Проверка гипотез однородности в случае нескольких групп

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

Однофакторный дисперсионный анализ

Пусть имеется выборка наблюдений, которая разбивается на r групп, Каждая группа содержит n_i величин $\mathcal{N}(\mu_i, \sigma)$ с одинаковой дисперсией, $n_1 + \ldots + n_r = n$, x_{ij} — j-я величина $(j = 1, \ldots, n_i)$ в i-й группе $(i = 1, \ldots, r)$. Модель

$$x_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$$

Нулевая гипотеза о равенстве внутригрупповых средних
$H_0 : \mu_1 = \mu_2 = \ldots = \mu_r = \mu$
Оценки параметров модели

- $\hat{\mu}_i = \bar{x}_i = \frac{1}{n_i} \sum\limits_{j=1}^{n_i} x_{ij}$ — среднее в i-й группе;
- $\hat{\mu} = \bar{x} = \frac{1}{n} \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{n_i} x_{ij}$ — общее среднее.

Поскольку разбиение на группы обусловлено одним фактором (методика обучения), анализ называется однофакторным.

Пример данных для однофакторного дисперсионного анализа

Три группы водителей обучались по различным методикам. После окончания обучения был проведен контрольный тест. Получены следующие результаты:

Номер группы i	Число ошибок x_{ij}	Сумма	Число водителей n_i	Среднее в группе $\bar{x}_i$
1	1 3 2 1 0 2 1	10	7	$10/7 = 1.43$
2	2 3 2 1 4	12	5	$12/5 = 2.4$
3	4 5 3	12	3	$12/3 = 4$
		34	$n = 15$	

Общее среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} = \frac{34}{15} = 2.27\,.$

Значимо ли различие между результатами разных методик?

Разложение суммы квадратов отклонений от выборочного среднего

- $Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ общий источник вариации
- $Q_1 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i (\bar{x}_i - \bar{x})^2$ источник вариации, обусловленный различием между группами
- $Q_2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ источник вариации, обусловленный различием внутри групп

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 =$$

$$= \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}_{Q_2} + \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2}_{Q_1}, \text{ так как}$$

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) = \sum_{i=1}^{r} (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0\,.$$

Теорема Фишера-Кочрена

Theorem

(Fisher-Cochran) Пусть $Y = (y_1, \ldots, y_n)$ с независимыми компонетами, $y_i \sim \mathcal{N}(0,1)$, квадратичные формы $Q_1, \ldots, Q_k$ рангов $n_1, \ldots, n_k$, такие что

$$Y^T Y = Q_1 + \ldots + Q_k \,.$$

Тогда $n = n_1 + \ldots + n_k$ равносильно $Q_i \sim \chi^2(n_i)$ и $Q_1, \ldots, Q_k$ независимы.

Доказательство т.Фишера-Кочрена

Пусть $Q_i = Y^T A_i Y$, где матрица A_i ранга n_i. Тогда существует матрица B_i размерности n_i на n, при помощи которой квадратичная форма приводится к диагональному виду, то есть $Q_i = (B_i Y)^T \Delta_i (B_i Y)$, где Δ_i диагональная матрица с элементами ± 1. Соберем k матриц B_i в одну матрицу B размерности n на n.

$$[B_1^T(n, n_1)| \dots |B_k^T(n, n_k)] \cdot \begin{bmatrix} \Delta_1 & \dots & 0 \\ \vdots & \dots & \vdots \\ 0 & \dots & \Delta_k \end{bmatrix} \cdot \begin{bmatrix} B_1(n_1, n) \\ \vdots \\ B_k(n_k, n) \end{bmatrix} = \sum_i B_i^T \Delta_i B_i$$

Следовательно,

$$Y^T Y = \sum_i Q_i = \sum_i (B_i Y)^T \Delta_i (B_i Y) =$$

$$= Y^T (\sum_i B_i^T \Delta_i B_i) Y = Y^T B^T \Delta B Y, \ \Rightarrow \ B^T \Delta B = I\,.$$

Ранг матрицы B должен быть равен n, так как с одной стороны, он не может быть больше n, а с другой ранг произведения матриц не превышает минимального ранга сомножителей,

$$n = \mathrm{rank}((B^T \Delta)B) \leq \min(\mathrm{rank}(B^T \Delta), \mathrm{rank}(B)).$$

Тогда $\Delta = (B^T)^{-1} B^{-1}$ положительно определенная матрица, следовательно, $\Delta = I$ и матрица B является ортогональной.

Из ортогональности B следует, что компоненты вектора $X = BY$ независимы и нормальны, кроме того $Y^T Y = X^T X$.

$$\begin{aligned} Q_1 &= x_1^2 + \dots x_{n_1}^2, \\ Q_2 &= x_{n_1+1}^2 + \dots x_{n_1+n_1}^2, \\ Q_3 &= x_{n_1+n_2+1}^2 + \dots x_{n_1+n_2+n_3}^2, \\ &\dots \end{aligned}$$

Осюда все Q_i независимы и имеют распределение $\chi^2(n_i)$ соответственно. Таким образом достаточность $n = n_1 + \dots + n_k$ установлена. Необходимость очевидна.

Распределение Q_1, Q_2

Ранги квадратичных форм Q, Q_1 и Q_2 равны $n-1$, $r-1$ и $n-r$ соответственно.

Пусть A ортогональная матрица ($A^{-1} = A^T$), у которой последняя строка имеет вид $\frac{1}{\sqrt{n}} + \ldots + \frac{1}{\sqrt{n}}$, и введем преобразование $y = Ax$, $x = A^{-1}y = A^T y$.

$$\sum_{ij} x_{ij}^2 = x^T x = (A^T y)^T A^T y = y^T A A^T y = y^T y = \sum_{k=1}^{n} y_k^2 ,$$

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{ij} x_{ij}^2 - n\bar{x}^2 =$$

$$= \sum_{ij} x_{ij}^2 - \left(\frac{x_1}{\sqrt{n}} + \ldots + \frac{x_n}{\sqrt{n}} \right)^2 = \sum_{k=1}^{n} y_k^2 - y_n^2 = \sum_{k=1}^{n-1} y_k^2$$

$$\Longrightarrow \quad \operatorname{rank}(Q) = n - 1.$$

Источник вариации, обусловленный различием между группами, можно представить в виде

$$Q_1 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} L_i^2, \quad \text{где} \quad L_i = \sqrt{n_i}(\bar{x}_i - \bar{x}),$$

$$\sum_{i=1}^{r} L_i\sqrt{n_i} = 0, \quad \text{так как} \quad n\bar{x} = \sum_{i=1}^{r} n_i\bar{x}_i\,.$$

Следовательно, $\mathrm{rank}(Q_1) \le r - 1$.

Источник вариации, обусловленный различием внутри групп, можно представить в виде

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i,j} L_{ij}^2, \text{ где } L_{ij} = x_{ij} - \bar{x}_i,$$

$$\sum_{j=1}^{n_i} L_{ij} = 0 \ \ \forall i = 1, \ldots, r. \Longrightarrow \ \ \operatorname{rank}(Q_2) \leq n - r.$$

Воспользуемся свойством рангов квадратичных форм

$$\operatorname{rank}(Q_1 + Q_2) \leq \operatorname{rank}(Q_1) + \operatorname{rank}(Q_2)\,.$$

$$n - 1 \leq (r - 1) + (n - r) = n - 1,$$

и ранги Q_1 и $Q - 2$ в точности равны $r - 1$ и $n - r$. По теореме Фишера-Кочрена Q_1, Q_2 независимы, $\frac{Q_1}{\sigma^2} \sim \chi^2(r - 1)$, $\frac{Q_2}{\sigma^2} \sim \chi^2(n - r)$.

Проверка гипотез в однофакторном дисперсионном анализе

При справедливости гипотезы $H_0 : \mu_1 = \ldots = \mu_r = \mu$

- $Q \sim \sigma^2 \chi^2(n-1)$
- $Q_1 \sim \sigma^2 \chi^2(r-1)$
- $Q_2 \sim \sigma^2 \chi^2(n-r)$
- Q_1 и Q_2 независимы.

Для проверки гипотезы H_0 используется статистика

$$F = \frac{\frac{1}{r-1} Q_1}{\frac{1}{n-r} Q_2} \sim F(r-1, n-r).$$

Пример вычислений

$r = 3,\ n = 15$

- $Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}^2 - 2x_{ij}\bar{x} + \bar{x}^2) =$

  $= \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 - 2\bar{x}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} + n\bar{x}^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 - n\bar{x}^2 =$

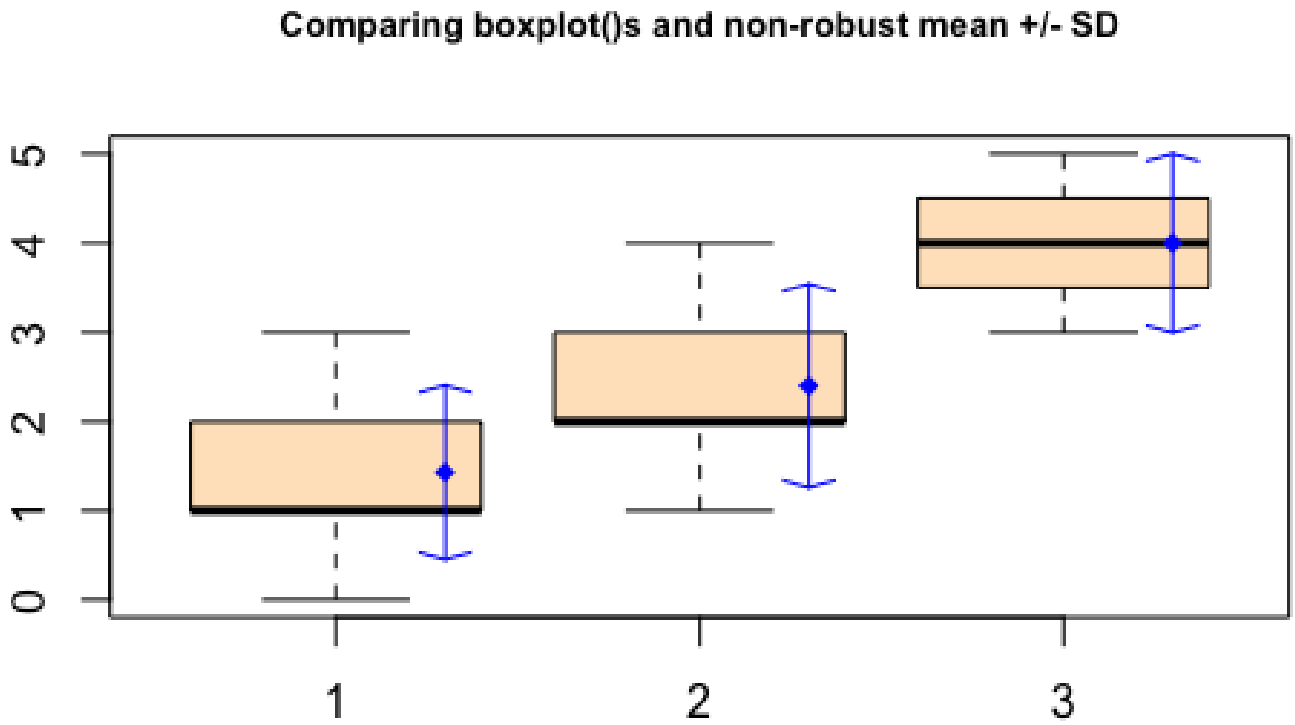
  $= 104 - 15\left(\frac{34}{15}\right)^2 \approx 26.93$

- $Q_1 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i(\bar{x}_i^2 - 2\bar{x}_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{r} n_i\bar{x}_i^2 - n\bar{x}^2 =$

  $= 91.086 - 15\left(\frac{34}{15}\right)^2 \approx 14.02$

- $Q_2 = Q - Q_1 = 12.91$
- $F = \frac{Q_1/(r-1)}{Q_2/(n-r)} = \frac{14.02/2}{12.912/12} \approx 6.513$
- $\alpha_* = P\{F > 6.513\} = \mathrm{FDIST}(6.513, 2, 12) = 0.012 < 0.05$

Методики дают значимо различные результаты тестового контроля.


Comparing boxplot()s and non-robust mean +/- SD
0
1
2
3
4
5
1
2
3

Попарные сравнения

Для проверки гипотез $H_0 : \mu_i = \mu_k$ значимости отклонений внутригрупповых средних используем статистки

$$\tau = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{\sqrt{Q_2} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_k}}} \sqrt{n - r} \ \sim \ T(n - r)\,.$$

$$\xi_0 = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{\sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_k}}} \sim \mathcal{N}(0, 1), \qquad \eta = \frac{Q_2}{\sigma^2} \sim \chi^2(n - r)$$

$$\tau = \frac{\xi_0}{\sqrt{\frac{1}{n-r}\eta}} = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{\sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_k}} \sqrt{\frac{Q_2}{\sigma^2}}} \sqrt{n - r} \ \sim \ T(n - r)$$

$$H_0 : \mu_1 = \mu_2, \quad \tau_{1-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{Q_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sqrt{n - r} =$$

$$= \frac{(1.43 - 2.4)}{\sqrt{12.91} \cdot \sqrt{\frac{1}{7} + \frac{1}{5}}} \sqrt{15 - 3} = -1.6,$$

$$p_{1-2} = P\{\tau < -1.6\} = 0.068$$

$$H_0 : \mu_1 = \mu_3, \quad \tau_{1-3} = -3.59, \quad p_{1-3} = P\{\tau < -3.59\} = 0.0018$$

$$H_0 : \mu_2 = \mu_3, \quad \tau_{2-3} = -2.11, \quad p_{2-3} = P\{\tau < -2.11\} = 0.028$$

При альтернативе $H_1 : \mu_1 \neq \mu_1$ имеем
$p_{1-2} = P\{|\tau| > |-1.6|\} = 2 \cdot 0.068 = 0.136$. Вычисление доверительных уровней вероятности
$P\{\tau < -\tau_*\} = \mathrm{TDIST}(\tau_*, df, 1))$,
$P\{\tau > \tau_*\} = \mathrm{TDIST}(\tau_*, df, 1))$, $df = n - r = 15 - 3 = 12$

Поправки Бонферони, частный случай

Мы проверили три гипотезы о $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_2 = \mu_3$.
Число проверяемых гипотез обозначим через $\mathrm{m} = 3$.

- Пусть α_0 вероятность ошибки первого рода (отвергнуть правильную гипотезу) в каждой из этих гипотез.
- Обозначим через $\tilde{\alpha}$ вероятность ошибиться хотя бы в одной из них, $\tilde{\alpha} = 1 - (1 - \alpha_0)^{\mathrm{m}}$.
- Задача найти такое α_0, что $\tilde{\alpha} = 1 - (1 - \alpha_0)^{\mathrm{m}} < \alpha$.

$$(1 - \alpha_0)^{\mathrm{m}} > 1 - \alpha \iff 1 - \alpha_0 > (1 - \alpha)^{\frac{1}{\mathrm{m}}} \iff$$
$$\iff \alpha_0 < 1 - (1 - \alpha)^{\frac{1}{\mathrm{m}}} = \alpha_*$$

Например, при $\mathrm{m} = 3$, $\alpha = 0.05$ имеем $\alpha_0 < \alpha_* = 0.0169$. Сравниваем $\mathrm{p}_{1-2} = 0.068, \mathrm{p}_{1-3} = 0.0018, \mathrm{p}_{2-3} = 0.028$. В результате поправок получаем значимое различие только между 1 и 3 группами.

Поправки Бонферони, упрощение граничного значения

На практике предлагается вместо неравенства $\alpha_0 < 1-(1-\alpha)^{\frac{1}{\mathrm{m}}}$ рассматривать $\alpha_0 < \frac{\alpha}{\mathrm{m}}$. Например, $\frac{\alpha}{\mathrm{m}} = 0.01667 < \alpha_* = 0.0169$. Поэтому если рассматривать значения $\alpha_0 < \frac{\alpha}{\mathrm{m}}$, то из $\frac{\alpha}{\mathrm{m}} < \alpha_* = 1-(1-\alpha)^{\frac{1}{\mathrm{m}}}$ следует, не увеличится ошибка $\tilde{\alpha} = 1-(1-\alpha_0)^{\mathrm{m}} < \alpha$. Осталось убедиться

$\frac{\alpha}{\mathrm{m}} < 1-(1-\alpha)^{\frac{1}{\mathrm{m}}}$

$1-\frac{\alpha}{\mathrm{m}} > (1-\alpha)^{\frac{1}{\mathrm{m}}} \iff (1-\frac{\alpha}{\mathrm{m}})^{\mathrm{m}} > 1-\alpha$, так как $\mathrm{C}_{\mathrm{m}}^2\frac{\alpha^2}{\mathrm{m}^2} - \mathrm{C}_{\mathrm{m}}^3\frac{\alpha^3}{\mathrm{m}^3} + \mathrm{C}_{\mathrm{m}}^4\frac{\alpha^4}{\mathrm{m}^4} + \ldots > 0$, например, обозначим $\alpha' = \frac{\alpha}{\mathrm{m}}$ и при $\mathrm{m} = 3$ имеем $3\alpha'^2 - \alpha'^3 = \alpha'^2(3-\alpha') > 0$.

Поправки Бонферони

Пусть имеются нулевые гипотезы $\mathcal{H}_1, \ldots, \mathcal{H}_m$ с соответствующими p-значениями: $p_1, \ldots, p_m$. Например, $m = C_r^2$, если проверяются гипотезы о равенстве всех попарных средних. Определим $\tilde{\alpha}$ как вероятность совершить хотя бы одну ошибку в выводе из m выводов. Если осуществляется m независимых испытаний с вероятностью ошибки α_0 в каждом, то $\tilde{\alpha} = 1 - (1 - \alpha_0)^m$. В случае зависимых испытаний $\tilde{\alpha} \leqslant 1 - (1 - \alpha_0)^m$. Заметим, что по неравенству Буля

$$\tilde{\alpha} = 1 - (1 - m\alpha_) + C_m^2 \alpha_0^2 - \ldots) \leqslant m \cdot \alpha_0.$$

Таким образом, если мы хотим, чтобы вероятность наличия хотя бы одного неверного вывода из m была равна $\tilde{\alpha} = 0.05$, то достаточно установить вероятность неверного отклонения нулевой гипотезы равной $\alpha_0 = \tilde{\alpha}/m$ для каждого вывода.

Неравенство Буля: $P(A_1 + \ldots A_n) \leq P(A_1) + \ldots P(A_n)$ следует из определения вероятности суммы событий

$P(A + B) = P(A) + P(B) - P(AB)$.

Задача на множественные сравнения

Проверим однородность данных о тяжести социальной зависимости asi2 emp у наркоманов в зависимости от фактора educat образования: 1 - неполное среднее образование, 2 - среднее, 3 - неполное высшее, 4 - высшее. Нужно вычислить средние с ошибками среднего в каждой группе и заполнить таблицу соответствующими значимостями. Принять $\alpha = 0.1$. $Q_2 = 19.299$.

p	1	2	3	4
1				
2				
3				
4				

Задача на множественные сравнения, проверка

Проверим однородность данных о тяжести социальной зависимости asi2 emp у наркоманов в зависимости от фактора educat образования: 1 - неполное среднее образование, 2 - среднее, 3 - неполное высшее, 4 - высшее. Заполнить таблицу соответствующими значимостями.
$Q_2 = 19.299$, $n = 278$, $r = 4$.

$$\tau_{1-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{Q_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sqrt{n - r}$$

p	1	2	3	4
1 $0.760 \pm 0.073(21)$		0.21	0.1	0.11
2 $0.807 \pm 0.017(221)$			0.0038	0.028
3 $0.662 \pm 0.062(27)$				0.39
4 $0.634 \pm 0.108(9)$				

$\alpha_0 < \frac{0.05}{6} = 0.008.$

Пример вычислений в R

```
x <- c(1, 3, 2, 1, 0, 2, 1, 2, 3, 2, 1, 4, 4, 5, 3)
gr <- c(rep(1, 7), rep(2, 5), rep(3, 3))
df <- data.frame(x = x, gr = as.factor(gr))
ao <- aov(x ~ gr, df)
summary(ao)
```

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
gr	2.000	14.019	7.010	6.513	0.012
Residuals	12.000	12.914	1.076		

Для парных сравнений используем pairwise.t.test(dfx, dfgr, "none")

Попарные сравнения

```
pairwise.t.test(df$x, df$gr, "none")
```

```
Pairwise comparisons using t tests with pooled SD
data : df$x and df$gr
     1       2
2 0.1358  -
3 0.0037  0.0563
P value adjustment method :  none
```

```
pairwise.t.test(df$x, df$gr, "bonf")
```

```
Pairwise comparisons using t tests with pooled SD
data : df$x and df$gr
     1       2
2    0.407   - 3    0.011   0.169
P value adjustment method :  bonferroni
```