Отчет о ДЗ №2 по Прикладной математической статистике
Исполнительница: Смолкина Ю.А
Группа: АДБМ

Задача 1

В лесу случайным образом было выбрано 7 участков одинаковой площади. На каждом участке был посчитано число взрослых сосен, росших на нём. Эти числа оказались такими: 7, 12, 9, 17, 10, 13, 15. Существенно ли варьирует число сосен?

Для данного задание используется Критерий Согласия Пирсона. Вычисляется статистика χ, после чего по таблице определяется соответствующее значение p для данного количества степеней свободы.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Подсчитанные значения:

Переменная	значение	пояснение
Obs (O)	[7,12,9,17,10,13,15]	Наблюдаемое количество деревьев на каждом участке
Exp (E)	11.86	Ожидаемое среднее количество деревьев на участке при верности нулевой гипотезы
chi^2	6.145	хи-квадрат статистики chi^2 = sum((Obs-Exp)**2/Exp)

Дано: 7 наблюдаемых элементов, по формуле вычисления степеней свободы -> степень_свободы = 6 (Это нужно, чтобы найти необходимое значение П-значения в таблице распределения Хи_квадрат).
П-значения находится между 0.3 и 0.5.

Вывод : Число сосен варьируется несущественно.

Использовалось в работе: Библиотеки scipy.stats, numpy

Задача 2

В каждом из двух прудов было поймано по 50 прудовиков. В 20 прудовиках из первого пруда и 32 прудовиках из второго были обнаружены личинки печёночных сосальщиков. На каком уровне значимости можно утверждать, что пруды различаются по заражённости прудовиков сосальщиком?

Возможные методы решения: приближение биномиального распределения к нормальному, тест Хи-квадрат, точный тест Фишера.

Выбран способ - приближения к к нормальному. Т.к число событий достаточно велико. Следующий шаг : ПОлучить Z-статистику и обратиться к таблице. Был выбран двусторонний критерий,тк в условии используется мн-ое число в постановке вопроса.

Переменная	значение	пояснение
n	20	
N	50	
m	32	
M	50	
p	(n+m)/(N+M)	Вероятность успеха
D1	p*(1-p)/N	Дисперсия
D2	p*(1-p)/M	

Z = -2.40
Далее сравниваем с таблицей нормального распределения
Получаем П-значение слегка меньше 0.0082*2=0.0164

Вывод: На уровне значимости 0.0164 можно утверждать о том, что пруды различаются по зараженности

Использовалось в проверке: Библиотека statsmodels

(-2.401922307076307, 0.016309171877754974)

Задача 3

Геном одного из штаммов вируса SARS-CoV-2 содержит 29903 нуклеотида, которые распределены так:

T	9594
A	8954
G	5863
C	5492

(замечание: носителем генома является РНК, которая содержит урацил (U) вместо тимина (T), но по сложившейся традиции в базах данных используется буква T и для тимина, или урацила).

В этом геноме 2377 раз встречается слово TA. Определите, имеется ли достоверное ($\alpha = 0,001$) отличие частоты этого слова от ожидаемой при предположении независимого появления букв в геноме (равновероятность букв не предполагается, рассматриваем наблюдаемые частоты отдельных букв).

Ход решения:

Т.к число испытаний велико и число успехов велико.
Можно использовать нормальное приближение биномиального распределения.
Имеем 9594 "T" после которых может попасться "A" с вероятностью A/N. То есть 9594 испытаний с вероятностью успеха = p.
Осталось вычислить ожидаемое число комбинаций " TA", при условии независимости появления букв в геноме:
μ = T*A/N , а также p и σ.
После этого, я проверю на сколько стандартных отклонений отличается наблюдаемое значение комбинации "TA" от среднего значения. Из-за того, что мы ищем отличие наблюдаемого значения от ожидаемого в любую сторону, то имеем двухсторонний случай.

Формулы: Z = (TA-mu)/sigma
mu = T*A/N
p = A / N
sigma = (N * p * (1-p)) ** 0.5

Z = -6.26

Из таблицы : Z = -3.0 достаточно для одностороннего случая
Z = -3.3 достаточно для двухстороннего случая
(чтоб отвергнуть нулевую гипотезу)

Вывод : При пороге значимости
α = 0.001 значимое отличие ожидаемой частоты TA от наблюдаемой имеется.

Задача 4

Из многолетних наблюдений известно, что средняя температура воды некоторого горячего источника составляет 61,5°С. В районе, где расположен этот источник, недавно произошло землетрясение и геологи хотят выяснить, не повлияло ли оно на температуру источника. В файле `task2_4.txt` находятся результаты измерений температуры источника, проведённые вскоре после землетрясения. На каком уровне значимости можно утверждать, что землетрясение повляло на источник?

Использование библиотек: все предыдущие и scipy.stat

Дано : количество наблюдений мало.

Ход решения: Задачу можно свести к биномиальному распределению.
Я буду предполагать, что распределение новых температур - после землетрясения - нормальное, проверить на нормальность (критерий Шапиро-Уилка).
После этого будет необходимо сравнить насколько отличается среднее до землетрясения от новоподсчитанного среднего нормального распределения после землетрясения.
T-test для n-1=18 степеней свободы.
По выбранному критерию statistic=0.9289318323135376,
П-значение =0.166 Критерий стьюдента, П-значение < 0.0803

Вывод: на уровне значимости 0.0803 мы утверждаем, что землетрясение повлияло на источник.

Задача 5

(Пример взят из книги: Бочаров П. П., Печинкин А. В. Теория вероятностей. Математическая статистика. 2-е изд. М.: ФИЗМАТЛИТ, 2005)

Для сравнительного анализа надежности крепёжных болтов, выпускаемых двумя заводами, были проверены на разрыв $m = 24$ изделия первого завода и $n = 20$ изделий второго. Силы натяжения ($\times 10^5$ Н), при которых произошли разрывы изделий первого и второго заводов, приведены в файле `task2_5.txt`.

Сравните эти две выборки по крайней мере одним (а лучше всеми) из известных вам методов и сделайте выводы.

Ход решения: У двух выборок предполагается одинаковая дисперсия и происхождение из нормального распределения.
средние двух выборок (2.61875, 3.6660000000000004)
Использовать буду метод - ANOVA, как эквивалент двухстороннему t-тесту.
Посчитаем Ф-статистику(F-stat) и ассоциированное П-значение

11.052098565131523, 0.0018451142120997227 - соответственно (функция f_oneway)

Вывод: значимое различие средних выборок при при пороге значимости ~0.002 или более.
Проверка: H_0, что обе выборки происходят из нормальных распределений, критерием Шапиро-Уилка не отклоняется (shapiro (массив выборки))
для 1-ой Изделия первого завода
(ShapiroResult(statistic=0.9654770493507385, pvalue=0.5577559471130371),
для 2-ой Изделия второго завода
(statistic=0.9843093752861023, pvalue=0.9770391583442688))

Посчитано с помощью shapiro(X), shapiro(Y)