

## Домашнее задание по статистике №2

Это задание является необязательным, но если вы его сделаете, то его обязательно проверят и укажут на недостатки, если таковые будут.

1. Британские учёные выдвинули гипотезу о зависимости между мощностью голоса ( $X$ , в децибеллах) и арсеналом тяжёлых предметов ( $Y$ , в килограммах) на кухне у лондонских домохозяек

|   |    |     |    |    |     |    |    |    |    |    |    |    |    |     |    |
|---|----|-----|----|----|-----|----|----|----|----|----|----|----|----|-----|----|
| X | 87 | 135 | 65 | 80 | 112 | 77 | 93 | 91 | 55 | 58 | 63 | 90 | 76 | 105 | 34 |
| Y | 10 | 40  | 43 | 22 | 9   | 15 | 34 | 21 | 32 | 80 | 27 | 12 | 26 | 14  | 41 |

Проверить гипотезу о независимости этих двух факторов на уровне значимости 0, 1.

2. В исследовании оценивается эффективность поведенческой терапии для лечения анорексии. Для 50 пациентов известен вес до начала терапии и по её окончании. Проверить, была ли терапия эффективной, на уровне значимости 0.05. Данные содержатся в файле `weight.txt`.
3. Загрузите данные из набора Forest Fires (файл `forestfires.csv`) о лесных пожарах в Португалии. Задача состоит в том, чтобы с помощью линейной регрессии научиться предсказывать координату `area` (площадь пожара) в виде линейной комбинации других данных.

*Преобразование данных.* Чтобы работать с числовыми координатами, нечисловые координаты (`month`, `day`) нужно перевести в числовые. Для простоты можно заменить координату `month` на индикатор летнего сезона, а координату `day` не использовать вообще. По желанию можно сделать преобразование другим способом. Так же желательно добавить координату, тождественно равную единице.

Разбейте выборку на две части в соотношении 7:3 (перемешав её с помощью `random.shuffle`). По первой части постройте регрессионную модель. Примените модель ко второй части выборки и посчитайте по ней среднеквадратичную ошибку.

Сделайте для `area` преобразование  $f(x) = \ln(x + c)$  и постройте для нее новую регрессионную модель. Посчитайте среднеквадратичную ошибку для преобразованных значений. При каком  $c$  предсказания получаются лучше всего?

При выбранном  $c$  сделайте разбиение выборки в соотношении 7:3 разными способами (перемешивая каждый раз). Сильно ли зависит качество от способа разбиения? Сделайте выводы.

4. Выданы данные (`prostate.txt`) о связи между уровнем антигена (`lpsa`), специфичного при наличии рака простаты, с рядом клинических показателей (столбцы 1-8), которые были измерены у мужчин непосредственно перед проведением операции. Выбрав наилучшую модель линейной регрессии, предсказать `lpsa` на тестовой выборке. Описать и объяснить проделанные процедуры. Будьте внимательны и не используйте для обучения модели значения `lpsa` на тестовой выборке. Их можно использовать только для демонстрации качества проделанных процедур.