Домашняя работа 2

Прикладная статистика, НИУ ВШЭ

3 марта 2023 г.

Срок выполнения: утро 7 марта 2023. Отчёт о решении каждой задачи должен включать краткое описание хода решения и результат. Обязательно достаточно полное описание материалов и методов: если использовались таблицы, то какие и откуда, если функции, то из какого пакета и т. п.

Отчёт можно представлять в форматах pdf, docx, odf и т.д. Можно и в ipynb (и аналогичных), но в таком случае обязателен комментарий к каждому действию: что делает используемая функция и почему используется именно она и именно с такими параметрами. В конце решения каждой задачи должен быть понятными словами сформулирован и обоснован ответ.
Отчёты отправляйте на адрес `sspirin@hse.ru`.

Система оценки

За правильное решение каждой задачи будет даваться 1 балл. Кроме того, ещё до 1 балла будет даваться за точность описания решения, решение несколькими способами и т.д.

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?

Задача 2

Препарат, призванный увеличить продолжительность жизни дрозофил, испытывался в четырёх разных концентрациях, отдельно на самцах и самках. При сравнении экспериментальных и контрольных групп были получены следующие p-значения: самцы: 0,045; 0,029; 0,0062; 0,52; самки: 0,26; 0,017; 0,0081; 0,14.

- На каком уровне значимости можно утверждать, что препарат вообще имеет эффект?
- Сколько вариантов эксперимента можно считать удачными (показавшими эффект) при допущении уровня ложных предсказаний (FDR) в 5%?

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 испольуется буква T и для тимина, или урацила).

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

Задача 4

Из многолетних наблюдений известно, что средняя температура воды некоторого горячего источника составляет 61,5°C. 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?

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.