20th IMEKO TC4 International Symposium and
18th International Workshop on ADC Modelling and Testing
Research on Electric and Electronic Measurement for the Economic Upturn
Benevento, Italy, September 15-17, 2014

Repeatability and Reproducibility Techniques for Measurement System Analysis

Marcantonio Catelani, Lorenzo Ciani, Bianca Sereni and Andrea Zanobini

[1] *Department of Information Engineering, University of Florence,*
via S. Marta 3, 50139, Florence, Italy
marcantonio.catelani@unifi.it, lorenzo.ciani@unifi.it, andrea.zanobini@unifi.it

***Abstract* – This research was conducted with the aim of analyzing two of the main metrological characteristics of any measurement system: Repeatability and Reproducibility. Both of these features play an important role in the analysis of the measurements and they can give us a lot of information about who and what influences any measuring system. The analysis of Repeatability and Reproducibility is generally carried out through the use of the study Gage R&R. This study is very useful because it permits us to understand which are the decisive factors in a measurement system, and, definitively, if the process is stable, that is under statistical control or out of statistical control.**

I. INTRODUCTION

The statistical control [1-2] of processes consists in a set of techniques of analysis concerning the quality of products and services, and in this particular case, of the measures. To define the concept of quality is not simple, in one of its definitions it is inversely proportional to the variability: in fact a decrease of quality corresponds to an increase in variability.

Many statistical tools can be used in the analysis of the measurement system [3-9]. In particular, the SPC or Statistical Process Control is a process of analysis of variability, or rather to its reduction; and it uses some methods or techniques such as the Gage R&R to achieve that. The Gage R&R is a study on the variability observed in a measurement and due to the measurement system itself. R&R denotes Repeatability and Reproducibility, that are two characteristics of each measurement system.

In particular, Repeatability is the variation caused by the instrumentation or the variation observed when the same operator measures the same part more times with the same instrumentation.

Reproducibility is the variation caused by the measurement system or the variation observed when different operators measure the same part with the same instrumentation.

A small variability of a series of measurements is a good indicator of repeatability, meantime the reproducibility is colligated to the stability of a measurement process. The ANOVA or ANalysis Of VAriance [10-12] and the DOE or Design Of Experiment are very powerful methods to conduct a study Gage R&R. The Gage R&R studies determine how much of variability of processes is due to the variation of the measurement system and they uses technique as ANOVA to estimate Repeatability and Reproducibility.

II. ANOVA

The *ANOVA* consists in a series of techniques originating by the inferential statistics theory, that are interested in the comparison of data variability [13]. These methods are used when there are two or more populations to estimate the differences between their sample means; analyzing the respective variances to achieve its purpose. By the evaluation of two or more different distributions, *ANOVA* allows to determine if the differences are random or not.

ANOVA is based on a technique that compares the sample means to estimate the data variation. It obtains this aim decomposing the variability in *between* and *within*.

ANOVA is a process of statistical inference, in particular it is a technique of parametric statistical inference based on an hypothesis test.

If we suppose to have only a factor characterized by a levels or treatments and n observations for each level we must consider that the answer to each of the a levels is a random variable. The observed data can be represented in the following table:

ISBN-14: 978-92-990073-2-7

Table 1: Table of data detection

LEVELS	*OBSERVED VALUES*				*TOTAL*	*EXPECTED VALUES*
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\overline{y}_{2.}$
…	…	…	…	…	…	…
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\overline{y}_{a.}$

In general the observations can be described through a statistical linear model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases}$$

Considering that $\mu_i = \mu + \tau_i$, this model becomes:

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases}$$

Each level can be considered belonging to a population with mean μ_i and variance σ^2 .

As we have just said, the experiment must to be completely randomized, or rather the observations are extracted in a completely random way.

So the analysis of variance consists to perform the following hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a = 0$$
$$H_1 : \mu_i \neq 0$$

The analysis of variance checks if the means of the *a* populations are equal.

It is possible to consider τ_i as the first order deviation by the general mean μ , so:

$$\sum_{i=1}^{a} \tau_i = 0 = \sum_{i=1}^{a} (\mu_i - \mu)$$

Consequently trying the equality between the means is trying the equality of levels effects. So the hypothesis test can be written as:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1 : \tau_i \neq 0$$

Moreover the following relationship are valid:

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \qquad \overline{y_i} = \frac{y_{i.}}{n}$$

$$y. = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \qquad \overline{y} = \frac{y.}{N}$$

with: *N=an* and *i=1,2,…,n*

Regarding the total variability of the samples, we can decompose in a variability due to the treatments and in a variability relative to the errors. The variability can be described by the sums of the squares of the deviations of the average values, then we can write [1]:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}\right)^2 = n\sum_{i=1}^{a}\left(\overline{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_i\right)^2$$

We can therefore write:

$$SS_T = SS_{Treatments} + SS_E$$

Where:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}\right)^2$$

$$SS_{Treatments} = n\sum_{i=1}^{a}\left(\overline{y}_i - \overline{y}\right)^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_i\right)^2$$

Considering the expected values we can write the following equality:

$$E(SS_{Treatments}) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$

If the null hypothesis is true, $\tau_i = 0$, so:

$$E\left(\frac{SS_{Treatments}}{a-1}\right) = \sigma^2$$

While it is true the alternative hypothesis:

$$E\left(\frac{SS_{Treatments}}{a-1}\right) = \sigma^2 + \frac{\sum_{i=1}^{a}\tau_i^2}{a-1}$$

Then we can define:

$$MS_{Treatments} = \frac{SS_{Treatments}}{a-1}$$

And:

$$MS_E = \frac{SS_E}{a(n-1)}$$

In the final analysis, if we assume, as said, that each of the a populations is described by a normal distribution, if the null hypothesis is true, then the ratio:

$$F_0 = \frac{SS_{Treatments}/(a-1)}{SS_E/[a(n-1)]} = \frac{MS_{Treatments}}{MS_E}$$

has a Fisher distribution with *(a-1)* and *a(n-1)* degrees of freedom.
In conclusion, the answer to the hypothesis testing will be based on the *p-value*: observing the *p-value* relative to the value of *F*, obtained from statistical tables or software, and comparing it with the level of significance α, we can decide whether to accept or reject the null hypothesis. In particular, we accept the null hypothesis if the *p-value* is greater than the significance level, while we accept the alternative hypothesis if the *p-value* is less than α. Refusing, or rather not accept the null hypothesis means that we can conclude that among the observed data there may be significant differences.

Intuitively it can be stated that, if the null hypothesis is true, so the data differ for the effect of random factors, and the changing the levels of the factor does not change the response, but if the null hypothesis is false means that the variability total phenomenon can be attributed to systematic factors. So the method allows to value if the difference between the variances of two distributions, namely the total variability, is due to chance or is significant.

III. GAGE R&R

The Gage R&R is a study on the ability of a measuring system. The *Gage R&R* techniques are used to determine how much of the variability of the processes observed is due to the variability of the measurement system. To do this, the study is able to isolate the various components of variability and estimate them individually, with the aim to assess whether the measuring system is capable or less (a measurement system is defined *capable* when it is appropriate to application which is predetermined). The purpose of *Gage R&R* is to verify that a measurement system is acceptable. The *Gage R&R* studies the accuracy, parameter which indicates the variability of a measurement and that can be decomposed into Repeatability and Reproducibility. The total variability can be, according to this method, divided into:

$$\sigma^2_{total} = \sigma^2_{measurement\ system} + \sigma^2_{repeatability} + \sigma^2_{reproducibility}$$

Consider a instruments and b operators, both randomly selected. Suppose further that each operator measures each part for n times. We assume that i = instrument, j = operator and k =measurement.
Then, similar to the ANOVA method, the measurements can be represented by the following model:

$$y_{ijk} = \mu + I_i + O_j + (PO)_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,b \\ k = 1,2,\cdots,n \end{cases}$$

This is a standard model for a *Gage R&R* experiment, but it is only a model of variability analysis with casual effects.
If we hypothesize that $I_i, O_j, (PO)_{ij}, \varepsilon_{ijk}$ have a distribution with null mean and variances defined as

$$Var(I_i) = \sigma_I^2 \qquad Var(O_j) = \sigma_O^2 \qquad Var(IO_{ij}) = \sigma_{IO}^2$$

and:

$$Var(y_{ijk}) = \sigma_I^2 + \sigma_O^2 + \sigma_{IO}^2 + \sigma^2$$

Now we can estimate the components $\sigma_I^2, \sigma_O^2, \sigma_{IO}^2$ and, to do this, we can apply the methods of analysis of variance. This implies, however, the decomposition of the total variance of the measurement:

$$SS_{Total} = SS_{Instrumentss} + SS_{Operators} + SS_{PO} + SS_{Error}$$

Dividing each sum of squares by its degrees of freedom we get:

$$MS_I = \frac{SS_{Instruments}}{a-1}$$
$$MS_O = \frac{SS_{Operators}}{b-1}$$
$$MS_{PO} = \frac{SS_{PO}}{(a-1)(b-1)}$$
$$MS_E = \frac{SS_{Error}}{ab(n-1)}$$

and:

$$E(MS_O) = \sigma^2 + n\sigma_{IO}^2 + bn\sigma_I^2$$

$$E(MS_O) = \sigma^2 + n\sigma_{IO}^2 + bn\sigma_O^2$$
$$E(MS_{PO}) = \sigma^2 + n\sigma_{IO}^2$$
$$E(MS_E) = \sigma^2$$

Using the software Minitab it is possible to perform the analysis of variance; the components of variance can be estimated equating the means with their expected values and solving in the variance components. Obtaining:

$$\hat{\sigma}^2 = MS_E\,, \qquad \hat{\sigma}_{IO}^2 = \frac{MS_{IO} - MS_E}{n}\,,$$
$$\hat{\sigma}_O^2 = \frac{MS_O - MS_{IO}}{an}\,, \qquad \hat{\sigma}_I^2 = \frac{MS_I - MS_{IO}}{bn}$$

We generally consider that the *Repeatability* of the variance corresponds to the component of variance due to the random error σ^2, while the *Reproducibility* is due to the sum of the components of the variance of operator and operator by parts: $\sigma_O^2 + \sigma_{IO}^2$. So:

$$\sigma_{Instrument}^2 = \sigma_{Repeatability}^2 + \sigma_{Reproducibility}^2$$

And the estimate is:

$$\hat{\sigma}_{Instrument}^2 = \hat{\sigma}^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2$$

IV. A PRATICAL EXAMPLE

The test phase configurations use the oscilloscope *LeCroy SDA 6000* to measure 10 values of both rise time and fall time by setting a sampling frequency of 10GS/s. We have used two different waveform generators to generate the signal: *Agilent 33220A* and *Tektronix AWG 420* [14].

Starting with the analysis of the **Two-Way ANOVA** data, we can look the values of the *p-value*. Considering the confidence interval of 95%, its level of significance is α = 0.05. Considering this value, we can analyze the values of *p-value* there are in the table:

- the *p-value* relative to components (rise / fall) is 0.542, so greater than α. Then in this case we can say that it is reasonable to accept the null hypothesis, ie that there are no significant differences between the rise and fall times.
- The *p-value* relative to operators (*generator*) is 0.008, so less than α. In this case is therefore permissible to reject the null hypothesis in favor of the alternative and affirm that there are significant differences in the use of two different waveform generators.
- The *p-value* relative to components and operators interaction (*generatore*rise/fall*) is null, ie 0.000 and that means that we reject the null hypothesis and that the interaction between the two factors influence the quality of the measurements of times.

Analyzing the table *Gage R & R* we can pay attention on the columns for *% Contribution* and *%StudyVar*. The first indicates which are the percentages of contribution of each source of variation of the process, while the second is the standard deviation expressed as a percentage and relative to the various sources of process variation.

- for *%Contribution*:
  - the value of *Repeatability* is 0.02%, so <1% and this means that the measurement system is acceptable, in fact measuring rise and fall times by the same generator there were not significant differences.
  - the value of *Reproducibility* is instead very high, ie 99.98%, that is >9%, so the measurement system is inacceptable. This happens because measuring rise and fall times by different generators there are significant differences. In fact this value is due, for 99.95%, by the used operator.
- for *%StudyVar* we can observe the same results (thought the threshold values change):
  - the value of *Repeatability* is 1.44%, so <10% and this means that the measurement system is acceptable, in fact measuring rise and fall times by the same generator there were not significant differences.
  - the value of *Reproducibility*, as the previous case, is instead very high, ie 99.99%, that is >30%, so the measurement system is Inacceptable. This happens because measuring rise and fall times by different generators there are significant differences. In fact this value is due, for 99.97%, by the used operator.

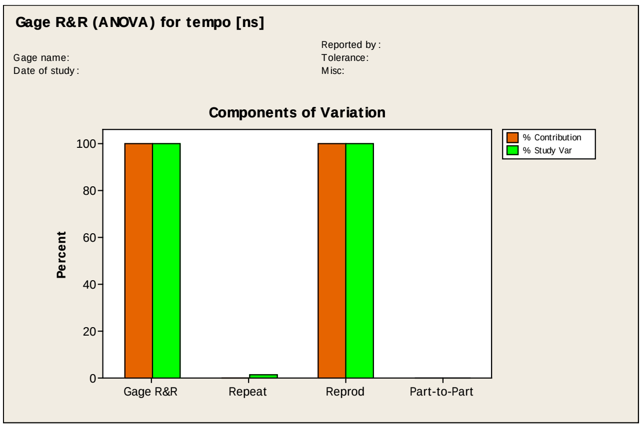

Figure 1. Graph Components of Variation

The graph ***Components of Variation*** (Figure 1) is an histogram that express graphically that is present in the columns *%Contribution* e *%StudyVar* of *ANOVA* table. In fact it shows the percentages of contribution and standard deviation relatives to various sources of variation. As we have analyzed previously this graph show that the greater part of total variation of measurement system is due by reproducibility and, in general to the measurement system itself.

So the new measurement system adds a lot of variability to the overall variation (over 90%) and the fact that also the reproducibility is very high, it means that the operators, or the generators, they measure the components, ie, the rise and fall times, in a very different way. This graph reveals that the variation of the measurement system is the largest source of variability observed in the experiment, with different operators (*reproducibility*) as major contributors to the variation.

V. CONCLUSIONS

The Analysis of Variance (ANOVA) and the Gage R&R (Repeatability & Reproducibility) techniques, little known if compared to the other existing, are essentially the only ways to identify any problem in a measurement system. In a Gage R&R Study, measurements of one part measured by one operator are analyzed to estimate the level of variation, the gage repeatability, and the accuracy of the measurements. By analyzing the measurements of one reference part by one operator, it is always possible to determine whether a measuring device is capable of measuring a particular characteristic under conditions with relatively small variation.

In fact any time we want to measure the results of a process we will see some variation.

This variation comes from two sources:

1) There are always differences between parts made by any process

2) Any method of taking measurements is imperfect - thus, measuring the same part repeatedly does not result in identical measurements.

Even though in a Gage R&R study it is important that the measurements are done in a random order and that the parts selected provide a representative sample across the possible range of responses, the ANOVA methods, that form the basis of Gage R&R studies, determine what portion of the variability in measurements may be due to the measurement system.

We conclude with a sentence of one of the greatest of all statisticians who defined our problem for us: "A statistics is a value calculated from an observed sample with a view to characterizing the population from which is drawn (Fisher R.A., *The Design of Experiments*. (6th ed.). New York: Hafner, 1950). The "observed sample" of which Sir Ronald Fisher speaks is a part of the membership of a total set of elements of which we desire knowledge. This total set of elements is called "population". The point of Fisher's definition is that the sample mean is an estimate of the population mean, and its uncertainty in that role is a function of the sample size and of the variability on the measured characteristics of the elements of the population, and of the care with which the sample members are selected from population. The many subtleties which can be introduced into the design and analysis of experiments are treated in the ANOVA, one of the greatest of Fisher's inventions, where is possible to partition variance in outcomes of an experiment into many independent components assigned to different influences or sources. In the subject of this multivariate analysis the interrelationships among numerous variables are taken into consideration with the powerful and essential computer support.

REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control*, 6th Edition, 2009, Wiley.

[2] W.Navidi, *Statistics for Engineers and Scientists*, McGraw Hill, NY, USA, 2007.

[3] M.Catelani, L.Ciani, A.Zanobini, "Uncertainty Interval Evaluation Using The Chi-Square And Fisher Distributions In The Measurement Process", *Metrology and Measurement Systems*, Vol. XVII (2010), No. 2, pp. 195-204.

[4] E. Balestrieri, M. Catelani, L. Ciani, S. Rapuano, A. Zanobini, "Word Error Rate measurement uncertainty estimation in Digitizing Waveform Recorders", *Measurement* 46 (2013), pp. 572-581, DOI 10.1016/j.measurement.2012.08.016.

[5] E. Balestrieri, M. Catelani, L. Ciani, S. Rapuano, A. Zanobini, "The Student's t distribution to measure the word error rate in analog-to-digital converters", *Measurement*, Volume 45, Issue 2, February 2012, Pages 148-154, ISSN 0263-2241, DOI 10.1016/j.measurement.2011.03.002.

[6] A. Zanobini, L. Ciani, M. Catelani, "Hypothesis test on the ratio between two measurement uncertainties", *Proc. Of 19th IMEKO TC4 Symposium - Measurements of Electrical Quantities 2013 and 17th International Workshop on ADC and DAC Modelling and Testing*, pp. 400-404.

[7] M.Catelani, L.Ciani, S.Giovannetti, A.Zanobini, "Uncertainty Analysis in High-Speed Multifunction Data Acquisition Device", *Proc. Of 2011 IMEKO IWADC & IEEE ADC FORUM*, Orvieto (PG), pp. 260-265.

[8] M. Catelani, L. Ciani, A. Zanobini, "Some remarks on a bivariate analysis in the propagation of measurement uncertainty as an alternative approach to the Monte Carlo method", *Proc. Of 17th Symposium IMEKO TC 4 on Instrumentation for the*

ICT Era, Sept. 8-10, 2010, Kosice, Slovakia, pp. 288-292.

[9] M.Catelani, A. Zanobini, L.Ciani, “Introduction to the t and chi-square distribution for a more accurate evaluation of the measure of the Word Error Rate in Analog-to-Digital Converters”, *Metrology and Measurement System*, Issue 4/2008, vol. XV, Number 4 (2008), pp. 483-488, ISSN: 08608229.

[10] M. D’Arco, A. Liccardo, N. Pasquino, "ANOVA-Based Approach for DAC Diagnostics," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 7, pp. 1874-1882, July 2012.

[11] Abbas Al-Refaie, Nour Bata, Evaluating measurement and process capabilities by GR&R with four quality measures, *Measurement*, Volume 43, Issue 6, July 2010, Pages 842-851, ISSN 0263-2241, http://dx.doi.org/10.1016/j.measurement.2010.02.016.

[12] Binnur Sagbas, M. Numan Durakbasa, Modelling and optimization of biotribological testing parameters on frictional heating of vitamin E blended UHMWPE with the help of statistical techniques, *Measurement*, Volume 54, August 2014, Pages 249-255, ISSN 0263-2241, http://dx.doi.org/10.1016/j.measurement.2014.01.022.

[13] M. Catelani, L. Ciani, B. Sereni, A. Zanobini, "ANOVA Techniques for Reliability Analysis", *Proc of 13th IMEKO TC10 Workshop on Technical Diagnostics - "Advanced measurement tools in technical diagnostics for systems' reliability and safety"*, June 26-27, 2014, Warsaw, Poland, pp.201-205.

[14] E.Balestrieri, M.Catelani, L.Ciani, S.Rapuano, A.Zanobini, “Uncertainty evaluation of DAC Time Response -Parameters”, *Proc. Of 2011 IMEKO IWADC & IEEE ADC FORUM*, Orvieto (PG), pp.278-281.