Домашнее задание № 5

Корреляционный анализ. Статистические тесты для сравнения групп

Крайний срок сдачи: 13 декабря 2020 г. (до конца дня).

Максимальные баллы:
N1 -2.5 балла,
N2 - 3 балла,
T1-T3 - 1.5 балла,
T4 - 2 балла*

1

N1 Пусть X - случайная величина с равномерным распределение на $[0, 1]$. Рассмотрим 3 модели

(a) монотонная зависимость: $Y = \sin(\pi X/2)$;

(b) монотонная зависимость с маленьким шумом: $Y = \sin(\pi X/2) + \varepsilon$, где ε имеет нормальное распределение с нулевым средним и дисперсией $\sigma^2 = 0.04$;

(c) немонотонная зависимость с маленьким шумом: $Y = \sin(\pi X) + \varepsilon$.

Целью задач, перечисленных ниже, является изучение возможностей коэффициентов корреляции "поймать" разные виды зависимости.

(i) Просимулируйте выборку $(X_1, Y_1), ...(X_n, Y_n)$ для каждой модели M раз (например, $n = 1000$ и $M = 20$). По каждой выборке вычислите коэффициенты корреляции Пирсона, Кендалла и Спирмена и соотвествующие уровни значимости. Если для каких-то моделей и методов получаются существенно различные уровни значимости для разных выборок, отобразите эти уровни на диаграмме размаха.

(ii) В модели (b), увеличите дисперсию шума - рассмотрите $\sigma^2 = 0.5, 1, 1.5, 2, ...$ Для каждого σ^2, вычислите коэффициенты корреляции Кендалла. Найдите наибольшее значение σ^2, при котором коэффициенты являются значимыми.

(iii) Для σ^2, полученного на предыдущем шаге, сравните коэффициенты корреляции Кендалла и Спирмена. Проверьте гипотезу, что сдвиг (медиана разностей) равна нулю. Постройте соответствующий доверительный интервал.

N2 Рассмотрим базу данных "swiss" , включающую в себя показатели рождаемости и различные социально-экономические индикаторы для 47 франкоговорящих провинций Швейцарии в 1888 году, см. `https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/swiss.html`.

Целью данной задачи является анализ зависимости между рождаемостью и социально-экономическими индикаторами. Особенность данных состоит в том, что большинство переменных представляют собой процентные соотношения.

(i) Вычислите непараметрические коэффициенты Спирмена и Кендалла между показателем рождаемости и каждой переменной. Для коэффициента корреляции Кендалла вычислите уровни значимости на основе точных распределений соответствующих статистики, а для коэффициента корреляции Спирмена - на основе метода AS-89. Сделайте выводы.

(ii) Рассмотрим более подробно рождаемость в первых 10 провинциях. Для каждой из этих провинции, найдите провинцию в остальной базе данных (наблюдения с 11 по 47) с наиболее близкими социально-экономическими показателями. В качестве меры близости используйте евклидово расстояние между стандартизованными показателями. Протестируйте гипотезу, что показатели рождаемости одинаковы между провинциями 1-10 и похожими на них (по социально-экономическому развитию) провинциями 11-47.

(iii) Разделите все провинции на 3 группы, которые мы будем обозначать C,P,M: C ("catolic") - более 80 % населения католики ; P("protestant") - более 80 % протестанты; M ("mixed")-

"смешанные" провинции (не менее 20 % католики и не менее 20 % протестанты). Протестируйте гипотезу, что во всех трёх провинциях уровень рождаемости имеет одно и тоже распределение. Рассмотрите также гипотезу попарно (то есть, для групп C и P, C и M, P и M), используя наиболее подходящую альтернативу для каждой пары.

(iv) Для каждой группы, полученной на предыдущем шаге, разделите провинции на 4 группы

1. более 50% мужчин работают в сельском хозяйстве и низкий уровень детской смертности (менее 1-ого квартиля детской смертности по всем провинциям);
2. менее 50% мужчин работают в сельском хозяйстве и низкий уровень детской смертности;
3. более 50% мужчин работают в сельском хозяйстве и высокий уровень детской смертности (более 1-ого квартиля детской смертности по всем провинциям);
4. менее 50% мужчин работают в сельском хозяйстве и высокий уровень детской смертности.

Вычислите средние значения показателя рождаемости в каждой подгруппе. Если в какой-то из подгрупп не будет наблюдений, замените медианным значением этой подгруппы по всем наблюдениям (без деления на C, P, M). Протестируйте гипотезу, что средний показатель рождаемости в каждой подгруппе одинаков для групп C, P, M.

2

T1 Совместное распределение случайных величин X и Y определено равенством

$$\mathbb{P}\{X = x,\ Y = y\} = \frac{x+y}{21}, \qquad x = 1, 2, 3, \ \ y = 1, 2.$$

Вычислите коэффициент корреляции τ Кендалла между этими случайными величинами.

T2 Задано N объектов, разделённых на k групп, причём группа номер $j = 1..k$ состоит из n_j элементов ($n_1 + ... + n_k = N$). Для каждого объекта известно значение x_{ij} некоторой характеристики этого объекта. Докажите, что

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\big(x_{ij} - x_{..}\big)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\big(x_{ij} - x_{\cdot j}\big)^2 + \sum_{j=1}^{k} n_j\big(x_{\cdot j} - x_{..}\big)^2,$$

где $x_{..}$ - среднее значение по всем $x_{ij}, i = 1..n_j, j = 1..k$, и $x_{\cdot j}$ - среднее значение по j-ой группе $x_{ij}, i = 1..n_j$.

Комментарий. Данное равенство можно прочитать как "мера изменчивости всех объектов есть сумма меры изменчивости внутри групп и меры изменчивости между группами".

Указание. Данная задача имеет элегантное решение, основанное на теореме Гюйгенса-Штейнера: если в пространстве $\mathbb{R}^p$ заданы точки $\vec{z}_i$ с массами $m_i, i = 1..Q$, то момент инерции

$$I_{\vec{a}} := \sum_{i=1}^{n} m_i \|\vec{z}_i - \vec{a}\|^2$$

относительно любой точки $\vec{a}$ может быть подсчитан как

$$I_{\vec{a}} = I_{\vec{c}} + M\|\vec{c} - \vec{a}\|^2,$$

где $M = \sum_{i=1}^{Q} m_i$ - суммарная масса системы и $\vec{c} := (\sum_{i=1}^{Q} m_i\vec{z}_i)/M$ - центр масс системы.

T3 Рассмотрим 2 независимые выборки размеров m и n. Допустим, что в данных нет повторяющихся наблюдений, и мы приписали по объединённым выборкам ранги от 1 до $(m+n)$. Обозначим соответсвующие ранги $R_1, ..., R_m$ и $S_1, ..., S_n$. Обозначим через $W = \sum_{j=1}^{n} S_j$ статистику Уилкоксона.

(i) Найдите наибольшее и наименьшее значения статистики W.

(ii) Предположим, что распределения выборок совпадает. Докажите, что распределение W является в данном случае симметричным относительно своей медианы, то есть

$$\mathbb{P}\Big\{W = \min_W + x\Big\} = \mathbb{P}\Big\{W = \max_W - x\Big\}, \qquad \forall x > 0,$$

где $\max_W$ и $\min_W$ - наибольшее и наименьшее значения W, вычисленные в п. (i).

T4* Докажите, что если коэффициент корреляции Спирмена $\hat{\rho}_S$ вычислен по независимым выборкам размера n, полученных из непрерывных распределений, то

$$\operatorname{Var} \hat{\rho}_S = 1/(n-1).$$

Указание. Покажите, что

$$\hat{\rho}_S = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} (iS_i) - 3\frac{n+1}{n-1},$$

где в случае независимости исходных выборок $(S_1, ..., S_n)$ является случайным вектором, состоящим из чисел $1, .., n$ и имеющим равномерное распределение на множестве $n!$ перестановок этих чисел.