Практикум 8. Критерий хи-квадрат.

Сегодня мы с вами обучаемся проверять различные виды гипотез с помощью критерия хи-квадрат.

Основным файлом, с которым мы будем работать, будет Priem.csv, содержащим информацию о годах рождения абитуриентов механико-математического факультета.

1. **Вариант 1.**

   1.1 Начнем с проверки простой гипотезы.

   Найдите первые 1000 цифр числа π после запятой. С помощью критерия хи-квадрат проверьте, можно ли при уровне значимости 0.05 считать эти цифры случайными?

   1.2 Переходим к параметрической гипотезе. Т

   Среди 2020 семей, имеющих 2 детей, 527 семей, в которых 2 мальчика, и 476 - две девочки. Можно ли при уровне значимости 0.05 считать, что количество мальчиков – биномиальная случайная величина? (Для этой задачи $\hat{\theta} = (N_1 + 2N_2)/n$, где N_i – число семей, в которых ровно i мальчиков, n – общее число семей.)

   1.2* Исследовать работу критерия для проверки сложной гипотезы на модельных данных. Рассмотреть следующие распределения:

   a) $Binom(2, 1/2)$,
   b) $Binom(2, 1/10)$,
   c) равномерное распределение $\mathcal{R}\{0, 1, 2\}$,
   d) $\mathbf{P}(X = 0) = \mathbf{P}(X = 2) = 3/8, \mathbf{P}(X = 1) = 1/4$ (X – число мальчиков).

   Из каждого распределения сгенерировать по 100 выборок. К каждой выборке применить критерий, построенный в задаче 1.2, получить p-value. Построить графики p-value для каждого распределения, сравнить их.

   1.3 Используем данные из файла Priem.csv. Давайте ответим на злободневный вопрос – отличаются ли мальчики и девочки в плане успешности сдачи ЕГЭ? Сравните на однородность суммарные баллы, баллы по русскому, баллы по математике.

2. **Вариант 2.**

   2.1 Закон Бенфорда в его житейской интерпретации утверждает, что если выбрать набор каких-то чисел из реальных данных, то распределение первых цифр этих чисел будет иметь вид $log_{10}(1 + 1/k)$, $k = 1, \ldots, 9$. Проверить его на списке стран по населению из файла Countries.csv

   2.2 Переходим к параметрической гипотезе.

   В понедельник в госпиталь поступило 37 человек, во вторник – 53, в среду – 35, в четверг – 27, в пятницу – 30, в субботу – 44, в воскресенье – 28. Проверить гипотезу о том, что пациенты попадают в госпиталь во вторник в два раза чаще, чем в четверг при $\alpha = 0.05$. (Для этой задачи ОМП $\hat{p}_i = N_i/n, i = 1, 3, 5, 6, 7$, $\hat{p}_4 = (N_2 + N_4)/(3n)$, $\hat{p}_2 = 2\hat{p}_4$ где N_i – число госпитализированных в i-й день недели, n – общее число пациентов.)

   2.2* Исследовать работу критерия для проверки сложной гипотезы на модельных данных. Рассмотреть следующие распределения:

   a) ОМП из предыдущей задачи,
   b) равномерное распределение по дням недели,
   c) распределение с $p_2 = 3p_4$,
   d) распределение с $2p_2 = p_4$.

   Из каждого распределения сгенерировать по 100 выборок. К каждой выборке применить критерий, построенный в задаче 2.2, получить p-value. Построить графики p-value для каждого распределения, сравнить их.

   2.3 Используем данные из файла Priem.csv. Давайте ответим на злободневный вопрос – насколько ЕГЭ по русскому вообще связан с ЕГЭ по математике. Проверить на независимость оба вида данных и сделать соответствующие выводы.

3. **Вариант 3.**

3.1 Начнем с простой гипотезы. Проверить гипотезу о равномерном распределении дней рождения в файле Priem.csv. Правда ли, что все дни месяца равновероятны? Правда ли, что все месяцы года равновероятно. Объясните результаты.

3.2 Продолжим параметрической гипотезой.

Из 1000 посетителей кафе 282 пили чай и 579 – кофе. С помощью критерия χ^2 проверить гипотезу о том, что кофе в среднем выбирают в 2 раза чаще чая при $\alpha = 0.05$.. (Для этой задачи ОМП $\hat{p}_t = (N_t + N_c)/3n$, $\hat{p}_c = 2\hat{p}_t$ где N_t и N_c – число выбравших чай и кофе соответсвенно, n – общее число посетителей.)

3.2* Исследовать работу критерия для проверки сложной гипотезы на модельных данных. Рассмотреть следующие распределения:

a) ОМП из предыдущей задачи,

b) равномерное распределение $\mathcal{R}\{1, \ldots, 3\}$

c) $1 + Binom(2, \hat{p})$.

Из каждого распределения сгенерировать по 100 выборок. К каждой выборке применить критерий, построенный в задаче 3.2, получить p-value. Построить графики p-value для каждого распределения, сравнить их.

3.3 Посмотрим на данные приема из Priem.csv. Животрепещущий вопрос: а насколько влияют баллы за ГТО, отличный аттестат и сочинение.

Сравните на однородность в плане суммы баллов каждую пару: имеющих значок ГТО и не имеющих, с отличным аттестатом и без, с сочинением и без.