

## Testing and measuring the differences among methods:

The basic philosophy of the text is that some different methods must be exactly equivalent.

My philosophy is that two methods are never exactly equivalent. So what one wants is a measure of how different the methods are. Are the differences important?

Standard transportation practice is to let a t-test (or related test) determine whether or not methods differ. This is logically strange since:

1. Assumptions never exactly hold.
2. Even unimportant differences will be detected if the sample sizes are big enough.

In collaboration or consulting doing something else requires explaining:

1. Reformulate the hypothesis-testing problem as:

$$H_0: |\mu_1 - \mu_2|$$

$$H_a: |\mu_1 - \mu_2| \quad ,$$

and use say a maximum likelihood ratio test. This is a different test than the engineers are using. Whether or not the different test is important depends upon  $\alpha$  and the sample sizes.

Alternatives to standard hypothesis tests:

Suppose  $F_1$  is the maximum likelihood density estimate under  $H_0$  and  $F_2$  is the maximum likelihood density estimate under  $H_a$ . We can take a convex combination of the two  $pF_1 + (1-p)F_2$ , and estimate  $p$  from the data by pseudo MLE techniques.

The constant  $p$  measures the strength of the data supporting each hypothesis. Instead of rejecting or accepting either hypothesis both hypotheses are supported to a degree.

Estimates of differences among the methods should probably be made using robust estimators (probably the median should be used) as transportation data is often plagued by outliers.

### **Producing a common value using different methods.**

Producing a common value using different methods, examples taken from chemistry and transportation.

Suppose a SRM is to be made and certified using measurements from several techniques. What should be its certified value? Is the value to be a consensus value or an absolute value? (Interval or ratio scale?)

$$X_{ij} = \mu + r_i + e_{ij}, \quad i = 1, \dots, k; \quad j = 1, \dots, n_i. \quad 2.$$

Where the  $r_i$  represents the bias for group  $i$  and the measurement errors,  $e_{ij}$ , are all independent with  $\text{Var}(e_{ij}) = \frac{2}{i}$ .

If the  $r_i$  is random then we have a random effects model (see references that in the hand out).

There are several ways in which a SRM can be certified. An obvious way is with method dependent certification: a value for ICP, a value for gamma spectroscopy, etc.. A single value assuming the methods were chosen at random with an uncertainty statement or a single value assuming the methods were not chosen at random and an uncertainty statement.

Ways to proceed if the  $r_i$ 's are regarded as fixed:

1. Test the null hypothesis that all the  $r_i$ 's are equal. If the null hypothesis is accepted then ignore the  $r_i$ 's (at your peril) under the assumption that all  $r_i = 0$  if this occurs.

2. Assume that at least two of the  $r_i$ 's have different sign:  $r_i r_j < 0$ . If so then the union of confidence intervals provide valid coverage (at what level) and the center of this set is a point estimate. What properties does this estimate have?

3. The  $r_i$ 's are known to lie within specified bounds:  $|r_i| \leq M_i$  where the  $M_i$  are given bounds. Then we can hope to measure the maximum mean squared error.

$$E \left( \sum c_i \bar{x}_i - \mu \right)^2 = \sum c_i^2 \frac{S_i^2}{n_i} + \left[ \sum (|c_i| M_i) + \mu \left( 1 - \sum c_i \right) \right]^2.$$

It is clear that  $1 - \sum c_i = 0$  or the maximum mean squared error =

Using calculus or a quadratic programming routine the optimal  $c$ 's can be found and CI's made. If, as in the usual case, the standard deviations are not known but instead estimates are available we use these but then the CI's are harder to form.

The approximate degrees of freedom of the variance estimate

$$DF \left( \sum c_i^2 \frac{S_i^2}{n_i} \right) = \frac{\left( \sum c_i^2 \frac{S_i^2}{n_i} \right)^2}{\sum \frac{\left[ c_i^2 \frac{S_i^2}{n_i} \right]^2}{n_i - 1}}.$$

Comparisons to other estimators:

## Calibration Curves

Calibration relates instrument response to standard values

Data is collected in two stages:

1. Training stage (calibration experiment)
2. Measurement stage (instrument response)

Step 1. Collecting the calibration data and modeling the calibration curve

The data is  $(x_i, Y_i)$ ,  $i = 1, \dots, n$ .

In the usual statistics tradition  $x_i$  denotes the chosen standards (Predictors) and  $Y_i$  denotes the instrument response. Engineers and scientists sometimes interchange the notation so beware. In addition  $c$  (concentration) is often substituted for  $x$ .

In the beginning we assume a straight-line homoskedastic model.

$$Y_i = a + b x_i + e_i$$

and our goal from the first stage is to

1. Choose the  $x_i$
2. Estimate a, b, and  $s$ .

Second stage

We collect instrument responses  $Z_i$  of unknown  $x_i$  values.

Provide point and interval estimates of the  $x_i$  values.

There are many possibilities.

Absolute calibration:

A quick or nonstandard method is calibrated against a standard or defined measurement.

Comparative calibration:

One instrument or measurement technique is calibrated against another. Neither is inherently standard. Last chapter.

How is the Calibration Curve Used?

Single use: Here separation of 2 calibration steps is artificial.

Multiple use: Lots of dependent measurements.

Used in combination with other measurements?

Who produces the calibration curve or does the calibrator do both steps 1 and 2?

Designed or natural calibration experiment?

Designed calibration experiments are typical of laboratory settings.

Natural calibration occurs in observational areas such as traffic monitoring.

Are the  $Z_i$  arbitrary or natural?

Arbitrary implies no prior knowledge about where the  $Z$ 's come from and is typical of a frequentist set up.

Natural implies some prior distribution about the  $Z$ 's.