Determination of probability distribution of customer input at post office

Silvia Ďutková[1], Dominika Hoštáková[1], Tomáš Mišík[1], Iwona Rybicka[2]

[1]Department of Communications, Faculty of Operation and Economics of Transport and Communications, University of Zilina, Univerzitná 8215/1, 010 26 Zilina, Slovakia
[2] Faculty of Mechanical Engineering, Lublin University of Technology, Lublin, 20-618, Poland

Abstract If we want to analyze a real system with a large number of input data, it is very convenient to determine a probability distribution that best fits to given input data. There are many statistical methods to determine the correct probability distribution and one of them is Chi-Square Goodness of Fit Test. This statistical test can be also used to find out a probability distribution of time intervals between arrivals of customers at post office. Intervals between arrivals of customers occur in continuous time and therefore we consider continuous probable distributions.

Keywords probability distribution, customer input, Chi-Square Goodness of Fit Test

JEL L87, L97

1. Introduction

In real systems such as queuing systems at post offices are based on random events. A system is a set of elements that are arranged in a certain way. Models of systems that are affected by random events show random variables of different form. The result of random event is a random variable [10, 12]. These random variables acquire different values and according to the type of these values we divide random variables to discrete and continuous random variables. Discrete random variables are usually integer values. Continuous random variables are values from closed or non-closed interval.

When we examine a particular system, we work with a number of data that represent the values of a random variable. In this case, it is advantageous to determine laws of probability that are attached to the given data. One of them is a probability distribution that describes the probability of the random variable in each value. In other words, probability distribution is the probability of occurrence of each outcome and in the context of queuing system at post office the outcome represents the event - the customer's arrival at the post office.

2. Background

The development of probability theory had a significant advance at the beginning of the 18th century with a predominantly normal distribution. The rapid development of probability theory probably began with DeMoivre's Dootrine of Chances (1713) and continued with Laplace´s and Gauss´s studies at the beginning of the 19th century and even more increased the dominance of normal distribution in statistics. The development of the exponential distribution came later. In 1931 T. Kondo in devoted his article in Biometrika to exponential distribution and Pearson's type X curve. In 1937 Sukhatme for the first time mentioned the idea that exponential distribution may be an alternative to normal distribution in the cases where the form of variation in the population is known and is not normal. In the 19th century Rényi, Epstein and Sobel made a significant contribution to the development of the exponential distribution. Also, very important was the paper by W. Weibull in 1951 in which he examined the expansion of the exponential distribution which now has his name. The first characterization of the exponential distribution was elaborated by Ghurey (1960) and Teicher (1961) which modified the characterization of normal distribution to the exponential distribution. In the main studies of exponential distribution began in the later years when the bases of statistics were basically built. [7]

In 1900, Pearson introduced Chi-Square Goodness of Fit Test that is universally applicable to determine the probability distribution of a given random variable. Pearson found that for a certain amount of data is a distribution approximately chi square with k -1 degrees of freedom. [3] The point of the test that number of classes are fixed, and test is asymptotically chi-square distributed. [2]

3. Continuous distribution

With respect to this kind of system, which is based on events over time, we consider continuous distributions. A continuous random variable is a random variable with a set of possible values that is infinite and uncountable.

3.1. Uniform distribution

Uniform distribution is defined by two parameters, a is the minimum and b is the maximum. The probability density of the uniform distribution from the interval (a, b) is: [1] [4]

$$f(x)=\frac{1}{b-a}, \qquad x\in(a,b) \tag{1}$$

$$= 0, \qquad \text{otherwise} \tag{2}$$

Uniform distribution R (a,b) has distribution function:

$$F(x)\begin{cases} 0, & x<a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ 1, & x\ge b \end{cases} \tag{3}$$

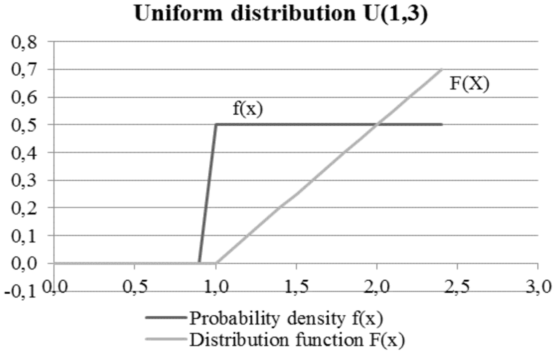

Figure 1. Continuous uniform distribution

3.2. Normal distribution

Regarding to normal distribution random errors are often mentioned as measurement errors caused by a large number of unknown and mutually independent causes. Probability density of normal distribution is given by the following formula: [1] [3]

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp-\frac{(x-\mu)^2}{2\sigma^2}, \qquad \begin{matrix}-\infty<x<\infty,\\ -\infty<\mu<\infty,\\ \sigma>0\end{matrix} \tag{4}$$

Normal distribution N (μ,σ) has distribution function:

$$F(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt, \qquad -\infty<x<\infty \tag{5}$$

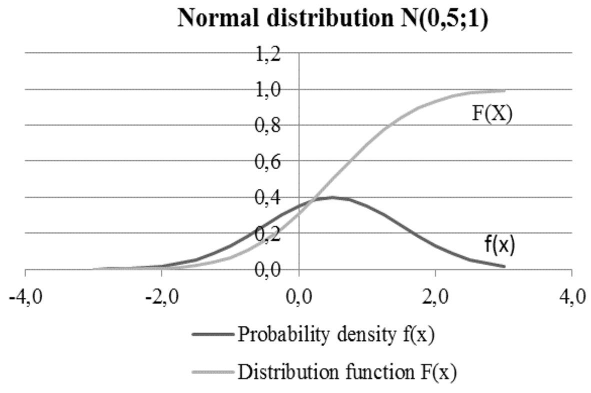

Figure 2. Continuous normal distribution

3.3. Exponential distribution

Exponential distribution reflects the time between randomly occurring events. Probability density of exponential distribution is given by the following formula: [1] [4]

$$f(x)=\lambda e^{-\lambda x}, \qquad x>0 \tag{6}$$

$$=0, \qquad x\le 0 \tag{7}$$

Distribution function of exponential distribution Exp (λ) is following:

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases} \tag{8}$$

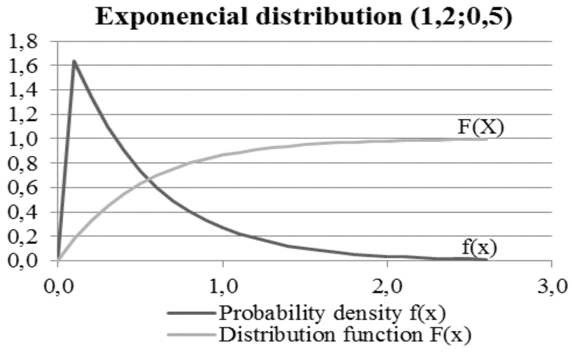

Figure 1. Continuous exponential distribution

3.4. Gama distribution

Probability density of exponential distribution is given by the following formula: [1] [4]

$$f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right), \qquad x>0 \tag{9}$$

$$=0, \qquad x\le 0 \tag{10}$$

While for parameters α and β apply α > 0, β > 0. If the parameter α natural number than gama distribution is called Erlang distribution. The distribution function of the gamma distribution does not exist.

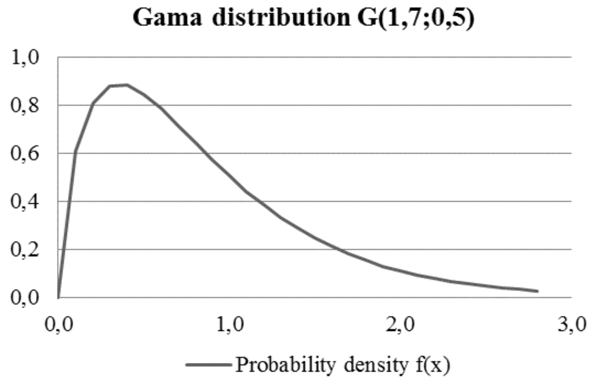

Figure 4. Continuous gama distribution

Working with data in this way is efficient if we want to perform the simulations where we generate data from a given probability distribution. There is a lot of different algorithms for this generation of random values. Using those algorithms, it is possible to transform random variables of uniform distribution from the interval (0,1) into the appropriate distribution.

It is essential to realize that random values are independent values of the uniform distribution from the interval (0,1). There are many mathematical generic formulas that can be used to analyse a particular system. Queuing theory is one of many mathematical sciences that offer such mathematical formulas where using it means obtaining results analytically by fitting specific parameters into the given formulas. If we decide to analyse queuing system this way, it is necessary to select the correct model, model that is the closest to the real model. The individual models offered by the queuing theory are characterized by the basic parameters. To specify the mathematical model of queuing system, it is necessary to specify: [10]

- customer input
- network of service lines,
- average service time,
- rules of entering and exiting into the system,
- other specific elements of the system.

The arrival of customers (customer input) is a stochastic process which probability distribution reflects the length of time intervals between customer arrivals. Customer input, which meets three properties:

- stationarity,
- unconsciousness,
- regularity,

we call the elementary flow. The flow is stationary if the probability of the arrival of the customer does not depend on the particular time placed on the numerical axis. The property unconsciousness is fulfilled if the events occur independently or respectively. It means that customer enters the service system independently of other customers. The regularity of customer input is based on principle there are not two events happening at the same time, we always find a small-time interval in which only one customer enters the system.

4. Objective and methodology

The objective of this paper is to determine the probability distribution of measured data. The probability distribution was determined to examine the random variable that is in our case the customer's arrival at the post office. Intensity of customer arrivals is one the parameter of queuing system at post office. In the order determine the probability distribution of variable and to create model we used Chi-Square Goodness of Fit Test as a tool of inductive statistics. This method allows us to determine the probability distribution that fits, and work predict further behaviour of system.

To determine the quantitative side of the system we used the empirical method such a measurement. The object of the measurements were time intervals between the arrivals of the customers at the Bytča Post office and the measurement was done using timers directly at Bytča Post Office during different part of opening hours of the post office. The basic statistical set is potentially infinite. The required standard deviation is ± 0,05 and the required confidence level is 95%. For the calculation of the sample, we used a relationship for calculating the minimum sample:

$$\sigma = \sqrt{p \cdot (1 - p)} \tag{11}$$

$$\sigma = \sqrt{0{,}5 \cdot (1 - 0{,}5)}$$

$$n \geq \frac{t^2_{1-\frac{\alpha}{2}} \cdot \sigma}{\Delta^2} \tag{12}$$

$$n \geq \frac{3{,}84 \cdot 0{,}25}{0{,}0025}$$

$$n \geq 384$$

where $t_{1-\alpha/2}$ is the critical value determined from the tables, σ is variance calculated from the standard deviation, p is variability of the base file and Δ is maximum allowable error range.

The measured values are divided into intervals. To determine the number of intervals and their length, we used the formulas from statistics. Determining the number of classes:

$$k = 1 + 3{,}3 \log n = 1 + 3{,}3 \log 384 = 10 \ .$$

Calculate the interval length:

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{4{,}6 - 0{,}1}{10} = 0{,}5 \text{ min} \ .$$

For graphical representation of the measured data, we used column graph. We also used the indicative statistics tool. Inductive statistics are concerned with statistical hypothesis testing. Testing is based on verifying the null hypotheses versus alternative hypothesis. Chi-Square Goodness of Fit Test is appropriate for determination of probability distribution. We used this test to verify the correspondence of measured data with exponential distribution.

5. Results

Since the customer´s requests handling system of post office mirror a queuing system with two basic input parameters the average interval between customer arrivals λ and the average service time *1/μ*, it was necessary to obtain customer input data. Intervals between customer arrivals at post office are defined in continuous time. Customer arrival process represents stochastic process, meaning that each customer's arrival is random, and no rule is attached to it. In the order to examine the properties of the system at post Office in Bytča we made 7 measurements of customer input. After that, we divided the measured data into interval classes as you can see in the table below. [8]

Table 1. Measured data divided to time intervals

Class *i*	Class interval	Central of interval	Absolute frequency	Relative frequency in %	Cumulative absolute frequency	Cumulative relative frequency in %
1	(0;0,5>	0,25	1100	49	1100	49
2	(0,5;1>	0,75	558	25	1658	74
3	(1;1,5>	1,25	267	12	1925	86
4	(1,5;2>	1,75	134	6	2059	92
5	(2;2,5>	2,25	87	4	2146	96
6	(2,5;3>	2,75	44	2	2190	98
7	(3;3,5>	3,25	26	1	2216	99
8	(3,5;∞>	3,75	12	1	2228	100

In order to create a system model approaching the real system, it is necessary to find out what probability distribution belongs to the measured data. There are many ways and tests for verifying the probability distribution applied in practice. We chose Chi-Square Goodness of Fit Test that verifies if empirical distribution is statistically identical to any of the theoretical probability distribution and this test is generally applicable to discrete and continuous distributions with a sufficient amount of data. In order to determine what probability distribution is could be considered we plotted the measured data into a graph. [8]

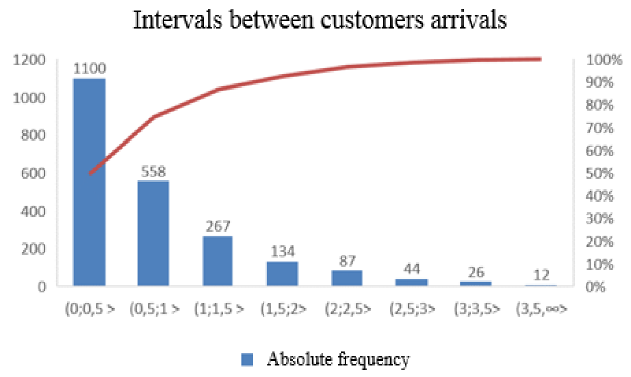

Figure 5. Intervals between customers arrivals

Figure 5. showed us that it could be potentially an exponential or Erlang distribution. Intervals between customer arrivals generally behave according exponential distribution in systems similar queuing system at Post Office [11]. To prove or disprove hypotheses about exponential distribution we decided to verify if measured data fits to exponential distribution: [5] [6]

- Null hypothesis H_0 = Intervals between arrival of customer is modeled by exponential distribution.
- Alternative hypothesis H_1 = Intervals between arrival of customer is not modeled by exponential distribution.

Level of significance reflect probability that we reject the true hypothesis. In general, this probability must be low and therefore we have chosen α = 0,05.

A goal of the Chi-Square Goodness of Fit Test is to compare the calculated test criterion with the critical value that can be found in the table Chi-Square distribution. Calculation of the test criterion is given by the mathematical relationship:

$$T = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} \tag{13}$$

where p_i represents the probabilities of individual class intervals. Those probabilities can be calculated using the following formula:

$$p_i = F(b_i) - F(a_i) \tag{14}$$

$$p_i = 1 - e^{-\lambda b_i} - \left(1 - e^{\lambda a_i}\right) \tag{15}$$

where *α* and *β* are class interval boundaries, and parameter λ is $1/\bar{x}$ average customer flow. In the table below, we can see probability classes with test criteria values for each class interval. There is also probability condition says that probability of class can not be small than value *5/n* otherwise we have to merge interval classes until condition is not respected. For this testing the condition is following:

$$p_i = \frac{5}{n} = \frac{5}{2228} = 0{,}0022 \tag{17}$$

Table 2. Calculation of test criteria

Class *i*	**(a_i,b_i>**	**x_i**	**n_i**	**x_i*n_i**	**p_i**	**T_i**
1	(0;0,5>	0,25	1100	275	0,4791	0,9915
2	(0,5;1>	0,75	558	418,5	0,2496	0,007
3	(1;1,5>	1,25	267	333,75	0,1300	1,7678
4	(1,5;2>	1,75	134	234,5	0,0677	1,8848
5	(2;2,5>	2,25	87	195,75	0,0353	0,9017
6	(2,5;3>	2,75	44	121	0,0184	0,2299
7	(3;3,5>	3,25	26	84,5	0,0096	1,0268
8	(3,5,∞>	3,75	12	45	0,0104	5,3922
∑	-	-	**2228**	**1708**	**1**	**12,2017**

After calculating the test criterion, we found the critical value χ^2-distribution of the distribution corresponding to the chosen significance level and the degree of freedom f:

$$\chi^2_{0,05}(8-1-1) = \chi^2_{0,05}(6) = 12{,}5916 \ .$$

If the test criterion value is < critical value, we do not reject the null hypothesis:

$$T < \chi^2_{0,05} \tag{18}$$

$$12{,}2017 < 12{,}5916$$

This inequality is true, meaning that we accept a null hypothesis - the intensity of the customer input at post office in Bytča corresponds to the exponential distribution.

6. Conclusions

Determining a probability distribution of measured data allows deeper analysis of data and to use the mathematical relationships that relate to particular probability distribution. If it´s known the probable distribution of data, it is also possible to simulate the data and predict its evolution based on the characteristics of the theoretical probability distribution. Probability distributions also play a very important role in generating random numbers in simulation models built on algorithms. Simulation models have a wide range of uses in many areas also in postal processes. The determination of exponential distribution that fits to time intervals between customer arrivals at Post Office Bytča can be useful in building a simulation model of queuing system of Post Office Bytča. In this research we were also able to calculate the average customer input, which is one of the basic parameters of queuing system.

REFERENCES

[1] Husek, R., Lauber, J.: Simulačné modely, ALFA, 1987, Praha, ISBN 04-326-87.

[2] Rui, H., Hengjian, C.: Consistency of Chi-Squared Test with Varying Number of Classes, J Syst Sci Complex, vol. 28, 439-450, 2015.

[3] Plackett, R.L.: “Karl Pearson and the Chi-squared Test”, International Statistical Review, 51, pp. 59-72., 1983.

[4] Achimská V.: Modelovanie systémov. Žilina, Žilinská univerzita v Žiline, 2011, ISBN 978-80-554-0450-9.

[5] Lyócsa, Š., Baumohl, E., Výrost, T.: Kvantitatívne metódy v ekonómii II., ElFA, 2013, ISBN 978-80-8086-210-7.

[6] Mrkvička, T., Petrášková, V.: Úvod do štatistiky, České Budejovice, 2006, ISBN 80-7040-894-4.

[7] Galambos, J., Kotz, S.: Charakterizations of Probability Distributions, Springer, 1978, New York, ISBN 3-540-08933.

[8] Ďutková, S.: Využitie teórie hromadnej oblsuhy na vybranej poštovej prevádzkarni, Diplomová práca, Žilinská univerzita v Žiline, 2017.

[9] Achimský K.: Simulácia činností poštovej prevádzky, Zborník, Žilina, 1990, str. 127-133.

[10] Madleňák, R., Madleňáková, L., Pavličko, M. Poštové prepravné siete, EDIS, Žilina, 214, ISBN 978-80-554-0903-0

[11] Madleňák, R., Madleňáková, L.: Comparison of regional postal transportation networks in Zilina region. In: Transport means 2015, Kaunas: Kaunas University of Technology, 2015. - S. 277-280.

[12] Matúšková, M., Madleňáková, L.: The impact of the electronic services to the universal postal services. In: Procedia Engineering, Vol. 178 (2017), s. 258-266.