

# Likelihood Basis for Multiple Data Sources

Keith O'Rourke

Visiting Assistant Professor,

Dept of Statistical Science, Duke University

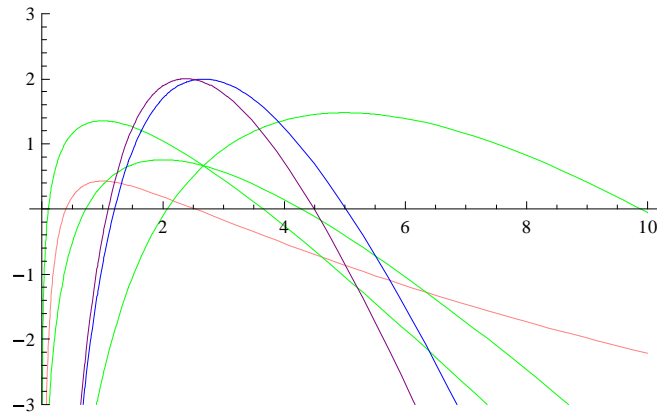
Likelihood for multiple observations in a single data source, focusing on one observation at a time.

A perhaps novel, but at first distracting means to expound the likelihood basis for multiple data sources is to first focus on individual observation likelihoods in a single data source. These individual likelihoods are defined as

$$L(\theta_i; y_i) = c(y_i) \Pr(y_i; \theta_i)$$

preferably with the choice of  $c(y_i)$  to make

$$L(\theta_i; y_i) = \Pr(y_i; \theta_i) / \Pr(y_i; \hat{\theta}_i).$$



You really always want (lost of) such a plot(s)

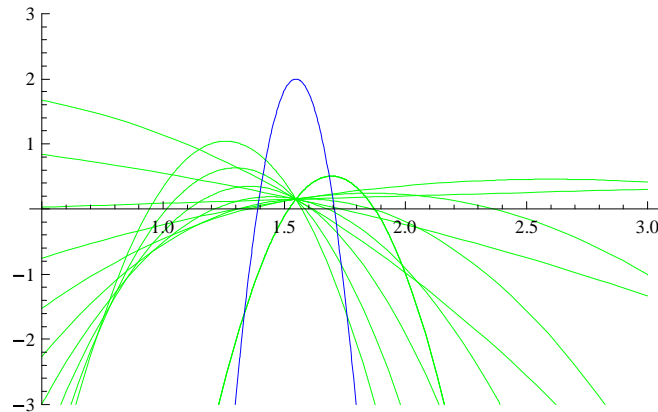
To contrast prior versus combined likelihood

To contrast combined likelihood with individual likelihoods

To transparently add up everything

If all quadratic - just generalized least squares picture

Evasions of nuisance parameters really challenging (impossible?)



Drop one green curve, other green ones change and add to something different

Change prior and green curves (sometimes) change

More usual regression diagnostic plots - drop one curve rest add to  $(n-1)$  case

Using likelihood to combine Tibshirani and Efron's cross-validation Logistic regression coefficients and contrast this with "Pre-validation".

At the 2003 Joint Statistical Meeting in San Francisco, Robert Tibshirani gave a talk entitled "Pre-Validation and Inference in Microarrays", joint work with Bradley Efron, in which he reported on work from their earlier paper[14]. The abstract (which was the same for both the paper and talk) was as follows -

“In microarray studies, an important problem is to compare a predictor of disease outcome derived from gene expression levels to standard clinical predictors. Comparing them on the same dataset that was used to derive the microarray predictor can lead to results strongly biased in favor of the microarray predictor. We propose a new technique called "pre-validation" for making a fairer comparison between the two sets of predictors. We study the method analytically and explore its application in a recent study on breast cancer.”

The new technique called "pre-validation" had been intuitively motivated and their work involved a more rigorous evaluation of its properties - in particular determining if the degrees of freedom for the microarray predictor using the pre-validation technique were correct. Both in the talk and the paper, an alternative method was first described as the usual k-fold cross-validation approach, where the microarray predictor and clinical predictor were compared on subsets of data that omitted data on which the microarray predictor was developed, and then these comparisons on subsets were averaged (unweighted) in the paper and "somehow to be combined" in the talk. When asked about the "somehow to be combined", Robert Tibshirani made it clear that likelihood combination had not been considered.

First, the details from the paper are convenient to quote and provide a concise summary -

“The microarray predictor was constructed as follows:

1. 70 genes were selected, having largest absolute correlation with the 78 class labels
2. Using these 70 genes, a nearest centroid classifier (described in detail in Section 6) was constructed.
3. Applying the classifier to the 78 microarrays gave a dichotomous predictor  $z_j$  for each case  $j$ .

It was of interest to compare this predictor to a number of clinical predictors ...

In order to avoid the overfitting problem ... we might try to use some sort of cross-validation:

1. Divide the cases up into say  $K$  approximately equal-sized parts
2. Set aside one of parts. Using the other  $K - 1$  parts, select the 70 genes having the largest absolute correlation with the class labels, and form a nearest centroid classifier.
3. Fit a logistic model to the  $k$ th part, using the microarray class predictor and clinical predictors
4. Do steps 2 and 3 for each of the  $k = 1, 2, \dots, K$  parts, and average the results from the  $K$  resulting logistic models.

The main problem with this idea is step 3, where there will typically be too few cases to fit the model. In the above example, with  $K = 10$ , the 10th part would consist of only 7 or 8 cases. Using a smaller value of  $K$  (say 5) would yield a larger number of cases, but then might make the training sets too small in step 2. Use of multiple random splits can help cross-validation a little in this case.

Pre-validation is a variation on cross-validation that avoids these problems. It derives a “fairer” version of the microarray predictor, and then this predictor is fit along side the clinical predictors in the usual way.

1. Divide the cases up into  $K = 13$  equal-sized parts of 6 cases each.
2. Set aside one of parts. Using only the data from the other 12 parts, select the genes having absolute correlation at least .3 with the class labels, and form a nearest centroid classification rule.
3. Use the rule to predict the class labels for the 13th part
4. Do steps 2 and 3 for each of the 13 parts, yielding a “pre-validated” microarray predictor  $\tilde{z}_j$  for each of the 78 cases.
5. Fit a logistic regression model to the pre-validated microarray predictor and the 6 clinical predictors.”

In their particular example Tibshirani and Efron used a linear logistic model. Now the separate  $K$  within group linear logistic models with parameters  $\alpha, \beta$  and  $\gamma$  within the separate  $K$  groups would be

$$f \left( y_{g(i)} \mid \frac{\exp(\alpha_{g(i)} + \tilde{z}_{g(i)}\beta_{g(i)} + c_{g(i)}\gamma_{g(i)})}{1 + \exp(\alpha_{g(i)} + \tilde{z}_{g(i)}\beta_{g(i)} + c_{g(i)}\gamma_{g(i)})} \right)$$

averaging the coefficients  $\alpha_{g(i)}, \beta_{g(i)}$  and  $\gamma_{g(i)}$  "somehow" presumes common (or common in distribution)  $\alpha, \beta$  and  $\gamma$  across the groups

$$f \left( y_{g(i)} \mid \frac{\exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)}{1 + \exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)} \right)$$

and their combination via likelihood multiplication together (assuming just common  $\alpha$  and  $\gamma$ ) is

$$\prod_{g(i)} f \left( y_{g(i)} \mid \frac{\exp(\alpha + \tilde{z}_{g(i)}\beta_{g(i)} + c_{g(i)}\gamma)}{1 + \exp(\alpha + \tilde{z}_{g(i)}\beta_{g(i)} + c_{g(i)}\gamma)} \right).$$

and further adventurously assuming common  $\beta$  is

$$\prod_{g(i)} f \left( y_{g(i)} \mid \frac{\exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)}{1 + \exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)} \right).$$

On the other hand, pre-validation is defined by Tibshirani and Efron, starting with an expression predictor  $z = (z_1, z_2, \dots, z_n)$  which is adaptively chosen from the data  $X$  and  $y$

$$z_j = f_{X,y}(x_j)$$

where their notation indicates that  $z_j$  is a function of the data  $X$  and  $y$ , and is evaluated at  $x_j$ . Rather than fit  $f_{X,y}$  using all  $X$  and  $y$  Tibshirani and Efron instead divide the observations into  $K$  roughly equal-sized groups, and denote by  $g(k)$  the observations composing each part  $k$ . For  $k = 1, 2, \dots, K$ , Tibshirani and Efron form the pre-validated predictor

$$\tilde{z}_{g(k)} = f_{X-g(k), y-g(k)}(x_{g(k)}); \text{ for } k = 1, 2, \dots, K$$

where the notation indicates that cases  $g(k)$  have been removed from  $X$  and  $y$ . Finally, Tibshirani and Efron fit the model to predict  $y$  from  $\tilde{z}$  and the clinical covariates  $c$ , and compare the contributions of  $\tilde{z}$  and  $c$  in this prediction - i.e. the  $\tilde{z} = (\tilde{z}_{g(1)}, \tilde{z}_{g(2)}, \dots, \tilde{z}_{g(n)})$  are included with  $c$  in a multivariate statistical model.

The linear logistic model here is

$$f \left( y \mid \frac{\exp(\alpha + \tilde{z}\beta + c\gamma)}{1 + \exp(\alpha + \tilde{z}\beta + c\gamma)} \right).$$

Now this the multiple of the individual observation likelihoods

$$\begin{aligned} & f \left( y \mid \frac{\exp(\alpha + \tilde{z}\beta + c\gamma)}{1 + \exp(\alpha + \tilde{z}\beta + c\gamma)} \right) \\ &= \prod_j^n f \left( y_j \mid \frac{\exp(\alpha + \tilde{z}_j\beta + c_j\gamma)}{1 + \exp(\alpha + \tilde{z}_j\beta + c_j\gamma)} \right) \end{aligned}$$

and recall these can be grouped within the separate cross validation groups

$$\begin{aligned} & f \left( y_{g(i)} \mid \frac{\exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)}{1 + \exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)} \right) \\ &= \prod_{j \in g(i)} f \left( y_j \mid \frac{\exp(\alpha + \tilde{z}_j\beta + c_j\gamma)}{1 + \exp(\alpha + \tilde{z}_j\beta + c_j\gamma)} \right) \end{aligned}$$

and then multiplied together

$$\begin{aligned} & f \left( y \mid \frac{\exp(\alpha + \tilde{z}\beta + c\gamma)}{1 + \exp(\alpha + \tilde{z}\beta + c\gamma)} \right) \\ &= \prod_{g(i)} \prod_{j \in g(i)} f \left( y_j \mid \frac{\exp(\alpha + \tilde{z}_j\beta + c_j\gamma)}{1 + \exp(\alpha + \tilde{z}_j\beta + c_j\gamma)} \right) \end{aligned}$$

to show that pre-validation is just the likelihood combination assuming common parameters  $\alpha, \beta$  and  $\gamma$

$$\begin{aligned} & f \left( y \mid \frac{\exp(\alpha + \tilde{z}\beta + c\gamma)}{1 + \exp(\alpha + \tilde{z}\beta + c\gamma)} \right) \\ &= \prod_{g(i)} f \left( y_{g(i)} \mid \frac{\exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)}{1 + \exp(\alpha + \tilde{z}_{g(i)}\beta + c_{g(i)}\gamma)} \right) \end{aligned}$$

But  $\beta$  surely would almost never be common - the  $\tilde{z}_{g(i)}$  are different functions of the data for different cross-validation groups! Each of the components of

$$(\tilde{z}_{g(1)}, \tilde{z}_{g(2)}, \dots, \tilde{z}_{g(n)})$$

are different functions of  $X$ . What about the assumption of common  $\alpha$  and  $\gamma$  now? Recall that their meaning is determined by the other variables that are fit in the model. How about the  $\beta_j$  being drawn from a common distribution? Which distribution? How about the  $\alpha_j$  and  $\gamma_j$  also being drawn (dependently) from a common distribution? How should we pool or partially pool to avoid arbitrary parameters for each  $(\alpha_{g(i)}, \beta_{g(i)}, \gamma_{g(i)})$  cross-validation group?

The main problem identified by Tibshirani and Efron with the first method - "there will typically be too few cases to fit the model" - does not apply to likelihood combination as likelihoods are defined for single observations - i.e. even "leave one out cross-validation" is feasible. The pre-validation technique, being an incorrect combination, should not be expected to have good properties. Tibshirani and Efron did determine that pre-validation did not provide the correct degrees of freedom, but did not report on the performance of the unweighted average of the cross-validations.

Pairs of outcomes - Neyman and Scott meat-analysis examples.

Interestingly, meta-analysis problems from astronomy (Neyman & Scott)[10] originally drew attention to the challenge of dealing with common and non-common parameters via parametric likelihood with a relatively small number of observations per non-common parameter. In particular they looked at pairs of observations. It may be important to keep in mind that meta-analyses in other areas seldom, if ever, face the same degree of challenge as was faced in astronomy where a large number of very small studies were encountered, but the problems are still instructive. The most important lesson is perhaps that it is the number of studies that can be foreseen, not just the number in hand, that needs to be considered when evaluating methods of summarizing for future analysis and eventually undertaking that final analysis. Real problems with likelihood as a general estimation approach were perhaps first encountered (or at least written about) by Neyman & Scott[10] (Stigler 2006 suggests that Wald's

correspondence to Neyman in 1938 lead to the Neyman & Scott paper, i.e. Neyman waited until he had an applied problem so that Fisher could not dismiss the issue as being only theoretical).

In two of the three problems Neyman & Scott addressed, there were repeated studies that all had two observations. In the first problem the mean was considered common and the variance non-common, and in the second the variance common and the mean non-common. For both, they assumed the observations were Normally distributed. In the first, the likelihood-based estimate is consistent but its asymptotic variance is not minimum (where the asymptotics fixes the number of observations per study and allows the number of studies to go to infinity), while in the second, the likelihood-based estimate is not even consistent. Various approaches have been offered to address the second situation but no approach is yet fully satisfactory? for the first (which is also known as the Fisher-Berhans problem for common mean).

According to Barndorff-Nielsen and Cox[3], essentially the approaches to salvage the likelihood separate into two, one is to find an exact or approximate factorization of the likelihood so that one factor contains all or most of the information about the common parameter, sometimes utilizing conditional or marginal probability models

and the second replaces the specification of arbitrary non-commonness of the non-common parameter with a common distribution for that parameter. A common parameter then resides in the marginal (over the non-common parameters) level 2 distribution and difficulties presented by having to (separately) estimate the non-common parameters disappear.

## Example 1 - common mean, arbitrary variance

Quoting from Neyman and Scott "Let  $\alpha$  be some physical constant such as the radial velocity of a star or the velocity of light. Assume that  $s$  series of measurements are to be made and let  $y_{ij}$  stand for the result of the  $j$  th measurement of the  $i$  th series ( $i = 1, 2, \dots, s; j = 1, 2, \dots, n_i$ ). We will assume that the measurements follow the normal law with the same mean  $\alpha$  and an unknown standard error  $\sigma_i$  which may and probably does vary from one series of observations to another. Thus the probability density function of  $y_{ij}$  is

$$f(y_{ij}; \alpha, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(y_{ij} - \alpha)^2 / 2\sigma_i^2}$$

This is exactly the case when  $\alpha$  stands for the radial velocity of a star and the  $y_{ij}$  are its measurements obtained from  $n_i$  different spectral lines

on the  $i$  th plate. ... This is also the situation in all cases where it is desired to combine measurements of physical quantities, made in different laboratories, by different experimenters, etc."

The log-likelihood in general is

$$-\frac{1}{2} \sum_i n_i \log(\sigma_i^2) - \frac{1}{2} \sum_i \sum_j (y_{ij} - \alpha)^2 / \sigma_i^2$$

and the score function (differentiation of above with respect to  $\alpha$ )

$$U_\alpha = \sum_i \sum_j (y_{ij} - \alpha) / \sigma_i^2 = \sum_i n_i \bar{y}_i / \sigma_i^2 - \alpha \sum_i n_i / \sigma_i^2.$$

As is well known with  $\sigma_i^2$  replaced by  $\hat{\sigma}_{i\hat{\alpha}}^2$  (the maximum likelihood estimate of  $\sigma_i^2$  for  $\hat{\alpha}$  not  $\hat{\sigma}_{i\alpha}^2$  - see below) the expectation of  $U_\alpha$  is zero, so the estimated likelihood of  $\alpha$ ;  $L_p(\alpha; y_{ij}) = c(y_{ij}) f(y_{ij}; \alpha, \hat{\sigma}_{i\hat{\alpha}}^2)$  is consistent for  $\alpha$ .

But if the  $\sigma_i^2$  were known,  $\sigma_i^2 = \sigma_{i0}^2$ , say, the inverse variance weighted mean

$$\frac{\sum_i \bar{y}_i. (\sigma_{i0}^2/n_i)^{-1}}{\sum_i (\sigma_{i0}^2/n_i)^{-1}}$$

would be normally distributed with mean  $\alpha$  and variance  $1/\sum_i (\sigma_{i0}^2/n_i)^{-1}$ . Now the estimated likelihood MLE for  $\alpha$  is

$$\hat{\alpha} = \frac{\sum_i \bar{y}_i. (\hat{\sigma}_{i\hat{\alpha}}^2/n_i)^{-1}}{\sum_i (\hat{\sigma}_{i\hat{\alpha}}^2/n_i)^{-1}}$$

where

$$\hat{\sigma}_{i\hat{\alpha}}^2 = \sum_j \frac{(y_{ij} - \hat{\alpha})^2}{n_i} = \frac{\sum_j (y_{ij} - \bar{y}_i.)^2 + n_i(\bar{y}_i. - \hat{\alpha})^2}{n_i}$$

has (Barndorff-Nielsen and Cox[3]) asymptotic (for fixed  $n_i$  and  $s \rightarrow \infty$ ) normal distribution with mean  $\alpha$  and variance

$$\frac{\sum_i n_i / \{(n_i - 2)\sigma_i^2\}}{(\sum_i n_i / \sigma_i^2)^2}$$

exceeding that of  $1/\sum_i (\sigma_{i0}^2/n_i)^{-1}$  with  $\sigma_i^2$  known. Additionally, different weights can result in smaller asymp-

otic variances with  $\sigma_i^2$  unknown but it is unclear as to the best estimator for all  $\sigma_i^2$ .

Example 2 - common variance, arbitrary mean

This is the same set up as example 1, but now the precision of measurements does not change from one series to another yet the quantity measured does.

$$f(y_{ij}; \alpha_i, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_{ij}-\alpha_i)^2/2\sigma^2}.$$

With  $n_i = 2$  for all  $i$  the log-likelihood is

$$-\frac{1}{2}2s \log(\sigma^2) - \sum_i^s \left\{ \frac{(y_{i1} - \alpha_i)^2 + (y_{i2} - \alpha_i)^2}{2\sigma^2} \right\}$$

and the score function (differentiation of above with respect to  $\sigma^2$ )

$$U_t = -\frac{2s}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_i^s \left\{ (y_{i1} - \alpha_i)^2 + (y_{i2} - \alpha_i)^2 \right\}.$$

As is well known with  $\alpha_i$  replaced by  $\hat{\alpha}_{i\sigma}$  (the maximum likelihood estimate of  $\alpha_i$  for a given  $\sigma$ ) the expectation of  $U_t$  is not zero but  $-s/2\sigma^2$  so the profile likelihood of  $\sigma^2$ ,  $L_p(\sigma; y_{ij}) = c(y_{ij})f(y_{ij}; \hat{\alpha}_{i\sigma}, \sigma)$ , is not consistent for  $\sigma^2$  (note that here  $\hat{\alpha}_{i\sigma}$  does not in fact depend on  $\sigma$  so that it is also the estimated likelihood here).

Example 1 recast - common mean, common distribution of variance

With assumptions that  $\sigma_i^2$  are independently inverse gamma distributed as

$$p_0(\sigma^2) = \left(\frac{1}{2}d_0\sigma_0^{2'}\right)^{\frac{1}{2}d_0}(\sigma^2)^{-\frac{1}{2}d_0-1}e\left(-\frac{1}{2}d_0\sigma_0^{2'}/\sigma^2\right) / \Gamma\left(\frac{1}{2}d_0\right)$$

where  $d_0$  is an effective degrees of freedom and the "prior" mean is  $\sigma_0^{2'}d_0/(d_0 - 2)$ . For one sample of size  $r$  the likelihood would be

$$\int_0^\infty \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}r}} e\left(-\sum \frac{(y_i-\alpha)^2}{2\sigma^2}\right) p_0(\sigma^2) d\sigma^2.$$

The full log-likelihood is

$$\frac{1}{2} \sum_i (r_i + d_0) \log \left\{ 1 + \frac{r_i (\bar{y}_i - \alpha)^2}{(y_{ij} - \bar{y}_i)^2 + d_0 \sigma_0^2} \right\}.$$

Note here that there are now just two parameters  $\alpha$  and  $\sigma_0^2$  for all the observations.

Example 2 recast - common variance, common distribution of mean

With assumptions that  $\alpha_i$  are independently normally distributed with mean  $\nu$  and variance  $\omega$ . The pairs  $(y_{i1}, y_{i2})^T$  are now independently bivariate normal with mean  $(\nu, \nu)^T$  and covariance matrix

$$\begin{bmatrix} \omega + \sigma^2 & \omega \\ \omega & \omega + \sigma^2 \end{bmatrix}$$

It follows either via the bivariate normal form or by integrating the joint density with respect to the  $\alpha_i$  that the log-likelihood is

$$-\frac{1}{2} s \log(\omega + \frac{1}{2} \sigma^2) - \frac{\sum_i (\bar{y}_i - \nu)^2}{2\omega + \sigma^2} - \frac{1}{2} s \log(2\sigma^2) - \frac{\sum_i (y_{i2} - y_{i1})^2}{4\sigma^2}$$

The maximum likelihood estimate of  $\sigma^2$  is  $\frac{1}{2} \sum_i (y_{i2} - y_{i1})^2 / s$  unless

$$\frac{1}{2} \sum_i (y_{i2} - y_{i1})^2 / s > 2 \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 / (s - 1)$$

then it is

$$\left( \frac{1}{2} \sum_i (y_{i2} - y_{i1})^2 + 2 \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \right) / (2s - 1).$$

The complication arising because the parameter space is  $\nu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ , so that if

$$\frac{1}{2} \sum_i (y_{i2} - y_{i1})^2 / s > 2 \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 / (s - 1)$$

then the maximum likelihood is achieved on the boundary  $\omega = 0$ . Except for this complication, the usual properties of likelihood-based procedures hold[3] and note that only the parameters  $\omega, \sigma^2$  and  $\nu$  are involved.

In terms of combination of observations, the common means and arbitrary variance problem arises in that the correct combination for the mean, on its own, is not quite known because it depends on the variance and this is not well estimated. Given the assumptions of Normality if one knew the relative variances, this could be fixed by multiplying the individual likelihoods by the ratio of variances and then combining by multiplication.

On the other hand, treating the non-common variances as random variables changes the probability specification to one where this problem no longer remains – combination for the common mean is simply by the multiplication of the marginal (over the unknown variances) likelihoods that involve just common parameters. Of course, the mis-specification of the random distribution of these non-common variances raises additional if not more serious problems - the form of the distribution can be a quite problematic nuisance parameter both in Classical and Bayesian approaches.

In terms of combination of observations, the common variance and arbitrary means problem arises in that in the full likelihood, the likelihood component for the variance depends on the unknown value of the means – if the values of the means were known the correct combination would simply be accomplished by the multiplication of the likelihoods given the true means. But by considering just the marginal observations of pair differences, there is only one marginal (over the sample) likelihood that does not depend on the mean (and there is no loss of information) and a combination of it is immediate. Alternatively, treating the means as random variables changes the specification to one where this problem no longer remains – combination is simply by the multiplication of the marginal (over the unknown mean) likelihoods.

Likelihood for multiple data sources a new problem - being stuck with reported summaries, perhaps even informatively reported and censored summaries. Even if unselectively reported, they will not be the sufficient statistics for all the probability model one might want to consider and their joint marginal probability distributions (which you need to get the likelihood) are just not available except for special cases such as order statistics

$$\left\{ \{ \min, \max, n \}, p(\min)[P(\max) - P(\min)]^{n-2}p(\max) \right\}$$

Recall though, as reported order statistics are almost never sufficient, other reported outcomes such as a p\_value would in principle need to be included.

Monte-Carlo methods to (try to) get observed summary likelihoods

First we need to look closely at marginal likelihoods for summaries rather than outcomes. Suppose we have a probability model  $(\Psi, A, P)$ , where  $\Psi$  is the sample space,  $A$  the class of events expressed as subsets, and  $P$  the probability measure. Now suppose we observe only events based on a reported function  $S(y) = s$ . Let  $\Omega$  be the sample space for  $S(y) = s$ ; then  $S$  is a function that maps  $\Psi$  into  $\Omega$ . We assume in general that  $\Psi$  is a partition so that  $P(y_i \cup y_j) = P(y_i) \cup P(y_j)$  for all  $i, j$   $i \neq j$  and  $S$  is a measurable function.

Consider the assertion that the function  $S$  has the value  $s_0$ ; this is equivalent to the assertion that the original sample point is in the set

$$S^{-1}(s_0) = \{y : S(y) = s_0\}$$

called the preimage of  $s_0$ . The probability that attaches to the set  $s_0$  on the new space  $\Omega$  must be the probability of the preimage set  $S^{-1}(s_0)$  on the space  $\Psi$ . Hence, the marginal likelihood - the probability of observing  $S(y) = s_0$  - is given by the probability of

$$S^{-1}(s_0) = \sum_{\{y:S(y)=s_0\}} P(y.)$$

. So then, letting

$$y^* = \{y : S(y) = s_0\},$$

the marginal likelihood is defined as

$$c(s_0) \int_{y^*} \text{Pr}(y; \theta) dy$$

where  $\text{Pr}(y; \theta)$  is the probability and  $y$  the possible individual observations. The  $\int$  is really always  $\sum$  (recall the likelihood is the probability of an *observed* event) but

depending on the probability model  $\Pr(y; \theta)$  for the outcomes  $y$ ,  $\int$  may be a very convenient approximation. See also formula 2 of Copas and Eguchi along with examples and additional technical reference[5]

Thus to estimate probability of observing the reported summary  $s_0$ , we could pick a probability specification and a particular point in the parameter space and then calculate the percent of the draws where the summary equaled  $S(y) = s_0$ . Alternatively, if  $y^*$  can be uniformly sampled from,  $\sum_{y^*} \Pr(y; \theta)$  will provide an estimate[12]. In general, one can do this by uniformly sampling  $y$  unconditionally and just keeping those in  $y^*$ . So for each point in the parameter space, we need to do rejection sampling to get an estimate of the marginal likelihood at that point...

A formula given by Barndorf-Nielsen [1] for the analytical derivation of marginal likelihoods, suggested the calculation of a likelihood surface would be possible given conditional samples were drawn from an "opportunistically" chosen single point in the parameter space. The formula was given (in different notation) as

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int \frac{f_X(x | \theta)}{f_X(x | \theta_0)} f_{X|U}(x | u, \theta_0) dx$$

Now, the marginal distribution is simply

$$f_U(u | \theta) = \int_{x^*} f_X(x | \theta) dx \text{ where } x^* \text{ is the level set given by}$$

(or more formally  $x \in \{\mathbf{x} : U(\mathbf{x}) = u\}$ ) but only the (relative) likelihood  $\frac{f_U(u|\theta)}{f_U(u|\theta_0)}$  is needed. Now

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int_{x^*} f_X(x | \theta) dx \frac{1}{f_U(u | \theta_0)}$$

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int_{x^*} \frac{f_X(x | \theta)}{f_X(x | \theta_0)} \frac{f_X(x | \theta_0)}{f_U(u | \theta_0)} dx$$

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int \frac{f_X(x | \theta)}{f_X(x | \theta_0)} \frac{f_X(x | \theta_0)}{f_U(u | \theta_0)} f_{U|X}(u | x, \theta_0) dx$$

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int \frac{f_X(x | \theta)}{f_X(x | \theta_0)} f_{X|U}(x | u, \theta_0) dx$$

Should be familiar to everyone?

Alternatively, starting out as importance sampling - Green, Geyer/Thompson, Griffith's formula

$$f_U(u | \theta) = \int \frac{f_X(x | \theta)}{f_{X|u}(x | u, \theta_0)} f_{X|u}(x | u, \theta_0) dx$$

$$\frac{f_U(u | \theta)}{f_U(u | \theta_0)} = \int \frac{f_X(x | \theta)}{f_X(x | \theta_0)} f_{X|u}(x | u, \theta_0) dx$$

Conditional samples can simply generated by rejection sampling.

Formula can be naturally split into model fit and likelihood (Normal example with  $\min=1$ ,  $\text{median}=4$ ,  $\max=125$ ). In order to acquire the needed conditional samples within reasonable computing time, an opportunistic value in the probability models parameter space is "guesstimated" or searched for and then unconditional samples, each of size  $n$ , are drawn (using only these parameter values). Only those that are within a given tolerance are kept to get the conditional sample (rejection sampling). This can be done in subsets (randomly or perhaps based on varying tolerances) to give some sense of the accuracies being obtained. Usually tens of thousands of samples are drawn and rejected to meet chosen tolerances but extremely high rejection rates (in the millions) suggests either poorly chosen values in the parameter space or probability model/reported summary conflict.

Sensitivity analyses for possibly informative choice of reported summary

Dawid's[7] approach is presented here for possibly informative choices of a reported summary as distinct from the approach in this thesis which assumed the summary would always be chosen. Using Dawid's notation, let  $\{S_n(y)\}$  be a finite set of possible summaries, one of which is reported. Recall, it has been shown that the observed summary likelihood - if that summary is always reported is

$$c(\overline{S_n(y)}) * \int_{y:S_n(y)=\overline{S_n(y)}} f(y|\theta) dy \text{ where } \overline{S_n(y)} \text{ is the value reported}$$

The alternative formulation of the likelihood-based on the distribution of the summary reported and the fact that the summary was chosen to be reported on the basis of  $\overline{S_n(y)}$  and  $\theta$  is.

$$p(R_n|S_n(y)) = \frac{c f'(\overline{S_n(y)}, \theta) * c(\overline{S_n(y)}) * \int_{y:S_n(y)=\overline{S_n(y)}} f(y|\theta) dy}{c(\overline{S_n(y)}) * \int_{y:S_n(y)=\overline{S_n(y)}} f(y|\theta) dy}$$

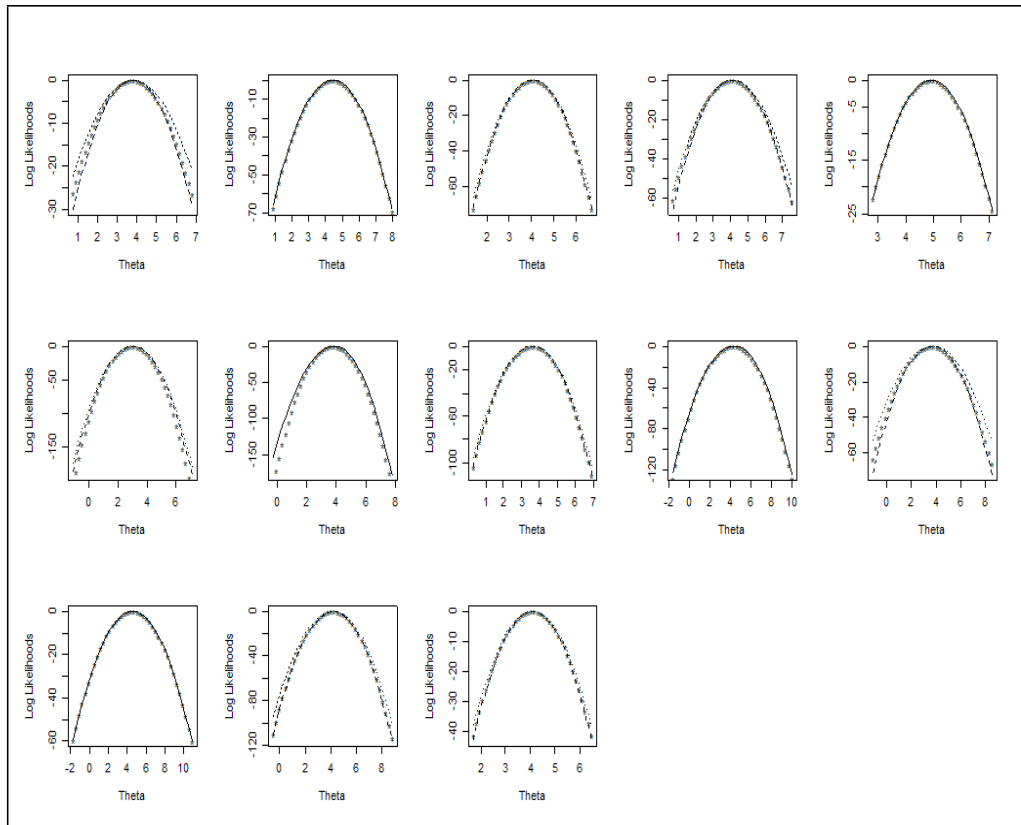


Figure 1: Simulated versus exact observed summary log-likelihoods for 13 studies that reported minimum, median and maximum. Simulated = \* , Exact = solid, dotted or dashed line (odd, even +1, even - 1)

There might be non-data influences of  $\theta$  on the choice of what is reported. Non-data influences on the choice of reported summary could arise easily from data from previous trials. These would require careful thought - including the possibility of interaction between within and outside trial data.

## Computational strategies and tactics

Ideally, one would wish to have a few procedures that would facilitate the meta-analysis of differing examples with the setting of options or minor re-programing, starting with a simple inverse weighted approach i.e. general least square (Guass-Markov), and then crawling to various likelihood based and full probability based (Bayesian) analyses.

A list structure was used to represent the reported group summaries such as

$$g. = \{mean, sd, n\} \text{ or } g. = \{\min, mean, \max, n\}$$

within the  $m$  multiple studies. For single group studies, the list would be

$$\{g_1, \dots, g_m\}$$

, for two group studies

$$\{\{g_{11}, \dots, g_{1m}\}, \{g_{21}, \dots, g_{2m}\}\}$$

and more generally for  $k$  group studies

$$\{\{g_{11}, \dots, g_{1m}\}, \dots, \{g_{k1}, \dots, g_{km}\}\}.$$

Next, a list of probability models, one for each of these same groups, was defined. For single groups studies, a simple specification could be

$$\{\Pr[\mu, \sigma_1], \dots, \Pr[\mu, \sigma_m]\}$$

with  $\Pr$  being for instance the *Normal* distribution. More generally, the specification could vary much more from study to study as in

$$\left\{ \Pr_1[\mu, \sigma_1], \dots, \Pr_m[\mu, \nu_m, \gamma] \right\}.$$

Random effects probability specifications would be nested as in

$$\{\Pr[\mu_1 \sim \Pr[0, \sigma_b], \sigma_1], \dots, \Pr[\mu_m \sim \Pr[0, \sigma_b], \sigma_m]\}$$

though one may more easily start with

$$\{\Pr[\mu_1, \sigma_1], \dots, \Pr[\mu_m, \sigma_m]\}$$

and leave the specification regarding the commonness in distribution of  $\mu$  to a later step. Initially, any default or canonical parameterization will suffice.

For instance, for two group randomized studies, a common (starting) specification would be

$$\left\{ \begin{array}{l} \{\Pr[\mu_{11}, \sigma_1], \dots, \Pr[\mu_{1m}, \sigma_m]\}, \dots, \\ \{\Pr[\mu_{21}, \sigma_1], \dots, \Pr[\mu_{2m}, \sigma_m]\} \end{array} \right\}$$

, representing arbitrary control means, arbitrary treatment means and arbitrary but common within study standard deviations.

A list of marginal likelihood approximations for all these groups can then be generated from these two lists, as appropriate. For reported group summaries that are sufficient for their associated probability models, a single sample of the appropriate size  $n$  with exactly the same sufficient summaries will provide a fully accurate "recreation" of the original data likelihood - i.e.  $\{y_1, \dots, y_n\}$ . For special closed form marginal likelihoods, a sublist of necessary probability specifications to provide the marginal likelihood directly from the reported summaries can be created - i.e. for reported minimums and maximums

$$\left\{ \{\min, \max, n\}, p(\min)[P(\max) - P(\min)]^{n-2}p(\max) \right\}$$

. Finally for the general case, a sample of  $k$  samples of size  $n$  are generated that have approximately the same summaries as the reported summaries and the importance sampling observed summary likelihood approximation formula is applied to these, i.e.

$$\{\{y_1, \dots, y_n\}_1, \dots, \{y_1, \dots, y_n\}_k\},$$

to get the approximate observed summary likelihood. In a given meta-analysis a resulting list such as the following can result, illustrating all three cases of sufficiency, closed form observed summary likelihood and non-sufficient summary

$$\begin{aligned} & \{\{y_1, \dots, y_n\}, \dots, \\ & \left\{ \{\min, \max, n\}, p(\min)[P(\max) - P(\min)]^{n-2}p(\max) \right\}, \dots \\ & \{\{y_1, \dots, y_n\}_1, \dots, \{y_1, \dots, y_n\}_k\} \end{aligned}$$

Reparameterizations are then required to highlight the arbitrariness, commonness or commonness in distribution of the various parameters amongst all the groups. A rewrite of the list of probability models is perhaps most convenient for this. For instance, for the two group randomized study one such rewrite could be

$$\left\{ \begin{array}{l} \{\Pr[\mu_{11}, \sigma_1], \dots, \Pr[\mu_{1m}, \sigma_m]\}, \dots, \\ \{\Pr[\mu_{21}, \sigma_1], \dots, \Pr[\mu_{2m}, \sigma_m]\} \end{array} \right\}$$

↕

$$\left\{ \begin{array}{l} \{\Pr[\mu_1, \sigma_1], \dots, \Pr[\mu_m, \sigma_m]\}, \dots, \\ \{\Pr[\mu_1 + \delta, \sigma_1], \dots, \Pr[\mu_m + \delta, \sigma_m]\} \end{array} \right\}$$

for fixed effect and

$$\left\{ \begin{array}{l} \{\Pr[\mu_1, \sigma_1], \dots, \Pr[\mu_m, \sigma_m]\}, \dots, \\ \left\{ \begin{array}{l} \int \Pr[\mu_1 + \delta_1 \sim \Pr[0, \sigma_b], \sigma_1] d\delta_1, \dots, \\ \int \Pr[\mu_m + \delta_m \sim \Pr[0, \sigma_b], \sigma_m] d\delta_m \end{array} \right\} \end{array} \right\}$$

for random treatment effects (note only the treatment group has a common in distribution parameter).

At this point various simple techniques for getting guesstimated study treatment effect estimates and their variances (for instance mean = report median and variance = (max - mean)/6 ) may be a useful starting point. Then optimizations (or integrations) will then need to be carried out to get profile (or integrated) likelihoods to focus on a common parameter of interest. The needed optimization is succinctly given as another rewriting of the list of probability models. For instance, for focussing on  $\delta$

$$\left( \begin{array}{c} \left\{ \Pr[\hat{\mu}_{1(\delta)}, \hat{\sigma}_{1(\delta)}], \dots, \Pr[\hat{\mu}_{m(\delta)}, \hat{\sigma}_{m(\delta)}] \right\}, \dots, \\ \left\{ \Pr[\hat{\mu}_{1(\delta)} + \delta, \hat{\sigma}_{1(\delta)}], \dots, \Pr[\hat{\mu}_{m(\delta)} + \delta, \hat{\sigma}_{m(\delta)}] \right\} \end{array} \right)$$

for fixed effect and similarly for random effects. Fortunately, these optimizations can sometimes be factorized by study. In general, a meta-analysis likelihood from  $m$  studies is given as

$$\prod_i^m L(\gamma(\theta), \gamma(\lambda), \chi(\lambda)_i; y_i)$$

with  $\theta$  representing the interest parameters,  $\lambda$  the nuisance parameters and  $\gamma(\cdot)$ ,  $\chi(\cdot)_i$  isolating the common

and non-common parameters. Here there is no parameter based factorization and the full likelihood must be used.

If however there are no common nuisance parameters

$$\prod_i^m L(\gamma(\theta), \chi(\lambda)_i; y_i)$$

the profile likelihood value for a given  $\gamma(\theta)'$  becomes factorized as

$$\sup_{\chi(\lambda)_i \in \Omega} \prod_i^m L(\chi(\lambda)_i; y_i, \gamma(\theta)')$$

and as long as the  $\chi(\lambda)_i$  are variation independent components (i.e.  $\chi(\lambda)_i \in \Omega_i$  and  $\Omega_1 \times \Omega_2 \times \dots \times \Omega_n = \Omega$ ), the profile likelihoods can be obtained separately since

$$\begin{aligned} & \sup_{\chi(\lambda)_i \in \Omega} \prod_i^m L(\chi(\lambda)_i; y_i, \gamma(\theta)') \\ &= \prod_i^m \sup_{\chi(\lambda)_i \in \Omega_i} L(\chi(\lambda)_i; y_i, \gamma(\theta)'). \end{aligned}$$

This simplifies the required numerical optimization considerably. Unfortunately, the random effects meta-analysis likelihoods have common nuisance parameters, and this simplification is not available - these optimizations need to be carried out jointly over studies or the information lost by ignoring common elements is somehow argued to be unimportant[6].

It may always be useful to start with fixed effect assumptions where there are no common nuisance parameters. This might also provide good starting values for the joint optimizations required for the random effects model. Once the desired likelihood based meta-analysis formulation has been successfully plotted in a KOR plot, one may well wish to add priors.

Combination of observations: A parametric likelihood approach

A brief summary of the review of the parametric likelihood-based approach to statistics, conceptualized as the investigation and synthesis of individual observations, is as follows:

1. A descriptively appealing and transparent definition of likelihood is “the probability of re-observing exactly what was observed” under a given probability model – in notation  $L(\theta; \textit{observed}) = f(\textit{observed}; \theta)$  considered as a function of  $\theta$  an  $n$ -dimensional vector of reals for fixed *observed*, where *observed* can never really be a continuous number but instead some interval and, more often than not, in meta-analysis is a reported summary rather than actual individual observations.
2. Likelihoods from different observations multiply (after appropriate conditioning if observations are dependent).

3. If there is more than one likelihood and something is common in these likelihoods - i.e. some  $\gamma(\theta)$  is repeated in the likelihoods - the multiplication of them provides a combination for that (the  $\gamma(\theta)$ ), and under the probability model, that multiplication provides the “best” combination. A common parameter reparameterization  $(\gamma, \chi_i)$  may help make this more apparent.
4. It does not matter if likelihoods are based on  $n = 1$  (a single observation) or  $n = k$  (a single study) for 1, 2 & 3 and the order of multiplication also does not matter.
5. There usually is something common -  $\gamma(\theta)$  and something non-common  $\chi(\theta)_i$  in the probability model entertained -

$$f(y_i; \gamma(\theta), \chi(\theta)_i)$$

The something common may be a particular parameter or a distribution of a particular parameter (often referred to as a random effects model) and that commonness will be with regard to a particular transformation  $\gamma$ . Perhaps a good example of this would be a non-common treatment effect parameter transformable into a common direction parameter and a non-common magnitude parameter.

6. When it is the distribution of a parameter that is common, the multiplication referred to above for combining applies only to the marginal or expected likelihood with respect to the distribution of  $\gamma_i^* \sim p(\gamma)$  i.e.  $\prod_i E_\gamma L(\gamma_i^*; y_i)$  as  $\prod_i L(\gamma_i^*; y_i)$  does not provide a combination.
7. In meta-analysis, usually only a summary of  $y$  is available, say  $s(y)$ , so the required likelihood is  $L(\theta; s(y))$  i.e. a marginal likelihood with respect to the distribution of individual observations, which is “forced” upon the meta-analyst.

8. Naming conventions are problematic here, marginal in 7 reflecting unobserved (missing) observations and in 6, an unobserved random parameter. We suggest the first marginal likelihood is also called the observed data likelihood and for data only available as reported summaries, observed summary likelihood. For the marginal over the unobserved random effects likelihood, in meta-analysis, there would usually be only a level 1 and level 2 distribution. Following this, the level 1 likelihood is conditional on the value of unobserved random effect while the level 2 likelihood only involves parameters of the common distribution (from which the unobserved random effect was drawn), obtained by integrating over the unobserved random effect.
  
9. As the integrals in 6 and 7 will not be tractable in general, numerical methods will be required. For 6, the dimension is usually equal to one (i.e. to avoid Simpson's paradox) and numerical integration methods may suffice. For 7, the dimension is usually high

(number of observations in  $s(y)_o$ ) but importance sampling techniques may suffice. The EM algorithm and other systematic approximations might also be considered when the exact marginal distribution is not available.. Assuming estimated parameters as known, such as with-in study variances, should be carefully considered (i.e. check what happens when it is treated as unknown specifically does the combined Log likelihood become multi-modal?)

10. In general, it is not known how to get exact confidence intervals from  $\prod_i L(\theta; y_i)$  - in particular for a common mean with arbitrary variances, e.g. where the combined likelihood is equal to  $\prod_i L(\mu, \sigma_i; y_i)$ , they are known not to exist even under assumptions of Normality, (see page 77 of Spratt)[13]. Alternatively, while it is known how to get credible intervals, in general it is unknown as to how to (conveniently) get the meaningful priors required[8], especially if  $\theta$  is of high dimension, or how the shape of these intervals should be chosen [personal conversation, Mike

Evans]. However if  $\log \prod_i L(\theta; y_i)$  is approximately quadratic in the region of its maximum, at least from a practical point of view, confidence intervals and regions based on the likelihood ratio (using first order results relating to the likelihood ratio being distributed approximately as a chi-square random variable [2]) and credible intervals using non-informative or reference priors are non-controversially obtainable and are usually quite similar.

11. It may also be useful to think of the various values of the nuisance parameters as generating various likelihoods for the parameter(s) of interest, and the issue being again the investigation and synthesis of what is possibly common in these various likelihood functions which differ for unknown values of the nuisance parameters. One could think of the unknown nuisance parameters as being like unknown sources of measurement error that caused observations of something common to differ from each other. Early

astronomers debated about using un-weighted versus weighted averages as well as other methods of combining the differing observations. Integrated likelihoods  $\int L(\theta_i; y_i) d\theta_i$  are an obvious "un-weighted" combination of likelihoods over nuisance parameters[4]. "Parameter" marginal likelihoods in random effects approaches weight likelihoods by the probability of the unobserved random nuisance parameter -i.e.

$$E_{\theta} L(\theta_i^*, y_i)$$

or

$$\int L(\theta_i^*; y_i) \Pr(\theta_i^*; \theta) d\theta_i^*$$

. Profile likelihood could also be viewed as a particular combination of differing likelihoods - where the combination is to choose from the possible values of the nuisance parameters those that are most likely ("best") for each value of the common parameter.

This parametric likelihood-based approach was to some extent anticipated in the “likelihood menu” sketched out in O’Rourke[11] and the strategy was originally suggested in the appendix of L’Abbe, Detsky and O’Rourke[9]. One can greatly extends its scope and generality to arbitrary summary statistics and general probability models as well as general random effects models (and priors).

## References

- [1] Barndorff-Nielsen, O. *Information and exponential families: In statistical theory*. Chichester, New York, 1978.
- [2] Barndorff-Nielsen, O. Likelihood theory. In *Statistical Theory and Modelling*, D. V. Hinkley, N. Reid, and E. J. Snell, Eds. Chapman and Hall, London, 1990.

- [3] Barndorff-Nielsen, O., and Cox, D. R. *Inference and asymptotics*. Chapman and Hall, London, 1994.
  
- [4] Berger, J. O., Liseo, B., and Wolpert, R. L. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* 14 (1999), 1–28.
  
- [5] Copas, J., and Eguchi, S. Local model uncertainty and incomplete-data bias. *Journal of the Royal Statistical Society B* 67, 4 (2005), 459–513.
  
- [6] Cox, D. R. *Principles of statistical inference*. Cambridge University Press, Cambridge, 2006.
  
- [7] Dawid, A. P., and Dickey, J. M. Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association* 72 (1977), 845–850.

- [8] Gelman, A. The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions. Tech. rep., Columbia University, 2006. To appear in *The American Statistician*.
  
- [9] L'Abbe, K. A., Detsky, A. S., and O'Rourke, K. Meta-analysis in clinical research. *Ann.Intern.Med.* 107, 2 (1987), 224–233.
  
- [10] Neyman, J., and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica* 16 (1948), 1–32.
  
- [11] O'Rourke, K. Meta-analysis: Conceptual issues of addressing apparent failure of individual study replication or “inexplicable” heterogeneity. In *Empirical Bayes and likelihood inference* (2001), pp. 161–183.
  
- [12] Ross, S. *Simulation*. Academic Press, 2002.

- [13] Sprott, D. A. *Statistical inference in science*. Springer-Verlag, New York, 2000.
- [14] Tibshirani RJ, Efron, B. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* 1, 1 (2002), 1–18.