Course: Computer statistics

Lecture 5: One-factor variance analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 5. Однофакторний дисперсійний аналіз.

I. Теоретичні відомості

II. Приклад застосування

I. Теоретичні відомості

Однофакторний дисперсійний аналіз вирішує задачу перевірки гіпотези про наявність розбіжностей між групами обє'ктів, які об'єднані у відповідні групи за рівнями певного фактору. Перевіряється розбіжність між середніми у групах, але робиться це на підставі аналізу дисперсій. Задачу однофакторного дисперсійного аналізу можна сформулювати наступним чином.

Нехай спостерігається вплив певного фактора на певну числову змінну. Приміром, вплив на тривалість життя людини певної марки цигарок. Результатом спостережень є l вибірок обсягами:

$$n_k, k = \overline{1,l}, n = \sum_{k=1}^{l} n_k,$$

де n - загальна кількість спостережень.

Кожна з вибірок формується за одного певного сталого рівня фактору. Позначимо

x_{ik} i-те значення k - тої групи, $k = \overline{1,l}, i = \overline{1,n_k}$.

Тоді, в основі однофакторного дисперсійного аналізу лежить наступна теоретико- ймовірнісна схема:

$$x_{ik} = m_k + \varepsilon_{ik}, k = \overline{1,l}, i = \overline{1,n_k},$$

де $m_k, k=\overline{1,l}$ - середні кожній групі, а ε_{ik} - нормально розподілені випадкові величини з нульовим математичним сподіванням та однаковою невідомою дисперсією σ^2 .

Основна гіпотеза - $H_0 = \{m_1 = m_2 = ... = m_l\}$,тобто, групові середні є рівними між собою.

Позначмо x_{ik} i-те значення k - тої групи вибірки, $k=\overline{1,l}, i=\overline{1,n_k}$.

Введемо групові середні

$$\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}, k=\overline{1,l} .$$

Також загальне вибіркове середнє:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{l} n_k \bar{x}_k = \frac{1}{n}\sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik} .$$

Введемо також наступні вибіркові характеристики :

$Q = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2$ - загальна сума квадратів відхилень спостережень від загального середнього ;

$Q_1 = \sum_{k=1}^{l} n_k(\bar{x}_k - \bar{x})^2$ - сума квадратів відхилень вибіркових середніх по групах $\bar{x}_k$ від загального середнього $\bar{x}$ (між групами);

$Q_2 = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2$ - сума квадратів відхилень спостережень від групових середніх.

Основна тотожність дисперсійного аналізу має вигляд:

$$Q = Q_1 + Q_2 .$$

Співвідношення між Q_1 і Q_2 може свідчити про те, чи є розбіжність між середніми у групах, чи ні. Перша Q_1 пов'язана з розкидом даних у вибірці, а друга Q_2 – з відмінністю середніх значень.

Якщо гіпотеза H_0 є **правдивою**, то :

1) Статистики $S_1^2 = \frac{Q_1}{l-1}$ та $S_2^2 = \frac{Q_2}{n-l}$ є незсуненими оцінками дисперсії σ^2;
2) Їх відношення $\frac{S_2^1}{S_2^2} = \frac{Q_1/(l-1)}{Q_2/(n-l)} \sim F(l-1, n-l)$ має розподіл **Фішера** з (l-1) та (n-l) степенем свободи.

Отже, критерій будується наступним чином:

Основна гіпотеза $H_0 = \{m_1 = m_2 = ... = m_l\}$

Статистка критерію $\pi(x) = \frac{Q_1/(l-1)}{Q_2/(n-l)}$.

Тоді, при заданому рівні значущості(помилки першого роду) α критерій працює так:

- гіпотеза H_0 приймається, якщо $\pi(X) < F_{1-\alpha}(l-1, n-l)$;
- гіпотеза H_0 відхиляється, якщо $\pi(X) \geq F_{1-\alpha}(l-1, n-l)$,

де $F_{1-\alpha}(l-1, n-l)$- квантіль розподілу Фішера з (l-1) та (n-l) степенем свободи рівня $1-\alpha$.

Відхилення гіпотези означає, що серед середніх є, принаймні, два не рівних між собою.

II. Приклад застосування

Дисперсійний аналіз реалізовано R функцією:

aov(formula,…)

Розглянемо приклад, у якому є дані про масу кущів томатів (**weight**, у кг), які вирощували протягом 2 місяців при трьох різних експериментальних умовах, рівнях фактору **trt**, (treatment), - при поливі **водою**, у середовищі з додаванням **добрива**, а також у середовищі з додаванням **добрива** та **гербіциду**:

Приклади:

```
#  Вага томатів залежить від трьох факторів
>tomato <- data.frame(weight=
            c(1.5, 1.9, 1.3, 1.5, 2.4, 1.5,   # вода
              1.5, 1.2, 1.2, 2.1, 2.9, 1.6,    # добрива
              1.9, 1.6, 0.8, 1.15, 0.9, 1.6), # гербіцид
             trt = rep(c("вода", "добр", "герб"),
                  c(6, 6, 6))) # складаємо таблицю даних згрупованих за
                               # фактором
> tomato$trt<-as.factor(tomato$trt)

>tomato$trt <- relevel(tomato$trt, ref = "вода") #  базовий рівень
>attach(tomato)
```

Щоб краще зрозуміти властивості наявних даних, зобразимо їх за допомогою одновимірної діаграми розсіювання (Рис.1):

Приклади:

```
>stripchart(weight ~ trt, xlab = "Вага, кг", ylab = "Фактори")
```

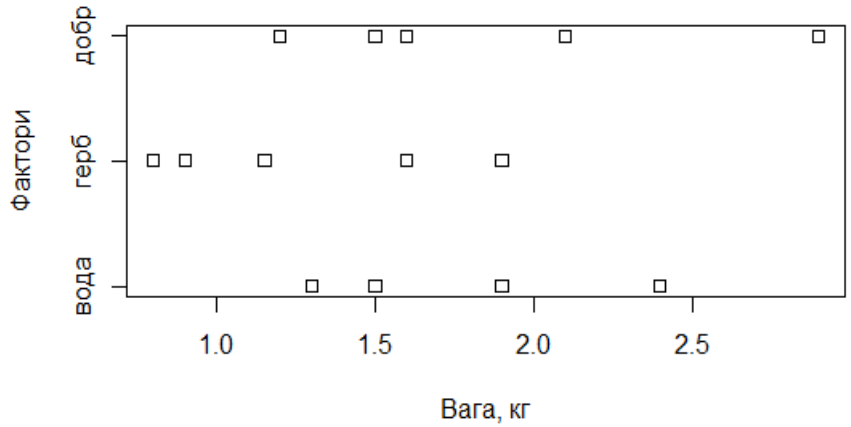

Рис. 1 (Рисунок згенеровано **RStudio**).

Обчислимо середні для кожної групи.

```
> Means <- tapply(weight, trt, mean)  # середні значення
> Means
     Water      Nutrient     Nutrient+24D
  1.683333     1.750000     1.325000
```

Проведемо дисперсійний аналіз.

```
> summary(aov(weight ~ trt, data = tomato))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	**2**	**0.627**	**0.3135**	**1.202**	**0.328**
Residuals	**15**	**3.912**	**0.2608**		

Оскільки **P- value=0.328**, то немає підстав відхиляти гіпотезу про рівність середніх.

Стовпець **Sum Sq** містить значення Q_1 *і* Q_2,

а стовпець **Mean Sq** - S^2_1 *і* S^2_2.

Про відсутність розбіжності між середніми свідчить рисунок 2, на якому зображені «скрині з вусиками», які очевидно перетинаються.

Приклади:

```
> a<-weight[1:6]     # вода
> b<-weight[7:12]  # добр
> c<-weight[13:18] # герб
> x<-list(a,b,c)
> boxplot(x,names=c("вода","добр","герб"))
```

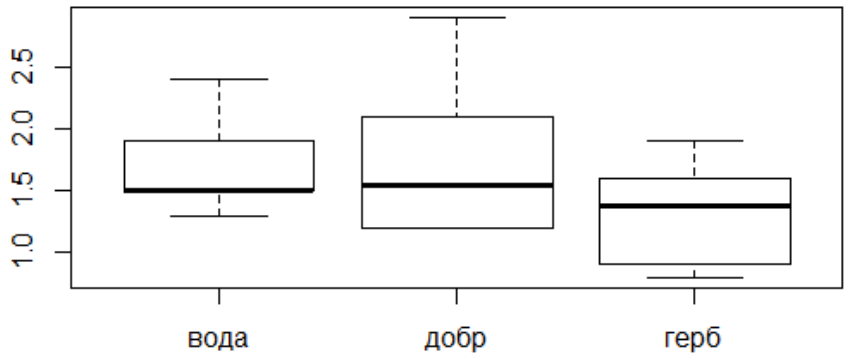

Рис.2 (Рисунок згенеровано **RStudio**).

Також про відсутність розбіжностей між середніми свідчать і довірчі інтер вали, які зображено на рисунку 3.

Приклади:

```
>install.packages("plotrix")
> library("plotrix")
> install.packages("rcompanion")
> library("rcompanion")
> CI<-groupwiseMean(weight ~ trt,data=tomato,conf=(0.95)^(1/3)) # групо
#ві середні
> plotCI(1:3,y=CI$Mean,ui=CI$Trad.upper,li=CI$Trad.lower,
+        xlab=" ",ylab="counts",xaxt="n") #довірчі інтервали
>  axis(1,at=1:3,labels=levels(CI$trt))
```

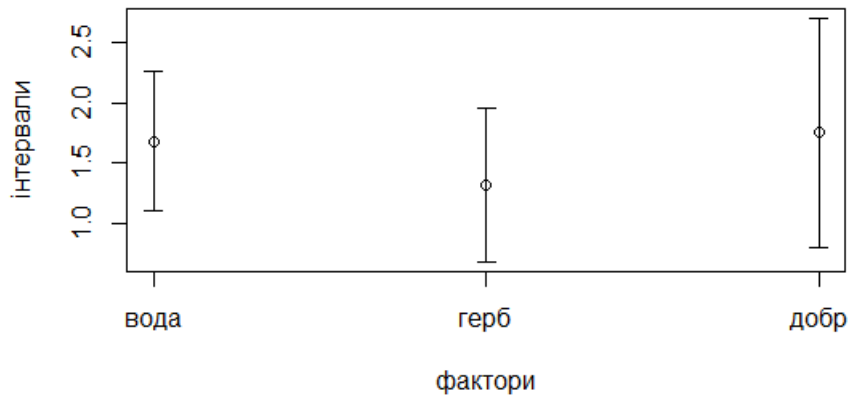

Рис. 3 (Рисунок згенеровано **RStudio**).

Як бачимо з рисунку, 3 довірчі інтервали перетинаються. Тобто, середні можна вважати рівними.