strength of the evidence against a null hypothesis. The fail-safe sample size can likewise be considered as measuring the weight of evidence. Thus, $n(0)$ is the number of *hypothetical* studies, conducted under the null hypothesis, that need to be added to the database to just offset a significant result. Here "offset" is used in the sense that the updated test statistic should have its expected value, *conditional on the published studies*, just equal to the critical point.

The authors' equation (4) modifies the foregoing in two respects: (i) the hypothetical studies are replaced by actual and unpublished studies and (ii) a weight function is made available for modeling the publication bias. Now we have difficulty interpreting the numerator of (4) as a conditional expectation. Granted the values reported in the published studies are noninformative for the unpublished studies, but the number k of published studies, when coupled with the weight function, is informative and should be reflected in the conditional expectation. Thus, does $n(\alpha)$ in equation (4) refer to actual studies (in file drawers) or to hypothetical studies?

As an alternative, we suggest using the weight function to estimate the number of unpublished studies. For example, letting k and k_0 be the number of published and unpublished studies and $N = k + k_0$,

$$k_0 \approx N \int (1 - \omega(x)) f(x)\, dx,$$

$$k \approx N \int \omega(x) f(x)\, dx,$$

and k_0 can be estimated as the solution of

$$\frac{k_0}{k} = \frac{E[1 - \omega(X)]}{E[\omega(x)]}. \tag{3}$$

Note that the righthand side of (3) reduces to a $(1 - \alpha)/\alpha$ in case of a dichotomous weight function. Also, $\omega(x)$ must have a phenomenological interpretation as a probability. It will not do to replace ω by a scalar multiple or to use an unbounded ω.

The solution of (3) is sensitive to the weight function, and one may prefer to look upon the estimate of k_0 as shedding light upon the reasonableness of the chosen ω instead of upon the "true" significance of the published results.

ADDITIONAL REFERENCES

GLYNN, R. J., LAIRD, N. M. and RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.). Springer, New York.

HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **41** 517–529.

LITTLE, R. J. A. (1983). The nonignorable case. In *Incomplete Data in Sample Surveys* (W. G. Madow, I. Olkin and D. B. Rubin, eds.). Academic, New York.

PETO, R. (1987). Why do we need systematic overviews of randomized trials? *Statist. Med.* **6** 233–240.

RUBIN, D. B. (1977). Formalizing subjective notions about the effect of nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **72** 538–543.

SIMES, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statist. Med.* **6** 11–29.

Comment

M. J. Bayarri

"Meta-analysis, like rock and roll, is here to stay" claim the authors of this interesting and stimulating paper, and they are right. Similar experiments are conducted and replicated, providing informaiton about the same unknown quantity, and statisticians have to face the challenge of providing methods for pooling this information. In a sense, the problem is similar to that of combining a set of expert opinions. Unfortunately, results from experiments are not, in general, expressed as distributions of the unknown quantity. If they were, then not only would the publication bias due to statistical significance be greatly reduced, but also the techniques for combining probability distributions would be available (for an excellent summary and comprehensive annotated bibliography about these techniques see Genest and Zidek, 1986). Moreover, meta-analysis is usually based on results that got published in the scientific literature. Due to the overabuse of hypothesis testing as a statistical methodology and to the overappreciation of statistical significance, it does not come as a surprise that publications are highly biased toward studies showing statistically significant results. This publication bias should be taken into account when carrying out a

M. J. Bayarri is Titular Professor at the University of Valencia (Spain). Her mailing address is Departamento de Estadística e IO, Facultad de Matemáticas, Av. Dr. Moliner 50, 46100 Burjasot, Valencia, Spain.

meta-analysis. The authors present an elegant solution that explicitly models the selection bias when building the model governing the statistical behavior of the observables. Also, their general formulation contemplates the use of different weight functions for each study to be included in the meta-analysis, thus allowing the consideration of different publication policies.

In spite of my general agreement with the approach taken by the authors, I don't quite understand why the fail-safe sample size (FSSS) approach and the maximum likelihood (ML) approach are treated as if they were two different complementary or exclusive approaches to dealing with the publication bias effect. As a matter of fact, both approaches are just trying to draw conclusions based on two aspects that are *always* present in a meta-analysis, namely what is observed (the published studies) and what is not observed (the unpublished ones).

The FSSS approach proposes to analyze reported results as if they were generated by the underlying density $f(x \mid \theta)$ to decide whether or not data are significant at a certain level α (0.05 for instance). If data are declared significant, then it proposes to use the subjective opinion that the reader of that conclusion might have about the number of unpublished, nonsignificant results in a rather limited way, namely whether or not this number is believed to be smaller than the FSSS. If it is, then the original conclusion would still be the same. Notice though that even if the conclusion (significance) remains unchanged, this should no longer be true with other characteristics, such as p-values, of the statistical analysis. Indeed, for a given set of data, the larger the reader thinks the number of unpublished null results is, the larger the overall p-value should be. This part is usually forgotten in the FSSS approach: it merely takes into account the influence of the unobserved, unpublished results on the decision about significance or nonsignificance of the observed data, but not on the general statistical analysis of the data.

On the other hand, the ML approach taken by the authors in Section 3 somehow does the opposite. That is, it carefully takes into consideration the modified behavior of the observed data due to selection bias, but it does not provide explicit information about the number of unobserved studies, which may or may not be of interest and may or may not influence the conclusions of the analysis. In other words, once we have agreed on the weight function to be used in a given problem (the model for the *observables*) it looks like we don't have to worry about what we don't get to see. When a Bayesian approach is used, it is of course true that, in some problems, the explicit consideration of the (unobserved) number of performed experiments has no effect on the statistical analysis of the observed data. There are problems, however, in which the opinions about this unobserved number have a dramatic influence on the conclusions of the analysis, as we will shortly show. Moreover, this influence cannot be avoided by using a ML approach. On the contrary, the mere consideration of this unobservable can change a ML analysis of the problem (Bayarri, DeGroot and Kadane, 1987).

My main point is that, when analyzing the results from published experiments that are suspected to be subject to statistical selection bias, the analysis should take into account *both* the effect that the selection has in the model for the observables and the influence that the number of unobservable, unpublished studies might have in the final conclusions of this analysis.

To illustrate this point, I will consider a simple situation in the framework of a one-sided test of hypotheses on the mean of a normal distribution with known variance. Assume that independent experiments are carried out by the same or different experimenters around the world. In each of them a random sample of size n is taken from the normal distribution with unknown mean μ and known variance σ^2 and the uniformly most powerful test is used for testing

$$(1) \qquad H_0: \mu \leq 0 \quad \text{vs.} \quad H_1: \mu > 0$$

at some level α. In this case, the distribution of the test statistic $T = (n^{1/2}\bar{X}_n)/\sigma$ is normal with mean $\theta = (n^{1/2}\mu)/\sigma$ and variance 1, where $\bar{X}_n$ represents, as usual, the sample mean of a given experiment. The restriction of equal sample sizes n and variances σ^2 is, of course, quite unrealistic and it is used here just to ease the presentation.

Assume that the results of one such experiment appear in some scientific journal, declaring the data significant at the level $\alpha = 0.05$ and yielding a p-value of $p = 0.033$. (Here the number k of studies in the meta-analysis is taken to be 1 just for simplicity. In the simple scenario that we are considering, an analogous argument would apply with any other value of k.) How should these significant results be interpreted? If we use the FSSS approach, even Rosenthal's FSSS is strictly less than one. What should be concluded from the published experiment is just not clear, apart from the caution that "the finding is not resistant to the file drawer threat" (Rosenthal, 1984, page 108). It does not seem possible to draw further conclusions or make inferences about θ. Furthermore, even a reported p-value as small as 0.0001 would yield a FSSS smaller than 4. Thus, under a FSSS approach a single published study will generally be regarded as highly unreliable, with no indications about what can be concluded from it.

A ML approach, which explicitly includes the selection bias, seems more appropriate. More information is then needed in order to assess the weight function. We will assume that we have read about this experiment in a journal that we know, with certainty, only publishes experiments yielding statistically significant results at the level $\alpha = 0.05$. In this case, we will only observe values of T greater than or equal to 1.645 so that the density of any value of T that we will actually get to observe is given by the selection model

$$(2) \quad g(t \mid \theta) = \frac{\phi(t - \theta)}{1 - \Phi(1.645 - \theta)} \quad \text{for} \quad t \geq 1.645,$$

and $g(t \mid \theta) = 0$ otherwise, where ϕ and Φ denote the standard normal pdf and df, respectively. Here, the weight function is known and it is simply given by the indicator function of the set $[1.645, \infty)$. From (2) it can be seen that the MLE of θ is the unique solution to

$$(3) \qquad (t - \theta)M(1.645 - \theta) = 1,$$

where $M(\cdot)$ stands for Mills' ratio. In our example, with a reported p-value of 0.033 (so that $t = 1.844$) the solution to (3) is $\theta = -3$, an estimate which is at least three standard deviations away from the parameter values in H_1. Thus, the significant results are in fact providing strong support to the null hypothesis, and the MLE of θ would no longer be 1.844 as it would have appeared in that published experiment, but rather -3 as derived from (3). This is an improvement over the FSSS approach because it does allow a modified statistical analysis of the reported data. Notice though that, in order to reach this conclusion, we have implicitly assumed that the experiment has been performed repeatedly (not necessarily by the same experimenter, but possibly by different experimenters working on the same problem) until one significant report was obtained and got published. The unknown number N of performed experiments does not provide further information about θ and it should not change our inferences about θ even if we explicitly introduce it into the analysis in order to learn about it. We will turn to this point later on.

The situation would be completely different if we knew that only one experiment with the characteristics of the published one had been performed in the entire world. In this case all we learn when reading about it in the journal is that it turned out to yield significant results and thus got published. Then the reported statistical analysis should be accepted at face value, including the p-value and the MLE. Notice that, in this situation, the value of $k (k = 1)$ itself carries information about θ. Indeed, if we did not observe this experiment to get published then we would have known that it had turned out to be nonsignificant.

A general Bayesian approach to the problem of analyzing our significant result could proceed as follows. For simplicity of notation we will use the symbol p to denote an arbitrary generalized density, without any implication that it is the same for all variables whose distribution it is representing. Also, as the number k of published studies can provide information about θ, it will be regarded as an observable random variable and explicitly introduced in the notation. Our observation then is the pair (t, k) where $t = 1.844$ and $k = 1$. Under the Bayesian approach, a joint posterior density $p(N, \theta \mid t, k)$ is obtained, which can be represented in the following convenient way:

$$(4) \quad p(N, \theta \mid t, k) \propto p(t \mid k, N, \theta)p(k \mid N, \theta)p(N \mid \theta)p(\theta).$$

Here N and θ are unobservables, so that $p(N \mid \theta)p(\theta) \equiv p(N, \theta)$ represents the joint prior density of N and θ. In our example, $p(t \mid k = 1, N, \theta) = p(t \mid k = 1, \theta)$ is given by the selection model (2), but we will alter the notation a little bit (in order to mimic the one in the paper) and represent by $f(\cdot \mid \theta)$ and $F(\cdot \mid \theta)$ the pdf and df, respectively, of the normal distribution with mean θ and variance 1. Thus, $g(\cdot \mid \theta)$ represents the selection model and $f(\cdot \mid \theta)$ the underlying density. Also, for convenience, the known value 1.645 will be represented by τ. Then, (4) becomes

$$(5) \quad p(N, \theta \mid t, k) \propto \frac{f(t \mid \theta)}{1 - F(\tau \mid \theta)} p(k \mid N, \theta)p(N \mid \theta)p(\theta).$$

The two situations above in which data should be analyzed according to $g(t \mid \theta) = f(t \mid \theta)/[1 - F(\tau \mid \theta)]$ in the first one and according to $f(t \mid \theta)$ in the second one, are just particular cases of (5). In the first case, when the experiment was repeatedly performed until one significant report was obtained, $p(k = 1 \mid N, \theta) = 1$ (irrespective of N) in (5) so that the data are analyzed according to $g(t \mid \theta)$. Notice that, independent of whether or not we are also interested in N, the marginal posterior density of θ would be

$$(6) \qquad p(\theta \mid t, k) \propto g(t \mid \theta)p(\theta),$$

for every $p(N \mid \theta)$. Thus, posterior inferences about θ would be the same independently of whether or not we explicitly introduce N in our analysis. This result is to be expected because, in this case, the unobservable N does not provide any information about θ, although we could learn about it if we wished. This is no longer the case with a ML approach. Indeed, $N - 1$ being the number of unpublished results, it seems natural to assume that it has a geometric

distribution with parameter $P = Pr(T \geq \tau \mid \theta)$ so that

$$p(N \mid \theta) = [1 - F(\tau \mid \theta)][F(\tau \mid \theta)]^{N-1} \quad \text{for} \quad N = 1, 2, \ldots. \tag{7}$$

If the MLEs of N and θ are now derived from a likelihood function of the form $g(t \mid \theta)p(N \mid \theta)$ then the MLE of θ will not be the same as the one derived from $g(t \mid \theta)$. Thus, just contemplating the possibility of learning about this uninformative, unobserved N will change our estimate of θ.

In the second case, when we knew that just one experiment was performed, we have $p(N = 1 \mid \theta) = 1$ and

$$p(k = 1 \mid N = 1, \theta) = Pr(T \geq \tau \mid N = 1, \theta) = 1 - F(\tau \mid \theta).$$

Then

$$p(\theta \mid t, k) \propto f(t \mid \theta)p(\theta) \tag{8}$$

so that we analyze the data (t, k) using the original underlying density $f(t \mid \theta)$.

Intermediate situations between these two can occur, and knowledge about N more vague than that considered above can be incorporated in a natural way into the analysis. The example presented here is just a particular case in which two selection mechanisms (geometric and Bernoulli) have been considered to generate $k = 1$ published studies out of N performed ones. In Bayarri and DeGroot (1987) more general selection mechanisms are considered, as well as conditions under which the selection mechanism can be ignored in the analysis of the data, either because it does not provide additional information about θ or because, even if it does, the particular form of the prior distribution makes it ignorable when making inferences about θ.

To conclude this comment, I would like to stress my personal opinion that meta-analysis is one of the areas in statistics that really calls for a Bayesian analysis. As we have seen, conclusions from a meta-analysis rely very heavily on the prior information; even the assessment of the weight function can be highly subjective. All these subjectivities must be incorporated in an explicit form into the analysis. In this way, different readers can judge whether or not the different components of the analysis agree with their own particular beliefs on the subject and, if not, reach their own particular conclusions.

ADDITIONAL REFERENCES

BAYARRI, M. J. and DEGROOT, M. H. (1987). Selection models and selection mechanisms. Technical Report 410, Dept. Statistics, Carnegie Mellon University.

BAYARRI, M. J., DEGROOT, M. H. and KADANE, J. B. (1987). What is the likelihood function? In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 3–27. Springer, New York.

GENEST, C. and ZIDEK, J. V. (1986). Combining probability distributions: A critique and annotated bibliography (with discussion). *Statist. Sci.* **1** 114–148.

Comment

C. Radhakrishna Rao

Meta-analysis is an important area of research and any contribution to its methodology is welcome. I am glad to see that Iyengar and Greenhouse extended the scope of meta-analysis by modeling selection bias using simple classes of weight functions that cover a variety of situations. However, some caution is necessary in pooling information from different sources. Often the parameter under estimation like θ in the example of Table 4 may not be the same in all studies. So modeling must take into account variations in θ also. In that case one must specify what exactly is being estimated by meta-analysis. If θ is considered as a variable, it would be of interest to estimate its mean value and variance. The anomalies noted by the authors in the estimation of θ can be explained by the possibility that θ is not the same in all the studies.

Perhaps a preliminary test for homogeneity of different studies with respect to the parameters of underlying distributions is one way of approaching the problem. Of course, in constructing such a test, one must take into account selection bias. If the test reveals inhomogeneity, then other problems arise, such as comparison of estimates between studies and possible explanation of the differences. A more satisfactory method may be to introduce a prior distribution on θ; the problem in such a case is the estimation of the prior distribution of θ, which provides all the information.

C. Radhakrishna Rao is University Professor at the University of Pittsburgh and National Professor in India. His mailing address is Mathematical Statistics Department, Thackeray Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.