2005 WSEAS Int. Conf. on DYNAMICAL SYSTEMS and CONTROL, Venice, Italy, November 2-4, 2005 (pp131-135)

Inspection performance's estimation using McNemar statistical test

CARMEN SIMION, SORIN BORZA
Manufacturing Engineering Department
University "Lucian Blaga" from Sibiu, Faculty of Engineering "Hermann Oberth"
4 Emil Cioran Street, 550025 Sibiu
ROMANIA

Abstract: - The benefit of using a data based procedure is largely determined by the quality of the data obtained from a measurement system. If the measurement system has too much variation, it will affect the ability to make the right decision. Examining measurement system elements that may cause measurement errors, is a good approach to troubleshooting. One of the error sources may be the inspector, so the inspection capability must be evaluated. The paper deals with inspection capability problem and one first aspect must be to determine if there is a significant difference in the performance characteristics of the inspectors. Because one approach to test for difference between inspectors is to use the McNemar statistical test, the paper presents a case study where this method was applied for comparing inspection performance from two inspectors.

Key-Words: data, measurement system, inspection performance, hypothesis testing, McNemar statistical test.

1 Statistics in quality control work

Statistical concepts have many uses in controlling and improving quality. For example, data on the number of customer complains and returns in a department store may be gathered and described using statistical methods as a basis for future decisions. Statistical methods may be used to determine the capability of machines to produce parts consistently so that a critical dimension (one that will prevent performance of the function of the product) will fall within a specified range, thus ensuring that parts may be properly assembled with mating parts produced elsewhere. In food processing, statistical methods are used to ensure that filling machines do not put too much or too little product in the containers passing along the line.

To collect data on which to assess the current state of quality and to make decisions, some type of inspections and measurements are necessary, because they form the foundations for both product and process control. The data may be dimensions of bolts being produced on a production line, order entry error per day in an order entry department or number of flight delays per week at an airport.

Raw data such as the individual lengths of bolts does not provide information necessary for quality control or problem solving. Data must be organised, analysed and interpreted in a meaningful fashion and statistics provides an efficient and effective way of obtaining meaningful information from data.

In figure 1 [3] are summarised the statistical processes and methods commonly used in quality assurance.

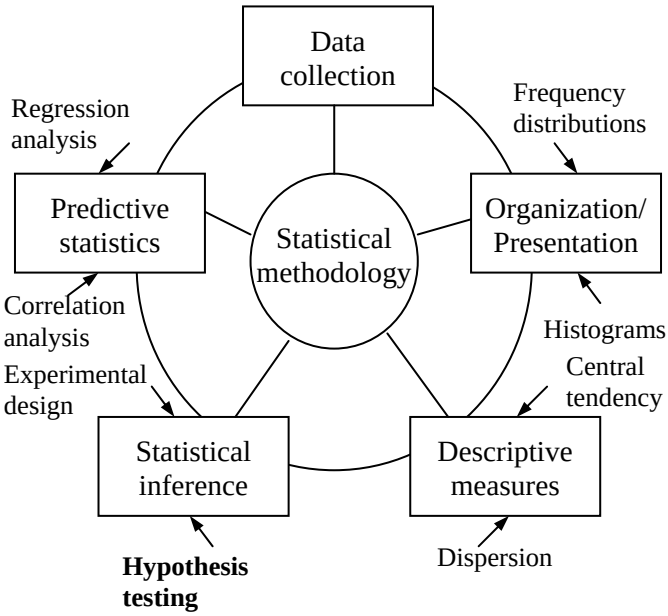

Fig.1 [3]

Drawing conclusions from a process data involves an assessment of the confidence one has in statistics generated form these data. Probability is a useful tool used in qualifies the degree of confidence associated with process statistics.

Because measurement data are used more often and in more ways than ever before, statistical methods and applied probabilities are the bases to the understanding and implementation of quality

assurance; their purpose is to assist managers, supervisors and operators in controlling and improving the quality of products.

2 Elements of a measurement system

The benefit of using a data based procedure is largely determined by the quality of the measurement data obtained from a measurement system. A measurement system is the collection of instruments or gages, standards, operations, methods, fixtures, software, personnel, environment and assumptions used to quantify a unit of measure or fix assessment to the feature characteristics being measured. From this definition is follows that a measurement system may be viewed as a manufacturing process that produces numbers (data) for its output. Viewing a measurement system this way is useful because it allows us to bring to bear all the concepts, philosophy and tools that have already demonstrated their usefulness in the area of statistical process control.

Measurement systems are used every day in manufacturing, research and development, sales and marketing. Measurement systems are essential to the quality of a manufacturing process, characteristics that may be measured include distances, temperatures, strengths, sales, hardness, sweetness, electrical resistance, frequency, viscosity just to name a few.

Similar to all processes, the measurement system is impacted by both random and systematic sources of variation. These sources of variation are due to common and special (chaotic) causes. Although the specific causes will depend on the situation, a general error model can be used to categorise sources of variation for any measurement system. A useful model for defining a measurement system by its basic sources of variation is represented by the acronym P.I.S.M.O.E.A. from table 1 [10]. It is not the only model, but does support universal application.

Whenever there is an apparent gauging problem, examining these elements, and their characteristic influencing factors that may cause measurement errors, is a good approach to troubleshooting.

3 Problem formulation

As we mentioned, in order to understand, control and improve a measurement system, the potential sources of variation ought to first be identified and then, eliminated (whenever possible) or monitored.

Since quality decisions are based on inspection (the judging of a product's conformance to specifications with feedback on the quality provided to the producer), undesirable consequences may result if this task is not performed properly. Consequently, one of the error sources may be the inspector, so the inspection capability must be evaluated. In the paper, the term inspector refers to any inspection instrument (human or otherwise) whose evaluation of products and processes is given in terms of attribute measurements (count data). Traditionally, human inspectors have been responsible for generating most attribute data, but with the increasing dependence on ATE (automatic test equipment), much inspection data are now machine generated.

Two questions arise when one compares the capabilities of two or more inspectors. First, is there a significant difference in the performance characteristics of the inspectors ? Second, if a difference does exist, how should the inspection data from both inspectors be combined to give a better view of the process ?

3.1 Testing for differences between two inspectors-paired data

The paper deals with the first question and applies a statistical method for comparing inspection performance from two inspectors.

One approach to test for difference between two inspectors is to simply select any group of "n" production items and to ask both inspectors to examine them. This way, both inspectors can be compared since they examine the same group of "n" items. Suppose that, these "n" production items are first submitted to inspector 1 and then the same "n" items are submitted to inspector 2. Figure 2 shows a convenient format for displaying the results of this inspection. For any item, the inspectors are said to "agree", if they both rate the item as conforming (T) or if both rate it is non-conforming $(\overline{T})$. Otherwise, they are said to "disagree" on the item.

So, the notation (T) and $(\overline{T})$ is used to denote the inspector's decision:

- (T) means that the inspector classifies the item as conforming and
- $(\overline{T})$ indicates a classification as non-conforming.

Furthermore, the direction of the disagreements can be either fairly evenly split between $(\overline{T},T)$ and $(T,\overline{T})$ or heavily one-sided (e.g., if one inspector consistently classifies items as T, while the other

Table 1 [11]

Error source		Alias or Component	Factor or Parameter
P	**Part**	Production part, sample, measurand, Unit Under Test (UUT), artefact, check standard	Unknown
I	**Instrument**	Gage, unit of M&TE, master gage, measuring machine, test stand	Means of comparison
S	**Standard**	Scale, reference, artefact, check standard, intrinsic standard, consensus, Standard Reference Materials (SRM), class, acceptance criteria	Known value accepted as "truth" (actual or physical "true value" are unknown), reference value or acceptance criteria
M	**Method**	On-the-job training, verbal, work instruction, control plan, inspection plan, test program, part program	How
O	**Operator**	Appraiser, calibration or test technician, assessor, inspector	Who
E	**Environment**	Temperature, humidity, contamination, housekeeping, lighting, position, vibration, power, Electromagnetic Interference, noise, time, air	Conditions of measurement, noise
A	**Assumptions**	Statistical, operational, calibration, constants, handbook values, thermal stability, modulus of elasticity, laws of science	Criteria, constant or supposition for reliable measurement

Item	**Inspector 1 classifies item as:**	**Inspector 2 classifies item as:**	
1	(T)	(T)	← Inspectors "agree"
2	(T)	($\overline{T}$)	← Inspectors "disagree"
3	($\overline{T}$)	($\overline{T}$)	← Inspectors "agree"
.	.	.	.
.	.	.	.
n	($\overline{T}$)	(T)	← Inspectors "disagree"

		Inspector 2		
		(T)	($\overline{T}$)	Inspector 1 totals
Inspector 1	(T)	a	b	a + b
	($\overline{T}$)	c	d	c + d
	Inspector 2 totals	a + c	b + d	n

(T) - inspector classifies the item as conforming

($\overline{T}$) - inspector classifies the item as non-conforming

a – the number of agreement of the form (T, T)

d – the number of agreement of the form ($\overline{T}$, $\overline{T}$)

b – the number of disagreement of the form (T, $\overline{T}$)

c – the number of disagreement of the form ($\overline{T}$, T)

Fig.2

classifies them as $\overline{T}$ whenever they disagree).

The following procedure provides a test for the second of these forms of disagreement. In statistical terms, this procedure compares two sets of paired attributes data (the same items are examined by both inspectors) and in the literature, goes under the name of the McNemar test of correlated proportions or test of "change" [1], [2], [4], [5], [6], [7], [8], [9], [11], [12], [13], [14], [15], [16].

➔ For a set of "n" items inspected by each of two inspectors, count the number of pairs $(T,\overline{T})$ and $(\overline{T},T)$ in which the inspectors disagreed on the classification of an item. Denote these numbers by "b" and "c" respectively.

➔ Calculate the chi-square statistic χ^2, using the more common formula for McNemar test:

$$\chi^2 = (b-c)^2/(b+c) \qquad (1)$$

and conclude, at significance level "α", that the inspectors disagree in the direction of their classification if $\chi^2 \rangle \chi^2_\alpha$, where χ^2_α is the upper $\alpha\cdot(100\%)$ point of the chi-square distribution with 1 degree of freedom.

Some authors recommend a version of the McNemar test with a correction for discontinuity, calculated as:

$$\chi^2 = (|b-c|-1)^2/(b+c) \qquad (2)$$

but this is controversial [11], [12], [14].

3.2 Case study

Human visual inspection of solder joints on printed circuit boards (PCBs) can be very subjective. Part of the problem stems from the numerous ways in which a joint can be non-conforming, e.g. pad non-wetting, knee visibility, voids to name but a few and the degree to which a joint may exhibit any of these problems. Consequently, even highly trained inspectors tend to disagree when examining the same PCB.

The accompanying data (figure 3) show the results of two inspectors who examined a number of n=233 solder joints for the particular problem of "pad non-wetting".

To test for a possible difference in the inspector work, the chi-square statistic was calculated:

$$\chi^2 = \frac{(b-c)^2}{(b+c)} = \frac{(3-11)^2}{(3+11)} = \frac{-8^2}{14} = \frac{64}{14} = 4{,}57 \qquad (3)$$

From the chi-square distribution table, for a significance level of $\alpha = 0{,}05$ and 1 degree of freedom, we obtained the critical value χ^2_α=3,841. Since $\chi^2 = 4{,}57 \rangle \chi^2_\alpha = 3{,}841$, for these data it can be concluded that there exists a statistically difference in the disagreements between the inspectors, but not a significant one.

In particular, disagreement of the $(\overline{T},T)$ form occur more often than those of the form $(T,\overline{T})$; that is, when the inspectors disagree, it is usually because inspector 1 thinks an item is non-conforming while inspector 2 believes it is conforming. This could mean that inspector 1 is using a more stringent definition of non-conforming than inspector 2.

4 Conclusion

When we base decisions on production data, the accuracy of that data is critical. If our measurement system has too much variation, it will affect our ability to make the right decision. Examining measurement system elements that may cause measurement errors, is a good approach to troubleshooting.

Because one of the error sources may be the inspector, the inspection performance must be evaluated and statistical methods are useful tools for the reduction of variation and in controlling and improving quality.

		Inspector 2		
		(T)	$(\overline{T})$	Inspector 1 totals
Inspector 1	(T)	23	3	26
	$(\overline{T})$	11	196	207
	Inspector 2 totals	34	199	233

Fig.3

Since it is very unlikely that two human inspectors will always produce identical inspection results, testing for differences between inspectors often only confirms this intuitive fact.

It can be concluded that this type of statistical analysis is usefully to test if there is a significant difference in the disagreements between the inspectors and consequently a certain inspector may be a source of variation in the measurement system. Remedial action, e.g. human inspector training, is a necessity in cases where large differences exist between inspectors.

References:

[1] A. Agresti, *Categorical data analysis*, John Wiley & Sons, New York, 1990.

[2] M. L. Berenson, D. M. Levin, *Statistics for Business and Economics*, Prentice Hall, New Jersey, 1990.

[3] J. R. Evans, W. M. Lindsay, *The Management and Control of Quality*, West Publishing Company, St. Paul, MN, 1993.

[4] B. S. Everitt, *The analysis of contingency tables*, Chapman & Hall, London, 1977.

[5] N. R. Farnum, *Modern Statistical Quality Control and Improvement*, Wadsworth Publishing Company, Belmont, California, 1992.

[6] J. Fleiss, *Statistical Methods for Rates and Proportions*, 2nd edition, John Wiley & Sons, New York, 1981.

[7] W. L. Hays, *Statistics*, 5th edition, Harcourt Brace, Orlando, 1994.

[8] G. V. Glass, K. D. Hopkins, *Statistical Methods in Education and Psychology*, 3rd edition, Allyn and Bacon, Needham Heights, 1996.

[9] J. R. Levin, C. S. Ronald, Changing Students' Perspectives of McNemar's Test of Change, *Journal of Statistics Education*, Vol.8, No.2, 2000, found at http://www.amstat.org/publications/jse/secure/v8n2/levin.cfm

[10] *Measurement Systems Analysis Reference Manual*, 3rd edition, Daimler Chrysler Corporation, Ford Motor Company, General Motors Corporation, 2002.

[11] W. Mendenhall, T. Sincich, *Statistics for engineering and the sciences*, 4th edition, Prentice Hall, Englewood Cliffs, NJ, 1995.

[12] D. J. Sheskin, *Handbook of parametric and non-parametric statistical procedures*, 2nd edition, Chapman & Hall, Boca Raton, 2000.

[13] S. Siegel, J. J. Castellan, *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, New York, 1988.

[14] http://ourworld.compuserve.com/homepages/jsuebersax/mcnemar.htm

[15] http://www.biostat.umn.edu/~john-c/5421/n54703.003

[16] http://www.angelfire.com/wv/bwhomedir/notes/categorical.pdf