Лекция 4

Ковариация и корреляция. Тестирование гипотез

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

18 апреля 2022 г.

В предыдущих сериях

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

Обобщение: векторы

$X = [34, 35 \ldots, 108] \longrightarrow X \in \mathbb{R}^n$

n чисел

Ковариация и выборочная ковариация

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

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

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

Корреляция и выборочная корреляция

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

$$\widehat{\mathrm{Corr}}(X, Y) = \mathrm{sCorr} = \frac{\mathrm{sCov}}{\mathrm{std}_X \mathrm{std}_Y}$$

$$\| \quad \| \\ se_x \quad se_y$$

Примеры: доходы регионов и цвет

Cov, Corr – только для непрерыв. перем. (веществ.)

→ дискр., бинарных

Для категор. перем – нельзя;

	цвет	цвет_код
1	"к"	1
2	"с"	2
3	"з"	3
4	"с"	2

"к" = 1
"с" = 2
"з" = 3

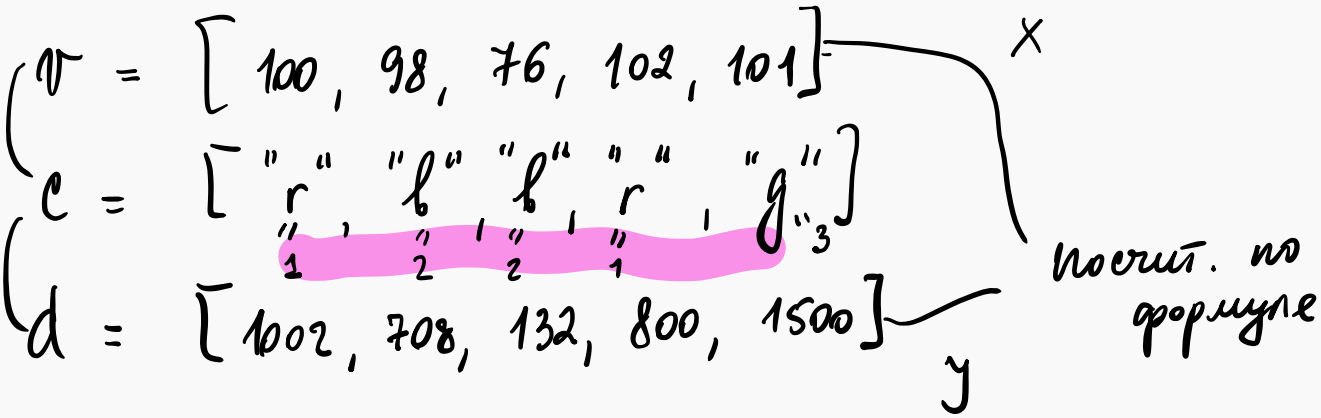
Ковариация / корреляция и типы данных

$X = [1, 2, 3]$
$Y = [4, 5, 6]$
} вещ. перем.

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

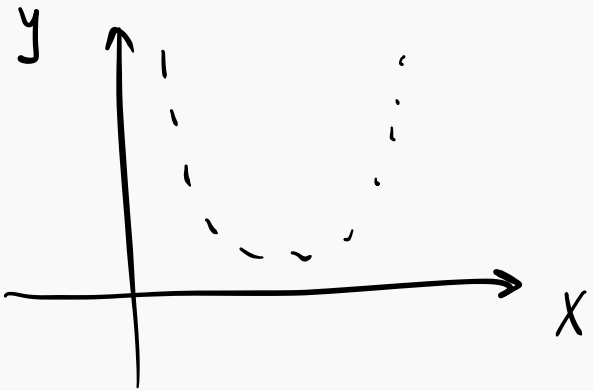
$$\hat{cov} = \frac{1}{3-1} \sum_i (X_i - 2)(y_i - 5)$$

Пример: задача из экзамена

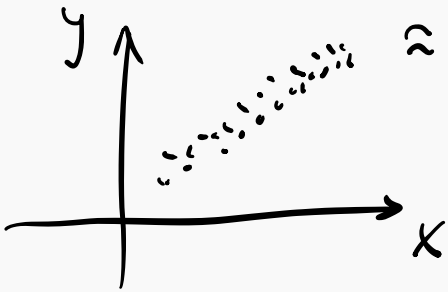


Корреляция и причинность

$cov(X, Y) = 0 \Rightarrow$ X и Y нет лин. связи



$Y = X^2$



$\approx cov(X, Y)$ большая

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.

Гипотезы

$$\begin{cases} H_0: \mu_{\text{ам. акита}} = \mu_{\text{акита-ину}} \\ H_1 / H_A: \mu_{\text{ам. акита}} \neq \mu_{\text{акита-ину}} \end{cases}$$

Цель – отвергнуть H_0

H_0 → отверг.

→ не отверг.

~~принять~~.

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: отверг. верная Н$_0$.

II: не отверг. неверн. Н$_0$

1. Нулевая гипотеза: дождя не будет.
   - Александр вышел на улицу без зонта, но пошёл дождь. ⇒ промок — II
   - Александр вышел на улицу с зонтом, но дождя не было. → носил зонт — I
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 → сгорело

Статистический тест и распределения