

# Лекция 4

Ковариация и корреляция. Тестирование гипотез

---

**Курс:** Введение в DS на УБ и МиРА (весна, 2022)

**Преподаватель:** Владимир Омелюсик

18 апреля 2022 г.

- Функция плотности.
- Некоторые меры центральной тенденции и меры разброса.

## Обобщение: векторы

$$X = [ \underbrace{34, 35 \dots, 108}_{n \text{ чисел}} ] \rightarrow X \in \mathbb{R}^n$$

# Ковариация и выборочная ковариация

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Handwritten annotations: A bracket above the first term  $(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$  is labeled with  $+/-$ . A bracket above the second term  $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$  is labeled with  $+/-$ .

$$\hat{\text{Cov}}(X, Y) = \text{sCov} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Handwritten annotations: A bracket under  $\hat{\text{Cov}}(X, Y)$  points to the formula below. The entire sum term is highlighted in pink.

$$\frac{1}{N-1} (X_{\max} - \bar{X})(Y_{\max} - \bar{Y}) //$$

Handwritten annotations: A bracket under the sum term in the previous block points to this formula. The formula is enclosed in large square brackets.

# Корреляция и выборочная корреляция

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\hat{\text{Corr}}(x, y) = s\text{Corr} = \frac{s\text{Cov}}{\text{std}_x \text{std}_y}$$

$\parallel \quad \parallel$   
 $se_x \quad se_y$

## Примеры: доходы регионов и цвет

Cov, Corr — только для непрерыв. перемен. (веществ.)

↳ дискр., бинарных

Для категорию. перемен. — нельзя;

	цвет	цвет-код
1	"к"	1
2	"с"	2
3	"з"	3
4	"с"	2

"к" = 1

"с" = 2

"з" = 3

## Ковариация / корреляция и типы данных

$$\begin{array}{l} X = [1, 2, 3] \\ Y = [4, 5, 6] \end{array} \left. \vphantom{\begin{array}{l} X \\ Y \end{array}} \right\} \text{векст. перемен.}$$

$$\bar{X} = \frac{6}{3} = 2 \quad \bar{Y} = \frac{15}{3} = 5$$

$$\hat{cov} = \frac{1}{3-1} \sum_i (X_i - 2)(Y_i - 5)$$

## Пример: задача из экзамена

$$\begin{cases} r = [100, 98, 76, 102, 101] \\ c = ["r", "l", "l", "r", "g"] \\ d = [1002, 708, 132, 800, 1500] \end{cases}$$

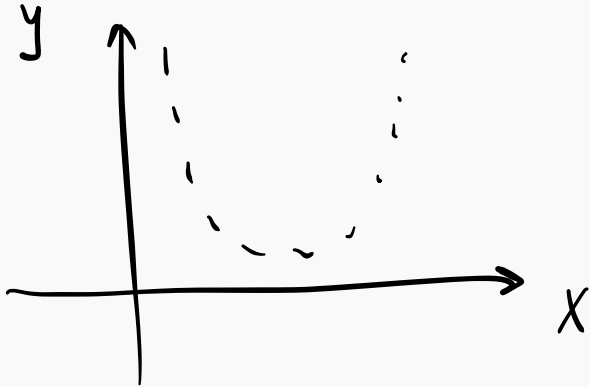
Handwritten annotations:

- A bracket labeled  $x$  connects the  $r$  array to the  $c$  array.
- A bracket labeled  $y$  connects the  $d$  array to the  $c$  array.
- The  $c$  array is annotated with indices:  $"_1"$  under "r",  $"_2"$  under "l",  $"_2"$  under "l",  $"_1"$  under "r", and  $"_3"$  under "g".
- A pink highlight is under the first four elements of the  $d$  array:  $[1002, 708, 132, 800]$ .
- The text "исчисл. по формуле" (calculated by formula) is written next to the  $d$  array.

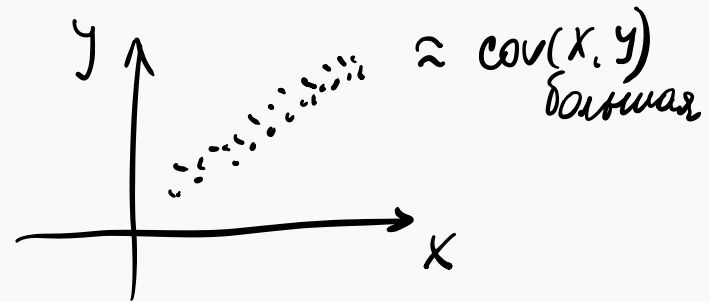


# Корреляция и причинность

$\text{cov}(X, Y) = 0 \Rightarrow X$  и  $Y$  нет лнн. связи



$$y = x^2$$



## Пример: рост собак

- Выборка из 1000 независимых измерений роста собак породы американская акита. Средний рост – 66 см.
- Выборка из 1000 независимых измерений роста собак породы акита-ину. Средний рост – 65 см.

$$\begin{cases} H_0 : \mu_{\text{ам.актор}} = \mu_{\text{актиса-ику}} \\ H_1 / H_A : \mu_{\text{ам.актор}} \neq \mu_{\text{актиса-ику}} \end{cases}$$

Цель - отвергнуть  $H_0$

$H_0 \rightarrow$  отверг.  
 $\searrow$  не отверг.  
~~прим.~~

# Ошибки I и II рода

## Ошибка I рода (False Positive)

Ситуация, когда отвергнута верная нулевая гипотеза.

## Ошибка II рода (False Negative)

Ситуация, когда не отвергнута неверная нулевая гипотеза.

# Пример: дождь и пожарная тревога

I : отвергн. верная н.

II : не отвергн. неверн. Но

1. Нулевая гипотеза: дождя не будет.

- Александр вышел на улицу без зонта, но пошёл дождь. II
- Александр вышел на улицу с зонтом, но дождя не было. I

⇒ промок

⇒ носил зонт

2. Нулевая гипотеза: пожара нет.

- Сработал датчик пожарной тревоги. По приезде оказалось, что это ошибка. I ⇒ ложный приезд
- Датчик пожарной тревоги не сработал. Оказалось, что был настоящий пожар. II

II ⇒ сгорело

# Статистический тест и распределения

II рода более страшны, чем ош. I рода

Стат. тест  
(проц-а)

↓  
(1)

$$H_0: \mu_1 = \mu_2$$

↓  
(2)

100 пр.

50 пр.  
(наши. и  
одни. вер. (II рода))

→

↓  
1  
с наши.  
вер. ош. I  
рода