Лабораторная работа №1
Выполнил: Кузнецов Павел М3207
осваиваю LaTeX:)

Задача 1. Пусть выборка $X_1, \ldots, X_n$ соответсвует классу распределений F_θ, $\theta \in E \subset \mathbb{R}$. При каком минимальном объёме выборки n равномерно для $\theta \in E$ выборочное среднее отличается от математического ожидания μ_θ не более чем на $\varepsilon > 0$ с вероятностью, не меньшей $1-\delta$, $\delta \in (0,1)$? Сгенерировать 500 выборок найденного объема при $\varepsilon = 0.01$ и $\delta = 0.05$ из указанного распределения F_θ при конкретном параметре θ и посчитать, сколько раз выборочное среднее отличается от математического ожидания μ_θ более чем на ε.
$F_\theta = Bern(p)$, $p \in (0,1)$, $p = 2/3$

Решение:

sum-up: у нас есть некоторая выборка $X_1, \ldots, X_n$, которая соответсвует классу распределений $F_\theta = Bern(p)$, $p \in (0,1)$, нужно найти $\min n : P(|\bar{X}_n - \mu| \leq \varepsilon) \geq 1-\delta$

$P(|\bar{X}_n - \mu| \leq \varepsilon) = P(\frac{\sqrt{n}|\bar{X}_n - \mu|}{\sigma} \leq \frac{\sqrt{n}\varepsilon}{\sigma}) \approx 2\Phi(\frac{\sqrt{n}\varepsilon}{\sigma})$ (из центральной предельной теоремы, плотность ст. нормального распределения симметрична)

$2\Phi(\frac{\sqrt{n}\varepsilon}{\sigma}) \geq 1-\delta$

$\Phi(\frac{\sqrt{n}\varepsilon}{\sigma}) \geq \frac{1-\delta}{2}$

(так как функция Лапласа нормального распределения возрастает от аргумента, минимальное n будет при минимальном значении Φ)

$\Phi(\frac{\sqrt{n}\varepsilon}{\sigma}) = \frac{1-\delta}{2}$

$\frac{\sqrt{n}\varepsilon}{\sigma} = \Phi^{-1}(\frac{1-\delta}{2})$

$$n_{min} = (\frac{\sigma\Phi^{-1}(\frac{1-\delta}{2})}{\varepsilon})^2$$

Найдём конкретные значение n_{min} при заданных значениях $\theta, \varepsilon, \delta$:

```
import numpy
from scipy.stats import norm

#mathematical expectation
p = 2/3

eps = 0.01
delta = 0.05

#standard deviation from the Bernoulli distribution
sigma = numpy.sqrt(p * (1 - p))

n = pow(sigma*norm.ppf(1 - delta/2)/eps,2)
n_min = round(n)
print(n_min)
```

Получившееся значение $n_{min} = 8537$.

Сгенерируем 500 выборок найденного объема при $\varepsilon = 0.01$ и $\delta = 0.05$ из указанного распределения F_p при конкретном параметре p и посчитаем, сколько раз выборочное среднее отличается от математического ожидания μ_p более чем на ε:

```
for i in range(500):
    X = numpy.random.binomial(n, p)/n
    if (abs(X - p) > eps):
        count += 1
```

Вывод:

Выборочное среднее отличается от мат. ожидания более чем на epsilon 22 раз. Это соответсвует 4.4 % от 500 выборок.

Задача 2-4. В файле mobile_phones.csv приведены данные о мобильных телефонах. В сколько моделей можно вставить 2 сим-карты, сколько поддерживают 3G, каково наибольшее число ядер у процессора? Рассчитайте выборочное среднее, выборочную дисперсию, выборочную медиану и выборочную квантиль порядка 2/5, построить график эмпирической функции распределения, гистограмму и box-plot для емкости аккумулятора для всей совокупности и в отдельности для поддерживающих/не поддерживающих Wi-Fi.

Решение: (код)

```
import numpy
from scipy.stats import norm
import pandas
import matplotlib.pyplot as plt
import statsmodels.api as statsm
```

Листинг 1: Imports

```

data = pandas.read_csv("/Users/frogge/proggs/mathstatLabs/lab-1/mobile_phones.csv")
# how many models are available to install 2 sim cards
print(data['dual_sim'].sum())

# how many models support 3-G
print(data['three_g'].sum())

# maximum number of cores
print(data['n_cores'].max())
```

Получившиеся значения:

available to install 2 sim cards = 1019

support 3-G = 1523

maximum numb of cores = 8

Далее считаем значения (указанные в задании), строим график, гистограмму и box-plot для емкостей аккумулятора при разных выборках:

```
new_data = data['battery_power']

# selective mean
print(round(new_data.mean(), 4))

# dispersion
print(round(new_data.var(), 4))

# median
print(round(new_data.median(), 4))

# quantile 2/5
print(round(new_data.quantile(q = 0.4), 4))

x = numpy.linspace(min(new_data), max(new_data))
y = statsm.distributions.ECDF(new_data)(x)
plt.step(x, y)
plt.title("Graph of ecdf")
plt.show()
#plt.hist(new_data, histtype='step', cumulative=True, bins=len(
    sample))     -     another way to plot a graph

plt.hist(new_data)
plt.title("Histogram")
plt.show()

plt.boxplot(new_data)
plt.title("Box-plot")
plt.show()
```

Листинг 2: for all data

Получившиеся значения:

Выборочное среднее = 1238.5185

Выборочная дисперсия = 193088.3598

Выборочная медиана = 1226.0

Выборочная квантиль порядка 2/5 = 1076.0

Графики в приложении

Далее считаем для выборки из даты, только для моделей поддерживающих WI-FI и наоборот, код остаётся прежним меняем только значение new_data

```
    new_data = data[data['wifi'] == 1]['battery_power'] #with wf
    new_data = data[data['wifi'] == 0]['battery_power'] #without wf
```

Получившиеся значения:
Выборочное среднее with wf = 1234.9043
Выборочная дисперсия with wf= 190296.4005
Выборочная медиана with wf= 1233.0
Выборочная квантиль порядка 2/5 with wf = 1077.8
Выборочное среднее without= 1242.2353
Выборочная дисперсия without= 196128.438
Выборочная медиана without= 1222.0
Выборочная квантиль порядка 2/5 without= 1076.8

Приложение

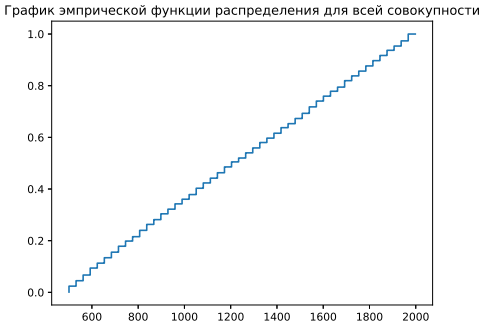
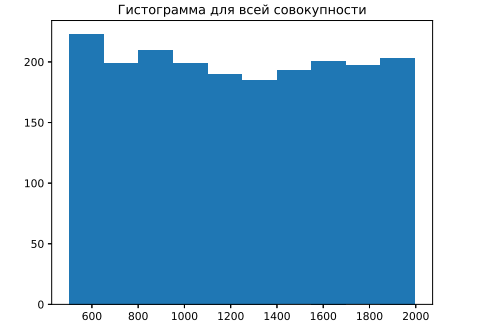


График эмпрической функции распределения для всей совокупности
1.0
0.8
0.6
0.4
0.2
0.0
600
800
1000
1200
1400
1600
1800
2000


Гистограмма для всей совокупности
200
150
100
50
0
600
800
1000
1200
1400
1600
1800
2000


Box-plot для всей совокупности
2000
1800
1600
1400
1200
1000
800
600
1

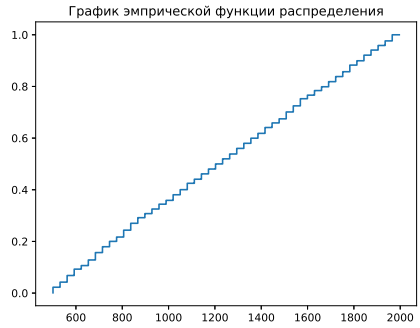
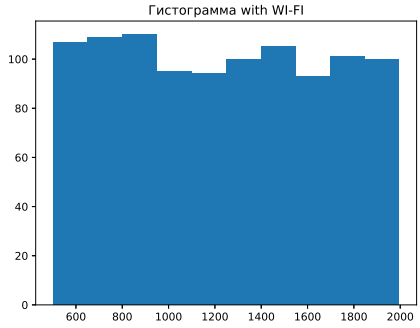
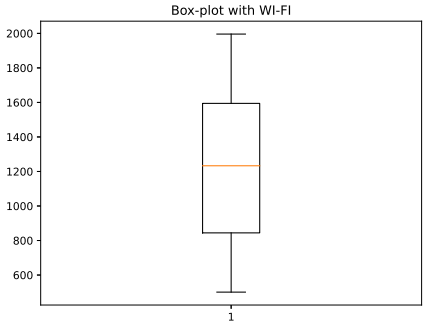


График эмпрической функции распределения
1.0
0.8
0.6
0.4
0.2
0.0
600
800
1000
1200
1400
1600
1800
2000


Гистограмма with WI-FI
100
80
60
40
20
0
600
800
1000
1200
1400
1600
1800
2000


Box-plot with WI-FI
2000
1800
1600
1400
1200
1000
800
600
1

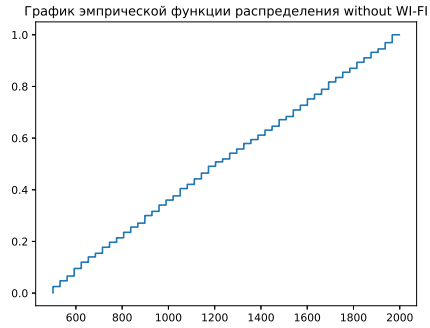
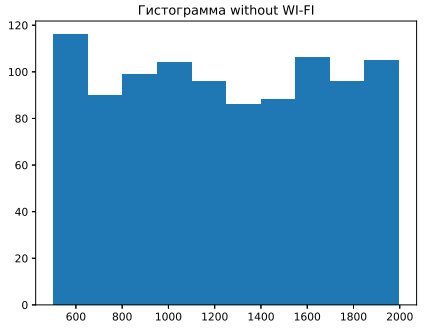
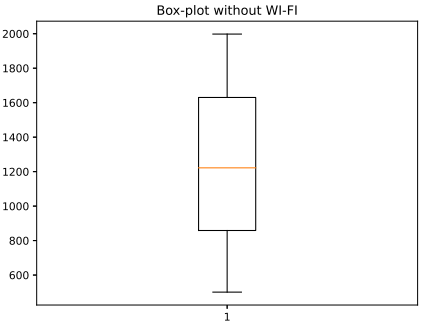


График эмпрической функции распределения without WI-FI
1.0
0.8
0.6
0.4
0.2
0.0
600
800
1000
1200
1400
1600
1800
2000


Гистограмма without WI-FI
120
100
80
60
40
20
0
600
800
1000
1200
1400
1600
1800
2000


Box-plot without WI-FI
2000
1800
1600
1400
1200
1000
800
600
1