Иван Солоухин

In [1]:
```
import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
import scipy.stats as sts

import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
# зафиксируем random state для всех последующих случайных функций
RS = 1234
# и уровень значимости, одинаковый для задач
alpha = 0.05
```

№ 4.

In [3]:
```
# подготовка данных
df = pd.read_excel('https://docs.google.com/spreadsheets/d/e/2PACX-1vR19Hh8Jv
df = df[['Last name', 'Экзамен']].copy()
```

In [4]:
```
df.drop(index=df.loc[df['Экзамен'] == 'неявка'].index, inplace=True)
df.dropna(axis=0, inplace=True)
```

In [5]:
```
vowels = 'АЕЁИОУЫЭЮЯ'

crunch = np.ones((len(df), 1))
crunch[:] = np.nan
```

In [6]:
```
# создаем колонку: если фамилия начинается с гласной - единица, иначе ноль

crunch = np.ones((len(df), 1))
crunch[:] = np.nan
df['Vowel'] = crunch

for i in range(len(df)):
    if df.iloc[i, 0][0] in vowels:
        df.iloc[i, -1] = 1
    else:
        df.iloc[i, -1] = 0

df.tail()
```

Out[6]:

	Last name	Экзамен	Vowel
327	Сенников	13	0.0
328	Ся	30	0.0
329	Сятова	30	0.0
330	Темиркулов	30	0.0
331	Эшмеев	13	1.0

In [117…

```
# размер выборок
df.Vowel.value_counts()
```

Out[117…

```
Vowel
0.0    158
1.0     30
Name: count, dtype: int64
```

4а

In [7]:

```
# 4а - реализуем внутренню функцию scipy.stats
sts.ttest_ind(df.loc[df.Vowel == 0, 'Экзамен'],
              df.loc[df.Vowel == 1, 'Экзамен'],
              equal_var=False,
              alternative='two-sided')
# pvalue > alpha, гипотеза о равенстве матожиданий (результатах экзамена) не отвергает
```

Out[7]:

```
Ttest_indResult(statistic=0.42784294504977305, pvalue=0.671119924586115)
```

4б

In [8]:

```
# 4б
np.random.seed(RS)

obs_x = pd.to_numeric(df.loc[df.Vowel == 0, 'Экзамен']).to_numpy()
obs_y = pd.to_numeric(df.loc[df.Vowel == 1, 'Экзамен']).to_numpy()
n_x, n_y = len(obs_x), len(obs_y)

mean_diff = np.mean(obs_x)- np.mean(obs_y) # разница выборочных средних

inds_x = np.random.choice(np.arange(n_x), (10**4, n_x)) # случайные индексы для а
inds_y = np.random.choice(np.arange(n_y), (10**4, n_y))
mean_x_bs, mean_y_bs = np.mean(obs_x[inds_x], axis=1), np.mean(obs_y[inds_y],
# выборочные средние бутстрапированных выборок и их разница
means_diff_bs = mean_x_bs - mean_y_bs

# квантили для доверительных интервалов и p-value
q_l, q_r = np.percentile(means_diff_bs, 100*alpha/2), np.percentile(means_dif
p_value = 2*(np.min([np.mean((mean_diff < means_diff_bs)),
                     1 - np.mean(mean_diff >= means_diff_bs)]))

print(p_value)
# получается слегка больше единицы, чего вообще быть не должно; видимо, дело в численных
if p_value > alpha:
    print(f'Для уровня значимости {alpha} гипотеза НЕ отвергается')
else:
    print(f'Для уровня значимости {alpha} гипотеза отвергается')
```

```
1.0148
Для уровня значимости 0.05 гипотеза НЕ отвергается
```

№ 6

In [19]:

```
# 6а
df['len'] = pd.to_numeric(df['Last name'].apply(len)) # новая колонка в датафрей
b_score = df['Экзамен'].mean() / df.len.mean() # бета через выборочное среднее (пере
corr = np.corrcoef(pd.to_numeric(df['Экзамен']), df['len'])[1, 0]
# оценка методом моментов и очень слабая положительная корреляция
print(f'Beta={b_score}, коэффициент корреляции={corr}')
```

```
Beta=2.196877121520706, коэффициент корреляции=0.011115389336859618
```

In [32]:

```python
# 6б
# реализация перестановочного теста и проверка гипотезы beta=0

corrcoefs = []
for i in range(10**4):
    perestanovka = np.random.choice(pd.to_numeric(df['Экзамен']), size=len(df)
    corr_perestanovka = np.corrcoef(perestanovka, df['len'])[1, 0]
    corrcoefs.append(corr_perestanovka)
corrcoefs = np.array(corrcoefs)
q_l, q_r = np.percentile(corrcoefs, 100*alpha/2), np.percentile(corrcoefs, 10
p_value = 2 * min(1 - (corrcoefs < 0).sum()/len(corrcoefs),
                  (corrcoefs < 0).sum()/len(corrcoefs))
print(p_value)
if p_value > alpha:
    print(f'Для уровня значимости {alpha} гипотеза НЕ отвергается')
else:
    print(f'Для уровня значимости {alpha} гипотеза отвергается')
```

```
0.9950000000000001
Для уровня значимости 0.05 гипотеза НЕ отвергается
```

№ 8

https://arxiv.org/pdf/2106.00871.pdf \ Доказательство главной теоремы всего матстата, которое можно понять после вышкинского курса математического анализа и ТВиМС. \ Помогло бы разобраться с тем, откуда берется ЦПТ, если бы я свободно ориентировался в уже изученном материале. Поскольку условие не очень выполняется, помогло разобраться с планами на это лето.