Leveraged Measurement System Assessment

Ryan Browne, Jock MacKay and Stefan Steiner

Dept. of Statistics and Actuarial Sciences

Business and Industrial Statistics Research Group

Dept. of Statistics and Actuarial Sciences

University of Waterloo

Wateroo, N2L 3G1, Canada

rpbrowne@uwaterloo.ca

To assess a measurement system, we propose an alternative to the traditional Gauge Repeatability and Reproducibility (R&R) investigation. With the new plan called Leveraged Measurement System Assessment (LMSA), we select parts with extreme initial measured values to measure repeatedly. We assume these parts are available from regular production or a prior baseline investigation. We also assume the total variation and the process mean are known or well estimated. The substantial benefit of the LMSA is quantified. Finally, we discuss planning and implementation of a LMSA study.

KEY WORDS: Gauge R&R; Components of Variation; Intraclass correlation; Sample size

1 INTRODUCTION

Verifying the quality of a measurement system is crucial to any manufacturing process, scientific investigation, or clinical trial because all measurements are subject to error. As discussed in Shrout and Fleiss (1979), measurement errors can seriously affect decision making, statistical analysis and its interpretation. It is important to quantify such errors by assessing the measurement system.

In this article, we use manufacturing terminology but the proposed plan and conclusions apply in a wide variety of settings. For example, in a medical context we can replace part-to-part variability by person-to-person variability when assessing measurement variation in a medical instruments.

To assess a measurement system, we repeatedly measure the same parts. A commonly adopted model is

$$Y_{ij} = P_i + E_{ij} \quad i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, n \tag{1}$$

where P_i is a random variable representing the possible values for the true dimension of part i, E_{ij} is a random variable representing the measurement error, n is the number of repeated measurements on each part and k is the number of parts. We assume that the part effects $\{P_i\}$ are independent and identically distributed normal random variables with mean μ and variance σ_p^2, the measurement errors $\{E_{ij}\}$ are independent and identically distributed normal random variables with mean zero and variance σ_m^2, and $\{P_i\}$ and $\{E_{ij}\}$ are mutually independent. The variance of Y_{ij}, called the total variation, is $\sigma_t^2 = \sigma_p^2 + \sigma_m^2$.

In the standard R&R study, operators are an additional source of variation. For our purposes, we assume that there is no operator effect (as is the case with many automated measurement systems).

To quantify the contribution of the measurement system to the total variation, there are three commonly used coefficients:

- Gauge Repeatability and Reproducibility (R&R) defined as the proportion of measurement variation relative to the total variation on the standard deviation scale

$$GRR = \frac{\sigma_m}{\sqrt{\sigma_m^2 + \sigma_p^2}} = \frac{\sigma_m}{\sigma_t}.$$

- The intraclass correlation coefficient ρ defined as the proportion of process variation relative to the total variation,

$$\rho = \frac{\sigma_p^2}{\sigma_m^2 + \sigma_p^2} = \frac{\sigma_p^2}{\sigma_t^2}.$$

- The discrimination ratio D defined as the process standard deviation divided by the measure-

ment system standard deviation

$$D = \frac{\sigma_p}{\sigma_m}.$$

The three measures are one-to-one functions of each other: $\rho = 1 - GRR^2$ and $\rho = \frac{D^2}{1+D^2}$. This means they all convey the same information about the contribution of the measurement system to the total variation. It also means a test of hypothesis phrased in terms of one measure can be re-phrased in terms of another. Table 1 shows common cut off values of the GRR (see Automotive Industry Action Group, 2002) and discrimination ratio (see Steiner and Mackay, 2005) used to determine if a measurement system is acceptable.

Table 1: Cutoffs for Acceptable and Unacceptable Measurement Systems in Terms of ρ

	Measurement System		
	acceptable	needs improvement	not acceptable
GRR	$< 10\%$	$= .99 > \rho > .91 =$	$> 30\%$
Discrimination Ratio	> 3	$= .90 > \rho > .80 =$	< 2

With the standard Gauge R&R plan, we measure a sample of k parts n times each. The parts may be selected at random or haphazardly from the process. We compare the standard plan to a new plan that we call a Leveraged Measurement System Assessment (LMSA) study. In a LMSA study, k parts are selected based on their initial measurement from a large sample of measured parts and each selected part is then re-measured n times. We recommend choosing parts with extreme initial values.

Note that selecting an extreme part provides substantially more information about ρ than the 1 degree of freedom gained if the true dimension of a selected part is known. An initial extreme value can give more information than the true part dimension because we can use both the mean and variance of repeated measurements, given the initial value, to estimate ρ. When the true dimension is known only the variance of the repeated measurements can be used.

For either plan, we can test the hypothesis

$$H_0 : \rho \leq \rho_0 \text{ versus } H_A : \rho > \rho_0 \tag{2}$$

to determine whether the measurement system is acceptable. To compare study plans, we compare the power of these tests over all values of ρ when each test has equivalent size and the same number of total observations.

For the purpose of this paper, we assume the total process variation σ_t^2 and mean μ are known (or estimated with negligible error). We called these known values the baseline information. This assumption is reasonable in many applications because for many processes, such as those using 100% final inspection, baseline information is readily and cheaply available. In addition, for variation reduction exercises, a first step is to get a good estimate of the current process performance, i.e. to estimate the baseline information.

2 GAUGE R&R

In the automotive industry, the standard measurement system assessment study is a Gauge R&R plan (see Automotive Industry Action Group, 2002). To analyze the data from such a plan, AIAG recommends an Analysis of Variance (ANOVA) (see also Burdick et al., 2003) that ignores any baseline information (the known values of μ and σ_t^2). This analysis is outlined in Section 2.1. In Section 2.2, we make use of the baseline information in the analysis of a Gauge R&R plan with one method based on ANOVA and a second using Maximum Likelihood Estimation (MLE). In Section 2.3, we present a summary and comparison of the three analysis methods.

2.1 Standard Gauge R&R

The notation and standard ANOVA summary are presented in Table 2.

If we follow standard practice and ignore the baseline information, we estimate ρ by

$$\frac{\text{MSA} - \text{MSW}}{\text{MSA} - (n-1)\text{MSW}}$$

Table 2: Analysis of Variance for Gauge R&R Study

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Among parts	$k-1$	SSA	MSA = SSA/$(k-1)$	MSA/MSW
Within parts	$k(n-1)$	SSW	MSW = SSW/$[k(n-1)]$	

$$SSA = \sum_{i=1}^{k} n\,(\overline{y}_{i.} - \overline{y}_{..})^2 \qquad SSW = \sum_{i=1}^{k}\sum_{j=1}^{n}\,(y_{ij} - \overline{y}_{i.})^2$$

$$\overline{y}_{i.} = \sum_{j=1}^{n} y_{ij}/n \qquad \overline{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}/nk$$

the sample intraclass correlation coefficient.

To carry out the test of hypothesis (2) using ANOVA, we compare the value of F in Table 2 to the quantity $c_0 F_\alpha; v_1, v_2$, where $c_0 = 1 + [n\rho_0/(1-\rho_0)]$, and $F_\alpha; v_1, v_2$ is the $(1-\alpha)^{th}$ percentile of an F distribution with $v_1 = k-1$ and $v_2 = k(n-1)$ degrees of freedom. To find the power of the test when $\rho = \rho_1$, we use equation (3)

$$P\,(F \geq c\,F_\alpha; v_1, v_2) \tag{3}$$

where

$$c \;=\; \frac{(1 + n\frac{\rho_0}{1-\rho_0})}{(1 + n\frac{\rho_1}{1-\rho_1})}$$

Details of the derivation can be found in Donner and Eliasziw (1987) who use (3) to determine appropriate sample sizes for a measurement system assessment . They first rearrange the equation, expressing ρ_1 as a function of n and k when testing (2) with $\alpha = 0.05$ and power 0.80. Then, they plot contours of ρ_1 as a function of n and k. In Section 5, we use the same type of display to calculate the sample sizes for a LMSA plan.

2.2 Gauge R&R Analysis using Baseline Information

Next, we look at two ways of incorporating the baseline information, i.e. the known values of σ_t and μ, in the analysis of Gauge R&R to estimate ρ. One difference between the two methods is that the likelihood-based approach requires the parts to be chosen randomly whereas the ANOVA method does not depend on how the parts are selected.

2.2.1 ANOVA using Baseline Information

Using the ANOVA method when μ and σ_t^2 are known, we estimate ρ by utilizing the relationships $\rho = \frac{\sigma_p^2}{\sigma_p^2+\sigma_m^2}$ and $\sigma_t^2 = \sigma_p^2 + \sigma_m^2$. Estimating σ_m^2 using MSW from Table 2, the estimate for ρ is

$$1 - \frac{\text{MSW}}{\sigma_t^2}. \tag{4}$$

The corresponding estimator is unbiased and has variance

$$\frac{2(1-\rho)^2}{k(n-1)}. \tag{5}$$

A benefit of this estimator is that its properties do not depend on how the parts were selected. This means the parts selected for the measurement system investigation do not have to be representative of the process. This estimator uses the results from the gauge R&R investigation to estimate only σ_m^2, rather than both σ_m^2 and σ_p^2 as done with the standard analysis for a gauge R&R.

Since the error terms $\{E_{ij}\}$ in model (1) are assumed to be normally distributed with mean zero and variance σ_m^2, the statistic

$$W = \frac{k(n-1)\text{MSW}}{\sigma_t^2(1-\rho)}. \tag{6}$$

follows a $\chi^2_{k(n-1)}$ distribution. To test the hypothesis $\rho \leq \rho_0$, we compare W to $\chi^2_{\alpha,k(n-1)}$, the $(1-\alpha)^{th}$ percentile of a χ^2 distribution with $k(n-1)$ degrees of freedom.

The power when $\rho = \rho_1$ is

$$P\left(W \leq \frac{1-\rho_0}{1-\rho_1}\chi^2_{\alpha,k(n-1)}\right) \tag{7}$$

2.2.2 Maximum Likelihood

If we assume the parts selected for the measurement investigation are randomly sampled from the process, then we can use the maximum likelihood method to estimate ρ. To do this, we need to determine the distribution of the repeated measurements.

The n measurements on the same randomly selected part have the joint distribution:

$$\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,n} \end{pmatrix} \backsim N\left(\mu \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & & \\ \vdots & & \ddots & \vdots \\ \rho & & \dots & 1 \end{bmatrix} \right) \tag{8}$$

Observations from the k different parts are independent of each other. Thus, combining the likelihoods across k parts, we obtain the log-likelihood

$$\begin{aligned} l_1(\rho) &= -\frac{k}{2}\left\{n\log\sigma_t^2 + (n-1)\log(1-\rho) + \log\left[\rho(n-1)+1\right]\right\} \\ &\quad -\frac{1}{2\sigma_t^2(1-\rho)\left[\rho(n-1)+1\right]}\left(\left[\rho(n-1)+1\right]SSW^* - \rho n^2 SSA^*\right) \end{aligned} \tag{9}$$

where $SSW^* = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{i,j}-\mu\right)^2$ and $SSA^* = \sum_{i=1}^{k}\left(\overline{y}_{i,\cdot}-\mu\right)^2$.

There is no closed form solution for the maximum likelihood estimate of ρ. We can however examine the asymptotic properties of the estimator with the Fisher information of the parameter. The Fisher information is

$$J_1(\rho) = E\left[-\frac{\partial^2}{\partial\rho^2}\, l_1\left(\rho\right)\right] = \frac{1}{2}\frac{nk(n-1)(\rho^2(n-1)+1)}{(1-\rho)^2(1+n\rho-\rho)^2}. \tag{10}$$

The Fisher information is the inverse of the asymptotic variance of the maximum likelihood estimator. This means larger is better.

To test the hypothesis (2), we use the Wald test based on the asymptotic distribution of the MLE. Suppose that $\tilde{\rho}$ is the maximum likelihood estimator of ρ. Then, approximately, we have

$$W(\rho_0; \tilde{\rho}) = (\tilde{\rho} - \rho_0) \left[J(\rho_0)\right]^{1/2} \sim N(0, 1).$$

To perform the test, we calculate the maximum likelihood estimate $\hat{\rho}$ and then reject the hypothesis $\rho \leq \rho_0$ if the calculated value of $W(\rho_0; \hat{\rho}))$ is greater than $Z_{1-\alpha}$, the $(1-\alpha)^{th}$ quantile of the N(0,1) distribution.

The approximate power of the Wald test when $\rho = \rho_1$ is

$$P\left\{Z \geq \left(Z_{1-\alpha}\left[J_1(\rho_0)\right]^{-1/2} + \rho_0 - \rho_1\right)\left[J(\rho_1)\right]^{1/2}\right\}. \tag{11}$$

2.3 Value of the Baseline Information

We use statistical power to compare the three methods of testing the hypothesis $\rho \leq \rho_0$: maximum likelihood and ANOVA that use the known baseline parameters and standard ANOVA that does not. When comparing power curves from different tests, we ensure they all have the same size $\alpha = 0.05$. Figure 1 shows power curves, calculated from formulas (3), (7) and (11) when the sampling plan has ten parts with six repeated measurements on each part, the default choice in Gauge R&R studies (see Automotive Industry Action Group, 2002).

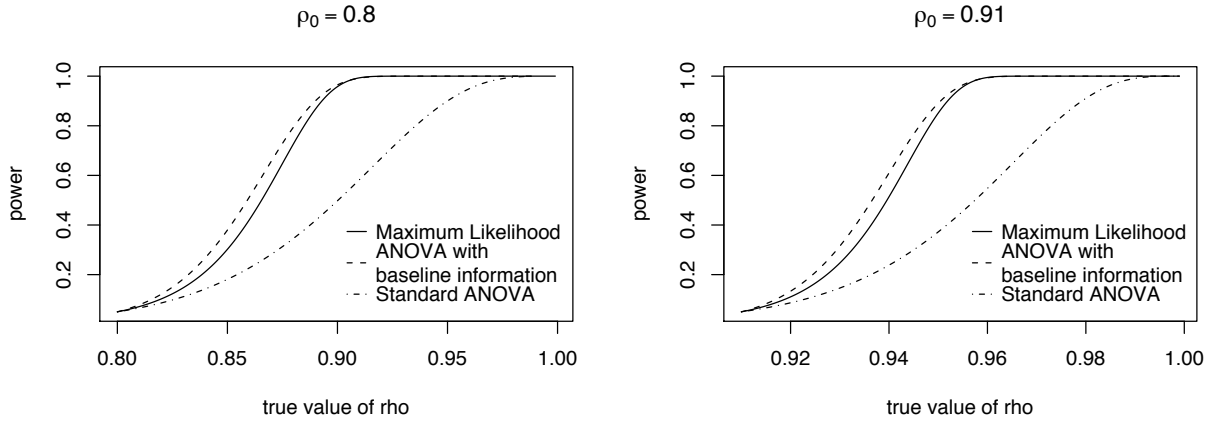

Figure 1: Power Curves for Testing (2) when $\rho_0 = 0.80$ and $\rho_0 = 0.91$

Figure 1 indicates that Maximum Likelihood and ANOVA with baseline information are significantly more powerful than the standard ANOVA analysis when ρ_0 is 0.80 or 0.91. This strongly suggests that if μ and σ_t^2 are known, this information should be used. The values of $\rho_0 = 0.80$ and 0.91 are important because they correspond to cutoff values for unacceptable measurement systems as suggested by the GRR and Discrimination Ratio in Table 1.

Under the assumption that the k parts are sampled randomly from the process, Maximum Likelihood should be the most efficient method of estimation. However, Figure 1 shows that ANOVA with baseline information has a higher power curve than the MLE. This likely happened because the power of the Wald test used in Figure 1 is based on an approximation using the asymptotic distribution of the Maximum Likelihood estimator. We verified, using simulation for $\rho \geq 0.80$ and a range of sample sizes, that the Wald and ANOVA tests are virtually equivalent in terms of power. This is true because most of the information about ρ is in MSW when ρ is close to one.

In summary, when the baseline parameters are known, we recommend the analysis using ANOVA. This approach has a closed form estimate for ρ and performs as well as the Wald test when ρ is larger than 0.80, the typical values of interest. We do not have to find the MLE numerically. In addition, since in the ANOVA method we only estimate σ_m^2 (and not σ_p^2) using the measurement assessment data the method does not require the selected parts to be representative of the process. Also, there is no need to select as many as 10 parts. In line with this observation, Steiner and Mackay (2005), Chap. 7, suggests using the ANOVA estimator with 3 parts.

3 LEVERAGED MSA

The LMSA plan is similar to the standard Gauge R&R study but instead of using a random sample, we sample parts based on their initial measurements as observed from regular production or in a baseline investigation. In particular, we sample parts that are extreme relative to the process mean μ. Again, we present three approaches for testing the hypothesis of interest. One method uses the same approach as Section 2.2.2, Maximum Likelihood. The second is based on a regression estimate. A third is based on a weighted average of the regression estimate and the ANOVA

estimate using baseline information (from Section 2.2.1).

3.1 Maximum Likelihood and Fisher Information

For a single part, the joint distribution of the initial measurement and the n repeated measurements is given by (8).

The distribution of the repeated measurements $\{Y_1, \ldots, Y_n\}$ on a single part, given that the part is selected to have initial measurement $Y_0 = y_0$ is

$$\left(\begin{array}{c|c} Y_1 & \\ \vdots & Y_0 = y_0 \\ Y_n & \end{array} \right) \backsim N \left(\mu + \rho(y_0 - \mu) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \Sigma = \sigma_t^2 \begin{bmatrix} 1-\rho^2 & & \rho(1-\rho) \\ & \ddots & \\ \rho(1-\rho) & & 1-\rho^2 \end{bmatrix} \right).$$

where $\sigma_t^2 = \sigma_p^2 + \sigma_m^2$ is the total variation. The covariance matrix Σ has a special form which allows us to obtain the following well known properties:

$$\Sigma^{-1} = \frac{1}{\sigma_t^2(1-\rho)(1+\rho n))} \begin{bmatrix} 1+\rho(n-1) & & -\rho \\ & \ddots & \\ -\rho & & 1+\rho(n-1) \end{bmatrix} \tag{12}$$

$$|\Sigma| = \sigma_t^{2n}(1-\rho)^n(1+n\rho) \tag{13}$$

Using these properties of Σ, we can write down the conditional likelihood (conditional on y_0) for the repeat measurements on a single part. The repeated measurements for one part are independent of the repeated measurements from another part, thus the conditional likelihood for k parts each with n measurements is the product of their likelihoods. The log-likelihood for n repeated measurements on k parts is

$$\begin{aligned} l_2\left(\rho \,|y_{10}, \ldots, y_{k0}\right) &= -\frac{nk}{2}\log\sigma_t^2 - \frac{nk}{2}\log(1-\rho) - \frac{k}{2}\log(1+n\rho) \\ &\quad -\frac{1}{2}\frac{1}{\sigma_t^2(1-\rho)(1+n\rho)}\left\{(1+n\rho)SSW + n\sum_{i=1}^{k}\left[\overline{y}_{i\cdot} - \mu - \rho(y_{i0}-\mu)\right]^2\right\} \end{aligned} \tag{14}$$

where SSW and $\overline{y}_{i\cdot}$ are the same as in Table 2 and y_{i0} is baseline measurement for the i^{th} selected

part.

We can calculate the Fisher information directly from the log-likelihood function.

$$J_2(\rho) = E\left[-\frac{\partial^2}{\partial\rho^2}\, l_2\,(\rho)\right] = \frac{1}{2}\frac{kn^2}{(1+n\rho)^2} + \frac{kn\rho(n+1)}{(1+n\rho)(1-\rho)^2} - \frac{1}{2}\frac{kn}{(1-\rho)^2} + \frac{n}{(1-\rho)(1+n\rho)}\frac{\sum(y_{i,0}-\mu)^2}{\sigma_t^2} \tag{15}$$

To estimate ρ, we can maximize (14) numerically. The Wald test used to test the hypothesis in (2) is found as in Section 2.2.2.

The Fisher information (15) depends on the initial values of the chosen parts through the sum of the squared standardized values (SSS) of the initial measurements,

$$SSS = \sum_{i=1}^{k} z_{i,0}^2 = \sum_{i=1}^{k}\left[\frac{y_{i,0}-\mu}{\sigma_t}\right]^2. \tag{16}$$

This means that on the standardized scale, increasing the z_{i0}^2's will increase the Fisher information. Figure 2 shows the effect of selecting 10 parts, with specified SSS value on the power. It is based on the asymptotic power of the Wald test as shown in (11) with $J_1(\rho)$ replaced by $J_2(\rho, y_0)$, as given in (15).

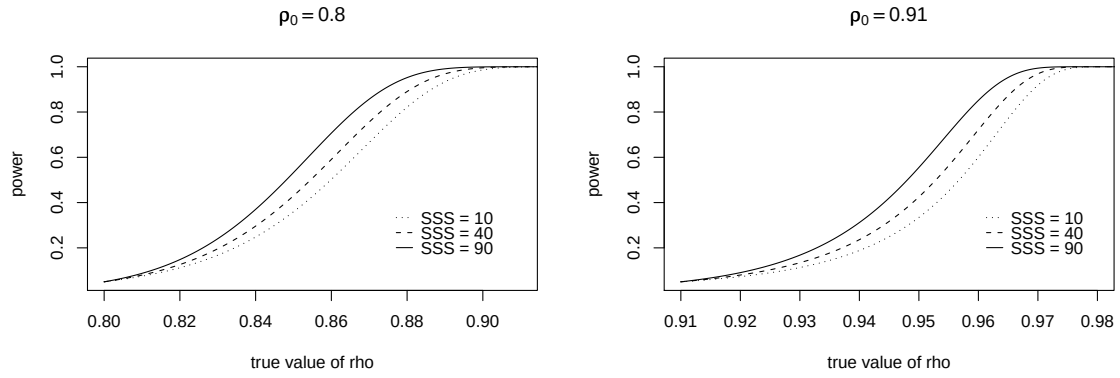

Figure 2: Power Curves for Testing the Hypothesis (2) when $\rho_0 = 0.80$ and $\rho_0 = 0.91$ at $\alpha = 0.05$, with six repeated measurements and ten parts having SSS = 10, 40 and 90

This could correspond to 10 parts with initial values equal $\mu \pm 2\sigma_t$ or any other set of $\{y_{10}, \ldots, y_{k0}\}$

such that $\sum_{i=1}^{k}\left[\frac{y_{i0}-\mu}{\sigma_t}\right]^2 = \sum_{i=1}^{k} z_{i,0}^2 = 40$.

3.2 Regression Estimator for ρ

Maximum Likelihood is an efficient method of estimation but, when an explicit expression of the MLE cannot be found, it can be useful to look for simple, efficient estimators to avoid complex calculations. The distribution of the mean of the repeated measurements on single part i given the initial measurement y_{i0} is

$$\overline{Y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij} \backsim N\left(\mu + \rho(y_{i,0}-\mu),\ \ \sigma_t^2(1-\rho)\frac{(n\rho+1)}{n}\right). \tag{17}$$

Notice that the $\overline{Y}_{i.}$'s are independent and have constant variance. This justifies the use of regression to estimate ρ. Rearranging (17) we have

$$R_i \quad \backsim N\left(\ \rho\ z_{i,0},\ \ \sigma_r^2\ \right), \tag{18}$$

where $R_i = \left[\frac{\overline{Y}_{i.}-\mu}{\sigma_t}\right]$, $z_{i,0} = \left[\frac{y_{i,0}-\mu}{\sigma_t}\right]$ and $\sigma_r^2 = (1-\rho)\frac{(n\rho+1)}{n}$. An appropriate linear model to estimate ρ is

$$R_i \ = \ \rho z_{i,0} \ + \ \epsilon \qquad \text{where} \qquad \epsilon \sim N\left(\ 0\ \ \sigma_r^2\ \right). \tag{19}$$

The estimate of ρ (Montgomery, Peck, and Vinning, 2001) based on the deviations between the observed part averages and their predicted values is

$$\hat{\rho}_{reg} = \frac{\sum_{i=1}^{k} r_i z_{i,0}}{\sum_{i=1}^{k} z_{i,0}^2} \tag{20}$$

where $r_i = \left[\frac{\overline{y}_{i.}-\mu}{\sigma_t}\right]$. The estimator is unbiased with variance

$$Var\,(\rho) = \frac{(1-\rho)(\rho+1/n)}{\sum_{i=1}^{k} z_{i,0}^2}. \tag{21}$$

We see from the denominator of (21) that this estimator has smaller variance when we choose parts that increase the sum of squares of the standardized initial measurements.

Using this estimate, we reject the hypothesis $\rho \leq \rho_0$ if

$$\frac{\hat{\rho} - \rho_0}{V(\rho_0)^{1/2}} > Z_{1-\alpha} \tag{22}$$

The power of the test when $\rho = \rho_1$ is given by

$$P\left\{Z \geq \left(Z_{1-\alpha}\left[Var(\rho_0)\right]^{1/2} + \rho_0 - \rho_1\right)\left[Var(\rho_1)\right]^{-1/2}\right\} \tag{23}$$

Note that this test does not use the information from the variability of the repeated measurements to help to estimate ρ.

3.3 Combining the ANOVA and Regression Estimators

As shown in the next subsection, the test based on the regression estimator does not perform well when compared to the MLE. To improve the power, we propose to estimate ρ using a weighted average of the regression and ANOVA estimators.

The variances of the estimators (5) and (21) depend on ρ which makes finding optimal weights impossible since ρ is unknown. We can, however, find optimal weights for a given value of $\rho = \rho_0$, the hypothesized value in (2).

If we have two unbiased independent estimators of ρ, $\tilde{\rho}_1$ and $\tilde{\rho}_2$ with known variances σ_1^2 and σ_2^2, the optimal linear combination is

$$\tilde{\rho} = w_1\,\tilde{\rho}_1 + w_2\,\tilde{\rho}_2 = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}\,\tilde{\rho}_1 + \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\,\tilde{\rho}_2 \tag{24}$$

The variance of the combined estimator is

$$Var\left(\tilde{\rho}\right) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{25}$$

Now, if we obtain $\tilde{\rho}_1$ using (20) and $\tilde{\rho}_2$ with (4), it can be shown that the two estimators are

independent. Applying (24), we obtain the combined leveraged estimate

$$\hat{\rho}_{comb} = w_1 \left(\frac{\sum_{i=1}^{k} r_i z_{i,0}}{\sum_{i=1}^{k} z_{i,0}^2} \right) + w_2 \left(1 - \frac{\text{MSW}}{\sigma_t^2} \right) \tag{26}$$

$$\begin{aligned} \text{with} \quad w_1 &= \frac{2n \sum_{i=1}^{k} z_{i,0}^2 (1-\rho_0)}{2n(1-\rho_0) \sum_{i=1}^{k} z_{i,0}^2 + k(n-1)(\rho_0 n + 1)} \\ \text{and} \quad w_2 &= \frac{k(n-1)(\rho_0 n + 1)}{2n(1-\rho_0) \sum_{i=1}^{k} z_{i,0}^2 + k(n-1)(\rho_0 n + 1)} \end{aligned}$$

Using (25), the variance of $\tilde{\rho}_{comb}$ when $\rho = \rho_0$ is

$$Var\left(\tilde{\rho}_{comb}; \rho = \rho_0\right) = \frac{2(1-\rho_0)^2(\rho_0 n + 1)}{2n(1-\rho_0) \sum_{i=1}^{k} z_{i,0}^2 + k(n-1)(\rho_0 n + 1)}. \tag{27}$$

When $\rho = \rho_1$, the variance of $\tilde{\rho}_{comb}$ is

$$Var\left(\tilde{\rho}_{comb}; \rho = \rho_1\right) = 2(1-\rho_1) \frac{\left[2n(1-\rho_0)^2(\rho_1 n + 1) \sum_{i=1}^{k} z_{i,0}^2 + k(n-1)(\rho_0 n + 1)^2(1-\rho_1)\right]}{\left[2n(1-\rho_0) \sum_{i=1}^{k} z_{i,0}^2 + k(n-1)(\rho_0 n + 1)\right]^2}. \tag{28}$$

We construct the test of the hypothesis (2) using the normal approximation for the estimate. The approximate power of the test is

$$P\left\{ Z > \left(Z_{1-\alpha} \left[Var(\tilde{\rho}_{comb}; \rho = \rho_0) \right]^{1/2} + \rho_0 - \rho_1 \right) \left[Var(\tilde{\rho}_{comb}; \rho = \rho_1) \right]^{-1/2} \right\}. \tag{29}$$

Here we use a normal approximation of the distribution of MSW. Actually, MSW follows a chi-square distribution with $k(n-1)$ degrees of freedom. A normal approximation of a chi-square distribution is reasonable if $k(n-1)$ is larger than say, 30. The default choice for Gauge R&R studies is $k = 10$ and $n = 6$ which means $k(n-1) = 50$.

3.4 Comparison of Leveraged MSA Analysis Methods

We compare the three estimation methods for a LMSA using the power of the tests. We use the power from the Wald tests from Section 3.1, (23) and (29) to generate the power curve for the

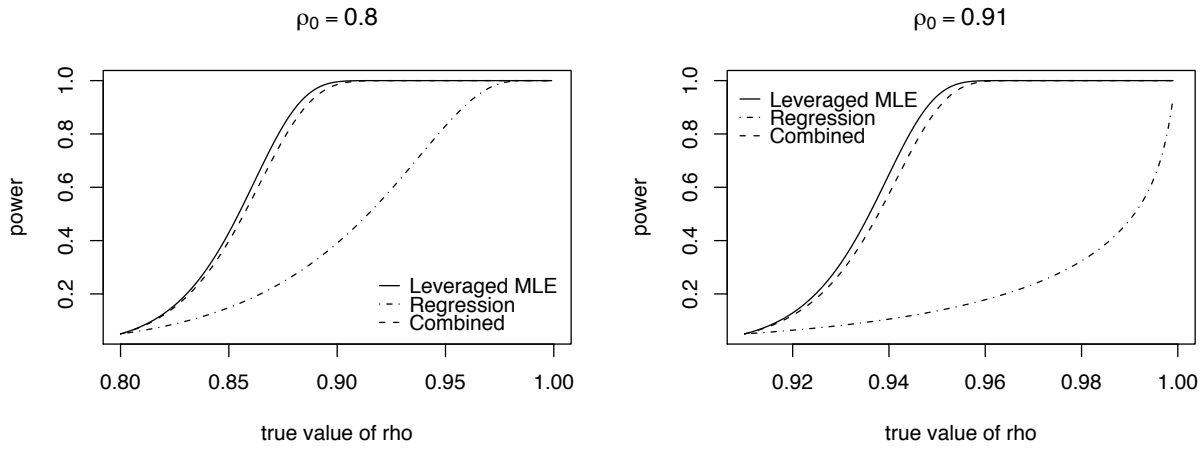

Figure 3: Power Curves for Testing the Hypothesis (2) when $\rho_0 = 0.80$, $\rho_0 = 0.91$, ten parts having $\sum z_{i,0}^2 = 40$ and $k = 10$, $n = 6$.

Leveraged MLE. For the regression and combined estimators, the power is given by (23) and (29), respectively. In all the comparisons, we use a sample of ten parts with $\sum_{i=1}^{k} z_{i,0}^2 = 40$ and six repeated measurements on each part.

Figure 3 shows that the test based on the regression estimator performs poorly and that the test based on the combined estimator performs almost as well as the Wald test based on the MLEs, with the advantage of a closed form solution.

4 THE VALUE OF LEVERAGING

We can use leveraging whenever there is a supply of parts with measured (and recorded) values so that we can select extremes. Such a supply will be available if the measurement system is used routinely in production or if the measurement system assessment is preceded by a baseline study to assess the performance of the process. There would be a small cost to find the extreme parts. What is the gain? To address this question we compare power curves for leveraged and unleveraged measurement assessment studies with the same number of parts and repeated measurements.

The most common sample size used in Gauge R&R studies is 10 parts and 6 repeated measurements. Suppose for the leveraged plan, we select the 10 parts so that $\sum_{i=1}^{k} z_{i,0}^2 = 40$. For the standard plan, we select the parts at random and assume μ and σ_t are known. We generated the

power curves for testing the hypothesis in (2) using the Maximum Likelihood Wald test for LMSA and gauge R&R studies. The power of the Wald test for Gauge R&R studies is found in (11). The power of the Wald test for Leveraged MSA uses the same formula but instead of $J_1(\rho)$ it uses $J_2(\rho)$ which is given in (15). We used Maximum Likelihood for both power curves because this way, the two designs can be evaluated fairly.

Figure 4, shows that a Leveraged MSA study is substantially more powerful than a Standard Guage R&R plan when the 10 parts are used with $\sum_{i=1}^{k} z_{i,0}^2 = 40$ and six repeated measurements on each part.

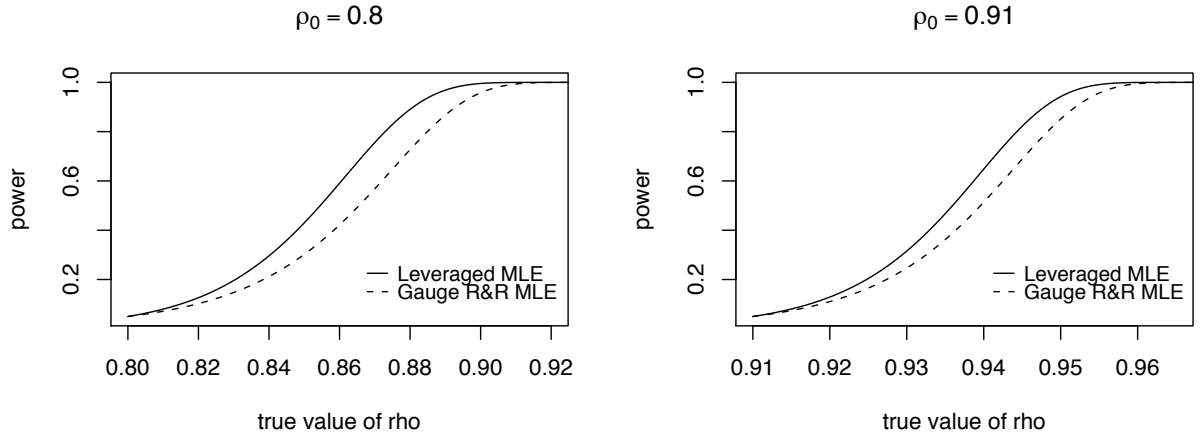

Figure 4: Power Curve for Leveraging and Standard Gauge R&R from Testing (2) when $\rho_0 = 0.80$, $\rho_0 = 0.91$, ten parts having $\sum z_{i,0}^2 = 40$ and 6 repeated measurements.

We can also quantify the effects of leveraging by comparing sample sizes. Suppose we select 5 parts with initial measured values about two standard deviations from the mean and measure each 5 times. Using the test based on the combined estimate, the power of the test for $\rho_0 = 0.8$ at $\rho = 0.9$ is about 0.80. To get the same power with the unleveraged plan and the test based on the ANOVA estimate, we would need to measure each of the 5 randomly selected parts 7 times.

The conclusion is that, in many situations, leveraging can increase power or correspondingly, the precision of the estimate of ρ, with little increase in cost.

5 PLANNING & IMPLEMENTING A LMSA STUDY

To plan a measurement investigation, we need to specify ρ_0 in (2) and the value of ρ that we want to detect at the desired power. Figure 5 shows the values of ρ that have 0.80 power when testing ρ_0 at various values of k and n. Thus, these contours show the (n, k) combinations that satisfy the required size and power for the assumed values of ρ.

To obtain the contours in Figure 5, we determined, for all discrete points (n, k), the values of ρ that have power 0.80 using the asymptotic Wald test when testing ρ_0 in (2). To determine the power of the Wald test for Leveraged MSA, we use (11) but instead of $J_1(\rho)$, we substituted $J_2(\rho)$ as given in (15).

Figure 5 can be used to determine the necessary sample sizes for testing (2) when $\rho_0 = 0.80$ and $\rho_0 = 0.91$ with size 0.05 and power 0.80. For example, suppose we wish to verify if the process variability is 80% or more of the total variation (i.e. $\rho_0 = 0.80$) and it is currently thought that ρ is around 0.95. The point on Figure 5, with $\rho_0 = 0.80$, that corresponds to five parts with $\sum_{i=1}^{5} z_{i,0}^2 = 20$ and five repeated measurements on each part satisfies the needs of this investigation because it is above the $\rho = 0.95$ contour.

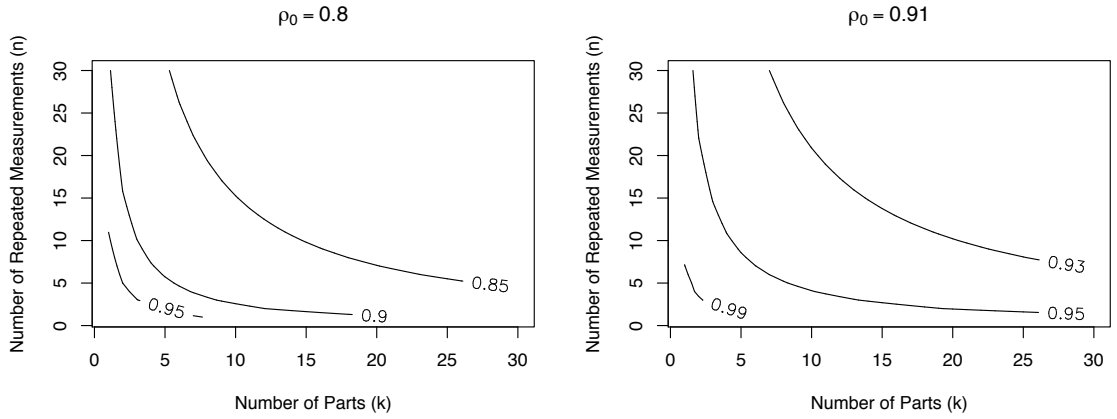

Figure 5: Contours of ρ for testing (2) when $\rho_0 = 0.80$ and $\rho_0 = 0.91$ at $\alpha = 0.05, \beta = 0.80$, with $\sum_{i=1}^{k} z_{i,0}^2 = 4k$

For another example, suppose we are planning an investigation to determine if the discrimination ratio is greater than 2 and we want to detect a ratio greater than 3 with 0.80 power. As shown

in Table 1, these discrimination ratio values correspond to $\rho = 0.80$ and 0.90, respectively. Any combination of a number of parts and repeated measurements which is above the 0.90 contour on Figure 5, with $\rho_0 = 0.80$, will be satisfactory.

Examining Figure 5 shows that theoretically a design with $k \leq 2$ can satisfy the power requirements. This type of design, though possible, is not recommended. A more desirable design is to use several extremes on either side of the mean and to choose a couple initial measurements close to the mean. This type of design will also help to check if the measurement variance is constant across the range of true part dimensions.

In practice one will not be able to find k parts with initial measures of predetermined value, for example, $\mu \pm 2\sigma_t$. Fortunately, the analysis can be done conditionally on any set of initial measurements $\{y_{01}, y_{02}, \ldots, y_{0k}\}$ because the distribution of these measurements does not depend on ρ. The conditionality principle is regularly used in regression when the analysis is conditional on the covariates. See Cox and Hinkley (1974) for a discussion on the conditionality principle. As seen in (27) and (28) we get the gain as long as $\sum_{i=1}^{k} z_{i,0}^2$ is large. The results presented in Figure 4 illustrate the effectiveness of leveraging and Figure 5 is useful to assist in planning a LMSA.

6 CONCLUSIONS AND DISCUSSION

We compare two approaches for assessing a measurement system: Gauge R&R and Leveraged MSA. We assume the total variation σ_t^2 and the process mean μ are known. Under this assumption, we have shown that incorporating baseline information into the analysis is simple and greatly improves the precision of estimation and power of the hypothesis test $H_0 : \rho \leq \rho_0$ versus $H_A : \rho > \rho_0$ over methods which ignore baseline information.

Also, we showed that leveraging, i.e. sampling and repeatedly measuring parts which have initial measurements far from the mean of the process, increases the precision of estimation of ρ, the intraclass correlation coefficient and increases the power of the hypothesis test in (2) compared to the standard Gauge R&R plan. Complicated analysis is not required to realize this gain in power. Section 3.4 shows that the estimator (26) which combines the regression (20) and ANOVA using baseline information (4) estimators achieves almost the same power as the MLE. The benefit

of leveraging increases when more extreme parts are chosen.

Leveraging can be applied to other variance component problems where one source of variation can be held fixed (the process) and the other varies (the measurement system). One example is an assembly-disassembly experiment (see Steiner and Mackay, 2005, 10.4) where the goal is to determine which is the greater source of output variation, the components or the assembly operation. In this experiment, $k = 2$ products with several components are measured, disassembled, and then reassembled several times. Performing the experiment in this way ensures the variation from the components is held fixed while the assembly contribution varies in each run. We use leveraging by selecting the two products for the experiment to have extreme and opposite performance when initially measured.

In this work we assumed the process mean μ and overall process variance σ_t^2 are known. These assumptions may not be realistic. To relax this assumption we can consider taking a baseline sample of b randomly chosen parts. From the baseline sample we can estimate μ and σ_t^2. If b is large, then the estimation errors will have negligible effects on the results presented here. Furthermore there will be a number of extreme parts available. A more interesting case, to be considered in a subsequent paper occurs when we simultaneously plan the baseline and measurement system study. Now a key issue is how to allocate resources between the two phases of the study.

References

Automotive Industry Action Group. *Measurement Systems Analysis*. Southfield MI., third edition, 2002.

R. K. Burdick, C. M. Borror, and D. C. Montgomery. A review of methods for measurement systems capability analysis. *J. Qual. Technol.*, 35:342–354, 2003.

D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

A. Donner and M. Eliasziw. Sample size requirements for reliability studies. *Statistics in Medicine*, 6:441–448, 1987.

Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vinning. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2001.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86:420–428, 1979.

Stefan H. Steiner and R. Jock Mackay. *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes*. ASQ Quality Press, 2005.