Combining Information

Session 1 – Thursday, September 26, 2002
Session 2 – Friday, September 27, 2002

Administration Building – Lecture Room **D**
1:30 p.m. – 4:30 p.m.

James Yen
Statistical Engineering Division

yen@nist.gov
x2843

Outline

1. Combining information for parameter estimation
    - Exploring the Data
    - Mean of Means
    - Weighted Means
    - Multi-Method Problem
        - Schiller-Eberhardt Method
        - BOB method (Levenson)
        - Mande-Paule method
        - Maximum Likelihood (Vangel, Rukhin)
    - Borrowing Strength: Empirical Bayes analysis
2. Combining information for decision making.
    - Simultaneous statistical inference
        - Bonferroni method
    - Hypothesis testing and p-values
    - Combining p-values
    - Effect sizes

3. Estimation of Consensus Function

- Linear Case
  - Melting Pot Regression
  - Average Coefficients
- Non-linear case
  - Loess–Localized Regression
  - Splines
- Uncertainty Estimation
  - Residual Samples Method

$$\begin{tikzcd}
E_1 \arrow[d] & E_2 \arrow[d] & \cdot & \cdots\cdots & \cdot & \cdot E_{k-1} \arrow[d] & E_k \arrow[d] \\
R_1 \arrow[drrr] & R_2 \arrow[drr] & \cdot \arrow[dr] & \cdots\cdots & \cdot \arrow[dl] & R_{k-1} \arrow[dll] & R_k \arrow[dlll] \\
 & & & \hat{R} & & &
\end{tikzcd}$$

Some Goals :

- Learning from Combining Studies
  - Can we combine the studies to get an overall estimate or conclusion ?
- Borrowing Strength
  - Can we use data from all the studies to help analyze individual studies?

Combining Information: Simplest Case

- $x_1, \ldots, x_n$ are a sample of independent observations from a $N(\mu, \sigma^2)$ distribution= the Normal Distribution with center μ and variance σ^2.

- $\bar{x} = \sum_{i=1}^{n} x_i/n$ is an estimate of μ. The variance of $\bar{x}$ is σ^2/n, and the standard deviation of the mean (also known as the standard error) is $\sigma/\sqrt{n}$.

- This simplest case highlights that the amount of 'information' you have is proportional to the number of replicates–things that when averaged together will 'even things out.'

Why we do what we do

- Simple case of the Law of Large Numbers: (Under regularity conditions) As the sample size gets very large, the average should converge towards the true mean.
  - As a coin is flipped more and more, the proportion of heads converges to 1/2.
- Central Limit Theorem: (Under regularity conditions) The distribution of a statistic that is an average of n numbers becomes approximately normal as n grows large.
- How do we average together numbers from different experiments, studies, etc.?

Melting Pot

What about taking all the original data from k studies and throwing it into a "single pot" ?

- Original data may not be available.
- Data from different studies may be on different scales or constructs.
- Loses information about the experimental conditions of the data.
- Can be dominated by one or two experiments with huge sample sizes.
- Heterogeneity may preclude combining.
- Your effective sample size is not as big as advertised.
  - If the sample sizes are $n_1, \ldots, n_k$, then the denominator of the variance of the mean is $n_1 + \cdots + n_k$.
  - You do not really have $n_1 + \cdots + n_k$ independent replicates that average together.

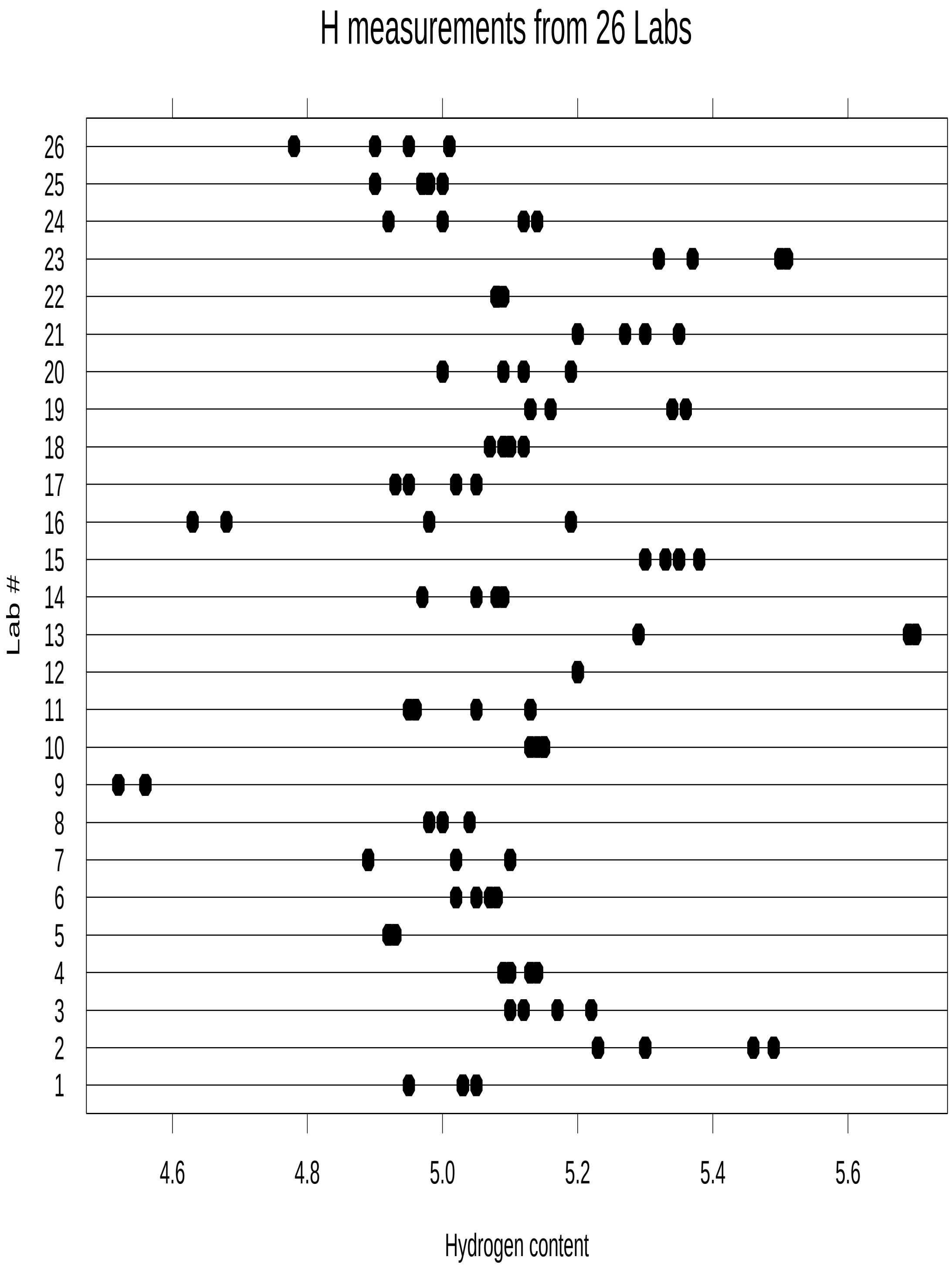
Round Robin of H data

The following pages show a measurements of 26 laboratories of H in a reference material and a convenient graphical representation (called a dotplot) of the data.

There seems be much more between-lab variation than within-lab variance; you do not have 4*26 (=104) effective replicates.

Lab				
1	4.95	5.03	5.03	5.05
2	5.46	5.49	5.23	5.30
3	5.22	5.17	5.12	5.10
4	5.09	5.14	5.10	5.13
5	4.92	4.92	4.93	4.92
6	5.02	5.05	5.07	5.08
7	5.02	4.89	4.89	5.10
8	5.04	5.00	4.98	5.00
9	4.52	4.56	4.56	4.56
10	5.13	5.14	5.15	5.15
11	4.95	4.96	5.13	5.05
12	5.20	5.20	5.20	5.20
13	5.70	5.69	5.29	5.29
14	5.09	5.05	5.08	4.97
15	5.33	5.30	5.35	5.38
16	4.68	4.63	4.98	5.19
17	4.93	4.95	5.02	5.05
18	5.07	5.09	5.10	5.12
19	5.16	5.13	5.36	5.34
20	5.19	5.09	5.12	5.00
21	5.27	5.20	5.35	5.30
22	5.08	5.09	5.08	5.08
23	5.32	5.51	5.37	5.50
24	5.12	5.14	5.00	4.92
25	4.90	5.00	4.97	4.98
26	4.90	4.78	4.95	5.01

Table 1: H Measurements from 26 Labs


H measurements from 26 Labs
Lab #
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
4.6
4.8
5.0
5.2
5.4
5.6
Hydrogen content

Mean of Means Approach

Sometimes it makes sense (especially when you have 26 studies) to treat the collection of study means $\bar{x}_1, \ldots, \bar{x}_k$ as you would treat a sample (Natrella 1963, Levenson et al 2000):

- Use the mean of means $\bar{\bar{x}} = \sum_{i=1}^{k} \bar{x}_i$ as the overall mean.

- Use the sample standard error of the mean of the k means $\bar{\bar{x}}$

$$\hat{u} = \sqrt{\frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2}{k(k-1)}}$$

  (as from a sample of k) as the standard uncertainty.

- The quantity $\hat{u}$ captures both within– and between–sample variability.

- Use as a 95 percent “confidence interval” $\bar{\bar{x}} \pm t^* \, \hat{u}$, where t^* is obtained from the Student’s t distribution with $k - 1$ degrees of freedom.

Mean of Means (Cont'd)

Some issues with this method:

- If k is small, then t^* can be very large (e.g. 12.7 for $k = 2$ and 4.3 for $k = 3$); it's very hard to get good variance estimates with only a few observations.

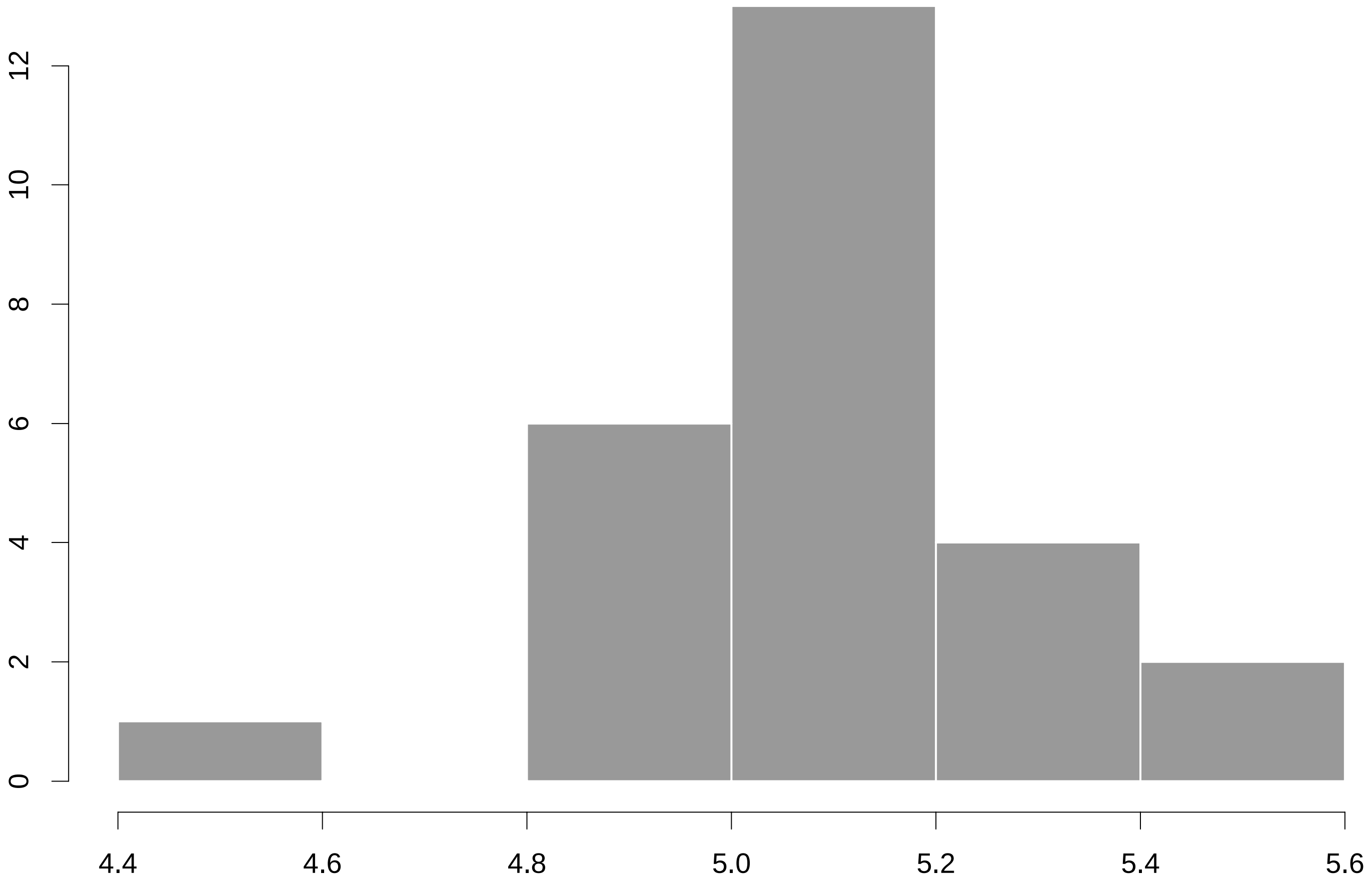
- "It rests on the assumptions that there is a population of methods whose biases are centered around zero and that the chosen methods are a random sample from the population."

  – (Levenson 2000)

- Ignores possible heterogeneities of within-lab variances and other factors.

Exploratory analysis

- It helps to examine the data graphically to see how and if things really do go together.
- Ideally the study means to be combined should like a sample from a normal distribution.
- The following page shows a histogram of the Lab Means.
- It looks ‘sort of’ normal, but not as normal as we’d like.



Histogram of 26 Lab Means of H

Stem-Leaf diagram of 26 Lab Means of H

```
4.5|5
4.6|
4.7|
4.8|7
4.9|12689
5.0|1225558
5.1|00245
5.2|058
5.3|47
5.4|39
```

A **Stem-Leaf** plot, which is essentially a hand-written histogram (with more information) is often a convenient way to examine the distribution of a modestly sized sample.

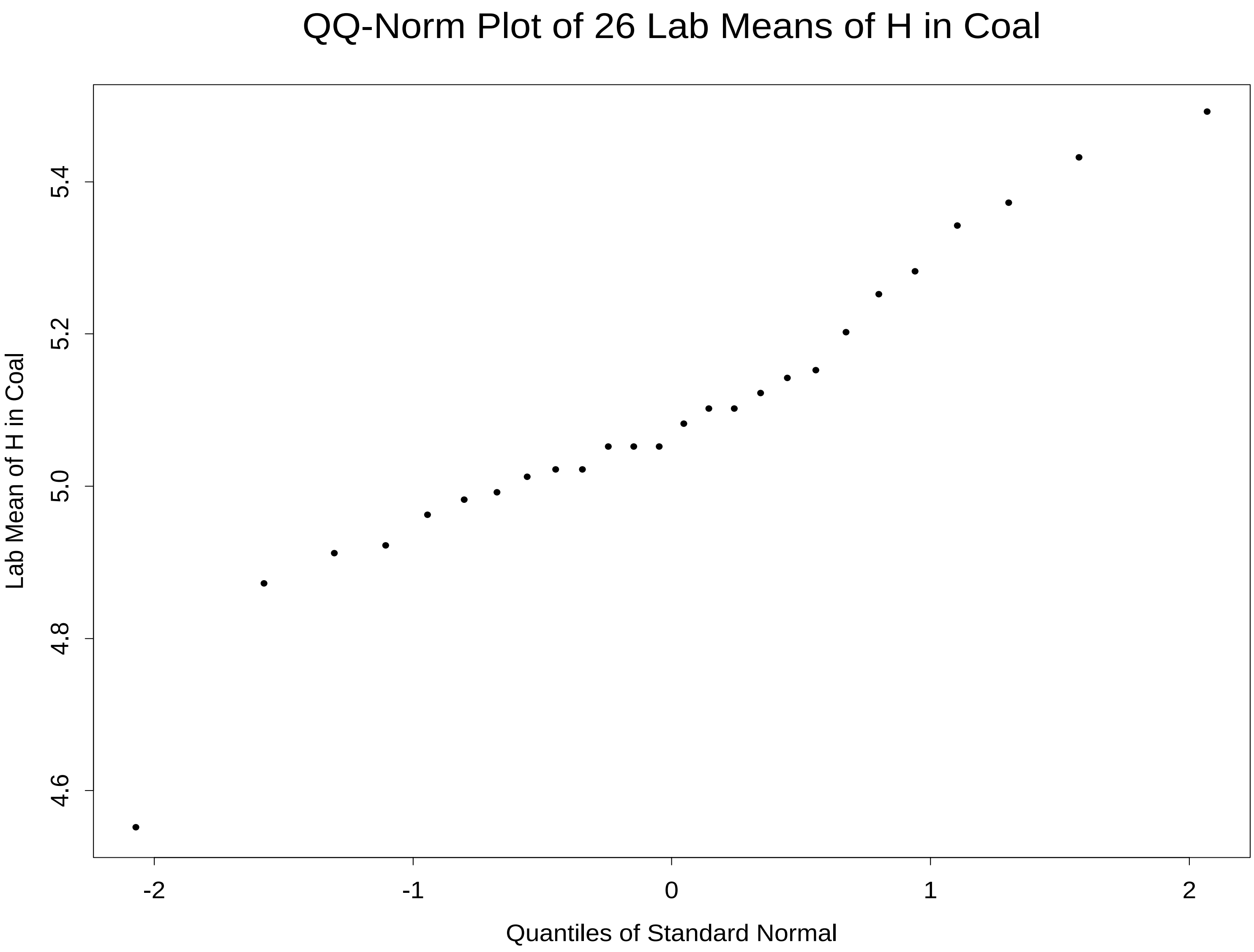
Here, we see that the distribution of lab means is in between normal and skewed to the right, with the 4.55 value sticking out.

QQ-Norm Plots

A **QQ-Norm** plot (available in some computer packages) is another way to check the normality of your data.

Deviations from a straight line indicate deviations from normality.

In the QQ-Norm plot on the next page, most of the data looks approximately normal except for at the ends; again, the lowest value is much lower than it should be under normality.


QQ-Norm Plot of 26 Lab Means of H in Coal
Lab Mean of H in Coal
4.6
4.8
5.0
5.2
5.4
Quantiles of Standard Normal
-2
-1
0
1
2

Hypothesis tests for Homogeneity?

There are hypothesis tests to for homogeneity..

- For two samples, there is the 2-sample t-test to check if the two studies have the “same” mean.
- One can do an ANOVA (Analysis of Variance) to test for inhomogeneity.
- A common test of homogeneity of effect sizes in the meta-analytic literature is a large sample test based on the Q statistic:

$$Q = \sum_{i=1}^{k} \frac{(\bar{x}_i - x_+)^2}{\hat{\sigma}^2(\bar{x}_i)}, \tag{1}$$

where

- $\bar{x}_i$ is the estimated mean for study i,
- $\hat{\sigma}^2(\bar{x}_i)$ is the estimated variance for $\bar{x}_i$,
- x_+ is the weighted mean where the weights are inversely proportional to the estimated variances.

Homogeneity Tests (Cont'd)

If all k studies have the same population mean and *all have large sample sizes*, then Q has an asymptotic effect size of a chi-square distribution with $k-1$ degrees of freedom.

The Q test is liable to blow up when some estimated variances are very small.

These tests should be treated with caution because:

- accepting the null hypothesis of homogeneity does not imply you should pool, and
- rejecting the null hypothesis does not imply that you should not pool.

Back to H

- Note: the H data all tested as having significant differences, especially the Q test, which is very susceptible to small variances. This test was not a factor in the ensuing analysis.

- For this particular data set, it was known that the labs used several different measurement techniques among them to measure coal.

- It was decided to group the labs according to two groups via types of measurement technique, and use a weighted average of the two technique means.

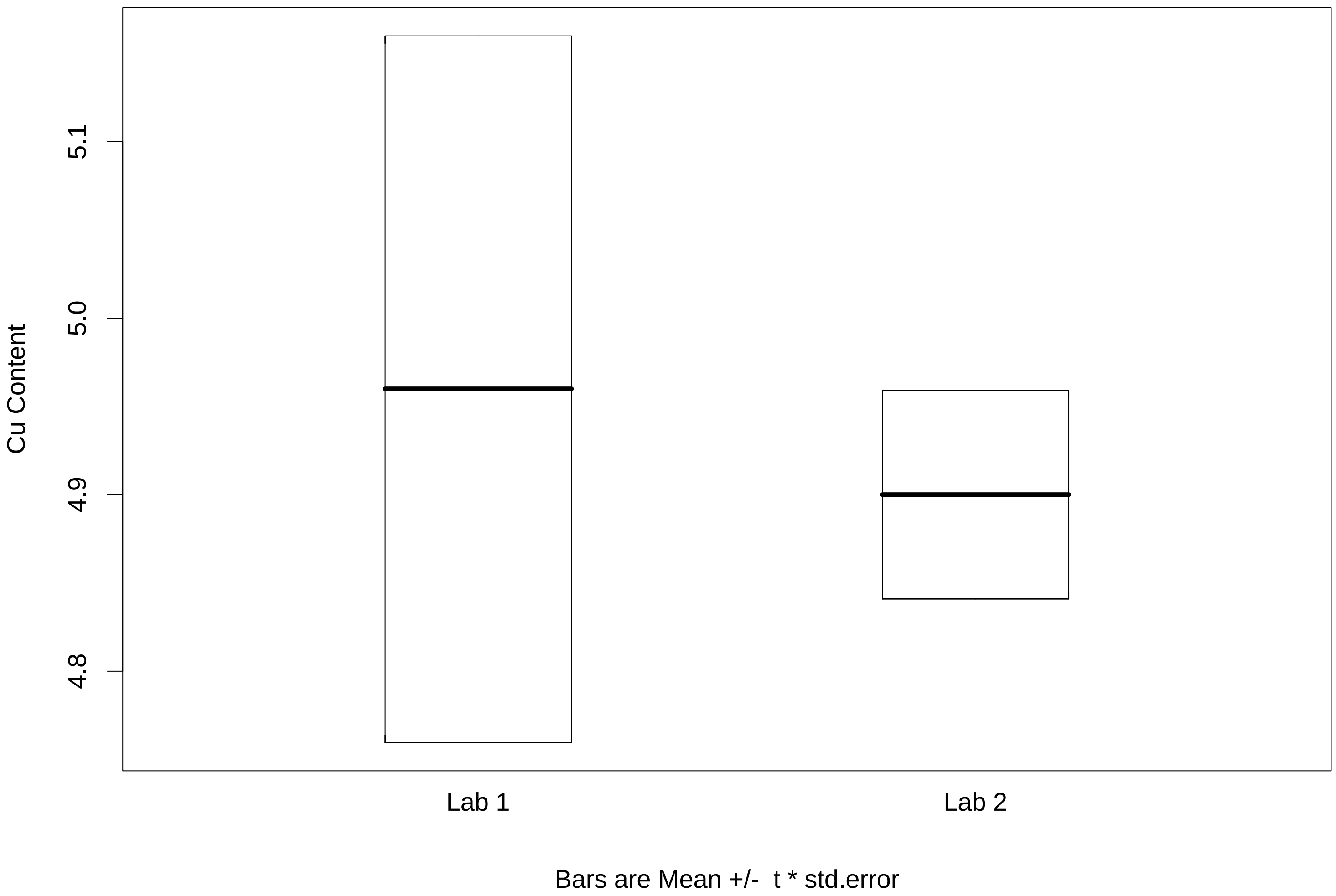
Reference Material Example: Cu

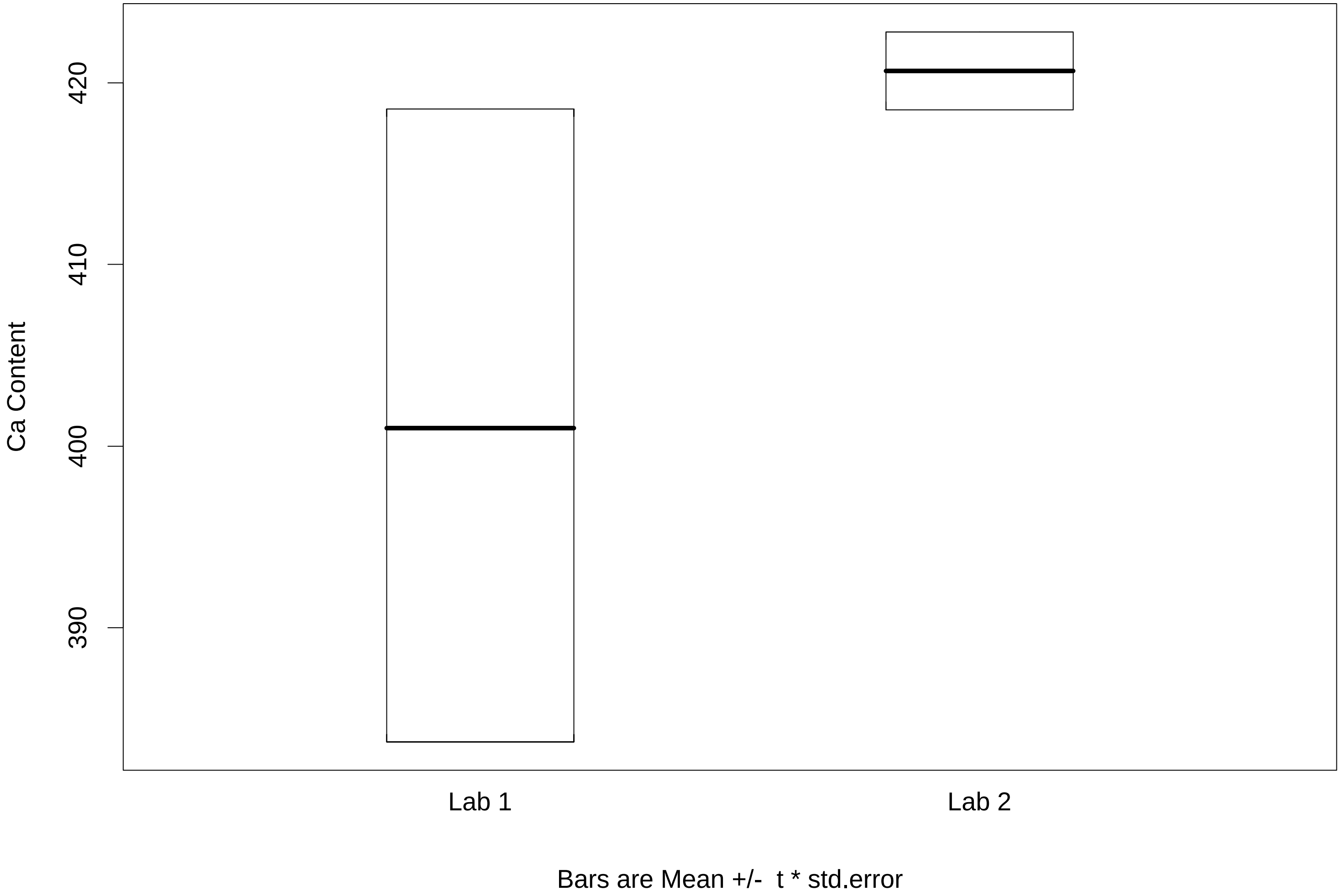
The following example is of Cu concentration in a reference material.

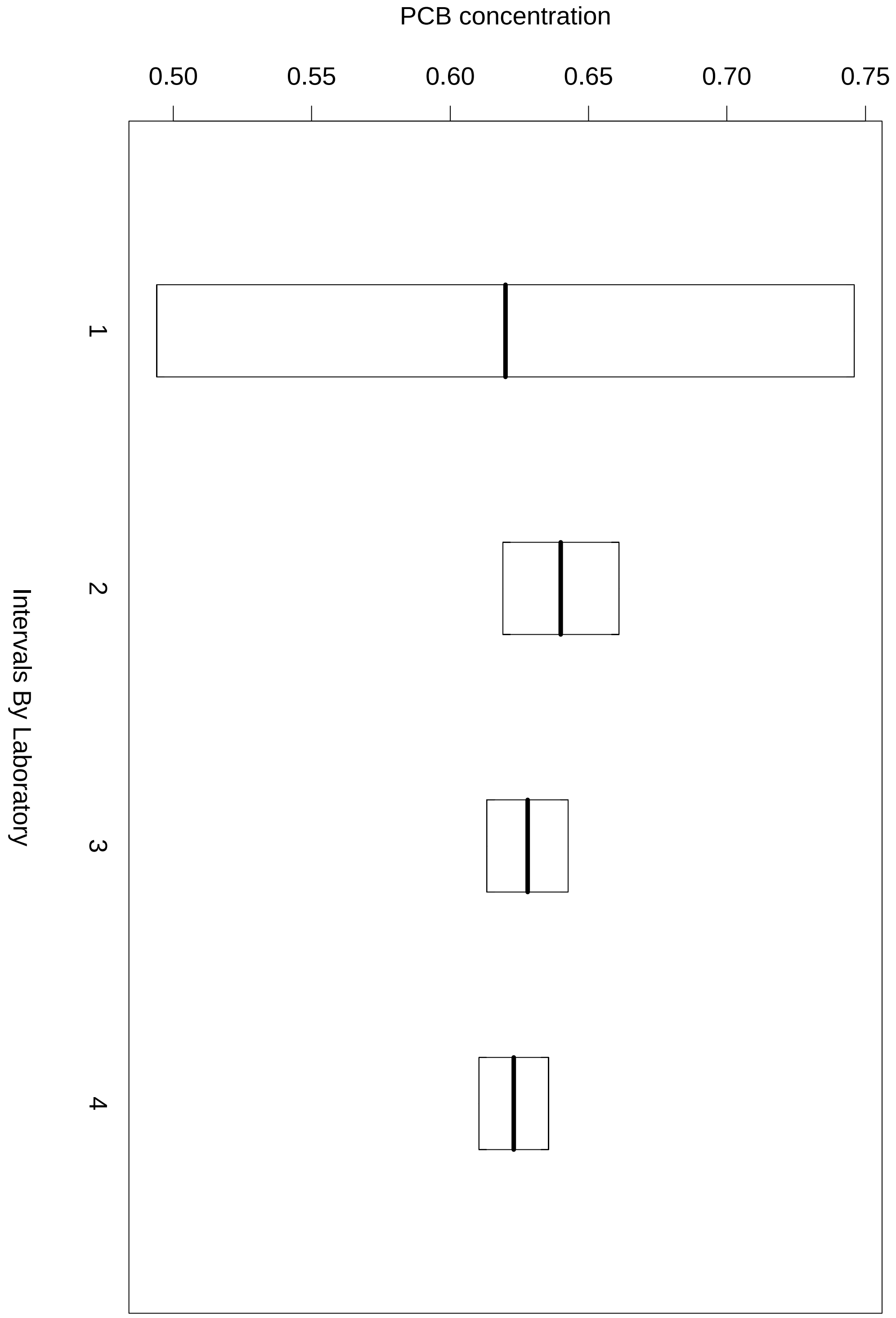
- Method 1 (from Lab 1) has mean $\bar{x}_1 = 4.96$ with a standard uncertainty of $u_1 =.09$ with 10 degrees of freedom.
- Method 2 (from Lab 2) has mean $\bar{x}_2 = 4.90$ with a standard uncertainty of $u_2 = .025$ with 7 degrees of freedom.

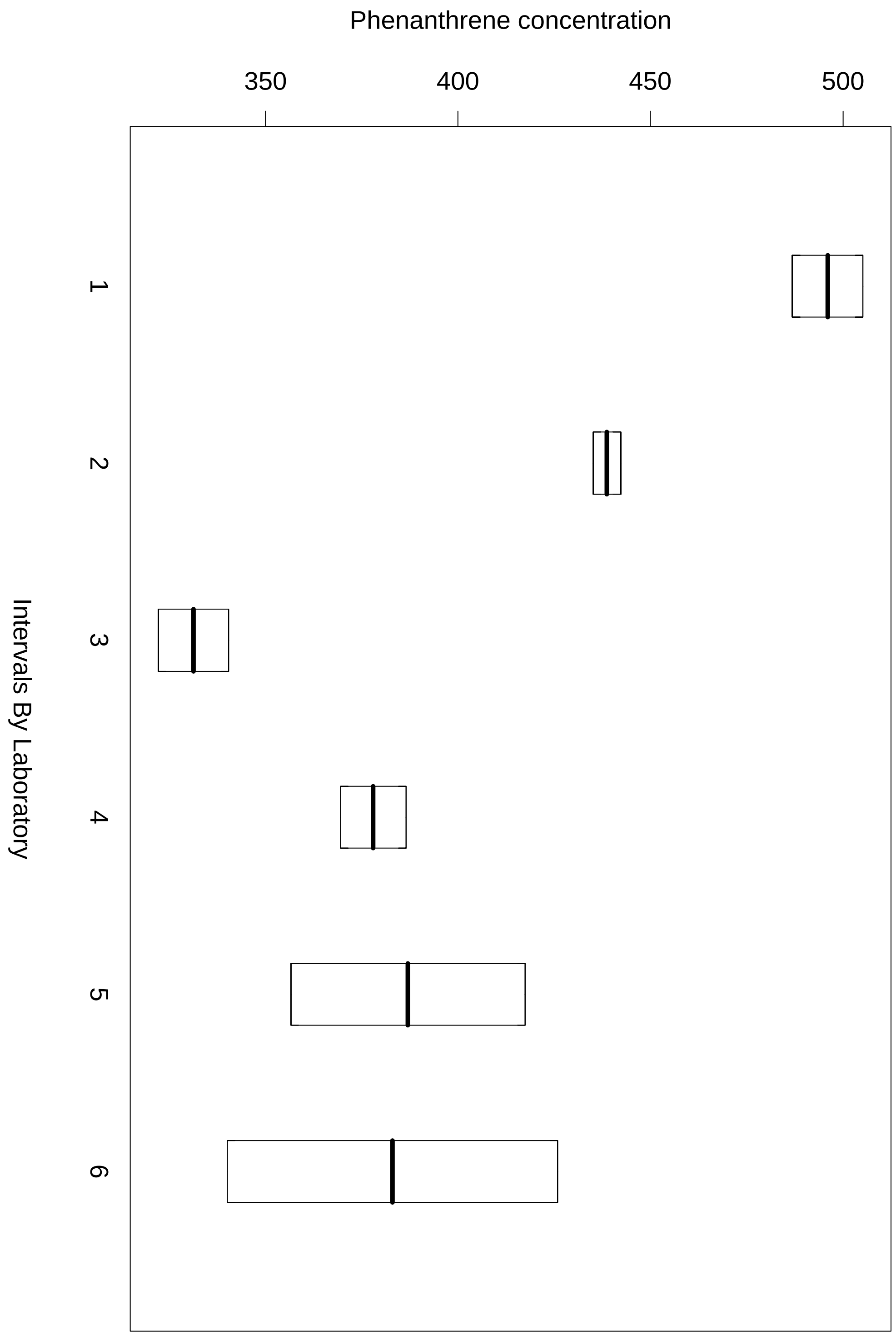
How the two sets of results compare with each other is diagrammed in the schematic on the next page.

Following that are diagrams of 3 other cases where we seek to comine the results. Think about what you would do in those cases and why.


Cu Content
5.1
5.0
4.9
4.8
Lab 1
Lab 2
Bars are Mean +/- t * std.error


Ca Content
420
410
400
390
Lab 1
Lab 2
Bars are Mean +/- t * std.error


PCB concentration
0.50
0.55
0.60
0.65
0.70
0.75
1
2
3
4
Intervals By Laboratory


Phenanthrene concentration
350
400
450
500
1
2
3
4
5
6
Intervals By Laboratory

Weighted Means

Suppose that there are summary statistics $y_1, \ldots, y_k$ (e.g. sample means) from k individual studies, with respective raw weights $w_1^*, \ldots, w_k^*$.

The weighted mean is

$$y_+ = \sum_{i=1}^{k} w_i^* \; y_i \; / \sum_{i=1}^{k} w_i^* .$$

If we define $w_i = w_i^* / \sum_{j=1}^{k} w_j^*$ to be the ith standardized weight, then

$$y_+ = \sum_{i=1}^{k} w_i \; y_i .$$

If we assume that the $y_1, \ldots, y_k$ are independent, and the $w_1, \ldots, w_k$ are fixed, then

$$Var(y_+) = \sum_{i=1}^{k} w_i^2 \; Var(y_i).$$

Without independence, one needs to account for the covariances between the y_i's:

$$Var(y_+) = \sum_{i=1}^{k} w_i^2 \, Var(y_i) + 2 \sum_{i<j} w_i w_j Cov(y_i, y_j).$$

Optimal Weighting by Inverse Variance

A standard result is that if $y_1, \ldots, y_k$ are independent and have respective known variances the $\sigma_1^2, \ldots, \sigma_k^2$, then the variance of the weighted mean is minimized by $w_i \propto 1/\sigma_i^2$.

If we denote the 'optimal' weights by

$$\hat{w}_i = (\frac{1}{\sigma_i^2}) \;/ (\sum_{j=1}^{k} \frac{1}{\sigma_j^2}),$$

then the variance of the weighted mean $\sum_{i=1}^{k} \hat{w}_i \; y_i$ is

$$\sum_{i=1}^{k} \hat{w}_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^4} \; \sigma_i^2}{(\sum_{j=1}^{k} \frac{1}{\sigma_j^2})^2} = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}{(\sum_{j=1}^{k} \frac{1}{\sigma_j^2})^2}$$

$$= (\sum_{j=1}^{k} \frac{1}{\sigma_i^2})^{-1} \qquad (2)$$

We will discuss later some problems we might face with using this formula.

Cu Example–Unweighted Mean

Remember that

- from Lab 1 we have mean $\bar{x}_1 = 4.96$ and standard uncertainty u_1=.09.
- from Lab 2 we have mean $\bar{x}_2 = 4.90$ and standard uncertainty $u_2 = .025$.

The unweighted (equally weighted) mean of $\bar{x}_1$ and $\bar{x}_2$ is 4.93. If we assume independence of $\bar{x}_1$ and $\bar{x}_2$, then because $w_1 = w_2 = \frac{1}{2}$, the standard deviation of the unweighted mean is

$$\sqrt{(\tfrac{1}{2})^2\ .09^2 + (\tfrac{1}{2})^2\ .025^2} = .047,$$

which is slightly more than half of the larger standard error. This is a common occurrence for the mean of 2 averages, because the sum of squares will be dominated by the large standard error.

Minimum Variance Weighted Mean–Cu Example

The weighted mean with minimum variance has weights inversely proportional to the respective variances:

$$\hat{w}_1 = \frac{1}{.09^2} / (\frac{1}{.09^2} + \frac{1}{.025^2}) = .07$$

$$\hat{w}_2 = \frac{1}{.025^2} / (\frac{1}{.09^2} + \frac{1}{.025^2}) = .93$$

The minimum variance weighted mean $\hat{w}_1 y_1 + \hat{w}_2 y_2 =$ 4.904, and using Formula (2), we find its standard deviation to be

$$(\frac{1}{.09^2} + \frac{1}{.025^2})^{-1/2} = .024.$$

This weighted mean is dominated by the more precise y_2, and it shows in the standard error, which is much smaller than for the unweighted mean.

Later we will discuss why we don't use this estimate.

Advantages of Weights

- Minimizes variance (optimal precision).
- Utilizes more available information.
- Gives more weight to “better” studies.
- Very widely used.

Disadvantages of Weights

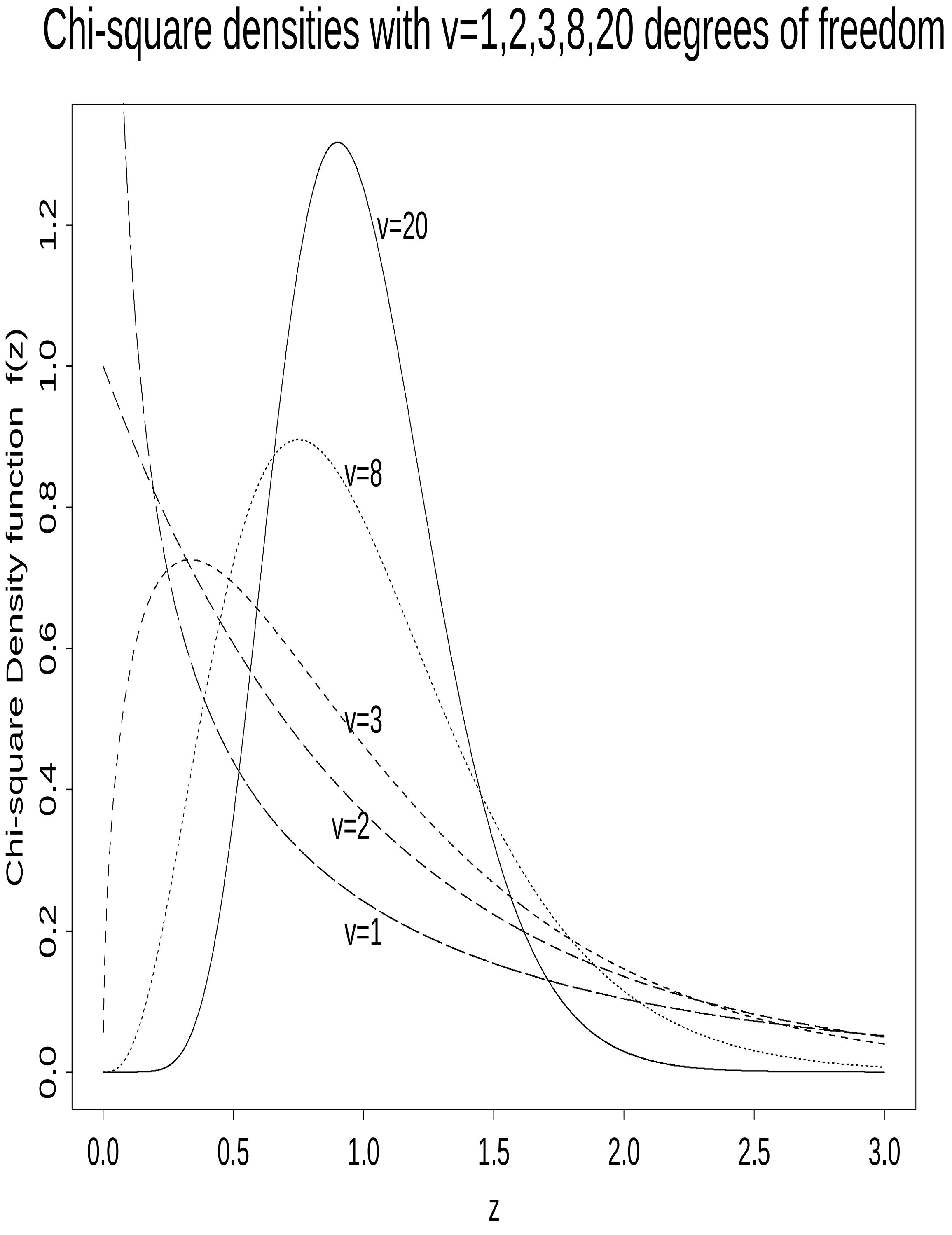
- Simplicity of unweighted mean vs. complex formulas for weighted mean.
- Precision may not indicate accuracy (low bias).
- Relative variances may not indicate relative precision:
  - Some experimenters may not include all sources of uncertainty.
  - Other experimenters may "pad" uncertainty to be "safe."
- Estimates of variance are notoriously inaccurate, especially for sample sizes.
- Weighted mean may be dominated by one or a few values with very small variances (lack of robustness).

On Chi-square distributions

Given n normally distributed observations, the sample variance will be distributed as a constant times a $\chi^2_{n-1}/(n-1)$ random variable (where χ^2_{n-1} is a chi-square with $n-1$ degrees of freedom).

The diagram on the next page depicts the probability densities of χ^2_ν/ν for $\nu = 1, 2, 3, 8, 20$. Some things to note:

- χ^2_ν/ν has mean 1 and variance $2/\nu$.
- The distributions are quite skewed for small degrees of freedom.
- Thus, variance estimates from low sample sizes have a good chance of being much too small (which is one reason by the t^* multiplier is so large for small degrees of freedom).
- Variance estimates that are much too small lead to estimated weights that are much too big.


Chi-square densities with v=1,2,3,8,20 degrees of freedom
Chi-square Density function f(z)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
v=20
v=8
v=3
v=2
v=1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
z

Inverse Variance Weighting (Cont'd)

If x_+ is the optimally weighted mean, and recall

$$Var(x_+) = (\sum_{j=1}^{k} \frac{1}{\sigma_i^2})^{-1}$$

- Thia formula holds when the weights w_i are known and fixed.
- Hence, it underestimates $Var(x_+)$ because in practice the σ_i^2's and thus the w_i's are not known but are estimated.
- Hence, there are errors in the estimated weights, which can be quite large for small sample sizes.

Inverse Estimated Variance Weighting

- If $s_i/\sqrt{n_i}$ is the estimated standard error for the ith sample mean, let the estimated weights be $\hat{w}_i^* = 1/(s_i^2/n_i)$, and
- $\hat{W}^* = \sum\limits_{i=1}^{k} \hat{w}_i^*$,
- then from (2), $Var(x_+)$ can be estimated by $1/\hat{W}^*$.
- Even if the estimates of the individual variances are themselves unbiased, then $1/\hat{W}^*$ as an estimator of $Var(x_+)$ is biased downward.
- A more conservative estimate that takes into account the variability of the weights is

$$\frac{\sum\limits_{i=1}^{k} n_i \; \hat{w}_i^*/(n_i - 4)}{\hat{W}^{*2}}. \tag{3}$$

- This formula is predicated on relatively large sample sizes.

Inverse Variance Weighting (Cont'd)

- Cochran (1954) has a more accurate (and quite complicated) estimate of $Var(x_+)$, also only for large sample sizes ($n_i > 8$).
- For smaller sample sizes, Cochran and Carroll (1953) and Cochran (1954) contain tables of empirical results showing tables much $1/W$ needs to be inflated to be a good estimate of $Var(x_+)$. (These tables don't seem to match?)
- They show that $1/\hat{W}$ not only underestimates $Var(x_+)$ for small sample sizes, but it can vastly underestimate $Var(x_+)$ for small sample sizes and a large number of studies k.

Weighting for Cochran

William Cochran (1909-1980) made many contributions to statistics, especially applied to agricultural experiments. Here are some comments from Cochran (1954) as highlighted by Rao (1984).

- The weighted estimator can be recommended when the individual studies have large sample sizes.
- Weighting may be preferable if the effect size variances are heterogeneous and are based on substantial degrees of freedom.
- His "experience with actual data has been that often there is little to choose between $\bar{x}$ *(unweighted mean)* and $\bar{x}_w$ *(weighted mean)*, but occasionally $\bar{x}_w$ wins handsomely."
- The unweighted mean is "preferable on account of its simplicity unless $\bar{x}_w$ brings a worthwhile gain in precision."

Weighting for Cochran (Cont'd)

- The unweighted mean is optimal when standard errors are equal, and is "highly efficient for small (sample sizes) when the (standard errors) do not vary much."
- "When the numbers of degrees of freedom in the individual experiments are less than 8, the weighted mean will seldom be more precise than the unweighted mean."

Partial Weighting

Cochran (1934,1954) suggests *partial weighting* as a way to weight for precision without allowing a single experiment from dominating the result:

- "The same weight is given to all experiments with relatively low values of s_i^2, this weight being $\bar{w}_p = 1/\bar{s}_p^2$, where $\bar{s}_p^2$ is the mean of the s_i^2 over those experiments that are chosen to have equal weight. Each of the remaining experiments receives its individual weight $w_i = 1/s_i^2$."
- For partial weighting, "the choice of the number of experiments that are to receive equal weight is to some extent arbitrary. A good working rule is to give equal weight to between 1/2 and 2/3 of the experiments."

While partial weighting may give sensible answers for combining (more than a few) study effects, it may be too complicated (politically as well?).

What are we really combining?

For the previous discussion of weighted means, we have often implicitly supposed that a group of k studies satisfies the following:

- The ith study has n_i measurements.
- If x_{ij} is the jth measurement of the ith study,

$$x_{ij} = \mu + e_{ij}, \tag{4}$$

- where $e_{ij} \sim N(0, \sigma_i^2)$.
- The $\{e_{ij}\}$ are mutually independent.

If $x_i = \Sigma_{j=1}^{n_i} x_{ij}/n_i$ is the ith study mean, then

- $\bar{x}_i \sim N(\mu, \tau_i^2)$, where
- $\tau_i^2 = \sigma_i^2/n_i$.

Here, all the study means are estimating the **same thing**; they would all converge towards the same number as sample sizes grow, which means:

- We don't include between–study variation.
- Weighting more precise estimates makes more sense.

Fixed and Random Effects Models

An often more realistic model for k summary statistics (e.g. Lab Means) $y_1, \ldots, y_k$:

$$y_i = \mu + b_i + \epsilon_i, \tag{5}$$

- The $\{b_i\}$ are independent can be interpreted as the biases.
- Each ϵ_i has mean 0 and variance τ_i^2 .
- The $\{b_i\}$ and $\{\epsilon_i\}$ are all independent.
- The $\{\epsilon_i\}$ are often assumed to have normal distributions.

- How do we interpret the b_i's?
  - In a random effects model, the $\{b_i\}$ come from a distribution with mean 0 and variance σ_b^2. The σ_b^2 is the between-study variance.
  - In a fixed effects model, the $\{b_i\}$ are not random, but fixed.

Fixed or Random Effects Models

Some statisticians feel the distinction between fixed and random effects is not worth worrying about, but here it is useful to think about what you're really estimating.

- In a random effects models, the b_i's will jump around (0), and may change next month.
- In a fixed effects model, the b_i's stay the same.
- In the random effects models, you need to figure in the extra uncertainty caused by the variation in the random b_i's.
- For some fixed effects model, you may not need to figure in the variation in the b_i's if what you are really after is some theoretical center μ, but merely the average $\mu + \sum_{i=1}^{k} b_i/k$.

Fixed and Random Effects (cont'd)

- For a NIST SRM, we presume the b_i's are close to 0, and ideally they average out (whether they are fixed or random).
- However, for some fixed effects models, the b_i's do not have to be close to each other or to 0.
- For example, there may be stratified studies where we may be able to estimate a quantity for individual groups very precisely. If we then want an overall average, then the weighted mean of the group means with a (possibly) very small variance is appropriate.
  - E.g. an insurance company may have payout data on various demographic groups to be combined into an overall payout.

When combining multiple methods, it would be nice if ...

Some desirable properties for techniques to combine multiple methods :

- Reasonable intervals and coverage probabilities for real problems
- If from 2 methods, should include both method means
- Interval should at least intersect the component intervals
- Follows, or agrees with ISO G.U.M.
- Results can be used for secondary and future analyses
- Solution should depend continuously on the data
- Weights should properly reflect relative scientific status of component methods.
- Scales up from 2 methods nicely (approach smoothly case with large number of methods)

- Not penalized for a larger number of methods
- Good whether or not there is between-method variance
- Accounts for dependence between methods
- Provision for using prior information
- Face validity
- Statistically justifiable
- Recognize that the input uncertainties are themselves uncertain (random), especially when used to determine weights

Unfortunately, no single technique uniformly satisfies all these conditions.

Note: The results combined can be from multiple methods, or from multiple labs. One "method" may an average from a round robin, or be itself a combination of different methods or labs.

Random Effects Models Refresher

Random Effects model: k summary statistics (e.g. Lab Means) $y_1, \ldots, y_k$ follow the model:

$$y_i = \mu + b_i + \epsilon_i, \tag{6}$$

- The $\{b_i\}$ are independent and come from a distribution with mean 0 and variance σ_b^2. The b_i can be interpreted as the biases, and σ_b^2 as the between-study variance.
- ϵ_i has mean 0 and variance τ_i^2 .
- The $\{b_i\}$ and $\{\epsilon_i\}$ are all independent.
- The $\{\epsilon_i\}$ are often assumed to have normal distributions.

Mandel-Paule procedure

In the random effects model, $Var(y_i) = \sigma_b^2 + \tau_i^2$. Define

$$w_i = 1/(\sigma_b^2 + \tau_i^2),\ i = 1, \ldots, k \tag{7}$$

Then, the w_i's are the weights that minimize the variance of the weighted mean because w_i is inversely proportional to $Var(y_i)$.

To estimate σ_b^2, Paule and Mandel (1982) note that if y_+ is the weighted mean with such weights $\{w_i\}$, then

$$E[\sum_{i=1}^{k} w_i(y_i - y_+)^2] = k - 1. \tag{8}$$

Solving (8) iteratively for σ_b^2 provides a solution $\tilde{\sigma}_b^2$ for σ_b^2 and thus for the w_i's and y_+ as well.

Schiller-Eberhardt

Schiller-Eberhardt (1991) procedures produce uncertainty statements that take into account the differences between method means; they use expanded uncertainty intervals of the form $y_+ \pm (k\ u_c + B)$, where

- y_+ is the Mandel-Paule estimate (weighted) or the unweighted avearage of the method means,
- k is an appropriate t multiplier garnered from use of the Satterthwaite approximation.
- u_c is the standard uncertainty calculated as if the weights w_i are constant.
- B is a bias adjustment that ensures that the interval covers all of the y_i's.
- The Schiller-Eberhardt grows larger with each method added, hence it is best for only 2-3 methods.

“Type B on Bias” Approach

- The BOB (Type B on Bias) approach of Levenson produces intervals for the mean similar to those of Schiller-Eberhardt. However, it follows the ISO Guide by incorporating the uncertainty of the bias into the combined standard uncertainty.

- It does so by modeling the bias $\bar{b}$ of the unweighted mean as a Type B distribution. Usually, that bias is modeled as a rectangular (uniform) distribution centered at 0 and having the same range R as the collection of study means; such a distribution has variance $R^2/12$.

- Thus, for $k = 2$ studies with study means $\bar{x}_1, \bar{x}_2$, $R = |\bar{x}_1 - \bar{x}_2|$; the type B uncertainty $u_B = R/\sqrt{12}$ is then incorporated with the other components of uncertainty in the usual way.

BOB: Example

Back to the Cu example:

- Method 1 has mean $\bar{x}_1 = 4.96$ with a within standard uncertainty of u_1 =.09 with 10 degrees of freedom.

- Method 2 has mean $\bar{x}_2 = 4.90$ with a within standard uncertainty of $u_2 = .025$ with 7 degrees of freedom.

- $R = |4.96 - 4.90| = .06$, so $u_B = .06/\sqrt{12} = .017$.

- The within variance of $\bar{x} = (\bar{x}_1 + \bar{x}_2)/2$ is $(\frac{1}{2})^2[.0025^2 + 0.09^2] = .045^2$.

- The combined uncertainty is $u_c = \sqrt{.045^2 + .017^2} =$ .048.

- Procedures to calculate the appropriate degrees of freedom using the Welch-Satterthwaite formula are in Levenson et al (2000). Here, the approximate d.f. is calculated to be around 13, leading to a multiplier of $k = t_{13,.025} = 2.16$.

- The expanded uncertainty is $U = k * u_c = 2.16 * .048 = .104$, leading to the interval 4.93 ± 0.104.

BOB Comments

- Distributions other than the rectangular can be used.
- Levenson et al (2000) also contains a Bayesian justification for BOB.
- Both the Schiller-Eberhardt and the BOB method are appropriate only for small $k \leq 5$ (ideally 2 or 3?).

Normal Random Effects

We go back to the previous random effects model and add the assumption of normally distributed biases.

Suppose that there are k labs, and the ith study does n_i measurements. If x_{ij} is the jth measurement of the ith study, then

$$x_{ij} = \mu + b_i + e_{ij}. \tag{9}$$

- Suppose that the bias b_i come from a $N(0, \sigma_b^2)$ distribution.
- The e_{ij} come from a $N(0, \sigma_i^2)$ distribution.
- The b_i's and e_{ij}'s are all mutually independent.

Rukhin and Vangel (1998) produce the Maximum Likelihood Estimator of μ under the above model and show that the Mandel-Paule estimator x_+ is a good approximation to the MLE under that model.

Rukhin-Vangel–Mandel-Paule

Under the normal random effects model (9), let

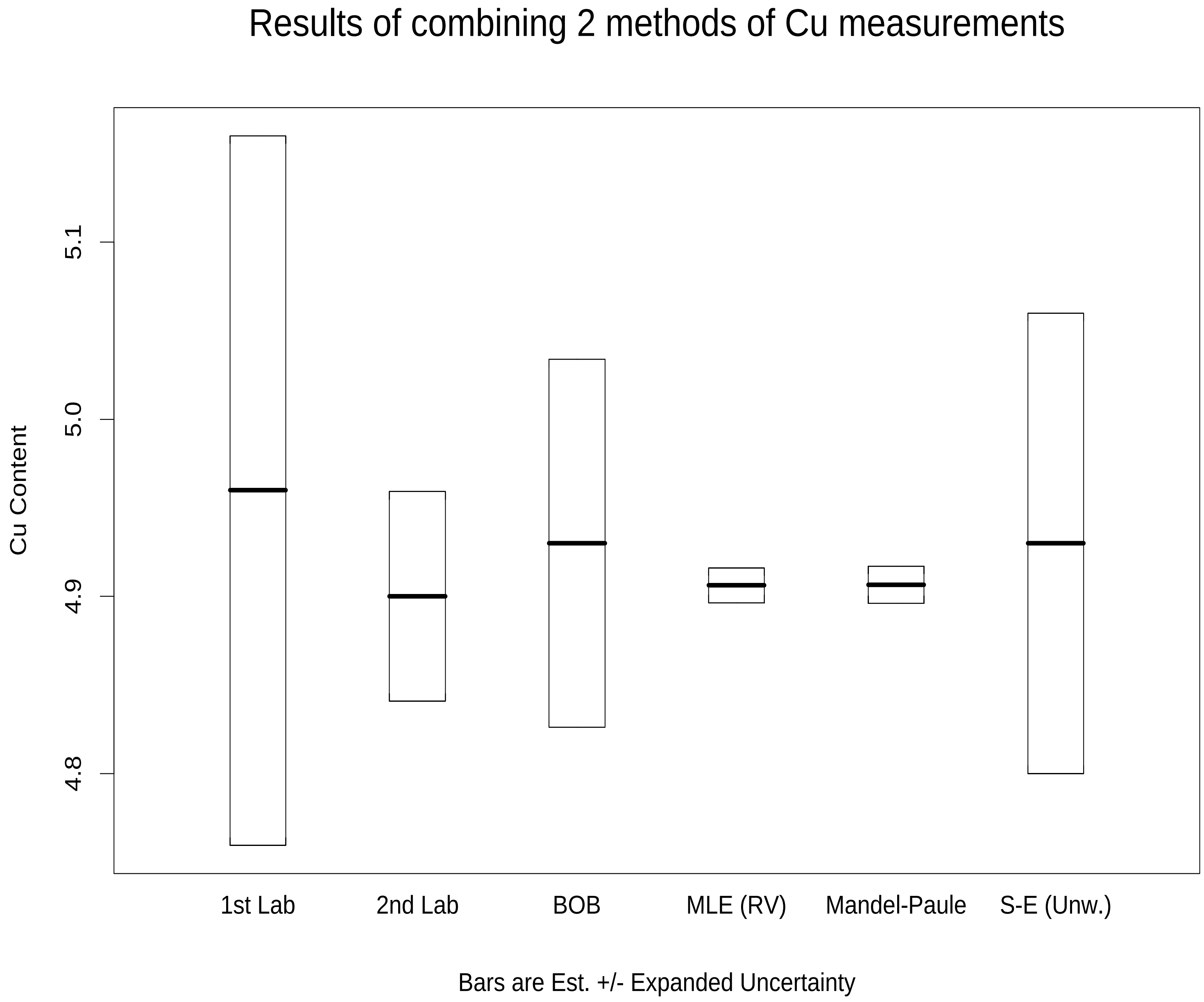
- $\bar{x}_i = \Sigma_{j=1}^{n_i} x_{ij}/n_i$ (ith study mean).
- $t_i^2 = s_i^2/n_i$ (estimated standard error of $\bar{x}_i$).
- $\tilde{\sigma}_b^2$ is Mandel-Paule estimate of σ_b^2.
- $w_i = 1/(\tilde{\sigma}_b^2 + t_i^2),\ i = 1, \ldots, k$ are the Mandel-Paule weights.
- $x_+ = \sum\limits_{i=1}^{k} w_i \bar{x}_i$ is the Mandel-Paule estimator.

Rukhin and Vangel (1998) provide an approximate estimator of the asymptotic variance of x_+ and the MLE (predicated on a large k):

$$\{ \sum_{i=1}^{k} \frac{(\bar{x}_i - x_+)^2}{(\tilde{\sigma}_b^2 + t_i^2)^2} \} [\sum_{i=1}^{k} \frac{1}{\tilde{\sigma}_b^2 + t_i^2}]^{-2} \qquad (10)$$

A confidence interval based on that asymptotic variance may be appropriate for large k (at least 5–6). For smaller k, such intervals may tend to be small.

Don't worry–the Dataplot function "Consensus Means" does all the work!

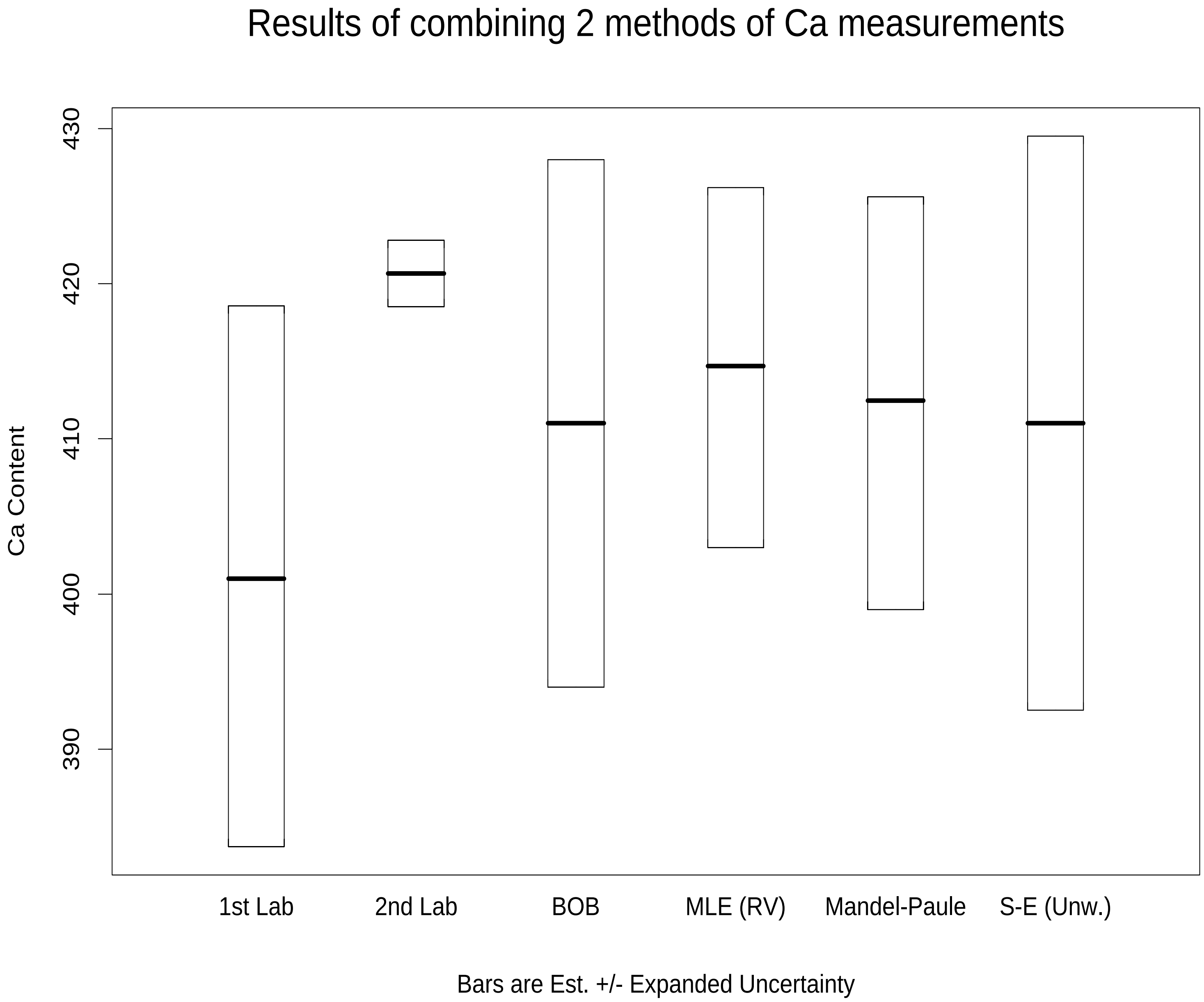

Results of combining 2 methods of Cu measurements
Cu Content
5.1
5.0
4.9
4.8
1st Lab
2nd Lab
BOB
MLE (RV)
Mandel-Paule
S-E (Unw.)
Bars are Est. +/- Expanded Uncertainty

Combining procedures: Cu Example

The previous page diagrams the first of several examples with respective intervals with 95% confidence.

- This example has just **two** methods, which favors methods like BOB rather than the Mean of Means or MLE (Maximum Likelihood method) method.
- The Mean of Means interval would be too big to fit into this picture because having only 1 degree of freedom would lead to a t multiplier of over 12.
- The two labs show relative agreement with each other (one interval sort of nested within the other), which is rather rare; in such situations:
  - the Mandel-Paule and MLE feel confident that the two methods are indeed estimating the same thing, and thus give more weight to the more precise estimate, leading also to a tighter estimate (remember the formula for variance of a weighted mean).

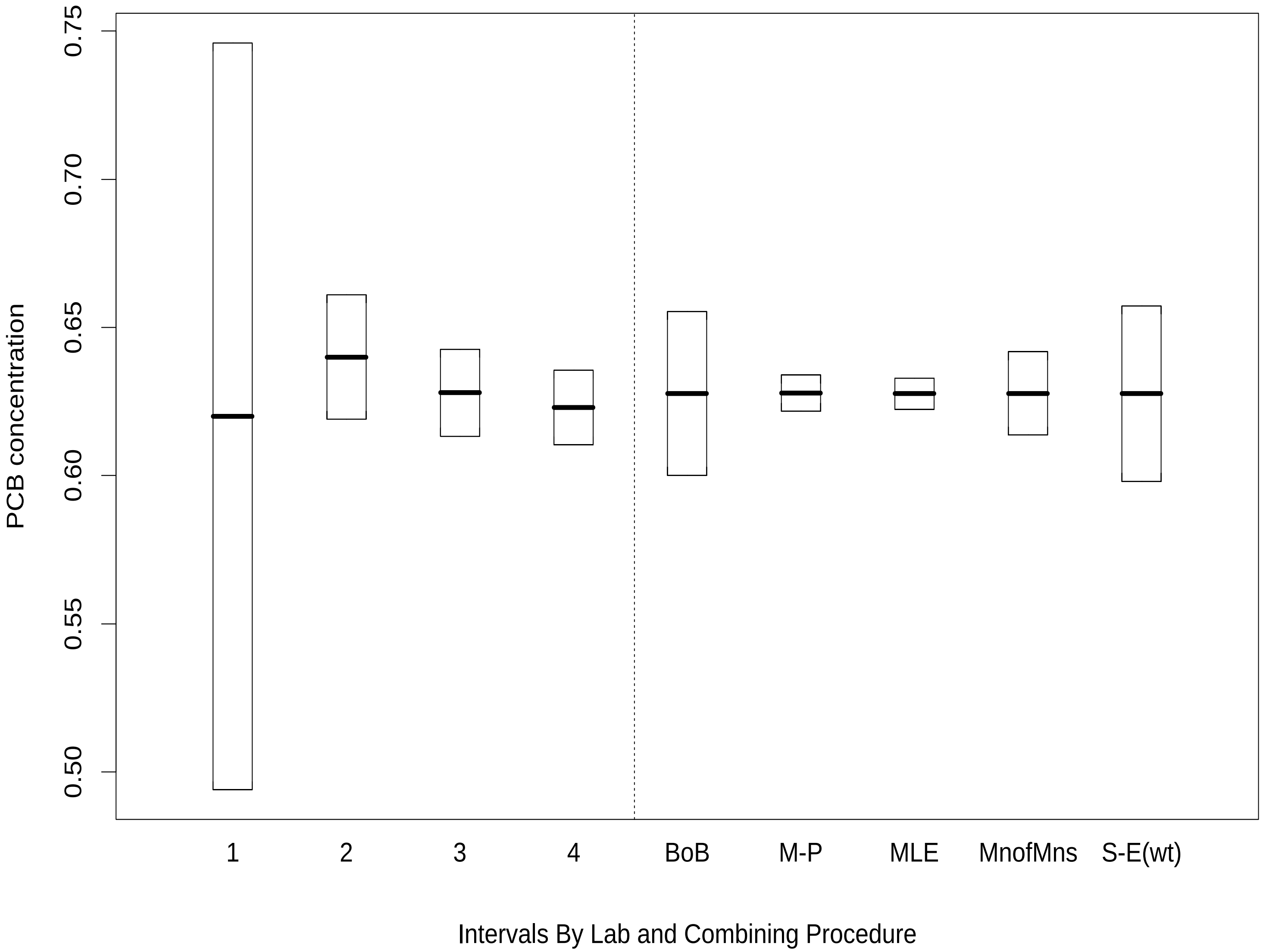
- The BOB and unweighted Schiller-Eberhardt, by design, equally weight the estimates (the weighted Schiller-Eberhardt would give more weight to Lab 2 but with similarly wide intervals).
- The BOB and unweighted Schiller-Eberhardt are quite similar, and their intervals are so large not only to cover each lab mean, but also because the large uncertainty of Lab 1 is given equal weight as the smaller one in Lab 2.


Results of combining 2 methods of Ca measurements
Ca Content
430
420
410
400
390
1st Lab
2nd Lab
BOB
MLE (RV)
Mandel-Paule
S-E (Unw.)
Bars are Est. +/- Expanded Uncertainty

Combining procedures: Ca Example

The previous graph shows relative disagreement between the two labs; in such cases, the intervals look more similar because:

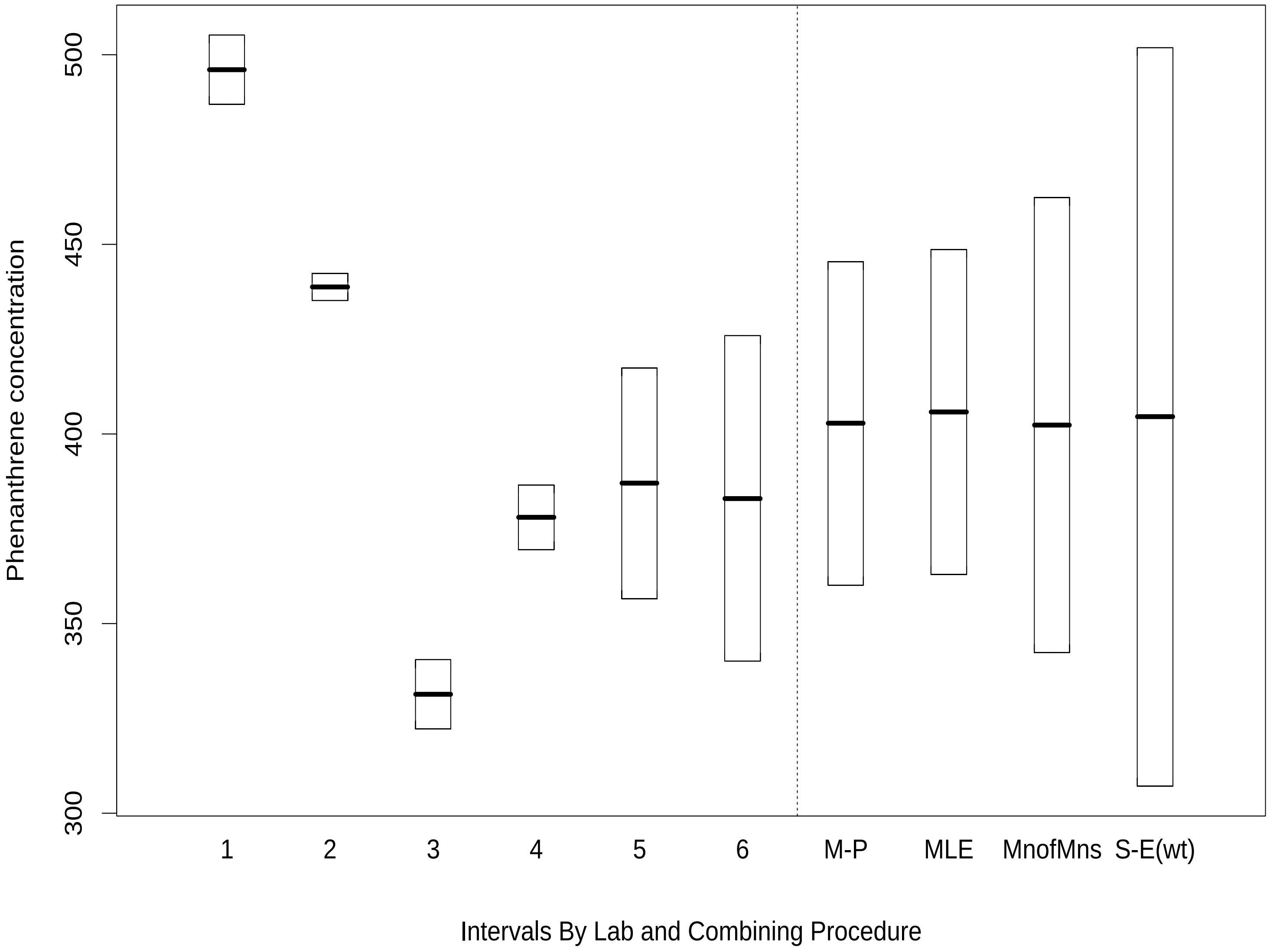
- All the combining procedures see the disagreement and include a large between-method variance in the uncertainty
- Although the Mandel-Paule and MLE give more weight to the more precise measurement, the weight distribution is much more even than the previous case.


PCB concentration
0.75
0.70
0.65
0.60
0.55
0.50
1
2
3
4
BoB
M-P
MLE
MnofMns
S-E(wt)
Intervals By Lab and Combining Procedure

Combining procedures: PCB Example

There are 4 methods here, which is reaching BOB's limit for usefulness.

- It is relatively rare to see this degree of agreement with this many methods.
- The MLE and Mandel-Paule procedures give more weight to the most precise estimates, producing tighter intervals.
- The BOB and Schiller-Eberhardt have to include the large uncertainty from Lab 1.


Phenanthrene concentration
300
350
400
450
500
1
2
3
4
5
6
M-P
MLE
MnofMns
S-E(wt)
Intervals By Lab and Combining Procedure

Combining procedures: Phenanthrene Example

- There are 6 methods here, which is too many for BOB.
- The disagreement among labs results in sizeable intervals for all.
- The Schiller-Eberhardt interval has to extend itself to cover all the Lab means. The S-E intervals tend to get larger as more labs are added, unless they all agree.

Borrowing Strength–Empirical Bayes

When the quantities are not the same, but similar in some ways, information can be combined to make each individual estimate better.

Suppose that

$$X_i \sim N(\theta_i, V_i), \quad i = 1, \ldots, k.$$

The usual maximum likelihood estimator for $\theta = (\theta_1, \ldots, \theta_k)$ is $X_i, \ldots, X_k$.

When the variances V_i are equal and known ($V_i = V$), it can be advantageous to estimate each θ_i by by an estimator of the form

$$\hat{X} = [1 - B]X_i + B\bar{X}, \tag{11}$$

where $\bar{X} = \sum_{i=1}^{k} X_i$, and

$$B = \min(1, (k-3)\ V/S),$$

with $S = \sum_{i=1}^{k} (X_i - \bar{X})^2$.

The estimator $\hat{X}$ is called a James-Stein estimator or a **Shrinkage** estimator because it shrinks each estimate of θ_i towards the group mean (it is a linear combination of the individual mean and the group mean).

It is an example of an **Empirical Bayes** procedure.

Note: Empirical Bayes procedures can also be helpful even in situations where the V_i are not all equal and known.

1970 Baseball Example

Player	batting average (1st 45 AB)	batting average (rest of season)	Empirical Bayes Estimate
Clemente	.400	.346	.290
F. Robinson	.378	.298	.286
F. Howard	.356	.276	.281
Johnstone	.333	.222	.277
Berry	.311	.273	.273
Spencer	.333	.279	.273
Kessinger	.289	.263	.268
Alvarado	.267	.210	.264
Santo	.244	.269	.259
Swoboda	.244	.230	.259
Unser	.222	.264	.254
Williams	.222	.256	.254
Scott	.222	.303	.254
Petrocelli	.222	.264	.254
E. Rodriquez	.222	.226	.254
Campaneris	.200	.285	.249
Munson	.178	.316	.244
Alvis	.156	.200	.239

Baseball Example–Discussion

- Example (taken from Efron and Morris (1975)). Listed on the table of the previous page are the batting averages of the 18 players who were listed as having 45 at-bats in the April 26 or May 3 (1970) editions of the *New York Times*.
- The conventional MLE estimates (given no other information) of θ_i= the ith player's batting batting average for the rest of the season is simply their current batting average (after 45 at bats).
- The Empirical Bayes estimate pools information from all the players to help estimate each player's future batting average.
- Let y_i be the ith players batting average, or rather, batting proportion during his first 45 at-bats (e.g. $y_1 = .400$).
- For technical reasons, we find it advantageous to

use a variance-stabilizing arcsin transformation

$$x_i = \sqrt{45}\ \arcsin(2y_i - 1),$$

so we now have a sample $x_1, \ldots, x_{18}$ (with approximately approximately unit variance).

- We shrink the individual estimates back toward the group means with

$$\hat{x}_i = [1 - B]x_i + B\bar{x}, \qquad (12)$$

with

$$B = \min(1, (k - 3)\ /S),$$

with $S = \sum\limits_{i=1}^{k} (x_i - \bar{x})^2$.

- Transforming $\hat{x}_i$ back to the original units give the Empirical Bayes estimates of future batting averages given in the table.

- The Empirical Bayes estimates are closer to the conventional estimates for 15 of the 18 batters!

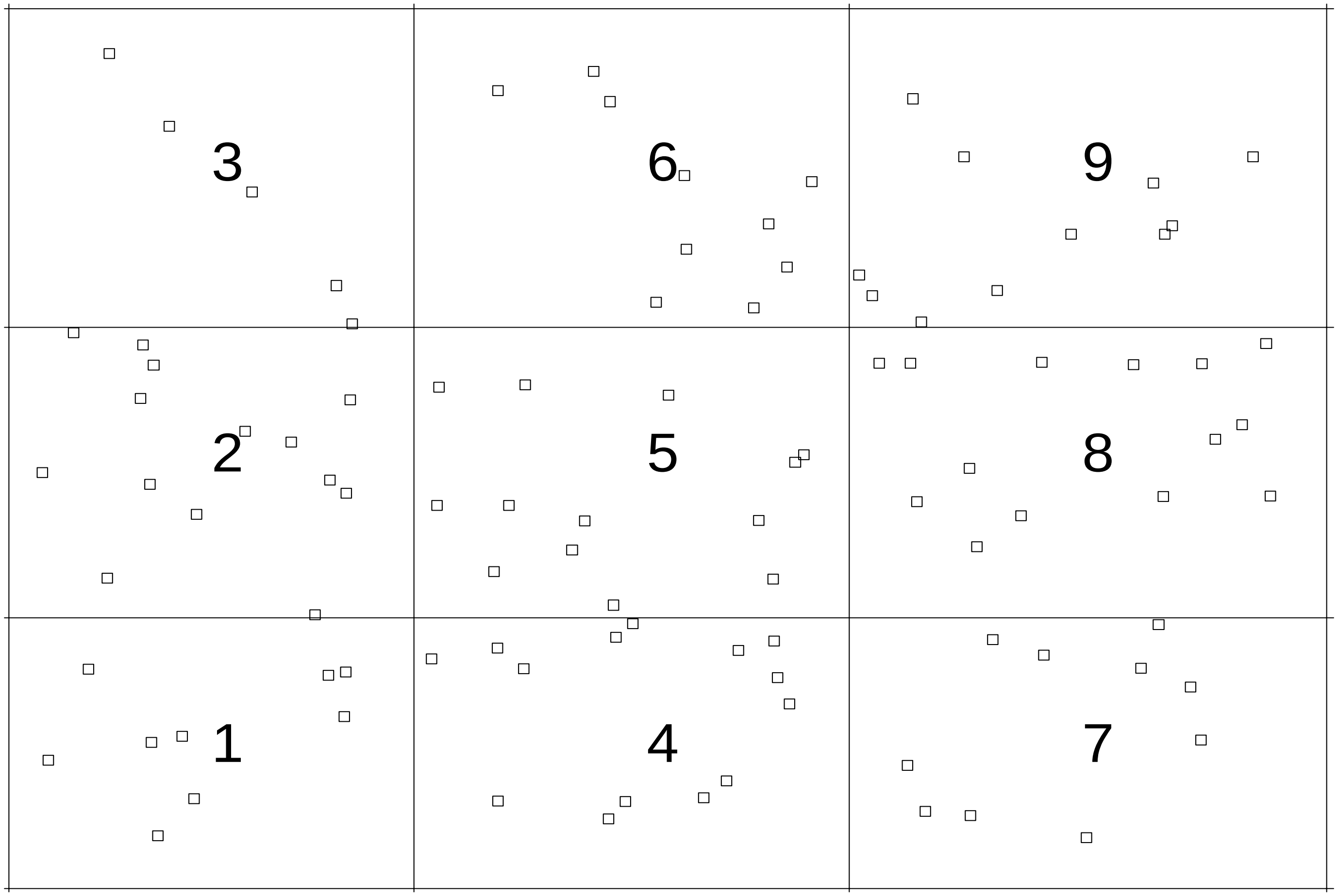
Bayes isn't just Empirical

Baseball fans already know that better estimates of future batting average would incorporate prior knowledge of the player's career.

In fact, far better estimates of a player's future batting average would be
A * [Career Average] + B*[Current Average] + C*[Pooled Average of players],
with A having by far the largest weight.

This indicates the value of a **Bayesian** approach that can incorporate **prior information** in a rigorous way.

Hypothetical Location of Fire Alarm Boxes in 9 Districts



An Alarming Example

- A more serious application of Empirical Bayes methods was done by the RAND Corporation for the New York City Fire Department in the 1970's.
- The previous page shows a (very) simplified schematic of the possible location of fire alarm boxes in a region of the Bronx.
- The region is divided into 9 neighborhoods considered to be approximately homogeneous within each region.
- The Fire Department wanted to know the likelihood that a box-reported alarm indicates a structural (serious) fire given the alarm box location (in allocating how many fire engines to send on an initial response).
- Each estimated probability was a weighted estimate of the neighborhood average probability and the average for that particular alarm box.

- The relative few previous alarms for most alarm boxes precludes producing a good estimate using data from that box only. Pooling data from surrounding boxes produce much better estimates.

Empirical Caveats

- “Although blind applications of these methods would gain little in most instances, the statistician who uses them sensitively and selectively can expect major improvements.” (Efron and Morris, 1975)
- Use of empirical Bayes estimated means would of course be silly in many cases (e.g. estimating the concentrations of 30 different analytes in an SRM).
- However, they can be useful where pooling makes sense, and can help make up for lack of information in individual cases.
- Borrowing strength can also be useful for helping estimating uncertainties where such information may be lacking for individual cases.

Simultaneous Intervals– SRM Example

(Simplified, with some numbers changed)

- There are two lots of a proposed reference material. To check that the two lots are the same, 10 different elements are measured for each lot.
- If we do two-sample t-tests for each element separately, there are statistically significant differences in 2 of the elements. 95 percent confidence intervals for these elements are:
  - Element 1: (-.035 , -.003)
  - Element 2: (-.0067, -.0017)
- Can we adjust these results to take into account doing 10 comparisons at once?

Significance Levels and P-values

- A Hypothesis test with a significance level of .05 means that if it sees the Null Hypothesis for real many times, it would reject it 1 out of 20 times on the average.
- (If you throw 4 dice many times, some of your throws will be 6666.)
- Of course, the more individual significance tests you do, the more likely that you will turn up a false positive by chance.
- Multiple comparisons were developed so that an entire set of hypotheses or statments would have a desired significance.
- They are useful for guarding against "data dredging" (and finding spurious significant effects).

Confidence intervals and hypothesis tests

- Note: There is often a duality between confidence intervals and hypothesis tests such that the statement that $H_0 : \mu = 0$ is not rejected by a hypothesis test with significance level .05 is equivalent to that the statement that a 95% Confidence interval for μ includes 0.
  - Confidence intervals give more information than mere hypothesis tests, e.g. direction and magnitude of the effect.
- Thus, one sometimes desires a set of simultaneous confidence intervals with a desired probability that they *all* cover their respective estimands, e.g. this set of confidence intervals include all the right values 95% of the time.

Bonferroni Method

A simple way of doing multiple comparisons is use the Bonferroni rule.

- Suppose you have a set of k statements $S_1, \ldots, S_k$ (e.g. k different null hypotheses $H_{01}, \ldots, H_{0k}$) who have respective significance levels $\alpha_1, \ldots, \alpha_k$.
- If H_{0i} is true, let S_i be the event that its test statistic comes out significant; $P(S_i|H_{0i}) = \alpha_i$.

Suppose all the null hypotheses are true; the probability then that they are all accepted is
P(No S_i's true)=

$$1-P(\text{at least one of the } S_i \text{ true}) = 1-P(S_1\cup\cdots\cup S_k)$$

$$\geq 1-P(S_1)-\ldots-P(S_k) = 1-\alpha_1-\ldots-\alpha_k, \quad (13)$$

which follows from

$$P(S_1 \cup \cdots \cup S_k) \leq P(S_1) + \cdots + P(S_k).$$

(13) is known as the **Bonferroni inequality**.

- A Bonferroni's inequality is not as crude as it looks, and it is especially tight for k not too large (~ 5) and α small (e.g .01). Miller (1981)
- More exact methods can provide improvement but often not much.

Using Bonferroni

A typical $100(1-\alpha)\%$ confidence interval is of the form

$$\bar{x} \pm t_{m,\alpha/2}\, S,$$

where S is an estimate of the appropriate standard deviation, and $t_{m,\alpha/2}$ is the $100(1-\alpha/2)$th percentile point of the Student's t distribution with m degrees of freedom.
To obtain k confidence intervals with $100(1-\alpha)\%$ simultaneous confidence, these intervals would be of the form

$$\bar{x}_i \pm t_{m,\alpha/2k}\, S,\ i = 1, \ldots, k$$

These intervals aren't as huge as you might think at first glance, because $t_{m,a}$ grows relatively slowly as a shrinks closer to 0.

SRM Example with Bonferroni Method

The individual 95 percent confidence intervals for the average lot differences that did not contain 0 were:

- Element 1: (-.035 , -.003)
- Element 2: (-.0067, -.0017)

These intervals take the form

$$\bar{x}_1 - \bar{x}_2 \pm t_{k,\alpha/2}\, \tilde{S} \tag{14}$$

where in this case $k = 10$ is the number of degrees of freedom (calculated from sample sizes, not number of elements), $\alpha = .05$ is the significance level, and $\tilde{S}$ is an estimate of the standard deviation of the difference.

The Bonferroni simultaneous intervals are

$$\bar{x}_1 - \bar{x}_2 \pm t_{k,\alpha/20}\, \tilde{S},$$

which use $t_{10,.0025} = 3.58$ instead of $t_{10,.025} = 2.23$.

- The 8 other confidence intervals that contained 0 still do so.

- The 95 percent simultaneous confidence intervals for the other two elements are
  - Element 1: (-0.044 , 0.007)
  - Element 2: (-.0082 , -0.0002)
- Note that the interval for Element 1 now includes 0, and the interval for Element 2 is even closer to including 0 (within detection limits).
- As with all such statistical techniques, multiple comparison and simultaneous inference procedures need to be used not blindly but only when appropriate!

Other methods

There are many other procedures for simultaneous inference and multiple comparisons, which we won't go into. Miller (1981) explains all of them.

- Tukey's Studentized Range
- Scheffe projections
- Fisher's Least Significant Difference method
- Maximum Modulus techniques

In addition to problems like the example above, these procedures are useful for answering other questions like: I have k different sample averages (e.g. measurements of the same material using k methods); an F Test tells me they are not all the same, but *which* of these averages are different from each other?

Significance Levels and P-values

- Under the null hypothesis, a hypothesis test with a significance level of $\alpha = .05$ has a 5 percent chance of incorrectly rejecting the null hypothesis (an innocent person has probability 0.05 of being convicted).
- The p-value of a hypothesis test result is the proportion of such tests that would have a result as extreme or more **if the Null Hypothesis were really true**. By "as extreme," we mean as favorable to the alternative hypothesis.

P-value Examples

- **Example 1.** The P-value of a DNA match is $2e10^{-N}$; the probability that a random person who have DNA that matched the sample so exactly is $\epsilon < 2e10^{-N}$.

- **Example 2.** You flunked a lie detector test with a P-value of .25. One of four random honest persons would have flunked the test as badly as you did or worse.

- A p-value of .04 would be significant for a .05 level test but not for for a .01 level test.

Power of Hypothesis Test

- Note that the Significance Level doesn't tell you how effective the test is when the null hypothesis is not true (i.e. at discerning guilt when guilt is present).

- The probability of correctly discerning the Alternative Hypothesis when it is present is the **Power** of the test.

- Since most hypothesis tests are based on a statistic with a threshold value, the choice of threshold involves the conflict between wanting high power and very low significance level.

- Traditional practice has fixed the significance level at 0.05 (or .01 or .10), and let the power be whatever it is. If the true effect and the sample size are not large, then the test may have rather low power (perhaps not even 50 percent!).

Empowering the Small and Weak

It may be desirable to bring together studies that point in the right direction but which lack power. This is ideal for bringing small effects to light:

- Psychotherapy is useful!
  - Glass, Smith, Rosenthal, etc.
- Peto: In medicine, incremental improvements are more likely than revolutionary changes. These incremental improvements can save thousands of lives.
- "Chalmers (1991): 10,000-20,000 deaths a year in the United States alone could have been prevented if the recommendations of a meta-analysis conducted in the mid-1970's on the effectiveness in aspirin in preventing heart attacks had been fully implemented." Although that estimate is likely somewhat inflated, it demonstrates the appeal of combining studies.

- Unfortunately, a poorly done meta-analysis can make bad results even worse. (Combining apples and oranges, Garbage in–Concentrated Garbage Out, etc.).

On Combining P-values

- What if you have k p-values all between .11 and .15? While none are individually significant, the group of p-values provides considerable evidence of some effect.
- If the null hypothesis is true, then (theoretically) the p-value of a hypothesis test has a Uniform (Rectangular) distribution on the interval (0,1).
- There are many methods of combining k p-values (for experiments testing the "same" thing) that say, "If this Null Hypothesis were really true, then these k p-values would be like a sample of k observations from a uniform (0,1) distribution. Let's test if that's the case!"

Fisher's Method

- If p_i is from a uniform (0,1) distribution, then $-2 \log p_i$ has the χ^2_2 distribution (with 2 degrees of freedom).
- Thus, if $p_1, \ldots, p_k$ are from a uniform (0,1) distribution, then $-2 \sum_{i=1}^{k} \log p_i$ has a χ^2_{2k} distribution (with $2k$ degrees of freedom).
- Reject H_0 if

$$-2 \sum_{i=1}^{k} \log p_i > C_\alpha, \tag{15}$$

where C_α is the $100(1-\alpha)$ percentile point of the χ^2_{2k} distribution.

Example *(part of a larger data set from Hedges and Olkin (1985))* Four studies of Sex Differences in Conformity produced p-values of .0029, .0510, .6310, and .3517.

Fisher's Statistic is $(-2) \times (-5.84 - 2.98 - 4.6 - 1.04) = 20.65$, which is greater than $C_{.05} = 15.5$ (and $C_{.01}$) for the χ^2_8 distribution.

Inverse Normal (Stouffer's) Method

- Let $z_i = \Phi^{-1}(1 - p_i)$, which then has a $N(0, 1)$ distribution.
- Then

$$\tilde{z} = (\sum_{i=1}^{k} z_i)/\sqrt{k}$$

also has a $N(0, 1)$ distribution.

- Reject H_0 if $z > C_\alpha$, where C_α is the $100(1-\alpha)$ percentile point of the $N(0, 1)$ distribution.

Sex Differences Example

Stouffer's statistic is $\tilde{z} =$

$(\Phi^{-1}(.9971)+\Phi^{-1}(.949)+\Phi^{-1}(.369)+\Phi^{-1}(6483))/\sqrt{4}$

$= 2.2$, which is greater than$1.96 = C_{.05}$ for the standard normal distribution.

Other Methods of Combining P-Vaues

- There are many others schemes for combining p-values. Most involve transformations of the p-values (e.g. Fisher's, Stouffer's method) or based on order statistics of the set of p-values (the smallest, largest, median, etc.).
- In addition, there are modified schemes that involve weighting and trimming the p-values according to their magnitude, quality, etc.

Weaknesses of Combining P-values

- Nowadays, just combining p-values is something you should only for those cases where you have no other information.
- There may be considerable information about a study's results that are not captured in the p-value; e.g. a very low p-value may be the result of a very large effect or a very large sample size and a small effect.
- "If more complete summary information about a study is available, it makes good sense to use it and avoid P-values." (Gaver, et al, 1992)

Effect Sizes

- Some data, especially in educational and psychological research often consists of artificial constructs for possibly nebulous quantities (e.g. self-concept, attitude towards school).
- Glass (1976) advocated the use of **effect sizes**, or **effect magnitudes**, that did not depend on the arbitrary scaling of the dependent variable.
- In this way, a series of studies that used different but “approximately equatable” outcome measures could have comparable effect magnitudes that can be combined.
- Effect sizes usually are combined in a weighted mean (weighted by inverse variance or sample size).

Two Families of Effect Sizes

There have been two main families of effect sizes in traditional meta-analysis:

- The "r family": based on correlation coefficients and related quantities. It is suited for cases when seeking relationships between 2 continuous variables. See the text by Rosenthal (1981) for more.
- The "δ family": based standardized mean differences and is suited for comparing two groups (e.g. treatment and control), e.g.

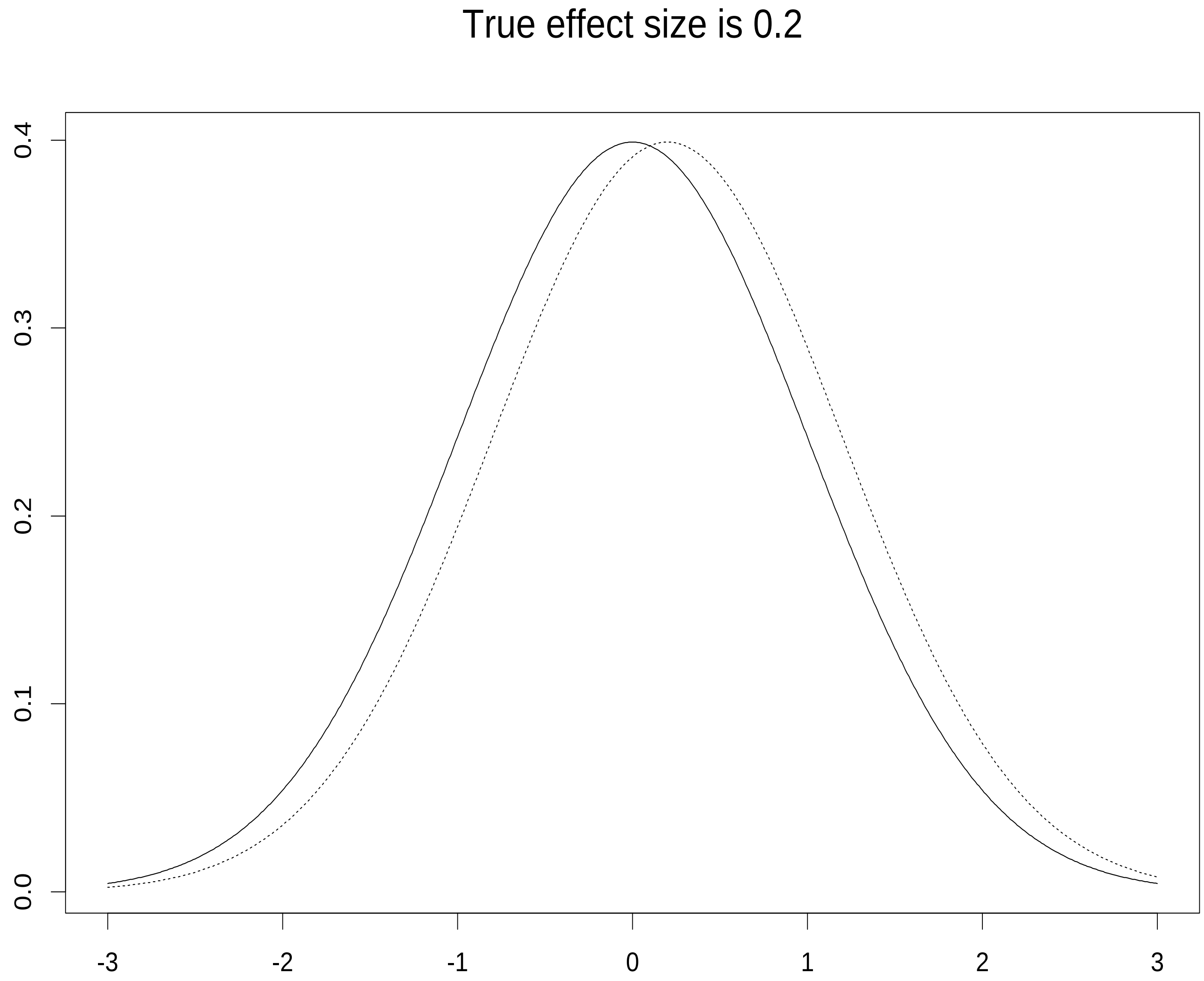
$$\delta = (\mu_T - \mu_T)/\sigma,$$

where μ_T and μ_C are the means of the treatment and control groups, respectively, and σ is the (often common) *population* standard deviation.

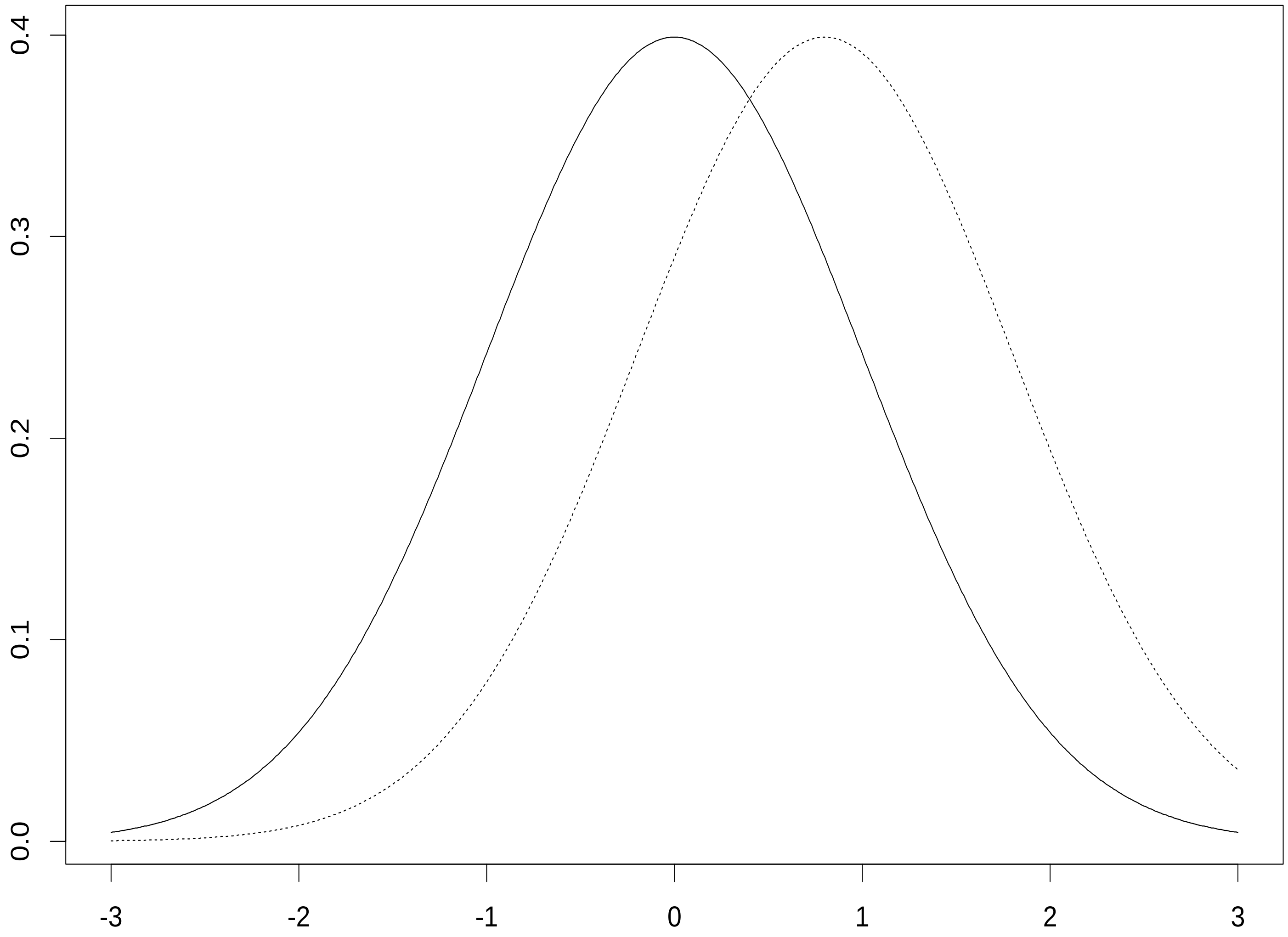
- **Note:** δ is **not** what a t-statistic measures! (What happens to a t statistics as the sample sizes go to ∞?)

How big is my effect size?

- “Cohen (1977) defines a large effect size as one visible to the naked eye.”
  - For normal distributions, Cohen says $\delta = .2$ is small, and $\delta = .8$ is large. The following diagrams depict these cases.


True effect size is 0.2
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3

True effect size is 0.8



Blood Pressure Example

Control:	135	148	155	162
Treatment:	131	124	166	127

An commonly used example involves a data set of systolic blood pressure readings from persons with hypertension (from Kraemer and Andrews (1981) and other sources). For illustrative purposes we will look at only a tiny portion of that data.

An Estimate of δ

The usual estimators of δ are of the form

$$(\bar{x}_T - \bar{x}_C)/S,$$

where $\bar{x}_T$ and $\bar{x}_C$ are the sample means for the treatment and control groups, respectively, and S is some measure of the (possibly common) standard deviation.

Glass advocated the estimator $g' = (\bar{x}_T - \bar{x}_C)/S_C$, where S_C is the sample standard deviation of the Control group.

According to Glass, g' is especially appropriate for comparing several treatment groups to a single control group, because the different treatment groups would probably not have the same variances.

Blood Pressure Example

$\bar{x}_T = 137$, $\bar{x}_C = 150$, and $S_C = 11.5$, so

$$g' = (137 - 150)/11.5 = -1.13.$$

g estimates δ

When there is only one treatment group and the variances of the treatment and the control groups are equal, it is advantageous for an estimate to utilize both samples to obtain a *pooled* estimate of the common standard deviation.

Suppose that the Treatment sample has m observations and the Control sample has n observations.

Then, an estimator of δ is

$$g = (\bar{x}_T - \bar{x}_C)/s,$$

where

$$s = \sqrt{\frac{(m-1)\,(S_T)^2 + (n-1)\,(S_C)^2}{m+n-2}}.$$

Blood Pressure Example

$\bar{x}_T = 137$, $\bar{x}_C = 150$, $S_T = 19.5$, and $S_C = 11.5$, leading to $s = 16$ and

$$g = (137 - 150)/16 = -.81.$$

Variance of g

The asymtotic variance of g is

$$\sigma_{\infty}^{2}(g) = \frac{m+n}{mn} + \frac{\delta^2}{2\,(m+n)}.$$

Note that it depends only on the sample sizes m and n and on δ.

An estimate of the variance of g is thus

$$\hat{\sigma}^2(g) = \frac{m+n}{mn} + \frac{g^2}{2\,(m+n)}.$$

Blood Pressure Example

Using an asymptotic expression for sample sizes $m = n = 4$ is clearly questionable, but for illustrative purposes we have

$$\hat{\sigma}^2(g) = \frac{4+4}{4 \;\cdot\; 4} + \frac{(-.81)^2}{2\,(4+4)} = 0.54$$

Probability of Superiority

Another type of effect size has been called the “Probability of Superiority”; given $x_1, \dots, x_m \sim F$ and $y_1, \dots, y_n \sim G$, what is the probability that a random observation from F will be greater than one from G?

Define

$$U_{ij} = \begin{cases} 1, \text{ if } & y_j < x_i \ , \\ \frac{1}{2}, \text{ if } & y_j = x_i, \\ 0, \text{ if } & y_j \geq x_i \ \ , \end{cases}$$

and $U = \sum_{i=1}^{m} \ \sum_{j=1}^{n} \ U_{ij}$.

Use $\widehat{PS} = U/mn$ as an estimator of PS= the Probability of Superiority of X over Y.

This is based on the Mann-Whitney and Wilcoxon nonparametric tests.

- If $U = 0$, use $\widehat{PS} = 1/(mn+1)$, and if $U = mn$, then put $\widehat{PS} = mn/(mn+1)$.
- We can transform $\widehat{PS}$ into a form comparable with the standardized mean difference d:
  - Assume approximate normality and equivariance of the distributions with a real effect size of $d = \delta$.
  - Then $\hat{\delta} = \sqrt{2}\, \Phi^{-1}(\widehat{PS})$ is an estimator of δ
  - This is based on the relation

$$P(y_j < x_i) = \Phi(\delta/\sqrt{2}\ ), \qquad (16)$$

which follows from the distribution of $(x_i - y_j)$.

Blood Pressure Example

Control:	135	148	155	162
Treatment:	131	124	166	127

To calculate $\hat{\delta}$, consider that there are $4 \times 4 = 16$ possible different pairings of a control observations with an treatment observations. Of these 16 pairings of observations, there were only 4 pairs in which the control observation was smaller than the Treatment observation (the pairs involving 166 from the Treatment group). Thus, $U = 4$, and $\widehat{PS} = 4/16 = 0.25$.

$$\hat{\delta} = \sqrt{2}\, \Phi^{-1}(0.25) = -.95.$$

Pros and Cons of $\hat{\delta}$

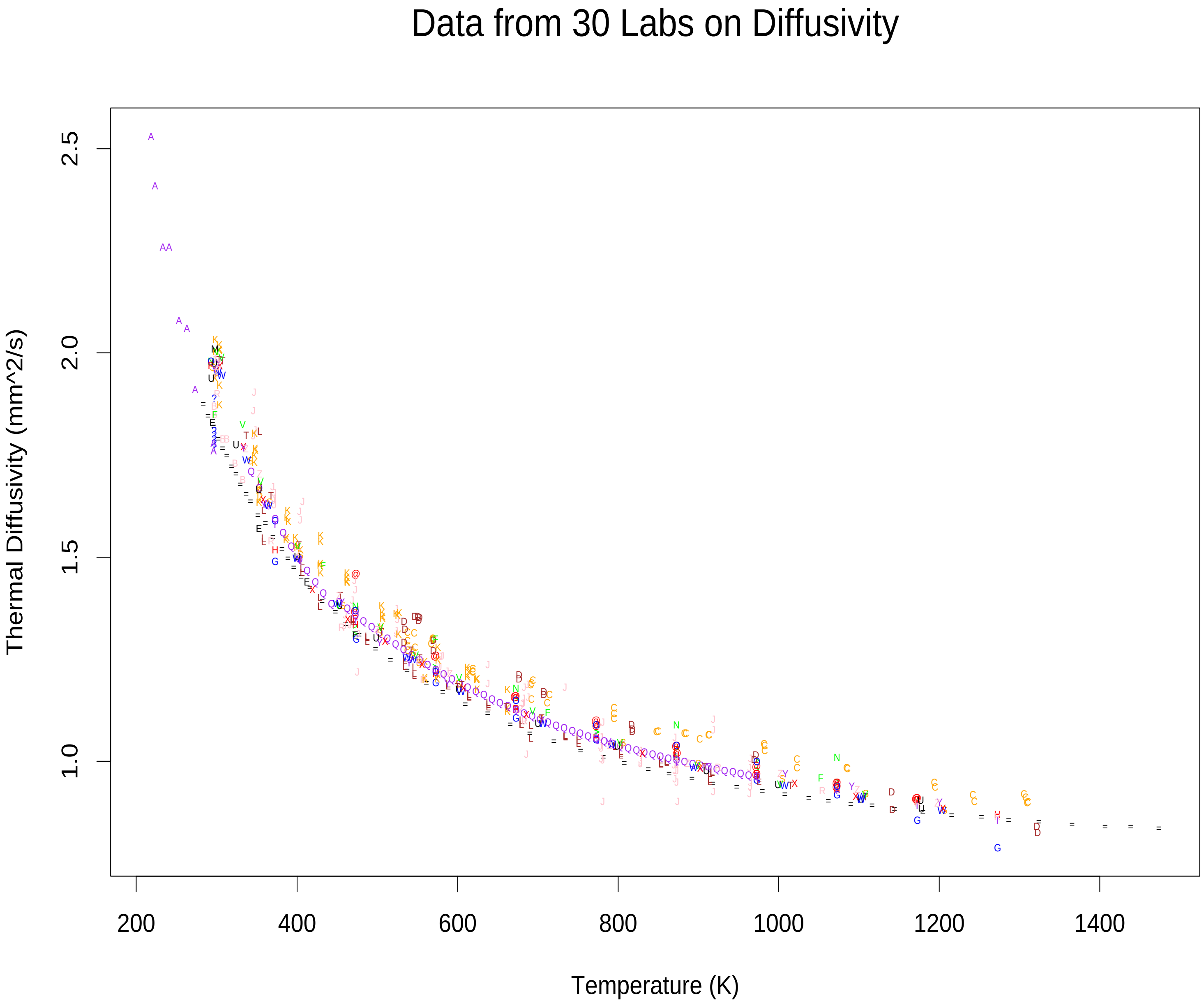
- The estimate $\hat{\delta}$ is much more robust than traditional estimators of d and almost as efficient.
- Unfortunately, they are much more difficult to calculate, requiring a computer for all but small data.
- Those not wanting to do their own programs can use the relation

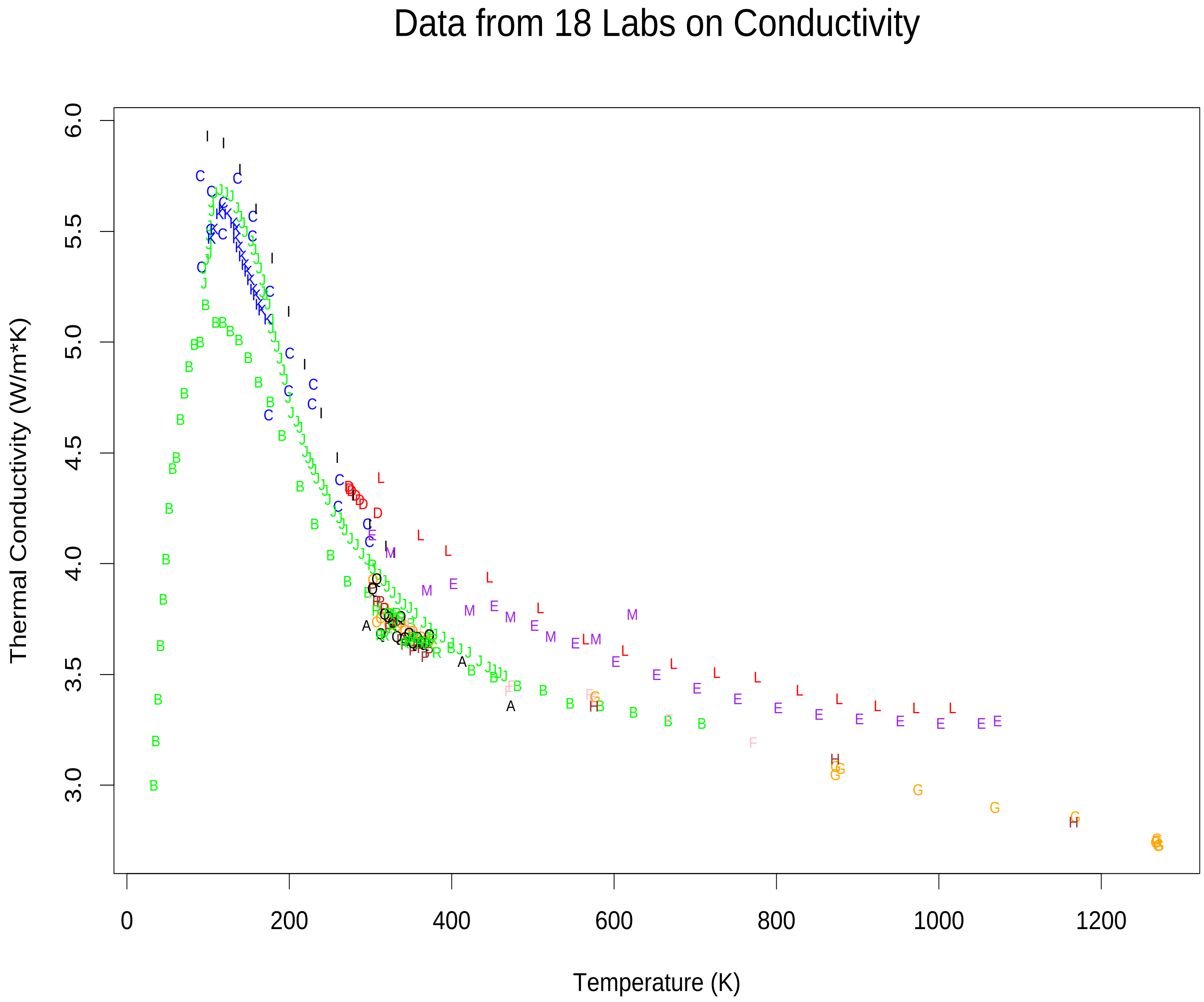
  $$U = W - m(m+1)/2,$$

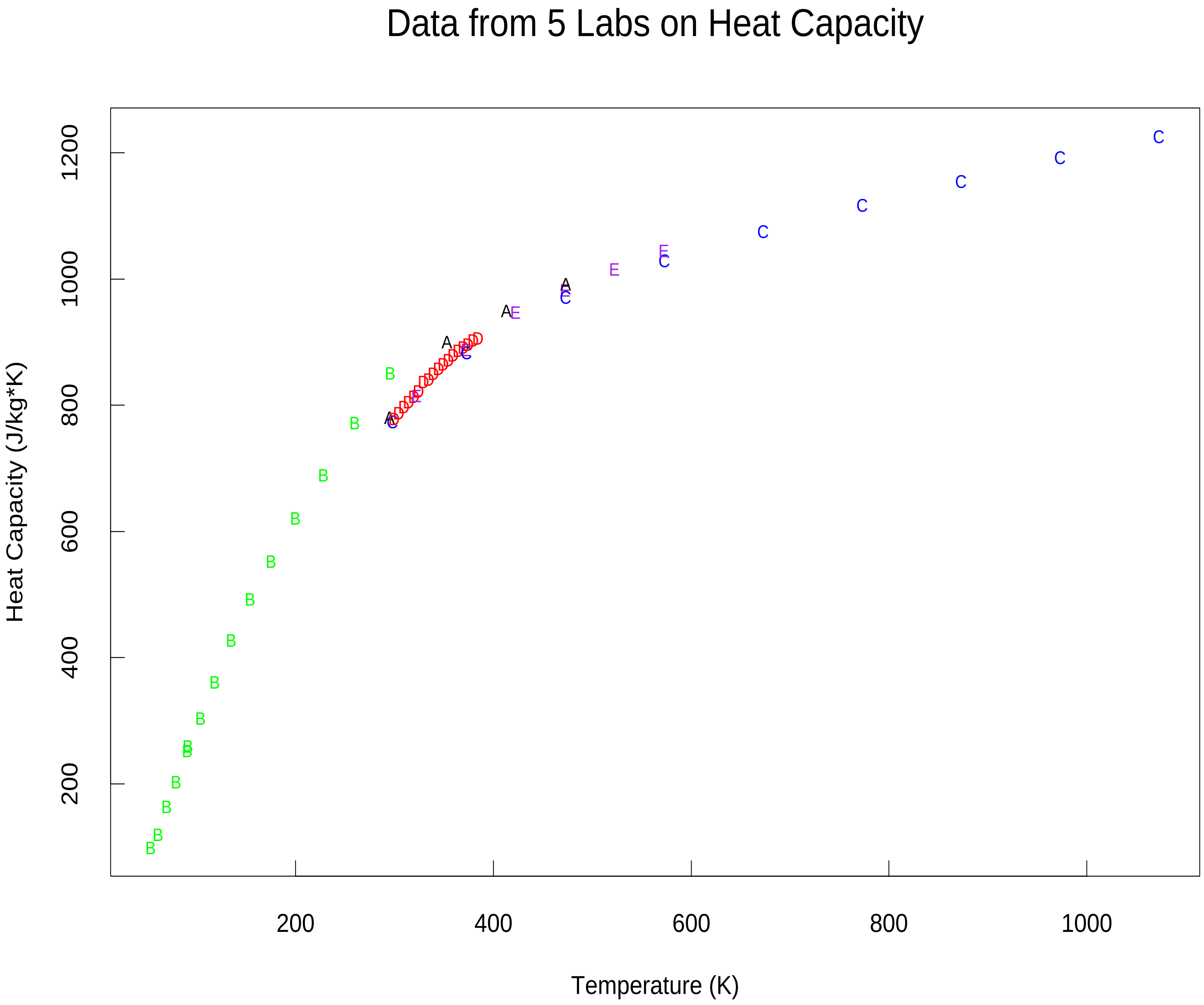
  where the Wilcoxon statistic W is the sum of the ranks of the treatment observations in the combined sample (Randles and Wolfe (1979)) and is also available in some computer packages.
- A computer program is required for variance estimation.

Function Estimation

- Often the desired result is not just a number, but a function.
- As an example, NIST is considering using a certain ceramic material as a reference material for thermal properties such as thermal diffusivity, thermal conductivity, and heat capacity.
- Scientists around the world have performed experiments measuring these properties.
- These experiments have varied greatly in several factors: temperature range of the observations, number of observations, and general quality of the experiment.
- The next pages show data for the three mentioned thermal properties, with each different experiment's measurements depicted by different symbols. Some data sets have already been excluded from the figures by subject matter experts.


Data from 30 Labs on Diffusivity
Thermal Diffusivity (mm^2/s)
2.5
2.0
1.5
1.0
200
400
600
800
1000
1200
1400
Temperature (K)


Data from 18 Labs on Conductivity
Thermal Conductivity (W/m*K)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
Temperature (K)
0
200
400
600
800
1000
1200
A
B
C
D
E
F
G
H
I
J
K
L
M
O
P
Q
R


Data from 5 Labs on Heat Capacity
Heat Capacity (J/kg*K)
1200
1000
800
600
400
200
Temperature (K)
200
400
600
800
1000
A
B
C
E

Consensus Functions?

- How do we combine the data from the various labs into a "consensus function"?
- In the past, the "authoritative" functions were sometimes results from a regression but also sometimes from an expert taking a pencil and drawing a line through the data.
- The quality of those results is a tribute to the expertise and experience of the scientists involved.
- Although we would like to move towards more computational ways of function estimation, scientific judgment will always be needed.

Scope of Discussion

- This class concentrates on the issues involved in combining data sets.
- In some cases it will be practical to combine the data sets and then proceed as if dealing with a single data set.
- We will not go into these single data set techniques in great detail because it is already the subject of a vast literature on regression, smoothing, and fitting, as well as previous (and future?) NIST-SED classes, e.g.
  - Regression Methods (Will Guthrie)
  - Functiona Data Analysis (Walter Liggett)
  - Exploratory Data Analysis (Jim Filliben)
- Since most procedures like regressions and splines are now carried out by computer packages, I won't detail the algorithms (can be found in references).

Simulation of Linear Case

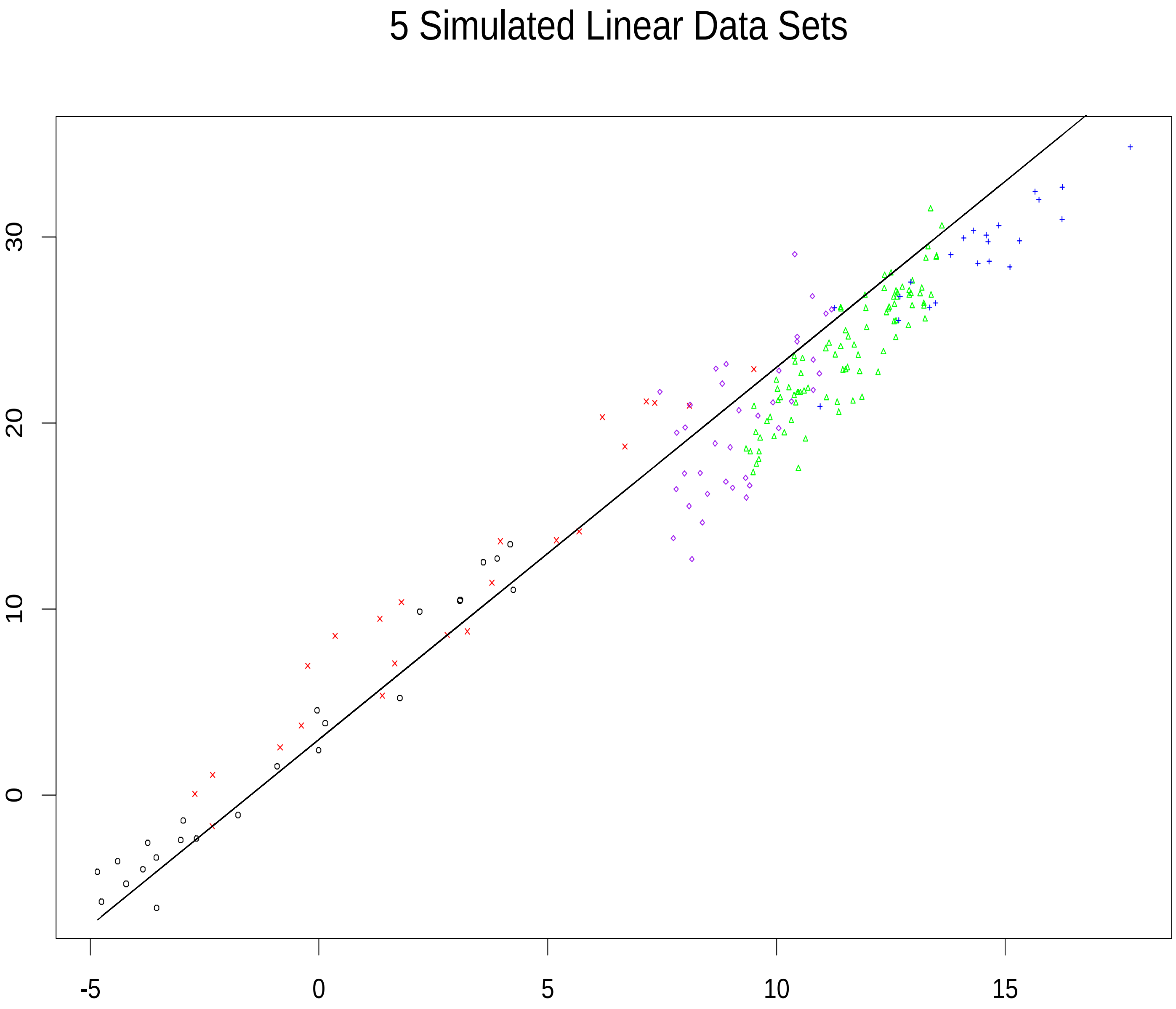
- A standard way to evaluate estimation procedures is on simulated data where we know what the "real answer" is.
- The next graph plots one example of the very special case of simulated linear data sets.

The simulation scenario is the following:

- The "real answer" is $y = m\ x + b$.
- The ith of k data sets has n_i points scattered on a random subset of the real line.
- Each point is generated as:

$$y_{ij} = m\ x_{ij} + b + Bias_i + \epsilon_{ij},$$

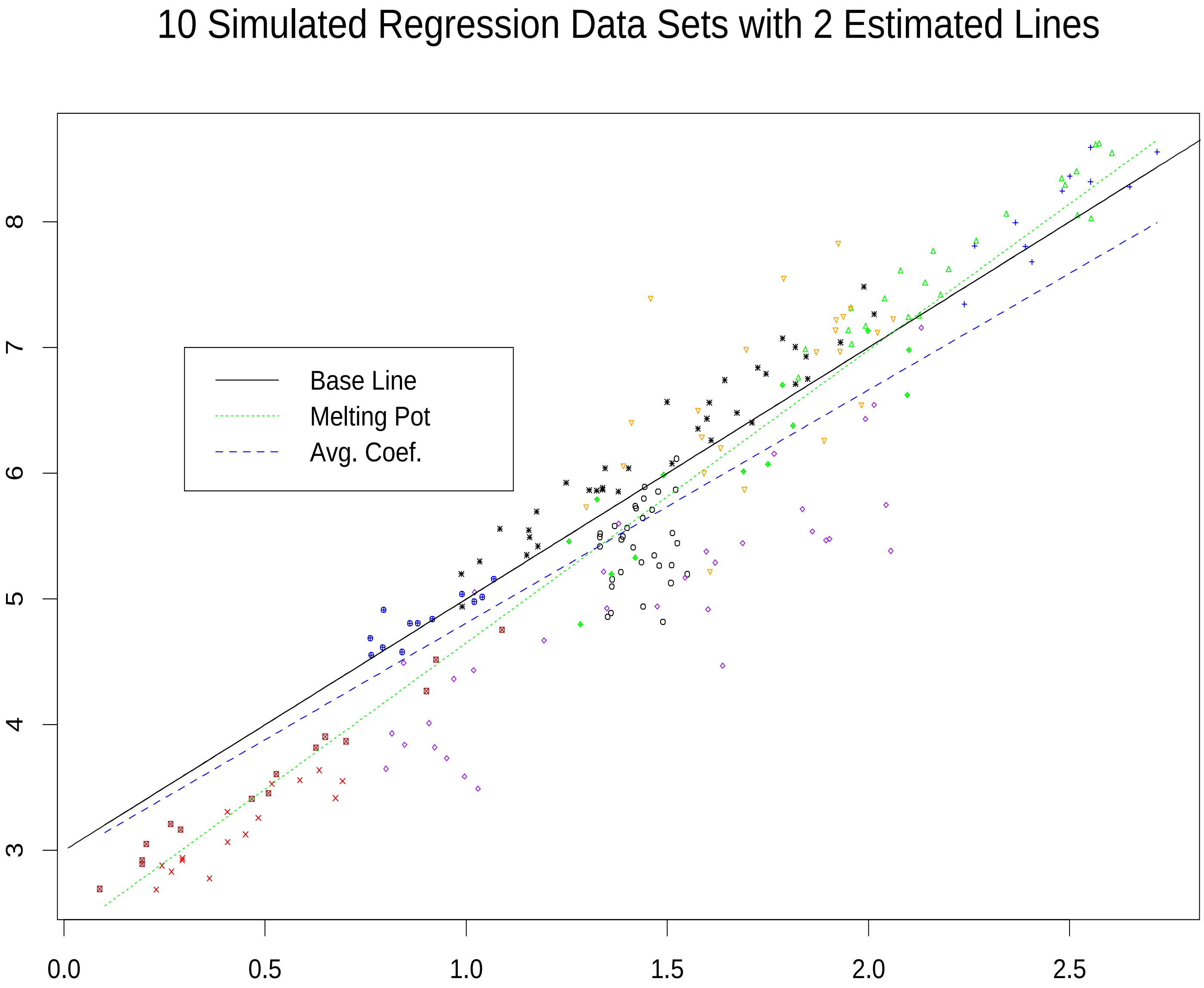
- $Bias_i$ comes from a uniform distribution on $(-b, b)$; the error term ϵ_{ij} comes from a $N(0, \sigma_i^2)$ distribution.
- Each variance σ_i^2 is generated from a *Gamma* distribution.


5 Simulated Linear Data Sets
0
10
20
30
-5
0
5
10
15

Some Combining Methods

This is a very special (and simple) case where you know the model is linear. We will look at the following methods of combining work for this case:

- Melting Pot Method
- Parameter Averaging
- Line Averaging
- Non-linear smoothing (not covered)



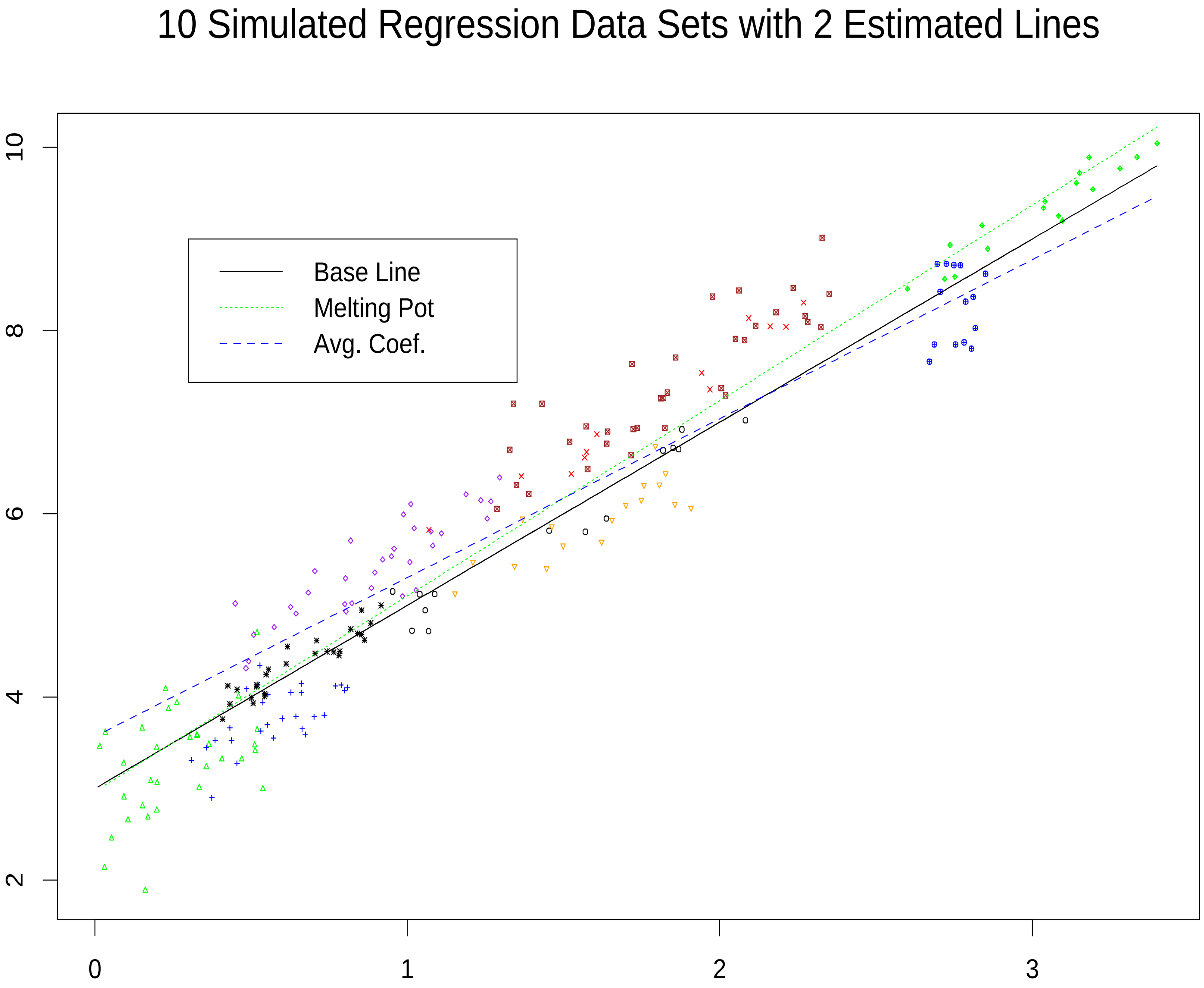
10 Simulated Regression Data Sets with 2 Estimated Lines

Melting Pot Method

- A relatively easy strategy is to put all the points in one large dataset and run a linear regression.
- This option may not be available if data are missing or not available.
- Weighting the observations from the different data sets is an option.
- This is the linear case, but for the general case, if treating it as just one data, can utilize general regression and function (functional) estimation techniques.
- For this simulation situation, the estimates were quite good.

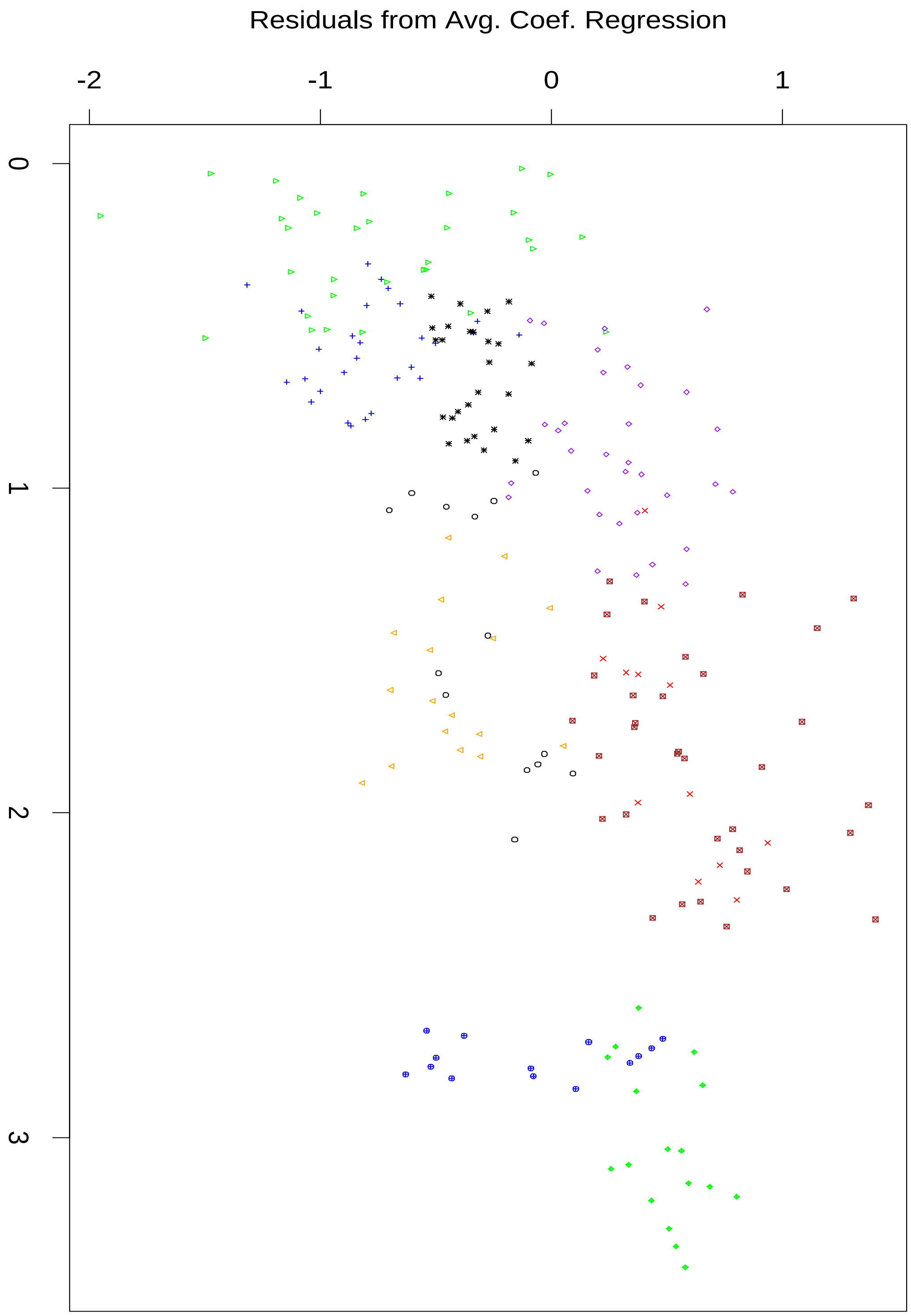
Parameter Averaging and Line Averaging

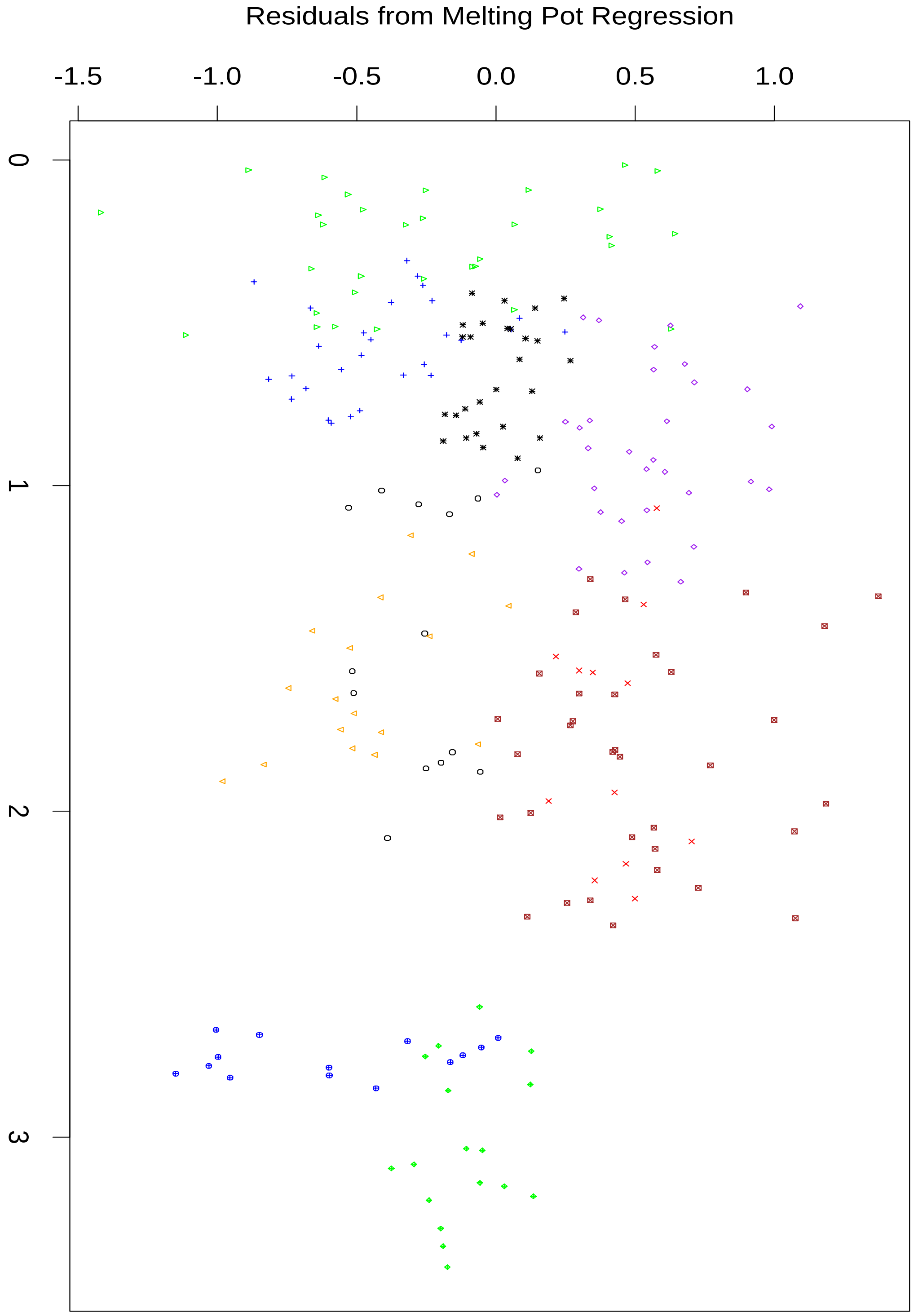
- In the linear case, taking the set of estimated parameters and averaging them gives the same result as taking the set of lines and averaging them.
- They would not be the same in most other situations.
- There are many situatios where neither method would make sense. In the linear case, if the theory truly holds that estimated parameters are normally distributed around the true parameters, then the estimates can be quite good.
- In the simulation, the estimates were good most of the time, but occasionally bad–a stray line can mess it up.
- A data set with points only over a small interval influences predictions far away.


10 Simulated Regression Data Sets with 2 Estimated Lines
Base Line
Melting Pot
Avg. Coef.
10
8
6
4
2
0
1
2
3

Importance of Residual Analysis

- Residual analysis is a very important part of the fitting process.
- For instance, when looking at a graph of the x's vs. the residuals, a random pattern of residuals (that fit the model) indicates a good fit, while patterns in the residuals indicate that your work isn't done.
- The next 2 graphs show the x's vs. the residuals for the one example shown of the Melting Pot regression and Average Coefficient lines.
- Other sorts of residual plots as well as hypothesis tests (e.g. F-tests) are useful for model validation.
- See regression texts and also Will Guthrie's Regression class for more details.


Residuals from Avg. Coef. Regression
-2
-1
0
1
0
1
2
3


Residuals from Melting Pot Regression
-1.5
-1.0
-0.5
0.0
0.5
1.0
0
1
2
3

Residual plots for multiple studies

- When there are multiple studies, there tends to within-study agreement (correlation), as all the data from certain labs may show the same bias and variability.
- Hence you are more likely to see clumps of residuals from the same labs that are all high or all low rather than the ideal totally random patttern.
- Residual plots that look more random or fit the residual model better may indicate a better fit, e.g. in the last graph, the residuals for the Melting Pot regression look better.

Simulation Model for a Non-Linear Function

- The next graph shows a simulation of several data sets in a decidedly non-linear situation.
- The "real" function is

$$f(x) = \frac{1}{x}\, exp(-[\log(x)]^2)\ ,\ x > 0.$$

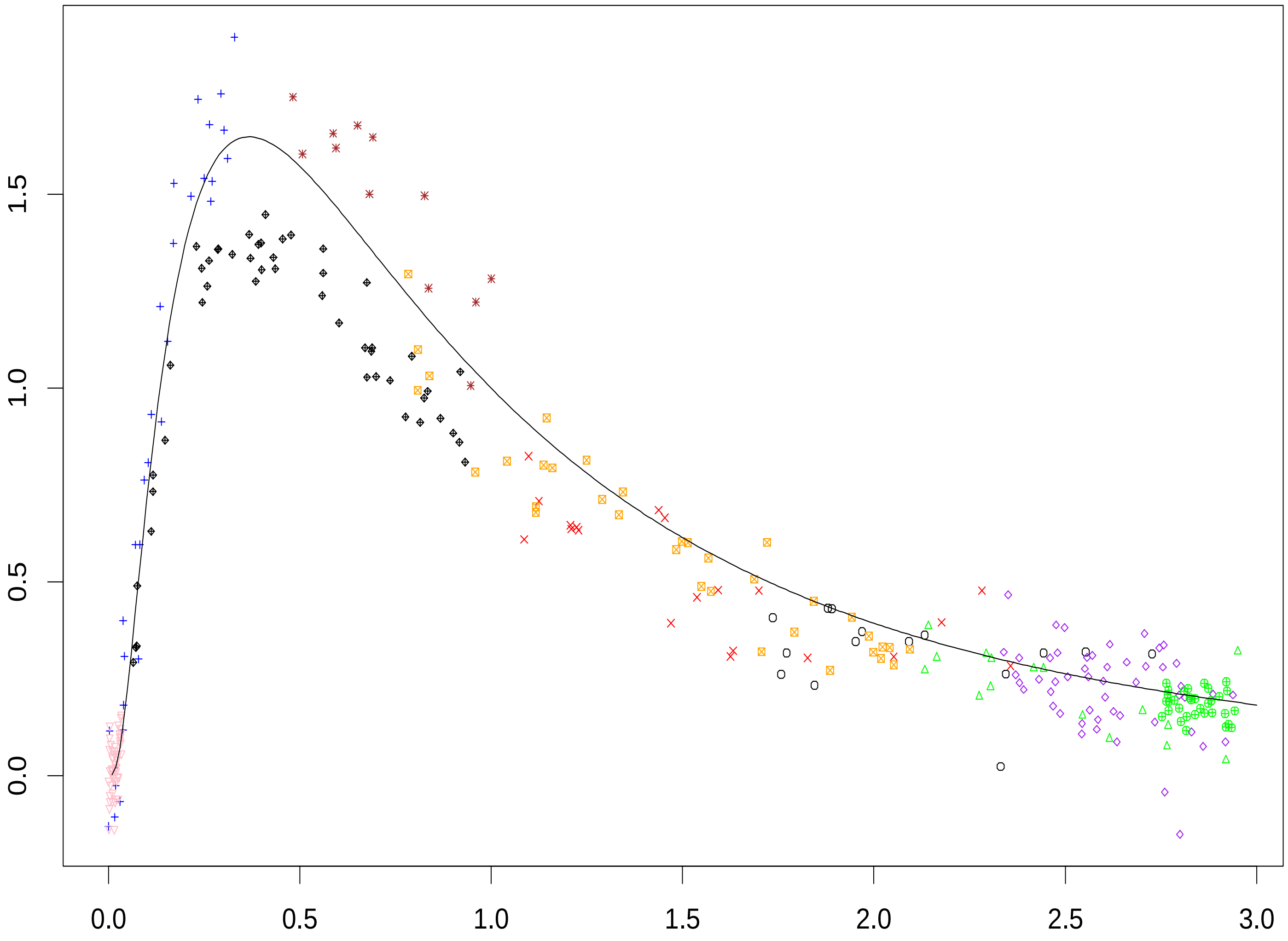
- [Some people may notice the resemblance to the density function of the lognormal distribution.]
- The ith of k data set has n_i points scattered from a uniform distribution on a random subset of part of the real line.
- Each point is generated as

$$y_{ij} = Bias_i * f(x_{ij}) + \epsilon_{ij},$$

- The $Bias_i$ comes from a uniform distribution on (-0.8,1.2), and the error term ϵ_{ij} comes from a $N(0, \sigma_i^2)$ distribution.

- Each variance σ_i^2 is generated from a *Gamma* distribution.
- We also did runs with an additive bias term, but we felt that with this particular function, a multiplicative bias factor made the most sense.

True Function pictured with 10 simulated datasets



Experience Desired

- Obviously, the best case is if you already know the form of function, i.e. insight from physical theory. Then some sort of linear or non-linear least squares (or weighted least squares) would probably be appropriate.
- Polynomials and other suitable functions cam be used "less knowledgeable" fitting.
- If you don't know and don't need an explicit function, then smoothing methods are very useful for fitting and making predictions.
- Two very useful procedures are Localized Regression and Splines.

Loess–Localized Regression

- When fitting (predicting) $f(x)$, points in the regression are weighted by their closeness to x. Thus, far away data points will have little influence on each other.

- As implemented on most statistical packages, the user tunes the span (window), which is the proportion of points used in each fit.
  - A Loess with small span will fit the data more closely; too small a span will give a jagged overfit
  - A larger span gives a smoother line, but won't follow the data as well.
  - It may be awkward at times to balance fit and smoothness; one may want to fit the data in separate pieces.

Also, each observation can be given a general weight for a weighted Loess.

- Does not give an explicit formula, so suitable for fitting and prediction rather than a physical interpretation.
- Details can be found in the article by Cleveland and Devlin, or in the documentation to your computer package.

Splines

- The most commonly known splines are cubic splines, which approximate a function with a collection of polynomials of degree less than or equal to 3 on subintervals such that the 2nd derivatives agree at the "knots" between subintervals.

- Splines have good structural properties and tend to give good fits.

- One can tune the balance between smoothness and fit desired.

- Does not give an easy explicit formula, so more suitable for fitting and prediction rather than physical interpretation.

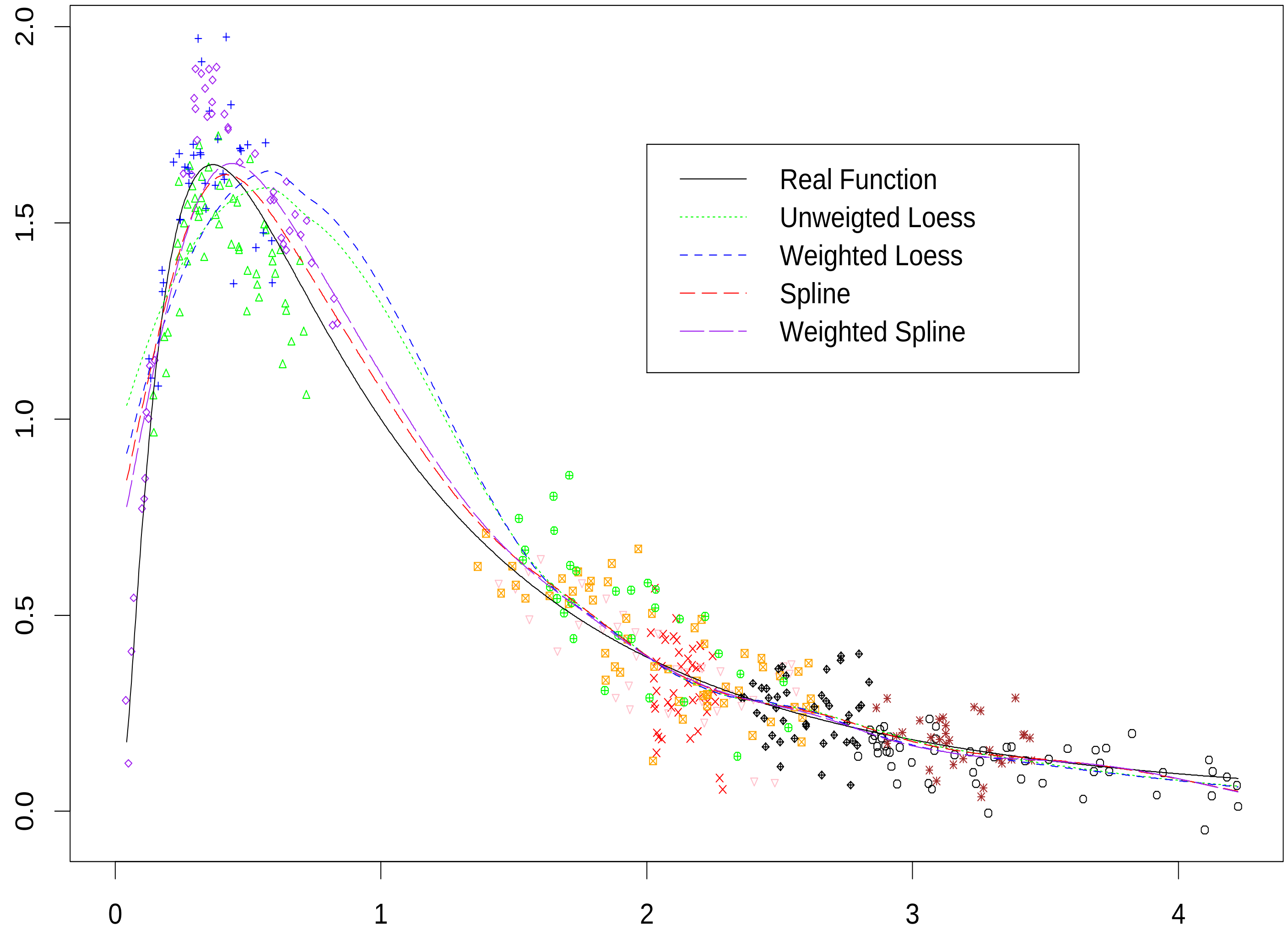
- There are many references for splines, including the text by Green & Silverman.

Weighty Concerns

The use of weights in Loess, splines, regressions, and other procedures can address several concerns:

- There may be know differences in the quality of data from different sources.
- Weighting by the Lab uncertainties may be desirable.
- There have been some data sets that are not really data, but are discretized values from a previously estimated curve.
  - In this case, the number of points in this dataset (which may be very large) is not meaningful.
  - Weighting can adjust the influence of such a set to a proper level, e.g. so that its effective number of points or its density of points is equal to that of a typical lab.
  - Such sets also have an artificial lack of variability which may come into play.

- Weighting points by its distance away from the other points, analogous to trimmed means and M-estimation, may be useful (this is something we haven't implemented).

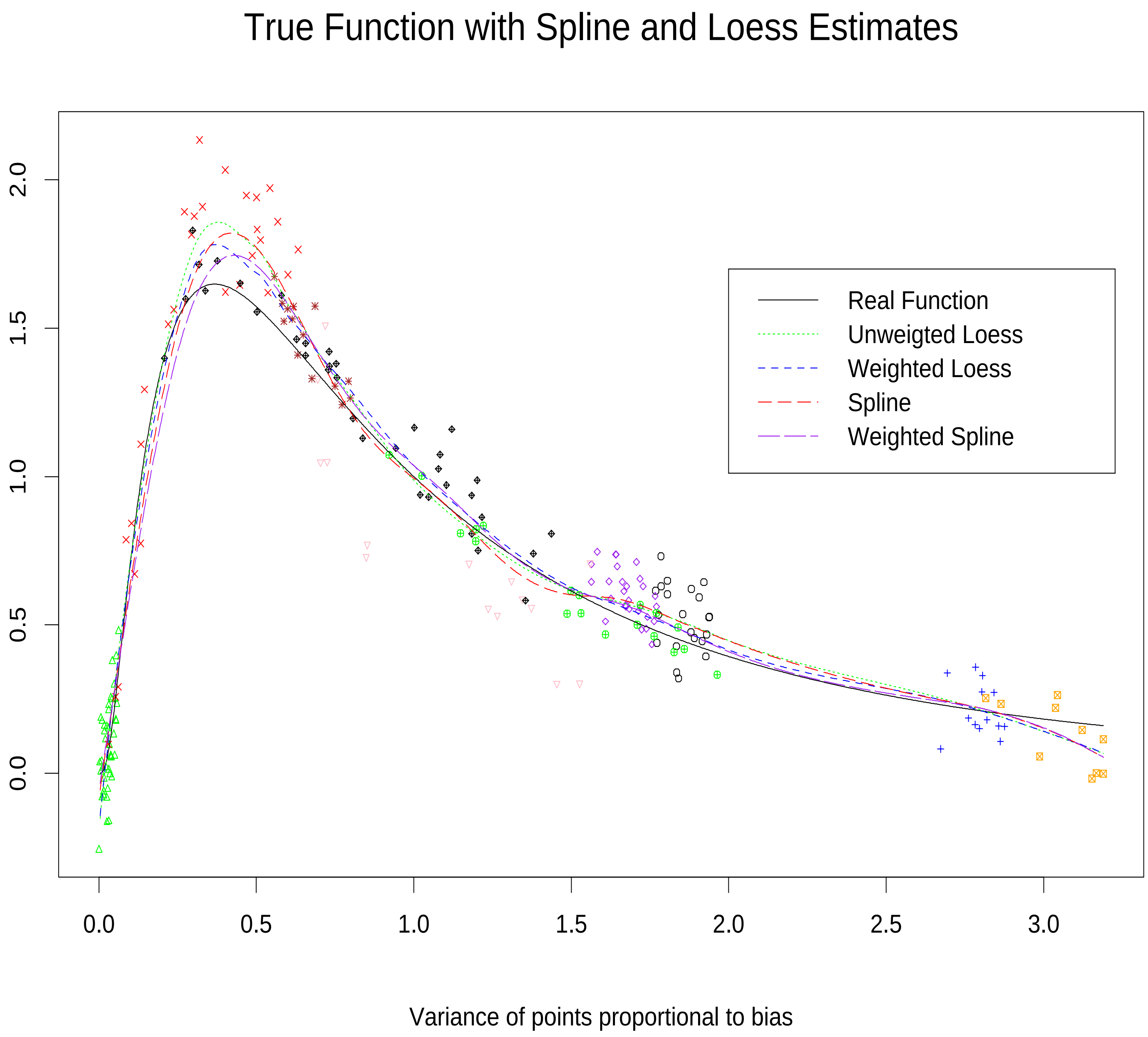

True Function with Spline and Loess Estimates

Simulation Results

- There tended to be very little difference between weighted and unweighted cubic splines, and also between weighted and unweighted Loess estimates.
- Perhaps the minimal effect of weighting was due to the simulation scenarios.
- Splines seemed to do better than Loess, but that could be due to better tuning for these scenarios.

Variance proportional to bias

- It is possible that lower quality experiments would have both larger bias bias and larger variance. In this case, weighted procedures should be advantageous.
- Simulations were run using the non-linear function in which the variance was proportional to the magnitude of the bias. The following page graphs one result.
- Here, weighted Loess and weighted splines had an advantage over their unweighted counterparts, but it was not as large as expected.
- Perhaps the dependence of variance on bias used were too mild to get much of an effect.


True Function with Spline and Loess Estimates
Real Function
Unweigted Loess
Weighted Loess
Spline
Weighted Spline
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Variance of points proportional to bias

Uncertainties of Estimated Functions

- Normally, computer packages will provide each fitted value $\hat{y}$ from Loess, etc. with a standard error.
- However, this standard error will be manifestly too small, probably because the package treats the data as one huge data set with a huge number of observations (remember the $1/n$ rule for the variance).
- Because the observations come from multiple data sets, you don't have as many effective replications as the computer thinks you do.

Residual Samples Method

An ad hoc method of summarizing the amount of uncertainty:

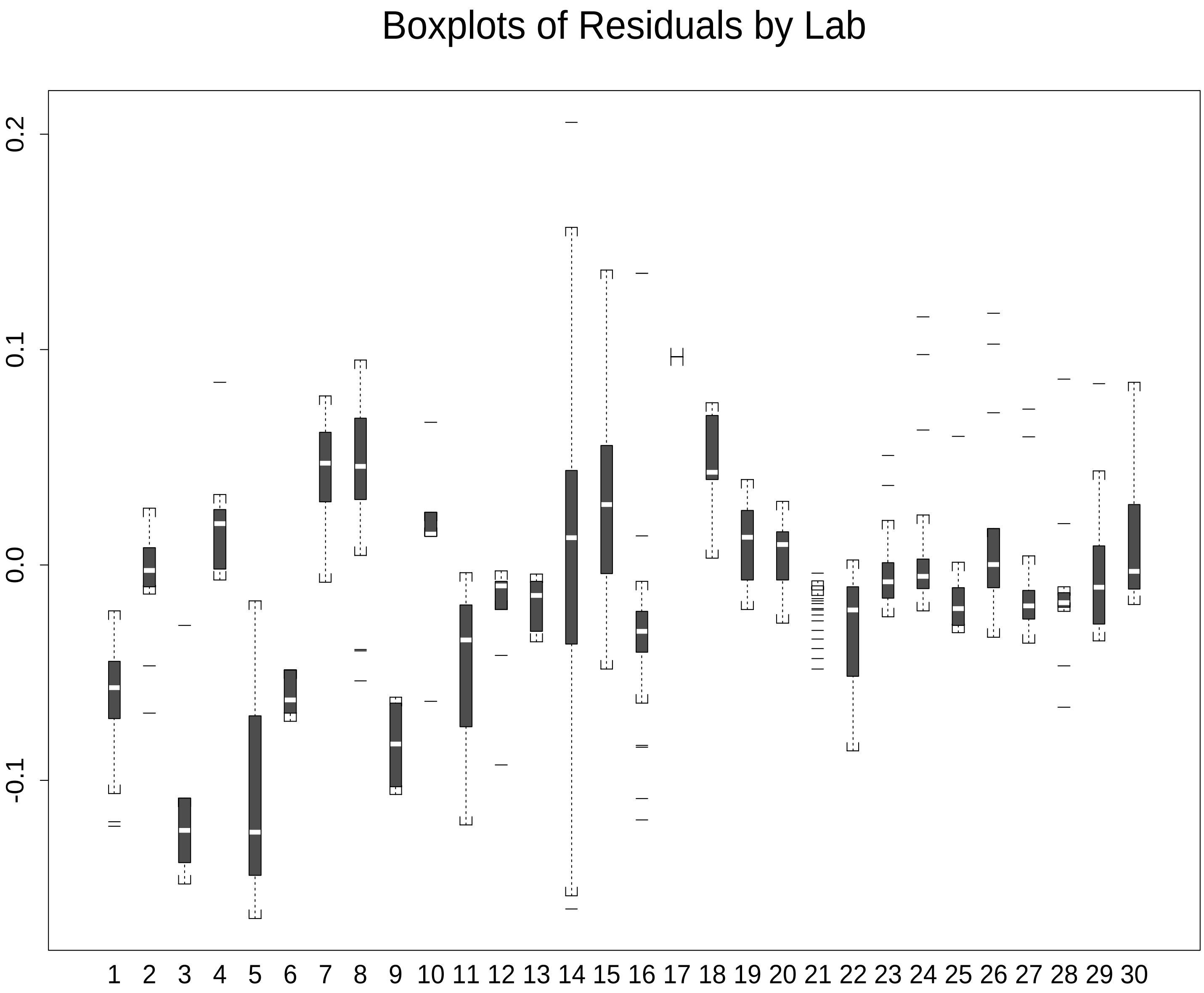
- Partition the residuals by study, so that we a have a different set of residuals for each experiment. The next graph does this for the diffusivity example.

- Treat the multiple sets of residuals as you would multiple sets of data, via the Mean of Means, BOB, etc.

- Use the intervals obtained for summarizing pointwise uncertainty.

- There may be a problem in having the same pointwise uncertainty on the real line.

- If the fit is done on the log scale (as in the diffusivity case), then after transforming back, the pointwise uncertainties are then proportional to the fitted values, which is often appropriate.


Boxplots of Residuals by Lab
0.2
0.1
0.0
-0.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Other Uncertainty Methods

This is still an area of ongoing research.

- Multiplying the standard errors given by Loess and other programs by an appropriate (still unknown) factor may give reasonable estimates.
- Perhaps a Moving Window version of the Residual Samples Method: use only the fits or points that are close by?
- Current intervals are pointwise (rather than covering the function as a whole).
- One needs to decide whether confidence intervals or prediction intervals are more appropriate.

Topics Left Out (or only briefly touched upon)

- Trimmed estimators and Robustness issues
- Estimating a bevy of physical constants simultaneously using extended least squares.
- “Meta-analysis”–the glamour and the naysayers
- Literature search and evaluation
- File drawer problem
- Theory of Hierarchical Models
- Bayesian Analysis

References

Box, Hunter and Hunter. 1978. Statistics for Experimenters. NY: John Wiley.

BUGS software for doing Bayesian analyses see:
http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

Carter, G. and Rolph, J. (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. Journal of the American Statistical Association, 69: 880–885.

Chalmers, T. 1991. “Heart attacks were avoidable.” Boston Globe (May 6): 3.

Chatterjee & Price (1991) Regression analysis by example. John Wiley.

Cleveland, W.S., and Devlin, S.J., (1988) Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association., Vol. 83, pp 596-610.

Cochran (1934). See instead Yates and Cochran (1938) below.

Cochran, W. 1937. Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society (Supplement), 4: 102-118.

Cochran, W. 1954. The combination of estimates from different experiments. Biometrics, 10: 101-129.

Cochran, W. 1982. Contributions to Statistics. John Wiley.

Cochran, W. and Carroll, S. 1953. A sampling investigation of the efficiency of weighting inversely as the estimated variance. Biometrics, 9: 447-459.

Cooper, Harris M. 1984, The Integrative Research Review: A Systematic Approach. Sage.

Cohen, J. 1977. Statistical power analysis for the behavioral sciences. San Diego: Academic Press.

Cook et al (1992) Meta-analysis for Explanation: A Casebook. Sage.

Cooper and Hedges (ed.) (1994), Handbook of Research Synthesis. Sage Publications.

Efron, B. and Morris, C. (1975) Data Analysis Using Stein’s Estimator and Its Generalizations. Journal of the American Statistical Association, 70: 311–319.

Fisher, R.A. (1932) Statistical Methods for Research Workers, London: Oliver and Boyd.

Gaver, et al. (NRC Panel) (1992) Combining Information: Statistical Issues and Opportunities for Research.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. and Smith, M. (actually Smith and Glass) (1977) Meta-analysis of psychotherapy outcome studies. American Psychologist, 32: 752-760.

Graybill & Iyer. Regression analysis : concepts and applications. Duxbury.

Green, P. J. and Silverman, B. (1994) Nonparametric regression and generalized linear models : a roughness penalty approach. Chapman & Hall.

Grissom, R.J. (1994). Probability of the superior outcome of one treatment over another. Journal of Applied Psychology, 79, 314–316.

Guthrie, Will (2000) Regression Models (NIST-SED Course Notes)

Hedges, L.V., and Olkin, I. (1985). Statistical Methods for Meta-analysis. New York: Academic Press.

Hunter, J. and Schmidt, F. 1990. Methods of meta-analysis. Newbury Park: Sage.

Levenson et al (2000). An approach to Combining results from multiple methods motivated by ISO GUM, NIST Journal of Research, 105, 571-579.

Li, Y., Shi L., and Roth, H. (1994) The Bias of the Commonly-Used Estimate of Variance in Meta-analysis. Communications in Statistics–Theory and Methods, 23(4): 1063-1085.

Liggett, W. (2001) Functional Data Analysis (NIST-SED Course Notes).

Mandel-Paule (see Paule, R. and Mandel, J. (1982) below).

Miller, Rupert. (1981). Simultaneous Statistical Inference. Springer-Verlag.

Mosteller, F. 1948. On pooling data. Journal of American Statistical Association. 43: 231-242.

Mullen, Advanced BASIC Meta-analysis. (1989).

NRC Panel (1992) see Gaver et al.

Natrella, M. 1963. Experimental Statistics, Handbook 91, NBS.

Paule, R. and Mandel, J. (1982) Consensus values and weighting factors, NIST Journal of Research, 87, 377-385. Randles, R.H., & Wolfe, D.A. (1979). Introduction to the theory of nonparametric staitistics, John Wiley.

Rao, Poduri S. R. S. (1983/1984) Cochran's Contributions to Variance Components Models for Combining Estimates, in W. G. Cochran's Impact on Statistics, ed. by Rao, Poduri S. R. S. and Sedransk, Joseph, John Wiley, 1984

Rosenthal, R. 1984. Meta-analysis procedures for social research. Beverly Hills: Sage.

Rukhin, A. L. and Vangel, M. G. (1998). "Estimation of a Common Mean and Weighted Means Statistics," Journal of the American Statistical Association, 93, 303-309.

Schiller, S. and Eberhardt, K. "Combining Data from Independent Chemical Analysis Methods," Spectrochimica Acta, 46B No. 12, pp. 1607-1613, (1991).

SED website: http://www.itl.nist.gov/div898/
contains links to the software Dataplot with its "Consensus Means" function, and also the NIST/Sematech Statistics Handbook and previous SED courses.

Snedecor, George Waddel, (1946) Statistical methods, Iowa State Press. [newer editions also available].

Vangel, M. G. and Rukhin, A. L. (1999). "Maximum-Likelihood Analysis for Heteroscedastic One-Way Random Effects ANOVA in Interlaboratory Studies," Biometrics, 55, 302-313.

Vangel, M. G., (1998). "ANOVA Estimates of Variance Components for Quasi-Balanced Mixed Models," Journal of Statistical Planning and Inference, 70, 139-148.

Yates, F. and Cochran (1938) The analysis of groups of experiments. Journal of Agric. Science, 28 : 556-580.

Epilogue

LORD OF THE FILES

Three Studies by the Scholar-kings under the sky,
   Seven by the Professors in their halls of stone,
Nine by Junior Faculty doomed to die,
   One for the Meta-analyst in his dark room
In the Land of Meta-analysis where the Studies do not lie.
   One Study to to rule them all, One study to find them,
   One Study to bring them all and in the darkness bind them
In the Land of Meta-analysis where the Studies do not lie.

Please take a few minutes to comment on the class on the Feedback Form.