

Statistical Analysis of Data for Modeling The Behavior of An Indeterministic Software

Ghodrat Moghadampour

Department of Information Technology and Production Economics,

University of Vaasa

PO Box 700, FIN-65101 Vaasa, Finland

Fax:+358+6+3248467

E-mail: mg@uwasa.fi

Abstract

In this work statistical analyses were used in order to model the behavior of a protection relay and to investigate relationships among variables affecting the functioning of the tested software. The objective was to test the performance of the protection relay under high load.

The results showed that statistical analyses were particularly a useful means for modeling the behavior of software, which work indeterministically. Moreover, a detailed picture of the behavior of the software under test can be achieved by statistical analyses.

Keywords : Software testing, testing of real-time systems, statistical analysis, statistics-based software testing, load testing, software modeling.

1. Introduction

Systems are generally required to be reliable and expected to operate without failures. Systems are also expected to be available when required. When software is safety-critical or has very high reliability requirements the program must be designed so that catastrophic failures never happen.

1.1 Testing of real-time systems

There is a definite need for verifying the correct behavior of embedded (real-time) systems, a system in which the software must operate at the speed demanded by the system inputs and outputs [O'Co94].

Some testing problems with real-time systems are [Aue97]:

1. The system may react differently to the same event in different states. The tester has only a limited control over the execution of the program; the system should be reset and observed as a black box; it is difficult to verify the behavior of the system according to the related pair of inputs and outputs.

2. Real-time systems behave periodically and perform some operations in the same way. So, it is difficult to identify what state the real-time system should be returned, before starting a new similar sequence.

2. Experimentation and statistical analysis

In order to identify the behavior of the tested software, the relay response time was taken as the *dependent* variable and it was divided into two parts: InTime and OutTime. The following *independent* variables were also specified: OInt: the time interval between sending consequent ordinary messages, ODInt: the time interval from sending the ordinary message to sending disturbance message, AL: data length of ordinary message for A card of the relay, BL: data length of ordinary message for B card of the relay, DL: data length of disturbance message, ROL: data length of returning ordinary message, POT: processing time for previous ordinary message, DSent: indicates whether the disturbance message is sent or not, and if sent, at which stage.

By investigating extreme values it was tried to find factors whose interactions make response time values unexpectedly small or large. The correlation analysis for extreme values (in this case values larger than 90 percentile) were used for investigating the linear relationship between relay response time and other variables.

The principal component analysis was used to see if the data could be summarized in fewer dimensions. This was necessary to decrease the number of values, form new auxiliary variables and see how efficiently auxiliary variables cloud explain the variance of data.

Contingency tables were used to investigate the observed frequency of having long response time values and sending disturbance signals in different stages while the main messages are

processed. Furthermore, multiple-comparison tests were used to see if there are statistical differences between the mean time values of sending disturbance signals in different phases of the processing of the main signals. Canonical discriminant analysis were also used to see if there were any statistical differences between classes of the variable indicating the phase at which disturbance signals were sent.

The interaction effects of main independent variables on the response time variable were also investigated for extreme values. The results of the general linear model were used to see how significantly the proposed model could predict the variation of the response time variable.

3. Conclusions

Results showed that statistical analyses could be efficiently used to indicate the factors most responsible for the differences in the response time distributions.

The analyses indicated that differences in the relay response time distributions were mostly due to the differences in *InTime* distributions. The *InTime* part of the relay response time had more significant influence and caused more variation in the relay response time distribution. Although there was no general linear relationship between variables, some variables in the space of functionality of the relay were linearly related. No single variable could predict the relay response time or even one of the parts of it.

Statistical analysis showed that two or three components, written as linear combination of different variables, provided a good summary of the data for the whole data set. The results were the same for extreme low values. However, for extreme high response time values the overall effect of variables could be better squeezed in a few compact variables.

The analyses indicated that only a minor portion of variance of the relay response time could be explained by the stage, at which the disturbing message was sent. Furthermore, the results proved that sending ordinary messages just before sending an ordinary message had the most dramatic effect on the relay response time.

The test results also showed that the stage, in which disturbance message was sent, had a significant influence on the variables: *InTime*, *OutTime* and *ODI*. The results for extreme low and extreme high response time values gave conclusions in the same directions.

The results pointed out that for the whole data set a quite significant portion of the variation of

the response time variable could be predicted by the model constructed by the set of independent variables in use.

For extreme high response time values the set of independent variables could predict only a minor portion of the variation of the relay response time variable. This fact implies that in case of high loads the relay acts more unpredictably than in general.

However, the predictability of *InTime* by the set of independent variables in case of high loads was as much as in other cases. Moreover, the results proved that *OutTime* was as predictable as in cases of low loads, but significantly more predictable than in general. These results, in turn, give reason to believe that *OutTime* is more predictable than *InTime* in case of extreme loads.

3.1 Directions for future research

Statistical analysis of data was found useful in determining the relations between variables affecting the behavior of an indeterministic software working in a simulation environment. In the next steps, statistical analysis can be used particularly to model the behavior of more indeterministic software, reveal unknown relations between factors, find faults and develop the software more. Statistical analysis of the effects of real loads on the system in a real environment would give a more realistic concept of the relations between variables in the software.

References

- [Aue97] Auer, Antti (1997). *State Testing of Embedded Software*. University of Oulu. Department of Information Processing Science and Infotech Oulu. Research papers. Series A 25.
- [O'Co94] O'Connor, Patrick D.T. (1994). *Practical Reliability Engineering*. John Wiley & Sons Ltd. Third edition. Printed in Great Britain by Bookcraft (Bath) Limited.