Course: Computer statistics

Lecture 7: Correlation analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 7. Аналіз кореляцій.

I. Діаграми розіювання

II. Коефіцієнт кореляції Пірсона

III. Коефіцієнт кореляції Спірмена

IV. Коефіцієнт кореляції Кендалла

Анализ залежностіей, навіть у найпростішому його вигляді, є основою багатовимірного статистичного аналізу.

I. Діаграми розіювання

Первинне уявлення про структуру даних дають діграми розсіювання, на яких по одній з осей відкладають одну змінну, а по іншій – другу.
Як простий приклад розглянемо залежність ваги тварини від ваги споживання корму на день.

Приклади:

```
> cor<- data.frame(w=c(1.5,1.9,1.3,1.5,2.1,2.2,2.4,2.5,2.3,2.4),#вага
+                  g=c(1.5,1.2,1.2,1.4,1.5,1.7,1.9,1.8,2.0,2.2)#корм
+                 )
>
> plot(cor$w~cor$g,xlab="корм", ylab="вага",col="green")
> abline(lm(cor$w~cor$g),col="red")
```

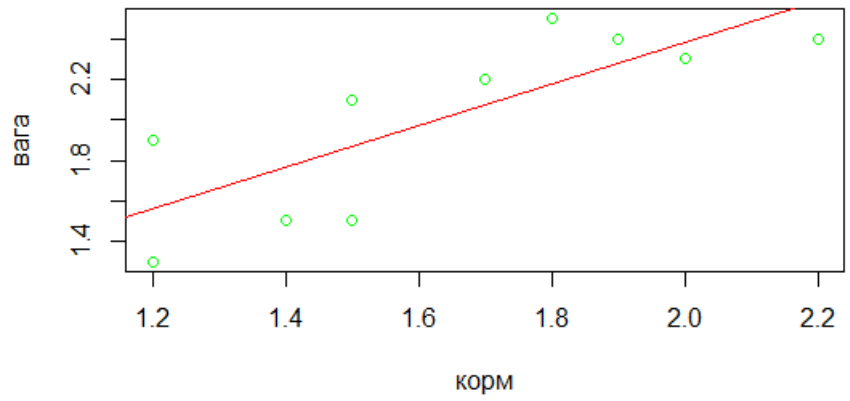

Рис.1(Рисунок згенеровано **RStudio**).

Як бачимо з рисунку 1, між змінними є очевидна лінійна залежність.

Можна розглядати зразу декілька змінних. Розглянемо стандартний набір **mtcars,** який містить числові технічні дані про автомобілі.

Приклади:

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Виберемо з цього набору декілька компонент та розглянемо між ними залежність (рис.2).

Приклади:

```
> df  <- mtcars
> df_numeric  <- df[, c(1,3:7)]
> pairs(df_numeric)
```

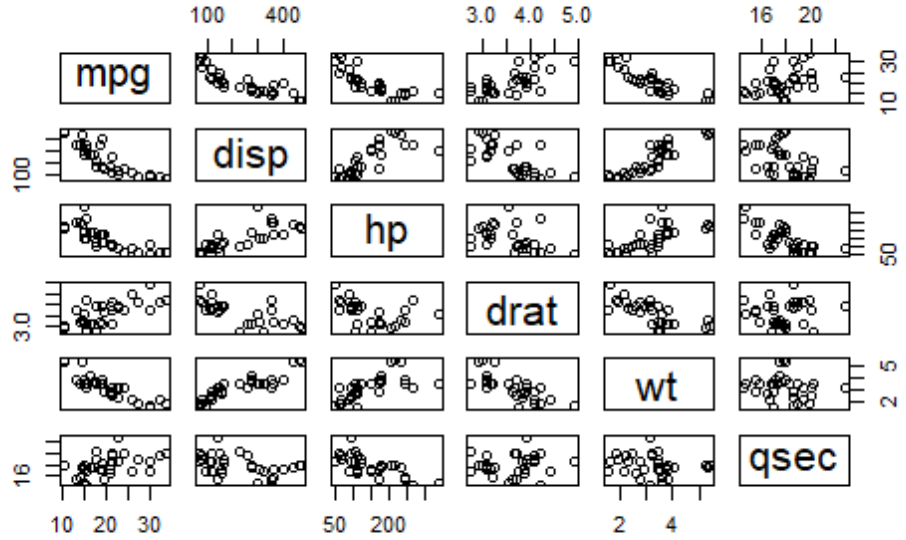

Рис.2 (Рисунок згенеровано **RStudio**).

II. Коефіцієнт кореляції Пірсона

Коефіцієнт кореляції Пірсона - це найбільш поширена кількісна міра залежності між двома числовими змінними.

Нехай є дві вибірки $X = (x_1, x_2, ..., x_n)$ та $Y = (y_1, y_2, ..., y_n)$ з двох генеральних сукупностей, які визначаються випадковими величинами x та y.

Введемо вибіркову коваріацію.

Вибіркова коваріація :

$$Cov_{xy} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$$

де $\bar{x}, \bar{y}$ — вибіркові середні значення відповідних змінних,

$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ — вибіркові дисперсії відповідних змінних,

n — об’єм вибірки.

Вибірковим коефіцієнтом кореляції Пірсона називають

$$\rho_{xy} = \frac{Cov_{xy}}{\sqrt{S_x^2} \cdot \sqrt{S_x^2}}.$$

Очевидно, що вибірковий коефіцієнт кореляції, як аналог теоретичного, задовольняє наступні умови:

1. $|\rho_{xy}| \le 1,$
2. $\rho_{xy} = \pm 1 \Leftrightarrow y_i = a_0 + a_1 x_i, i = \overline{1,n}$, тобто, має місце жорсткий лінійний зв'язок між елементами вибірки.

Обчислює коефіцієнт кореляції Пірсона функція

cor(x, y = NULL, use = "everything",
method = c("pearson", "kendall", "spearman")).

Для даних, які було розглянуто вище, коефіцієнт кореляції обчислюється так:

Приклади:

```
> cor(cor$w,cor$g,  method ="pearson")
[1] 0.7982573
```

Як бачимо, він є достатньо високим і додатним. Очевидно, що збільшення корму збільшує вагу. Водночас можна перевірити коефіцієнт кореляції на рівність нулеві. Критерій побудовано на підставі того, що за правдивості основної гіпотези, статистика $\sqrt{n-2} \cdot \frac{\rho_{xy}}{\sqrt{1-\rho^2_{xy}}}$ має розподіл Стьюдента с $n-2$ степенем свободи. Це робить функція **cor.test.**

Приклади:

```
> cor.test ( cor$w,cor$g,alternative="two.sided", method="pearson" )

        Pearson's product-moment correlation
```

```
data:  cor$w and cor$g
t = 3.7485, df = 8, p-value = 0.005636
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3390276 0.9502729
sample estimates:
      cor
0.7982573
```

Оскільки **p-value = 0.005636** є достатньо малим, то основна гіпотеза про рівність коефіцієнта кореляції нулю відхиляється.

Також можна зразу обчислювати кореляції для набору з декількох величин(кореляційну матрицю):

Приклади:

```
> cor(df_numeric)
            mpg       disp         hp        drat         wt
mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594
disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799
hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479
drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406
wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000
qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159
            qsec
mpg   0.41868403
disp -0.43369788
hp   -0.70822339
drat  0.09120476
wt   -0.17471588
qsec  1.00000000
```

Знайдені кореляції можна перевірити на значущість. Нижче таблиця значень **p-value**:

Приклади:

```
> fit  <- corr.test(df_numeric)
> fit$p
              mpg          disp            hp          drat            wt
mpg  0.000000e+00 1.219442e-08 1.966619e-06 1.243368e-04 1.811542e-09
disp 9.380327e-10 0.000000e+00 8.571214e-07 4.784260e-05 1.833479e-10
hp   1.787835e-07 7.142679e-08 0.000000e+00 4.994386e-02 2.487496e-04
drat 1.776240e-05 5.282022e-06 9.988772e-03 0.000000e+00 4.784260e-05
wt   1.293959e-10 1.222320e-11 4.145827e-05 4.784260e-06 0.000000e+00
qsec 1.708199e-02 1.314404e-02 5.766253e-06 6.195826e-01 3.388683e-01
             qsec
mpg  0.0525761455
disp 0.0525761455
hp   0.0000478426
drat 0.6777365683
wt   0.6777365683
qsec 0.0000000000
```

III. Коефіцієнт кореляції Спірмена

Якщо між величинами є нелінійна залежність, то в якості критеріїв оцінки незалежності можуть застосовуватися й інші коефіцієнти кореляції, а саме, коефіцієнт **рангової кореляції Спірмена**, що дозволяє оцінити нелінійну, але монотонну залежність: в цьому випадку обчислюється кореляції не самих значень, а їх рангів (порядкових номерів при упорядкуванні).

Якщо позначити R_j^X ранг j -го елемента (номер місця за зростанням) у вибірці X, то **коефіцієнт кореляції Спірмена** визначається наступним чином:

$$\rho s_{xy} = \frac{\sum_{i=1}^{n} \left(R_i^X - \overline{R^X}\right)\left(R_i^Y - \overline{R^Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(R_i^X - \overline{R^X}\right)^2 \left(R_i^Y - \overline{R^Y}\right)^2}},$$

де $\overline{R^X}$, $\overline{R^Y}$ - середні значення рангів.

Перевірка за кількома критеріями може бути використана для приблизної оцінки оцінки виду залежності: якщо ранговая кореляція велика (статистично значуща), а лінійна - маленька (статистично незначуща), то залежність нелінійна; якщо обидві кореляції великі, то залежність лінійна; якщо обидві кореляції маленькі, що якої залежності немає, або вона немонотонна.

Приміром, якщо розглянути наступні дві вибірки оцінок двох експертів 5-ти конкурсантів за десятибальною шкалою, то, як видно з рисунку 3, явної лінійної залежності немає, коефіцієнт кореляції Пірсона не є високим, а зв'язок між оцінками «відловлює» коефіцієнт кореляції Спірмена.

Приклади:

```
> cor<- data.frame(a=c(4.8, 3.0,  2.2 , 7  ,2.8 ),#оцінка А
+                  b=c(8.5, 3.7 , 2.5 , 5  ,3.4 )  #оцінка В
+                   )
> plot(cor$a~cor$b,xlab="оцінка А", ylab="оцінка В",col="green")
> cor(cor$a,cor$b,method ="pearson")
[1] 0.568674
> cor(cor$a,cor$b  ,method ="spearman")
[1] 0.9
```

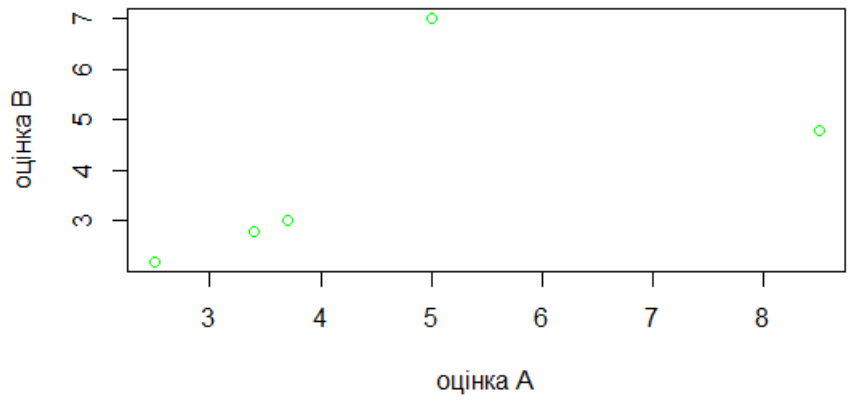

Рис. 3 (Рисунок згенеровано **RStudio**).

Аналогічно до перевірки значущості коефіцієнта кореляції Пірсона можна перевірити на значущість коефіцієнт кореляції Спірмена за допомогою функції **cor.test ()**.

IV. Коефіцієнт кореляції Кендалла

Нарешті, третій варіант коефіцієнта кореляції – **це коефіцієнт рангової кореляції Кендалла τ.** Обчислюється він у такий спосіб.

Припустимо, що ми маємо набір з парних спостережень двох змінних: $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$. Кажуть, що дві пари спостережень $(x_i, y_i),(x_j, y_j)$ є **конкордантними**, якщо є узгодженість між рангами відповідних елементів цих пар, тобто, якщо

- або $x_i > x_j$, $y_i > y_j$,
- або $x_i < x_j$, $y_i < y_j$.

Якщо знайти число конкордантних пар (nconc) і число дискордантних пар (ndiscord), то коефіцієнт кореляції Кендалла дорівнює

$$\tau_{xy} = 2\frac{(ncord - ndicord)}{n(n-1)}$$

Для тих самих даних коефіцієнт кореляції Кендалла складає:

Приклади:

> cor(cora,corb,method ="kendall")

[1] 0.8

Ше раз нагадаємо, що всі тіж самі обчислення коефіцієнтів кореляції можна було реалізувати за допомогою функції

cor.test (…),

яка водночас перевіряє гіпотезу про значущість коефіцієнта кореляції.