AUTOMATED APPLICATION FOR LARGE SAMPLE PROCESSING

Gusev S., Gudkov A., Sharonava A.
Department of Information Technologies Automated Systems,
Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: {st.al.gusev, gudkov.bsuir}@gmail.com, seizv@bsuir.by

The statement of the Central Limit Theorem is modeled on the example of large samples of the Laplace and the beta distributions using Python toolkit and the advanced capabilities of Jupyter Notebook.

Introduction

The law of Large Numbers and the Central Limit Theorem (CLT) are a generalization of the thought processes of mankind over the past two centuries [1]. Consequently, the authors decided to consider in detail the CLT which sometimes it is not always possible to find its graphical representation. In this regard, authors conducted a study to verify the following statement [2]: if there is a random variable (RV) X from almost any distribution, and samples of volume N are randomly formed from this distribution, then the distribution of sample averages can be approximated by a normal distribution with an average value that coincides with the expected value of the outcome population.

I. Toolkit

The task is to model the distribution of the RV's sample mean X at different sample volumes and estimate its approximation with a normal curve. To conduct the experiment authors chose Laplace and beta distributions from which the samples will be randomly formed [3]. The formation of samples, the calculation of their averages, the construction of graphs and histograms is carried out using the *Python* library toolkit: *scipy* module of statistical functions *scipy.stats*, *numpy*, *matplotlib*.

II. Laplace distribution

Let us first consider the Laplace distribution of a continuous RV X, the expected value and variance of which is calculated as follows:

$$E[X] = \lambda. \tag{1}$$

$$D[X] = 2 * b^2. \tag{2}$$

Here λ is a location parameter and $b > 0$ is a scale parameter. In our case $\lambda = 0$, $b = 50$. Using the statistical functions module of the *scipy* library, an instance of the *laplace_gen* class is created with parameters corresponding to the above. It's called *laplace_rv* and used for the entire study regarding the Laplace distribution.From this distribution, we will select 100 pseudo-random values.This is easy to do by calling the *rvs()* method with the sample size parameter. Also calling *plot()* function(from *matplotlib*) and *pdf()* method (from *scipy.stats*) we build Theoretical Probability Density Function (TPDF) and compare the obtained sample results in the form of a histogram with the TPDF graph of which corresponds to the blue line in Figure 1.

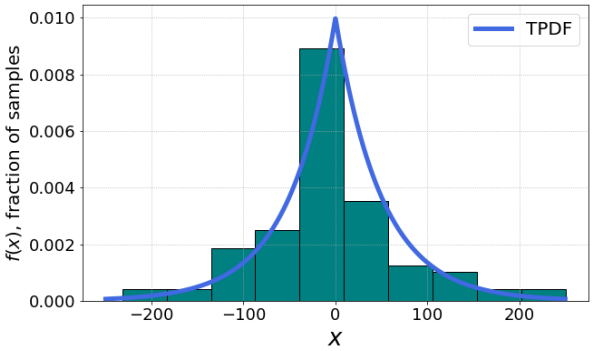

Рис. 1 – The Laplace distribution with sample volume of 100

Next, and this is the most important thing. For three or more values of N, 1000 samples of volume N are generated, the arithmetic mean is calculated for each sample. A histogram of the obtained sample means is constructed and a density graph of the corresponding normal distribution is superimposed on top of it with parameters:

$$\mu = E[X]. \tag{3}$$

$$\sigma = \sqrt{\frac{D(X)}{N}}. \tag{4}$$

Authors implemented the *buildHistNormCurve(ax, N, hCol)* function for generating a normal distribution and visualizing histograms with normal curve according to the sample size parameter N [4].

```
def buildHistNormCurve(ax, N, col):
    # list of sample means
    los = np.array([np.mean(l_rv.rvs(N)) for i in range(1000)])
    # normal distribution
    norm_rv = sts.norm(expected_value, (variance/N)**0.5)
    # normal curve construction
    x = np.linspace((-250 / N**0.4), (250 / N**0.4), 1500)
    pdf = norm_rv.pdf(x)
    ax.plot(x, pdf, lw=5, color='dodgerblue')
    # histogram
    ax.hist(los, bins=15, density=True, edgecolor='0', color=col)
    ax.grid(ls='dotted')
    # undername
    ax.set_xlabel('TPDF and Hist for N = ' + str(N), fontsize=14)
```

Рис. 2 – Build histogram with normal curve function

Let's call it 4 times with the following values of sample volume N: 3, 10, 50, 500. We get the results shown in figure 3.

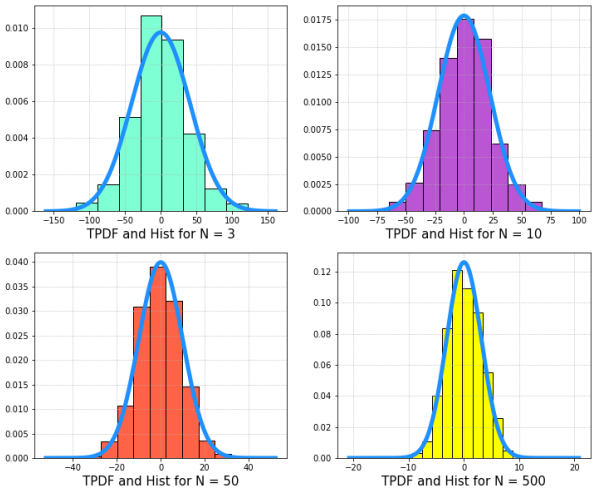

Рис. 3 – The result of CLT on the Laplace

III. Beta distribution

Also consider the beta distribution with the following numerical characteristics:

$$E[X] = \frac{\alpha}{\alpha + \beta}. \tag{5}$$

$$D[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{6}$$

Where $\alpha > 0$, $\beta > 0$. The appearance of the TPDF graph can vary greatly depending on the above parameters: from similar to an exponent and a parabola to an unusual curve, which we use, since it differs as much as possible from a normal curve. Let's define this distribution with $\alpha = \beta = 0.7$. In the case of a beta distribution, an instance of the *beta_gen* class is created. Then the TPDF and the histogram of the sample volume 100 takes the form as in figure 4.

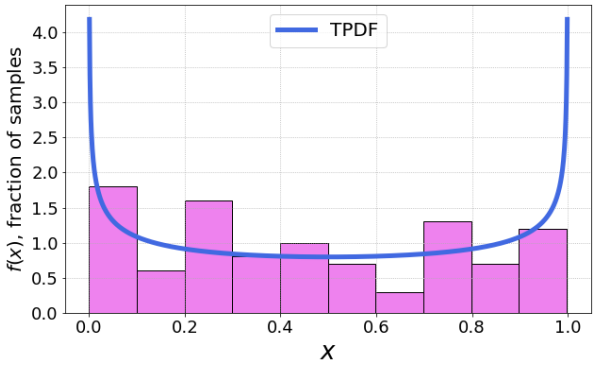

Рис. 4 – The beta distribution with sample volume of 100

Having done the same actions with the samples as the previous paragraph, we obtain the following distributions of sample means.

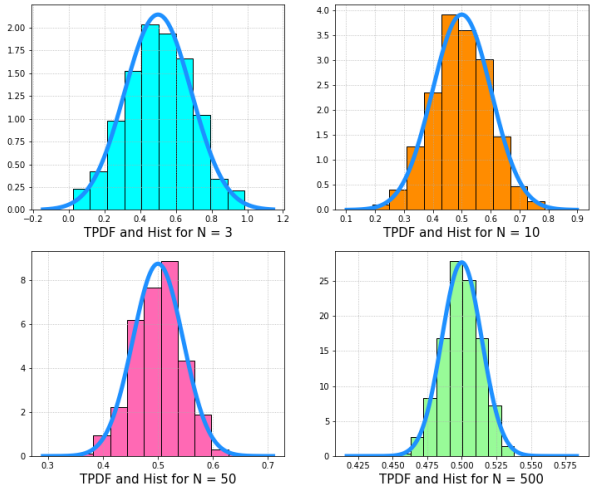

Рис. 5 – The result of CLT on the beta

If to compare the beta and the Laplace distributions by criteria of graphical representations (figures 1, 4) and formal ones (formulas 7, 8) they exactly differ.

$$f(x) = \frac{1}{2b}\exp\left(\frac{-|x - \lambda|}{b}\right). \tag{7}$$

$$f(x) = \frac{1}{B(\alpha, \beta)}x^{\alpha-1}(1 - x)^{\beta-1}. \tag{8}$$

But it doesn't matter how they're distant from the Gaussian, because the CLT gives an output distribution close to normal [5]. Moreover, the larger the sample size, the more accurate the approximation.

Conclusion

In accordance with the graphical representation of the results, the following pattern is well traced: with an increase in the sample size, the degree of approximation of the distribution of sample averages with a normal distribution also increases and there is a concentration of pseudo-random variables around the mathematical expectation of the initial distribution, which justifies the statement of the CLT, therefore principle of increasing entropy from a statistical point of view has led us to a fundamental conclusion: all closed macrosystems tend to move from less probable to more probable states.

References

1. Hawking, S. The Theory of Everything / S. Hawking. – Phoenix Books and Audio, 2005. – 148 p.
2. Вентцель, Е. В. Теория вероятностей / Е. В. Вентцель. – Москва: «Высшая школа», 2006. – 578 с.
3. Murphy, K. P. Machine learning: a probabilistic perspective / K. P. Murphy. – The MIT Press Cambridge, Massachusetts London, England, 2012. – 1098 p.
4. Chung, K. L. A Course in Probability Theory / K. L. Chung. – Academic Press, 2001. – 419 p.
5. Introduction to Probability and Statistics for Engineers [Electronic resource] / A. Shervine. – Stanford University, California, 2018. – Mode of access: https://stanford.edu/ shervine/teaching/cme-106. – Date of access: 22.02.2022.