sciendo QPI 2019, volume 1, issue 1, pp. 472-478

EFFECT OF NON-FULFILLMENT OF ASSUMPTIONS ON GAGE REPEATABILITY AND REPRODUCIBILITY STUDY EVALUATION

doi: 10.2478/cqpi-2019-0064
Date of submission of the article to the Editor: 29/04/2019
Date of acceptance of the article by the Editor: 17/05/2019

Pavel Klaput[1] – *orcid id: 0000-0003-2491-2277*
David Vykydal[1]
Jiří Plura[1]
[1]VSB-TU Ostrava, Department of quality management, 17. listopadu 2172/15 **Czech Republic**

Abstract: The evaluation of the measurement system quality has already become an integral part of quality planning activities in both the automotive and metallurgical industries. An important assumption for obtaining the most relia ble results is compliance with the basic assumptions for evaluating the variability of the measurement system. The main goal of this paper is to analyze, how the failure to meet the basic assumptions influences the evaluation of the measurement system's statistical properties. This goal is achieved by performing a detailed analysis of the latest developments in the field of measurement systems analysis aimed at verifying the assumptions of normality and uniformity. The evaluation of the effect of non-fulfillment of both assumptions on the values of the most important statistical properties of the measurement system is performed using simulated data. Suitable graphical tools are used for practical verification of both assumptions.
Keywords: normality, uniformity, repeatability, reproducibility

1. INTRODUCTION

The MSA methodology is the most used in the practice, it was created by the trinity of largest American automotive companies Chrysler Group LLC, Ford Motor Company and General Motors Corporation within standards QS 9000 (AIAG, 2010). The fundamental part of this methodology is made by gauge repeatability and reproducibility study (GRR). Average and range method (A&R) is most commonly used for GRR assessment in the practice. We can also use the evaluation by analysis of variance (ANOVA), but it requires an appropriate software and is more difficult to interpret (Mikulová and Plura, 2018). After the all indexes evaluation it is necessary to make the conclusion on the measurement system acceptability, based on the percentage share of the measurement repeatability and reproducibility of the total variability (%GRR) and on the number of distinct categories that can be discerned by the measurement system (ndc). Three situations may occur, as described in the table 1.

Table 1
GRR study acceptance criteria

%GRR < 10% and ndc ≥ 5	the measurement system is acceptable
10% < %GRR < 30% and ndc ≥ 5	the measurement system is conditionally acceptable owing to global variability of the process or the tolerance range, and it depends on the proportion of the remedy cost and importance of the quantity monitored.
%GRR > 30% or ndc < 5	the measurement system is unacceptable and it must be improved

Automotive industry suppliers are under increasing pressure to select more sophisticated methods, providing more detailed information about the analyzed measurement system. For this reason, it is necessary to specify the assumptions that are associated with these methods (Tošenovský, 2018). The basic assumptions are normal probability distribution of measured data, homogeneity of variance (uniformity) and Measurement independence (no autocorrelation). In general, every verification of data assumptions must consist of a numerical part, ie. testing of statistical hypotheses and consequently graphical analysis. The following part of this paper is focused on simulation of assumptions that lead to failure of the assumption of normality and homogeneity. Modified practice data (8 real GRR studies) will be evaluated in a standard way using both above-mentioned methods. It will be also found out, how these unfulfilled assumptions are reflected in the GRR analysis outputs (Sinay, 2018).

2. MEASURED DATA NORMALITY

The following was the procedure to simulate the failure to assume normality. For each study, the same range remained and there was only a change in the averages of the samples measured by the operators. Therefore, the first change involved adding more values of repeated measurements. For most studies, three operators were chosen to measure each sample three times. To increase the explanatory power of all studies, the number of measurements of each sample was increased to ten by each operator (Petrík, 2016). As a result, the new data set no longer came from the normal distribution. Since the normality of the data was not evaluated from the measured values, but from the deviations from the average of the measurements of the given samples by the individual operator, the remaining values had to be calculated. The results of the normality test for the simulated data (study 1) are shown in Figure 1.

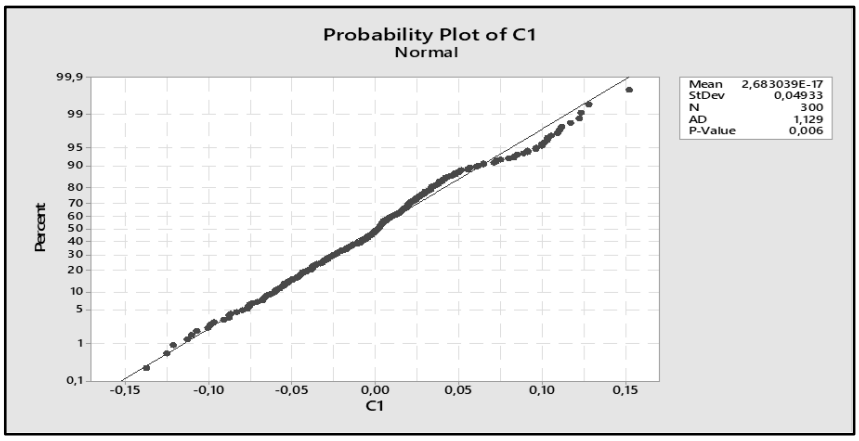

Fig. 1. Normally probability graph

The p-value (Fig.1) indicates that the null hypothesis can be rejected in favor of an alternative hypothesis, which means that this assumption of normality has not been met. In the next step, the simulated data was evaluated according to the average and range method (A&R) and also according to analysis of variance (ANOVA) method. The resulting values of %GRR and ndc (number of distinct categories) before and after the simulations are shown in Table 2. Minitab 17 and STATGRAPHICS Centurion 17 were used to evaluate data studies (Janošcová, 2012).

Table 2
Result of GRR studies before and after simulation

Study	GRR outputs	Before simulation		After simulation		A&R - %GRR difference	ANOVA -%GRR difference
		A&R	ANOVA	A&R	ANOVA		
Study 1	%GRR	12,62	12,70	7,55	11,16	-40,17 %	-12,13 %
	ndc	11	11	18	12	63,64 %	9,09 %
Study 2	%GRR	18,04	25,84	11,51	21,36	-36,20 %	-17,34 %
	ndc	7	5	12	6	71,43 %	20,00 %
Study 3	%GRR	1,32	1,63	1,02	1,38	-22,73 %	-15,34 %
	ndc	106	86	137	101	29,25 %	17,44 %
Study 4	%GRR	23,48	22,31	19,74	20,30	-15,93 %	-9,01 %
	ndc	5	6	7	6	40,00 %	0,00 %
Study 5	%GRR	73,44	71,75	54,69	61,38	-25,53 %	-14,45 %
	ndc	1	1	2	1	100,00 %	0,00 %
Study 6	%GRR	22,41	21,83	18,27	19,12	-18,47 %	-12,41 %
	ndc	6	6	7	7	16,67 %	16,67 %
Study 7	%GRR	35,07	48,14	31,64	52,79	-9,78 %	9,66 %
	ndc	3	2	4	2	33,33 %	0,00 %
Study 8	%GRR	60,06	66,43	32,34	58,47	-46,15 %	-11,98 %
	ndc	1	1	4	1	300,00 %	0,00 %

Table 2 is supplemented by the percentage differences of %GRR values achieved by each method. Fields that are marked in white indicate ± 5% difference in %GRR values. Lighter gray fields mean ± 15% difference and darker fields indicate 30% or higher

difference. The percentage changes in %GRR values calculated using both methods are better shown in Figure 2.

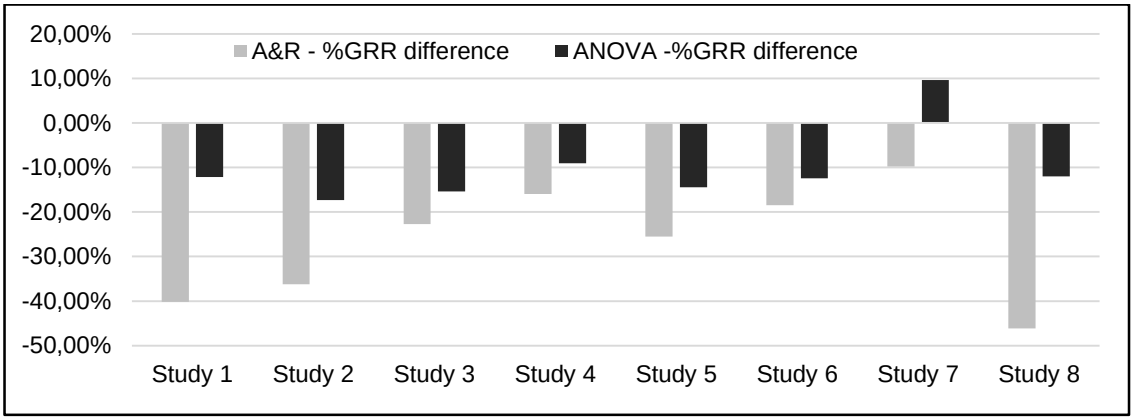

Fig. 2. Percentage difference of %GRR values achieved by different methods

Figure 2, which shows the percentage differences of both methods are shown in terms of %GRR values for all studies. In particular, we can see that the average and range method achieved bigger differences in all studies, compared to ANOVA. Therefore, the statistical method (ANOVA) is more robust (resistant) to failure to meet the assumption of normal distribution. This conclusion clearly demonstrates the direction in which organizations should go in choosing the method to be used for the measurement system evaluation.

3. MEASUREMENT UNIFORMITY

This part describes the results of simulation of insufficient data uniformity. It is therefore an insufficient variation of variance depending on the size of the measured samples. The simulation of failure to fulfill this assumption could not be realized on the basis of the above-mentioned studies with real data. This is because there are many different defects in the individual studies, or the differences in measurement systems, which would result in the mixing of several assumptions and thus a reduction of the possibility of a clear assessment of the effect of insufficient uniformity. For this reasons were simulated such data files, which allow to illustrate, how the evaluation methods of GRR study react just on the assumption of a homogeneous variance.

First, it is necessary to explain how the data simulation was performed. Different ranges (from 0.2 to 2) were simulated on samples with average values of 1 to 10. Thereby, two parameters (the same average range and the average of the measured values) were preserved, through which it is possible to show, how the evaluation methods respond to the different level of uniformity (Pačaiová et al., 2017). In order to ensure that no other influences are involved in this simulation, which would reduce the predictive power of this simulation, all simulations were performed in Minitab 17. The numerical results of the simulations are shown in Table 3.

Table 3
Result of GRR studies for uniformity simulations

Simulation	%GRR		Difference %
	A&R	ANOVA	
Simulation 1	22,37	15,76	41,94
Simulation 2	22,37	16,21	38,00
Simulation 3	22,37	16,86	32,68

Simulation 4	22,37	17,37	28,79
Simulation 5	22,37	17,71	26,31
Simulation 6	22,37	17,72	26,24
Simulation 7	22,37	18,85	18,67
Simulation 8	22,37	19,90	12,41
Simulation 9	22,37	20,70	8,07
Simulation 10	22,37	26,23	-14,72
Simulation 11	22,37	22,52	-0,67
Simulation 12	22,37	24,02	-6,87
Simulation 13	22,37	24,95	-10,34
Simulation 14	22,37	26,23	-14,72
Simulation 15	22,37	45,51	-50,85

Simulation 1 characterizes fully satisfactory uniformity, as all operators had the same range for all samples (Figure 3). In spite of this, the results of the GRR analysis indicators are different for both methods. These differences are caused due to the different way of repeatability calculations. The resulting difference for both methods differs by about 42% in this simulation.

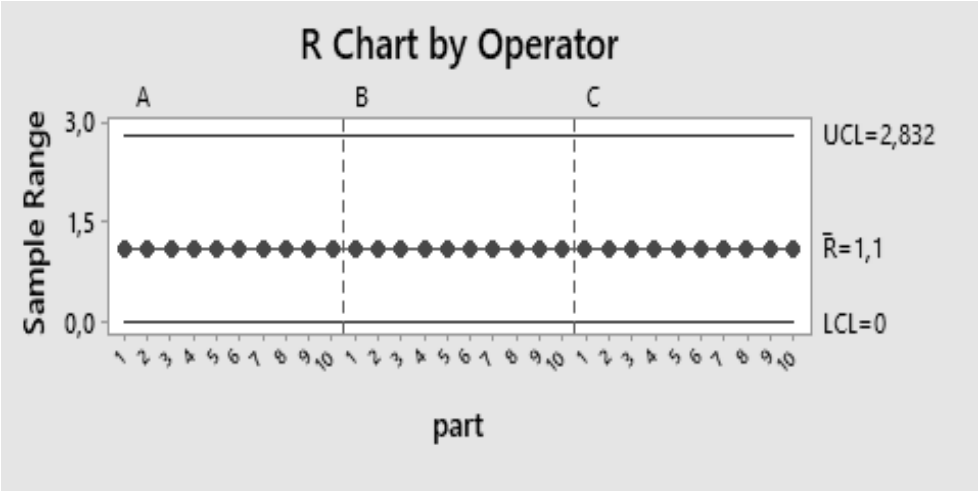

Fig. 3. R-chart for simulation 1

The average and range method is unable to react to the differences of range if uniformity occurs (for our simulation method). Some simulations also show relatively extreme situations in which statistically unstable assumptions have been achieved according to repeated measurements, since the values exceed the control limits and appropriate intervention should be made in this case. However, these simulations are important in terms of the overall impact of insufficient uniformity on GRR study evaluation. An overall summary of all insufficient uniformity simulations is shown in Figure 4.

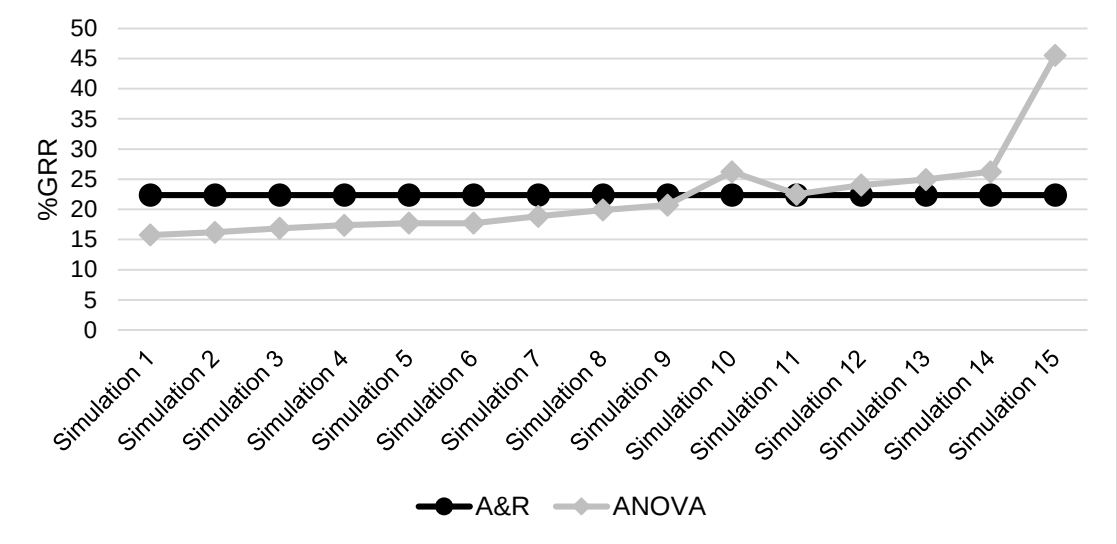

Fig. 4. Insufficient uniformity simulations results

Figure 4 shows that the average and range (A&R) method not registered the signal of increasing variance in dependence on the sample size. On the other hand, the ANOVA method clearly reflected the signal of insufficient uniformity. This was reflected in the increasing value of the %GRR indicator. Overall, the quality of the analyzed measurement system has declined.

4. CONCLUSIONS

A summary of all the above findings has shown that in the measurement system analysis it is appropriate to prefer the ANOVA method, before the average and range method. The average and range method failed mainly when the assumption of data uniformity was not met, where it did not capture the signal, that indicated, that there was a problem in the analyzed measurement system. Furthermore, this method has provided biased results for several studies, although it has been demonstrated by ANOVA that data are unbiased. Also, the analysis of variance method is more robust than the average and range method when the normality assumption is not met. The main difference between these methods cause the fact that the ANOVA method is a statistical method and is much more robust and provides more accurate results in case when data entering the measurement system analysis are affected by various deficiencies. These deficiencies can be easily identified in practice by designing and analyzing the appropriate graphical tools presented in this paper (Noskievičová, 2018).

ACKNOWLEDGEMENTS

This paper was elaborated in the frame of the specific research projects No. SP 2019/62 and SP 2019/129, which has been solved at the Faculty of Metallurgy and Materials Engineering, VŠB-TU Ostrava with the support of Ministry of Education, Youth and Sports, Czech Republic.

REFERENCES

Chrysler Group LLC, Ford Company, General Motors corporation. Measurement Systems Analysis, Reference Manual. 4th ed., 2010.

Janošcová, R., 2012. Evaluation of software quality. IMEA 2012, Hradec Králové, 24-30.

Mikulová, P., Plura, J., 2018. *Comparison of approaches to gauge repeatability and reproducibility analysis,* MATEC Web of Conferences. Volume 183, EDP Sciences.

Noskievičová, D., 2018. *APSS - Software Support for Decision Making in Statistical Process Control,* Quality Innovation Prosperity, 22, 19-26.

Pačaiová, H., Sinay, J., Turisová, R., Hajduová, Z., Markulík, Š., 2017. *Measuring the qualitative factors on copper wire surface*, Measurement, 109, 359-365.

Petrík, J., 2016. *On the Load Dependence of Micro-Hardness Measurements: Analysis of Data by Different Models and Evaluation of Measurement Errors,* Archives of Metallurgy and Materials, 61, 1819-1824.

Tošenovský, F., 2018. *In Search of an Appropriate Market Product Based on the Weighted-Average Multi-Criteria Decision Making Model*, Economic Computation and Economic Cybernetics Studies and Research, 52, 265-278.

Sinay, J., Balážiková, M., Dulebová, M., Markulik, Š., Kotianová, Z., 2018. *Measurement of low-frequency noise during CNC machining and its assessment*, Measurement, 119. 10.1016/j.measurement.2018.02.004.