Однофакторный дисперсионный анализ

Грауэр Л.В.

Дисперсионный анализ (ANOVA)

Однофакторный

Многофакторный

Одномерный

Многомерный

Модель

$$x_{i,j} = m_i + a_{i,j} + \varepsilon_{i,j}$$

Однофакторный дисперсионный анализ

1 фактор, $k \geq 2$ значений

Выборки случайны и взаимно независимы

Генеральные совокупности имеют нормальный з.р.

Генеральные совокупности имеют равные дисперсии

$$
\begin{array}{cccc}
X_{11} & X_{12} & \dots & X_{1k} \\
X_{21} & X_{22} & \dots & X_{2k} \\
\vdots & \vdots & & \vdots \\
X_{n_1 1} & X_{n_2 2} & \dots & X_{n_k k} \\
N(a_1, \sigma) & N(a_2, \sigma) & \dots & N(a_k, \sigma)
\end{array}
$$

Гипотезы

$$H_0 : a_1 = a_2 = \ldots = a_k$$

$$H_1 : \exists t, s : a_t \neq a_s$$

$$N = \sum_{j=1}^{k} n_j$$

Внутригрупповое среднее

$$\bar{X}_j =$$

Общее среднее

$$\bar{X} =$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 =$$

Статистика S_1

$$\sum_{i=1}^{n_j} \frac{(X_{ij} - \bar{X}_j)^2}{\sigma^2}$$

$$S_1 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{(X_{ij} - \bar{X}_j)^2}{\sigma^2}$$

$$\frac{(\bar{X}_j - a_j)}{\sigma} \sqrt{n_j}$$

Статистика S_2

Пусть H_0 верна

$$S_2 = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 =$$

$$= \sum_{j=1}^{k}\left[\frac{\sqrt{n_j}(\bar{X}_j - a)}{\sigma}\right]^2 - \left[\sum_{j=1}^{k}\sqrt{\frac{n_j}{N}}\frac{\sqrt{n_j}(\bar{X}_j - a)}{\sigma}\right]^2$$

Статистика критерия

S_1 и S_2 взаимно независимы.

$$F = \frac{S_2/(k-1)}{S_1/(N-k)}$$

Критическая область для α:

$$V_k = (u_{1-\alpha,k-1,N-k}; \infty)$$

Пример

The effect of iris color on critical flicker frequency. Journal of General Psychology

Colour	Flicker							
Brown	26.8	27.9	23.7	25	26.3	24.8	25.7	24.5
Green	26.4	26.2	28	26.9	29.1			
Blue	25.7	27.2	29.9	28.5	29.4	28.3		

Метод линейных контрастов

Линейный контраст Lk — линейная комбинация

$$Lk = \sum_{j=1}^{k} c_j a_j,$$

где c_j, $j = 1, \ldots, k$: $\sum_{j=1}^{k} c_j = 0$.

$\hat{Lk} =$

$D(\hat{Lk}) =$

Лемма (Метод Шеффе)

Для любой совокупности векторов $(c_1, \dots, c_k)$: $\sum_{j=1}^k c_j = 0$, вероятность одновременного выполнения неравенств

$$\left| \sum_{j=1}^{k} c_j (a_j - \bar{X}_j) \right| < S_{\hat{L}k} \sqrt{(k-1) u_{1-\alpha, k-1, N-k}}$$

не меньше $1 - \alpha$.

Проверка гипотезы

$$H_0 : Lk = 0$$

$$H_1 : Lk \neq 0$$

$$\Big(\hat{Lk} - S_{\hat{Lk}}\sqrt{(k-1)u_{1-\alpha,k-1,N-k}},$$
$$\hat{Lk} + S_{\hat{Lk}}\sqrt{(k-1)u_{1-\alpha,k-1,N-k}}\Big)$$

Пример линейных контрастов

$$H_0 : a_1 = a_2 = a_3$$

$$H_0^{12} : a_1 = a_2$$
$$H_0^{13} : a_1 = a_3$$
$$H_0^{23} : a_2 = a_3$$

Критерий Краскела-Уоллиса

1 фактор, $k > 2$ значений

- Выборки случайны и взаимно независимы
- Генеральные совокупности имеют непрерывные функции распределения $F_1,\ F_2, \dots,\ F_k$

$$\begin{array}{cccc}
X_{11} & X_{12} & \dots & X_{1k} \\
X_{21} & X_{22} & \dots & X_{2k} \\
\vdots & \vdots & & \vdots \\
X_{n_1 1} & X_{n_2 2} & \dots & X_{n_k k}
\end{array}$$

Проверяемая гипотеза

$$H_0: \; F_1(x) = F_2(x) = \ldots = F_k(x) \text{ для всех } x \in \mathbb{R}.$$

$$H_1: \; F_1(x) = F_2(x - \delta_2) = \ldots = F_k(x - \delta_k) \text{ для всех } x \in \mathbb{R}$$

$$N = \sum_{j=1}^{k} n_j$$

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(N)}$$

$$R_{ij} = rank(X_{ij})$$

Статистика критерия

$$\bar{R}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij} \quad , \bar{R} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} R_{ij} = \frac{N+1}{2}$$

$$Q_2 = \sum_{j=1}^{k} n_j(\bar{R}_j - \bar{R})^2$$

$$H = \frac{12}{N(N+1)} Q_2 = \frac{12}{N(N+1)} \left(\sum_{j=1}^{k} n_j \bar{R}_j^2\right) - 3(N+1)$$

$k = 3$ и $n_j \geq 5$

$k > 3$ и $n_j \geq 4$

Модификации

При наличии одинаковых значений величин из разных выборок

$$H^* = H \left\{ 1 - \left(\sum_{j=1}^{q} \frac{t_j^3 - t_j}{N^3 - N} \right) \right\}^{-1}$$

При больших N используют аппроксимацию Имана-Давенпорта

$$J = \frac{H}{2} \left(1 + \frac{N - k}{N - 1 - H} \right)$$

$$V_k = \{ J > \{ (k-1) F_{1-\alpha}(k-1; N-k) + \chi^2_{1-\alpha}(k-1) \} \}$$