Домашнее задание № 3

Тема: Общая теория построения статистических тестов. Следствия из теоремы Уилкса. Корреляционный анализ.

Крайний срок сдачи: 19 ноября 2021 г., 18:00.

1. (2 балла) Дана выборка $X_1, ...X_n$ из распределения с плотностью

$$p_\theta(x) = \frac{\theta}{x^2}\,\mathbb{I}\{x > \theta\}.$$

Требуется протестировать гипотезу $\theta = \theta_0$ против альтернативы $\theta = \theta_1 \neq \theta_0$. Постройте равномерно наиболее мощный рандомизированный тест для случаев $\theta_1 > \theta_0$ и $\theta_1 < \theta_0$.

2. (1 балл) В некоторый день было зафиксировано 96 звонков в колл-центр, причём время между последовательными звонками (в минутах) равнялось

$$1, 1, 7, 16, 8, 8, 11, 7, 5, 45, 13, 0, 36, 15, 4, 15, 7, 39, 6, 91, 28, 7, 0, 2,$$
$$9, 2, 6, 1, 4, 83, 2, 3, 5, 34, 1, 1, 2, 0, 11, 79, 2, 2, 4, 1, 3, 0, 2, 2, 17, 55, 8,$$
$$9, 20, 23, 16, 3, 5, 5, 4, 84, 1, 20, 1, 1, 20, 0, 19, 17, 5, 66, 0, 2, 5, 1, 26,$$
$$14, 1, 0, 9, 88, 4, 11, 4, 2, 1, 32, 21, 2, 15, 76, 44, 8, 16, 12, 1, 9$$

Возникает вопрос, можно ли моделировать количество поступивших звонков процессом Пуассона (по определению процесса Пуассона, время между последовательными событиями является набором независимых случайных величин с экспонециальным распределением). Для ответа на данный вопрос требуется проверить выборку на соответствие экспоненциальному распределению.

Для этого предлагается разделить всё множество положительных чисел на 5 интервалов (один из которых бесконечный) таким образом, что в соответствии с экспоненциальным распределением теоретическое количество элементов из выборки размера 96, попадающих в каждый из интервалов, превосходит число 5[1][2]. После этого нужно воспользоваться критерием хи-квадрат (критерий согласия). Во всех вычислениях предлагается заменить неизвестный параметр экспоненциального распределения на его оценку максимального правдоподобия.

3. (2 балла) Имеется набор из четырёх монет, вероятность выпадения орла i-ой монеты равна $p_i \in (0,1), i = 1..4$. При помощи статистических методов требуется проверить, являются ли данные монеты "настоящими" ($p_i = 1/2, i = 1..4$) или "фальшивыми" ($p_i \neq 1/2, i = 1..4$) . Для этого каждую монету подбрасывают 50 раз и записывают количество выпадений орла в каждой из 50 серий (если $X_i^1, X_i^2, X_i^3, X_i^4$ - результат выпадения орла для 1,2,3,4 монеты в серии номер $i = 1..50$, то записывается $S_i = X_i^1 + X_i^2 + X_i^3 + X_i^4$). Допустим, что выпало 0, 1, 2, 3, 4 орла 4, 12, 14, 11, 9 раз соответственно.

   (a) Примените хи-квадрат тест для проверки гипотезы $p_1 = ... = p_4 = 1/2$, (другими словами, гипотеза состоит в том, что все монеты являются "настоящими").

   (b) Предположим дополнительно, что вероятность выпадения орла у каждой монеты одинаковая, $p_1 = ... = p_4 = p$. С точностью до 0.01, вычислите оценку $\hat{p}$ параметра p , минимизирующую статистику критерия хи-квадрат. Проверьте гипотезу $p_1 = .. = p_4 = \hat{p}$ (другими словами, проверьте гипотезу, что все монеты одновременно являются "фальшивыми" с одинаковой вероятностью выпадения орла).

4. (2 балла) Датчик случайных цифр сгенерировал последовательность из 20 элементов

$$0, 1, 1, 4, 5, 8, 4, 9, 5, 1, 5, 5, 9, 6, 7, 2, 6, 2, 5, 4$$

Для проверки качества этого датчика предлагается 2 идеи:

[1]Как было объяснено на лекции, это ограничение, по всей видимости, было впервые описано в книге Г. И. Ивченко и Ю. И. Медведева "Введение в математическую статистику" и затем было использовано в большом количестве других пособий.

[2]Для этого деления рекомендуется использовать функцию qexp.

(i) Разделить последовательность на 4 подпоследовательности (цифры, стоящие на местах 1-5, 6-10, 11-15, 16-20) и проверить независимость фактора принадлежности цифры подпоследовательности и фактора принадлежности цифры множествам $\{0, 1, 2, 3, 4\}$ и $\{5, 6, 7, 8, 9\}$ (критерий хи-квадрат для таблиц сопряжённости).

(ii) Сравнить количество цифр из групп $\{0, 1, 2\}, \{3, 4, 5, 6\}, \{7, 8, 9\}$ с ожидаемыми количествами этих цифр в предположении равномерности распределения (критерий хи-квадрат, основанный на теореме Пирсона).

Имплементируйте эти методы и сделайте выводы.

5. (2 балла) Вычислите (без использования компьютера) точное распределение коэффициента корреляции Спирмена между двумя независимыми выборками размера $n = 4$, при условии, что в данных нет повторяющихся наблюдений.

6. (1 балл) Пусть (X_1, Y_1) и (X_2, Y_2) - две независимые пары случайных величин с плотностью

$$p_{(X,Y)}(x, y) = \begin{cases} \frac{1}{2} y^2 e^{-x-y}, & \text{если} \quad x > 0, \quad y > 0, \\ 0, & \text{иначе.} \end{cases}$$

Вычислите (теоретический) коэффициент корреляции Кендалла τ между (X_1, Y_1) и (X_2, Y_2) и (теоретический) коэффициент корреляции Пирсона между X_1 и Y_1.

7* (2 балла) Частой проблемой применения критериев согласия является то, что параметры распределения не известны. Методы исключения параметров всегда носят эвристический характер.

Пусть $X_1, ..., X_n$ - набор i.i.d. случайных величин с нормальным распределением с неизвестным средним значением μ и известной дисперсией σ^2. Предлагается 2 метода исключения неизвестного параметра μ.

(i) "Кустарный метод". Оценим параметр μ средним значением $\bar{X} = (X_1 + ... + X_n)/n$ и перейдём от X_i к $\widetilde{X}_i = X_i - \bar{X}, i = 1..n$. Докажите, что величины $\widetilde{X}_1, ..., \widetilde{X}_n$ являются зависимыми в вероятностно -статистическом смысле, но вектор $(\widetilde{X}_1, ..., \widetilde{X}_n)$ и $\bar{X}$ независимы.

(ii) "Профессиональный метод". Положим

$$A_m := \frac{1}{n+\sqrt{n}} \sum_{i=1}^{n} X_i + \frac{1}{1+\sqrt{n}} X_m,$$

где m - фиксированное число от 1 до n. Докажите, что случайные величины

$$X_1 - A_m, \quad ...., \quad X_{m-1} - A_m, \quad X_{m+1} - A_m \quad , ..., \quad X_n - A_m$$

независимы в совокупности и имеют нормальное распределение со средним 0 и дисперсией σ^2.