A Comparison of Testing Methodologies

James Berger
Duke University, Durham NC, USA

Abstract
This is a mostly philosophical discussion of approaches to statistical hypothesis testing, including p-values, classical frequentist testing, Bayesian testing, and conditional frequentist testing. We also briefly discuss the issue of multiplicity, an issue of increasing concern in discovery. The article concludes with some musings concerning what it means to be a frequentist.

1 Introduction

Because of the tradition in high-energy physics of requiring overwhelming evidence before stating a discovery, there has been limited attention paid to formal statistical testing. With the increasing cost of data, and issues involving simultaneous performance of a multitude of tests, there is likely to be an increasing interest in more formal testing. The main purpose of this article is to review the major approaches to testing, utilizing the basic high-energy physics problem as the vehicle for the discussion.

The following are some of the conclusions that will be argued:

- Tests are very different creatures than confidence intervals or confidence bounds, and it is often not correct to conclude an hypothesis is wrong because it lies outside a confidence interval.
- p-values are typically much smaller than actual error probabilities.
- Objective Bayesian and (good) frequentist error probabilities can agree, providing simultaneous frequentist performance with conditional Bayesian guarantees.

There will also be a brief discussion of multiplicity in testing in Section 3, highlighting the Bayesian approach to dealing with the problem. Section 4 contains some musings about the meaning of frequentism, motivated by presentations and discussions at the Phystat 07 conference.

2 Hypothesis testing

We review, and critically examine, p-values, classical frequentist testing, Bayesian testing and conditional frequentist testing. An ongoing example used in the discussion is a high-energy physics example described in the next section. For pedagogical reasons, a very stylized version of the problem will be considered here, ignoring most of the real physics.

2.1 The pedagogical testing problem and statistical model

Suppose the data, X, is the number of events observed in time T that are characteristic of Higgs boson production in an LHC particle collision experiment. The probabilistic model for the data is that X has density

$$\text{Poisson}(x \mid \theta + b) = \frac{(\theta + b)^x e^{-(\theta+b)}}{x!},$$

where θ is the mean rate of production of Higgs events in time T in the experiment and b is the (assumed known) mean rate of production of events from background sources in time T. Two specific values of X and b that we will follow through various analyses are

$$\textit{Case 1: } x = 7 \text{ and } b = 1.2; \qquad \textit{Case 2: } x = 6 \text{ and } b = 2.2.$$

The main purpose of the experiment is supposedly to determine whether or not the Higgs boson exists which, in terms of the probability model for the data, is typically phrased as testing $H_0 : \theta = 0$

versus $H_1 : \theta > 0$. Thus H_0 corresponds to 'no Higgs.' (Later we will discuss the issue of whether this statistical test is the correct representation of the desired scientific test.) There are many secondary issues that are of interest, such as "What is a lower confidence bound for the mass of the Higgs?" We will not discuss this issue in depth (noting that it has been the focus of many of the Phystat conferences), but will contrast the statistical analysis of the issue with the basic existence issue answered by the test.

2.2 Classical statistical analysis

There are two types of classical analysis: use of p-values, as recommended by Fisher [1], and use of fixed error probability tests, as recommended by Neyman [2].

2.2.1 p-values

The p-**value** in this example, corresponding to observed data x, is

$$p = P(X \geq x \mid b, \theta = 0) = \sum_{m=x}^{\infty} \text{Poisson}(m \mid 0 + b)\,.$$

This is the probability, under the null hypothesis, of observing data as or more extreme than the actual experimental data, and the tradition is to reject the null hypothesis if p is small enough. The part of the definition that may seem odd is the inclusion of *more extreme* data in the probability computation. Indeed, the oddity of doing so led to Jeffreys's [3] famous criticism of p-values "... a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." (It is worth spending the time to understand that sentence.) For the two cases,

Case 1: $p = 0.00025$ if $x = 7$ and $b = 1.2$; *Case 2:* $p = 0.025$ if $x = 6$ and $b = 2.2$.

There is general agreement that a small p-value indicates that something unusual has happened, but that the p-value does not have a direct quantitative interpretation as evidence against the null hypothesis. Thus Luc Demortier observed in his talk at the Phystat 07 conference:

> In any search for new physics, a small p-value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.

2.2.2 Fixed α-level testing

Under this approach, one pre-specifies the set of data for which one would reject the hypothesis – the *rejection region* – selecting the set so that the probability of rejection under the null hypothesis is the desired error probability α. Often, as in our example, one can formally state the rejection region in terms of the p-value, namely "reject if $p \leq \alpha$." Because X has a discrete distribution in our example, α should be limited to the possible values allowed by this discreteness; otherwise, one would have to artificially introduce some randomization which is unappealing. (That this rejection region indeed has probability α at the allowed values, follows from an easy computation.)

There are two major concerns with using fixed error probability testing. The first is that it does not properly seem to reflect the evidence in the data. For instance, suppose one pre-selected $\alpha = 0.001$. This then is the error one must report whether $p = 0.001$ or $p = 0.000001$, in spite of the fact that the latter would seem to provide much stronger evidence against the null hypothesis.

The second concern, as it applies to typical high-energy physics experiments, is more subtle: data naturally arrives, and is analyzed, sequentially and typical frequentist computations of fixed error probabilities must take this into account. For instance, suppose the experimental plan is to review the accumulated data at the end of each month, with there being a possibility of claiming a discovery at each

review. The rejection region is then a complicated set involving possible rejection at each of the time points (together with a lack of previous rejection); the frequentist error probability is the probability of this complicated rejection region and is typically much larger than the probability of the rejection region at a particular time. To achieve an error probability of $\alpha = 0.001$ for instance, the rejection region might have to be something such as "reject at each review if $p \leq 0.0001$", so that the frequent looks at the data require a higher standard of evidence to achieve the desired error probability. Note that p-values are affected by this same issue and in roughly the same way: much smaller p-values are needed in a sequential experiment to convey the same evidence as in a fixed sample size experiment.

Louis Lyons raised the interesting point that, with the LHC, declaration of a discovery would not stop the data gathering process, as is common in sequential experimentation in, say, clinical trials. (In clinical trials, claim of a discovery would ethically necessitate stopping the trial, in an attempt to save lives while, as Louis points out, no one we know of really cares if a few more particles are smashed.) So, in principle, a mistake made by this 'sequential look-elsewhere effect' could be corrected with later data.

In practice, however, declaration of a discovery often does have other effects – e.g., people stop research along lines that are incompatible with the discovery – so there is a serious cost to erroneous claims of discovery (in addition to having to return the Nobel prizes), even if there is a possibility of later correction. Also, we shall see that there are readily available reports (both Bayesian and frequentist) that can be made on an interim basis and which do not have difficulty with this sequential look-elsewhere effect, so the entire philosophical conundrum can be avoided.

2.3 Bayesian testing

2.3.1 Bayes factor

The **Bayes factor** of H_0 to H_1 in our ongoing example is given by

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid \theta + b)\pi(\theta)\, d\theta} = \frac{b^x\, e^{-b}}{\int_0^\infty (\theta + b)^x\, e^{-(\theta+b)}\pi(\theta)\, d\theta}\,;$$

in the *subjective Bayesian approach*, the prior density, $\pi(\theta)$, is chosen to reflect the beliefs of the investigators (e.g., it could reflect the standard model predictions pertaining to the Higgs) while, in the *objective Bayesian approach*, it is chosen conventionally and nominally reflects a lack of knowledge concerning θ.

A reasonable objective prior here (to be justified later, but note that it is a proper prior) is $\pi^I(\theta) = b(\theta + b)^{-2}$. For this prior, the Bayes factor is given by

$$B_{01} = \frac{b^x\, e^{-b}}{\int_0^\infty (\theta + b)^x\, e^{-(\theta+b)} b(\theta + b)^{-2}\, d\theta} = \frac{b^{(x-1)}\, e^{-b}}{\Gamma(x - 1, b)}\,,$$

where Γ is the incomplete gamma function. The result for the two cases is

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$); *Case 2:* $B_{01} = 0.26$ (recall $p = 0.025$)

2.3.2 Objective posterior probabilities of the hypotheses

The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}}\,.$$

For the two cases in the example,

Case 1: $\Pr(H_0 \mid x) = 0.0075$ (recall $p = 0.00025$); *Case 2:* $\Pr(H_0 \mid x) = 0.21$ (recall $p = 0.025$).

Of course, one can specify subjective prior probabilities of each hypothesis and determine the resulting posterior probabilities, but scientific communication is usually done through objective posterior probabilities or Bayes factors, since any individual can take either and easily convert it into the individual's personal subjective answer.

2.3.3 Complete posterior distribution

In addition to the uncertainty in the hypotheses, there is also uncertainty in θ, given that H_1 were true. The complete posterior distribution is thus determined by

- $\Pr(H_0 \mid x)$, the posterior probability of the null hypothesis;
- $\pi(\theta \mid x, H_1)$, the posterior distribution of θ under H_1.

For Case 1 in the example, Figure 1 presents these two parts of the full posterior distribution. One way of thinking of this is that the vertical bar gives the probability that one has just observed noise, while the density part says where θ is likely to be if there is a discovery.

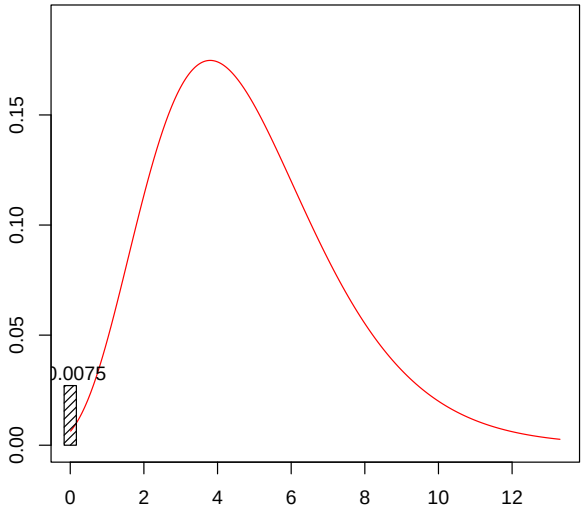

Fig. 1: For Case 1, $\Pr(H_0 \mid x)$ (the vertical bar), and the posterior density for θ given $x = 7$ and H_1.

A useful summary of the complete posterior is $\Pr(H_0 \mid x)$ and C, a (say) 95% posterior confidence interval for θ under H_1. For the two cases, and with C chosen to be an equal-tailed 95% posterior confidence interval (i.e., omitting 2.5% of the posterior mass on the left and the right)
Case 1: $\Pr(H_0 \mid x) = 0.0075$ and $C = (1.0, 10.5)$; *Case 2:* $\Pr(H_0 \mid x) = 0.21$ and $C = (0.2, 8.2)$.
C could, alternatively, be chosen to be a one-sided confidence bound, if desired.

Note that confidence intervals alone are *not* a satisfactory inferential summary. In Case 2, for instance, the 95% confidence interval does not include 0, and so many mistakenly believe that one can accordingly reject $H_0 : \theta = 0$. But, the full posterior distribution also has a probability of 0.21 that $\theta = 0$, which would hardly imply a confident rejection.

A Brief Aside: A precise null hypothesis, such as $H_0 : \theta = 0$, is typically never true *exactly*; rather, it is used as a surrogate for a 'real null' $H_0^\epsilon : \theta < \epsilon, \quad \epsilon$ small. In the Higgs example for instance, while the scientific null is real (i.e., the Higgs might not exist), the statistical null is based on the experimental measurements, and there is undoubtedly some small bias ϵ in the experiment. Berger and Delampady [4] show that, under reasonable conditions, if $\epsilon < \frac{1}{4}\sigma_{\hat{\theta}}$, where $\sigma_{\hat{\theta}}$ is the standard error of the estimate of θ, then $Pr(H_0^\epsilon|\mathbf{x}) \approx Pr(H_0|\mathbf{x})$, so that the point null is then a reasonable approximation to the real null.

2.4 The discrepancy between p-values and posterior probabilities

The Bayesian error probabilities given in the previous section differed from the corresponding p-values by factors of 30 and 10 in the two cases, respectively. What explains this?

It might be tempting to say that there is something wrong with the Bayesian analysis, but even a pure likelihood analysis (favored by many Fisherians) reveals the same effect. In particular (following Edwards, Lindeman and Savage [10]), note that a lower bound on the Bayes factor over all possible priors can be found by choosing $\pi(\theta)$ to be a point mass at $\hat{\theta}$ (the maximum likelihood estimate), yielding

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0+b)}{\int_0^\infty \text{Poisson}(x \mid \theta+b)\pi(\theta)\, d\theta} \geq \frac{\text{Poisson}(x \mid 0+b)}{\text{Poisson}(x \mid \hat{\theta}+b)} = \min\{1, \left(\frac{b}{x}\right)^x e^{x-b}\}\,. \qquad (1)$$

In 'likelihood language,' this says that, for the given data, the likelihood of H_0 relative to the likelihood of H_1 is at least the bound on the right hand side of (1). For the two cases, this bound is

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.11$ (recall $p = 0.025$),

so that a serious discrepancy remains even when the prior is eliminated. This can be traced to the fact that the p-value is based on the probability of the tail area of the distribution, rather than the probability of the actual observed data.

It is well known that Bayesian analysis utilizing suitable proper priors will automatically penalize more complex models (i.e., has an Ockham's razor effect – cf. Jefferys and Berger [5]), and it is useful to separate this effect from that observed above in explaining the difference between p-values and posterior probabilities or Bayes factors. Thus in Case 1, where the p-value ($\approx .00025$) and the objective posterior probability of the null (≈ 0.0075) differ by a factor of 30,

- a factor of $.0014/.00025 \approx 5.6$ is due to the difference between a tail area $\{X : X \geq 7\}$ and the actual observation $X = 7$ (as reflected through the likelihood ratio for the observation);
- the remaining factor of roughly 5.4 in favor of the null results from the Ockham's razor penalty resulting from the conventional proper prior that was used.

An Aside – Robust Bayesian Analysis: Robust Bayesian theory (cf. Berger [6] for references) takes a more sophisticated look at the type of bounding over priors that is done in (1). For instance, it might be deemed scientifically reasonable to restrict attention to priors $\pi(\theta)$ that are nonincreasing, in which case it is easy to see that

$$B_{01}(x) \geq \frac{b^x\, e^{-b}}{\sup_c \int_0^c (\theta+b)^x\, e^{-(\theta+b)} c^{-1}\, d\theta}\,.$$

For the two cases, this bound is

Case 1: $B_{01} \geq 0.0024$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.15$ (recall $p = 0.025$).

2.5 Conditional frequentist testing

There is a powerful (but, alas, largely overlooked) frequentist school called *conditional frequentist analysis*. This school was formalized by Kiefer [7] and Brown [8], and proceeds as follows:

- find a statistic S that reflects the "strength of evidence" in the data;
- compute the frequentist measure of error conditional on S.

Artificial example (from Berger and Wolpert [9]): Observe X_1 and X_2 where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

A classical (unconditional) 75% confidence set (here a point) for the unknown θ is

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2\,; \end{cases}$$

it is easy to compute that $P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75$. It is, however, clearly silly to report this; when $X_1 \neq X_2$, it is a certainty that the confidence set equals θ while, if $X_1 = X_2$, it is intuitively 50-50 as to whether the confidence set equals θ. The issue here is typically phrased in statistics as that of desiring good conditional performance (for relevant subsets of the actual data); in the Phystat literature it is more commonly phrased as desiring *Bayesian credibility:* for a reasonable prior, the Bayesian coverage of the confidence set should be reasonable. In this example, for instance, if one uses the objective prior $\pi(\theta) = 1$, then $C(X_1, X_2)$ has Bayesian credibility of 100% if $x_1 \neq x_2$ and 50% if $x_1 = x_2$, so that the report of 75% confidence in all circumstances would be seriously deficient from the viewpoint of Bayesian credibility.

The conditional frequentist approach here would

- measure the strength of evidence in the data by, say, $S = |X_1 - X_2|$ (either 0 or 2)
- compute the conditional coverage

$$P_\theta(C(X_1, X_2) \text{ contains } \theta \mid S) = \begin{cases} 0.5 & \text{if } S = 0 \\ 1.0 & \text{if } S = 2\,, \end{cases}$$

which is clearly the right answer.

Returning to the testing problem, Berger, Brown and Wolpert [11] for continuous data, and Dass [12] for discrete data, proposed the following conditional frequentist testing procedure for testing a simple hypothesis versus a simple alternative:

- Develop S, the measure of strength of evidence in the data, as follows:
  - let $p_i(x)$ be the p-value from testing H_i against the other hypothesis;
  - define $S = \max\{p_0(x), p_1(x)\}$; its use is based on deciding that data (in either the rejection or acceptance regions) with the same p-value has the same 'strength of evidence.'
- Accept H_0 when $p_0 > p_1$, and reject otherwise.
- Compute Type I and Type II conditional error probabilities as

$$\alpha(s) = P_0(\text{rejecting } H_0 \mid S = s) \equiv P_0(p_0 \leq p_1 \mid S(X) = s)$$
$$\beta(s) = P_1(\text{accepting } H_0 \mid S = s) \equiv P_1(p_0 > p_1 \mid S(X) = s),$$

  where P_i refers to probability under H_i.

The surprising feature of this conditional test is stated in the following theorem from those papers.

Theorem 1 *The conditional frequentist error probabilities, $\alpha(s)$ and $\beta(s)$, exactly equal the (objective) posterior probabilities of H_0 and H_1, so conditional frequentists and Bayesians report the same error probabilities.*

In our ongoing example, the conditional Type I error is thus $\alpha(s) = \Pr(H_0 \mid x) = B_{01}/(1 + B_{01})$ (=0.0075 in Case 1; =0.21 in Case 2). Some features of this:

- The conditional test can be viewed as a way to convert p-values into real frequentist error probabilities when there is an alternative hypothesis.
- The conditional error probabilities $\alpha(s)$ and $\beta(s)$ are fully data-dependent (being smaller when p is smaller, in contrast to the fixed α-level tests), yet are fully frequentist.
- The conditional test also applies without any change in sequential settings; since Bayesian error probabilities are known to ignore the stopping rule, so must the conditional frequentist test (Berger, Boukai and Wang [13]).

The conditional frequentist test thus overcomes all of the difficulties with the fixed α-level test that were discussed earlier, and so can be used happily by frequentists. Of course, one need not go through the formal conditional frequentist computation, since the theorem guarantees that the answer which would be obtained is the same as the objective Bayesian answer (which can be obtained much more directly).

There is the caveat that the above discussion was given only for the testing of two simple hypotheses. In our ongoing example, on the other hand, H_1 was a composite hypothesis (involving an unknown θ). The papers mentioned above do cover the extension of the theory to the composite alternative case, with the only modification being that the conditional Type II error that is obtained is a certain average Type II error over θ; the conditional Type I error is unaffected. Extensions to composite null hypotheses are considered in Dass and Berger [14] for composite null hypotheses that have an invariance structure to group operations; this class of composite null hypotheses includes most classical situations of testing. The nice feature of this class of composite null hypotheses is that the conditional Type I error is constant over the null hypothesis, and so no averaging over Type I error needs to be done. (There are other technical caveats to the conditional frequentist testing paradigm that are discussed in the mentioned papers, but they have essentially no practical impact.)

2.6 Implementing Bayesian testing

To implement objective Bayesian estimation (and confidence procedures) there are, in principle, excellent objective priors available, such as *reference priors* (see Bernardo [15] for a review and references). In practice, determination of such objective priors can be challenging but the goal is, at least, clear.

In Bayesian hypothesis testing and model selection, however, determination of suitable prior distributions is considerably more challenging, in part because it is typically the case that improper prior distributions cannot be used (or at least have to be used very carefully). Use of 'vague proper priors' (another staple of many Bayesians in estimation problems) is even worse, and will typically give nonsensical answers in testing and model selection. There has thus been a huge effort in statistics to derive objective (or at least conventional) priors for use in hypothesis testing and model selection. These issues and this literature can be accessed through Berger and Pericchi [16].

For our ongoing example, an appealing methodology for default prior construction is the *intrinsic* or *expected posterior* prior construction. For the situation where the data consists of i.i.d. observations from a density $f(x \mid \theta)$, and for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, the construction is as follows:

- let $\pi^O(\theta)$ be a good estimation objective prior, so that $\pi^O(\theta \mid \boldsymbol{x}) = [\prod_{i=1}^n f(x_i \mid \theta)]\pi^O(\theta)/m^O(\boldsymbol{x})$ is the resulting posterior, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $m^O(\boldsymbol{x}) = \int [\prod_{i=1}^n f(x_i \mid \theta)]\pi^O(\theta)\, d\theta$;
- then the intrinsic prior is $\pi^I(\theta) = \int \pi^O(\theta \mid \boldsymbol{x}^*)[\prod_{i=1}^q f(x_i \mid \theta_0)]\, d\boldsymbol{x}^*$, with $\boldsymbol{x}^* = (x_1, \ldots, x_q)$ being (unobserved) data of the minimal sample size q such that $m^O(\boldsymbol{x}^*) < \infty$.

Note that this will be a proper (not vague proper) prior.

The idea behind this prior is that, if one were handed the data $\boldsymbol{x}^*$ but allowed to use it only for prior construction, one would happily compute $\pi^O(\theta \mid \boldsymbol{x}^*)$ and use this proper prior to conduct the test. We don't have $\boldsymbol{x}^*$ available, but we could simulate $\boldsymbol{x}^*$ from the null model, and compute the resulting 'average' prior. There are many other justifications of this prior; see Pérez and Berger [17] for discussion and references. Note, however, that use of such conventional proper priors is inherently more contentious than use of objective priors for estimation problems. Indeed, it would be better to determine $\pi(\theta)$ from consensus scientific knowledge, providing the knowledge is relatively precise and quantifiable.

For our ongoing example, suppose we choose $\pi^O(\theta) = 1/(\theta + b)$. (Jeffreys prior, the square root of π^O, would probably be better, but leads to a much more difficult computation.) Following the ideas in Berger and Pericchi [18], we represent the Poisson observation, X, over the time period T from the distribution in the example as a sum of i.i.d observations from an exponential inter-arrival time process.

Indeed, for $i = 1, \ldots$, consider $Y_i \sim f(y_i \mid (\theta + b)/T) = (\theta + b)T^{-1} \exp\{-(\theta + b)y_i/T\}$; then $X \equiv \{\text{first } j \text{ such that } S_j = \sum_{i=1}^{j} Y_i > T\} - 1$. A minimal sample size for this exponential distribution can easily be seen to be $q = 1$. Computation then yields $\pi^I(\theta) = \int \pi^O(\theta \mid y_1) f(y_1 \mid 0) dy_1 = b/(\theta+b)^2$, which was the conventional proper prior used for Bayesian testing in the example.

3 Multiplicities

The issue of dealing with multiplicities in discovery is increasingly being recognized to be important. One type of multiple testing has already been discussed, namely sequential experimentation in which one periodically evaluates the incoming data to see if a discovery can be claimed. It is interesting that frequentist analyses often need to be adjusted to account for these 'looks at the data,' while Bayesian analyses (and optimal conditional frequentist analyses) do not. That Bayesian analysis claims no need to adjust for this 'look elsewhere' effect – called the *stopping rule principle* – has long been a controversial and difficult issue in statistics, as admirably expressed by Savage [19]: "I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that people resist an idea so patently right." See Berger and Berry [20] for discussion of this controversy, and note that the controversy is no longer a frequentist versus Bayesian issue, because of the fact that optimal conditional frequentist tests also obey the stopping rule principle.

Another common situation of multiple testing is when one is scanning many possible data sets for a discovery. For instance, suppose 1000 energy channels are searched for a signal expected from a non-standard theory. It is well known that one cannot proceed with separate testing of each data set, but the classical solution – the Bonferonni adjustment – is often viewed as being too harsh. The Bonferonni adjustment assumes each test is independent, in which case one divides the desired error probability α by the number of tests to determine the significance level that an individual test must achieve to declare a discovery. Thus if $\alpha = 0.001$ is desired for 1000 independent tests, the per-test significance level should be set at 0.000001 for declaring a discovery.

I have been told that the assumption of (at least approximate) independence of test statistics does hold for many high-energy physics experiments, in which case use of the Bonferonni correction is fine. When the various test statistics are dependent, however (as happens in most non-physics examples I know of), the Bonferroni correction can be much too conservative, so it's use would incur a dramatic loss of power for discovery. Finding appropriate correction for multiple testing under dependence is, unfortunately, quite difficult from the frequentist viewpoint. Note, also, that there are no shortcuts here; simple alternative methods such as the 'false discovery rate' are fine for screening purposes, but are not useful for claiming a discovery.

One of the highly attractive features of the Bayesian approach to multiple testing or model selection is that (if done properly) it will automatically adjust for multiplicities, and do so in a way that preserves as much discriminatory power as possible. The Bayesian adjustment for multiplicity occurs, somewhat curiously, directly through the prior probabilities assigned to the tests or models. Consider two illustrative cases:

Mutually exclusive hypotheses: Suppose one is testing mutually exclusive hypotheses $H_i,\ i = 1, \ldots, m$, where it is known that one is true. An objective Bayesian would choose $p_i = \Pr(H_i) = 1/m$. Suppose, for instance, that a signal is known to exist, but it is not known in which of 1000 energy channels it will manifest. Then each channel would be assigned prior probability 0.001 of containing the signal, an automatic penalization of each hypothesis.

Suppose instead that 1000 channels are searched for a signal expected from a non-standard theory that could manifest in only one channel. Then one should assign some prior mass – e.g. 1/2 – to 'no signal,' giving prior probability of 0.0005 to each channel. Note that these simple adjustments apply no matter what the dependence is between the test statistics, indicating why it is much easier to approach

multiplicity adjustment from the Bayesian perspective.

Independently occurring hypotheses: Consider, instead, the situation in which there are multiple possible discoveries, and that the signal from each would appear in a separate channel. If we knew nothing about these possible signals, we might choose to assign prior probabilities by first defining p as the probability that any given channel will manifest a signal. This would typically be unknown, and hence would need to be assigned a prior distribution $\pi(p)$. This could be chosen according to scientific knowledge, or set equal to a default prior such as the uniform distribution. That an assignment of prior probabilities such as this automatically deals with multiplicity is demonstrated in Scott and Berger [21].

There is a large and increasing literature on discovery techniques in the face of multiplicity. Two recent references are Storey, Dai and Leek [22] and Guindani, Zhang and Mueller [23].

4 Musings on the meaning of frequentism

4.1 Introduction and example

During the Phystat meeting, there were a number of interesting problems discussed that caused me to reflect on the meaning of frequentism. To facilitate the discussion here, consider the following version of the basic HEP problem, but now focusing on confidence bounds (see, e.g., Heinrich [24] for background).

Suppose $X_{s+b} \sim \text{Poisson}(X_{s+b} \mid s+b)$, where s is the unknown signal mean and b now an unknown background mean. The goal is to find an upper confidence limit for s. There is also information available about the nuisance parameter b, arising from either

- *Case 1:* independent sideband data $X_b \sim \text{Poisson}(X_b \mid b)$,
- *Case 2:* randomness in b from experiment to experiment arising from a known random mechanism,
- *Case 3:* agreed scientific beliefs.

4.2 Bayesian analysis

Suppose we have an agreed upon objective prior density $\pi^O(s \mid b)$ for s given b (the best objective priors will typically depend on nuisance parameters such as b here). The information about b would be encoded in a prior density $\pi(b)$. This density would be derived differently in each case:

- *Case 1:* With the sideband data X_b, a standard approach would be to chose an initial objective prior $\pi^O(b)$, and then choose the final $\pi(b)$ to be the posterior $\pi^O(b \mid X_b) \propto \text{Poisson}(X_b \mid b)\pi^O(b)$.
- *Case 2:* $\pi(b)$ describes the physical randomness of the (otherwise unmeasured) background from experiment to experiment.
- *Case 3:* $\pi(b)$ is chosen to encode accepted scientific beliefs.

In all three cases, Bayesian analysis would proceed in the same way, constructing a $100(1-\alpha)\%$ upper confidence limit U for s as the solution to

$$1-\alpha = \int_0^U \pi(s \mid X_{s+b})\, ds\,,$$

where $\pi(s \mid X_{s+b})$ is the posterior distribution

$$\pi(s \mid X_{s+b}) = \frac{\int \text{Poisson}(X_{s+b} \mid s+b)\pi^O(s \mid b)\pi(b)\, db}{\int\int \text{Poisson}(X_{s+b} \mid s+b)\pi^O(s \mid b)\pi(b)\, db\, ds}\,.$$

The point is that Bayesian analysis does not care about the nature of the randomness in the modeling of the information about b.

4.3 Frequentist analysis

Frequentist analysis can be quite different in the three cases.

4.3.1 Frequentist analysis in Case 1.

The natural frequentist goal: Frequentist coverage with respect to the joint distribution of X_{s+b} and X_b, i.e. control of

$$P(s \leq U(X_{s+b}, X_b) \mid s, b) = \sum_{X_{s+b}=0}^{\infty} \sum_{X_b=0}^{\infty} 1_{\{s \leq U(X_{s+b}, X_b)\}} \text{Poisson}(X_{s+b} \mid s+b) \text{Poisson}(X_b \mid b)\,,$$

where $1_{\{s \leq U(X_{s+b}, X_b)\}}$ is 1 if $s \leq U(X_{s+b}, X_b)$ and 0 otherwise.

This problem has been extensively studied in the Phystat literature. It is interesting that there is, as of yet, no solution which is agreed by all to be adequate in terms of both frequentist coverage and Bayesian credibility (conditional performance). The objective Bayesian holy grail in this problem would be to find the reference prior for (s, b), with s being the parameter of interest; the hope is that the upper confidence bound arising from such a prior would do an excellent job of balancing frequentist coverage and Bayesian credibility. Finding the reference prior is very challenging, however, as was discussed in the Phystat talk of Luc Demortier (and see Demortier [25]).

4.3.2 Frequentist analysis in Case 2.

The natural frequentist goal: Frequentist coverage with respect to the marginal density of X_{s+b}, given by $f(X_{s+b} \mid s) = \int \text{Poisson}(X_{s+b} \mid s+b)\pi(b)db$. The coverage target is then

$$P(s \leq U(X_{s+b}) \mid s) = \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} f(X_{s+b} \mid s)\,.$$

The reason this is the natural frequentist goal is because b changes from experiment to experiment according to $\pi(b)$, and *real* frequentism is about performance of a statistical procedure in actual repeated use of the procedure in differing experiments, as discussed in Neyman [2]. (The textbook definition of frequentism – in which one considers *imaginary* repetition of the same experiment – makes no sense in terms of reality; the standard definition has mathematical relevance, but the philosophical appeal of frequentism to scientists is presumably its relevance to real experimentation over time.)

Attaining this frequentist goal while achieving good Bayesian credibility is potentially rather straightforward, since the problem has been reduced to a one-parameter problem. Indeed, one simply computes the reference (Jeffreys) prior corresponding to $f(X_{s+b} \mid s)$, namely

$$\pi^J(s) = \sqrt{I(s)}\,, \qquad I(s) = -\sum_{X_{s+b}=0}^{\infty} f(X_{s+b} \mid s) \frac{d^2}{ds^2} \log f(X_{s+b} \mid s)\,.$$

The resulting Bayesian confidence bound will automatically have good Bayesian credibility (conditional performance), and the Jeffreys prior for one-parameter problems typically results in Bayes procedures with excellent frequentist coverage properties (except possibly at the boundary $s = 0$; see Bayarri and Berger [26] for discussion).

4.3.3 Frequentist analysis in Case 3.

The natural frequentist goal: Here the situation is quite murky. Since $\pi(b)$ is not physical randomness, but simply scientific opinion, a classical frequentist could insist that, for every given s and b, we control

$$P(s \leq U(X_{s+b}) \mid s, b) = \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} \text{Poisson}(X_{s+b} \mid s+b)\,.$$

This is actually not possible to control unless there is a known bound on b, but a classical frequentist would philosophically wish to control this coverage.

Alternatively, one could argue that, since $\pi(b)$ arises from consensus scientific opinion, it should be treated the same as when it arises from physical randomness, and so one should seek to control coverage as in Case 2, i.e.

$$\begin{aligned} P(s \leq U(X_{s+b}) \mid s) &= \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} f(X_{s+b} \mid s) \\ &= \int P(s \leq U(X_{s+b}) \mid s, b)\pi(b)db\,. \end{aligned}$$

The second expression for this coverage shows that the criterion can be interpreted as an average of the coverage for given s and b, averaged over the consensus prior distribution for b.

There are many situations in which it has been argued that a frequentist should use an average coverage criterion; see Bayarri and Berger [26] for examples and references. Here it seems clearly right because of necessity; what else can be done given the available information? The point worth pondering is – if average coverage is fine here, why should it be philosophically problematical in other cases?

Acknowledgements

This work was supported by NSF Grants AST-0507481 and DMS-0103265. Also very helpful was the extensive discussion by participants in the working group on Particle Physics (lead by Louis Lyons) during the program on Astrostatistics at the Statistical and Applied Mathematical Sciences Institute in March of 2006. Finally, thanks to Luc Demortier and Louis Lyons for their very helpful comments and discussions concerning this paper.

References

[1] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference (3rd ed.)*. Macmillan, London.

[2] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthèse*, **36**, 97–131.

[3] Jeffreys, H. (1961). *Theory of Probability*, London: Oxford University Press.

[4] Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with Discussion). *Statist. Science* **2**, 317–352.

[5] Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

[6] Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

[7] Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with Discussion). *J. Amer. Statist. Assoc.* **72**, 789–827.

[8] Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6**, 59-71.

[9] Berger, J., and Wolpert, R. L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications* (second edition, with Discussion). Hayward, CA: Institute of Mathematical Statistics.

[10] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

[11] Berger, J., Brown, L.D. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, **22**, 1787–1807.

[12] Dass, S. C. (2001). Unified Bayesian and conditional frequentist testing procedures for discrete distributions. *Sankhya Ser. B*, **63**, 251–269.

[13] Berger, J., Boukai, B. and Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, **86**, 79–92.

[14] Dass, S. and Berger, J. (2003). Unified Bayesian and conditional frequentist testing of composite hypotheses. *Scandinavian Journal of Statistics*, **30**, 193–210.

[15] Bernardo, J. M. (2005). Reference analysis. Handbook of Statistics 25 (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17–90.

[16] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 135–207.

[17] Pérez, J.M. and Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

[18] Berger, J. and Pericchi, L. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.

[19] Savage, L.J. (1962). *The Foundations of Statistical Inference.* London: Methuen.

[20] Berger, J. and Berry, D. (1988). The relevance of stopping rules in statistical inference (with Discussion). In *Statistical Decision Theory and Related Topics IV.* Springer–Verlag, New York.

[21] Scott, J. and Berger, J. (2006) An exploration of aspects of Bayesian multiple testing, Journal of Statistical Planning and Inference, Vol. 136, No. 7. (1 July 2006), pp. 2144–2162.

[22] Storey, J.D., Dai, J.Y and Leek, J.T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8***(2)*, 414–432.

[23] Guindani, M., Zhang, S. and Mueller, P.M. (2007). A Bayesian discovery procedure. Technical Report, MD Anderson Medical Center.

[24] Heinrich, J. (2008). Review of the Banff challenge on upper limits. These Proceedings.

[25] Demortier, L. (2005). Bayesian reference analysis for particle physics. *PHYSTAT05 Proceedings on "Statistical Problems in Particle Physics, Astrophysics and Cosmolgy"*. Oxford University Press.

[26] Bayarri, M.J. and Berger, J. (2004). The interplay between Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80.