Прикладная статистика и анализ данных

Задание 1

Правила:

- Дедлайн **17 февраля 16:30**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[asda] Фамилия Имя - задание 1"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `1.N.ipynb` и `1.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качестве основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.
- Задание выполняется на языке R.
- В каждой задаче не забывайте делать **пояснения и выводы**.

Баллы за задание:

- Задача 1 - 5 баллов
- Задача 2 - 5 баллов
- Задача 3 - 10 баллов

In [ ]:

```
options(repr.plot.width = 5, repr.plot.height = 4)
```

Задача 1.

Пусть $X_1, \ldots, X_n$ -- выборка из гамма-распределения $\Gamma(\theta, 5)$. Постройте равномерно наиболее мощный критерий для проверки гипотез $\mathsf{H}_0 : \theta = 1 \ vs. \ \mathsf{H}_1 : \theta < 1$.

a). Сгенерируйте выборку $X_1, \ldots, X_{50}$ из гамма-распределения $\Gamma(\theta, 5)$ для случаев $\theta = 1$, $\theta = 0.3$ и $\theta = 5$. В каждом случае посчитайте p-value. В каких случаях основная гипотеза отвергается?

b). На одном графике постройте кривые мощности для разных размеров выборки (см. лекцию 12 прошлого семестра).

Задача 2.

Данные классического эксперимента Майкельсона по измерению скорости света с помощью вращающегося зеркала, 100 наблюдений:

In [ ]:

```
speed <- scan("speed.txt")
print(speed)
```

Требуется исследовать данные на нормальность. При построении графиков подписывайте оси и сам график.

Постройте гистограмму по данным

In [ ]:

```
...
```

Постройте график ядерной оценки плотности, на который нанесите также график плотности нормального распределения, параметры которого соответствуют оценке максимального правдоподобия по данным.

Замечание. Функция `plot` создает новую фигуру и рисует линию/точки. Функция `lines` рисует линию/точки на уже *существующей* фигуре. Тип линии или точек определяется параметром `type` .

In [ ]:

```
...
```

Постройте график эмпирической функции распределения, на который нанесите также график функции распределения нормального распределения, параметры которого соответствуют оценке максимального правдоподобия по данным. Добавьте на график сетку.

Замечание. При отрисовке ЭФР установите параметры:

- `verticals = TRUE` -- рисовать вертикальные линии;
- `pch = NA` -- не рисовать точки.

In [ ]:

```
...
```

Простройте по данным Q-Q plot

In [ ]:

```
...
```

Примените к данным критерии Лиллиефорса, Андерсона-Дарлинга, Крамера-фон Мизеса, Жарка-Бера и Шапиро-Уилка

In [ ]:

```
#install.packages('nortest')
library('nortest')
#install.packages('normtest')
library('normtest')
```

In [ ]:

```
...
```

Сохраните p-value каждого критерия в вектор. Для этого у результата нужно взять поле `p.value`. Не забудьте, что в R символ `.` (точка) является частью имени переменной.

In [ ]:

```
...
```

Примените к ним процедуру множественной проверки гипотез по методу Холма

In [ ]:

```
...
```

Сделайте выводы:

...

Задача 3.

Загрузите набор маркетинговых данных о влиянии рекламных СМИ (youtube, facebook и газеты) на продажи. Значения признаков представляют рекламный бюджет, целевая метка (sales) указана в некоторых условных единицах товара. Датасет будет записан в переменную `marketing`.

In [ ]:

```
load('marketing.rda')
```

1. Разбейте случайно данные на обучающую и тестовую часть в соотношении 3:1.
2. Обучите линейную регрессию по всем признакам, используя обучающую часть данных, и напечатайте таблицу свойств полученной модели. Что по ней можно сказать? В чем практический смысл коэффициентов линейной регрессии.
3. Напечатайте матрицу ковариаций оценки вектора параметров в предположении гомоскедастичности.
4. Проведите отбор признаков, используя информационные критерии AIC и BIC.
5. Проведите отбор признаков, минимизируя ошибку на тестовой части данных по метрикам MSE, MAE, MAPE.
6. Рассмотрим оптимальную модель по метрике MAPE. Зафиксируйте некоторые значения параметров facebook и newspaper. Постройте график зависимости предсказания от параметра youtube. Нанесите на график предсказательный интервал.