Домашнее задание № 4

Тема: Goodness-of-fit. Таблицы сопряженности.

Крайний срок сдачи: 29 ноября 2020 г. (до конца дня).

1. Датчик случайных цифр сгенерировал последовательность

$$0, 1, 7, 9, 8, 4, 3, 6, 6, 6, 8, 9, 0, 1, 7, 9, 4.$$

   Для проверки качества этого датчика предлагается 2 идеи:

   (i) сравнить количество цифр 0, 1, 2, .., 9 с ожидаемыми количествами этих цифр в предположении равномерности распределения;

   (ii) сравнить выпадение такой же цифры / соседней цифры / другой цифры с ожидаемыми количествами таких выпадений (0 и 9 при этом удобно считать соседними).

   Имплементируйте эти методы и сделайте выводы.

2. Микробиолог проводит эксперимент. На каждую из 150 чашек Петри он помещает некоторое (одинаковое) количество бактерий и подсчитывает число колоний. Оказалось, что на 92 чашках колоний не образовалось, на 46 чашках была заметна только 1 колония, на 8 чашках - 2 колонии, на 3 чашках - 3 колонии, на 1 чашке - 4 колонии. При помощи критерия хи-квадрат протестируйте гипотезу, состоящую в том, что случайная величина "количество колоний на чашке Петри" имеет распределение Пуассона. При этом параметр распределения Пуассона оцените

   (i) методом максимального правдоподобия;

   (ii) минимизацией статистики критерия хи-квадрат.

3. Рассмотрим переменную eruptions из базы данных faithful (временные интервалы между последовательными извержениями Old Faithful Geyser, см. `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/faithful.html`). Предположим, что эта выборка имеет распределение смеси двух нормальных распределений,

$$p(x) = \lambda p_{(\mu_1,\sigma_1)}(x) + (1-\lambda)p_{(\mu_2,\sigma_2)}(x),$$

где $p_{(\mu_i,\sigma_i)}(x)$ - это плотность нормального распределения со средним μ_i и стандартным отклонением $\sigma_i, i = 1, 2$. Для оценки параметров $\lambda \in (0,1), \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0$ используются два метода:

1. "профессиональный" - EM - алгоритм;
2. "кустарный" :
   i. строится ядерная оценка плотности;
   ii. средние значения μ_1, μ_2 оцениваются x-координатами "горбов";
   iii. находится точка минимума оценки плотности, лежащая между "горбами" (далее будем называть эту точку разделительной);
   iv. стандартные отклонения σ_1, σ_2 оцениваются стандартными отклонениями подвыборок со значениями слева и справа от разделительной точки;
   v. наконец, λ оценивается как доля наблюдений слева от разделительной точки.

Найдите оценки параметров первым и вторым методом. Определите какой из этих методов лучше описывает переменную eruptions, применив критерий хи-квадрат при делении области значений переменной на 20 интервалов.

4. В страховую компанию поступает 2 типа заявок на возмещение ущерба по полисам автомобильного страхования: заявки, поданные после ДТП и по причине угона автомобиля. Для моделирования размеров выплат по этим заявкам используются экспонененциальные распределения со средними значениями 30 тыс. рублей и 500 тыс. рублей соответственно.

Сотрудник компании заметил, что количество заявок из второй группы уменьшается, и предложил для простоты вычислений считать, что все заявки с выплатой до 200 тыс. взяты из экспоненциального распределения со средним 30 тыс. рублей.

Выясните, при какой максимальной доле заявок из второй группы результаты применения критерия хи - квадрат не позволяют сделать вывод об отличии распределения всех заявок с суммой до 200 тыс от экспоненциального распределения со средним 30 тыс. рублей. Для проведения теста промоделируйте выплаты из смеси распределений 1000 раз для каждого возможного значения параметра смеси. Тест проводите при делении интервала [0,200 тыс] на 20 частей. Ответ на вопрос нужно дать с точностью до 0.01.

5. В течение курса студентам было задано 12 домашних заданий. Студенты A и B получили оценки

$$\begin{aligned} A &= (0, 2, 8, 6, 9, 10, 7, 8, 3, 10, 7, 5); \\ B &= (0, 7, 5, 6, 9, 2, 8, 7, 3, 10, 9, 6). \end{aligned}$$

При помощи критерия хи-квадрат

(i) проверьте гипотезу о равенстве распределений в этих группах;

(ii) проверьте гипотезу о том, что набор чисел, представляющий собой разности оценок A и B, образуют выборку.

Дайте интерпретацию полученных результатов. При проведении теста в (i) используйте деление оценок на 4 группы (0-3; 4-5; 6-7; 8-10).

6* 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.