Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

DESTRUCTIVE R&R STUDY - EVALUATION PROBLEMS

JAROŠOVÁ Eva (CZ)

Abstract. The paper deals with the estimation of variance components which is part of a study to assess the capability of a measurement system. The moment method based on the ANOVA model with random effects can yield negative estimates which leads to ambiguous results. Apart from this, example from a wire drawing laboratory shows how the excessive sample-to-sample variation in a destructive gage study may distort the assessment of the system's capability.

Keywords: variance components, ANOVA, approximate confidence intervals

Mathematics Subject Classification: 62P30, 62-07

1 Introduction

Sufficient quality of measured data is an important prerequisite for their correct evaluation. The measurement quality is influenced both by the measuring instrument used and by the conditions under which the measurements take place, including the operator who performs the measurements. A set of methods known as MSA (Measurement System Analysis) are used to verify the capability of the whole measurement system. Through the repeatability and reproducibility (R&R) study, the variability of measurements is analysed. Repeatability refers to the variation in repeated measurements made on the same part under identical conditions including the operator, while reproducibility refers to the variation in measurements made on the part under changing conditions, such as the measurements taken by different operators. The total variability of the measurement system is then compared to the process variability or to the specification limits and the suitability of the measurement system for a given application is assessed.
Typically, several operators participate in the study, who repeatedly measure several selected parts using the same device. Factorial design is used, the factors being *Part* and *Operator*. Both the experimental design and evaluation methods are described in many publications, see [1], [2], [8], and in the manuals of various statistical software products. The problem occurs if the characteristic measured on the same part changes over time, if its value is influenced by the measurement or if the measurement is destructive. These cases are referred to as non-repeatable or destructive measurements. The basic designs of the experiment for destructive R&R are listed, for example, in [5]. Instead of a part being measured repeatedly, a batch of

samples which are similar in terms of the measured characteristic is used in the experiment. It means that parts in the standard design are substituted with several such batches. The term repeatability becomes somewhat meaningless in this case, but it is still used. Then estimation of the variance representing repeatability reflects not only the variability due to the measuring device itself, but also the differences between samples. Some approaches aiming to reduce the impact of sample-to-sample differences are discussed in [4].
The paper deals with the problems associated with the design and evaluation of the R&R study. Only effects of two factors and their possible interaction are studied. Some methods for estimation of variance components and for constructing confidence intervals are described and the outputs of gage study analyses that are implemented in Minitab and Statgraphics are examined.

2 Experimental design for destructive gage study

The first problem concerns designs that are often referred to in connection with the destructive gage study - the crossed and nested designs [5]. Their schemes are illustrated in Fig. 1 and Fig. 2. In the illustrations two operators are participating in the experiment, each measuring nine samples. In the first case, there are three different batches of six samples and the samples from the same batch are measured by both operators. In the second case, the batches are six and the samples from one batch are measured by only one operator. An insufficient size of batches is stated as a reason for the nested design. Both Statgraphics and Minitab include nested designs on their gage study menu.
Without looking at the model equation, it is clear that in Fig. 2 the operator effect is confounded with differences between two groups of batches. Only the choice of similar batches could reduce this confounding, however, it would be contrary to the recommended procedure when the choice of parts (batches in the destructive study) should reflect the actual distribution of the measured characteristic. Consequently, only the crossed design is rationale for the destructive R&R study. The option of the nested design is the result of misunderstanding, see also [4]. Although the samples of the *j*th batch that are measured by the *i*th operator are nested within the *ij*th combination, batches are crossed with operators. Further, only the crossed design will be considered.

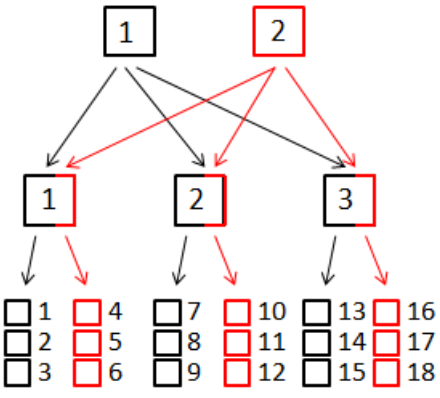

Fig. 1. Crossed design.

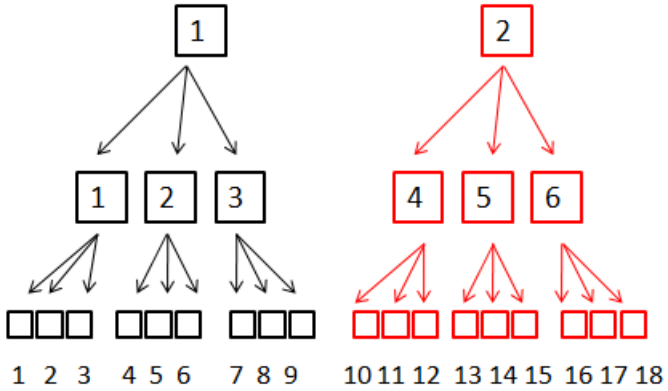

Fig. 2. Nested design.

By default, a complete factorial design with replications is used in non-destructive R&R studies and the factors being investigated are *Part* and *Operator*. The study includes *I* parts and *J* operators, each operator measures each of *I* parts *r*-times. Considering the recommended procedure for operators to proceed to the next measurement of the same part after all *I* parts have been measured, it would be appropriate to consider the layout of blocks made up of individual replicas [11] and the model should have the form

$$y_{ijk} = \mu + p_i + o_j + (po)_{ij} + b_k + e_{ijk}$$
$$i = 1, 2, ..., I;\ j = 1, 2, ..., J;\ k = 1, 2, ..., r \tag{1}$$

where μ denotes a constant (overall mean), $p_i, o_j, (po)_{ij}, b_k$ denote random effects of factors *Operator*, *Part*, *Operator***Part* interaction and blocks, e_{ijk} represents an error component. It is commonly assumed that the result is not influenced by the time of measuring, and therefore the model without the blocking factor is used

$$y_{ijk} = \mu + p_i + o_j + (po)_{ij} + e_{ijk} \tag{2}$$

The random effects of both factors correspond to the idea that both the parts and operators represent random samples. The aim of the study is not to compare the particular parts or the operators participating in the experiment; it is to measure the variability between different parts and different operators in general. It is assumed that $p_i, o_j, (po)_{ij}$ and e_{ijk} are mutually independent and normally distributed random variables with zero means and variances $\sigma_p^2, \sigma_o^2, \sigma_{po}^2$ and σ^2. Then the variance of response *Y* can be expressed in the form

$$Var(y_{ijk}) = \sigma_t^2 = \sigma_p^2 + \sigma_o^2 + \sigma_{po}^2 + \sigma^2 \tag{3}$$

The aim of the R&R study is to determine the variance components due to the measurement system

$$\sigma_{R\&R}^2 = \sigma_o^2 + \sigma_{po}^2 + \sigma^2 \tag{4}$$

and compare it with the total variance (3). Symbol σ^2 represents repeatability and $\sigma_o^2 + \sigma_{op}^2$ reproducibility.
In the destructive R&R study, where parts with repeated measurements are replaced by batches of similar samples, slight changes will be made in equations (2) to (4) - "*p*" will be replaced by "*b*".

3 Estimation of variance components

There are several methods to estimate variance components. ANOVA, maximum likelihood (ML) or restricted maximum likelihood (REML) methods belong to the best-known. For balanced data, ANOVA estimators of all variance components are unbiased and have the smallest variance of all estimators that are both quadratic functions of the observations and unbiased [10]. Under normality, they are minimum variance and unbiased. As the distribution

of estimated variance components with exception of the estimate of σ^2 cannot be described by any theoretical model, the exact confidence limits for $\sigma^2_{R\&R}$ cannot be found. Three methods for constructing approximate confidence limits can be used (see 3.2). The disadvantage of ANOVA is that it can yield negative estimates of variance components.
ML estimates of variance components are never negative but they are negatively biased. Under certain conditions of regularity, which are normally met in practice, they are consistent and asymptotically normal and efficient [7]. Using the asymptotic variance-covariance matrix, which is the inverse of the information matrix, confidence intervals for individual components and their sum $\sigma^2_{R\&R}$ can be constructed.
REML estimates, which are based on the logarithm of the restrictive maximum likelihood, take account of the number of the fixed effect parameters estimated (there is only one in the case of model (1)) and are recommended for mixed-effect models. They are usually approximately unbiased. For balanced data and if the ANOVA estimates are nonnegative, the REML and ANOVA estimates are identical. The REML estimates usually have higher mean-squared error (MSE) than the ML estimates.
Only the ANOVA method and three approaches to construction of approximate confidence are described; explanation of other two methods goes beyond the scope of this article.

3.1 ANOVA method

For balanced data, it is common practice to estimate variance components by the moment method based on ANOVA model (2). The mean squares in the ANOVA table are equated to their expected values (Tab. 1) and the resulting equations are solved for the variance components. The method is usually referred to as the ANOVA method. Sums of squares, degrees of freedom, and average squares are determined as in fixed effects models, see for example [8].

Source	Sum of squares	Degrees of freedom	Mean square	Expected mean square
Part	SS_1	$I-1$	M_1	$\theta_1 = \sigma^2 + r\sigma^2_{po} + Jr\sigma^2_p$
Operator	SS_2	$J-1$	M_2	$\theta_2 = \sigma^2 + r\sigma^2_{po} + Ir\sigma^2_o$
Interaction	SS_3	$(I-1)(J-1)$	M_3	$\theta_3 = \sigma^2 + r\sigma^2_{po}$
Residual	SS_4	$IJ(r-1)$	M_4	$\theta_4 = \sigma^2$

Tab. 1. Expected mean squares, ANOVA, random effects model.

The expected mean squares are also used to construct F-statistics and test whether the variance components differ significantly from zero (see [8] for details).
Estimates of the variance components are

$$\hat{\sigma}_p^2 = \frac{M_1 - M_3}{Jr} \quad \hat{\sigma}_o^2 = \frac{M_2 - M_3}{Ir} \quad \hat{\sigma}_{po}^2 = \frac{M_3 - M_4}{r} \quad \hat{\sigma}^2 = M_4 \tag{5}$$

Using $\psi = \sigma_{R\&R}^2$, we can write

$$\psi = \frac{\theta_2 + (I-1)\theta_3 + I(r-1)\theta_4}{Ir} \tag{6}$$

and

$$\hat{\psi} = \frac{M_2 + (I-1)M_3 + I(r-1)M_4}{Ir} \tag{7}$$

A negative estimate in (5) indicates that the true value of the corresponding variance component is zero. Even if it is replaced by zero, some problems remain. Replacing the negative estimate by zero disturbs the properties of estimates - they are not unbiased but their mean squared error is smaller [10]. Another approach is to drop the corresponding effect from the model. This approach is commonly used in connection with the interaction term but sometimes the factor *Operator* should be omitted. However, some statistical packages do not allow omitting this main effect while leaving the interaction term in the model.

3.2 Confidence limits

It can be shown that under the normality assumptions, the following notation applies

$$\frac{f_q M_q}{E(M_q)} \sim \chi^2(f_q), \; q = 1, 2, 3, 4 \tag{8}$$

and M_q are independent. Consequently, the exact confidence intervals for $\theta_q = E(M_q)$ can be determined. However, no theoretical model of the distribution of $\hat{\sigma}_o^2$ or $\hat{\sigma}_{po}^2$, which are linear combinations of mean squares, exists and therefore only approximate confidence limits for σ_o^2 and σ_{po}^2, as well as for $\sigma_{R\&R}^2$ can be determined.

Three methods for constructing approximate confidence intervals are presented in the paper: modified large sample method [3], Satterthwaite method [9] and AIAG method [1]. Due to the specific application in which small values of variance components are desirable, only upper one-sided confidence limits are considered.

3.2.1 Modified large sample method

For $\hat{\psi} = \sum_q c_q M_q$, the approximate upper 100(1–α)% one-sided confidence limit is given by the formula [6]

$$U_{MLS} = \hat{\psi} + \sqrt{\sum_q H_q^2 c_q^2 M_q^2} \tag{9}$$

where

$$H_q = \frac{f_q}{\chi^2_\alpha(f_q)} - 1 \quad \text{and} \quad c_2 = \frac{1}{Ir} \quad c_3 = \frac{I-1}{Ir} \quad c_4 = \frac{I(r-1)}{Ir}$$

3.2.2 Satterthwaite method

The distribution of $\frac{m\hat{\psi}}{\psi}$ is approximated by the chi-square distribution with m degrees of freedom, where m is the highest integer for which

$$m \le \frac{\hat{\psi}^2}{\sum_q \frac{c_q^2 M_q^2}{f_q}} \tag{10}$$

Consequently, the one-sided upper limit is

$$U_{satt} = \frac{m\hat{\psi}}{\chi^2_\alpha(m)} \tag{11}$$

3.2.3 AIAG method

The method recommended by AIAG is based on the exact confidence interval for $\theta_2 = E(M_2)$. The exact confidence limit for θ_2 or θ_2 / Ir is adjusted by adding the remaining terms in (6), with the expected mean squares θ_3 and θ_4 being replaced by their estimates M_3 and M_4. The approximate upper confidence limit is

$$U_{AIAG} = \frac{(H_2+1)M_2 + (I-1)M_3 + I(r-1)M_4}{Ir} \tag{12}$$

4 Example

In the R&R study samples of drawn wire were tested for tensile strength Rm. Due to the destructive nature of the measurements, twelve drawn wires with the final diameter of 2.5 mm were produced from 5.5 mm thick rolled rods by the drawing machine under different experimental conditions. From each wire, nine samples were prepared and divided between three operators.

4.1 Results in Minitab

Minitab uses the ANOVA method to estimate variance components and the modified large sample method to construct confidence limits. Two parts of the output yielded by Minitab are displayed in Tab. 2 and Tab. 3.

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Batch	260975	11	23725.0	14.1000	0.000
Operator	126	2	63.2	0.0376	0.963
Batch*Operator	37017	22	1682.6	2.0719	0.011
Residual	58471	72	812.1		
Total (corrected)	356590	107			

Tab. 2. ANOVA for full factorial design, Minitab.

Source	Var Comp	95% Upper Bound	% Contribution (of Var Comp)	95% Upper Bound
Total Gage R&R	1102.27	1547.689	31.04	48.05
Repeatability	812.10	1093.688	22.87	37.21
Reproducibility	290.17	695.486	8.17	19.08
Operator	0.00	33.460	0.00	0.00
Operator*Batch	290.17	730.997	8.17	21.52
Batch -to- Batch	2449.15	6147.498	68.96	85.60
Total Variation	3551.42	7267.499	100.00	

Tab. 3. Gage R&R, Minitab.

The variance component estimates in Tab. 3 indicate that no significant differences between operators exist ($\hat{\sigma}_o^2 = 0$). Based on the corresponding P-value in Tab. 2, we can consider $\sigma_o^2 = 0$. However, the estimate of the interaction variance component $\hat{\sigma}_{bo}^2 = 290.17$ differs significantly from zero. Reproducibility is given only by $\hat{\sigma}_{bo}^2$. The contribution of repeatability ($\hat{\sigma}^2 = 812.10$) to the total gage variation, the estimate of which is 1102.27, is considerable. Based on this value, the measurement system is found inadequate since the contribution of R&R is greater than 30 %, which is the maximum acceptable value according to [1]. However, since the repeatability cannot be separated from the within-batch variation and the same can be said about the interaction, it is more likely that the large variation results from non-uniform mechanical properties along the wire length due to an imperfect technology of the wire production.

Let us now look at the results of the estimation in more detail, see Tab. 4. Together with the confidence limits obtained by the modified large sample method, which are further discussed, the results of other two methods are displayed for information. Using formulas (5) we get

$$\hat{\sigma}_o^2 = -44.984 \quad \hat{\sigma}_{bo}^2 = 290.171 \quad \hat{\sigma}^2 = 812.099 \quad \text{and} \quad \hat{\sigma}_{R\&R}^2 = 1057.286$$

The upper confidence limit according to (9) is $U_{MLS} = 1502.705$, Minitab yields 1547.689.

Differences between the values of $\hat{\sigma}^2_{R\&R}$ and the values of the confidence limits correspond to replacing the negative estimate by zero. The question is whether the confidence limit should not be based on the formula $\psi = \sigma^2_{bo} + \sigma^2$. In this case we get $\hat{\sigma}^2_{R\&R} = 1102.27$ and $U_{MLS} = 1579.928$. Searle et al. [10] suggest reducing the original model, as is commonly done in the case of insignificant interaction, however, even the GLM procedure does not allow it in Minitab.

Estimation method	Estimate	CL method	95% upper CL	Note
ANOVA	1057.286	MLS	1502.705	Model (2) Negative estimate not replaced
	1102.270	MLS	1547.689	Model (2) Negative estimate replaced by 0, Formulas for CL unchanged
		SATT	1474.831	
		AIAG	1089.750	
		MLS	1579.928	Model (2) Negative estimate replaced by 0 CL based on $\psi = \sigma^2_{bo} + \sigma^2$
		SATT	1506.040	
		AIAG	1541.492	
	1088.776	MLS	1564.615	Model without *Operator* CL based on $\psi = \sigma^2_{bo} + \sigma^2$
		SATT	1442.292	
		AIAG	1527.999	
ML	1057.287	-	-	Calculation of ASYCOV matrix failed
REML	1057.233	-	-	

Tab. 4. Differences in estimates and confidence limits.

4.2 Results in Statgraphics

The output of R&R procedure displays estimated standard deviations and two-sided confidence limits. The formulas for confidence limits are not available in the Statgraphics manual, though. Tab. 5 shows the recalculated values of estimated variances and confidence limits for variance components to compare the results with Minitab.
The estimates in Tab. 5 agree with the results in Minitab while the confidence limits except for repeatability are strikingly different. Especially the confidence interval for R&R is suspiciously narrow and if the level of confidence is changed to 90 %, the upper limit is even lower than the estimate.
GLM procedure in Statgraphics can be used to omit the main effect of *Operator* in (2). Resulting estimates are (Tab. 4)

$$\hat{\sigma}^2_{bo} = 296.918\ \hat{\sigma}^2 = 791.858 \text{ and } \hat{\sigma}^2_{R\&R} = 1088.776$$

The procedure does not enable to calculate confidence limits.

	95% confidence			90% confidence		
	Lower CL	Estimate	Upper CL	Lower CL	Estimate	Upper CL
Repeatability	600.608	812.102	1159.498	630.020	812.102	1093.691
Reproducibility	0	0	0	0	0	0
Interaction	155.169	290.171	615.789	171.505	290.171	543.389
R & R	1056.010	1102.273	1124.858	1056.120	1102.273	1089.753
Parts	925.310	2449.151	7830.055	1084.405	2449.151	6431.783

Tab. 5. Variance estimates and confidence limits, Statgraphics, recalculated.

5 Conclusion

Based on the study, following conclusions can be drawn:

1. The procedure for nested design implemented in Minitab or Statgraphics is not suitable for R&R study because of confounding effects of *Operator* and differences between batch groups assigned to them.
2. The analysis of destructive R&R study can yield an evidence of satisfactory system capability if the total gage variance is small. On the other side, nothing can be concluded if the gage variance is large, because its components can be influenced by sample-to-sample variation.
3. From the statistical point of view, the way of estimation is not unambiguous if some of the estimates are negative. This is especially true for construction of the approximate confidence limits. Moreover, the confidence limits yielded by Statgraphics are not usable.
4. The methods for CL constructions should be compared as to how the stated confidence level is maintained. For example, based on their simulation study, Burdick and Larsen [3] claim that only MLS intervals met the stated confidence level - simulated confidence levels for other methods were lower. The lower confidence limits for SATT and AIAG in Tab. 4 do not contradict these findings.
5. Although the way of ML and REML estimation excludes the possibility of negative estimates, the resulting estimates of the total gage variance almost coincide with the unbiased estimates obtained by ANOVA (without replacement by 0).
6. Asymptotic confidence limits based on ML or REML estimation and the asymptotic variance-covariance matrix of estimates could not be determined because the calculation of the asymptotic variance-covariance matrix failed in SPlus.
7. Another approach could make use of the Bayes mixed model and credible intervals. This approach will be the subject of further study.

References

[1] AUTOMOTIVE INDUSTRY ACTION GROUP, *Measurement System Analysis* (*MSA*), 4th ed. Detroit, MI, 2010

[2] BURDICK, R.K., BORROR, C.M., MONTGOMERY, D.C., A Review of Methods for Measurement Systems Capability Analysis. *Journal of Quality Technology*, 35(4), 2003, pp. 342 –354.

[3] BURDICK, R.K., LARSEN, G.A. Confidence Intervals on Measures of Variability in R&R Studies. *Journal of Quality Technology*, 29(3), 1997, pp. 261–273.

[4] DE MAST, J., TRIP, A., Gauge R&R Studies for Destructive Measurements, *Journal of Quality Technology*, 37(1), 2005, pp. 40–49.

[5] GORMAN, D., BOWER, K.M., Measurement System Analysis and Destructive Testing. *Six Sigma Forum Magazine*, 1(4), 2002. Avail. from https://www.minitab.com/uploaded Files/Content/News/Published_Articles/measurement_system_analysis_destructive_ testing.pdf

[6] GRAYBILL, F.A.WANG, C.M., Confidence intervals on non-negative linear combinations of variances, *J. Amer. Stat. Assoc*. 75, 1980, pp. 869–873.

[7] HARVILLE, D.A., Maximum-likelihood approaches to variance component estimation and to related problems *J. Amer. Stat. Assoc*. 72, 1977, pp. 320–340.

[8] MONTGOMERY, D. C., *Design and Analysis of Experiments*, 8th ed. New York: Wiley, 2012,741 p.

[9] SATTERTHWAITE, F.E., An Approximate distribution of estimates of variance components. *Biometrics Bull.* 2, 1946, pp. 110–114.

[10] SEARLE, S.R., CASELLA, G., Mc CULLOCH, C.E., *Variance components*. John Wiley & Sons, New York, 1992, 501 p.

[11] SENOL, S.: Measurement System Analysis Using Designed Experiments with Minimum α-β Risks and *n*. *Measurement*, 36, 2004, pp. 131–141.

Current address

Jarošová Eva, doc. Ing., CSc.
ŠKODA AUTO University
Na Karmeli 1457, 293 01 Mladá Boleslav, Czech Republic
E-mail: eva.jarosova@savs.cz