Report on learning practice #1

Analysis of univariate random variables

Performed by
Alexander Razin
Mikhail Lovtskiy
Mark Evgrafov
Julia Pimkina
Ac. group J4132c

Saint-Petersburg
2022

Table of contents:

1. Substantiation of chosen subsample.

	steroid	systolic_blood_pressure	diastolic_blood_pressure	heart_rate	highest_heart_rate	lowest_heart_rate	lowest_urine_output
0	No	119.0	54.0	79.0	80.0	73.0	700.0
1	Yes	133.0	64.0	73.0	73.0	69.0	1351.0
2	Yes	140.0	74.0	70.0	72.0	65.0	1420.0
3	Yes	154.0	78.0	77.0	80.0	72.0	350.0
4	Yes	155.0	61.0	64.0	77.0	64.0	NaN
5	Yes	156.0	74.0	62.0	71.0	62.0	530.0
6	Yes	158.0	69.0	81.0	81.0	76.0	NaN
7	No	155.0	68.0	104.0	108.0	101.0	NaN
8	Yes	122.0	80.0	72.0	92.0	72.0	NaN
9	Yes	114.0	72.0	60.0	67.0	60.0	NaN

Fig.1. The first lines of the dataset.

Chosen dataset is selected from a curated collection of over 200 publicly available COVID-19 related datasets from sources like Johns Hopkins, the WHO, the World Bank, the New York Times, and many others. It includes data on a wide variety of health indicators data.

In learning practice #1 four variables have been used. In this work have been analyzed the main cardiac indicators

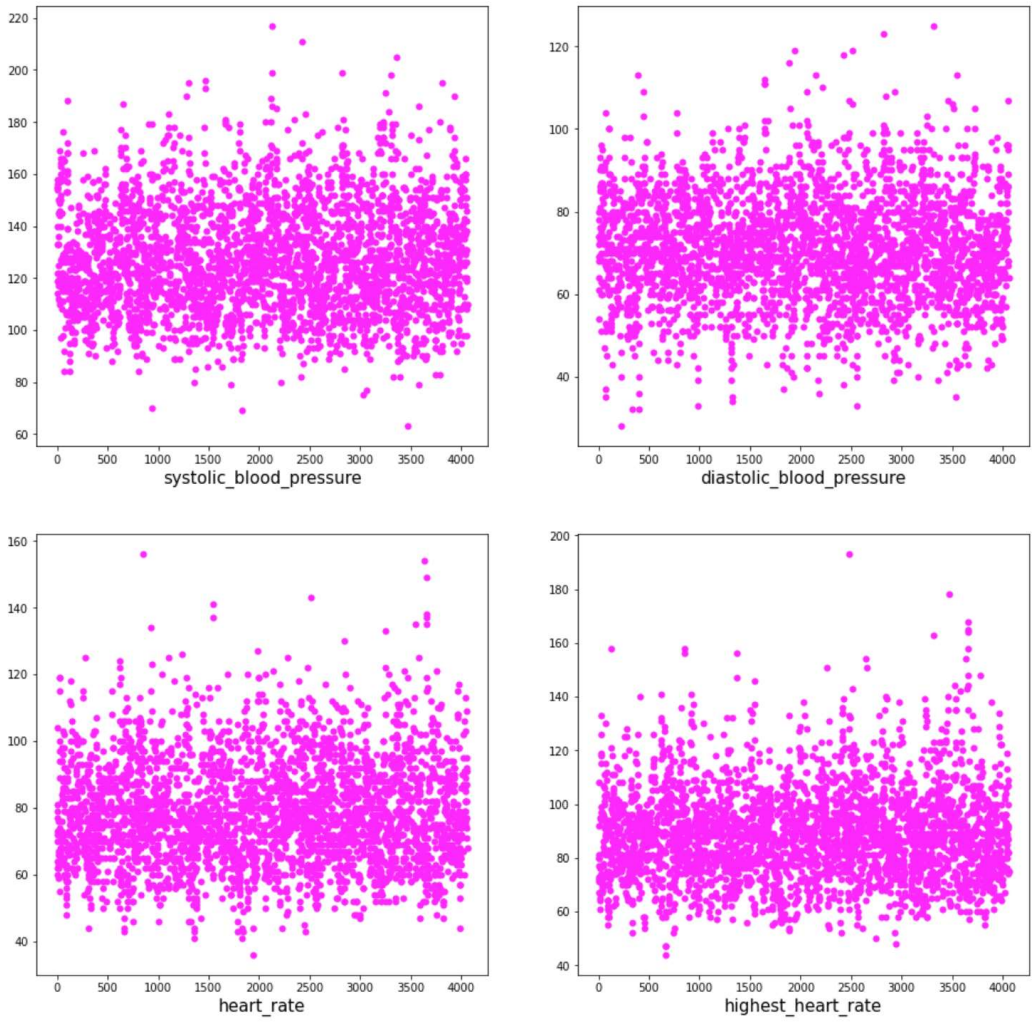

Fig.2. Display Lab 1 used data.

2. Plotting a non-parametric estimation of PDF in form of a histogram and using kernel density function (all chosen variables are continuous in our case).

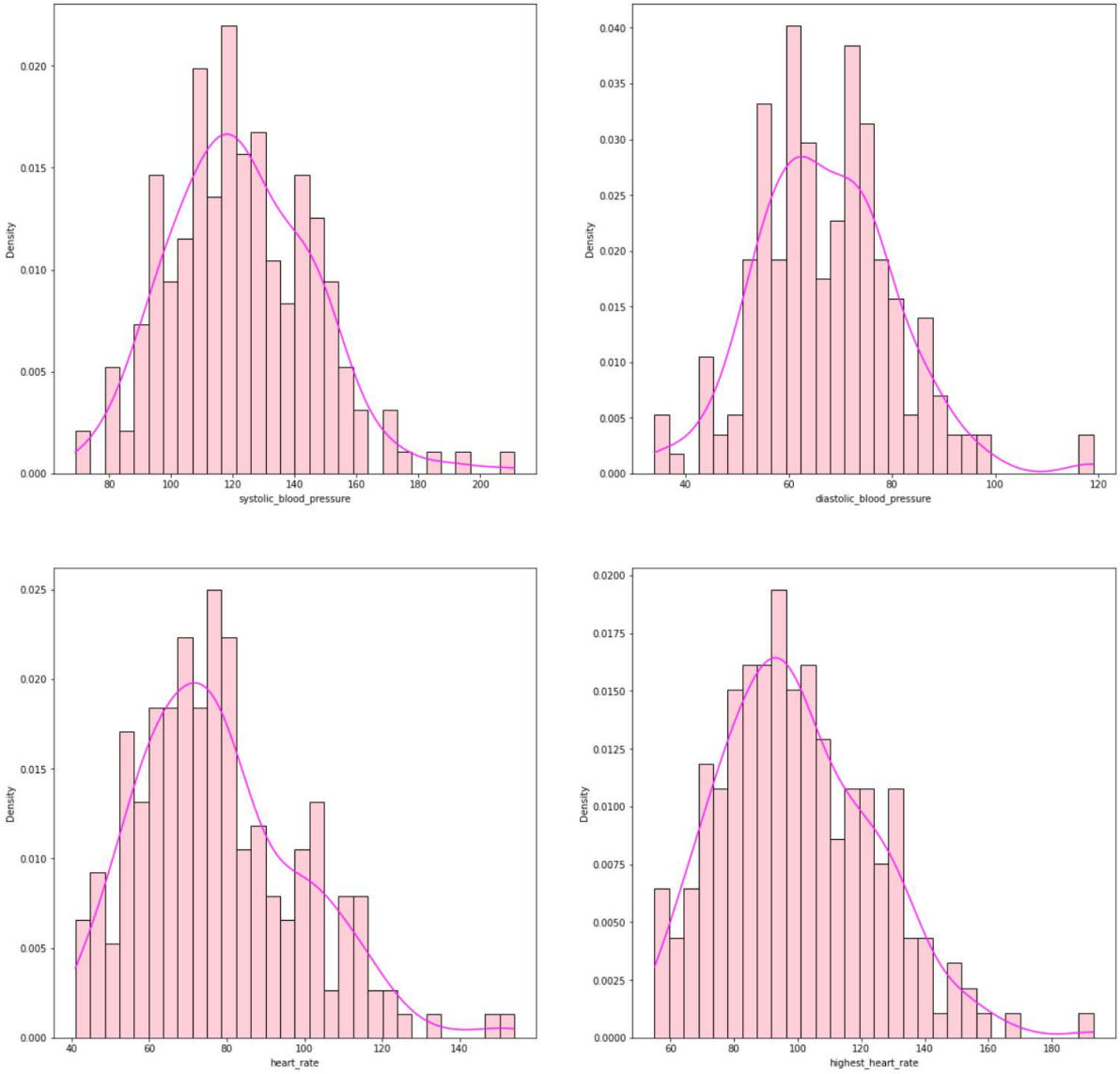

Fig.3. Histogram and KDE.

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

column name	m.expectation	median	variance	s.deviation
systolic_blood_pressure	122.70792079207921	121.0	538.7252105807596	23.210454768934614
diastolic_blood_pressure	67.20297029702971	66.0	182.33173242697404	13.503026787612251
heart_rate	78.10396039603961	76.0	439.70555637653314	20.96915726433786
highest_heart_rate	99.86633663366337	97.0	579.4397566622334	24.071554928218355

Fig.4. Statistics estimation.

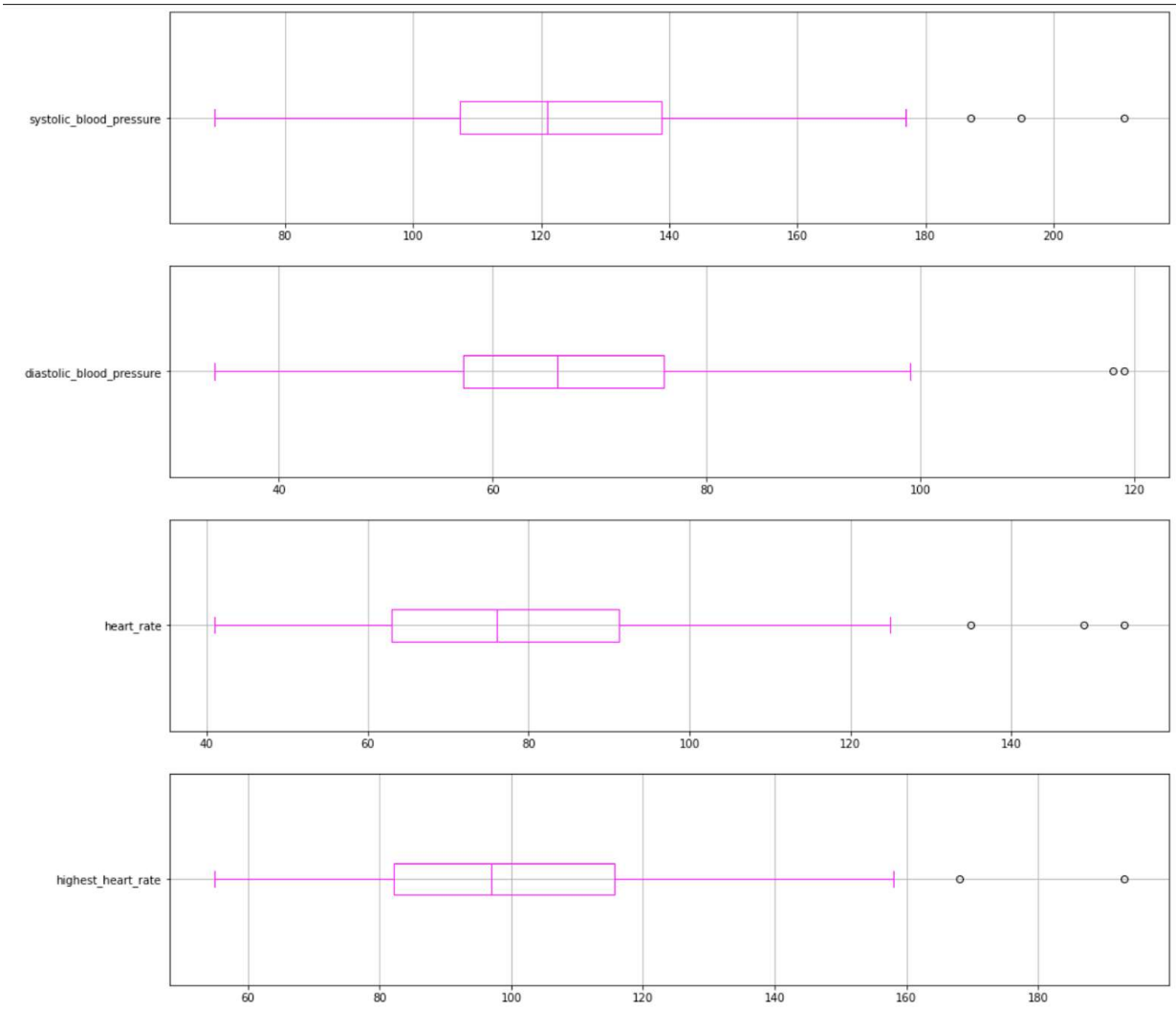

Fig.5. Variables' boxplot visualization.

Boxplots for all the variables indicate an observable number of outliers, as can be shown. But, as this number represents only roughly 1/1000 of all values, it has little impact on estimating the distribution parameters.

4. Selection of theoretical distributions that best reflect empirical data. and ***5. Estimation of random variable distribution parameters using maximum likelihood technique and least squares methods.***

column name	MLE	LSE
systolic_blood_pressure	(0.14367692232498774, -37.4539170111581, 158.51667565811556)	[7.32734315e-02 -1.95217371e+02 3.16269286e+
diastolic_blood_pressure	(0.11004879760245606, -54.37371862844321, 120.84161025382616)	[0.15543274 -16.23388392 82.87744334]
heart_rate	(20.719495086626452, 11.988636066170837, 3.175153042848409)	[205.7527838 -131.836868]
highest_heart_rate	(14.84931926041448, 33.74881597903993, 4.452561785696771)	[298.12621506 -201.64738218]

Fig.6. Parametric representation of MLE and LSE.

The distribution parameters estimated using the MLE and LSE methods are displayed in the table below. Scipy.stats distribution fitting functions were utilized for MLE of parameters. The minimize function from scipy.optimize was utilized for the least squares method. For the optimization task, histograms were used to obtain the X and Y data. For the highest heart rate and systolic blood pressure variables, lognormal distribution was chosen, as seen in the image below. Chi-squared one produced the best results when applied to heart rate and diastolic blood pressure.

Results from the Least Squares method are represented by the blue line on the image. The red line represents the PDF with parameters that were estimated by MLE.

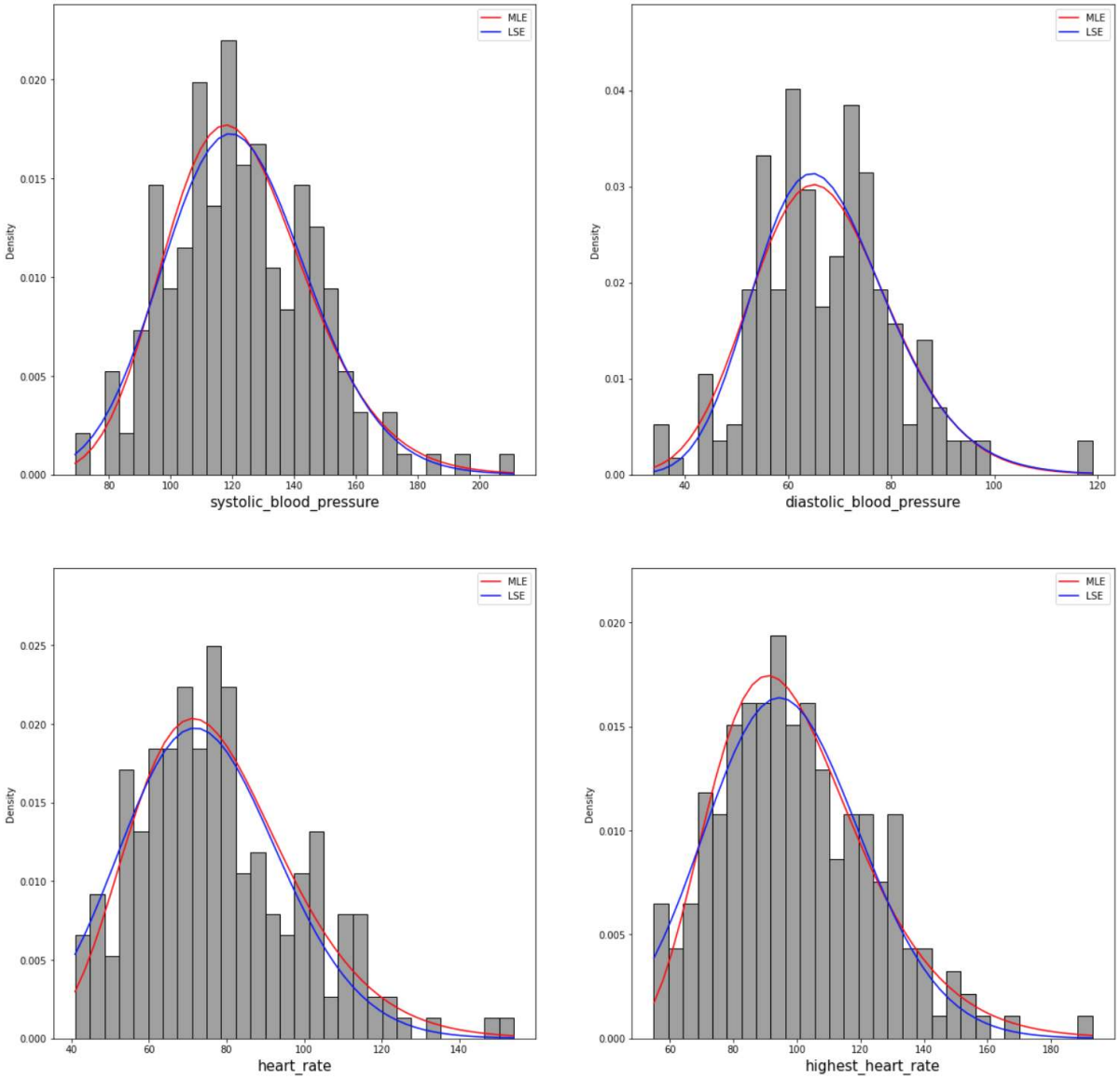

Fig.7. Hist MLE and LSE visualization.

6. Validation of empirical and theoretical distributions using quantile biplots.

This plot demonstrates that the tails of all QQ plots are generally good. At the same time, the lower percentiles of the heart rate distribution show observable variations. This is a result of the relevant histogram's relatively crisp shape.

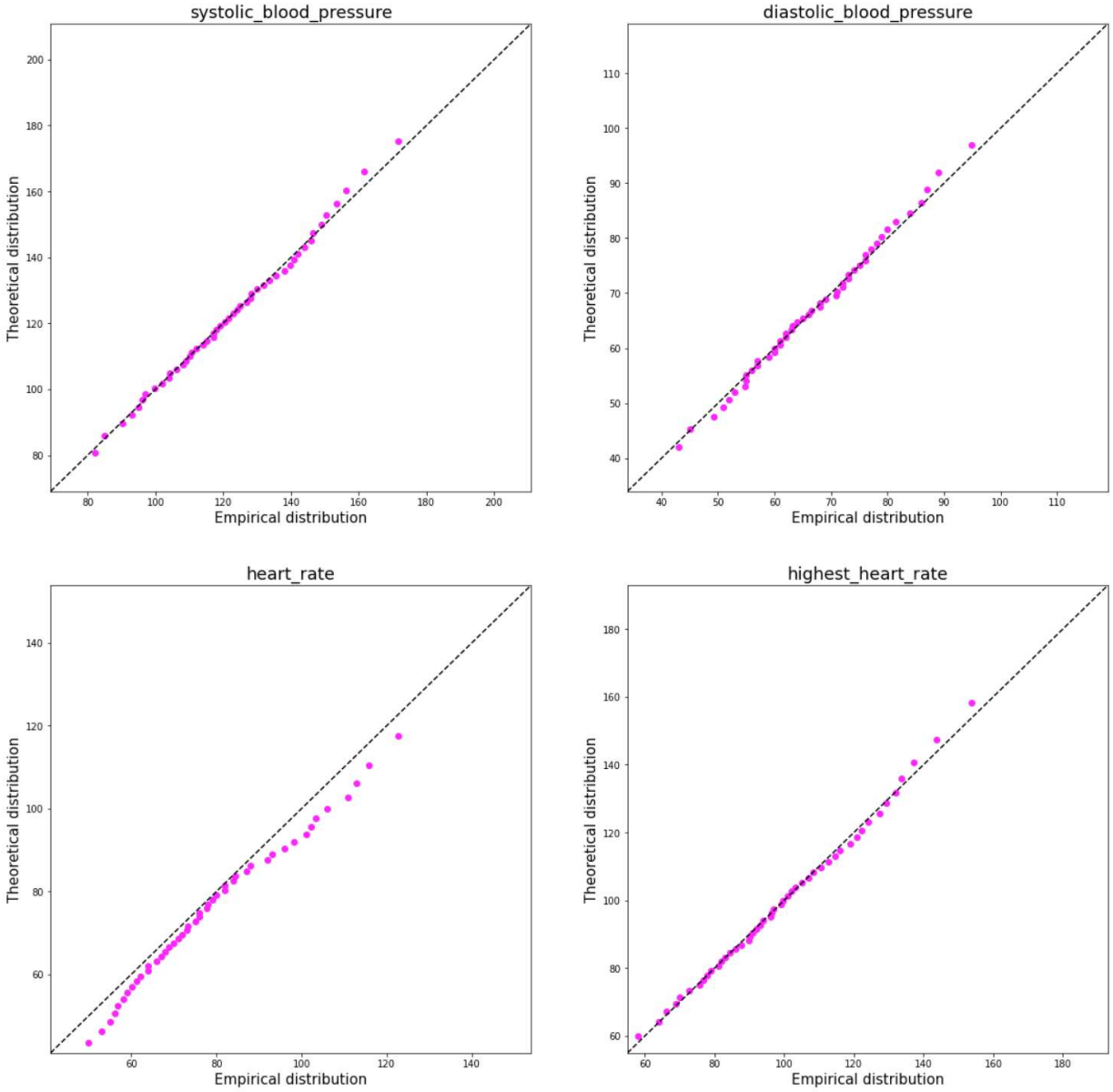

Fig.8. QQ biplots.

7. Statistical tests (2 at least).

Both the Kstest and CramerVonMises functions were utilized for statistical tests. Both MLE and LSE calculated parameters are used for these. The table below displays the obtained results. The acquired results argue against the claim that the theoretical distribution well explained the facts. The hypothesis can only be accepted with a good level of significance for heart rate.

column name	type	Kstest	CramerVonMises
systolic_blood_pressure	pvalue	0.9659741725492194	0.9697951478426843
diastolic_blood_pressure	pvalue	0.8549787045786086	0.8689697555174076
heart_rate	pvalue	0.13090688919589533	0.05677332325000928
highest_heart_rate	pvalue	0.4574990258402237	0.23197959751588315

Fig.9. Statistical tests.

Source code:

- The full repository with all the labs:
  https://github.com/RazinAleksandr/M-M-MSA-ITMO
- The repo with Datasets and additional used Data info:
  https://github.com/RazinAleksandr/M-M-MSA-ITMO/tree/main/Datasets
- The Lab1 ipynb file:
  https://github.com/RazinAleksandr/M-M-MSA-ITMO/tree/main/Lab_1/lab_1.ipynb

Furthermore, you can find README file with links for every lab folder on the main GitHub repository.