STATISTICAL PROCESS CONTROL IN SOFTWARE ENGINEERING

Šárka Květoňová, Zdeněk Martínek
Doctoral Degree Programme (2, 3), FIT BUT
E-mail: {kvetona, martine}@fit.vutbr.cz

Supervised by: Jitka Kreslíková
E-mail: kreslika@fit.vutbr.cz

ABSTRACT

This submission presents an improved/modified approach for statistical process control in software engineering. It is based on clasical statistical approach which is adjusted to specific requirements of software development processes. Within the scope of this article, statistical process control (SPC) is the key element. In general, SPC is a method for achieving quality control in manufacturing processes. It employs control charts to detect whether the process observed is under control. We assume that our suggested approach should improve security of software development through discovering more deviations and faults.

1. INTRODUCTION

Classical quality control was achieved by inspecting 100% of the finished product and accepting or rejecting each item based on how well the item met specifications. In contrast, statistical process control uses statistical tools to observe the performance of the production line to predict significant deviations that may result in rejected products.

This approach is applicable in software engineering, as well. However, this domain implies a slightly modified conception for performance and stability evaluation of given processes under examination.

1.1. IMPORTANT TERMS

In this section, we will clarify the following important terms from statistical process control (SPC):

Variation – common cause variation and assignable cause variation give together total variation.

Process stability – it means that process is in statistical control.

Testing for stability – can be performed with the next two perspectives in mind:

- variation of values within the subgroup at each measurement point,
- comparison of variability from one subgroup to another.

2. SUGGESTED APPROACH

Control limits

One of the most important principles associated with control charting involves using knowledge of the process to select subgroups that contain only common cause variation (in so far as possible). This is called *rational subgrouping* and the purpose is to permit the computation of estimates for standard deviations (values for sigma - σ) that are uncontaminated by assignable cause variations [4].

Variables and attributes data

Outputs of process measurements traditionally fall into one of two classes: variable data or attributes data. Variables data are usually measurements of continuous phenomena (elapsed time, amount of effort, years of experience, memory utilization, rework cost, etc.). On the other hand, attributes data have a different origin and serve to a different purpose. They occur when information is recorded only about whether an item conforms or fails to conform to a specified criterion or set of criteria. Attributes data almost always originate as counts: the number of defects found, the number of defective items found, the number of source statements of a given type, the percent of projects using formal code inspections, and so on.

Regarding extent of this paper, there will be presented suitable charts for variables data only. All described charts have corresponding alternative for attributes data too.

Detecting instabilities

Several procedures for detection of unusual patterns and nonrandom behaviour are available [5]:

- A single points falls outside the 3-sigma control limits.
- At least two out of three successive values fall on the same side of, and more than two sigma units away from, centreline.
- At least four out of five successive values fall on the same side of, and more than one sigma unit away from, the centreline.
- At least eight successive values fall on the same side of the centreline.

2.1. X-BAR AND R CHARTS FOR VARIABLES DATA

SCP gives us many tools and techniques to examine process stability. If there are subgroups of size n>1, then X-bar (average) chart is used in common to show the observed variation in central tendency from one subgroup to the next. To observe the dispersion in process performance across subgroups is usually used R chart. Control limits for X-bar and R charts can then be calculated from the range data and subgroup averages.

Charts for averages (X-bar) and range (R) charts are used to portray process behaviour when there is the option of collecting multiple measurements within a short period of time under basically the same condition. When the data are collected under basically the same conditions, measurements of product or process characteristics are grouped into self-consistent sets (subgroups) that can reasonably be expected to contain only common cause variation. The results of the groupings are used to calculate process control limits, which, are used to examine stability and control the process.

X-bar charts help us to find out the central tendency of the process and variation, which has occurred from subgroup to subgroup over time. The corresponding R charts indicate the variation (dispersion) within the subgroups. By using the stability detection rules discussed in previous section, it is possible to determine whether the subgroup averages (plotted on the X-bar chart) have been affected by assignable causes and if the subgroup ranges have been affected by assignable causes. Charts for averages and ranges are used together to identify points where a process has gone out of control.

For calculating control limits for X-bar and R charts is necessary to compute at first the average $\overline{X}$ and range R for each subgroup of size n, for each of the k subgroups. The equations are follows:

$$\overline{X_k} = \frac{X_1 + X_2 + .. + X_n}{n} \qquad R_k = |X_{MAX} - X_{MIN}|$$

In next step is computed the grand average $\overline{\overline{X}}$ by averaging each of the k subgroup averages:

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \overline{X}_2 + .. + \overline{X}_k}{k}$$

Similar step is made to compute the average range $\overline{R}$ by averaging each of the k subgroup ranges:

$$\overline{R} = \frac{R_1 + R_2 + .. + R_k}{k}$$

The equations for determining the centreline (CL) and the upper and lower limits (UCL and LCL) are following:

Average (X-bar) chart limits:	**Range (R) chart limits:**
$UCL_{\overline{X}} = \overline{\overline{X}} + A_2\overline{R}$	$UCL_{\overline{R}} = D_4\overline{R}$
$CL_{\overline{X}} = \overline{\overline{X}}$	$CL_{\overline{R}} = \overline{R}$
$LCL_{\overline{X}} = \overline{\overline{X}} - A_2\overline{R}$	$LCL_{\overline{R}} = D_3\overline{R}$

The terms A2, D3 and D4 are conventional symbols for factors that have been tabulated by statisticians for converting averages of subgroup ranges into unbiased estimates for 3-sigma limits.

2.2. XmR CHARTS FOR VARIABLES DATA

When measurements are spaced widely in time or when each measurement is used by itself to evaluate or control a process, a time-sequenced plot of individual values, rather than averages, may be all that is possible. The subgroup size n is then 1, and the formulas based on subgroup ranges that are used for plotting limits for X-bar and R charts no longer apply, so another way is needed to calculate control limits. Short term variation between adjacent observed values is used to estimate the natural variation of the process and this leads to a pair of charts – one of the individual values and another for the successive 2-point moving ranges (X and mR charts).

The idea behind XmR charts is that, when subgroups can easily include nonrandom components, we minimize the influence that nonrandom effects have upon estimates for sigma by keeping the subgroups as small as possible. The smallest possible subgroup size is 1. There is no way to estimate sigma from a single measurement, so changes that occur between successive values are attributed to inherent variability in the process. The absolute values of these changes are called 2-point moving ranges.

Before calculating process limits is necessary to specify moving ranges (r), ith moving range (mR_i), individuals average moving range ($\overline{mR}$).

Now is possible to calculate upper natural process limit (UNPL), centreline (CL) and lower natural process (LNPL).

$$r = k -$$

$$mR_i = |X_{i+1} - X_i| \text{ for } 1 \leq \text{i} \leq \text{k-1}$$

$$\overline{mR} = \frac{1}{r}\sum_{i=1}^{i=r} mR_i$$

$$UNPL_x = \overline{X} + \frac{3\overline{mR}}{d_2} = \overline{X} + 2{,}66\overline{mR}$$

$$CL_x = \overline{X} = \frac{1}{k}\sum_{i=1}^{i=k} X_i$$

$$LNPL_x = \overline{X} - \frac{3\overline{mR}}{d_2} = \overline{X} - 2{,}66\overline{mR}$$

Value for d_2 can be obtained from any statistical tables, namely table of dispersion adjustments factors (bias correction factors).

These factors enable us to calculate 3-sigma limits for both individual values and the average moving range:

$$3\sigma_x = \frac{3\overline{mR}}{d_2} = 2{,}66\overline{mR} \qquad \sigma_x = \frac{\overline{mR}}{d_2}$$

Although the impracticality of grouping may be one reason for charting individual values, there are other reasons that can motivate to usage of XmR charts. There are five types of conditions that may be detected more readily with individual charts than with X-bar and R charts [2]:

1. Cycles (regular repetitions of patterns)
2. Trends (continuous movement up or down)
3. Mixtures (presence of more than one distribution)
4. Grouping or bunching (measurements clustering in spots)
5. Relations between the general pattern of grouping and a specification.

XmR charts can also occur as a natural evolution of simple run charts, once sufficient points become available to compute reasonably reliable control limits.

Care should always be exercised when interpreting patterns on a moving range chart. All moving ranges are correlated to some degree, and this correlation can induce patterns of runs or cycles. A number of authors advise not to apply assignable cause tests 2, 3 and 4 to the moving range charts [3].

3. CONCLUSION

Several authors equate variables data to continuous phenomena and attributes data to counts. Unfortunately, there are many situations where counts get used as measures of size instead of frequency, and these counts should clearly be treated as variables data. Examples should be: counts of the total number of functions or modules, total lines of code, total bubbles in a data-flow diagram, total entities in ER-diagram, total people assigned to a project, etc. These counts represent all the entities in a population, not just the occurrences of entities with specific attributes. Counts of entities that represent the size of a total population should almost always be treated as variables data, even though they are instances of discrete counts.

It is clear to assume that classifying of data depends not so much on whether the data are discrete or continuous, but on how they are collected and used. The method of analysis that is chosen for any collected data depends on the questions, the data distribution model and the assumptions that were made with respect to the nature of the data.

In software environments measurements often occur only as individual values, so we are not able to get samples of size n>1 where the inherent variation within each sample can reasonably be assumed to be homogeneous.

This lead us toward individuals and moving range (XmR) charts, which are suitable for examining the time-sequenced behaviour of process data.

ACKNOWLEDGEMENTS

This research was supported by the Research Plan No. MSM 0021630528, Security-Oriented Research in Information Technology.

REFERENCES

[1] Florac W. A., Carleton A. D.: Measuring the Software Process, Addison-Wesley, 2001, ISBN: 0-201-60444-2

[2] Western Electric Co., Inc.: Statistical Quality Control Handbook, Indianapolis, AT&T Technologies, 1958

[3] Wheeler, Donald J., and chambers, David S.: Understanding Statistical Process Control, 2nd ed. Knoxville, Tennessee: SPC Press, 1992.

[4] Montgomery, Douglas C.: Introduction to Statistical Quality Control, 3rd edition, John Wiley & Sons, 1996

[5] Reijers, H. A.: Design and Control of Workflow Processes, Berlin, Springer, 2003, ISBN: 3-540-01186-2