Домашнее задание № 4

Статистические тесты для сравнения групп

Крайний срок сдачи: 3 декабря 2021 г., 18:00.

1. (1 балл) Даны 2 выборки одинакового размера n:

$$X_1, ..., X_n \sim \mathcal{N}(\mu_X, 400), \qquad Y_1, ..., Y_n \sim \mathcal{N}(\mu_Y, 225).$$

   Исследователю нужно проверить, можно ли считать, что средние значения в выборках совпадают, то есть $\theta := \mu_X - \mu_Y = 0$. Для проверки этой гипотезы было решено построить тест на основе критического множества $\{\bar{X} - \bar{Y} > C\}$, где $\bar{X}, \bar{Y}$ - средние арифметические элементов выборок. Найдите C и n такие, что ошибка первого рода этого теста равна 0.05, а ошибка второго рода при тестировании гипотезы против альтернативы $\theta = 10$ равна 0.1.

2. (3 балла) 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) 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.

(ii) 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.

(iii) 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.

3. (2 балла) 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 характери-

стики этого объекта. Докажите, что

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

где $x_{\cdot\cdot}$ - 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". Равенство играет ключевую роль для понимания сути дисперсионного анализа и критерия Краскела-Уоллиса.

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

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

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.

4. (2 балла) Пусть $(S_1, ..., S_n)$ - вектор рангов, имеющий равномерное распределение на множестве $n!$ перестановок чисел $1, .., n$. Покажите, что

   (i) $\mathbb{E}S_i = (n+1)/2, \qquad \forall\ i = 1..n,$

   (ii) $\operatorname{Var} S_i = (n^2-1)/12, \qquad \forall\ i = 1..n,$

   (iii) $\operatorname{cov}(S_i, S_j) = -(n+1)/12, \qquad \forall\ i, j = 1..n, \ \ i \neq j.$

5. (2 балла) 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).

6* (2 балла) Рассмотрим модель парных повторных наблюдений - набор независимых пар $(X_i, Y_i), i = 1..n$, таких, что X_i и Y_i зависимы. Обозначим разности $Z_i = Y_i - X_i$ и разложим их в сумму $Z_i = \theta + \varepsilon_i$, где ε_i - случайная ошибка такая, что

$$\mathbb{P}\{\varepsilon_i \leq x\} + \mathbb{P}\{\varepsilon_i \leq -x\} = 1, \qquad \forall x \in \mathbb{R}.$$

Обозначим R_i - ранг Z_i при расположении величин $Z_1, Z_2, ...$ в порядке возрастания их абсолютных значений.

Введём понятие антиранга: антиранг числа $k = 1..n$ равен A_k, если ранг числа Z_{A_k} равен k. Например, если

$$Z_1 = 2, \quad Z_2 = -2.6, \quad Z_3 = 1.8,$$

то

$$R_1 = 2, \quad R_2 = 3, \quad R_3 = 1,$$

и

$$A_1 = 3, \quad A_2 = 1, \quad R_3 = 2.$$

Другими словами,

$$A_k = s \qquad \Longleftrightarrow \qquad R_s = k.$$

Докажите, что если $\theta = 0$, то случайные величины $W_i = \mathbb{I}\{Z_{A_i} > 0\}$ образуют схему Бернулли с $p = 1/2$.

Комментарий. Нужно строго доказать независимость величин $W_1, W_2, .., W_n$.