USING THE METHODS OF PROBABILITY THEORY ANALYZING LOGS OF ELECTRONIC INFORMATION SYSTEMS

VALÓSZÍNŰSÉGELMÉLETI MÓDSZEREK ALKALMAZÁSA ELEKTRONIKUS INFORMÁCIÓS RENDSZEREK NAPLÓINAK ELEMZÉSE SORÁN

NYÁRI Norbert[1]

Abstract

Proper level of logging in various information systems and log analysis play a key role in achieving the desired level of security. The aim of the present study on one hand is to provide an overview of the known applications of mathematical statistical tools and methods in the field of information security, and, on the other hand, to demonstrate the applicability of probability theory in analysis of log files of IT systems through a specific example. Starting from the basic concepts of the topic, and the most important and relevant concepts related to logging, log analysis and statistics are discussed. After that comes a brief overview of the previous researches on the topic. Finally, a possible use of random variables and theoretical probability distributions in the analysis of logs of various IT systems is described. The raison d'être of the method is verified through a real-world example.

Keywords

probability theory, statistics, information security, electronic information system, log analysis

Absztrakt

A megfelelő naplózás a különféle információs rendszerekben és a naplóelemzés kulcsfontosságú szerepet játszik a megfelelő szintű biztonság elérésében. Jelen tanulmány célja egyrészt egy áttekintő képet adni arról, hogy az információbiztonság területén milyen alkalmazásai ismertek a matematikai statisztikai eszközöknek és módszereknek, másrészt egy konkrét példán keresztül bemutatni a valószínűség elmélet alkalmazhatóságát az informatikai rendszerek naplóállományainak elemzése során. A tématerület alapfogalmaitól elindulva tárgyalja a naplózással, naplóelemzéssel, statisztikával kapcsolatos legfontosabb és a téma szempontjából releváns fogalmakat, koncepciókat. Ezt követi a korábbi, a témához kapcsolódó kutatások rövid, áttekintő jellegű ismertetése. Végül pedig bemutatásra kerül a valószínűségi változók, elméleti valószínűségi eloszlások egy lehetséges felhasználása a különféle informatikai rendszerek naplóelemzésében. A módszer létjogosultsága egy, a való életből vett példán keresztül kerül igazolásra.

Kulcsszavak

valószínűségelmélet, statisztika, információbiztonság, elektronikus információs rendszer, naplóelemzés

[1] nyari.norbert@phd.uni-obuda.hu | https://orcid.org/0000-0003-0229-7584 | PhD student/doktorandusz | Óbudai Egyetem Biztonságtudományi Doktori Iskola

INTRODUCTION

An appropriate level of logging in various IT systems, including electronic information systems is a key factor in guaranteeing the security of these systems. The purpose of the systems and the security requirements placed on them must be taken to account when it comes to specifying the adequate level of logging. However, logging should not be seen only as goal to achieve, but rather as a tool in every day work, design, development, and operation of IT systems. It is absolutely necessary to use this tool in order to achieve and maintain a sufficient level of security, or even restoring security in the event of security incidents. Logging, and analyzing log data from a variety of perspectives, can serve many purposes like compliance, forensics, debugging, among other things.

The aim of the present study is to show how the methods of probability theory can be applied in log analysis in order to provide predictions of the occurrence of certain, sometimes costly, undesired events, and of course their negative consequences.

LITERATURE REVIEW

Let me start with a brief overview of the main concepts relevant to the topic of this study. Firstly, probability theory is a branch of mathematics dealing with the probability of random events based on the frequency of their occurrences. It stems from the work of Pascal, Fermat and Kolmogorov treating probability in an exact mathematical manner. [1] There are however other interpretations of probability such as Bayesian probability which highly relies on prior knowledge (personal experiences or beliefs etc.) of the phenomenon in question in the form of prior probability. [2] This approach however is out of the scope of this study.

One of the major concepts of probability theory is random variable. In formal mathematics a random variable is a measurable function that maps from the event space to the real numbers. Every random variable has a distribution. A random variable and its distribution can be discrete or continuous. The possible values of a discrete variable are countable, whereas the codomains of continuous variables are intervals of real numbers. [3]

Random variables can be described with many features like expected value, deviation, distribution etc. The expected value is the number around which the average of the observed values of the variable in question fluctuates. Deviation, as its name implies, characterizes the deviation of the values of a given variable from the expected value. [3] [1]

A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible values of a random variable. As I mentioned earlier, distributions can be also discrete or continuous. [3] Many distributions are important enough, theoretically or practically, to have their own name, including, but not limited to binomial distribution, Poisson distribution, exponential distribution. [3]

There is a strong connection between statistics and probability theory. The concepts of probability theory can be applied to series of data. I would like to highlight a particular use case: a theoretical distribution can be fitted to histogram of a series of data. [1] Fitting can be verified with a statistical hypothesis test. Hypothesis tests are used to verify a hypothesis based on observed, empirical data. Numerous methods exist for different purposes,

like testing for goodness of distribution fitting, homogeneity of two random variables, independence of two random variables. These tests can be performed with the Chi-squared test, and the Kolmogorov test. [3]

In the following paragraphs let me continue with computer logs and logging. Registering events and transactions occurring in IT systems into log files is the primary goal of logging. Events can be of many types such as user activities, errors, security events etc. [4] [5] The records in logs in themselves are only data though, the process of log analysis is used to extract valuable information from them.

Identification of major events, especial those that affect security in IT systems, can be achieved through an adequate logging policy and logging system. [5] Several different methodologies like Log Management (LM) and Security Information and Event Management (SIEM), provide guidelines helping IT professionals designing and implementing suitable logging. [5] Presenting these methodologies in depth would however go beyond the scope of the present study. I would like to stress that setting up a suitable logging facility is highly advisable for any enterprise scale IT system in order to provide sufficient data for subsequent log analysis.

As I already mentioned in the introduction, log analysis can serve many purposes like identifying policy violations, security incidents, fraudulence, and operational issues. Log analysis however can be a burden due to its repetitive and monotonous nature. No wonder it is considered boring by many system and security administrators. Providing the right tools for the staff performing log analysis is highly desirable. Occasionally rotating log analysis duties among staff members to prevent burnout can be a good idea as well. [6]

The basic techniques of log analysis are pattern recognition, normalization, classification and tagging, correlation analysis. [6]

Before presenting the subject of my research, I would like to give an overview of previous applications of statistical methods in the perspective of safety science.

In their study 'Mathematical and Statistical Opportunities in Cyber Security', Meza, Campbell and Bailey analyzed anonymized network traffic with a statistical model based on Poisson distribution. The model was used for attack detection. [7]

The next article is 'The usage Bayes-analysis in risk analysis' by Balogh and Hanka. They've used a Bayesian model to estimate the probability of a certain terrorist group committing a certain crime with a certain weapon. The model however, like any Bayesian models, needs input in form of prior possibilities, which can come from governmental agencies such as the intelligence agency. [8]

The study 'Statistical modelling and risk assessment' by Cuny, Lejeune is dealing with risk assessment, and occupational risks in particular. The study relies on the authors' previously introduced concept of risk curve, that is the combination of the two dimensions of quantitative risk assessment: frequency and severity. The authors provide a procedure for producing data suitable for statistical modelling of risks. [9]

As the previous paragraphs indicated, statistics provides a very useful toolset with which helpful information can be gained from even computer log data. Nevertheless, the article 'The use, misuse and abuse of statistics in information security research' points up statistics can also lead to false conclusions, intentionally or even accidentally. Ryan and Jefferson reviewed fourteen surveys on information security in terms of methodology and

results. They've found that in many cases, the research methodologies were unscientific and, in some cases, only partial results were reported in the press. [10]

The last study, I would like to mention here is not related to safety science or information security at all, but it uses statistics in a way which I have found useful in log analytics, so it contributed greatly to the basic idea of the present study. Lovasné performed calculations based on the mortality data of the Hungarian population. [11] I shall point out the similarities of her study and mine in the following section, where I also present the methodology I used.

RESEARCH METHODOLOGY

The aim of the current study is to present a method of how statistics can be used in log analysis. I would like to give predictions and recommendations related to undesired events in IT systems based on their log data. The log data I used for the following calculations are from a real-world scenario, but the data is not so important in this case. The method I shall present can be used on any kind of log data with the proper transformations. The concrete structure of log entries is also irrelevant to the matter.

I am well aware of the fact that using anonymous data can be misleading in some cases, but this is not an issue here, because technically I do not perform anonymization on the data. There are studies concerning anonymization like 'Anonymizing Research Data' by Clark, it states that websites are not willing to publicly release network data for a variety of reasons including confidentiality, privacy, and security issues [12]. Meza, Campbell and Bailey are also mentioning anonymization, in their aforementioned study, as an oftentimes necessary trade-off between security and usability [7].

The soon to be presented method, broadly speaking, takes only a few steps and can be done with basic tools like a spreadsheet application. The steps are the following, identifying a random variable which models the undesirable event, displaying its values on a histogram, finding a probability distribution, that fits to the data series and use that distribution to give predictions and recommendations related to undesired event and its consequences.

Let's consider an IT system with a client-server architecture. The clients are hardware units with embedded software, not necessarily in the same geolocation as the server. This system has an adequate logging facility on both server and client side, which registers operational events (e.g. installation, power on, power off), user activities, software and hardware related events, errors and such. The lifecycle of the clients is limited to a maximum of 3 years. The exact number of clients is 150. Occasionally an undesired event happens on the clients, after which they must be replaced, for becoming unusable. Every client is able to log this event before becoming unusable.

This scenario can be modelled with statistics as follows: X is a random variable; the value of X represents the number of the month from 1 to 36 in which the undesirable event happened and the client became unusable. This is the main idea that comes from Lovasné's aforementioned study, she mentioned a random variable which represented the number of the year in which a person died. [11]

The first step is to determine the value of X in the case of each client. The value of X can be easily calculated from the individual log files of the clients by subtracting the date

of the undesired event entry from the date of the installation event entry of the client, and take the month part of the result.

The next step is to plot a histogram of the values of variable X. I've used Microsoft Excel, but Libre Office Calc is also capable of any task used in this process.

Practically speaking a histogram can be considered as the empirical counterpart of the theoretical density function of a probability distribution (PDF, probability density function), more precisely, it can be verified with a statistical hypothesis test if a theoretical probability distribution estimates the empirical distribution of a data set with a predetermined confidence level. I will be using the so-called Chi-squared test. [3]

Every hypothesis test needs to have a null hypothesis (H_0) and an alternate hypothesis (H_a). These two must be mutually exclusive. In case of testing for goodness of fit H_0 states that the analyzed data series is consistent with the predetermined theoretical distribution. As the two hypotheses must be mutually exclusive H_a cannot state anything else but, that the data series is not consistent with the predetermined theoretical distribution. [3]

Validity of null hypotheses cannot be stated with 100% probability, because most of the times an analyst does not have information about the full population. Because of that, hypothesis tests can be performed having a pre-selected significance level which is usually 5%. [3]

After that the test statistics must be calculated as follows.

$$\chi^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i}$$

In the formula above r means the number of levels of the random variable. k_i is the observed value number i in the data series, n is the number of samples, p_i is value of the theoretical distribution for the corresponding level of the random variable. In the following I will refer to np_i as E_i, the expected value for level number i. [3]

$$E_i = np_i$$

The degrees of freedom (DF), and the test statistics must be calculated from the data series. The DF equals the number of levels (r) of the random variable minus one.

$$DF = r - 1$$

The next step is the determine the critical value (CV). It can be calculated from the Chi-squared distribution with two parameters: significance level and degrees of freedom. This test is a one-sided test, and H_0 is a one-tailed hypothesis, so the test statistics must be below or equal to the critical value. [3]

My H_0 hypothesis is that the data series from the log data conforms exponential distribution with a significance level of 5%. I've chosen exponential distribution because, in general, the service life of machines or their components follows this distribution, provided that they fail due to some random failure. In order to prove that, I needed to make some adjustments to the previously defined random variable X, it has to be the other way around. So, from now on random variable X means that how many months are left to the end of the maximum 3-year lifecycle of the clients when the undesirable event happened.

I would like to stress that, there is a connection between exponential distribution and Poisson distribution. Exponential distribution deals with the time between occurrences

of successive events as time flows by, whereas Poisson distribution deals with the number of occurrences in a fixed period of time. [1] [3]

Should H_0 prove to be valid, I shall use it to give recommendations and predictions related to the adverse events and their effects. Furthermore, applying Poisson distribution, the number of undesired events can be estimated for a time period, a month in this case.

Exponential distribution has a parameter denoted λ, which practically means the mean time between the occurrences of the event in question. The expected value of an exponentially distributed random variable is the following [3]:

$$E[X] = \frac{1}{\lambda}$$

Its variance is [3]:

$$Var[X] = \frac{1}{\lambda^2}$$

I shall estimate the expected value of the previously mentioned random variable X with the mean of the data series. From that λ can be easily estimated.

In the following paragraphs I shall present the concrete calculations I have performed in order to test my H_0 hypothesis. The values of the random variable X and the Chi-squared test calculations can be seen below in the 1. Table Chi-squared test for fitting of exponential distribution to random variable X. The columns in left to right direction contains the following data respectively, level number of the variable, the frequency of the undesired event (k_i), the theoretical values of the exponential distribution for the corresponding level (p_i) with the estimated λ parameter, the expected value for each level E_i, one element of the test statistics for each level S_i.

The mean of the frequencies is 4.167, that is the estimated expected value of variable X (E[X]). The estimate for λ equals reciprocal of E[X], specifically 0.240. Practically speaking, the value of λ means that we get one undesired event in every 0.240 month.

The value of p_i can be calculated with the EXP.DIST function of Microsoft Excel, which has three parameters, the value of the random variable, the value of λ, and a Boolean value, which determines if the EXP.DIST function should calculate the values of the distribution function or the density function (PDF). In this case the PDF is used.

Number of month	k_i	p_i	E_i	S_i
1	16	0.189	28.319	5.359
2	7	0.149	22.276	10.476
3	8	0.117	17.523	5.175
4	12	0.092	13.784	0.231
5	5	0.072	10.843	3.149
6	10	0.057	8.529	0.254
7	8	0.045	6.709	0.248
8	4	0.035	5.278	0.309
9	6	0.028	4.152	0.823

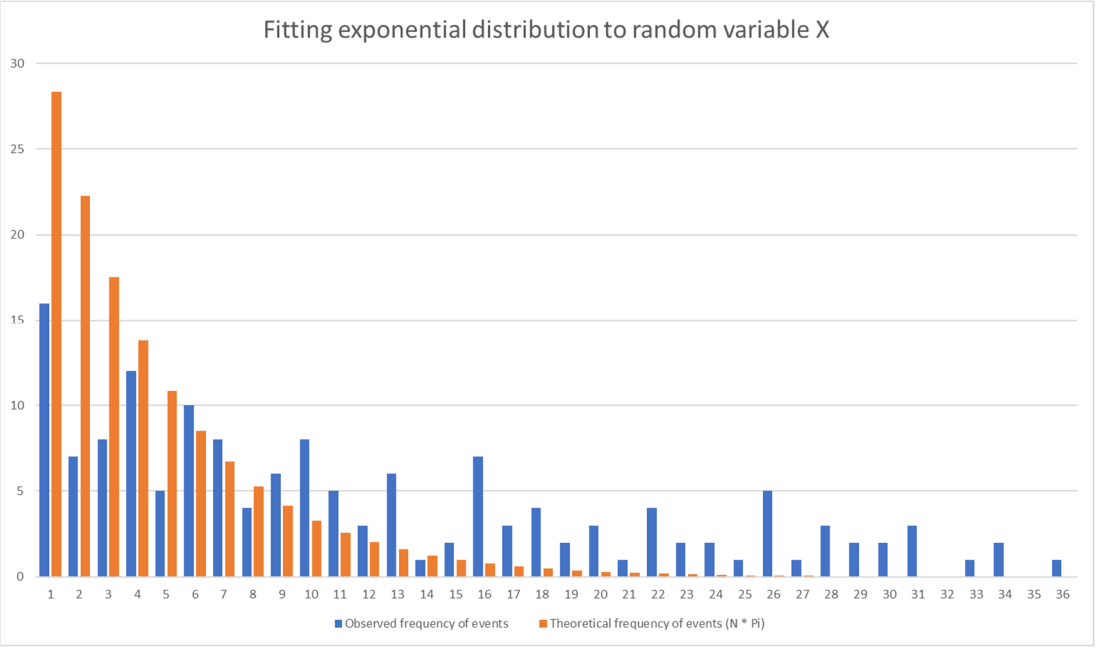
Number of month	k_i	p_i	E_i	S_i
10	8	0.022	3.266	6.863
11	5	0.017	2.569	2.300
12	3	0.013	2.021	0.474
13	6	0.011	1.590	12.236
14	1	0.008	1.250	0.050
15	2	0.007	0.984	1.050
16	7	0.005	0.774	50.100
17	3	0.004	0.609	9.395
18	4	0.003	0.479	25.896
19	2	0.003	0.377	6.997
20	3	0.002	0.296	24.674
21	1	0.002	0.233	2.524
22	4	0.001	0.183	79.459
23	2	0.001	0.144	23.881
24	2	0.001	0.113	31.374
25	1	0.001	0.089	9.296
26	5	0.000	0.070	346.222
27	1	0.000	0.055	16.166
28	3	0.000	0.043	201.248
29	2	0.000	0.034	113.105
30	2	0.000	0.027	144.853
31	3	0.000	0.021	419.709
32	0	0.000	0.017	0.017
33	1	0.000	0.013	74.451
34	2	0.000	0.010	384.698
35	0	0.000	0.008	0.008
36	1	0.000	0.006	155.043

1. Table Chi-squared test for fitting of exponential distribution to random variable X

The test statistics equals with the sum of the last column, that is 2168.111. The chosen significance level is 5% percent. The DF (degrees of freedom) value is the number of levels minus one, which is 35. The CV (critical value) can be calculated with the CHISQ.INV.RT Excel function. It requires two parameters: the probability associated with the chi-squared distribution, which equals 1-significance level, 0,95 and the value of DF. CV is 49,802.

Comparing the value of the test statistics and the CV, it turns out, that random variable X is not conform with exponential distribution, because the value of the test statistics is way above the CV.

On the 1. Figure Fitting exponential distribution to random variable X, below, the blue bars represent the empirical frequencies of variable X, and the orange bars the theoretical frequencies given the number of samples and parameter λ. The significant difference is clearly visible even to the naked eye.



1. Figure Fitting exponential distribution to random variable X

I have given it another try picking another time period instead of one month. I've merged every two months into one category, so let Y be a random variable; the value of Y represents the number of the 2-month period from 1 to 18 in which the undesirable event has happened and the client has failed.

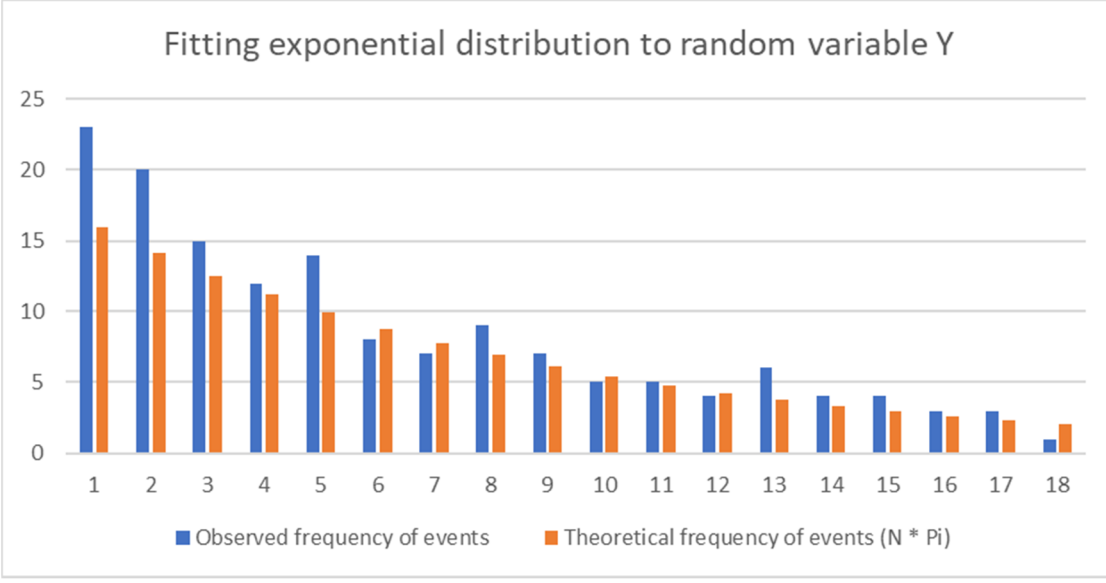
The 2. Table Chi-squared test for fitting of exponential distribution to random variable Y contains the merged data with the same columns and calculations as before. This way however, the model "loses precision", in other words estimates can only be made for only two-month periods.

Number of time period	**k_i**	**p_i**	**E_i**	**S_i**
1	23	0.106	15.965	3.100
2	20	0.094	14.159	2.409
3	15	0.084	12.558	0.475
4	12	0.074	11.138	0.067
5	14	0.066	9.879	1.719
6	8	0.058	8.762	0.066
7	7	0.052	7.771	0.076
8	9	0.046	6.892	0.645

Number of time period	k_i	p_i	E_i	S_i
9	7	0.041	6.113	0.129
10	5	0.036	5.421	0.033
11	5	0.032	4.808	0.008
12	4	0.028	4.265	0.016
13	6	0.025	3.782	1.300
14	4	0.022	3.355	0.124
15	4	0.020	2.975	0.353
16	3	0.018	2.639	0.049
17	3	0.016	2.341	0.186
18	1	0.014	2.076	0.558

2. Table Chi-squared test for fitting of exponential distribution to random variable Y

The results are the following: E[Y] is 8.333, λ equals to 0.120, CV is 27.587, test statistics is 11.314. The value of the test statistics is clearly below the CV, so this version of the random variable conforms exponential distribution. The 2. Figure Fitting exponential distribution to random variable Y, below also confirms the fact that there is no significant difference between the empirical and theoretical distribution in this case. Of course, the colors have the same meaning as before.



2. Figure Fitting exponential distribution to random variable Y

Now that it is proven that random variable Y can be described with an exponential distribution, and its λ parameter is 0.120, further calculations can be made.

As stated before, there is a connection between exponential and Poisson distribution. Poisson distribution also has a parameter named λ, but it has a different meaning than

it has in the case of the exponential distribution, so I shall use μ for referring to the parameter of Poisson distribution hereinafter. μ means the expected value of occurrences of the event in a period of time. The time period is 2 months, the expected value can be easily calculated from λ, as follows.

$$\mu = \frac{1}{\lambda}$$

The expected value and variance of a Poisson-distributed random variable are both equal to μ. [3] In this case $\mu = 8.333$

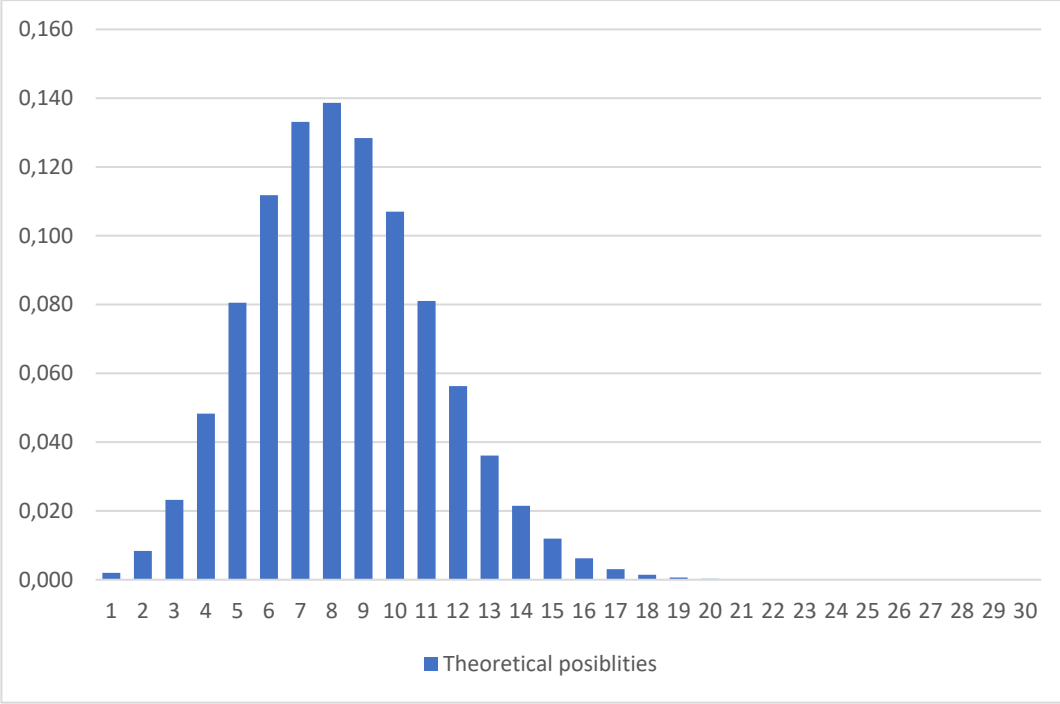
However, it is theoretically possible for each and every hardware unit to fail in a 2-month period of time, but it is very much not likely, given the observed values, among which the maximum number of failures was 23, so I shall only calculate the Poisson distribution for the maximum level of 30 using the POISSON.DIST function in MS Excel. The calculations can be seen in the 3. Table Poisson distribution of random varibale Y below.

Number of occurrences	Theoretical posiblities	P_b	P_{ea}
1	0.002		
2	0.008	0.002	0.998
3	0.023	0.010	0.989
4	0.048	0.034	0.966
5	0.080	0.082	0.918
6	0.112	0.162	0.837
7	0.133	0.274	0.726
8	0.139	0.407	0.593
9	0.128	0.546	0.454
10	0.107	0.674	0.325
11	0.081	0.781	0.219
12	0.056	0.862	0.137
13	0.036	0.919	0.081
14	0.021	0.955	0.045
15	0.012	0.976	0.024
16	0.006	0.988	0.012
17	0.003	0.994	0.005
18	0.001	0.997	0.002
19	0.001	0.999	0.001
20	0.000	0.999	0.000
21	0.000	1.000	0.000
22	0.000	1.000	0.000
23	0.000	1.000	0.000
24	0.000	1.000	0.000

Number of occurrences	Theoretical posiblities	P_b	P_{ea}
25	0.000	1.000	0.000
26	0.000	1.000	0.000
27	0.000	1.000	0.000
28	0.000	1.000	0.000
29	0.000	1.000	0.000
30	0.000	1.000	0.000

3. Table Poisson distribution of random varibale Y

The 3. Figure The Poisson distribution of random variable Y shows the PDF of the Poisson distribution which describes the data series.



3. Figure The Poisson distribution of random variable Y

The bars represent the probability of the number of occurrences for each two-month period of time. This PDF can be used for many purposes. For example, recommendations can be made for the number of backup hardware units (N_{BHU}) in order to ensure the quick and smooth handling of necessary replacements with the lowest cost possible. One way of doing this is the following. Calculate the accumulated possibility of the number of occurrences being below the current level (P_b) and the accumulated possibility of the number of occurrences being equal to or above the current level (P_{ea}). Find the lowest number where P_{ea} is lower than P_b, this will be the minimum recommended N_{BHU}, the smallest level where it is more likely that there are enough backup hardware units for replacing the broken ones.

Further increase of the N_{BHU} on one hand, enhances the chance of easy execution of replacements, on the other hand it increases the expenses on amassing backup hardware units, so determining the N_{BHU} requires individual consideration in specific cases. In the current case the minimum N_{BHU} is 9.

Furthermore, knowing the total cost of replacing a single hardware unit, the average maintenance expense can be calculated as well for the time periods. Completely independently of this model, the cost of eliminating the root cause of the error can be estimated. From the two results it can be shown how long will it take for the elimination of the root cause to pay off, helping management making the best choices.

SUMMARY

In my humble opinion this article proves that mathematical statistics should be taken into consideration when it comes to log analysis. Creating the appropriate model for the phenomenon in question may prove to be useful in many ways. Not only it provides a better understanding of the situation through descriptive statistics, but it may support the decision-making process with predictions formulated with mathematical precision using probability theory.

RESOURCES USED

[1] A. Prékopa, Valószínűségelmélet, Budapest: Műszaki Könyvkiadó, 1980.
[2] E. T. Janes, "Bayesian methods: General background," Maximum Entropy and Bayesian Methods in Applied Statistics, pp. 1-25, 1985.
[3] O. Lukács, Matematikai statisztika, Budapest: Műszaki Könyvkiadó Kft., 2006.
[4] L. Muha and C. Krasznay, Az elektronikus információs rendszerek biztonságának menedzselése, Budapest: Nemzeti Közszolgálati Egyetem, 2014.
[5] L. Muha and C. Krasznay, Az elektronikus információs rendszerek biztonságáról vezetőknek, Budapest: Nemzeti Közszolgálati Egyetem, 2018.
[6] K. Kent and M. Souppaya, "Guide to Computer Security Log Management," National Institute of Standards and Technology, Gaithersburg, 2006.
[7] J. Meza, S. Campbell and D. Bailey, "Mathematical and Statistical Opportunities in Cyber Security," Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 2009.
[8] Z. Balogh and L. dr. Hanka, "Bayes-analízis alkalmazása a kockázatelemzésben," Műszaki katonai közlöny, vol. különszám, pp. 57-72, 2012.
[9] X. Cuny and M. Lejeune, "Statistical modelling and risk assessment," Safety Science, vol. 41, pp. 29-51, 2003.
[10] J. J. Ryan and T. I. Jefferson, "The use, misuse and abuse of statistics in information security research," Management National Conference (ASEM 2003), 2003.
[11] J. Lovasné Avató, "A valószínűségszámítás egyik gyakorlati alkalmazása," Szakmai Füzetek, vol. 10., pp. 45-54., 2001.
[12] A. Clark, "Anonymising Research Data," Real Life Methods, Sociology, 2006.
[13] M. Keramati, "An Attack Graph Based Procedure for Risk Estimation of Zero-Day Attacks," 8th International Symposium on Telecommunications, 2016.