Course: Computer statistics

Lecture 6: Two-factor variance analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 6. Двофакторний дисперсійний аналіз.

I. Теоретичні відомості

Нехай на ознаку, що досліджується, роблять вплив одночасно два фактори, приміром, Ф і Z . Припустимо, що є p рівнів фактору Ф і q рівнів фактору Z. Отже, спостережуване значення x_{ijk} залежить від трьох змінних: i – рівня фактору Ф, j –рівня фактору Z і k – самого номеру спостереження.

Тоді модель набуває вигляду:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, i = \overline{1, p},\ j = \overline{1, q}, k = \overline{1, n},$$

де μ - загальне математичне сподівання;

α_i , β_j - частини впливу факторів Ф і Z;

γ_{ij} - частина їхньої взаємодії;

ε_{ijk} - похибка, або частина впливу неврахованих факторів (неспецифічна компонента), ε_{ijk} - нормально розподілені випадкові величини з нульовим середнім та однаковою невідомою дисперсією σ^2.

Двофакторний дисперсійний аналіз реалізує та ж сама функція, що й однофакторний:

aov(формула, data=….)

У випадку однофакторного аналізу залежність вказується так:

>aov(x ~factor1, data =…..) .

Для двох незалежних (без урахування взаємодії) факторів:

>aov(x ~factor1+ factor2, data =…..) .

Для двох факторів з урахуванням взаємодії :

>aov(x ~factor1+ factor2+ factor1: factor2 , data =…..) .

Зрозуміло, що формула розповсюджується й на випадок довільної кількості факторів та їх комбінацій:

>aov(x ~factor1+ factor2+ factor3+…+factor3: factor2 , data =…..) .

Для повторних спостережень(для однієї комбінації факторів декілька вимірювань ознаки):

>aov(x ~factor1+ Error(subject/IV1), data =…..)

II. Приклад застосування

Розглянемо, як це працює на прикладі. Візьмемо дані про вартість українських та імпортних продуктів харчування у великих і маленьких маркетах.

Приклади:

```
> mydata <- read.csv('C:... /shops.csv')
> mydata
       food    price      store           origin
1   chocolate 100.30  supermarket  ukraine
2   chocolate   55.57  minimarket    ukraine
3   chocolate  268.62  minimarket   import
4   chocolate  196.81  supermarket import
5      bread    10.91    minimarket   ukraine
6      bread    25.84    supermarket ukraine
7      bread    35.44    supermarket import
8  vegetables  64.93   minimarket   ukraine
9      bread    116.23  minimarket   import
10 vegetables 226.39 supermarket  ukraine
11 vegetables 209.15 supermarket  import
```

```
12 vegetables 359.00  minimarket   import
13     fruits   85.84   minimarket   ukraine
14     fruits  191.07   supermarket import
15     fruits  208.59   supermarket ukraine
16     fruits  400.17   minimarket   import
17     cheese  184.56  minimarket   ukraine
18     cheese  278.86  supermarket ukraine
19     cheese  333.21  supermarket import
20     cheese  504.00  minimarket   import
```

Проведемо двофакторний дисперсійний за двома незалежними факторами: «origin» і «store».

Приклади:

```
> fit2 <- aov(price ~ origin + store, data=mydata)
> summary(fit2)
            Df Sum Sq Mean Sq F value Pr(>F)
origin       1  94107   94107   6.355  0.022 *
store        1   2981    2981    0.201  0.659
Residuals   17 251749   14809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Як видно з резюме, різниця між цінами зумовлена фактором «origin» (P-value=0.022), а фактор «store» не впливає на ціну (P-value=0.659).

За допомогою команди

```
> model.tables(fit2,"means")
```

отримуємо середні значення

```
Tables of means
Grand mean
```

192.7745 – глобальне середнє

```
 origin
 import ukraine
 261.37  124.18
```

– групові середні по країнах

```
store
```

minimarket supermarket
204.98 180.57 - групові середні по маркетах

Очевидно, що різниця по країнах є суттєвішою.

За допомогою **ggplot2** збудуємо наступну діаграму (Рис.1).

Приклади:

```
> pd = position_dodge(0.1)
> ggplot(mydata, aes(x = store, y = price, color = origin, group = origin)) +
+   stat_summary(fun.data = mean_cl_boot, geom = 'errorbar', width = 0.2, lwd = 0.8, position = pd)+
+         stat_summary(fun.data = mean_cl_boot, geom = 'line', size = 1.5, position = pd) +
+         stat_summary(fun.data = mean_cl_boot, geom = 'point', size = 5, position = pd, pch=15) +
+         theme_bw()
```

Отримаємо:

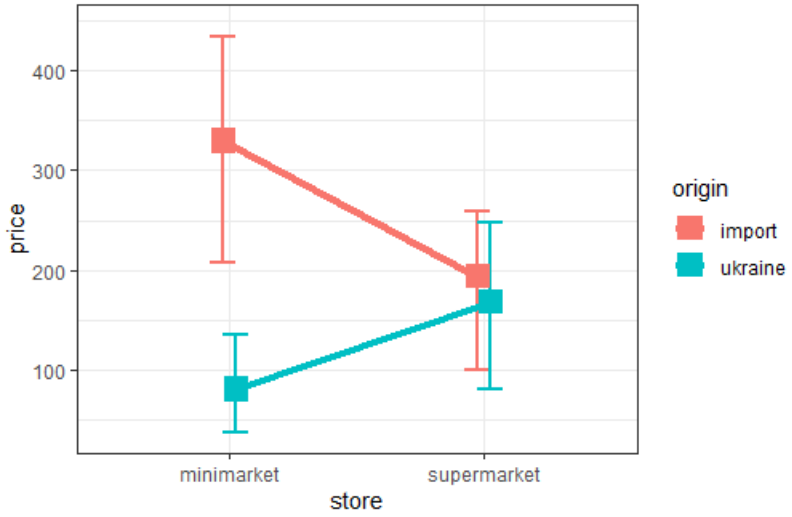

Рис.1 (Рисунок згенеровано **RStudio**).

На діаграмі зображено ціни вітчизняні та імпортні продукти. Як бачимо у цілому вітчизняні продукти дешевше імпортних, причому різниця у мінімаркетах є суттєвішою ніж у великих маркетах. Для з'ясування, чи є ця різниця значущою, скористаємось дисперсійним аналізом з урахуванням взаємодії.

Приклади:

```
> fit3 <- aov(price ~ origin + store + origin:store, data=mydata)
> summary(fit3)
                Df Sum Sq Mean Sq F value Pr(>F)
origin       1  94107   94107   7.968 0.0123 *
store        1   2981    2981   0.252 0.6222
origin:store  1  62777   62777   5.315 0.0349 *
Residuals    16 188971   11811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Знову бачимо, що на різницю цін впливає походження продукту та взаємодія між походженням та місцем продажу. Зауважимо, що ті ж самі результати можна було отримати за допомогою команди
> fit3 <- aov(store * origin, data=mydata).

Тепер з’ясуємо, як розрізняються ціни на продукти, якщо порівнювати їх попарно. Для цього побудуємо «скрині з вусами» для всіх видів продукції.

Приклади:

```
> ggplot(mydata, aes(x = food, y = price)) +
+   geom_boxplot()
```

Отримаємо:

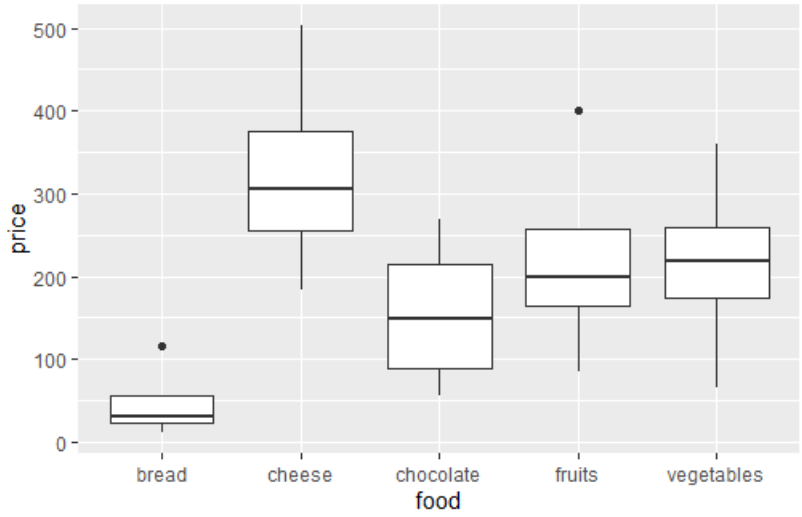

Рис.2 (Рисунок згенеровано **RStudio**).

Як бачимо з рисунку 2, найдорожчим є сир, а найдешевшим- хліб. Але наявність статистично значущої розбіжності можна зробити на підставі ди сперсійного аналізу. А саме,

```
> fit4 <- aov(price ~ food, data=mydata)
> summary(fit4)
             Df Sum Sq Mean Sq F value Pr(>F)
food          4 165823   41456  3.398 0.0362 *
Residuals    15 183013   12201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

III. Алгоритм Тьюкі.

Як бачимо, статистична розбіжність є (**P-value=0.0362**). Для того, щоб з’ясувати між саме якими продуктами є значуща розбіжність застосуємо ал горитм Тьюкі, який є найбільш поширеним(***honestly significant difference te st***), який реалізує функція **TukeyHSD().** HSD тест задає найменшу величин у різниці математичних сподівань в групах, яку можна вважати значущою, а також дозволяє розрахувати її довірчі інтервали з урахуванням числа вик онаних порівнянь.

Критерій Тьюки використовується для перевірки нульової гіпотези $H_0 = \{m_i = m_j\}$ проти альтернативної гіпотези $H_1 = \{m_i \neq m_j\}$, для будь-яки х пар індексів, які позначають будь-які дві порівнювані групи. При наявнос ті n груп треба виконати n (n - 1) / 2 попарних порівнянь. Порівняння відбу вається шляхом застосування модифікації критерія Стьюдента порівняння середніх.

Перший крок полягає в упорядкуванні всіх наявних групових середніх з начень по зростанню (від 1 до l). Далі виконують попарні порівняння цих середніх так, що спочатку порівнюють найбільше середнє з найменшим, то бто l -е з 1-им, потім l-е з другим, третім, і т.д. аж до (l - 1) -го. Потім перед останнє середнє, (l - 1) -е, тим же чином порівнюють з першим, другим, і т .д. до (l - 2) -го. Ці порівняння тривають до тих пір, поки не будуть перебра ні всі пари. Тоді, за умови рівності міжгрупових середніх статистика,

$$\frac{\left|\overline{x_i} - \overline{x_j}\right|}{\sqrt{\frac{Q_2}{2(n-l)}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

має розподіл Стьюдента з $\left(n_i + n_j - 2\right)$ степенем свободи.

Приклади:

```
> TukeyHSD(fit4)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price ~ food, data = mydata)

$food
                          diff        lwr       upr     p adj
cheese-bread           278.0525   36.86938 519.23562 0.0204058
chocolate-bread        108.2200 -132.96312 349.40312 0.6453667
fruits-bread           174.3125  -66.87062 415.49562 0.2209202
vegetables-bread       167.7625  -73.42062 408.94562 0.2512881
chocolate-cheese      -169.8325 -411.01562  71.35062 0.2413687
fruits-cheese         -103.7400 -344.92312 137.44312 0.6789317
vegetables-cheese     -110.2900 -351.47312 130.89312 0.6297401
fruits-chocolate        66.0925 -175.09062 307.27562 0.9117335
vegetables-chocolate    59.5425 -181.64062 300.72562 0.9375222
vegetables-fruits       -6.5500 -247.73312 234.63312 0.9999874
```

Як бачимо, статистично значущою є розбіжність між сиром і хлібом (останній стовпець, **P-value=0.0204058**).