Skewness and Kurtosis in Function of Selection of Network Traffic Distribution

Petar Čisar

Telekom Srbija, Subotica, Serbia, petarc@telekom.rs

Sanja Maravić Čisar

Subotica Tech – College of Applied Sciences, Subotica, Serbia, sanjam@vts.su.ac.rs

Abstract: The available literature is not completely certain what type(s) of probability distribution best models network traffic. Thus, for example, the uniform, Poisson, lognormal, Pareto and Rayleigh distributions were used in different applications. Statistical analysis presented in this paper aims to show how skewness and kurtosis of network traffic samples in a certain time interval may be criterions for selection of appropriate distribution type. The creation of histogram and probability distribution of network traffic samples is also discussed and demonstrated on a real case.

Keywords: skewness; kurtosis; network traffic; histogram; probability distribution

1 Introduction

Skewness characterizes the degree of asymmetry of a given distribution around its mean. If the distribution of the data are symmetric then skewness will be close to 0. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.

A measure of the standard error of skewness (SES) can roughly be estimated, according to [9], as the square root of 6/N, where N represents the number of samples. If the skewness is more than twice this amount, then it indicates that the distribution of the data is non-symmetric and it can be assumed that the distribution is significantly skewed. If the skewness is within the expected range of chance fluctuations in that statistic (i.e. ± SES), that would indicate a distribution with no significant skewness problem.

Kurtosis characterizes the relative peakedness or flatness of a distribution compared with the normal distribution. For normally distributed data the kurtosis is 0. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution. As with skewness, if the value of kurtosis is too big or too small, there is concern about the normality of the distribution. In this case, a rough formula for the standard error for kurtosis (SEK) is the square root of 24/N. Since the value of kurtosis falls within two standard errors (i.e. ± SEK), the data may be considered to meet the criteria for normality by this measure. These measures of skewness and kurtosis are one method of examining the distribution of the data. However, they are not definitive in concluding normality. What should also be examined is a graph (histogram) of the data; and further, one should consider performing other tests for normality such as the Shapiro-Wilk or the Kolmogorov-Smirnov test.

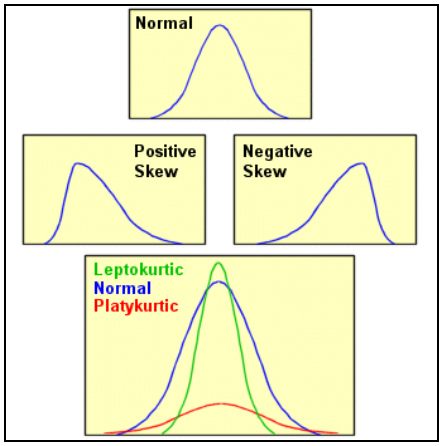


Figure 1
Illustration of skewness and kurtosis

Skewness is a network characteristic that can be successfully implemented in statistical anomaly detection algorithms, as is shown in [12]. When trying to distort the distribution parameters (mean, median, etc.), the attacker puts more weight into one of the tails of the probability density function, which leads to an asymmetric, skewed distribution. It is well-known that the standard characterizing parameters of a distribution are the mean (or median), the standard deviation, the kurtosis, and the skewness. It is reasonable to assume that there will be some empirical distribution of the skewness in the case when there is no attack is available, since most of the time, the system is not normally attacked. The statistical intrusion detection algorithm can compare the skewness of the sample to the expected value of the skewness in order to decide if an attack is taking place or not.

The authors have dealt with the topic of statistical intrusion detection in publications [1]-[8].

2 Different Types of Probability Distribution

Using the software package "Matlab", sequences of 40 random numbers with various types of probability distributions are generated (Figure 2 – function "Export"), as the way to simulate network traffic.

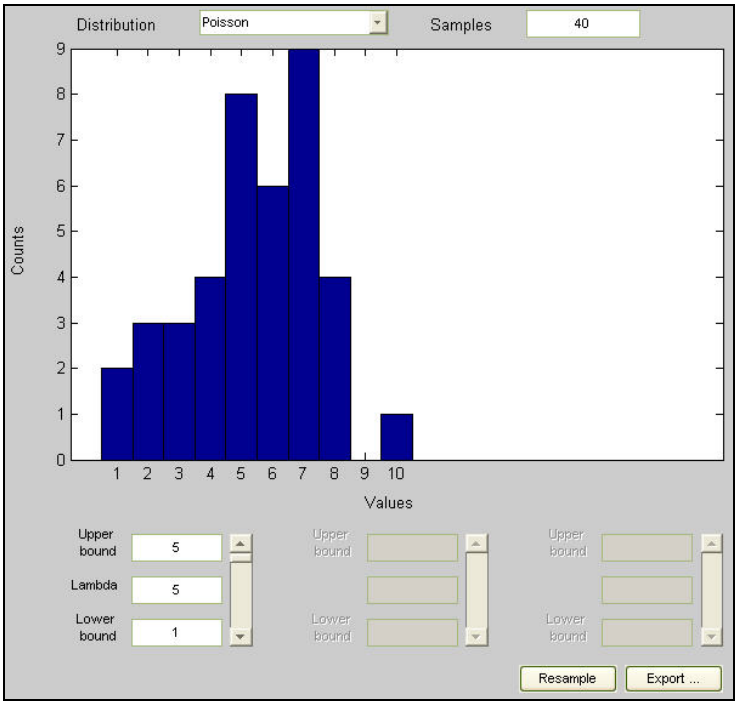
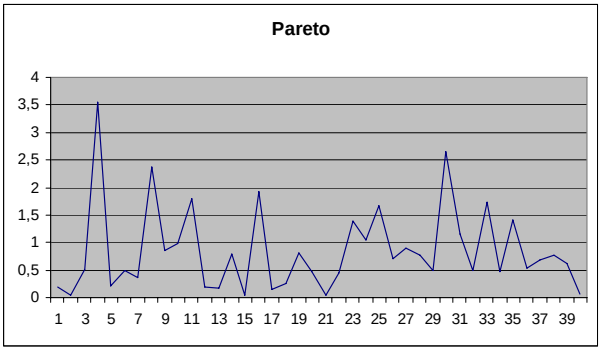
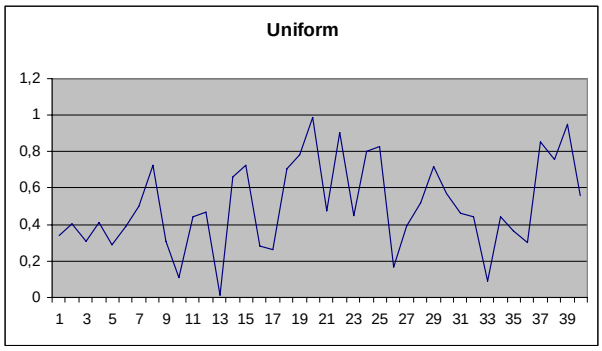
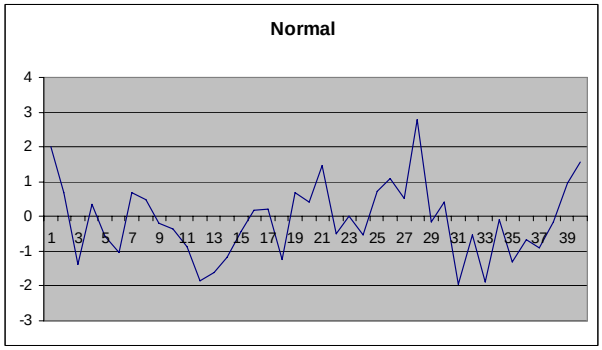


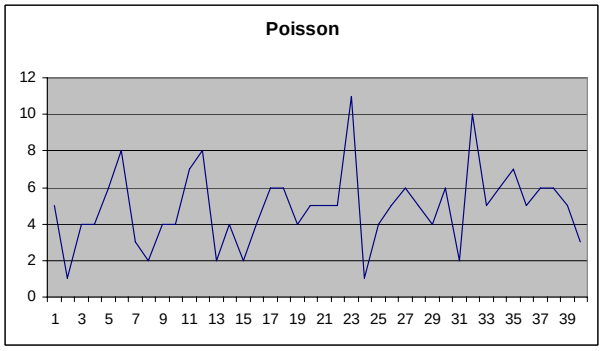
Figure 2
Random number generation (Poisson distribution)

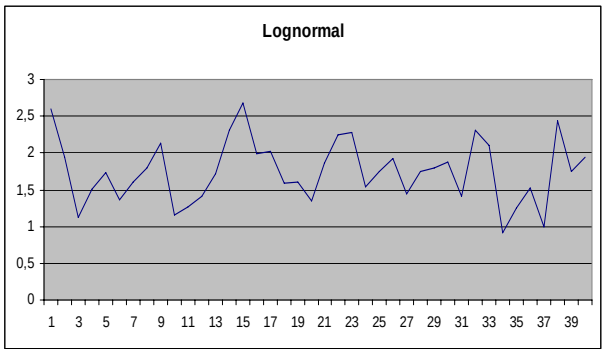
The following different types of distribution are examined: Pareto, normal, Poisson, lognormal, Rayleigh, chi-square and Weibull. Their graphical representation is shown in Figure 3.

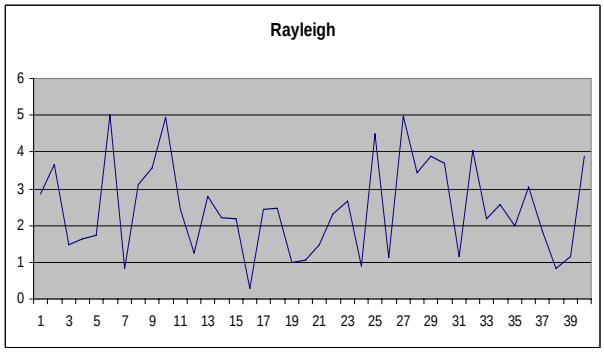


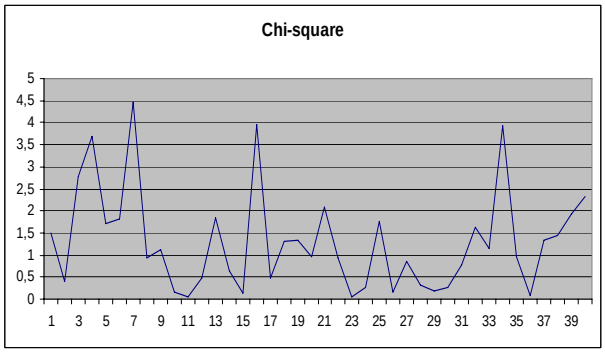

Uniform
1,2
1
0,8
0,6
0,4
0,2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39


Normal
4
3
2
1
0
-1
-2
-3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39


Poisson
12
10
8
6
4
2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39


Lognormal
3
2,5
2
1,5
1
0,5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39


Rayleigh
6
5
4
3
2
1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39


Chi-square
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39

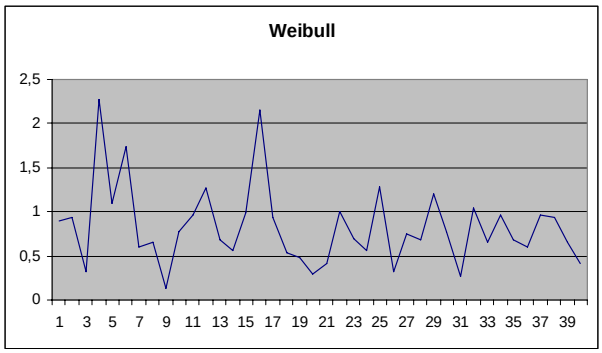


Figure 3
Examined types of distribution

For the purpose of the statistical comparison of generated curves, the authentic traffic samples y_t (local maxima) of a real ISP are also analyzed and graphically represented.

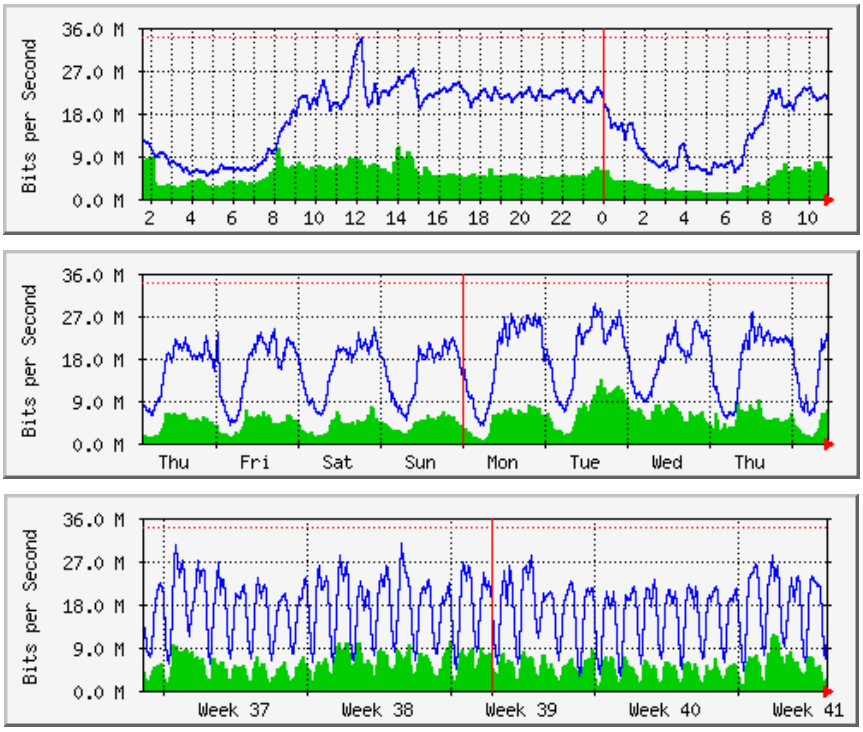


Figure 4
Network traffic curves

Table 1
Samples of network traffic

Sample	y_t (daily)	y_t (weekly)	y_t (monthly)
1	12	21	23
2	10.5	22.5	30
3	8.5	23	27
4	10.5	20	27
5	18	20.5	25
6	22	23.5	27
7	25.5	24	22
8	20	21	24
9	33.9	23	23
10	25	25	20
11	24	25.5	24.5
12	26.5	24.5	26.5
13	27.5	22	28
14	23	25.5	27
15	25	27	23
16	24	28	22.5
17	23	27	26.5
18	23	28	31
19	22	25.5	22.5
20	23	30	22.5
21	23	29	27
22	23	26.5	25
23	23	29	26
24	16	26.5	28
25	16	27.5	21
26	9	26	24
27	11.5	25	22
28	8.5	24	22
29	8.5	23.5	22
30	14	22	23
31	23	22.5	27
32	23	24	29
33	20	24	25
34	23	25	25
35	23	23	22

The numerical values of all samples are presented in the following table.

Table 2

Samples for various distributions

	Pareto	Uniform	Normal	Poisson	Lognormal	Rayleigh	Chi-square	Weibull	ISP (daily)
1	0.187	0.3418	1.9969	5	2.6035	2.8425	1.5095	0.8927	12
2	0.0449	0.4018	0.697	1	1.9416	3.6687	0.4015	0.9414	10.5
3	0.5183	0.3077	-1.3664	4	1.1205	1.489	2.7817	0.3218	8.5
4	3.5492	0.4116	0.363	4	1.5011	1.6232	3.6918	2.2777	10.5
5	0.2081	0.2859	-0.567	6	1.7366	1.7224	1.7237	1.1012	18
6	0.4941	0.3941	-1.0442	8	1.3615	5.012	1.8079	1.7357	22
7	0.3545	0.503	0.6971	3	1.6055	0.8327	4.4586	0.6054	25.5
8	2.3838	0.722	0.484	2	1.7944	3.1239	0.9332	0.6562	20
9	0.8559	0.3062	-0.1938	4	2.1348	3.5528	1.1275	0.131	33.9
10	0.9793	0.1122	-0.3781	4	1.1604	4.9411	0.1487	0.7701	25
11	1.7948	0.4433	-0.8864	7	1.2742	2.43	0.044	0.9571	24
12	0.1825	0.4668	-1.8402	8	1.4038	1.2626	0.4828	1.271	26.5
13	0.176	0.0147	-1.6282	2	1.7206	2.8012	1.8466	0.6853	27.5
14	0.7949	0.6641	-1.1738	4	2.3076	2.2195	0.6304	0.5569	23
15	0.0444	0.7241	-0.4154	2	2.6753	2.1782	0.1261	0.9892	25
16	1.9163	0.2816	0.1751	4	1.9844	0.294	3.9641	2.1489	24
17	0.1393	0.2618	0.2294	6	2.0198	2.4474	0.4758	0.9316	23
18	0.2621	0.7085	-1.2409	6	1.5909	2.4579	1.315	0.5318	23
19	0.8116	0.7839	0.7	4	1.6081	1.0003	1.332	0.4805	22
20	0.477	0.9862	0.4269	5	1.3498	1.0748	0.9661	0.2888	23
21	0.0495	0.4733	1.4548	5	1.8651	1.4749	2.0914	0.4108	23
22	0.4463	0.9028	-0.5102	5	2.2465	2.3204	0.947	1.0002	23
23	1.397	0.4511	-0.0067	11	2.2796	2.6605	0.044	0.6905	23
24	1.0421	0.8045	-0.5255	1	1.538	0.9073	0.26	0.5594	16
25	1.6721	0.8289	0.7177	4	1.7407	4.4974	1.771	1.2808	16
26	0.712	0.1663	1.0884	5	1.9303	1.1374	0.154	0.3163	9
27	0.8934	0.3939	0.5006	6	1.4374	4.9586	0.8429	0.7507	11.5
28	0.7689	0.5208	2.7718	5	1.7461	3.4222	0.3112	0.6776	8.5
29	0.4928	0.7181	-0.1603	4	1.8019	3.891	0.188	1.2044	8.5
30	2.6427	0.5692	0.4295	6	1.8782	3.6857	0.2644	0.7746	14
31	1.1575	0.4608	-1.9668	2	1.4134	1.1608	0.7791	0.2662	23
32	0.4969	0.4453	-0.546	10	2.3082	4.0349	1.6375	1.0455	23
33	1.7429	0.0877	-1.8884	5	2.1038	2.1679	1.1526	0.6499	20
34	0.4764	0.4435	-0.108	6	0.9098	2.5793	3.9352	0.9679	23
35	1.4026	0.3663	-1.3161	7	1.2547	1.9957	0.9714	0.6829	23
36	0.5321	0.3025	-0.6726	5	1.5198	3.0429	0.0734	0.5983	
37	0.6811	0.8518	-0.9024	6	0.9969	1.8625	1.3493	0.9609	
38	0.7662	0.7595	-0.1548	6	2.4398	0.8267	1.4533	0.9396	
39	0.6136	0.9498	0.9472	5	1.7478	1.1558	1.9199	0.6502	
40	0.0594	0.5579	1.5504	3	1.9379	3.8697	2.3274	0.421	

As emphasized in Chapter 1, if the skewness and kurtosis are within the expected ranges of chance fluctuations in that statistic (i.e. ± SES and ± SEK), this implies that the distribution has no significant skewness problem. The different distributions shown in the table above (N = 40) result in SEK = 1.55 and SES = 0.77. Also, for the ISP (N = 35) the values of SEK = 1.66 and SES = 0.83 were obtained. By calculating the kurtosis and skewness for each distribution, the values presented in the following table were obtained.

Table 3
Kurtosis and skewness for analyzed distributions

	Pareto	Uniform	Normal	Poisson	Lognormal	Rayleigh	Chi-square	Weibull	ISP
kurtosis	2.82	-0.64	0.11	1.09	-0.38	-0.69	1.08	2.96	-0.53
skewness	1.6	0.16	0.38	0.58	0.187	0.41	1.23	1.47	-0.44

In the table above, the values outside the calculated limits are marked in darker shading, indicating distribution types that have greater skewness and kurtosis than is allowed. In this sense, these distributions are not appropriate in applications that describe network traffic.

3 The Variations of Skewness and Kurtosis

Daily, weekly and monthly network traffic curves of the Internet user were analyzed in order to determine the extent of variations of skewness and kurtosis. Based on these curves, the appropriate descriptive statistics are calculated (Table 4).

Table 4
Descriptive statistics of network samples

	daily	weekly	monthly
Mean	19,75428571	24,68571429	24,85714286
Standard Error	1,090581087	0,430499863	0,455393185
Median	23	24,5	25
Mode	23	24	27
Standard Deviation	6,451964719	2,546871536	2,694142417
Sample Variance	41,62784874	6,486554622	7,258403361
Kurtosis	-0,526430851	-0,602769614	-0,614090489
Skewness	-0,441718333	0,172823583	0,343951523
Range	25,4	10	11
Minimum	8,5	20	20
Maximum	33,9	30	31
Sum	691,4	864	870
Count	35	35	35
Confidence Level (99.0%)	2,975535291	1,1745734	1,2424922

By analyzing the results obtained in the table above, it can be concluded that the values for skewness and kurtosis remain within allowable limits for all the time periods. Only in the case of skewness is a change of sign detected.

4 Histogram and Probability Distribution

According to [10], the purpose of a histogram is to graphically summarize the distribution of a univariate data set. The histogram graphically shows the following:

1 center of the data

2 spread (i.e. the scale) of the data

3 skewness of the data

4 presence of outliers

5 presence of multiple modes in the data

These features provide strong indications of the proper distributional model for the data.

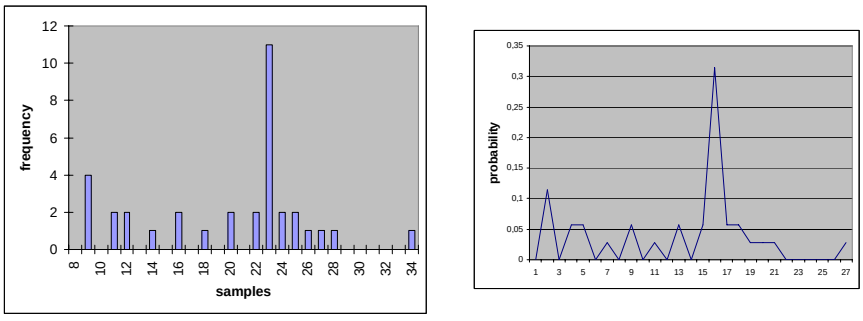
The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin is counted. That is

- vertical axis - frequency (i.e. counts for each bin)
- horizontal axis - response variable

The histogram of the frequency distribution can be converted to a probability distribution by dividing the tally in each group by the total number of data points to achieve the relative frequency [11]. In accordance with the aforementioned, the following graphical presentation related to network samples of ISP is obtained.

Based on Figure 5, it can be concluded that the most frequent samples approximately correspond to the mean of network traffic flow.

histogram **probability distribution**



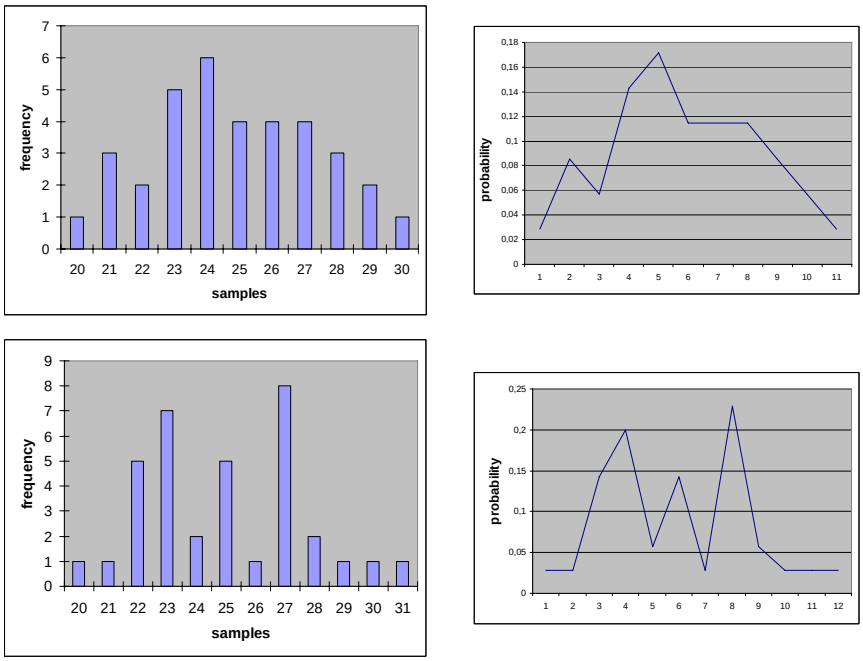


Figure 5
Histogram and probability distribution of network samples

Conclusions

Through analysis of the appropriate values, it can be concluded that the Pareto and Weibull distributions provide kurtosis and skewness that are significantly outside the permitted range. Looking at the skewness, it is necessary to note that the chi-square distribution gives values also higher than the allowed. Among the analyzed types of distribution, based on kurtosis and skewness as criteria for describing network traffic, the appropriate distributions are the uniform, the normal, the Poisson, the lognormal and the Rayleigh. It is also shown that the most frequent network sample in the histogram is approximately equal to the mean of the traffic flow.

In the next phase of research, the skewness and kurtosis of network traffic will be examined in terms of the reliability of network anomaly detection. The expectation is to try to find their application in the field of information systems security.

References

[1] Sorensen, S.: Competitive Overview of Statistical Anomaly Detection, White Paper, Juniper Networks, 2004

[2] Gong, F.: Deciphering Detection Techniques: Part II Anomaly-based Intrusion Detection, White Paper, McAfee Security, 2003

[3] SANS Intrusion Detection FAQ: Can You Explain Traffic Analysis and Anomaly Detection? Available at http://www.sans.org/resources/idfaq/ anomaly_detection.php

[4] Dulanović, N., Hinić, D., Simić, D.: An Intrusion Prevention System as a Proactive Security Mechanism in Network Infrastructure, YUJOR - Yugoslav Journal of Operations Research, Vol. 18, No. 1, pp. 109-122, 2008

[5] Lazarevic, A., Kumar, V., Srivastava, J.: Managing Cyber Threats: Issues, Approaches and Challenges, Chapter: A survey of Intrusion Detection techniques, Kluwer Academic Publishers, Boston, 2005

[6] Lončarić, S.: Osnove slučajnih procesa, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zavod za elektroničke sustave i obradbu informacija, Available at http://spus.zesoi.fer.hr

[7] Siris, V., Papagalou, F.: Application of Anomaly Detection Algorithms for Detecting SYN Flooding Attacks. Available at http://www.ist-scampi.org /publications/papers/siris-globecom2004.pdf

[8] CAIDA, the Cooperative Association for Internet Data Analysis: Inferring Internet Denial-of-Service Activity, University of California, San Diego, 2001

[9] Tabachnick, B. G., Fidell, L. S.: Using Multivariate Statistics (3rd ed.), New York: Harper Collins, 1996

[10] National Institute of Standards and Technology, Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook/pmc/pmc.htm

[11] NetMBA-Business Knowledge Center, Available at http://www.netmba.com/ statistics/histogram/

[12] Buttyan, L., Schaffer, P., Vajda, I.: Resilient Aggregation with Attack Detection in Sensor Networks, Second IEEE International Workshop on Sensor Networks and Systems for Pervasive Computing (PerSeNS), IEEE Computer Society Press, Pisa, Italy, 2006