).

3 Представьте результаты дисперсионного анализа в виде классической таблицы ANOVA.

4 Выполните дисперсионный анализ при помощи базовых функций aov() и lm() и сравните результаты [Мастицкий, 2020].

Алгоритм 1 Проверка значимости влияния изменений качественного признака на величину количественного признака с помощью методики однофкторного анализа

Вход: D, B, S – значения количественного признака, наблюдаемые на каждом из уровней качественного признака

```
1: D = c(4,0; 4,5; 4,3; 5,6; 4,9; 5,4; 3,8; 3,7; 4,0)
2: B = c(4,5; 4,9; 5,0; 5,7; 5,5; 5,6; 4,7; 4,5; 4,7)
3: S = c(5,4; 4,9; 5,6; 5,8; 6,1; 6,3; 5,5; 5,0; 5,0)
4: adhf = stack(data.frame(D, B, S))
5: print(anova(lm(values ~ind, data=adhf)))
```

Выход: Таблица ANOVA

Результат


```
        Analysis of Variance Table
Response: values
          Df Sum Sq Mean Sq F value   Pr(>F)
ind        2 4.9119 2.45593  7.7578 0.002519 **
Residuals 24 7.5978 0.31657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 1 – Результат выполнения алгоритма 1 в виде таблицы ANOVA

Алгоритм 2 Проверка одновременого влияния двух факторов и взаимодействия между этими факторами

Вход: *weightgain* – экспериментальные данные; *type*, *source* – проверяемые факторы.

```
1: install.packages("HSAUR2")
2: data(weightgain, package="HSAUR2")
3: M2 <- lm(weightgain ~type*source, data=weightgain)
4: anova(M2) # summary(M2)
```

Выход: Таблица ANOVA

Результат


```
Analysis of Variance Table
Response: weightgain
            Df Sum Sq Mean Sq F value  Pr(>F)
type         1 1299.6 1299.60  5.8123 0.02114 *
source       1  220.9  220.90  0.9879 0.32688
type:source  1  883.6  883.60  3.9518 0.05447 .
Residuals   36 8049.4  223.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 2 – Результат выполнения алгоритма 2 в виде таблицы ANOVA

Буховец, , Москалев, , Богатова, , and Бирючинская, (2010).
Статистический анализ данных в системе R. Учебное пособие.
ВГАУ, Воронеж.

Мастицкий, (2020).
Однофакторный дисперсионный анализ.
https://r-analytics.blogspot.com/2013/01/blog-post.html.
[Онлайн; дата доступа 15 апреля 2020].

Мастицкий, and Шитиков, (2014).
Статистический анализ и визуализация данных с помощью R.