ПРИКЛАДНАЯ СТАТИСТИКА

ДОМАШНЕЕ ЗАДАНИЕ №3

Доверительные интервалы. Критерии однородности.

Вам предстоит выполнить 4 задачи, в которых предстоит ответить на вопрос о наличии или отсутствии какого-либо эффекта, основываясь на имеющихся данных. Для ответа на заданный вопрос постарайтесь применить все возможные инструменты, которые мы изучали. Не пренебрегайте и таким инструментом как визуализация. Ваши решения необходимо сопроводить краткими комментариями и выводами, которые вы сделали на основе анализа. Для каждой задачи попытайтесь воспользоваться наиболее подходящим критерием из рассмотренных на лекции (важно использовать критерии для нормальных распределений там, где это возможно, потому что они точные и имеют большую мощность по сравнению с асимптотическими). Также в рамках визуализации как упражнение требуется построить не только гистограммы выборок, но и доверительные интервалы для параметров, которые вы собираетесь сравнивать (желательно, изобразив их на графике). Обязательно пишите формулу того доверительного интервала, который вы строили.

Напоминаем по мотивам прошедших семинаров - для того, чтобы выбрать, какой критерий использовать или как строить доверительный интервал, необходимо проверить нормальность распределения. К счастью, в **scipy.stats** вы можете найти функцию **normaltest** или же в **statsmodels** функцию **lilliefors**. Они реализуют одни из критериев проверки нормальности. Для того, чтобы ими воспользоваться нужно передать в них выборку и в ответе получите 2 числа: значение статистики (не интересует) и p-value. По p-value можно легко сделать вывод о нормальности выборки. Но не забывайте в явном виде проверять нулевую гипотезу!

1. **(20 баллов)** В файле **anorexia.txt** записан вес пациентов до начала терапии анорексии и после ее окончания. Была ли терапия эффективна?

```
data = pd.read_csv('anorexia.txt', sep = '\t')
```

2. **(20 баллов)** В файле **seattle.txt** записаны цены на недвижимость в одном из районов Сиэтла в 2001 году и в 2002 году (объекты недвижимости выбирались случайно). Изменились ли в среднем цены в этом районе за год?

```
data = pd.read_csv('seattle.txt', sep = '\t')
```

3. **(20 баллов)** В рамках исследования эффективности препарата метилфенидат пациенты с синдромом дефицита внимания и гиперактивности в течение недели принимали либо метилфенидат, либо плацебо. В конце недели каждый пациент проходили тест на способность к подавлению импульсивных поведенческих реакций. На втором этапе плацебо и препарат менялись, и после недельного курса каждый испытуемые прохо- дили второй тест. Был ли эффект от применения препарата? Данные находятся в файле **methylphenidate.txt**

```
data = pd.read_csv('methylphenidate.txt', sep = '\t')
```

4. В файле **mtcars.csv** находятся данные из американского журнала Motor Trend 1974 года. Они описывают расход топлива в зависимости от 10 характеристик автомобиля (все автомобили 1973-1974 года). Нас будут интересовать столбцы:

   mpg — расход топлива (миль/галлон);

   vs — тип двигателя (0 = V-образный, 1 = рядный);

   am — тип коробки передач (0 = автоматическая, 1 = ручная).

   Ответьте на следующие вопросы:

   (20 баллов) Влияет ли тип двигателя на расход топлива?

   (20 баллов) Влияет ли тип коробки передач?

```
data = pd.read_csv('mtcars.csv', index_col = 0)
```

5. *(Бонус 10 баллов) В предыдущих пунктах вам могло потребоваться проверить несколько гипотез для одной и той же выборки подряд. Допустим, вы проверяли четыре гипотезы: проверка нормальности каждой из подвыборок, равенство дисперсий и равенство средних. Каждую из гипотез вы проверяли на уровне значимости 0.05. Вопрос: какой уровень значимости получился у всей процедуры (то есть проверяя все четыре гипотезы вместе, какова вероятность сделать хотя бы одну ошибку первого рода)? Почему? Как думаете, как с этим можно бороться?

Замечания по проверке и выполнению заданий:

- В процессе использования normaltest может возникать предупреждение вида: *warnings.warn("kurtosistest only valid for n>=20 ... continuing "* Игнорируйте его в данной домашней работе. *(Бонус 1 балл) Почему ограничение именно на $n \geq 20$ и влияет ли это на результаты теста? Свой ответ обоснуйте.
- Отсутствие выводов: -30% за пункт.
- Отсутствие какой-либо визуализации: -20% за пункт.
- Отсутствие доверительных интервалов: -20% за пункт.
- Отсутствие проверок, которые необходимы для применения t-test'а (как самого мощного критерия из изученных на текущий момент): -10% за пункт.
- Использование не самого мощного из возможных критерия (а-ля можно было использовать t-test, а использовался z-test): -10% за пункт.
- Использование неподходящего критерия (а-ля вместо критерия для зависимых выборок использован критерий для независимых и наоборот): -20% за пункт.
- Красивый анализ, дополнительная визуализация, интересные наблюдения и выводы: до 10 дополнительных баллов за всё задание на усмотрение ассистента.