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

Ковариация и выборочная ковариация

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

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

Корреляция и выборочная корреляция

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

$$\mathrm{scorr} = \frac{\mathrm{sCov}}{\mathrm{std}_X \mathrm{std}_Y}$$

Примеры: доходы регионов и цвет

Ковариация / корреляция и типы данных

Пример: задача из экзамена

Корреляция и причинность

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

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

1. Нулевая гипотеза: дождя не будет.
    - Александр вышел на улицу без зонта, но пошёл дождь.
    - Александр вышел на улицу с зонтом, но дождя не было.
2. Нулевая гипотеза: пожара нет.
    - Сработал датчик пожарной тревоги. По приезде оказалось, что это ошибка.
    - Датчик пожарной тревоги не сработал. Оказалось, что был настоящий пожар.

Статистический тест и распределения

Статистическая значимость

$$\mathbb{P}(\text{ош. I рода}) \leqslant \alpha$$

Z-тест для одной выборки

$X_1, \ldots, X_N$ – выборка из N независимых, одинаково распределённых нормальных случайных величин с известной дисперсией.

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Тогда Z-статистка для гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma} \sim \mathcal{N}(0, 1)$$

t-тест для одной выборки

$X_1, \ldots, X_N$ – выборка из N независимых, одинаково распределённых нормальных случайных величин с неизвестной дисперсией.

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Тогда t-статистка для гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}} \sim t_{N-1}$$

NB!: $t_{N-1} \to Z$ при $N \to \infty$.

Пример: проверка гипотезы о среднем

$$X_1, \ldots, X_{300} \sim \mathcal{N}(\mu, \sigma)$$

$$\sum_i X_i = 60, \ \sum_i X_i^2 = 4000$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 0.3, \\ H_1 : \mu \neq 0.3 \end{cases}$$

на уровне значимости 5%.

Пример: проверка гипотезы о среднем

Проверка гипотезы о доле

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \to \mathcal{N}(0, 1)$$

p-value
Минимальный уровень значимости, при котором нулевая гипотеза не отвергается.

Посчитали Z_{obs} и $H_1 : \mu \neq \mu_0$. Тогда

$$p - value = 2\,\mathbb{P}(Z \leqslant Z_{obs})$$

Основной результат: p-value $< \alpha \Rightarrow$ нулевая гипотеза отвергается.

Проверка гипотез при помощи p-value

Доверительный интервал
Интервал со случайным границами, такой что

$$\mathbb{P}(T_l(X) < \theta < T_r(X)) \geqslant 1 - \alpha$$

Проверка гипотез при помощи доверительного интервала

Основной результат
Если наблюдаемое значение статистики попадает в $(1-\alpha)$-процентный доверительный интервал, то нулевая гипотеза не отвергается на уровне значимости α.

Тесты для двух независимых выборок

$X_1, \dots, X_{N1}$ и $Y_1, \dots, Y_{N2}$ – две независимые выборки. Предположим, что дисперсии этих выборок равны. Тогда для проверки гипотезы о равенстве средних можно использовать статистику

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}\sqrt{\frac{1}{N1} + \frac{1}{N2}}} \sim t_{N1+N2-2},$$

$$\hat{\sigma} = \sqrt{\frac{(N1-1)\hat{\sigma}_X^2 + (N2-1)\hat{\sigma}_Y^2}{N1+N2-2}}$$

Пример: доходы регионов

Тесты для парных выборок

Пусть есть N наблюдений над одним объектом ДО и ПОСЛЕ проведения эксперимента. Тогда гипотезу о разности средних ДО и ПОСЛЕ можно проверить при помощи статистики

$$t = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{N}}},$$

$$d = \sum_i d_i^{\text{ПОСЛЕ}} - d_i^{\text{ДО}}$$

Пример: доходы семьи