The Wish-lists: Some Comments

D.R. Cox and N. Reid
Nuffield College, Oxford and University of Toronto, Canada

Abstract
We provide brief comments on some common threads arising from the 'wish-lists' set out in some of the other papers in this volume. The discussion is necessarily incomplete: in particular we have dealt only with points for which a reasonably compact answer seems possible.

1 Introduction

The wish-lists are wide-ranging and raise issues of varying difficulties, ranging up to the seemingly impossible. The following comments concern just some of the topics raised.

2 Combination of independent sets of data

In the simplest situation there are m independent sets of data from each of which a common parameter θ can be estimated, representing for example some constant of interest. Separate analyses of the individual sets give estimates $t_1, \ldots, t_m$ with uncorrelated estimates of the variances $s_1^2, \ldots, s_m^2$. For an initial discussion ignore errors in the s_j^2.

If there are no additional sources of variation and the studies do indeed estimate the same unknown, there is the implicit model

$$t_j = \theta + \epsilon_j,$$

where the ϵ_j are independent Gaussian errors of zero mean and variances estimated by s_j^2. The parameter θ is estimated by weighted least squares or equivalently by ordinary least squares applied to a modified version of the model, namely

$$t_j/s_j = \theta/s_j + \epsilon_j/s_j,$$

where now the errors have unit variance. The estimate is

$$\tilde{\theta} = \frac{\Sigma t_j/s_j^2}{\Sigma 1/s_j^2},$$

with

$$\mathrm{var}(\tilde{\theta}) = \frac{1}{\Sigma 1/s_j^2}.$$

Importantly also the residual sum of squares from the modified model, namely

$$\Sigma(t_j/s_j - \tilde{\theta}/s_j)^2 = \Sigma t_j^2/s_j^2 - \tilde{\theta}^2 \Sigma 1/s_j^2,$$

has under the model a chi-squared distribution with $m-1$ degrees of freedom.

The argument can be refined, essentially by an empirical Bayes approach, to allow for errors in estimating the variances. The main practical point is that there can be major drawbacks to giving relatively high weight to individual estimates that have very small values of s_j^2 arising by chance.

Suppose now that the chi-squared test indicates clear heterogeneity, that is, the t_j vary too much. There are a number of possible explanations:

– the internal estimates of variance are unrealistically small

- there may be a small number of anomalous values
- there may be characteristics of the different studies which if entered into a regression equation for the t_j account for the additional variation.

If none of these is applicable and provided m is not too small, for example is at least, say, 10 it may be reasonable to suppose that there is an additional source of random error producing inter-study variation and to replace the starting model by

$$t_j = \theta + \eta_j + \epsilon_j,$$

where the ϵ_j are as before and the η_j are independent random variables of zero mean and unknown variance σ_η^2.

If in fact that variance were known, the least squares estimate of θ is

$$\frac{\Sigma t_j/(s_j^2 + \sigma_\eta^2)}{\Sigma 1/(s_j^2 + \sigma_\eta^2)}$$

a value intermediate between the simple weighted mean $\tilde{\theta}$ and the unweighted mean $\Sigma t_j/m.$ The component of variance σ_η^2 can be estimated by maximum likelihood or, slightly less efficiently, by equating a sum of squares to its expectation. The assumptions involved in this formulation are quite strong and estimation of σ_η^2 is fragile if m is small. When the estimates are based on Poisson-distributed counts, the variances are functions of the counts themselves and this involves some changes in any more refined formulae. The additive representation for additional variation may be better replaced by a multiplicative form.

These issues are treated in more depth in [2].

3 Comparison of fit of a small number of models

Suppose first there are just two models neither of which is nested within the other. Two broad approaches might be considered. There are formidable practical difficulties in most situations with a Bayesian discussion, not so much connected with specifying the prior probabilities of the two models as with the conditional densities of the (different) parameters within each model. Unless these priors can be specified at least approximately on external evidence there is difficulty in computing the posterior probabilities required for model assessment.

A frequentist approach is to test model 1 for departures in the direction of model 2 and compute a p-value. Then switch the roles of the two models. There results information about whether both models give an adequate fit in the respect tested, whether one but not the other fits or whether neither model is adequate. In the last case, further analysis to develop an improved model would normally be required. Note that such a possibility cannot be directly obtained from the formal Bayesian approach.

With three or more models the best procedure is usually to test model 1, say, in turn against model 2 and then model 3 and to take the smaller p-value adjusted for selection as an assessment of model 1, and so on.

4 Systematics

Most statistical analysis focuses on random errors, it being assumed that the impact of systematic errors has been eliminated by design, that is by arranging that the effects of interest are estimated by comparisons of groups of data equally affected by systematic errors. There is also a substantial literature on estimating sources of variability in complex measurement systems intended, in particular, to aid the standardization of measurement techniques.

In the present context these methods are largely not applicable and explicit consideration of systematic errors seems unavoidable. A common approach seems to be to estimate the effect of an estimated physical constant on the final result of interest by re-computing this final result with the physical constant changed by plus and minus one standard error. Half the difference between these two resulting values is then approximately the derivative of that quantity with respect to the physical constant. This could be combined with other estimated sources of error in a propogation of errors formula, but it is essential to note that the errors in estimating the constant must be independent of errors from other sources. A less formal approach would be to investigate the sensitivity to the result of interest to errors in the physical constant by re-computing the results over a range of plausible values for the constant.

If there are k sources of systematic error and these can be given bounds, taken without loss of generality as say $(-\Delta_j, \Delta_j)$ for $j = 1, \ldots, k$, a very cautious approach is to do 2^k possible analyses based on the set of extreme possibilities, each with its confidence limits for the effect of interest and to take the union of these intervals as the basis of inference. Assumptions that the Δ_j are random variables may be reasonable but the key issue will often concern the independence assumptions involved which may have very strong implications. Ref. [3] has given a careful account of these issues from a Bayesian perspective.

Systematic errors that are essentially nuisance parameters in a model that is fully specified, or even partially specified, can be eliminated from the full likelihood by either maximizing over them or by integrating over them, with respect to a weight function. The integration approach is emphasized in [4]. . The maximization approach results in a profile likelihood, discussed in [5], and is implemented as MINOS in MINUIT. The limiting distribution of statistics based on the profile likelihood is the same as that for a simple likelihood. However the approximations given by this limiting theory, such as the χ^2 approximation to twice the log-(profile) likelihood ratio, can be quite inaccurate, especially if there are large numbers of nuisance parameters. Several adjustments to profile likelihood have been suggested in the statistical literature (see for example [1], Chs. 2 and 3), to take account of the uncertainty in estimating the nuisance parameters. These adjustments are implicit in the weight function applied in the integration approach, although the weight function is best thought of as a prior distribution on the nuisance parameters. The choice of the prior is important, and a large body of evidence now indicates that flat priors on the nuisance parameters are not appropriate, and can lead to very poorly calibrated inference, especially if there are large numbers of nuisance parameters. In some applications it may be possible to construct an empirical prior distribution from previous observations or from simulations.

5 Comparison of alternative test statistics

Tests are conventionally assessed by the power curve. In the simplest case of testing a null hypothesis H_0 that a single parameter θ is equal to θ_0 against alternatives $\theta > \theta_0$, the power curve shows the probability that the test "rejects" H_0 at level α as a function of θ. Equivalently the power curve shows the probability of a p-value less than α versus θ. If correctly calibrated the curve should pass through (θ_0, α). It is often a good idea to plot $\Phi^{-1}(\text{power})$ against θ, where $\Phi(x) = \int_{\infty}^{x} \{1/\sqrt(2\pi)\} \exp(-y^2/2) dy$ is the standard Gaussian distribution function. This produces a series of roughly parallel curves, or even approximately lines, for different α. In comparing two tests the steeper the curves the better.

More mathematically for test statistics that are approximately normally distributed we may define the efficacy of a test T as

$$\{ \left. \frac{\partial E(T;\theta)}{\partial \theta} \right|_{\theta=\theta_0} \}^2 / \text{var}(T;\theta_0).$$

This measures the sensitivity of the expectation of T near the null hypothesis relative to the variance.

For two test statistics T_1, T_2 of the same null hypothesis the ratio of their efficacies is the asymptotic relative efficiency (ARE) of T_1 relative to T_2. Because efficacy usually scales as sample size, the ARE compares the sample sizes needed to achieve the same power with the two tests. Thus for testing the mean of a Gaussian distribution the ARE of the median relative to the mean is $2/\pi$ so that tests based on the median of n observations and on the mean of $0.63n$ observations have about the same power.

These ideas may be useful even if the properties of the tests are studied primarily by simulation.

6 p-values and limits

The CL_s or CL_{s+b} methods combine size and power in a very *ad hoc* way and are unlikely to have satisfactory statistical properties. As is emphasized in Neal [4], upper and lower one-sided confidence limits should replace confidence intervals, and a full plot of the log-likelihood function is better still. A related point is that the construction of a p-value for discoveries, i.e. for confirming the existence of a particular effect, should be treated as a separate problem from the establishment of limits on the magnitude of a well-established effect. When there are several parameters of interest, a decision is needed about whether they can be assessed separately, treating the other parameters as nuisance parameters for each of these assessments, or whether it is physically more relevant to consider two (or more) of the parameters as forming a single vector. In the latter case approximate p-values can be computed using the usual asymptotic theory of likelihood, or a more refined version, but the construction of confidence regions is considerably more difficult and often not very illuminating.

References

[1] A.R. Brazzale, A.C.Davison, N. Reid, *Applied Asymptotics.*, Cambridge University Press, 2007.

[2] D.R. Cox, *Encyclopedia of Statistical Science* N.L. Johnson and S.Kotz, eds. New York: Wiley **2**, 45-53, 1982.

[3] S. Greenland, *Journal of the Royal Statistical Society: Series A* **168**, 267-306, 2005.

[4] R. Neal, in this volume.

[5] N. Reid and D.A.S. Fraser, in *Proceedings of PHYSTAT2003*, L. Lyons, R. Mount, R. Reitmeyer, eds. SLAC e-Conf C030908, 265–271, 2003.