Combining Results in Collaborative Studies When Reported Uncertainties are Unreliable

By: Luke Benz
Statistical Engineering Division
Information Technology Laboratory
Mentor: Andrew Rukhin, PhD

What are Collaborative Studies?

- Studies that combine data from several independent sources reporting a common measurement.
- Allows for analysis on larger set of data, and thus in theory, more trustworthy results.
- Important part of *ANY* scientific field.
- The statistical analysis used to combine results is called a meta-analysis.

Collaborative Studies at NIST

- Inter-laboratory studies are an important part of establishing standards at NIST.
  - NIST researchers send for data to be collected by several labs/sources.
  - Upon the return of data, researchers must combine the each source's estimate of the measurement in question in order to determine a consensus value.

Collaborative Study Example

- Measure of Interest: Newton's Constant of Gravitation (G)
  - This problem has been of great interest to researchers at NIST. Dr. Antonio Possolo has also spent time examining this constant using several meta-analysis procedures.
- Experimental Data: 10 reported measurements (Mohr and Taylor, 2000)
  - There have been more recent studies, but this study has been chosen as a worked example for this presentation.
- Reported values (x_i) and uncertainties (s_i) in 10^{-11} $m^3kg^{-1}s^{-2}$

i	1	2	3	4	5	6	7	8	9	10
x_i	6.6699	6.6726	6.6729	6.6735	6.6740	6.6742	6.6754	6.6830	6.6873	6.7154
s_i	0.0007	0.0008	0.0005	0.0029	0.0007	0.0007	0.0015	0.0011	0.0094	0.0005

Decomposing Uncertainty

http://theendlessfurther.com/wp-content/uploads/2016/12/Uncertainty.jpg

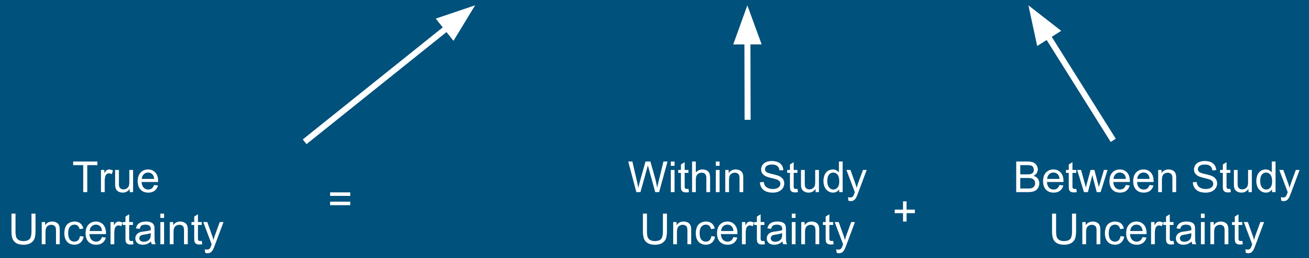
- The ISO *Guide to the expression of uncertainty in measurement* (GUM) outlines two evaluations of uncertainty:
  - Type A: Evaluation of uncertainty through means of a statistical analysis of observed data.
  - Type B: Evaluation of uncertainty that does not involve statistical analysis of observed data.
- Dark Uncertainty: Uncertainty that is not visible/not accounted for.

Decomposing Uncertainty

- In this project, two types of uncertainty are considered:
  - Within-Study Uncertainty (s^2)
    - Type A evaluation of uncertainty
    - Reported by each lab/source
  - Between-Study Uncertainty (τ^2)
    - Not perfect correspondence with Type B evaluation of uncertainty
    - Not reported. Must be estimated.
    - Ideally helps reduce “dark uncertainty”
    - In this project, between-study uncertainty is allowed to impact each source individually

Decomposing Uncertainty

- In some cases, reported within study uncertainties are unreliable.
    - The reported uncertainties s^2 are suggested as ***lower bounds*** for the true uncertainties σ^2
    - For the *k-th* lab, $\sigma^2_k = s^2_k + \tau^2_k$



True Uncertainty = Within Study Uncertainty + Between Study Uncertainty

Why We Care About Uncertainty

- Combining results is more complicated than simply averaging all estimates.
- Values with smaller variances (i.e. more certainty) should get more weight.
- Problem: What if some within study uncertainties are unreliable?

i	1	2	3	4	5	6	7	8	9	10
x_i	6.6699	6.6726	6.6729	6.6735	6.6740	6.6742	6.6754	6.6830	6.6873	6.7154
s_i	0.0007	0.0008	0.0005	0.0029	0.0007	0.0007	0.0015	0.0011	0.0094	0.0005

Low uncertainty for apparent outlier

The Project

- Test newly derived estimators for combining independent study results against existing estimators frequently employed in meta-analyses in hopes of establishing a standard technique.
- Estimators to compare:
  - Classical Maximum Likelihood Estimator $(\hat{x}_L)$ ← New!
  - Restricted Maximum Likelihood Estimator $(\hat{x}_{RL})$ ← New!
  - Bayes Posterior Mean Estimator (δ_B) ← New!
  - Arithmetic Mean $(\bar{x})$ ← Well Established Technique
  - Graybill-Deal Estimator (Weighted Mean) $(\hat{x}_{GD})$ ← Well Established Technique
  - DerSimonian-Laird Estimator $(\hat{x}_{DL})$ ← Well Established Technique

Maximum Likelihood Estimators

- Unlike the previously established “state of the art” estimator (DerSimonian-Laird estimator), the between study uncertainties τ^2 are allowed to affect each individual study differently.
- Calculated as follows:
  - For each possible subset of labs
    - Estimate $\hat{\tau}^2_k > 0$ for each lab in the subset, set remaining $\hat{\tau}^2_k = 0$
    - Estimate $\hat{\sigma}^2_k = s^2_k + \hat{\tau}^2_k$
    - Using new uncertainties, compute $\hat{\mu}$ and its corresponding likelihoods
  - Choose $\hat{\mu}$ estimates with maximum likelihood

Bayes (Posterior Mean) Estimator

- Noninformative prior distribution.
- As with Maximum Likelihood estimators, the between study uncertainty (τ^2) is allowed to affect each study differently.
- For δ_B , there is added constraint that $\tau^2_k \geq s^2_k$
- Calculated using techniques of numerical integration.
- Here γ(y) =γ(y, 1/2) = $\pi^{-1/2} \int_0^y e^{-u}u^{-1/2}du$ is an incomplete gamma-function (with the related error function Erf(z) = $\gamma(z^2)$).

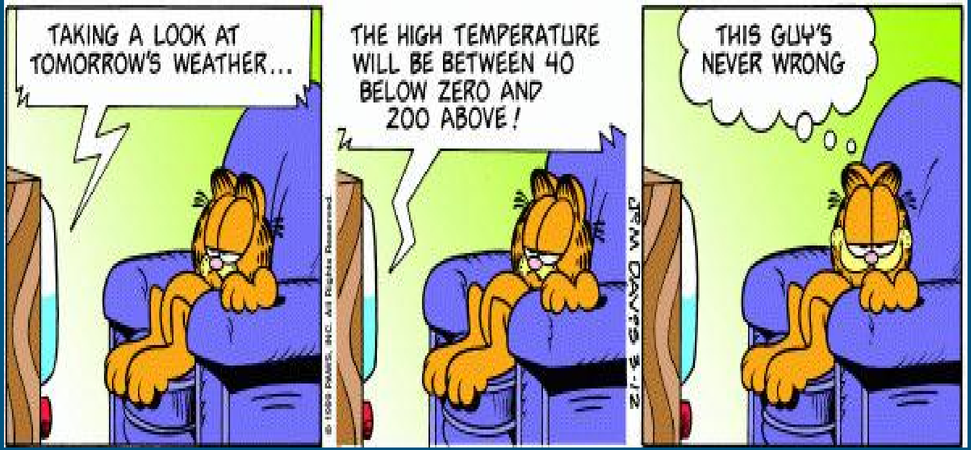
$$\delta_B = \int_{-\infty}^{\infty} \frac{\mu}{\prod_k |x_k - \mu|} \prod_k \gamma\left(\frac{(x_k-\mu)^2}{2s_k^2}\right) d\mu \left[\int_{-\infty}^{\infty} \frac{1}{\prod_k |x_k - \mu|} \prod_k \gamma\left(\frac{(x_k-\mu)^2}{2s_k^2}\right) d\mu\right]^{-1}$$

Procedure for Comparing Estimators

- Lab variances s^2 (within study uncertainties) are drawn at random from a uniform distribution between (0, 1) and fixed.
- Additional random noise multiplied by a factor $\lambda \in (0, 3]$ is added to some of the standard deviations to set “new” standard deviations for each simulation.
- Numbers are drawn at random from the standard normal distribution and multiplied by the corresponding “new” standard deviations, yielding lab means $\boldsymbol{x}$.
- For $\lambda \in (0, 3]$, 5000 random sets of lab means are generated. Estimators are computed for each set of lab means. λ increases spread between lab means.

Evaluating Estimator Efficacy

The true mean of simulated data is zero, as lab means were drawn from the standard normal distribution. Estimators were compared using the following three criteria:

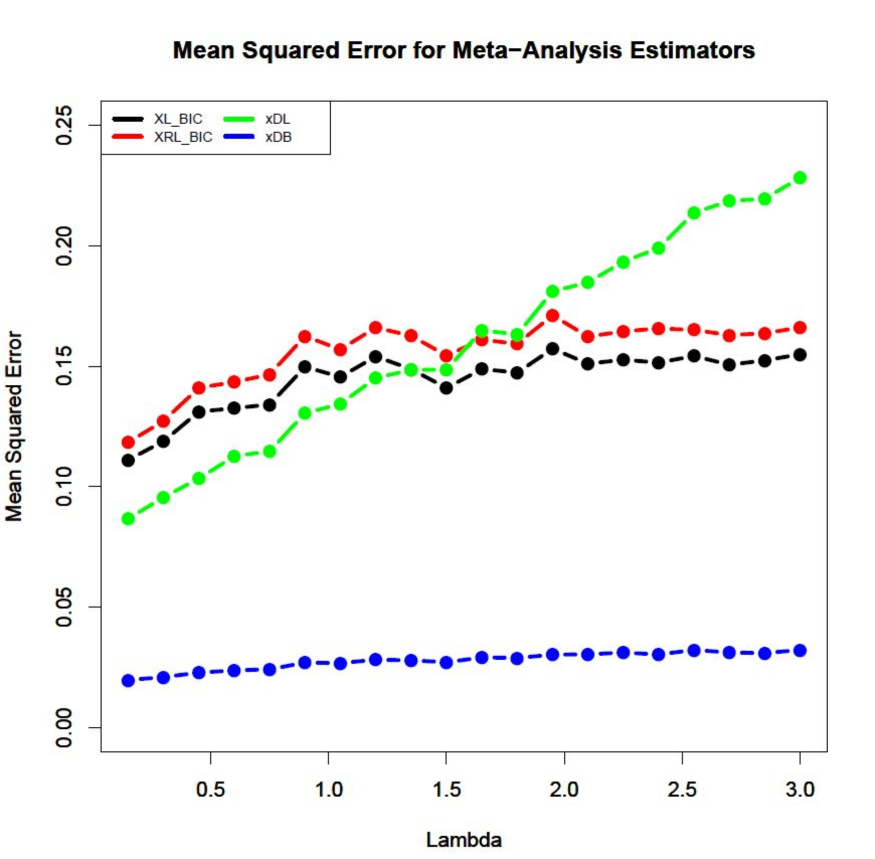

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
JPM DAVIS 3-12

http://nysora.com/files/Garfield_musing_CIs_533965604.jpg

- Mean Squared Error:
  - The average distance (squared) between each estimator and the true value.
- Coverage Probability:
  - The frequency with which each estimator's confidence interval contained the true value.
- Confidence Interval Width:
  - The average width of a 95% confidence interval for each estimator.
  - Ensures high coverage probability isn't result of excessively large interval width.

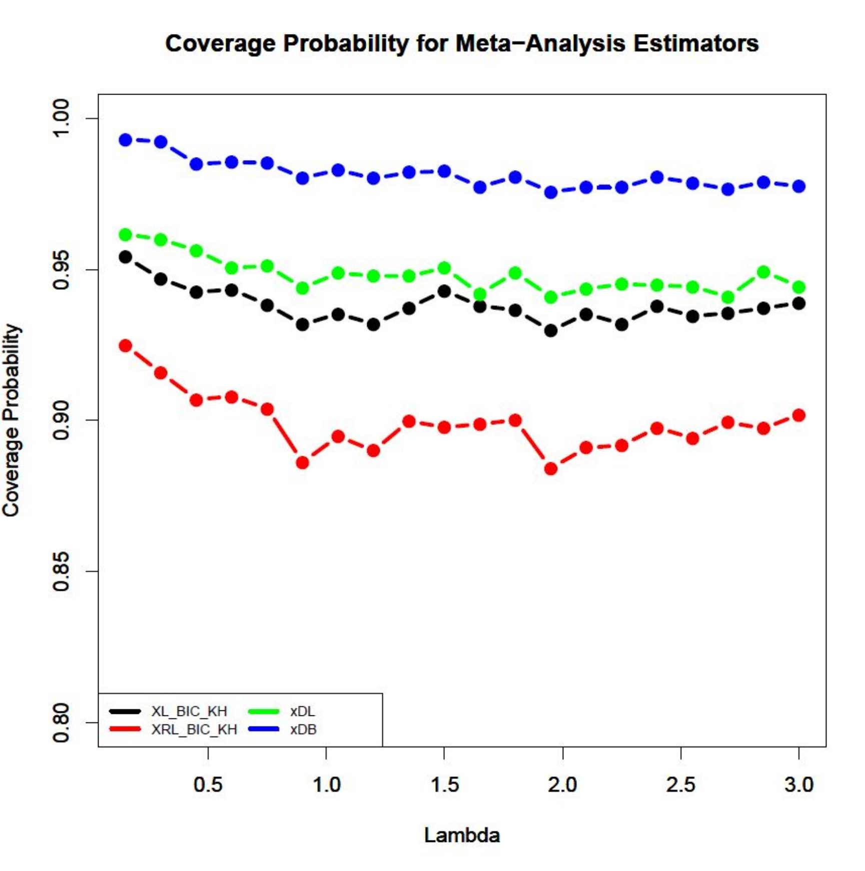
Results: Mean Squared Error

- The lower the Mean Squared Error, the closer the estimator is, on average, to the truth.
- Bayes δ_B clearly outperforms other estimators, even as λ gets larger.
- DerSimonian-Laird estimator seems to be most affected by increases in λ.
- The often used Graybill-Deal estimator and the simple arithmetic mean are much more affected by increase in λ than even $\hat{x}_{DL}$.



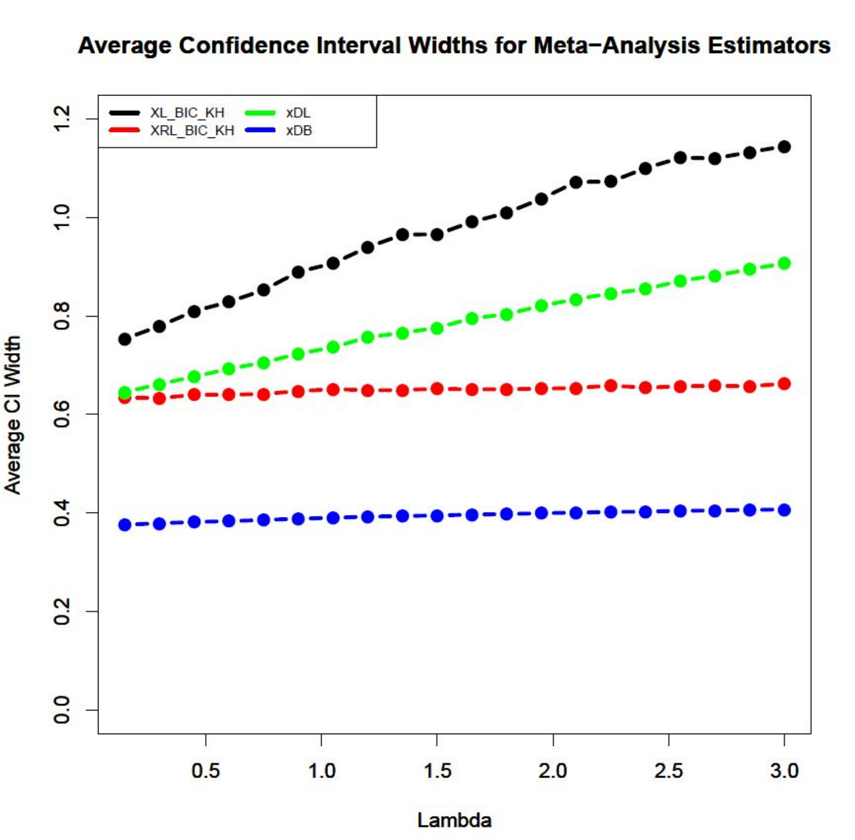
Results: Coverage Probabilities

- The larger the coverage probability, the greater the chance a random confidence interval for a given estimator will contain the truth.
- Bayes δ_B clearly outperforms other estimators, even as λ gets larger.
- Need to ensure that high coverage probability isn't the result of overly large confidence interval.



Results: Confidence Interval Widths

- Notice Bayes δ_B always has the smallest confidence interval width, on average.
- Unlike the DerSimonian Laird estimator, the width of confidence intervals for the Bayes δ_B estimator remains fairly constant as λ gets larger.



Back to Newton's Constant

- The new techniques outlined in this project were used to produce the following estimates for Newton's Gravitation constant using the data presented by Mohr and Taylor's 2000 study:
  - Maximum Likelihood (no penalty): $\hat{\mu}$ = 6.6736 10^{-11} $m^3kg^{-1}s^{-2}$
    - Increased uncertainty for labs 1, 3, 8, and 10
  - Maximum Likelihood (BIC penalty), REML, and Bayes δ_B: $\hat{\mu}$ = 6.6735 10^{-11} $m^3kg^{-1}s^{-2}$
    - Increased uncertainty for lab 10

i	1	2	3	4	5	6	7	8	9	10
x_i	6.6699	6.6726	6.6729	6.6735	6.6740	6.6742	6.6754	6.6830	6.6873	6.7154
s_i	0.0007	0.0008	0.0005	0.0029	0.0007	0.0007	0.0015	0.0011	0.0094	0.0005

Back to Newton's Constant

- Each of these estimators offers a better estimate of the current NIST standard value of μ = 6.67408 10^{-11} $m^3kg^{-1}s^{-2}$ than the value of 6.6833 reported by the 2000 study.
- Note that the current NIST standard value, created in 2014, uses more recent (and probably more reliable) data than was available at the time of the 2000 study worked in this presentation.
- A similar analysis on 2002 data produced maximum likelihood and REML estimates of $\hat{\mu}$ = 6.67413 10^{-11} $m^3kg^{-1}s^{-2}$

Conclusions and Notes

- These results suggest that Bayes δ_B may be better to use than the DerSimonian-Laird estimator when combining results in inter-laboratory studies (at least for small/medium sized studies).
- While the majority of simulations were run with the number of labs set at n = 9, simulations were also run with n = 5 and n = 12 to verify results. These numbers were chosen as they are plausible sizes for inter-laboratory studies as outlined by NIST.

Acknowledgements

- I would like to acknowledge my mentor Dr. Andrew Rukhin for his constant help along the way. I could not have done this project without him.
- I would also like to thank Dr. Antonio Possolo for providing me with code used to bootstrap confidence intervals for the DerSimonian-Laird Estimator. Additionally, I would like to thank Dr. Possolo for helping me understand his previous work examining the Newton Gravitation Constant.
- Finally, I would like to thank Will Guthrie and the entire Statistical Engineering Division for being so accommodating and helpful during my 11 weeks at NIST.
- Special thanks to the SURF program for funding my summer research.