FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
ITMO UNIVERSITY

# Report on learning practice №1 Analysis of univariate random variables

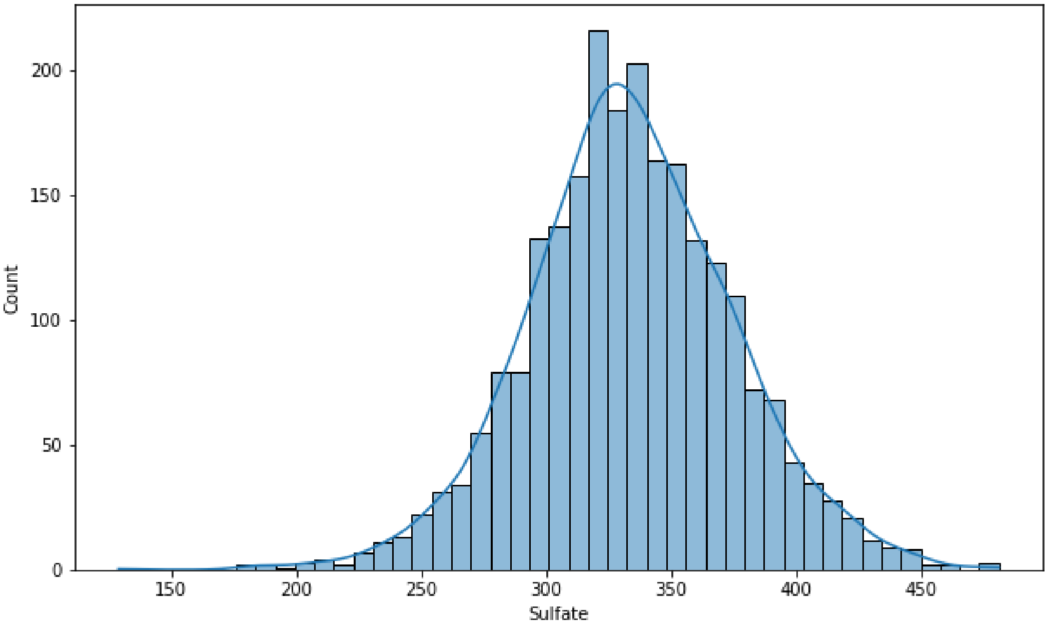
**Performed by:**
Roman Bezaev
Andrey Getmanov
J4133c

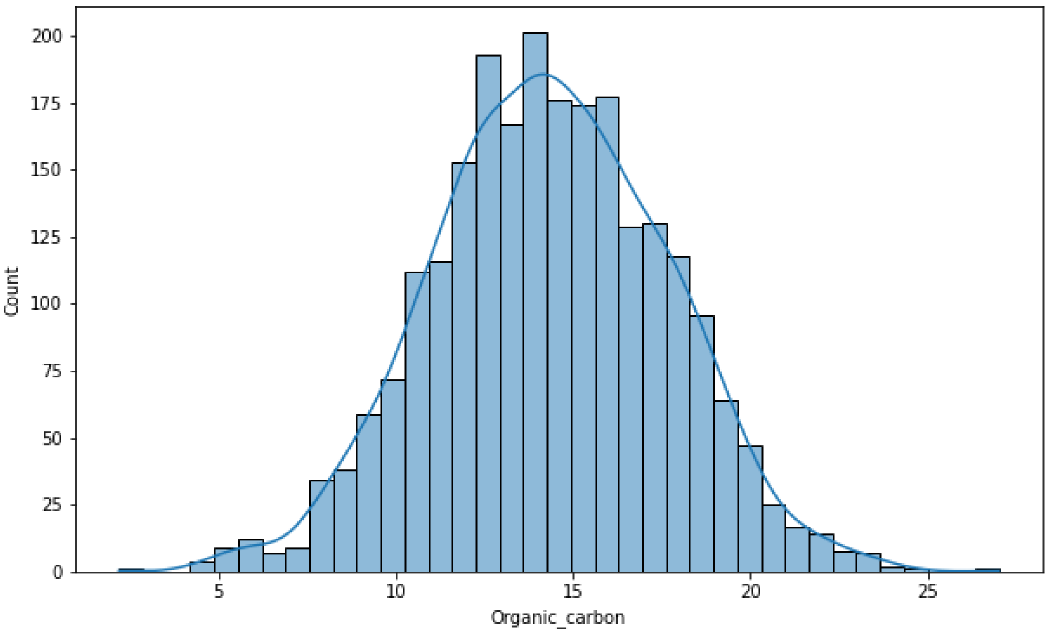
St. Petersburg
2021

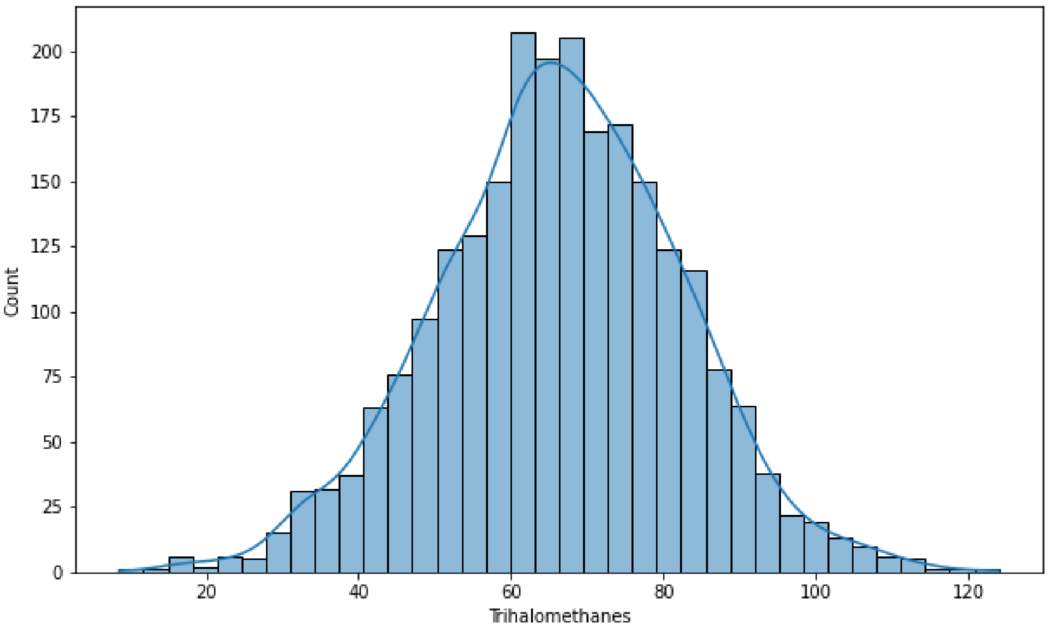
# 1 Substantiation of chosen subsample

As a dataset Water Potability was chosen. Subsample included such columns as Sulfate, Organic carbon and Trihalomethanes those defines sulfate concentration, amount of carbon and concentration of THMs, respectively.

# 2 Plotting a non-parametric estimation of PDF in form of a histogram and using kernel density function

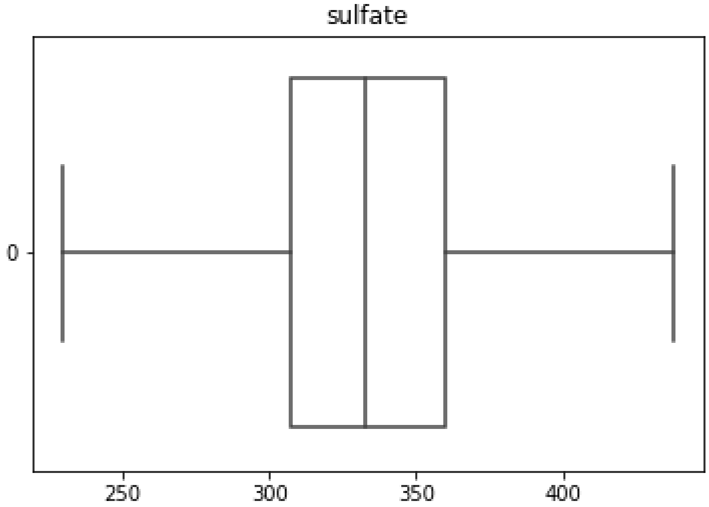


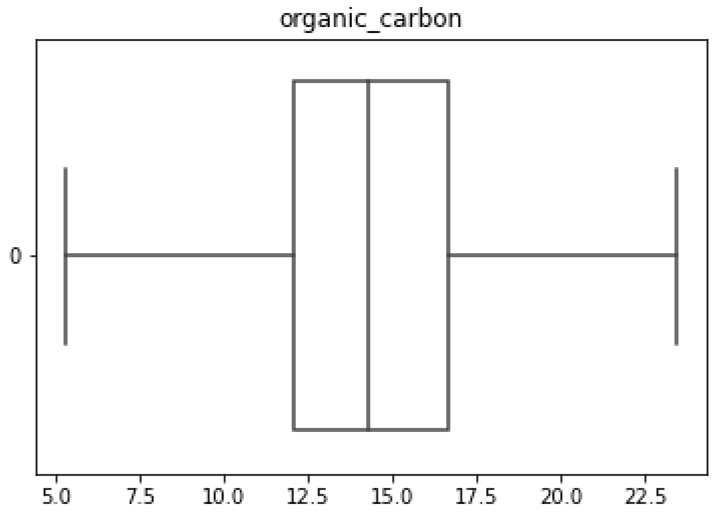

Count
0
25
50
75
100
125
150
175
200
5
10
15
20
25
Organic_carbon

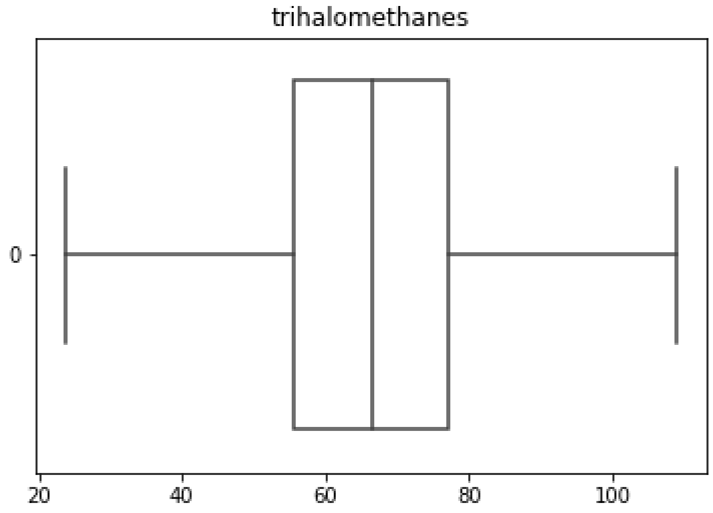

Count
0
25
50
75
100
125
150
175
200
20
40
60
80
100
120
Trihalomethanes

## 3 Order statistics estimation and its representation as "box with whiskers" plot

| | Sulfate | Organic carbon | Trihalomethanes |
|---|---|---|---|
| min | 129 | 2.2 | 8.57 |
| 25% | 307.56 | 12.08 | 55.7 |
| 50% | 332.89 | 14.26 | 66.54 |
| 75% | 359.94 | 16.64 | 77.15 |
| max | 481.03 | 27 | 124 |




organic_carbon


trihalomethanes

# 4 Selection of theoretical distributions that best reflect empirical data

According to the bell-shaped histograms, for every variable three different distributions was tried: normal, t-student and exponentiated Weibull.

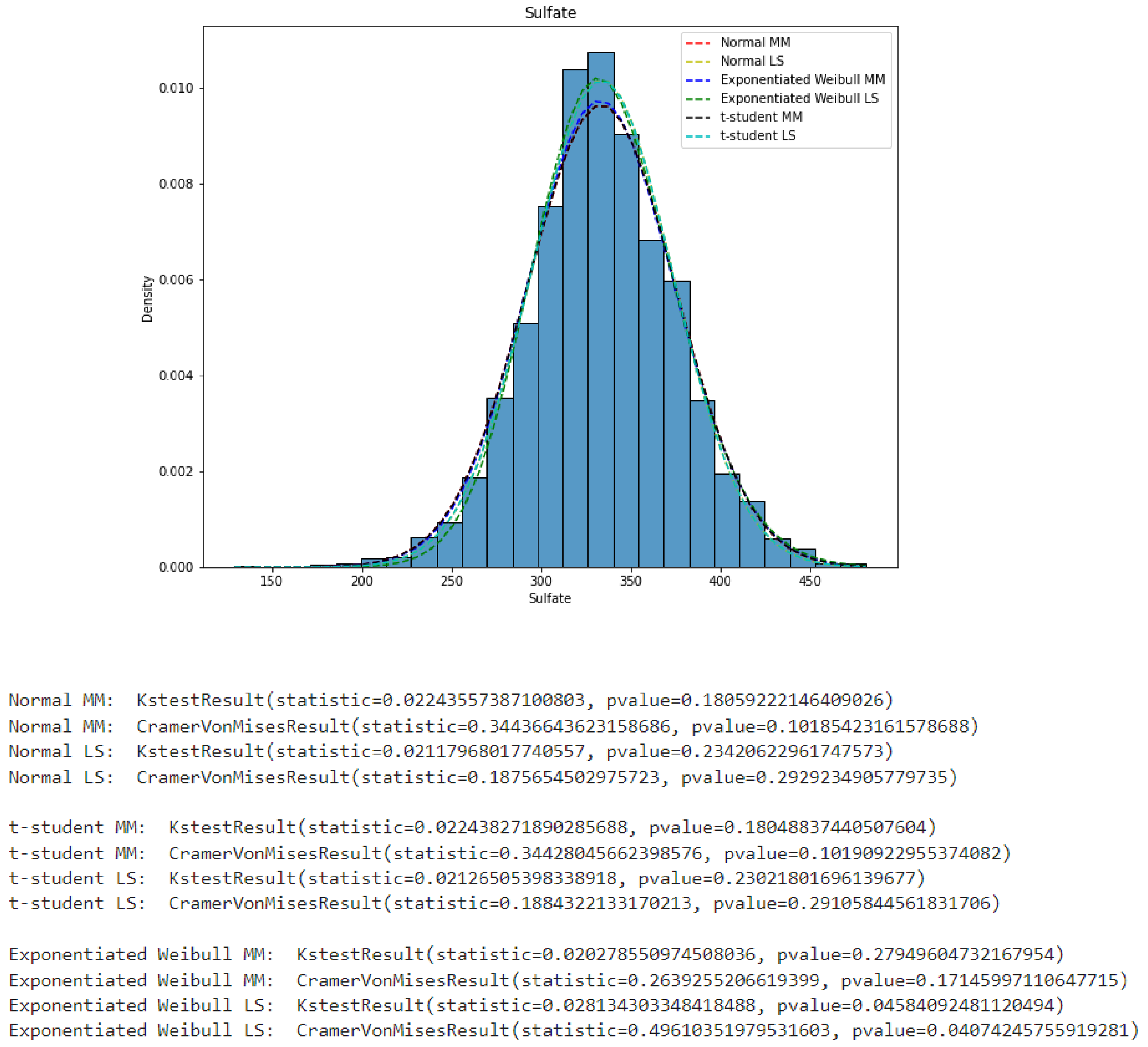
For estimating correctness of normal distribution we use two tests:

1. Kolmogorov-Smirnov test

2. Cramér–von Mises test

# 5 Estimation of random variable distribution parameters using maximum likelihood technique and LS methods

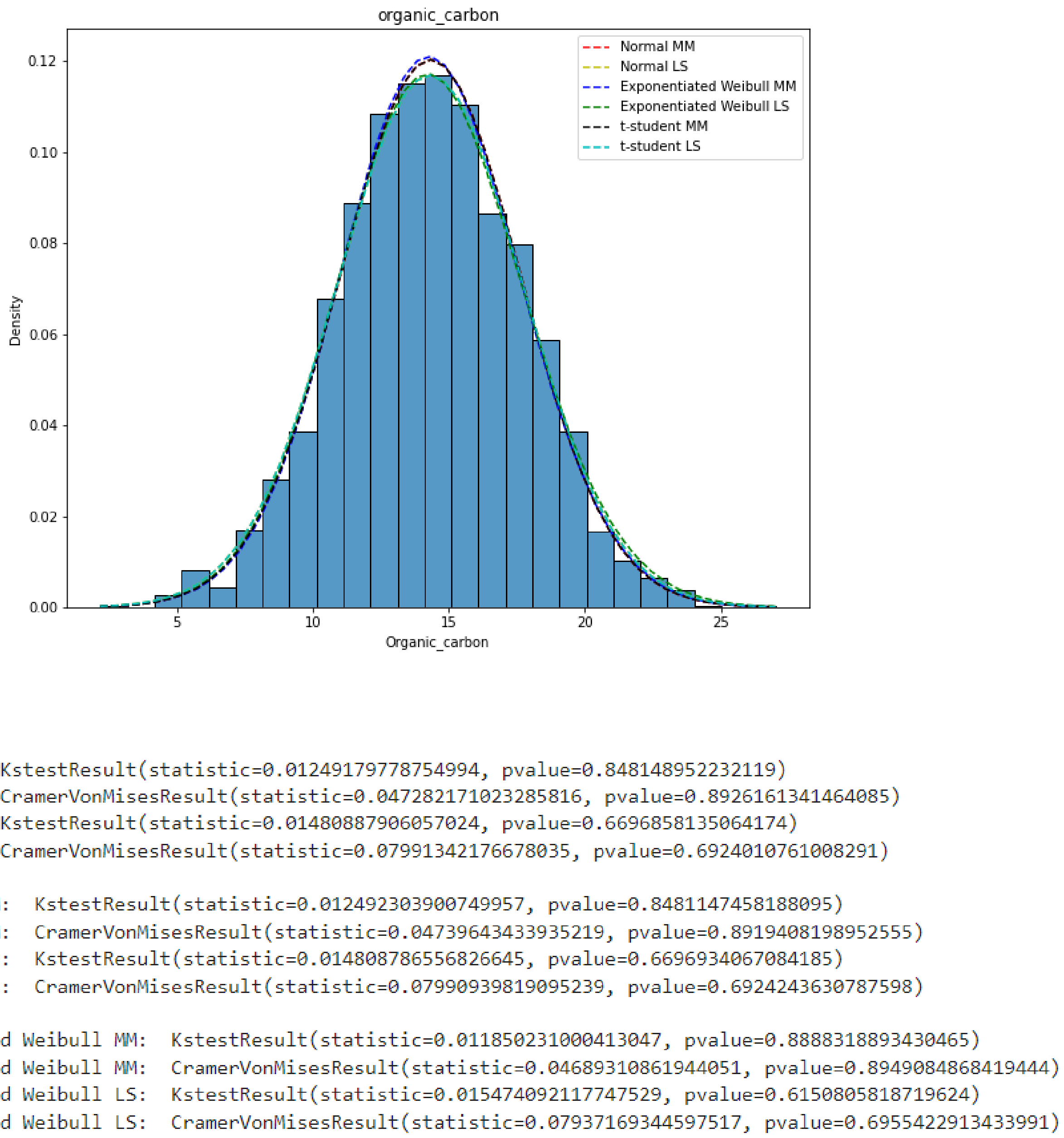
More often least squares method has better results in terms of p-value, than method of moments.



```
Normal MM:  KstestResult(statistic=0.02243557387100803, pvalue=0.18059222146409026)
Normal MM:  CramerVonMisesResult(statistic=0.34436643623158686, pvalue=0.10185423161578688)
Normal LS:  KstestResult(statistic=0.02117968017740557, pvalue=0.23420622961747573)
Normal LS:  CramerVonMisesResult(statistic=0.1875654502975723, pvalue=0.2929234905779735)

t-student MM:  KstestResult(statistic=0.022438271890285688, pvalue=0.18048837440507604)
t-student MM:  CramerVonMisesResult(statistic=0.34428045662398576, pvalue=0.10190922955374082)
t-student LS:  KstestResult(statistic=0.02126505398338918, pvalue=0.23021801696139677)
t-student LS:  CramerVonMisesResult(statistic=0.1884322133170213, pvalue=0.29105844561831706)

Exponentiated Weibull MM:  KstestResult(statistic=0.020278550974508036, pvalue=0.27949604732167954)
Exponentiated Weibull MM:  CramerVonMisesResult(statistic=0.2639255206619399, pvalue=0.17145997110647715)
Exponentiated Weibull LS:  KstestResult(statistic=0.028134303348418488, pvalue=0.04584092481120494)
Exponentiated Weibull LS:  CramerVonMisesResult(statistic=0.49610351979531603, pvalue=0.04074245755919281)
```

Figure 1: Sulfate tests

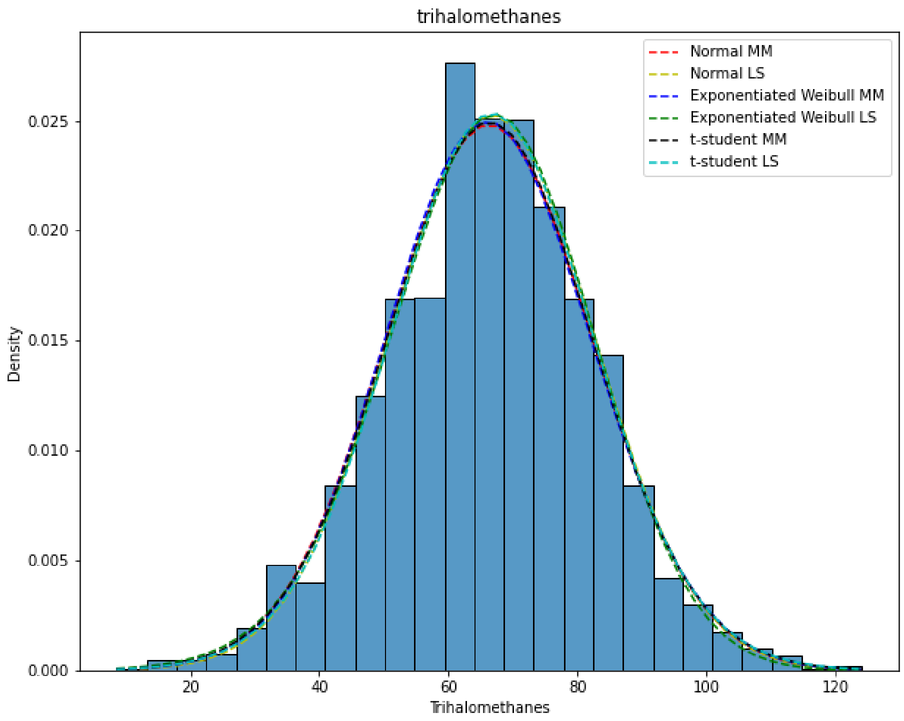
Based on the test results, we can conclude that this is more likely a normal or t-student distribution. For further analysis we'll assume that it is normal with least squares estimation parameters.

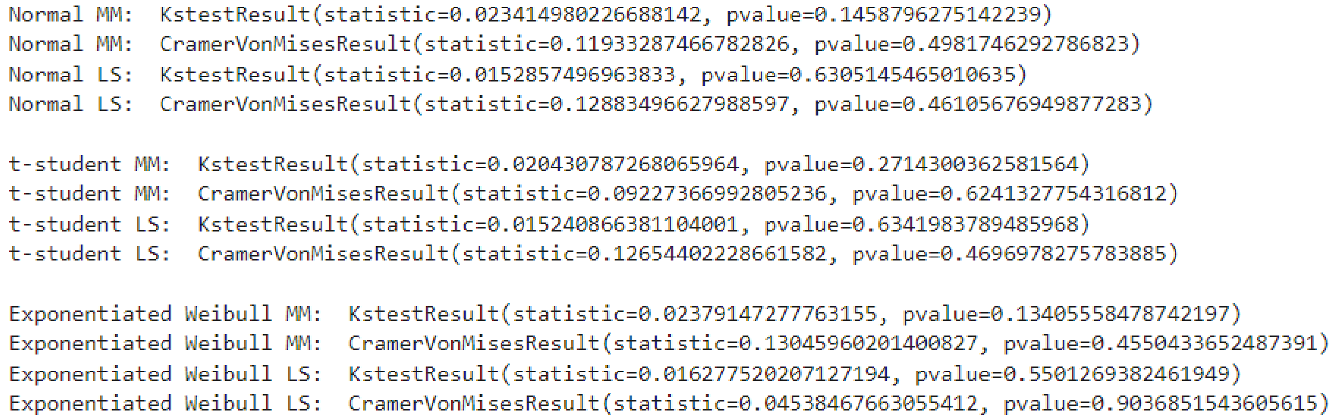


```
Normal MM:  KstestResult(statistic=0.01249179778754994, pvalue=0.848148952232119)
Normal MM:  CramerVonMisesResult(statistic=0.047282171023285816, pvalue=0.8926161341464085)
Normal LS:  KstestResult(statistic=0.01480887906057024, pvalue=0.6696858135064174)
Normal LS:  CramerVonMisesResult(statistic=0.07991342176678035, pvalue=0.6924010761008291)

t-student MM:  KstestResult(statistic=0.012492303900749957, pvalue=0.8481147458188095)
t-student MM:  CramerVonMisesResult(statistic=0.04739643433935219, pvalue=0.8919408198952555)
t-student LS:  KstestResult(statistic=0.014808786556826645, pvalue=0.6696934067084185)
t-student LS:  CramerVonMisesResult(statistic=0.07990939819095239, pvalue=0.6924243630787598)

Exponentiated Weibull MM:  KstestResult(statistic=0.011850231000413047, pvalue=0.8888318893430465)
Exponentiated Weibull MM:  CramerVonMisesResult(statistic=0.04689310861944051, pvalue=0.8949084868419444)
Exponentiated Weibull LS:  KstestResult(statistic=0.015474092117747529, pvalue=0.6150805818719624)
Exponentiated Weibull LS:  CramerVonMisesResult(statistic=0.07937169344597517, pvalue=0.6955422913433991)
```

Figure 2: Organic carbon tests

Based on the test results, we can conclude that this is more likely an exponentiated Weibull distribution. Parameters from method of moments will be used for QQ-plot.



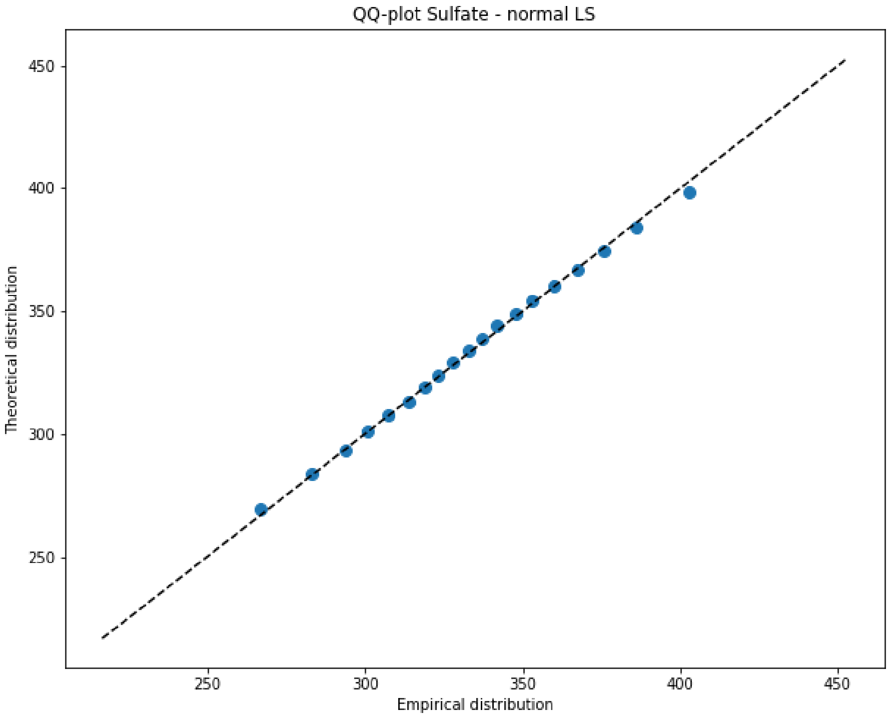


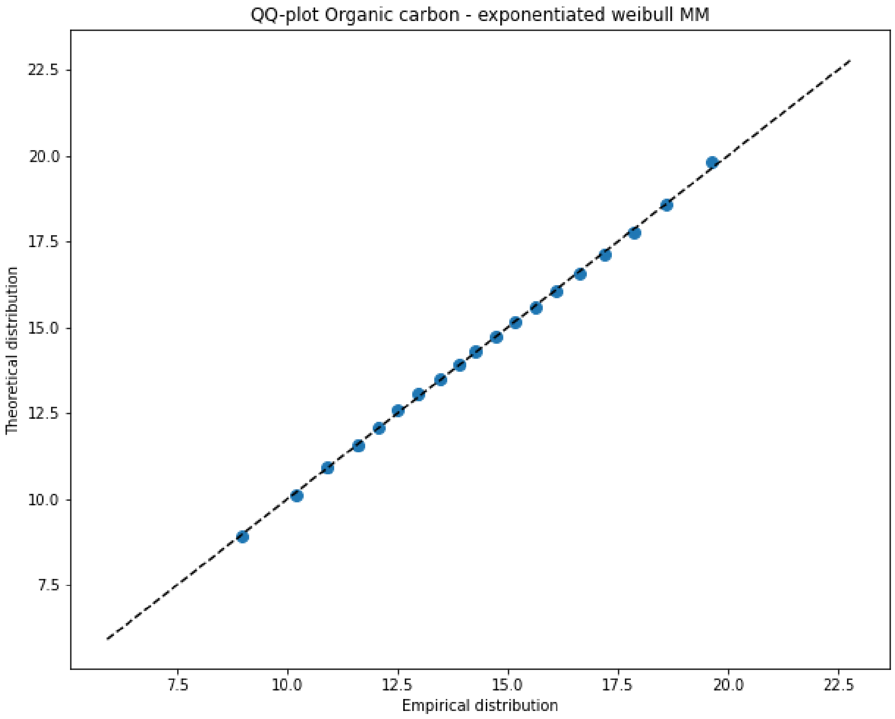
```
Normal MM:  KstestResult(statistic=0.023414980226688142, pvalue=0.1458796275142239)
Normal MM:  CramerVonMisesResult(statistic=0.11933287466782826, pvalue=0.4981746292786823)
Normal LS:  KstestResult(statistic=0.0152857496963833, pvalue=0.6305145465010635)
Normal LS:  CramerVonMisesResult(statistic=0.12883496627988597, pvalue=0.46105676949877283)

t-student MM:  KstestResult(statistic=0.020430787268065964, pvalue=0.2714300362581564)
t-student MM:  CramerVonMisesResult(statistic=0.09227366992805236, pvalue=0.6241327754316812)
t-student LS:  KstestResult(statistic=0.015240866381104001, pvalue=0.6341983789485968)
t-student LS:  CramerVonMisesResult(statistic=0.12654402228661582, pvalue=0.4696978275783885)

Exponentiated Weibull MM:  KstestResult(statistic=0.02379147277763155, pvalue=0.13405558478742197)
Exponentiated Weibull MM:  CramerVonMisesResult(statistic=0.13045960201400827, pvalue=0.4550433652487391)
Exponentiated Weibull LS:  KstestResult(statistic=0.016277520207127194, pvalue=0.5501269382461949)
Exponentiated Weibull LS:  CramerVonMisesResult(statistic=0.04538467663055412, pvalue=0.9036851543605615)
```

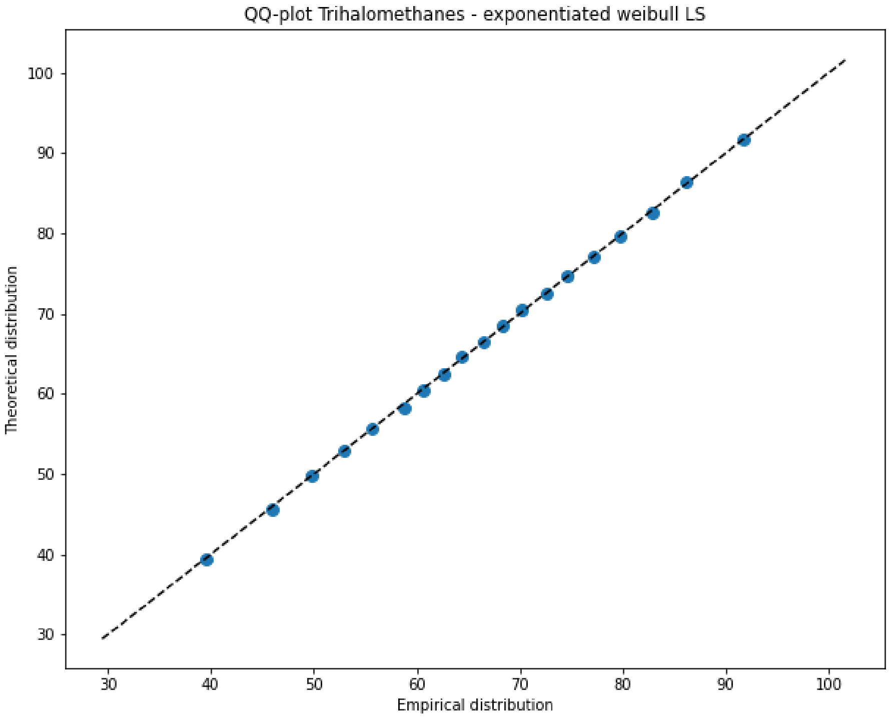
Figure 3: Trihalomethanes tests

Based on the test results, it's hard to choose the exact distribution, but according to the pvalue of CramerVonMises test of exponentiated Weibull method this distribution was chosen. Distribution with least squares estimation parameters will be used for QQ-plot.

# 6 Validation of empirical and theoretical distributions




QQ-plot Organic carbon - exponentiated weibull MM
Theoretical distribution
Empirical distribution
22.5
20.0
17.5
15.0
12.5
10.0
7.5
7.5
10.0
12.5
15.0
17.5
20.0
22.5


QQ-plot Trihalomethanes - exponentiated weibull LS
Theoretical distribution
Empirical distribution
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100

## 7 Source code

https://github.com/trixdade/ITMO-Methods-and-models-for-data-analysis/blob/master/Lab1/Lab1.ipynb