FULL TITLE
*ASP Conference Series, Vol. **VOLUME**, **YEAR OF PUBLICATION***
NAMES OF EDITORS

Statistics Perspective

James Berger

SAMSI and Duke University, Institute of Statistics and Decision Sciences, Durham, NC, 27708-0251, USA

Abstract. A wealth of fascinating statistical issues were raised at the conference and giving justice to them in a short comment is not possible. Hence, rather than attempting a systematic review, I will simply mention some of the issues that I personally found most interesting.

1. The SAMSI Astrostatistics Program

In Spring, 2007, the Statistical and Applied Mathematical Sciences Institute (SAMSI), held a research program in Astrostatistics. The program was in collaboration with **CASt**; indeed Jogesh Babu was the Program Leader at SAMSI during the spring. As a number of the presentations at the conference were outgrowths of the SAMSI program, it is worthwhile to review what occurred.

Tutorials were held from January 18-22, 2006, covering

- *Bayesian Astrostatistics:* This three-day session - taught by Tom Loredo, Bill Jefferys and Philip Gregory - covered the basic theory and practice of Bayesian statistics, using examples from astronomy.

- *Nonparametric statistics and machine learning:* This two-day tutorial, taught by Chad Schafer and Larry Wasserman, introduced astronomers to modern methods in nonparametric statistics.

- *Astronomy for statisticians:* This two-day tutorial, taught by Bill Jefferys and Eric Feigelson, reviewed modern understanding of our universe and raised key statistical issues underlying astronomical studies.

The tutorials are an excellent introduction to these topics, and can be viewed through the web at www.samsi.info/programs/2005astroprogram.shtml.

Working Groups and Intensive Study Sessions involved research, throughout the spring, by small teams of individuals. The groups and their relationships to this conference are as follows:

- *Exoplanets:* The results of this working group were discussed in the presentations of Eric Ford, Merlise Clyde and Bill Jefferys.

- *Surveys and Population Studies:* The results of this working group were discussed in the presentations of Tom Loredo and Woncheol Jang.

- *Source Detection and Feature Detection:* Aspects of the research of this working group were discussed in the presentations of Alanna Connors, David van Dyk, and Rebecca Willet.
- *Gravitational Lensing:* This working group, led by Arlie Petters, studied magnification of probability distributions with applications to dark matter substructures on galactic scales; statistical methods in cluster mass reconstruction; statistics of image counting in stochastic microlensing; and applications of spatial statistics to the distribution of dark matter structures.
- *Intensive Session on Statistical Issues in Particle Physics:* The results of this intensive study session were partly discussed in the presentations of Louis Lyons and Michael Woodroofe.
- *Intensive Session on Stellar Evolution:* This intensive study session, led by Bill Jefferys, studied improving MCMC sampling, handling field stars, and (very successfully) handling binary stars.

Workshops occurred before during (and possibly after) the program:

- *Planning meeting:* This was held at NASA-Ames from July 14-15, 2005, organized by Jeff Scargle, and set much of the research agenda for the SAMSI program. Some important themes that came up in that planning meeting and which were *not* pursued in the SAMSI program were:
  - How do we summarize data so that we can best use the conclusions for future scientific investigations, given that the raw data will not likely be available; some viewed this as an argument for summarizing the data via a full posterior distribution of all scientific unknowns (or perhaps a posterior sample).
  - How does one deal with the interface between complex computer models and data interface, including identifying and adjusting for systematic errors?
- *Opening Workshop:* This was held at SAMSI from January 23-25, 2006, with the primary purpose of forming and advising the working groups.
- *Closing Workshop:* The current workshop.
- *Reprise/Reunion Workshops:* These are possible over the next 2 years at SAMSI, if warranted by follow-up developments to program research.

2. The Current Status of Astrostatistics

During the SAMSI program and during this workshop, I was struck by how many astronomers and astrophysicists are doing excellent statistical work and even developing wonderful new statistical methodology. There seems to be no question that the need for sophisticated statistical analysis has taken hold among at least a significant segment of the astronomical community. Also, the opportunities for interaction between astronomers and statisticians have never been greater. Thus *astrostatistics* itself is clearly flourishing.

The involvement on the statistics side, however, is not nearly as dramatic. While the number of statisticians involved with astronomers and astrophysicists is definitely increasing, the rate of increase seems to be quite modest. What is perhaps lacking is a regular structure for forming a 'team environment,' where statisticians are involved in major astronomical projects from the beginning. Barriers include:

- Funding mechanisms in astronomy: little funding for statistics; typically none for statisticians.
- The shortage of statisticians (and 'easy funding' for them elsewhere).
- The fact that astronomers can often 'figure it out for themselves' and are used to trying to do so.

3. A Few Commonly Occurring Science Themes

In this section, several scientific themes that seemed to recur in the presentations at the conference are highlighted

3.1. 'Doing it right' versus 'facing up to reality'

There is always a tension between principled analysis and what has to be done in a practical messy situation to obtain answers. Both sides of the debate are right: one typically must produce an answer yet, if the analysis varies too far from being principled, the answers are not trustworthy. Thus Gary Hinshaw made the anecdotal comment that "half of all 3 sigma results seem to be wrong."

Alanna Connors and David van Dyk argued for the principled approach of including all physical/statistical models of structure at the beginning of the analysis, and carrying these structures through the entire data accumulation and processing stages. Tom Loredo called this "forward analysis" – you put in the models at the beginning, allowing for a principled likelihood analysis of the data at the end.

Proponents of 'facing up to reality' included Tim Axelrod, who called it "swimming upstream," the suggestion being that one is often limited to modeling locally at each step of the process, hoping one arrives safely at the end. While humorously calling this "sleazy analysis," Robert Lupton argued strongly that there is often no other way to proceed.

My own approach to problems is always to at least conceptualize the forward process, since that is the only real chance for getting the uncertainties right, and reduces the chance of bias creeping into the analysis. But, after the conceptualization in a complex problem, one must decide where to make 'sleazy approximations' to achieve progress. Note that use of more sophisticated statistics can often reduce the amount of sleaze that must be used.

This issue arose several times in discussion of Bayesian/MCMC methods. These were generally viewed as being principled efforts to account for all uncertainties in an analysis, but as rarely being implementable in complex situations (e.g., Gary Bernstein's weak gravitational lensing example).

While not being dogmatic about this, my view is that MCMC can be made efficient enough (with work) for use in more complex situations than one at

first might believe, and that there are a variety of approximations (e.g. the Laplace approximation mentioned by Jiayang Sun) that can help. 'Fiddles' are fair if absolutely needed (e.g., Christopher Kochanek did forward Bayes, but needed to use maximum likelihood estimates of nuisance parameters in MCMC loops). Finally, note that new very fast approximations to Bayesian answers are being developed, such as the *variational methods* from the machine learning community that can work with extremely large problems.

3.2. Prior distributions in Bayesian analysis

A number of discussions involving Bayesian analysis centered on issues involving the choice of prior distribution. Harrison Prosper defended use of subjective priors for important parameters in problems, but agreed that they are tough to come by for many parameters, especially nuisance parameters. Harrison observed that flat priors for such nuisance parameters are often tenable, even though "statisticians say flat priors are evil."

A major part of the difficulty of being an astronomer or physicist, and trying to interact with statisticians, is that there are so many kinds of statisticians. Everyone knows about the Bayesians and non-Bayesians, but less is known about the 'subjective Bayesians' and 'objective Bayesians (and Harrison was presumably reading the former). Objective Bayesian analysis has a wonderful 240 year history (starting with Bayes and Laplace), and is arguably the dominant practical Bayesian philosophy today. In this school, use of variants of flat priors is shown to be optimal in various scenarios. Two of the most prominent objective Bayesian approaches are

- The maximum entropy approach (especially useful when you know moments of the system); see Jaynes (2003).

- The reference prior approach: see Bernardo (2005).
  Example: During the conference, the question arose as to what is the best objective prior for a Poisson mean λ. The answer, according to reference prior theory, is $\pi(\lambda) = \lambda^{-1/2}$.

Hypothesis testing and model selection problems raise different issues in choice of prior distributions. The usual complaint one hears about Bayesian testing is that one has to specify prior probabilities of hypotheses and that such prior beliefs are particularly worrisome since testing is used to 'prove' scientific discoveries. For this reason, many Bayesians prefer use of *Bayes factors* for testing, since these do not involve prior probabilities of hypotheses. This is illustrated in the next section in the context of a particular example.

3.3. Bayesian hypothesis testing

There were several talks that directly addressed hypothesis testing, and there seemed to be some confusion as to the differences between classical and Bayesian testing. Because of the importance of formal testing of scientific theories, it is worthwhile to clarify the distinction. There is a vast literature on this topic (accessible through Sellke, Bayarri and Berger, 2001), but here we give only a simple illustration.

Before conducting a hypothesis test, a Bayesian must first answer the crucial question: Is the hypothesis that is being tested plausible?

Example 1. Glen Cowan discussed physics tests of

$$H_0 : s = 0 \quad \text{versus} \quad H_1 : s > 0\,, \tag{1}$$

where s is the mean signal arising from, say, a new particle (e.g., the Higgs) that is being sought. Here $H_0 : s = 0$ is clearly plausible (e.g., there is no Higgs).

Example 2. Gary Hinshaw, in discussing WMAP, posed the question of whether the inflation parameter n_s is < 1. Is $n_s = 1$ a plausible hypothesis so that one is testing

$$H_0 : n_s = 1 \quad \text{versus} \quad H_1 : n_s \neq 1\,,$$

or does $n_s = 1$ not correspond to any believable scientific theory, in which case one is, say, testing

$$H_0 : n_s < 1 \quad \text{versus} \quad H_1 : n_s \geq 1\,?$$

Discussion during my presentation clarified that the former is the case, i.e., that $H_0 : n_s = 1$ is a plausible scientific hypothesis.

The reason that this issue is crucial is that, in situations such as the two above, a Bayesian will have positive prior probability of H_0, i.e., will be putting positive prior probability on $s = 0$ or $n_s = 1$ (or, if there is possible experimental bias, will be putting positive prior probability on some very small intervals of values around these points). In this situation, formal Bayesian and frequentist testing will give very different answers than will use of common $p-$values. Consider the following version of the physics test, for instance.

Example 1 (continued). An experiment will yield an event count n, viewed probabilistically as arising from the Poisson$(s + b)$ distribution, where s is the mean signal rate of events and b is the (known for simplicity) background mean rate of events. It is desired to conduct the test in (1).

For known b, the p-value for this test is given by

$$p = P(N \geq n \mid b, s = 0)\,,$$

where we let N stand for the Poisson random variable and n for the actual observed event count. For instance, $p = 0.0014$ if $n = 5$ and $b = 0.8$.

The Bayesian approach is based on assigning s under H_1 a prior density $\pi(s)$, and computing the Bayes factor of H_0 to H_1, which is given (when b is known) by

$$B_{01} = \frac{b^n\ e^{-b}}{\int_0^\infty (s+b)^n\ e^{-(s+b)}\pi(s)\ ds}\,.$$

This can be thought of as the ratio of the data-based likelihood of the theory that says $s = 0$ (e.g., no Higgs) to the theory that says $s > 0$ (e.g., the Higgs exists). Different scientists might have different prior beliefs about the plausibility of these theories and these beliefs can be incorporated (if desired) by the simple device of multiplying the Bayes factor by the ratio of a scientist's prior

probability of the two hypotheses, obtaining (as a consequence of Bayes theorem) the scientist's 'posterior odds' of the hypotheses being true. Note that the Bayes factor itself does not depend on these prior beliefs about the hypotheses, and is hence the typical vehicle used for scientific communication.

To complete the Bayesian analysis, one must choose the prior distribution $\pi(s)$ for s under H_1. This could arise scientifically (e.g., from the *standard model* predictions of the mass of the Higgs), or be chosen in a default fashion. In the latter case, considerable care must be taken; for instance, one cannot use traditional vague proper priors for s.

Interestingly, an absolute lower bound on the Bayes factor exists for this problem: simply choose $\pi(s)$ to be a point mass at $\hat{s}$ (the maximum likelihood estimate of s), resulting in the lower bound being the observed likelihood ratio, i.e.,

$$B_{01} \geq \min\left\{1, \left(\frac{b}{n}\right)^n e^{n-b}\right\}.$$

Thus $B_{01} \geq 0.007$ if $n = 5$ and $b = 0.8$. A more reasonable 'objective' lower bound can be found by restricting to $\pi(s)$ that are nonincreasing, in which case it is easy to see that

$$B_{01} \geq \frac{b^n\ e^{-b}}{\sup_c \int_0^c (s+b)^n\ e^{-(s+b)} c^{-1}\ ds}.$$

This results in $B_{01} \geq 0.011$ if $n = 5$ and $b = 0.8$.

The point here is that, when testing plausible hypotheses, $p-$values are considerably smaller than direct Bayesian measures of evidence; compare $p = .0014$ above with $B_{01} \geq 0.011$. If, instead $n = 5$ and $b = 0.5$, then $p = 0.00017$, compared with $B_{01} \geq 0.0015$. And remember that the Bayesian numbers given here are lower bounds; the actual Bayes factor resulting from using a scientifically based $\pi(s)$ could be much higher.

Determining these objective lower bounds on Bayes factors is quite challenging in general but, for near-Gaussian problems, a very simple expression can be given:

$$B_{01} \geq -e\ p\ \log p\,.$$

When $p = .0014$, this yields $B_{01} \geq 0.025$ while, when $p = 0.00017$, this yields $B_{01} \geq 0.004$.

As a final comment, nearly identical results follow if one uses the frequentist approach and attempts to convert $p-$values to an 'error probability' scale. (It is mistakenly thought by many that $p-$values are frequentist error measures, but they are far from it in situations such as these.) An easy access to this literature is Sellke, Bayarri and Berger (2001).

3.4. Model Selection

Talks by Eric Ford, Bill Jefferys, Merlise Clyde and Graham Woan considered problems in Bayesian model selection. There was some discussion as to the domain of applicability of Bayesian model selection. In principle, it can be applied to any problem; in practice, the problems of selecting prior distributions and computing marginal likelihoods can be daunting. I will comment on two issues that arose in this regard.

Approximations: Approximate procedures are often employed in model selection because of the difficulty of implementing exact methods. For instance, Chris Genovese discussed the difference between AIC (often useful if the goal is simply choosing the best model for prediction) and BIC (generally viewed as most useful when trying to ascertain the 'true model'). BIC is given by

$$\mathrm{BIC} \equiv 2l(\hat{\boldsymbol{\theta}}) - p \log n \,,$$

where $l(\boldsymbol{\theta})$ is the likelihood of the model, p is the dimension of $\boldsymbol{\theta}$, n is the sample size, and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate.

BIC arises as an attempt to approximate the Bayesian marginal likelihood of a model, $m(\mathbf{x}) = \int l(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ (where $\pi(\boldsymbol{\theta})$ is the prior distribution for the parameters in the model), in the sense that, as $n \to \infty$,

$$m(\mathbf{x}) \to c_\pi e^{BIC/2} \,,$$

for some constant c_π depending on the prior. In defining this approximation, however, there are several issues:

- n is often ill-defined (e.g., with binning of the data, as Gary Hinshaw noted).
- p is often ill-defined, especially when parameters can be random effects.
- p often grows with n, invalidating the 'large sample' justification of BIC.
- The constant c_π is often very far from 1.

There is considerable ongoing work seeking to develop versions of BIC that overcome these problems.

Another aspect of approximation was raised by Harrison Prosper, who wondered if Bernardo's discrepancy measure (cf. Bernardo, 2005) is useful as a tool in model selection. This measure was developed to deal with the situation in which one is trying to decide if a smaller model is adequate *as an approximation* to a 'true' larger model. For instance, Newtonian mechanics is usually perfectly adequate as an approximation to relativistic mechanics. Bernardo's discrepancy measure is not, however, designed for discovering a true model.

Multiplicities: The issue of dealing with multiplicities in testing and model selection received quite a bit of discussion. Multiple testing can arise in a number of ways. For instance, it was noted that as 'events' come in sequentially in many experiments (e.g., the search for the Higgs), it is tempting to perform a new test after each new observation. It is well known, however, that this cannot be done within classical statistics; there must be an 'adjustment' for conducting multiple tests in such a sequential setting. Interestingly, this is not so in Bayesian analysis; one can recompute the Bayes factor after each new observation, and stop and report the result when desired. This controversial issue is discussed in Berger and Berry (1988).

Another common situation of multiple testing in astrostatistics is when one is scanning many possible data sets for a 'discovery.' (Istvan Szapudi's CMB

situation involved looking at as many as 1.6 million possible data sets, seeking to identify which were compatible with non-Gaussianity of the spatial process.) Again, it is well known that one cannot proceed with separate testing of each data set, but the classical Bonferonni adjustment for multiple testing is often viewed as being too harsh. There was thus considerable discussion as to whether the relatively new False Discovery Rate methodology is a useful tool for dealing with such situations in astronomy.

Some comments that arose in the discussion:

- As mentioned by Chris Genovese, FDR is only useful for controlling FDR, not for establishing a 'discovery.' FDR only purports to control the percentage of false discoveries among the claimed discoveries, and hence is not useful if one seeks to conclusively establish individual discoveries.
- The main use of FDR is thus in screening situations, where one is attempting to ascertain a set of possible discoveries that will subsequently be investigated in detail, with new experiments. Note, however, that FDR seems to lack a decision-theoretic justification in terms of screening (i.e., FDR does not seem to be derivable from first principles in screening experimentation).
- Perhaps surprisingly, controlling the Family Wide Error Rate by standard methods usually gives greater power for conclusively establishing a discovery (as in Chris Koen's method for testing for a significant peak in periodograms).
- FDR is quite conservative, unless one uses more recent (and more difficult) variants that incorporate estimates of the proportion of discoveries that exist (as discussed by Chris Genovese, Laura Cayon, and John Rice).
- Using FDR is okay for correlated data, in the sense that FDR control will usually still be achieved, but it is typically too conservative for correlated data (as noted by Istvan Szapudi).

As a final comment on multiplicity, one of the highly attractive features of Bayesian approaches to model selection (as mentioned in Graham Woan's talk) is that they automatically adjust for multiplicities, and do so in a way that preserves as much discriminatory power as possible. The nature of the multiplicity has to be carefully modeled through the prior distribution (see, e.g., Scott and Berger, 2006), but that is the best way to ensure that maximal discriminatory power is preserved.

3.5. Nonparametric analysis

Eric Feigelson noted that astrophysics yields parametric models in many situations, which can result in pleasantly straightforward analysis. He went on to argue, however, that not enough attention is paid to the very many situations in which parametric models do not exist. Several different cases of this were discussed in the talks:

- Nonparametric function estimation: often this is estimating a nice function, about which much may be known, but not an exact functional form.

- Dealing with a mess which does not appear nice in any sense, like cosmic structures (discussed by Vicent Martinez and Ofer Lahav).
- Nonparametric testing (discussed by William Romanishin and John Rice).

Let's look at each in turn.

Nonparametrics for nice functions: Usual nonparametric analysis is developed so that it performs well when there is a huge amount of data and a very nasty function to estimate. Such methods can be quite inefficient, however, when a nice (though of unknown shape) function is being estimated, as can be argued to be the case in many astrophysical problems. Imposing 'niceness' constraints on the functions (arising from astrophysics) – such as monotonicity or convexity – can give much more efficient procedures, as discussed in Michael Woodroofe's talk and Martin Hendry's comments.

Nonparametrics for messes: Here the goal is typically to find meaningful structure in very complicated data. There are many statistical tools for this, such as the spatial tools discussed in Adrian Baddeley's talk, clustering tools, ... A useful tool that was not discussed in other talks is *Dirichlet process mixtures* for clustering, a method becoming highly popular in the machine learning community because of its computational feasibility with very large problems.

As a simple example of this methodology, start with standard *finite normal mixtures:* vectors $\boldsymbol{y}_i, i = 1, \ldots, n$, are modeled by a mixture of k multivariate normal distributions, $\boldsymbol{y}_i \sim \sum_{j=1}^{k} w_j N(\boldsymbol{\mu}_j, \Sigma_j)$, where the mixture weights w_j are assigned a Dirichlet$(\alpha, \ldots, \alpha)$ distribution. Letting $k \to \infty$ and $\alpha \to 0$ at the right rates, one ends up with the nonparametric Dirichlet process mixture, a very flexible clustering method with fast computational implementations.

Nonparametric testing: The interesting talk by John Rice on this problem noted the impossibility of constructing a test statistic that will have power to detect any discovery that is present; one needs to have a fairly precise idea of what one is looking for to find it in a complex situation. It is thus tempting to look to Bayesian analysis for guidance in constructing test statistics to 'look in particular directions,' even if one is not a Bayesian.

Acknowledgements. This work was supported by NSF Grant AST-0507481. The conference was wonderful, and thanks to Eric and Jogesh for making it happen.

References

Berger, J., & Berry, D. 1988, in Statistical Decision Theory and Related Topics IV., ed. S. Gupta & J. Berger (New York: Springer-Verlag)

Bernardo, J. 2005, Handbook of Statistics 25, ed. D. Dey & C.R. Rao, 17

Jaynes, E. T. 2003, Probability Theory: the Logic of Science (London, Cambridge Unviersity Press)

Sellke, T., Bayarri, M. J., & Berger, J. 2001, American Statistician, 55, 62

Scott, J., & Berger, J. 2006, J. Statistical Planning and Inference, 136, 2144