

Graduate Lectures and Problems in Quality
Control and Engineering Statistics:
Theory and Methods

To Accompany

Statistical Quality Assurance Methods for Engineers

by

Vardeman and Jobe

Stephen B. Vardeman

V2.0: January 2001

© Stephen Vardeman 2001. Permission to copy for educational purposes granted by the author, subject to the requirement that this title page be affixed to each copy (full or partial) produced.

Chapter 4

Process Characterization and Capability Analysis

Sections 5.1 through 5.3 of V&J discuss the problem of summarizing the behavior of a stable process. The “bottom line” of that discussion is that one-sample statistical methods can be used in a straightforward manner to characterize a process/population/universe standing behind data collected under stable process conditions. Section 5.5 of V&J opens a discussion of summarizing process behavior when it is not sensible to model all data in hand as random draws from a single/fixed universe. The notes in this chapter carry the theme of §5.5 of V&J slightly further and add some theoretical detail missing in the book.

4.1 General Comments on Assessing and Dissecting “Overall Variation”

The questions “How much variation is there overall?” and “Where is the variation coming from?” are fundamental to process characterization/understanding and the guidance of improvement efforts. To provide a framework for discussion here, suppose that in hand one has r samples of data, sample i of size n_i ($i = 1, \dots, r$). Depending upon the specific application, these r samples can have many different logical structures. For example, §5.5 of V&J considers the case where the n_i are all the same and the r samples are naturally thought of as having a balanced hierarchical/tree structure. But many others (both “regular” and completely “irregular”) are possible. For example Figure 4.1 is a schematic parallel to Figure 5.16 of V&J for a “staggered nested data structure.”

When data in hand represent the entire universe of interest, methods of probability and statistical inference have no relevance to the basic questions “How much variation is there overall?” and “Where is the variation coming from?” The problem is one of descriptive statistics only, and various creative

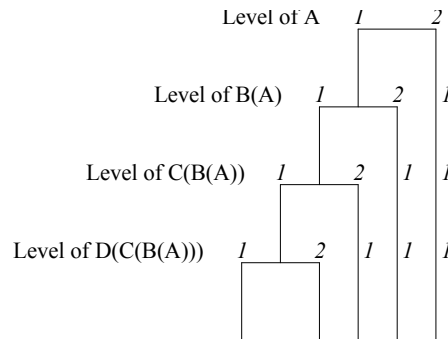


Figure 4.1: Schematic of a staggered Nested Data Set

combinations of methods of statistical graphics and basic numerical measures (like sample variances and ranges) can be assembled to address these issues. And most simply, a “grand sample variance” is one sensible characterization of “overall variation.”

The tools of probability and statistical inference only become relevant when one sees data in hand as representing something more than themselves. And there are basically two standard routes to take in this enterprise. The first posits some statistical model for the process standing behind the data (like the hierarchical random effects model (5.28) of V&J). One may then use the data in hand in the estimation of parameters (and functions of parameters) of that model in order to characterize process behavior, assess overall variability and dissect that variation into interpretable pieces.

The second standard way in which probabilistic and statistical methods become relevant (to the problems of assessing overall variation and analysis of its components) is through the adoption of a “finite population sampling” perspective. That is, there are times where there is conceptually some (possibly highly structured) concrete data set of interest and the data in hand arise through the application (possibly in various complicated ways) of random selection of *some* of the elements of that data set. (As one possible example, think of a warehouse that contains 100 crates, each of which contains 4 trays, each of which in turn holds 50 individual machine parts. The 20,000 parts in the warehouse could constitute a concrete population of interest. If one were to sample 3 crates at random, select at random 2 trays from each and then select 5 parts from each tray at random, one has a classical finite population sampling problem. Probability/randomness has entered through the sampling that is necessitated because one is unwilling to collect data on all 20,000 parts.)

Section 5.5 of V&J introduces the first of these two approaches to assessing and dissecting overall variation for balanced hierarchical data. But it does not treat the finite population sampling ideas at all. The present chapter of these notes thus extends slightly the random effects analysis ideas discussed in §5.5 and then presents some simple material from the theory of finite population

sampling.

4.2 More on Analysis Under the Hierarchical Random Effects Model

Consider the hierarchical random effects model with 2 levels of nesting discussed in §5.5.2 of V&J. We will continue the notations y_{ijk} , \bar{y}_{ij} , \bar{y}_i , and $\bar{y}_.$ used in that section and also adopt some additional notation. For one thing, it will be useful to define some ranges. Let

$R_{ij} = \max_k y_{ijk} - \min_k y_{ijk}$ = the range of the j th sample within the i th level of A ,

$\Delta_i = \max_j \bar{y}_{ij} - \min_j \bar{y}_{ij}$ = the range of the J sample means within the i th level of A ,

and

$\Gamma = \max_i \bar{y}_i - \min_i \bar{y}_i$ = the range of the means for the I levels of A .

It will also be useful to consider the ANOVA sums of squares and mean squares alluded to briefly in §5.5.3. So let

$$\begin{aligned} SSTot &= \sum_{i,j,k} (y_{ijk} - \bar{y}_.)^2 \\ &= (IJK - 1) \times \text{the grand sample variance of all } IJK \text{ observations ,} \\ SSC(B(A)) &= \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij})^2 \\ &= (K - 1) \times \text{the sum of all } IJ \text{ "level C" sample variances ,} \\ SSB(A) &= K \sum_{i,j} (\bar{y}_{ij} - \bar{y}_i.)^2 \\ &= K(J - 1) \times \text{the sum of all } I \text{ sample variances of } J \text{ means } \bar{y}_{ij} \end{aligned}$$

and

$$\begin{aligned} SSA &= KJ \sum_i (\bar{y}_i. - \bar{y}_.)^2 \\ &= KJ(I - 1) \times \text{the sample variance of the } I \text{ means } \bar{y}_i. . \end{aligned}$$

Note that in the notation of §5.5.2, $SSA = KJ(I - 1)s_A^2$, $SSB(A) = K(J - 1) \sum_{i=1}^I s_{B_i}^2$ and $SSC(B(A)) = (K - 1) \sum_{i,j} s_{ij}^2 = IJ(K - 1)\hat{\sigma}^2$. And it is an algebraic fact that $SSTot = SSA + SSB(A) + SSC(B(A))$.

Mean squares are derived from these sums of squares by dividing by appropriate degrees of freedom. That is, define

$$MSA \doteq \frac{SSA}{I - 1} ,$$

$$MSB(A) \doteq \frac{SSB(A)}{I(J-1)},$$

and

$$MSC(B(A)) \doteq \frac{SSC(B(A))}{IJ(K-1)}.$$

Now these ranges, sums of squares and mean squares are interesting measures of variation in their own right, but are especially helpful when used to produce estimates of variance components and functions of variance components. For example, it is straightforward to verify that under the hierarchical random effects model (5.28) of V&J

$$ER_{ij} = d_2(K)\sigma,$$

$$E\Delta_i = d_2(J)\sqrt{\sigma_\beta^2 + \sigma^2/K}$$

and

$$E\Gamma = d_2(I)\sqrt{\sigma_\alpha^2 + \sigma_\beta^2/J + \sigma^2/JK}.$$

So, reasoning as in §2.2.2 of V&J (there in the context of two-way random effects models and gage R&R) reasonable range-based point estimates of the variance components are

$$\hat{\sigma}^2 = \left(\frac{\bar{R}}{d_2(K)} \right)^2,$$

$$\hat{\sigma}_\beta^2 = \max \left(0, \left(\frac{\bar{\Delta}}{d_2(J)} \right)^2 - \frac{\hat{\sigma}^2}{K} \right)$$

and

$$\hat{\sigma}_\alpha^2 = \max \left(0, \left(\frac{\Gamma}{d_2(I)} \right)^2 - \frac{1}{J} \left(\frac{\bar{\Delta}}{d_2(J)} \right)^2 \right).$$

Now by applying linear model theory or reasoning from V&J displays (5.30) and (5.32) and the fact that $Es_{ij}^2 = \sigma^2$, one can find expected values for the mean squares above. These are

$$EMSA = KJ\sigma_\alpha^2 + K\sigma_\beta^2 + \sigma^2,$$

$$EMSB(A) = K\sigma_\beta^2 + \sigma^2$$

and

$$EMSC(B(A)) = \sigma^2.$$

And in a fashion completely parallel to the exposition in §1.4 of these notes, standard linear model theory implies that the quantities

$$\frac{IJ(K-1)MSC(B(A))}{EMSC(B(A))}, \frac{I(J-1)MSB(A)}{EMSB(A)} \text{ and } \frac{(I-1)MSA}{EMSA}$$

are independent χ^2 random variables with respective degrees of freedom

$$IJ(K-1), I(J-1) \text{ and } (I-1).$$

Table 4.1: Balanced Data Hierarchical Random Effects Analysis ANOVA Table (2 Levels of Nesting)

ANOVA Table				
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>EMS</i>
A	<i>SSA</i>	$I - 1$	<i>MSA</i>	$KJ\sigma_\alpha^2 + K\sigma_\beta^2 + \sigma^2$
B(A)	<i>SSB(A)</i>	$I(J - 1)$	<i>MSB(A)</i>	$K\sigma_\beta^2 + \sigma^2$
C(B(A))	<i>SSC(B(A))</i>	$IJ(K - 1)$	<i>MSC(B(A))</i>	σ^2
Total	<i>SSTot</i>	$IJK - 1$		

These facts about sums of squares and mean squares for the hierarchical random effects model are conveniently summarized in the usual (hierarchical random effects model) ANOVA table (for two levels of nesting), Table 4.1. Further, the fact that the expected mean squares are simple linear combinations of the variance components σ_α^2 , σ_β^2 and σ^2 motivates the use of linear combinations of mean squares in the estimation of the variance components (as in §5.5.3 of V&J). In fact (as indicated in §5.5.3 of V&J) the standard ANOVA-based estimators

$$\hat{\sigma}^2 = \frac{SSC(B(A))}{IJ(K-1)},$$

$$\hat{\sigma}_\beta^2 = \frac{1}{K} \max\left(0, \frac{SSB(A)}{I(J-1)} - \hat{\sigma}^2\right)$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{JK} \max\left(0, \frac{SSA}{(I-1)} - \frac{SSB(A)}{I(J-1)}\right)$$

are exactly the estimators (described without using ANOVA notation) in displays (5.29), (5.31) and (5.33) of V&J. The virtue of describing them in the present terms is to suggest/emphasize that all that was said in §1.4 and §1.5 (in the gage R&R context) about making standard errors for functions of mean squares and ANOVA-based confidence intervals for functions of variance components is equally true in the present context.

For example, the formula (1.3) of these notes can be applied to derive standard errors for $\hat{\sigma}_\beta^2$ and $\hat{\sigma}_\alpha^2$ immediately above. Or since

$$\sigma_\beta^2 = \frac{1}{K}EMS_B(A) - \frac{1}{K}EMSC(B(A))$$

and

$$\sigma_\alpha^2 = \frac{1}{JK}EMSA - \frac{1}{JK}EMS_B(A)$$

are both of form (1.4), the material of §1.5 can be used to set confidence limits for these quantities.

As a final note in this discussion of the what is possible under the hierarchical random effects model, it is worth noting that while the present discussion has been confined to a “balanced data” framework, Problem 4.8 shows that at least

point estimation of variance components can be done in a fairly elementary fashion even in unbalanced data contexts.

4.3 Finite Population Sampling and Balanced Hierarchical Structures

This brief subsection is meant to illustrate the kinds of things that can be done with finite population sampling theory in terms of estimating overall variability in a (balanced) hierarchical concrete population of items and dissecting that variability.

Consider first a finite population consisting of NM items arranged into N levels of A, with M levels of B within each level of A. (For example, there might be N boxes, each containing M widgets. Or there might be N days, on each of which M items are manufactured.) Let

y_{ij} = a measurement on the item at level i of A and level j of B within the i th level of A (e.g. the diameter of the j th widget in the i th box) .

Suppose that the quantity of interest is the (grand) variance of all NM measurements,

$$S^2 = \frac{1}{NM - 1} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y})^2 .$$

(This is clearly one quantification of overall variation.)

The usual one-way ANOVA identity applied to the NM numbers making up the population of interest shows that the population variance can be expressed as

$$S^2 = \frac{1}{NM - 1} (M(N - 1)S_A^2 + N(M - 1)S_B^2)$$

where

$$S_A^2 = \frac{1}{N - 1} \sum_{i=1}^N (\bar{y}_i - \bar{y})^2 = \text{the variance of the } N \text{ "A level means"}$$

and

$$S_B^2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{M - 1} \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2 \right) = \text{the average of the } N \text{ "within A level variances."}$$

Suppose that one selects a simple random sample of n levels of A, and for each level of A a simple random sample of m levels of B within A. (For example, one might sample n boxes and m widgets from each box.) A naive way to estimate S^2 is to simply use the sample variance

$$s^2 = \frac{1}{nm - 1} \sum (y_{ij} - \bar{y}^*)^2$$

where the sum is over the nm items selected and \bar{y}^* is the mean of those measurements. Unfortunately, this is not such a good estimator. Material from Chapter 10 of Cochran's *Sampling Techniques* can be used to show that

$$Es^2 = \frac{m(n-1)}{nm-1} S_A^2 + \left(\frac{n(m-1)}{nm-1} + \frac{m(n-1)}{nm-1} \left(\frac{1}{m} - \frac{1}{M} \right) \right) S_B^2,$$

which is not in general equal to S^2 .

However, it is possible to find a linear combination of the sample versions of S_A^2 and S_B^2 that has expected value equal to the population variance. That is, let

$$\begin{aligned} s_A^2 &= \frac{1}{n-1} \sum (\bar{y}_i^* - \bar{y}^*)^2 \\ &= \text{the sample variance of the } n \text{ sample means (from the sampled levels of A)} \end{aligned}$$

and

$$\begin{aligned} s_B^2 &= \frac{1}{n} \sum \left(\frac{1}{m-1} \sum (y_{ij} - \bar{y}_i^*)^2 \right) \\ &= \text{the average of the } n \text{ sample variances (from the sampled levels of A)}. \end{aligned}$$

Then, it turns out that

$$Es_A^2 = S_A^2 + \left(\frac{1}{m} - \frac{1}{M} \right) S_B^2$$

and

$$Es_B^2 = S_B^2.$$

From this it follows that an unbiased estimator of S^2 is the quantity

$$\frac{M(N-1)}{NM-1} s_A^2 + \left(\frac{N(M-1)}{NM-1} - \frac{M(N-1)}{NM-1} \left(\frac{1}{m} - \frac{1}{M} \right) \right) s_B^2.$$

This kind of analysis can, of course, be carried beyond the case of a single level of nesting. For example, consider the situation with two levels of nesting (where both the finite population and the observed values have balanced hierarchical structure). Then in the ANOVA notation of §4.2 above, take

$$s_A^2 = \frac{SSA}{(I-1)JK},$$

$$s_B^2 = \frac{SSB(A)}{I(J-1)K}$$

and

$$s_C^2 = \frac{SSC(B(A))}{IJ(K-1)}.$$

Let S_A^2 , S_B^2 and S_C^2 be the population analogs of s_A^2 , s_B^2 and s_C^2 , and f_B and f_C be the sampling fractions at the second and third stages of item selection. Then it turns out that

$$Es_A^2 = S_A^2 + \frac{(1-f_B)}{J}S_B^2 + \frac{(1-f_C)}{JK}S_C^2 ,$$

$$Es_B^2 = S_B^2 + \frac{(1-f_C)}{K}S_C^2$$

and

$$Es_C^2 = S_C^2 .$$

So (since the grand population variance, S^2 , is expressible as a linear combination of S_A^2 , S_B^2 and S_C^2 , each of which can be estimated by a linear combination of s_A^2 , s_B^2 and s_C^2) an unbiased estimator of the population variance can be built as an appropriate linear combination of s_A^2 , s_B^2 and s_C^2 .