Statistical Modelling of Network Communications based on Monitored Data

Radi Romansky
Technical University of Sofia (Bulgaria)

Contents

Abstract

The purpose of this article is to discuss a statistical approach for investigation of access to remote information objects in the digital space. The data are collected based on monitoring of a real communication parameters and processed by user application. Essence of statistical modelling applied in research is presented as a beginning of the discussion, which is used as a base for formulating the main analysis used in the investigation. The main part of the research is presentation of experimental result of selected factors of network communication study.

Keywords

statistics, statistical modelling, networking, communication investigation, assessments.

I. INTRODUCTION

The statistical approach allows a relatively simplified presentation of specific data through mathematical relationships, describing with a certain precision the relationships between the studied factors. It allows the classification of certain objects into separate groups based on statistical characteristics, which requires, before conducting the statistical analysis, to initially determine the main variables, to form a sample and to carry out filtering of atypical values.

Statistical modelling is applied in various fields of scientific research when it is necessary to represent causal relationships in the research object. It uses empirical values for the behaviour of an object, which is why it is often called empirical modelling or the empirical approach.

The paper deals with statistical model investigation of selected data obtained based on program monitoring of communication parameters. The objects are relations between user activity and reaction of the network spice. The goal is to determine the conditions for creating problematic situations in communications and to look for solutions to overcome them.

II. Mathematical Statistics and Modelling

A. Forming the sample and statistical estimations

Each sample is a certain part of a general population *W*, which is the complete set of elements *w* possessing the characteristic of carrying out the study. Events are defined as a subset of the selection space.

The choice of a statistical model must comply with the set goal of the research being conducted, and some traditional basic statistical models and estimates are summarized in Table 1.

The main methods and tools for mathematical statistics, applied when conducting statistical modelling are the following.

✓ *Method of stochastic approximation*, which is based on a general iterative procedure of a desired characteristic based on a previous value.

✓ *Method of parametric functions*, which is usually reduced to the above, but with the introduction of an additional parameter depending on the number of iterations.

✓ *Method of least squares (MLS)*, which is the most frequently used method in determining estimates for statistical models.

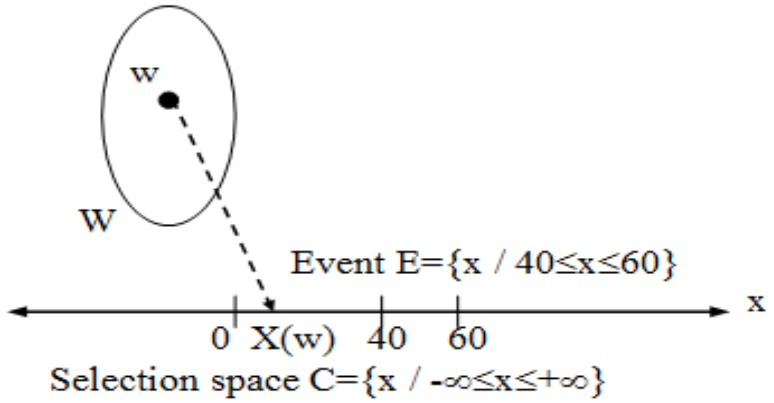
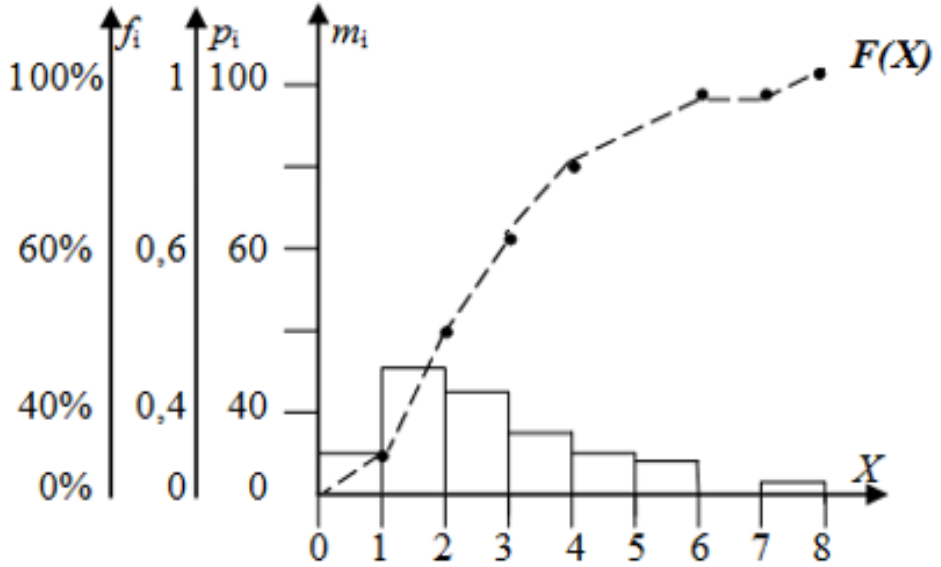


Table 1. Basic Statistical Assessments

Mathematical expectation	$E[X]=x_{av}=\frac{1}{N}\sum_{i=1}^{k}(m_i x_i)$
Variance	$V[X]=\sigma^2=\frac{1}{N}\sum_{i=1}^{k}[m_i(x_i-x_{av})^2]$
Standard deviation	$\sigma_X=+\sqrt{V[X]}$
j-th starting (initial) moment	$\mu_j=\frac{1}{N}\sum_{i=1}^{k}(m_i x_i^j)$
j-th central moment	$\sigma_j=\frac{1}{N}\sum_{i=1}^{k}[m_i(x_i-x_{av})^j]=\sigma^j$
Coefficients of asymmetry	$A_S=\frac{\mu_3}{\sigma^3}$
Coefficients of excess	$E_x=\frac{\mu_4}{\sigma^4}-3$
Coefficient of variation	$C_v=\frac{\sigma_i}{X_{i-AV}}$

B. Descriptive and multivariate statistical models

Statistical models from the first group of univariate (descriptive) models are formed by calculating univariate statistical estimates based on a formed sample $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ for a separate factor (variable) X .This basic statistical evaluation such as sample size N, range of variation (x_{min}, x_{max}), mean value x_{av}, variance σ^2, standard deviation σ, coefficient of deviation $C_v=\sigma/x_{av}$, etc. To these can be added formation of stratification of groups according to set limit values, formation of a histogram or of a frequency distribution.



The group of multivariate statistical models represents the relationship of two or more factors in a multivariate experiment. Basic statistical models in this group are as follows.

✓ Correlation analysis to study the relationship between two factors, with modelling based on the calculation of correlation and covariance coefficients for each defined pair of parameters.

✓ Regression statistical modelling, in which the existence of a dependence $Y=Y(X_1, \ldots, X_k)$ is assumed between one variable Y, chosen as dependent, and one (simple regression) or several (multiple regression) independent variables.

✓ Dispersion statistical models, where the variance of the general set is calculated by calculating statistical estimates based on the sample.

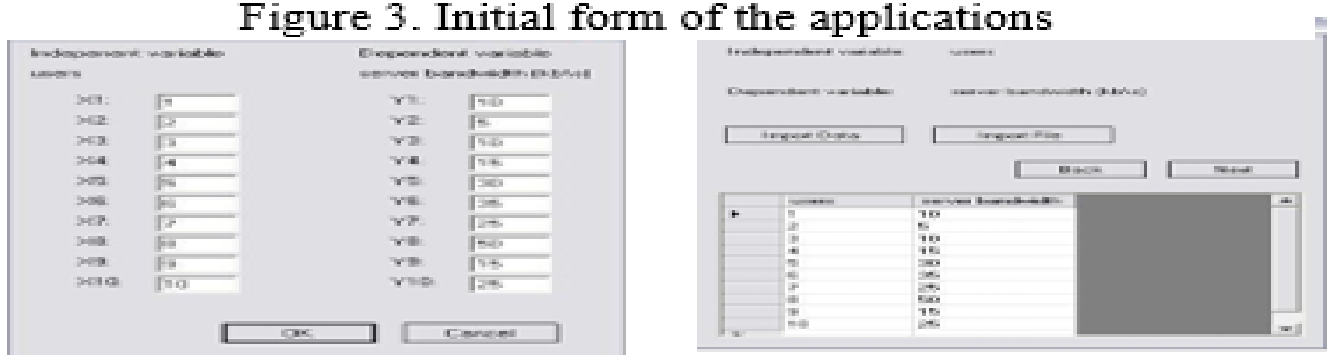
III. Statistical Modelling and Investigation

A. Program application for statistical modelling

"Network Analyser" is a Windows-based application developed in the Microsoft Visual Studio.Net environment and implemented in the C# language. The main purpose is to perform statistical modelling of accumulated empirical data regarding various applications. The main forms are presented in Figures 3 & 4, and example – in Figure 5.

(a) Dependence selection (b) Input of empirical data

Figure 3. Initial form of the applications



(a) Manual data entry (b) Import a data file

Figure 4. Empirical data entry form

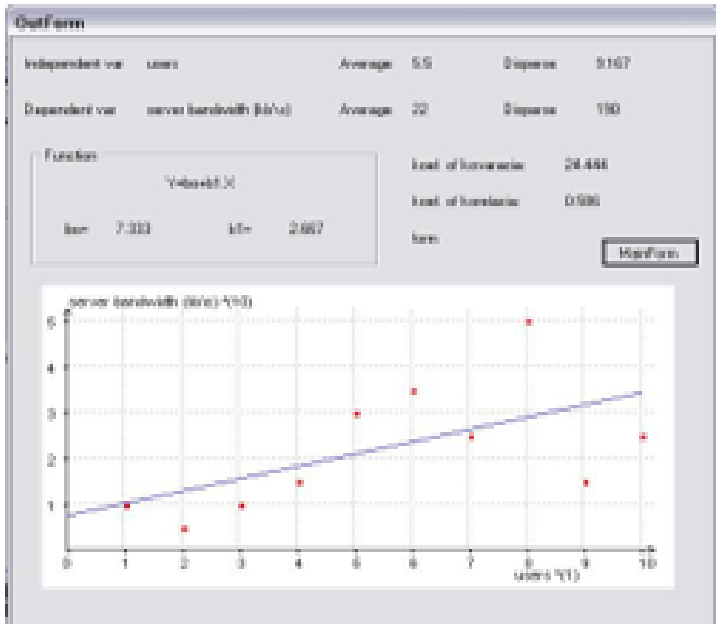


Figure 5. OutForm with a statistical regression model

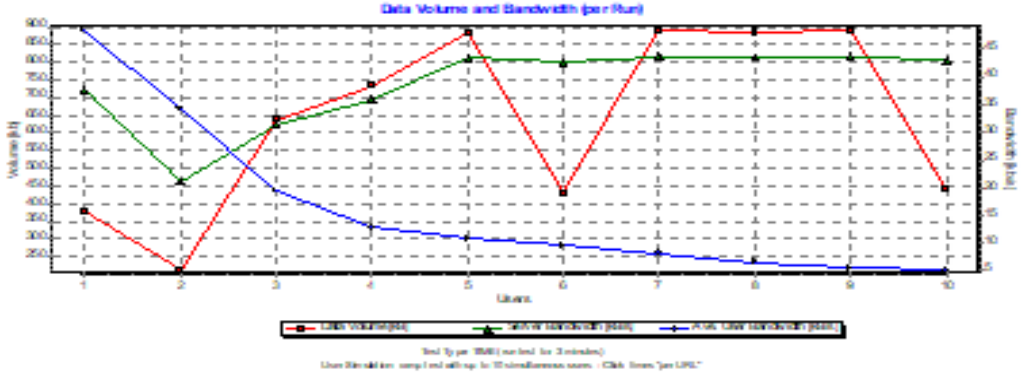
B. Experimental results

Statistical model experiments were conducted in several groups depending on the investigated factors and the set goal. Data collected by program monitoring is combined into work files (samples) for input into the "Network Analyser". Experimental results of a study of the mutual dependence of the factors '*server bandwidth', 'click time', 'average request time', 'time to first byte'* on the number of users ('*users'*), will be presented here, and the monitoring data is presented in Figure 6. Three objects are subjected to measurement – form (form), static site (stat) and dynamic site (dyna).

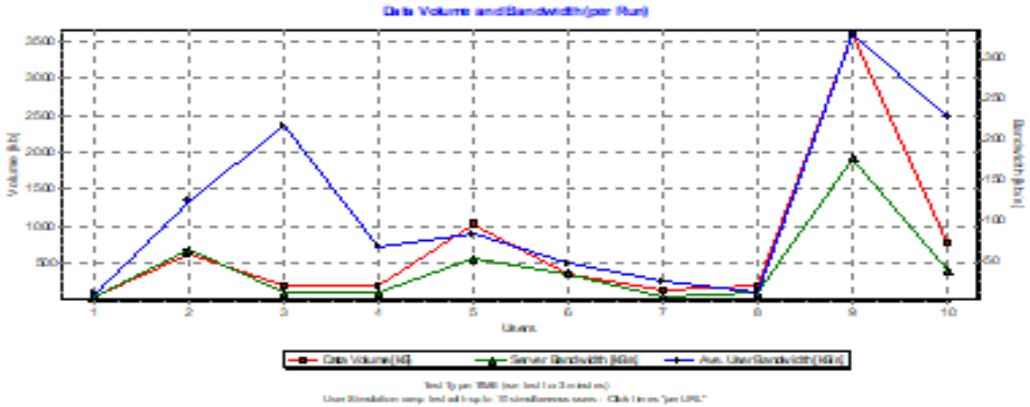
users	*server bandwidth (kb/s)*			*avg. click time (ms)*			*aver. request time (ms)*			*time to first byte(ms)*		
	form	stat	dyna	form	stat	dyna	form	stat	dyna	form	stat	dyna
1	10	37	0	1000	20	7500	20	20	7000	0	10	4500
2	5	20	65	12000	45	1000	45	45	1000	0	35	7500
3	10	32	12	0	60	3500	60	60	3500	0	50	500
4	15	37	12	3000	80	15500	80	80	15000	2500	65	1250
5	30	42	50	3000	90	4000	90	90	4000	500	75	1500
6	35	42	37	2500	100	3500	100	100	3500	750	85	1250
7	25	42	0	2000	120	3000	120	120	3000	1000	100	250
8	50	42	12	2000	150	6000	150	150	6000	1000	110	4000
9	15	42	175	4000	170	2000	170	170	2000	3250	160	500
10	25	42	37	7500	190	11000	185	185	11000	2500	170	1000

Figure 6. Empirical monitored data for group (1)

An initial hypothesis of a linear dependence of '*server bandwidth*' on the number of users '*users*' was accepted, and the analysis carried out and the statistical regression models formed confirm the assumption for the investigated objects. Figure 7 presents experimental data from the tests performed when measuring the dependence "server bandwidth/users“ and the regression models are presented in Figure 8.

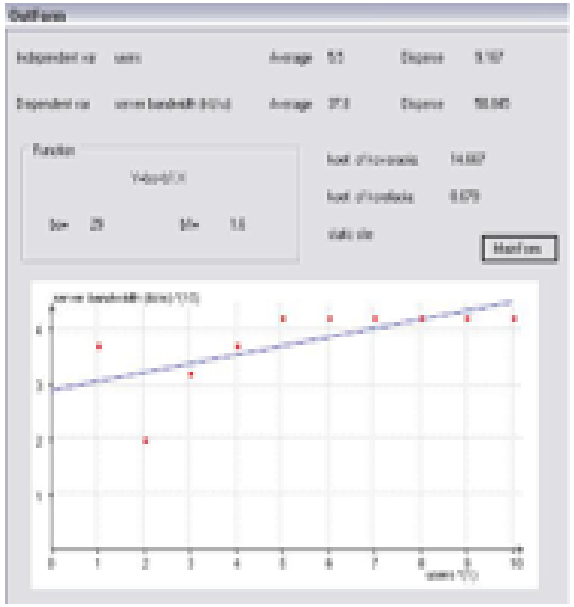


b) Statistical site

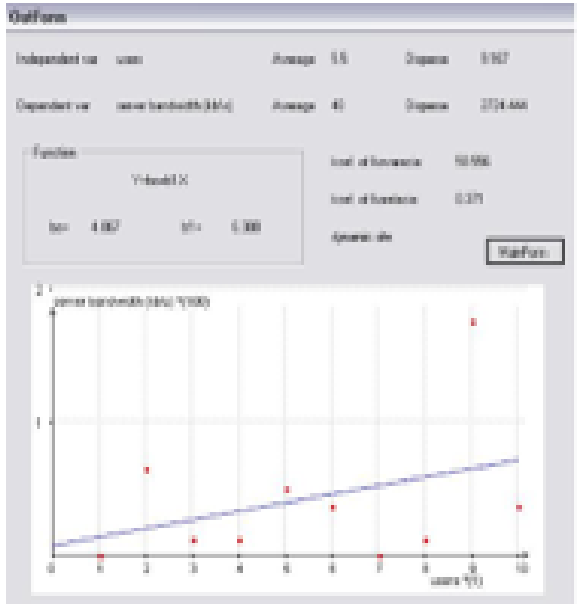


c) Dynamic site

Figure 7. Test results for ‘server bandwidth’ / ‘users’



(a) ‘Stat’



(b) ‘Dyna’

Figure 8. Regression models for ‘Stat’& ‘Dyna’

V. Conclusion

In principle, any experiment is a set of purposeful actions, allowing to reveal the essence of the state and functioning of an investigated object. The main components of a research experiment are its planning and execution, the accumulation of a representative sample of data, the selection of an appropriate method for processing the sample, and the conduct of analysis to form conclusions. In this reason, in the article, a method of program monitoring was chosen to form a sample of measured values for previously planned parameters (factors for analysis), and the analysis itself was made based on the method of statistical modeling and, in particular, conducting multifactorial regression and correlation analysis.

In principle, an extensive study of various groups of network parameters for remote access to resources has been carried out, of which only one group is presented here. This provides an opportunity for future expansion of experiments in the chosen direction, as well as application of other statistical methods for model research.

.

REFERENCES

15 publications are included in the list of references

Thank you for your attention

Radi Romansky, rrom@tu-sofia.bg