Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Расчётно-графическая работа №3

По дисциплине «Математическая статистика»

Михайлов Дмитрий Андреевич
P3206
368530
Медведев Владислав Александрович
P3206
368508

Санкт-Петербург
2025 год

Содержание

Задача №1

Условие задачи.

Для каждой проблемы нужно провести два статистических теста, если не сказано иное, причём первый из критериев нужно реализовать самостоятельно (считать и выводить значение статистики, критическое значение, p-value), в качестве второго можно воспользоваться готовой реализацией. Также нужно отдельно указывать, как формализуются H_0 и H_1 для выбранных тестов. Уровень значимости выбираете сами.

Вариант 1

В файле kc_house_data.csv приведены данные о цене на недвижимость где-то в окрестности Сиэтла.

1. Предположите с каким вероятностным законом распределена цена. С помощью статистического теста подтвердите/опровергните это предположение (первый тест - критерий согласия Колмогорова, если распределение абсолютно непрерывное, либо критерий согласия Пирсона хи-квадрат, если распределение дискретное).

2. Верно ли, что цена на старый и новый фонд распределена одинаково (порог возраста выбирайте сами) (первый тест - критерий однородности Смирнова или хи-квадрат, или f-тест + t-тест)?

3. Верно ли, что при увеличении “жилищной площади”растёт и цена (первый тест - критерий на один из коэффициентов корреляции)?

Решение.

Критерий согласия Колмогорова:

$$\begin{cases} H_0 = F(x) = F_0(x), \text{(данные имеют заданное теоретическое распределение)} \\ H_1 = F(x) \neq F_0(x), \text{(данные НЕ имеют заданное распределение)} \end{cases}$$

где:

- $F(x)$ — эмпирическая функция распределения
- $F_0(x)$ — теоретическая функция распределения

Функция распределения Колмогорова:

$$P(\sqrt{n}D_n \leq \lambda) \to 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2}$$

Тогда:

$$P(\sqrt{n}D_n > \lambda) \approx 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2}$$

Это и есть p-value - вероятность того, что наблюдаемая статистика D_n больше полученной.

Критерий согласия Андерсона-Дарлинга:
Для отсортированной выборки $x_1 \le x_2 \le \cdots \le x_n$, и теоретической CDF $F(x)$:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i-1) \cdot (\ln F(x_i) + \ln(1 - F(x_{n+1-i})))]$$

где:

- $F(x)$ — функция распределения предполагаемого закона (например, нормального)
- n — размер выборки

Проверим гипотезы с помощью статистических тестов:

Критерий Смирнова — сравнение распределений.
Критерий хи-квадрат — сравнения фактических данных в выборке с теоретическими результатами.
Возьмем за порог 1980 год.

KS-тест (критерий Колмогорова-Смирнова)
Цель: сравнить два эмпирических распределения $F_n(x)$ и $G_m(x)$

$$H_0 : F(x) = G(x), (\text{распределения одинаковы})$$

Статистика Колмогорова:

$$D = \sup_x |F_n(x) - G_m(x)|$$

где:

- $F_n(x)$ — эмпирическая функция распределения первой выборки
- $G_m(x)$ — эмпирическая функция второй выборки
- sup — супремум (максимальное расстояние между функциями)

Критерий хи-квадрат (на однородность)
Критерий однородности χ^2 используется для проверки, одинаково ли распределяется признак (в нашем случае — цена по квартилям) в двух или более группах (старый и новый фонд жилья).

Он сравнивает наблюдаемые частоты с ожидаемыми, которые рассчитываются при условии, что распределения в группах одинаковые.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

где:

- O_{ij} — наблюдаемое количество объектов в ячейке (i, j),

- E_{ij} — ожидаемое количество объектов в ячейке (i, j), вычисляемое по формуле:

$$E_{ij} = \frac{(\text{сумма по строке i}) \cdot (\text{сумма по столбцу j})}{\text{общая сумма}}$$

- $k = 2$ — количество групп (старый и новый фонд)
- $m = 4$ — количество интервалов (квартильные группы)

Порядок действий:

1. Разделить выборку по году постройки на старый и новый фонд.
2. Определить квартильные границы по всей совокупности цен.
3. Отнести каждое наблюдение к одному из квартилей.
4. Построить таблицу частот: по фондам и квартилям.
5. Вычислить статистику χ^2.
6. Сравнить результат с критическим значением χ^2 для уровня значимости (например, $\alpha = 0.05$) и нужного числа степеней свободы:

$$df = (k-1)(m-1) = (2-1)(4-1) = 3$$

Мы используем именно критерий однородности, а не независимости или согласия, потому что сравниваем распределения в разных группах, а не проверяем связь между двумя признаками.

Ожидаемые частоты считаются так:

$$E_{ij} = \frac{\text{сумма по строке i} \times \text{сумма по столбцу j}}{\text{общая сумма}}$$

Статистика χ^2
После этого мы считаем:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

В результате для наших данных получается:

$$\chi^2 \approx 444.91$$

Степени свободы считаются по формуле:

$$df = (r-1) \times (c-1)$$

где:

- r — количество строк в таблице (у нас 2 группы: старый и новый фонд)
- c — количество столбцов (у нас 4 квартиля)

Подставляем:

$$df = (2 - 1) \times (4 - 1) = 1 \times 3 = 3$$

Итого $df = 3$.

Когда известно χ^2 и степени свободы df, p-значение — это вероятность получить значение статистики ещё большее, чем наблюдаемое, при справедливости нулевой гипотезы.

p-значение считается через распределение хи-квадрат:

$$p = P(\chi^2 > 444.91)$$

$$p = 4.13 \times 10^{-96}$$

То есть вероятность случайно получить такие большие различия практически нулевая.

Вывод.
p-значение $(\chi^2) < 0.05 \rightarrow$ распределение отличается
p-значение KS $< 0.05 \rightarrow$ Форма распределения цен также различна.

Критерий на коэффициент корреляции Пирсона.
Гипотезы.

$$\begin{cases} H_0 : \rho = 0 & \text{(нет корреляции)} \\ H_1 : \rho \neq 0 & \text{(есть корреляция)} \end{cases}$$

Здесь ρ — истинный коэффициент корреляции Пирсона в генеральной совокупности.

Коэффициент корреляции Пирсона.
Оценка r вычисляется по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

где $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$ — выборочные средние.

Статистика критерия.
Если нулевая гипотеза верна и $\rho = 0$, то статистика:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

распределена по t-распределению Стьюдента с $n-2$ степенями свободы.

Правило принятия решения.

1. Найдём критическое значение t_{crit} для уровня значимости α и $n-2$ степеней свободы:

$$t_{\text{crit}} = t_{1-\alpha/2}(n-2)$$

2. Вычислим двустороннее значение p-value:

$$p = 2 \cdot P(T > |t_{\text{набл}}|)$$

3. Если $p < \alpha$, то отвергаем H_0 в пользу H_1.

Критерий на коэффициент корреляции Спирмена.

Оценка коэффициента Спирмена.
Коэффициент Спирмена r_s вычисляется по формуле:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

где $d_i = \text{rg}(x_i) - \text{rg}(y_i)$ — разность рангов соответствующих элементов выборки. (Ранг — это порядковый номер элемента в отсортированной выборке)

Или, как в ‘scipy‘, с помощью корреляции Пирсона между рангами:

$$r_s = \text{corr}(\text{rg}(x), \text{rg}(y))$$

Приближённая t-статистика.
Для больших n приближённо применяется t-распределение с $n-2$ степенями свободы:

$$t = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2}$$

Критическое значение и p-value.

- Критическое значение: $t_{\text{crit}} = t_{1-\alpha/2}(n-2)$
- Двусторонний p-value: $p = 2 \cdot P(T > |t| \mid H_0)$

Правило принятия решения.

- Если $|t| > t\_\text{crit}$, то отвергаем H_0 — есть статистически значимая корреляция.
- Иначе — не отвергаем H_0 — доказательств корреляции недостаточно.

Далее приведён код на языке Python.

Приложения

Задача №1

Ссылка на исходник с кодом программы, решающей эту задачу на языке Python. [1]

Список использованных источников

[1] Задача №1. *URL*: Исходник с кодом, решающий задачу №1.