Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

—

Мегафакультет трансляционных информационных технологий

**Институт дизайна и урбанистики**

**ОТЧЕТ ПО ЗАДАНИЮ №2**

Выполнил

Студент гр. **C4110** Егоров П.Н.

Преподаватель Леоненко В.Н.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 г.

Санкт-Петербург

2019

**Задание:**

1. Найти выборку непрерывных измерений (N>20).
2. Отобразить значения графически на одномерном интервале и проверить выборку на наличие грубых ошибок.
3. Написать скрипт для определения математического ожидания, дисперсии, моды, медианы.
4. Построить гистограмму и выдвинуть гипотезу о виде распределения соответствующей случайной величины.

**Выполнение**:

1. Была выбрана выборка возрастов пассажиров корабля «Титаник» из файла **«train.csv»** из сайта:[**https://www.kaggle.com/c/titanic/data**](https://www.kaggle.com/c/titanic/data).
2. Проверка выборки на выбросы была произведена с помощью интерквантильного расстояния IQR:

****

Тогда диапазон:

****

Как будет видно на финальном графике, все значения данных входят в этот интервал, значит выбросов нет.

1. Был написан следующий скрипт на языке Python:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Считывание данных

titanic\_data = pd.read\_csv('train.csv')

ages = titanic\_data.Age

# Подсчет параметров и их вывод

param = []

param.append(ages.mean())

param.append(ages.var())

param.append(ages.median())

param.append(ages.mode()[0])

param\_names = [ 'Mean (Yellow)', 'Variance', 'Median (Blue)', 'Mode (Green)' ]

for i in range(len(param)):

print(param\_names[i] + ' =', param[i], end='\n \n')

# Проверка на выбросы

Q3 = ages.quantile(0.75)

Q1 = ages.quantile(0.25)

IQR = Q3 - Q1

interval = [Q1 - 3 \* IQR, Q3 + 3 \* IQR]

# Отрисовка гистограммы, параметров и диапазона проверки на выбросы

ages.hist(bins=22)

plt.axvline(x=param[0], color='y', linestyle='-')

plt.axvline(x=param[2], color='b', linestyle='-.')

plt.axvline(x=param[3], color='g', linestyle='--')

plt.axvline(x=interval[0], color='r', linestyle='-')

plt.axvline(x=interval[1], color='r', linestyle='-')

plt.show()

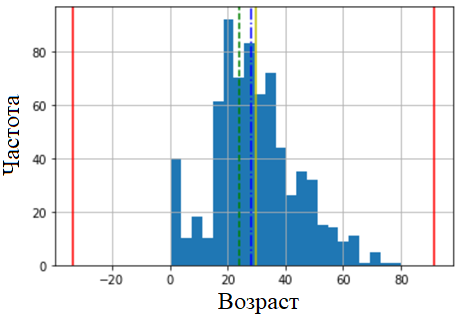
1. Была построена следующая гистограмма:

Рис. 1. Гистограмма

И следующие значения параметров:

Mean (Yellow) = 29.69911764705882

Variance = 211.01912474630802

Median (Blue) = 28.0

Mode (Green) = 24.0

На Рис.1 также обозначены:

* Математическое ожидание **желтой линией**.
* Медиана ‑ **синия линия**.
* Мода ‑ **зеленая линия**.
* Диапазон проверки на выбросы – **красные линии**.

Выдвигается гипотеза о том, что данная выборка имеет либо нормальное распределение с положительной асимметрией, либо отрицательное биноминальное распределение.