Interlaboratory Comparisons: A Review *

Haben Michael and Ingram Olkin
Stanford University

METRICS
March 28, 2016

*The authors wish to thank the Laura and John Arnold Foundation and the Meta-Research Innovation Center at Stanford for supporting this research.

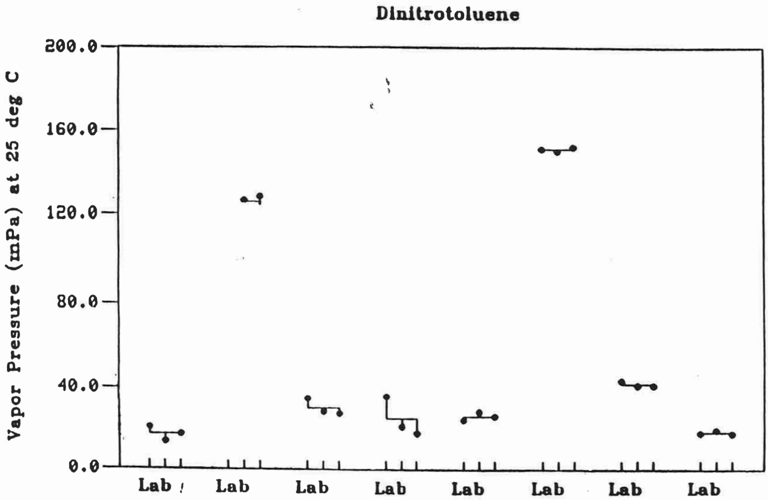
- Examples from Federal Regulatory Agencies
  - Food & Drug Administration (FDA): administer federal food purity laws, cosmetics
  - Food Safety and Inspection Service (FISIS): safety of meat, poultry, eggs
  - Environmental Protection Agency (EPA): monitor > 90 contaminants
- Standards Agencies
  - National Institute of Standards & Technology (NIST): weights and measures, smoke detection; performance standards for emissions
  - Design & construction standards
  - United States Military Standards

Scenario

- Similar samples undergo independent analyses by a number of labs
- Examples
  - Pesticides in plants
  - Contaminants in food
  - Percent nutrients in food
  - Setting of weights and measures
  - Standards for medical devices
  - Pathology analyses
  - Powder burning times of fuses

Figure 1

Vapor pressure readings from eight laboratories



	Obs. 1	**Obs. 2**	**Obs. 3**	**Obs. 4**	**Obs. 5**	**Obs. 6**	**mean**	**sd**
A	2.963	2.996	2.979	2.970	2.979	2.977	**2.977**	**0.011**
B	2.958	2.964	2.955	2.932	2.941	2.950	**2.95**	**0.012**
C	2.956	2.945	2.963	2.950	2.975	2.958	**2.958**	**0.01**
D	2.948	2.960	2.953	2.944	2.950	2.951	**2.951**	**0.005**
E	2.953	2.961	2.961	2.953	2.949	2.955	**2.955**	**0.005**
F	2.941	2.940	2.931	2.942	2.930	2.937	**2.937**	**0.005**
G	2.963	2.928	2.925	2.940	2.934	2.938	**2.938**	**0.014**
H	2.987	2.989	2.988	2.983	2.974	2.984	**2.984**	**0.005**
I	2.946	2.950	2.955	2.969	2.954	2.955	**2.955**	**0.008**
J	2.956	2.947	2.947	2.960	2.954	2.953	**2.953**	**0.005**

Table: Determination of Iron in Solution (%) (Olkin, Guttman, and Phillips 1995)

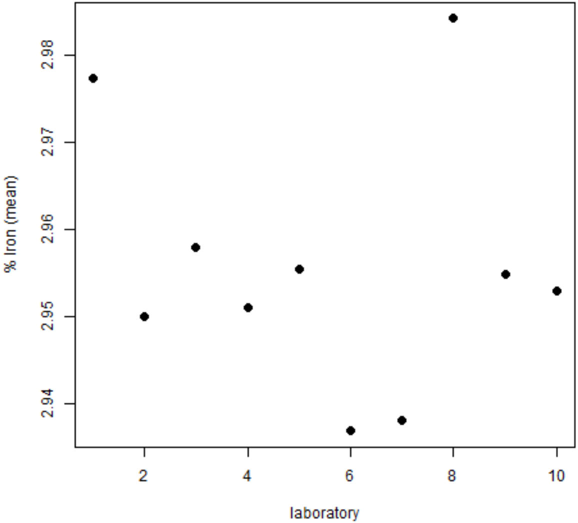


Figure: Determination of Iron in Solution (%). Laboratory means for 10 laboratories.

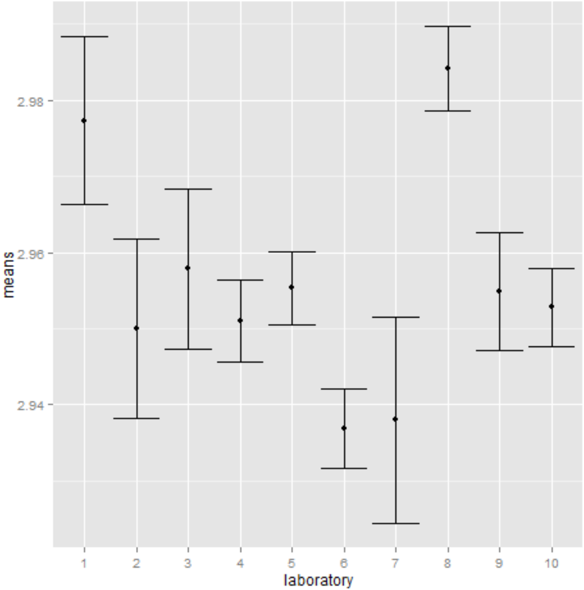


Figure: Iron in solution data with ± 1 standard deviation error bars

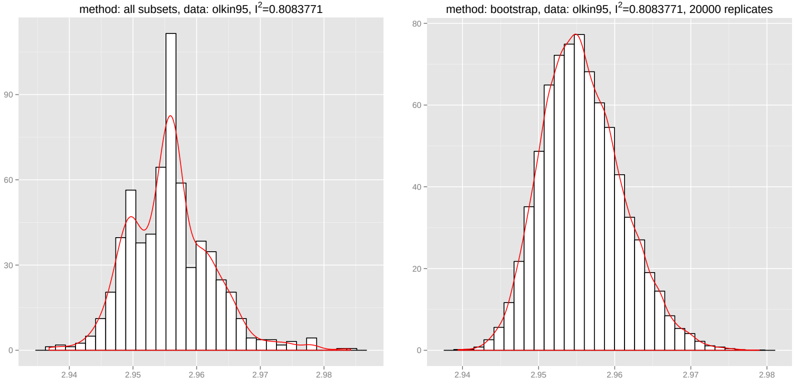


Figure: Histogram of summary statistic on samples drawn without replacement (left) and with replacement (right). (Iron in solution data.)

Potential Analyses: Searching for outliers

A. Nonparametric analyses

- Order observations: $X_{(1)} \leq \ldots \leq X_{(n)}$
- Define X_{median} and residuals

$$u_i = |X_{(i)} - X_{median}|,$$
$$u_{(1)} \leq u_{(2)} \leq \ldots u_{median} \leq \ldots \leq u_{(n)}$$

- Define outlier if $u_i \geq c u_{median}$, $c = 3.5, 4.5, 5.2$
- "Smallest proportion which may cause the test to fail"

B. Bayesian approach

- Set $C(j_1, \ldots, j_\alpha)$ as posterior probability labs $j_1, \ldots, j_\alpha$ aberrant
- Fix α and find $C^*(\alpha) = maxC(j_1, \ldots, j_\alpha)$
- Then $\underset{\alpha}{max} C^*(\alpha)$
- Example

$$\begin{aligned} C(8) &= 0.982 \\ C(1,8) &= 0.998 \\ C(1,6,8) &= 0.585 \\ C(1,6,7,8) &= 0.930 \end{aligned}$$

- *A variation on the Bayesian procedure*: minimize the within-cluster sum of squares. Set $C(J_1, \ldots, J_k)$ as the sum of cluster Q-statistics when the data is clustered as $J_1, \ldots, J_k$
- Set $C^*(k) = \min_{J_1, \ldots, J_k} C(J_1, \ldots, J_k)$
- Look for an "elbow" in a scree plot, or use a resampling procedure (gap statistic)

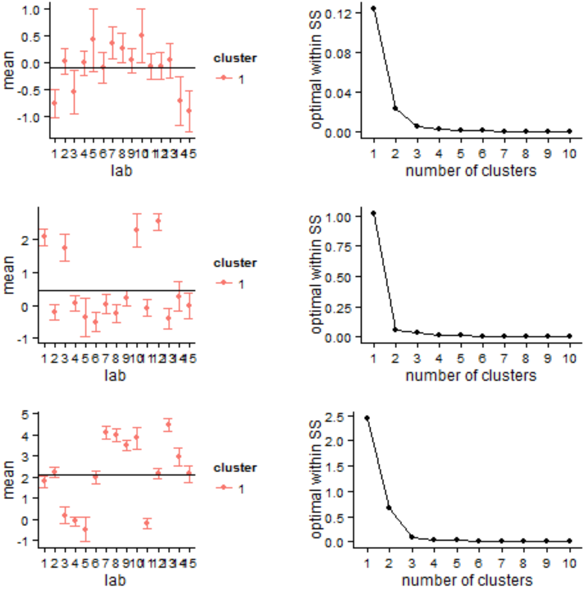


Figure: Scree plots for the minimum summed cluster Q statistic. The data is synthetic with 1 (homogenous), 2, and 3 clusters.

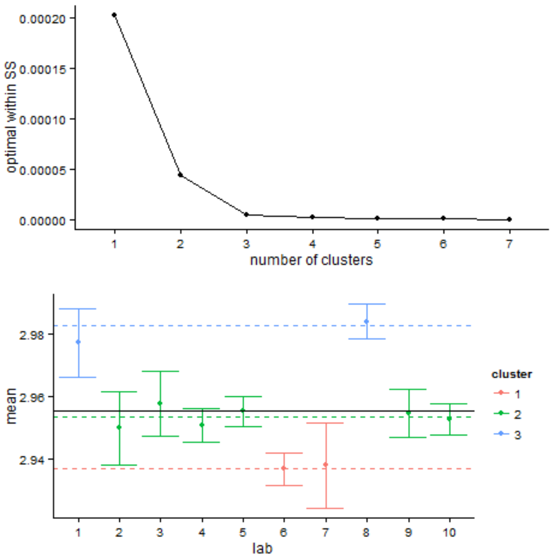


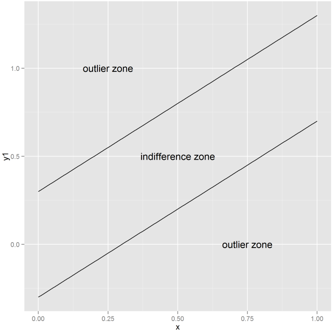
Figure: Minimum summed cluster Q statistic, for the iron in solution data. There appear to be 2 outlying clusters, one consisting of labs #s 1 and 8, the other labs #6 and #7.

C. Ranking and selection procedure

- Model $x_{ij} = \mu + \theta_j + \epsilon_{ij}$

$$\begin{aligned} \mu &= \text{true mean} \\ \theta_j &= \text{parameter of jth lab} \\ \epsilon_{ij} &= \text{independent, } \mathcal{N}(0, \sigma^2) \end{aligned}$$

- Screen out labs where $|\theta_j| \geq \delta$
- Indifference Zone: $|\theta_j| < \delta$ for some specified δ



- Rule: Assign each lab to a group so that the mean in each group is at least δ away from mean in any other group

D. Resampling-based graphical methods. These methods compute a summary statistic on resampled sets of the original data. The heuristic is that if the original data is homogenous, the typical resampled sets should be as well.

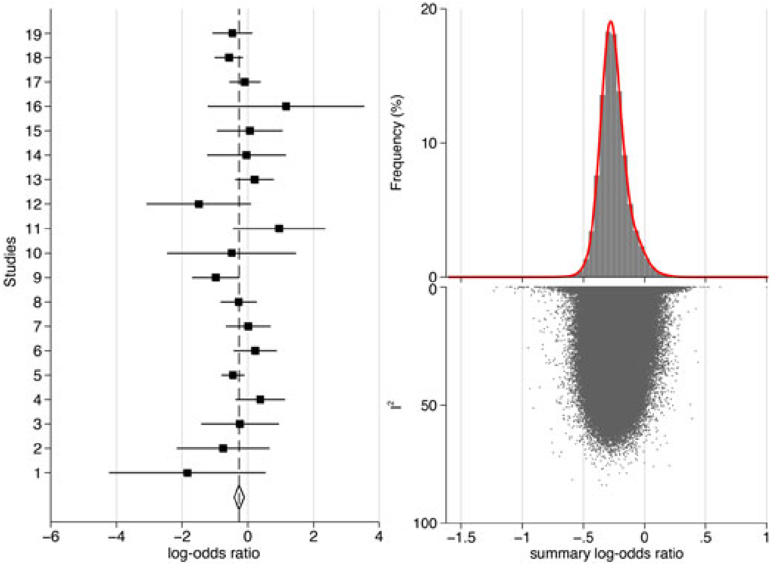


Figure: The histogram of summary statistics for all subsets is unimodal, suggesting homogeneity among the primary studies. (Data: Thrombolytics for acute myocardial infarction meta-analysis.)

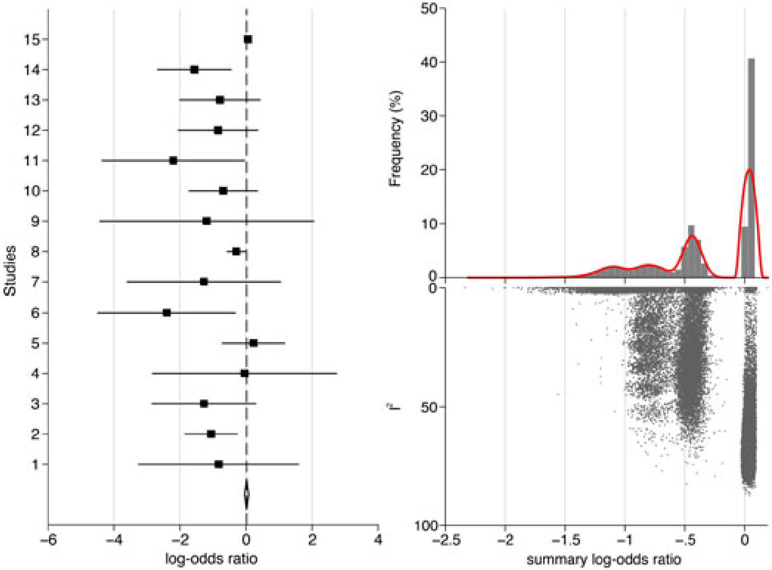


Figure: The histogram of summary statistics for all subsets is multimodal, suggesting heterogeneity among the primary studies. (Data: Magnesium for myocardial infarction meta-analysis.)

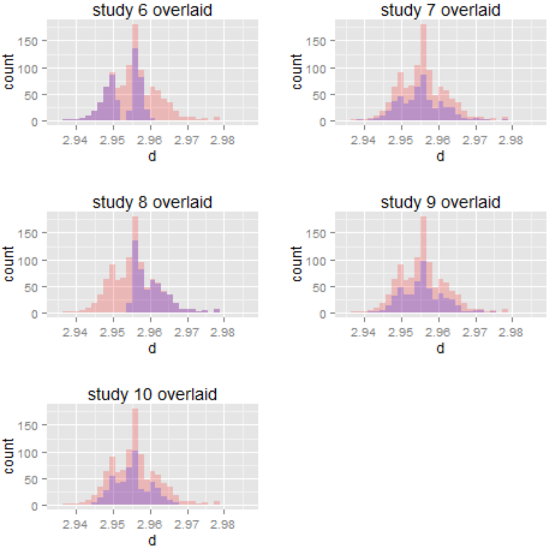


Figure: The histogram of summary statistics for subsets containing a given study is compared against the full histogram. Studies #6 and #8 are identified as potential outliers.

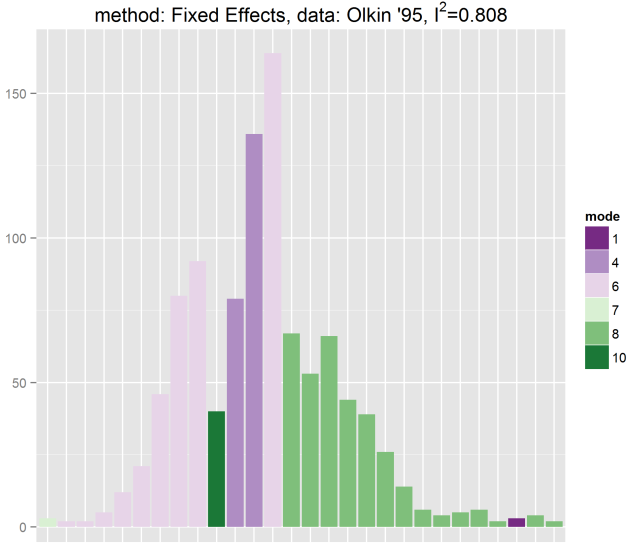


Figure: Coloring modes by bin (most frequently occurring lab among subsets in a bin). Lab #8 is overrepresented among the rightmost bins, lab #6 among leftmost.

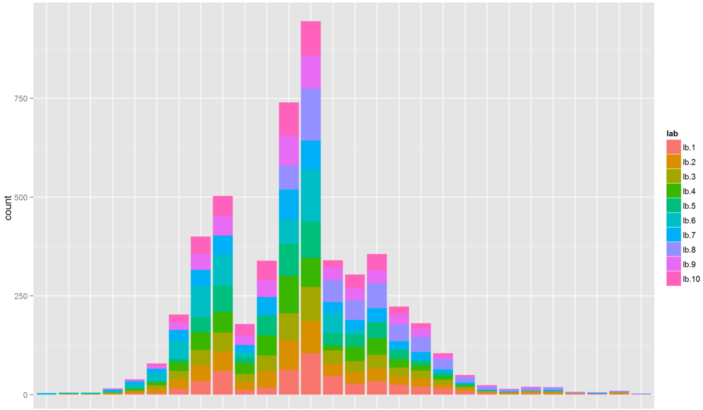


Figure: Stacked barplot of counts of labs per bin. Again, lab #8 is overrepresented among rightmost bins, lab #6 among leftmost.