Michelson's 1879 Determination of the Speed of Light: A Case Study in Bayesian Metrology

Bayesian Metrology Project
Statistical Engineering Division
National Institute of Standards and Technology
Gaithersburg, MD

Introduction

Michelson's 1880 article on the measurement of the speed of light (*Astronomical Papers*, **1**,1880, pp. 109-145) is a classic of both experimental physics and metrology. These notes consist of a partial re-analysis of the original data from this experiment. The analysis presented here is admittedly incomplete. It is intended primarily to illustrate the power of a Bayesian point of view in the analysis of high-quality data on a real-world measurement system.

Still, the results are of some historical interest. Some have argued that Michelson's published uncertainty was excessively pessimistic (e.g., McKay and Oldford, *Statistical Science*, **15**, 2000, p. 273). However, the uncertainty presented here is consistent with Michelson's original assessment. One important lesson to be learned from this is that it is not necessarily the case that a Bayesian approach will lead to smaller uncertainties. Rather, the Bayesian point of view provides a natural framework for accurate modeling of measurement systems, and, consequently, realistic assessments of uncertainty.

Outline: Part I

It is hoped that these notes will serve as a tutorial on the application of Bayesian methods to the analysis of a measurement system. So we will devote as much space to the preliminary model-building phase of the analysis as to the statement of the model and the discussion of results.

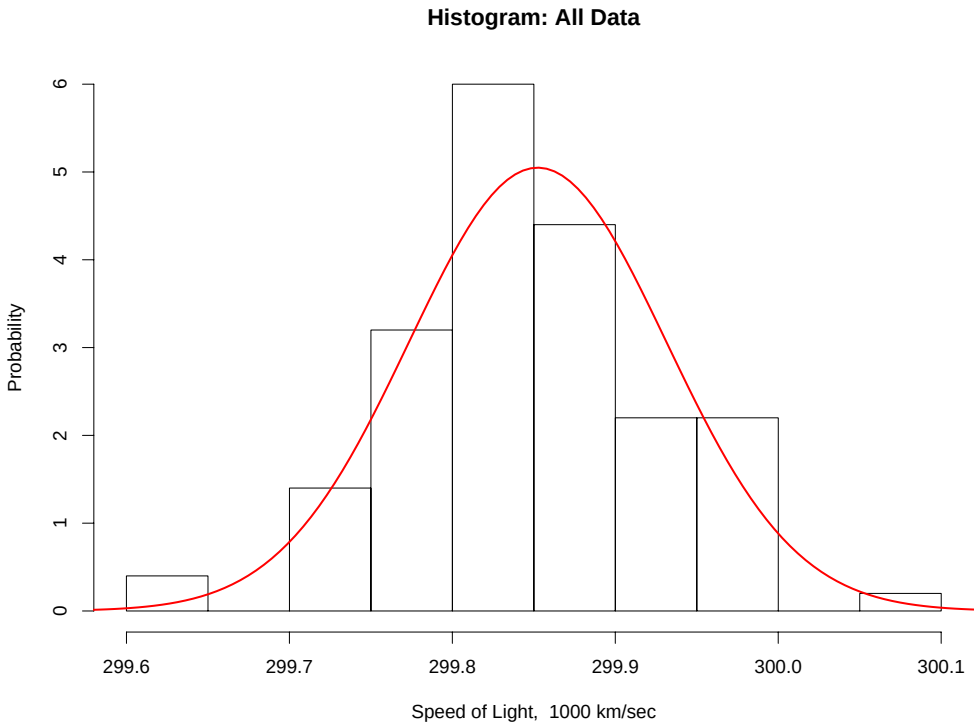
These remainder of these notes are organized in three sections. In **Part I**, we present the data, using various plots in order to informally examine its structure. This leads to four candidate statistical models, of which one is selected. This part of the analysis is not Bayesian. We could have applied Bayesian model selection techniques, but we did not in order to reduce the number of new ideas, and keep this introduction as straightforward as possible. The use of Bayesian model selection techniques would have been unlikely to change the results, at least for this example.

Outline: Parts II and III

In **Part II**, we use the results of Part I to formulate a reasonable Bayesian hierarchical model, incorporating temperature effects, variability among experimental runs, variability between data sets taken within the same run, and within-set measurement variability. We are also able to model the within-set variability itself, making use of information which Michelson provides on the relative quality of the image measurements.

Posterior inference on various quantities are obtained using the Gibbs sampler, and these results are interpreted in **Part III**. Here we estimate the overall uncertainty of Michelson's determination, we examine and compare the magnitudes and uncertainties regarding sources of variability, and we evaluate the "fit" of the model to the data.

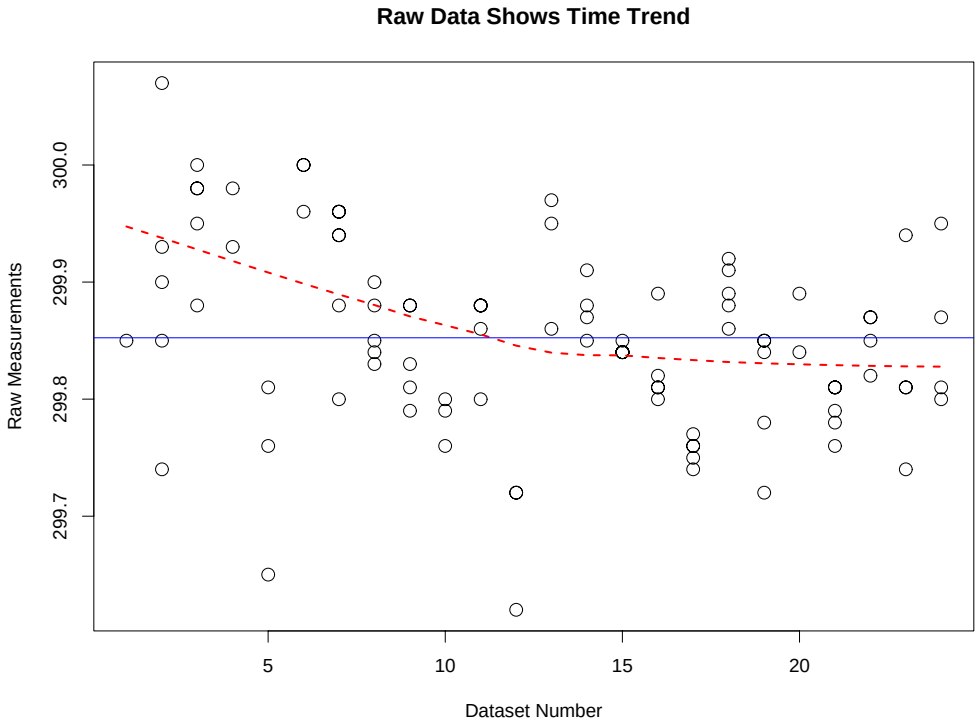
Part I: Exploratory Data Analysis



The data consist of 100 mesurements, made over 28 days. Each measurement was an average of several replicates. The following page shows a histogram of all of the data, with a superposed best-fitting normal (Gaussian) distribution.

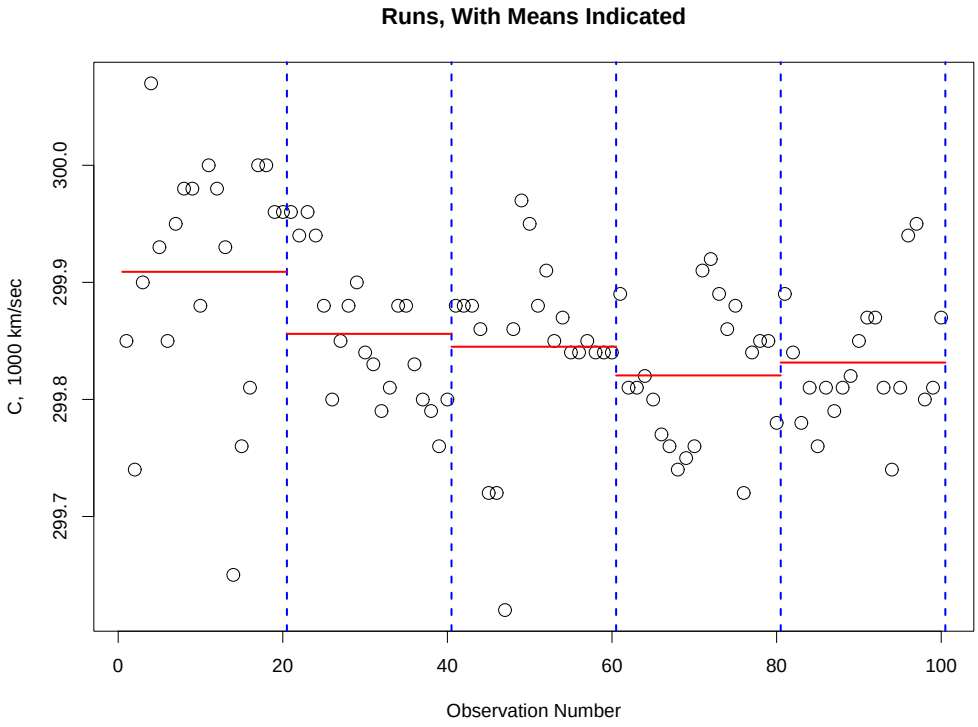
Although the simple bell-shaped curve provides a reasonable first approximation, it certainly leaves room for improvement. One of the objectives of the present analysis is to include details of the measurement process in a Bayesian statistical model. As a result, we will be able to better explain the variability in the measurements, and thus better characterize the measurement uncertainty.

The Effect of Measurement Day



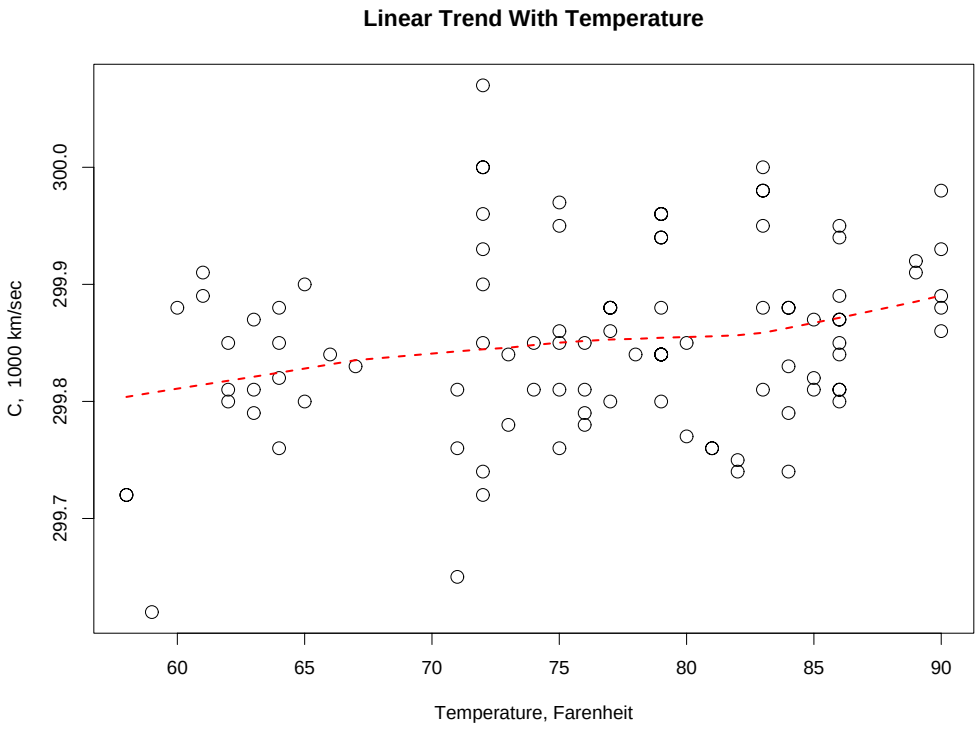
The 100 measurements were made on 18 distinct days. Michelson does not provide precise time-of-day information, but he does indicate whether the measurements were made in the morning or the evening. On some days there are both morning and evening measurements, for a total of 24 distinct combinations of day and time-of-day, which we will call 24 data *sets*. The sets are numbered chronologically, and displayed here along with the corresponding mean determinations. Note the slight, but probably not random, time trend, indicated here by the broken line (a *lowess* local regression smooth of the data).

Measurement Runs



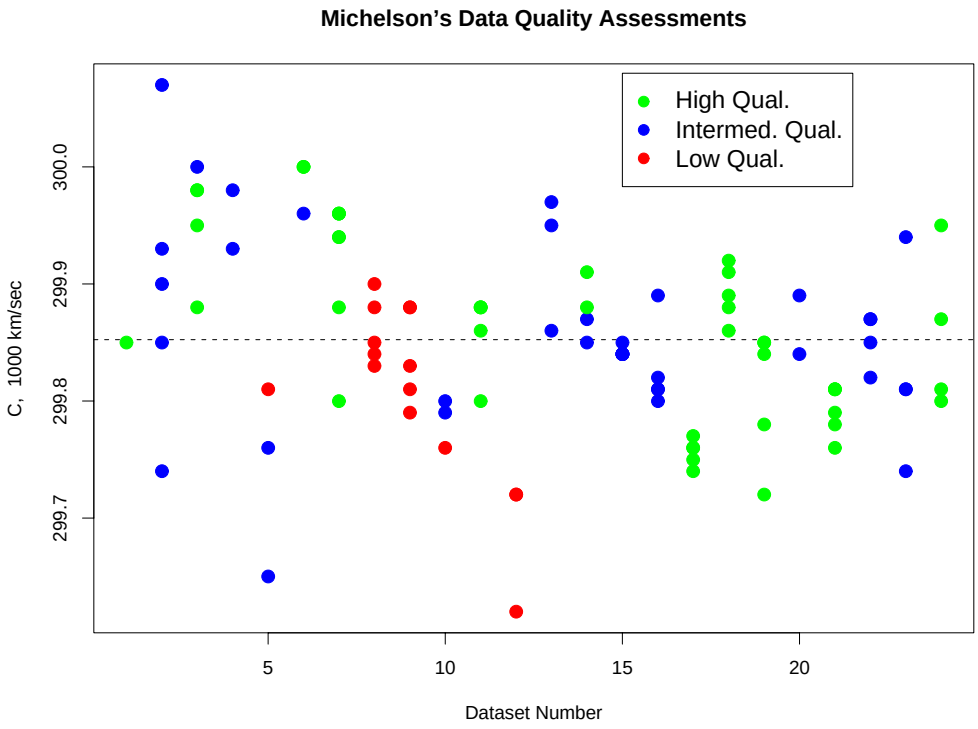
The 100 measurements were made in 5 *runs* of 20 measurements each. Presumably something was done to the measurement process which distinguishes these runs. Perhaps there was some "tuning" or calibration of the apparatus. In the above plot, the 100 measurements are shown chronologically, with vertical lines delimiting the runs. The horizontal lines indicate the average measurement within each run. It can be seen that there is evidence of possible variability between runs, and that this might explain the apparent trend with set number (equivalently, measurement day).

Temperature Effect



Michelson also provides the air temperature (to the nearest degree) at the time that each measurement was made. The above plot shows that temperature, in addition to run and set number, should be considered as an independent variable (covariate) in a model of the speed of light measurements. We now have three candidate covariates: dataset number (day/time of measurement), measurement run, and temperature. In order to see which of these factors are useful in explaining the apparently non-random behavior of these measurements, we will fit and compare least-squares regression models. Once we've decided which factors to use, then we will proceed with Bayesian modeling and model interpretation.

Data Quality



An essential part of the measurement involved the viewing of an image. Michelson was so concerned with errors introduced in this part of the measurement that he ranked the quality of the image viewed on a three-level scale: good, intermediate, and poor. (He included all of the data in his final analysis, though). This plot displays the measurements against data set number (i.e., "time"), with colors indicating the three grades of quality. We will make use of this information in our model by allowing the measurement standard deviation to be different for different image qualities.

Part II: Modeling

We consider 4 models in order of increasing complexity:

Model 0 Simple random sample, with a constant mean

Model 1 Model 0, and a linear dependence of the mean on temperature;

Model 2 Model 1, plus a random shift in the mean for each run;

Model 3 Model 2, plus an additional random shift in the mean for each set.

Models are fit by least squares (*regression*), and compared using 3 *P-values* comparing Model 3 with 2, 2 with 1, and 1 with 0. A P-value is a frequentist measure of statistical significance. Small values indicate high significance. Values of less than 0.05 would traditionally be taken to mean that the more complicated model offers a *statistically significant* improvement over the simpler model. There are Bayesian alternatives to using P-values which would be preferable here. But, as we will see, the P-values for the above comparison are so small that it is highly unlikely that a fully Bayesian approach would have changed the conclusion.

Selecting a Model for the Mean

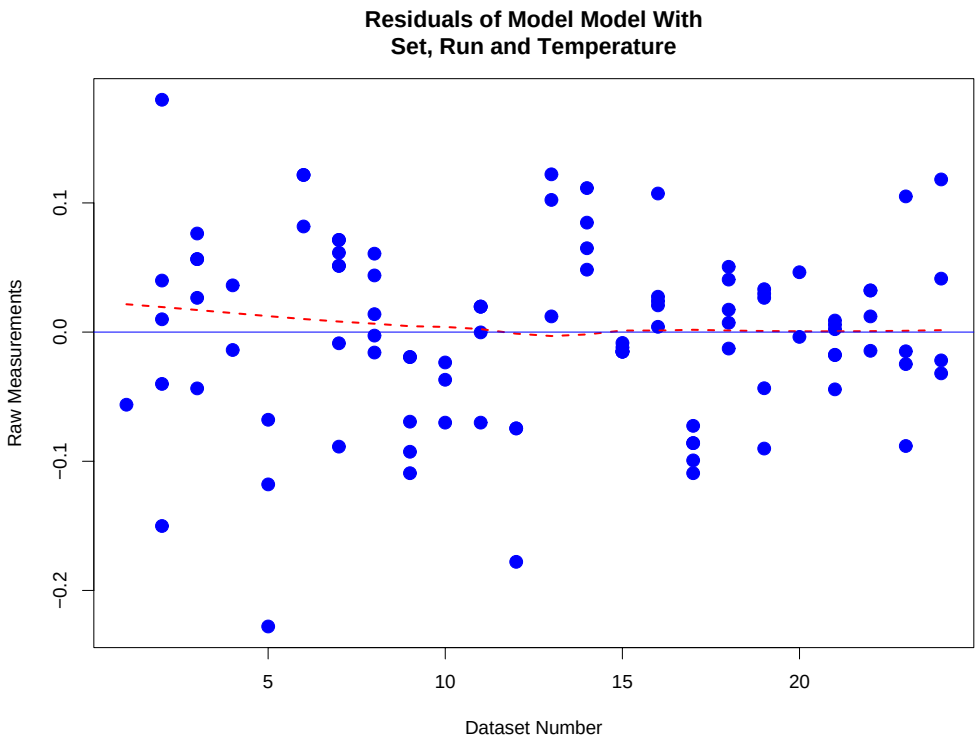
$$c_i = \alpha + \beta(T_i - \bar{T}) + r_i + s_i + e_i$$

where T_i is the temperature for measurement i, r_i and s_i are shifts in the mean due to the corresponding run and set, and e_i is the remaining measurement error.

Model	P-value	
Model 0	α	
Model 1	$\alpha + (T_i - \bar{T})\beta$	$P = 0.008$
Model 2	$\alpha + (T_i - \bar{T})\beta + r_i$	$P = 0.0003322$
Model 3	$\alpha + (T_i - \bar{T})\beta + r_i + s_i$	$P = 0.00000015$

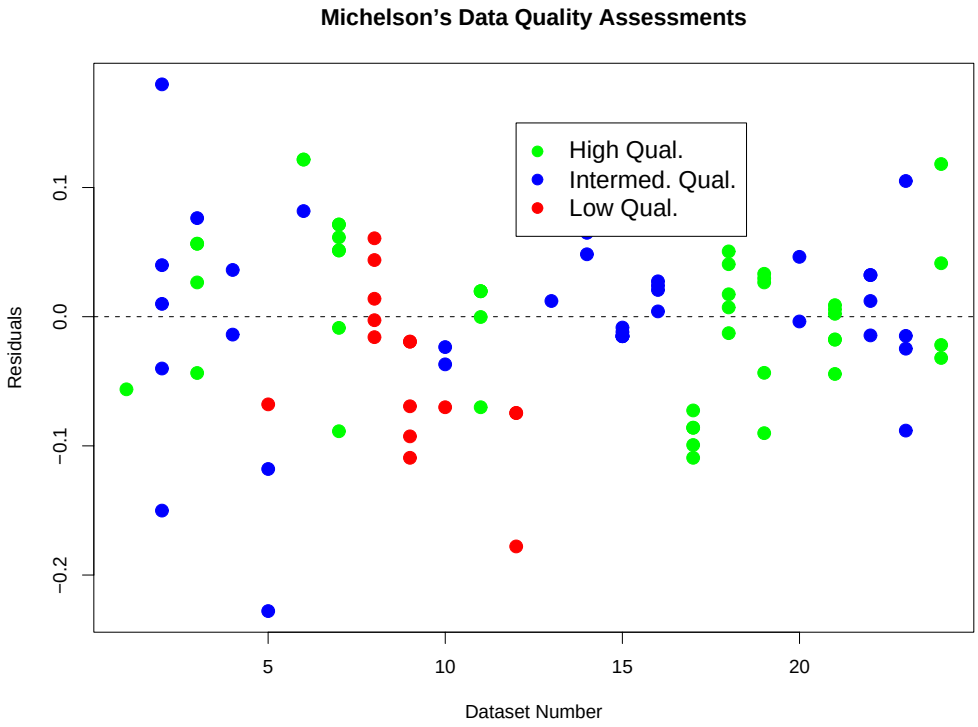
The P-values are measures of significance, comparing the each model with the one tabulated immediately above. All of them are substantially less than 0.05, suggesting that the complexity of Model 3 is justified. This model assumes that the mean speed of light measurement depends linearly on temperature, with additional random shifts due to run and set.

Residual Plot for Model 3



The residuals from the regression fit of Model 3 are displayed above. These residuals vary about zero, with no obvious pattern, suggesting that the model adequately explains the structure in the data. Consequently, a locally-linear *lowess* fit to the data, indicated by the broken line, is quite flat. This should be compared with plot against temperature shown previously, where the lowess line showed some trend.

Residuals with Michelson's Quality Indications



A second residual plot appears above, with the data points coded according to quality. Note that nothing remains of some of the patterns which were apparent in the previous plot of these quality indications (such as many of the low-quality "red" points being low). But there doesn't seem any reason to belive that the "best" points are in some way preferable, either in terms of mean or variability. Still, we will allow the within-set measurement standard deviation depend on these quality categories in our model, and compare them *a posteriori*.

Part 3: Bayesian Modeling and Analysis

In this section, we propose a Bayesian hierarchical model based on what we've learned during the data exploration and model selection phases of this study. In this model, the ith measurement is assumed to have a normal (Gaussian) distribution, centered at a mean μ_i, which depends on temperature, run and data set. The standard deviation for this normal distribution $\sigma_{e,j}$, will be allowed to differ, depending on the data quality class of the measurement ($j = 1, 2, 3$). The mean μ_i depends on random effects for run (r_i) and set (s_i), which will also be assumed to have normal distributions, centered at zero, and with standard deviations σ_r and σ_s, respectively. All of these parameters are also assumed to have prior distributions. These second-level (*hyperprior*) distributions are chosen to be "noninformative". That is, these hyperpriors are assigned functional forms which can be expected not to introduce information not present in the data. (These noninformative distributions are not even proper probability distributions, but this does not introduce any difficulties, either conceptually or computationally).

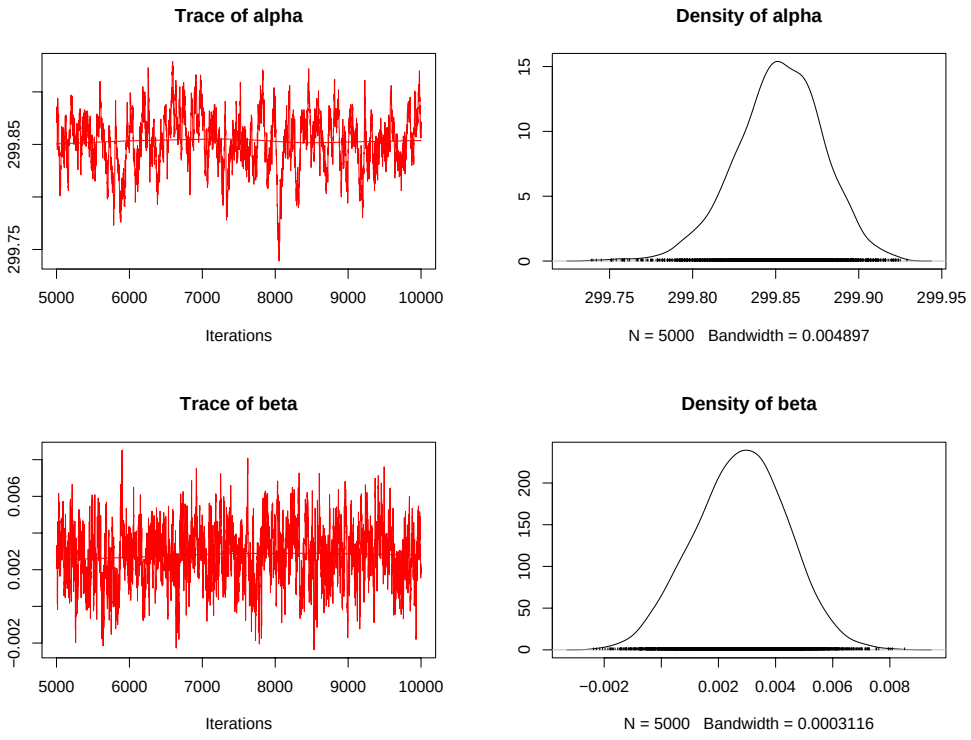
The model is fit using the *Gibbs sampler*, a special case of *Markov Chain Monte Carlo*, using the software package **Bugs**. Markov chains are checked for convergence, and posterior distributions are estimated and interpreted for various quantities of interest.

A Bayesian Hierarchical Model Motivated by Model 3

$$
\begin{aligned}
p(c_i|\mu_i, \sigma_{e,j}) &\sim \mathsf{N}\,(\mu_\mathsf{i}, \sigma_{\mathsf{e,j}}) \\
\mu_i &= \alpha + \beta(T_i - \bar{T}) + r_i + s_i \\
p(\alpha) &\propto \text{const.} \\
p(\beta) &\propto \text{const.} \\
p(r_i|\sigma_r) &\sim \mathsf{N}\,(0, \sigma_\mathsf{r}^2) \\
p(s_i) &\sim \mathsf{N}\,(0, \sigma_\mathsf{s}^2) \\
p(\sigma_r) &\propto 1/\sigma_r \\
p(\sigma_s) &\propto 1/\sigma_s \\
p(\sigma_{e,j}) &\propto 1/\sigma_{e,j}
\end{aligned}
$$

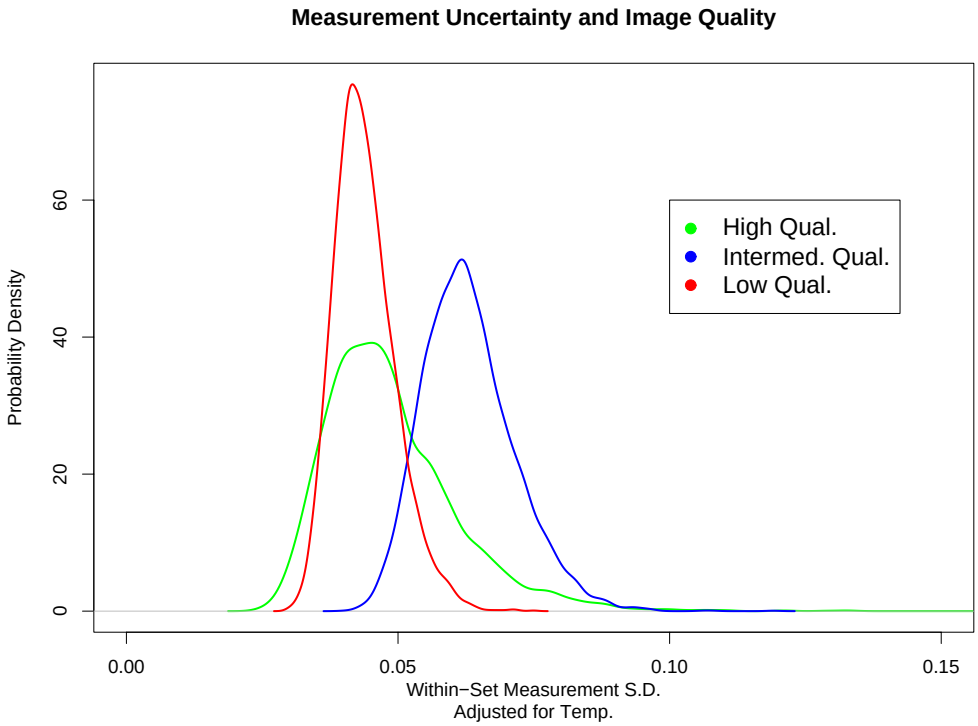
The likelihood, priors, and hyperpriors are given above for a Bayesian model which includes a "fixed effect" for temperature, and "random effects" for data set in run. Also, the measurement error $\sigma_{e,j}$ is allowed to depend on the data quality class.

Posterior Distributions of α and β



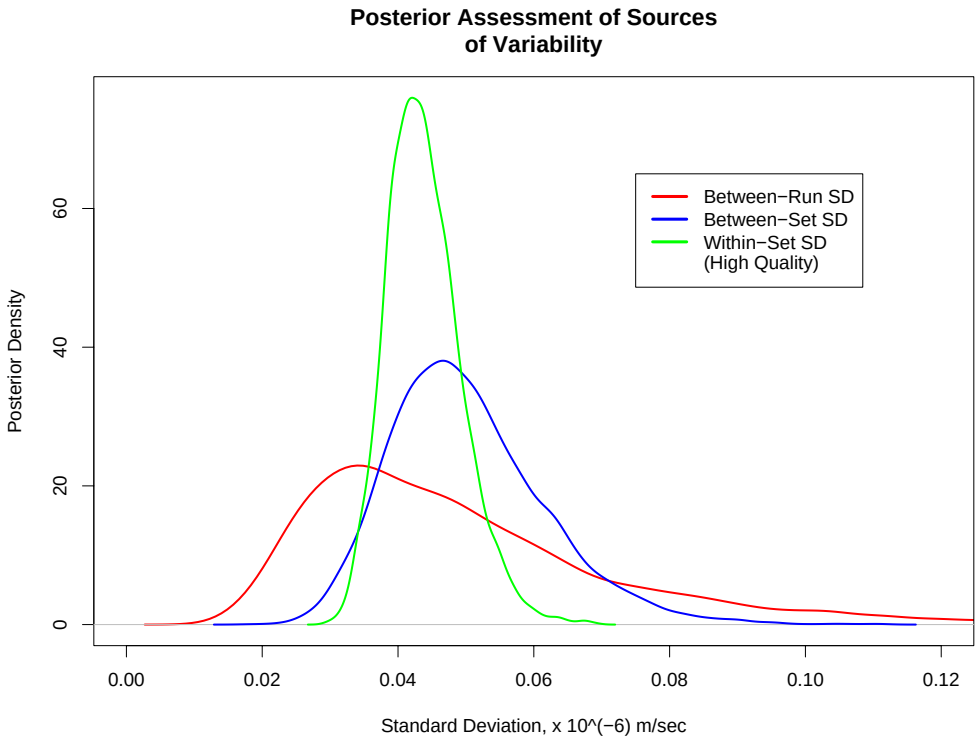
The Gibbs sampler approximates the posterior distributions of parameters by choosing random samples from *full conditional* distributions, which are univariate probability distributions in which all parameters are regarded as constant except one. There are as thus as many full conditional distributions as there are parameters in the model, and these are sampled from sequentially, updating the fixed parameters as each random draw is made. The figure above illustrates approximate posterior probability distributions for the slope and intercept in our model. The density estimates in the right hand figures were obtained from 5000 approximate random samples from the respective distributions, shown above to the left.

Posteriors of Measurement Error Standard Deviations



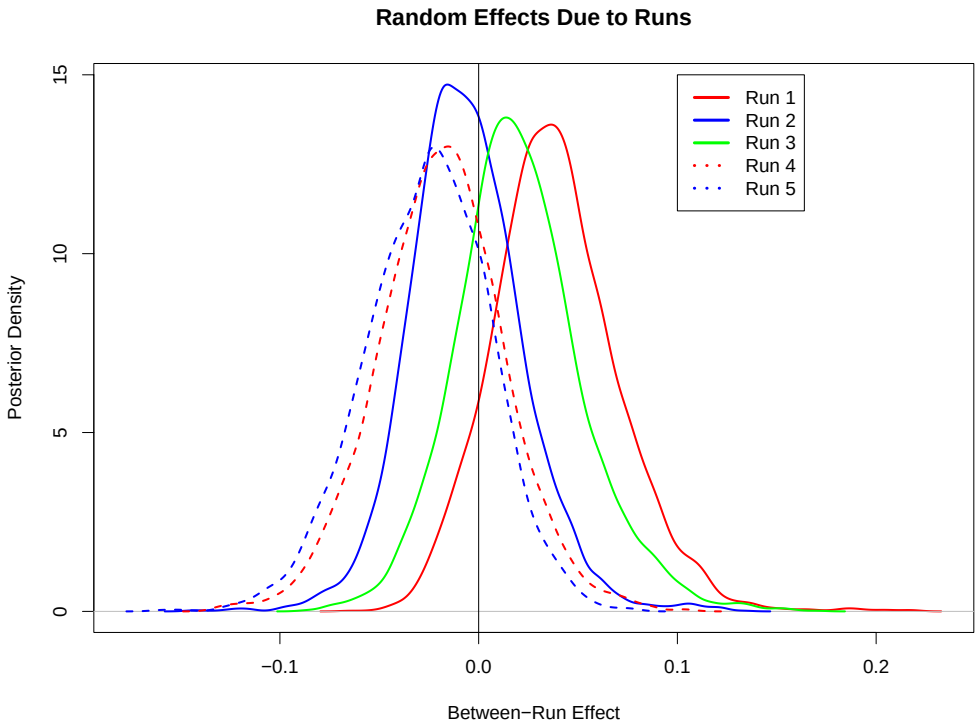
The figure above illustrates the three posterior distributions corresponding to the three within-set error variances $\sigma_{e,j}$ corresponding to the three data quality classes. Although these standard deviation distributions appear to be different, there is no obvious pattern. For example, the "best" measurements do not tend to have the smallest estimated measurement error. This could be because some of the things that made measurements good or bad are already accounted for in the between-set, between-run, and temperature effects in the model.

Posteriors of Run, Set and Measurement Error Standard Deviations



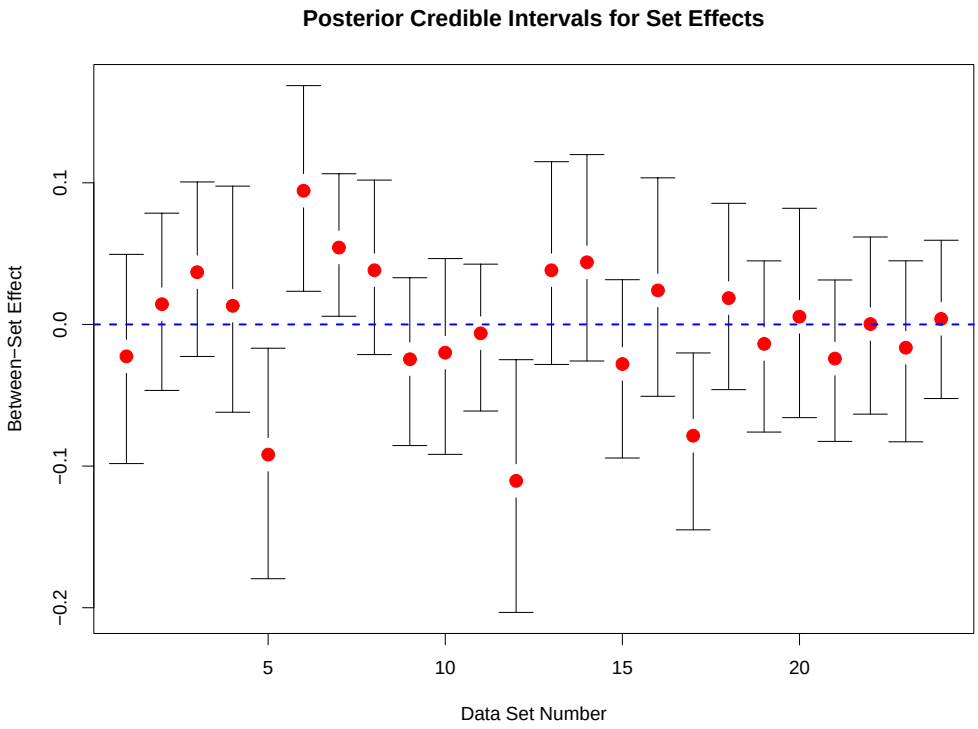
The above plots compare the posterior distributions of the standard deviations corresponding to three components of variability. The values of these standard deviations are comparable. In other words, the variability between-runs and between-sets are each of roughly the same magnitude as the within-set measurement uncertainty; hence it is important to include these sources of variability in any model of this measurement system. The more diffuse the posterior is, the more uncertainty there is in the corresponding quantity (eg., there were 100 measurements, but only 5 runs).

Posterior Distribution of Run Effects



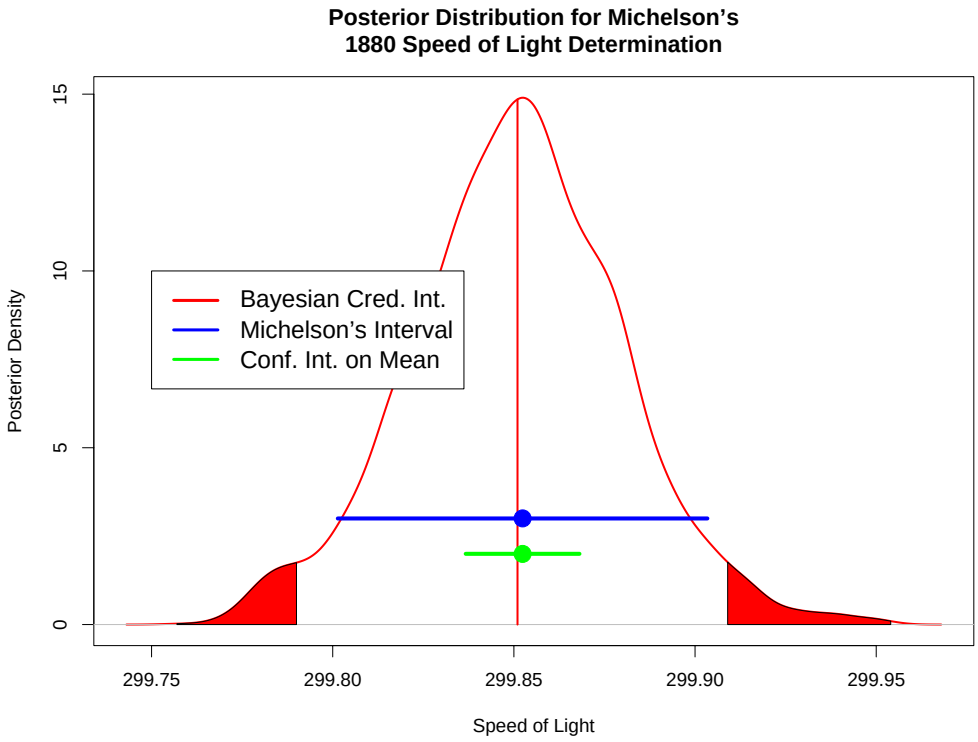
The posterior distributions for the shifts in the mean due to each of the five runs are displayed above. Note the trend in the peaks of these distributions, corresponding to the decreasing trend in measurements due to run which was observed graphically earlier.

Posterior Distribution of Data Set Effects



The above display provides 95% posterior probability intervals (*credible* intervals) for the shift in the mean measurement due to data set, adjusting for run and temperature. Note that many of these posteriors are more or less centered on the horizontal zero line, indicating set effects which are probably near zero. On the other hand, a few of the set effects differ substantially from zero. In a more complete analysis, one would look back at the measurements on those days, to see if there is evidence of anything out of the ordinary occurring.

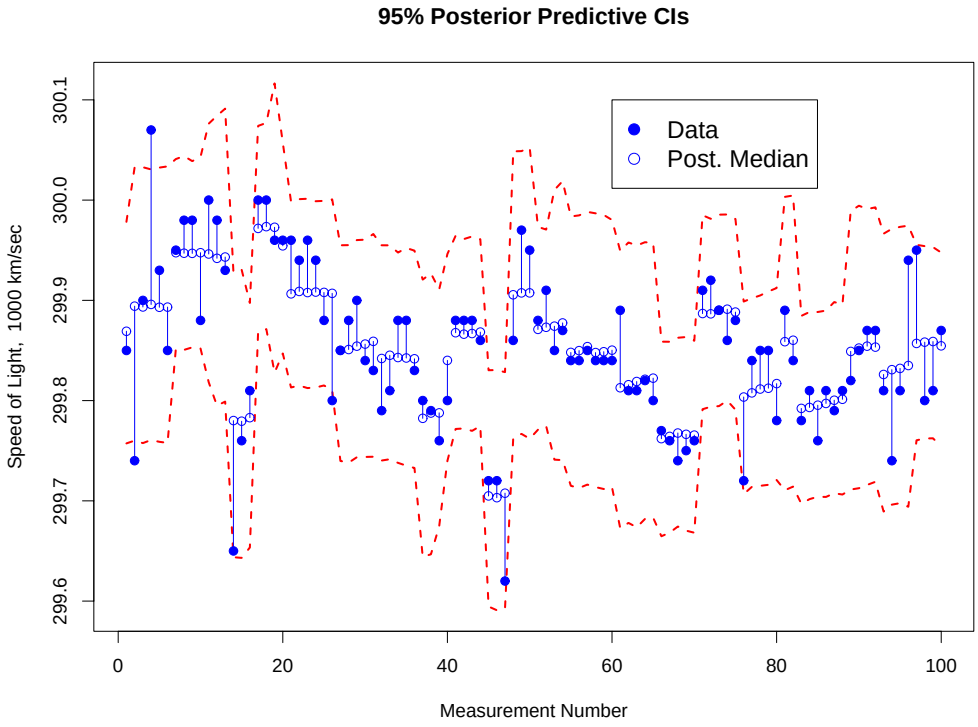
Posterior Uncertainty in Speed of Light from Michelson's Data



The above figure is probably the most interesting one in this analysis. The posterior distribution in α reflects the uncertainty in the speed of light in Michelson's data, based on our chosen model. The shaded areas cut off 2.5% probability in each tail, so that the interval between these shaded regions represents a 95% uncertainty interval on c. The two superposed intervals are for a simple random sample analysis, and for Michelson's published uncertainty. (Michelson had adjusted his result to correspond to the speed of light in a vacuum; we shift the center of his interval here so that the results can be more easily compared.)

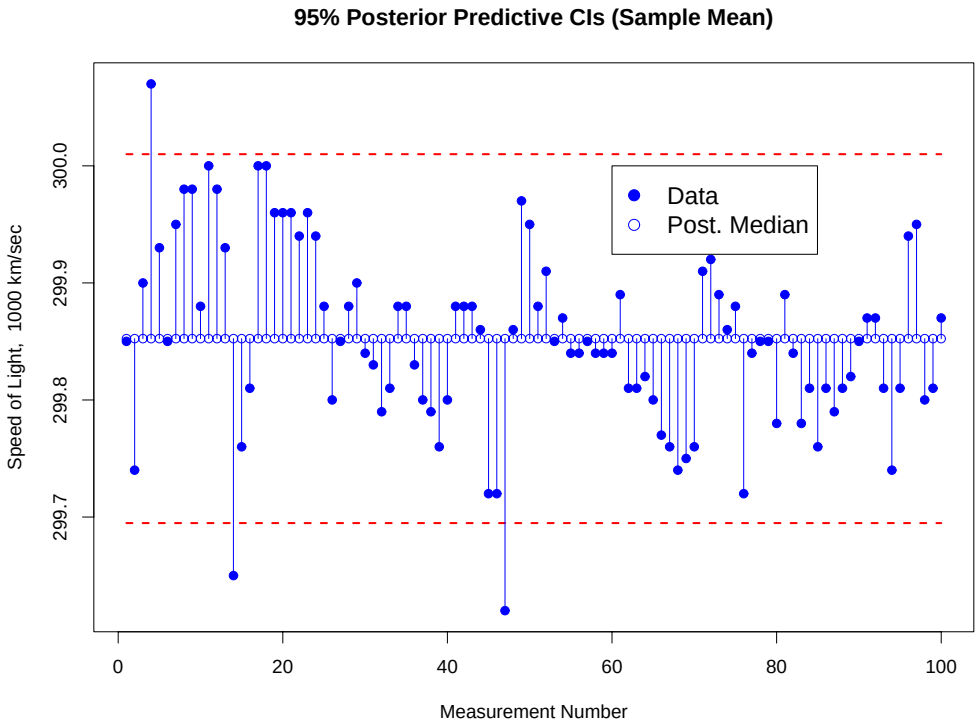
Michelson's knew that the simple random sample interval was too narrow. He included various sources of uncertainty in much the same way that a modern metrologist would compile an "uncertainty budget" of Type A and Type B uncertainties. We have reached a similar conclusion, without his detailed knowledge of the measurement system, by examining and modeling the apparent sources of variability in his measurement process. Although the striking agreement above is in part due to chance, it strongly suggests consistency with Michelson's original assessment.

Posterior Predictive Goodness-of-Fit: Bayesian Hierarchical Model



In order to see visually how well the model fits the data, posterior predictions for each of the 100 measurements were made. The above plot shows 100 95% posterior predictive probability intervals (the broken lines). The solid and open circles indicate the median of the posterior, and the data values. The fit seems to be adequate.

Posterior Predictive Goodness-of-Fit: Simple Random Sample



The above plot corresponds to the one on the previous page, except that the model is now our "Model 0", a common mean and variance for all observations. For this model, the posterior predictive distribution is a common t-distribution for all the observations; hence the parallel lines for the uncertainties and the median predictions. Although most of the observations fall within the prediction intervals, the fit is obviously inferior to that of Model 3.