CDF/MEMO/STATISTICS/PUBLIC/9426
Version 2.00
August 21, 2008

Some Statistical Issues in the Measurement of the Top Quark Charge

Luc Demortier[1]

Laboratory of Experimental High-Energy Physics

The Rockefeller University

Abstract

The measurement of the top quark charge at the Tevatron is currently set up as a frequentist hypothesis test comparing the standard model with an exotic model involving a heavy quark with charge $+4e/3$. A well known deficiency of this approach is that the confidence level of the test result must be chosen before the measurement and does not reflect the strength of evidence exhibited by the data. Other difficulties include the choice of null hypothesis and the choice of rejection threshold. After reviewing these issues we describe a *conditional* frequentist procedure which does take evidential strength into account and provides a new perspective on the above difficulties. We explain how this procedure not only satisfies standard frequentist desiderata, but can also be made to coincide with a Bayesian test, thereby enriching its potential for interpretation. Although the main motivation for this note is provided by the top charge analysis, the presentation is general enough to serve as an introduction to the main issues of frequentist hypothesis testing.

[1]luc@fnal.gov

Contents

1 Introduction

The standard reconstruction of top quark momenta in $t\bar{t}$ events does not attempt to identify the charge of the b quark jet associated with the W boson in each top quark decay. Thus, if we ignore background effects, the reconstructed quark could either be a standard model top quark decaying according to $t \to W^+ b$ and carrying electric charge $+2e/3$, or an exotic quark with charge $+4e/3$ decaying into $W^+\bar{b}$. In the latter case, consistency of the overall electroweak dataset can be maintained by assuming that the real top quark has a mass around 270 GeV/c^2 [1, 2, 3]. Such a scheme is of course beyond the standard model and will be referred to as "exotic model" in the remainder of this note. These theoretical considerations have motivated measurements of the charge Q of the particle currently assumed to be the top quark by both DØ [4, 5] and CDF [6, 7, 8, 9]. Although both collaborations conclude that said particle is much more likely to have charge $+2e/3$ than $+4e/3$, they have summarized their results in a way that frustrates direct comparison. This note is an attempt to lay out the statistical issues involved and point to possible solutions of the difficulties encountered.

The measurement of the charge of b jets, and therefore also that of top quarks [2], has rather poor resolution. However, there are two circumstances that help the experimenter in the present case. The first one is that it is sufficient to test the discrete hypothesis $H_0 : Q = +2e/3$ versus the discrete alternative $H_1 : Q = +4e/3$. There is no need to construct a confidence interval for Q, as if one had no idea of its true value. Thus, only the *sign* of the b jet charge needs to be determined for each top decay. The second circumstance is that both Tevatron collaborations now have at their disposal sizeable samples of top quarks, which helps the power of the test. Given a sample of reconstructed top quarks, what is done in practice is to calculate the fraction μ of quarks that can be classified as consistent with the standard model according to the sign of the associated b jet charge. The above hypothesis test can then be reformulated as a test of $H_0 : \mu = 1$ versus $H_1 : \mu = 0$.

Although there are several ways of performing such a test, the tradition in high energy physics favors frequentist approaches, and this will be our starting point; the standard frequentist test procedure and terminology are described in a separate subsection at the end of this introduction. We then discuss the choice of null hypothesis in section 2 and the interpretation of confidence levels (or error rates) in section 3. This is followed in section 4 by a discussion of the choice of the rejection threshold α and the effect this choice has on the interpretation of the results of the test. Sections 3 and 4 bring to light a couple of difficulties with the standard frequentist approach. The first one is that the level of confidence one has in the final test result is already known before looking at the data; in other words, the actual strength of evidence displayed by the data is not incorporated in the confidence level. Secondly, the choice of the rejection threshold of the test requires a subjective assessment of one's beliefs in the null hypothesis versus the alternative, and of the losses one is willing to incur when making an incorrect decision regarding which hypothesis is true. Whereas this second difficulty is unavoidable in any testing paradigm, the first one can be alleviated by

[2] In the interest of conciseness we will from now on use the expression "top quark" in lieu of "the particle currently assumed to be the top quark".

adopting a *conditional* frequentist approach, as described in section 5. With a careful choice of conditioning statistic, this approach has the additional advantage of allowing for a Bayesian interpretation. The handling of systematic uncertainties is considered in section 6. Finally, a summary of all the testing methods studied in this note is provided in section 7.

For reference, summaries of the measurements made by the CDF and DØ collaborations, as well as their interpretations of their measurements, are reviewed in Appendices A and B. A third appendix contains some technical details needed to justify the results described in sections 5 and 6.

Our concern throughout the note is to supplement frequentist error probability statements with evidential interpretations, which are more relevant in discovery situations. For example, the top quark discovery claim was not based on the expectation of many repetitions of the same experiment, but rather on careful inference from available evidence in the experiment at hand. The validity of this inference can be tested at any time with larger data samples and more probing analyses, and this is precisely what the top charge measurement is attempting. More than ten years after the top quark discovery announcement, we are still gathering evidence to complete the picture. It will often prove insightful to use Bayesian concepts to characterize this evidence.

1.1 Basic frequentist test setup: terminology and notation

As above, let μ be the true fraction of top quarks that have an electric charge consistent with the $+2e/3$ hypothesis in a given $t\bar{t}$ data sample, and let X be the measured fraction of those quarks. Although μ is bounded between 0 and 1, we assume that X can exceed those bounds due to resolution effects. If we believe that heavy quarks of charge $+2e/3$ can coexist with heavy quarks of charge $+4e/3$ (and have the same mass), there are two possible testing problems:

$$\underline{\text{Problem 1:}} \qquad H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1,$$

and

$$\underline{\text{Problem 2:}} \qquad H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1' : \mu < \mu_1,$$

where $\mu_0 = 1$ and $\mu_1 = 0$ in the top charge analysis. If coexistence is not possible, then only Problem 1 is relevant.

A frequentist test of H_0 starts by the selection of a test statistic T, such that extreme values of T are evidence against H_0 and in favor of H_1. Often a good choice for T is the likelihood ratio or a one-to-one function of it. The next step is to "calibrate" the evidence contained in T in order to facilitate its interpretation. This is done by calculating a p value, which is defined as the probability under H_0 of obtaining the observed value of T, *or a more extreme value*. In the case of the top charge analysis the likelihood ratio is a one-to-one function of X, so we will use X as test statistic in the following. Since small values of X are evidence against H_0 for either problem 1 or problem 2, we can calculate the p value as:

$$p_0 \;=\; \int_{-\infty}^{x_{obs}} f(x \,|\, \mu_0)\; dx, \tag{1.1}$$

where $f(x \,|\, \mu)$ is the probability density distribution of X and x_{obs} is its observed value. To test H_0 we must choose a rejection threshold $\alpha \in]0, 1[$ prior to the test. Letting x_α be the α^{th} quantile of $f(x \,|\, \mu_0)$:

$$\int_{-\infty}^{x_\alpha} f(x \,|\, \mu_0)\, dx \;=\; \alpha, \tag{1.2}$$

the basic frequentist test procedure is to reject H_0 whenever $x_{obs} \leq x_\alpha$, or equivalently, whenever $p_0 \leq \alpha$. By definition of p_0 we have:

$$\mathbb{P}(p_0 \leq \alpha \,|\, H_0) \;=\; \alpha. \tag{1.3}$$

In words, α is the probability of rejecting H_0 if H_0 is true, and is known as the Type-I error probability of the test. The quantity x_α is known as the *critical value* or *critical boundary*, and the rejection region $x_{obs} \leq x_\alpha$ is also referred to as the *critical region*. The power function of the test is $1 - \beta(\alpha, \mu)$, where $\beta(\alpha, \mu)$ is the Type-II error probability, namely the probability to accept H_0 when it is false:

$$\beta(\alpha, \mu) \;\equiv\; \mathbb{P}(p_0 > \alpha \,|\, H_1) \;=\; \int_{x_\alpha}^{+\infty} f(x \,|\, \mu)\, dx. \tag{1.4}$$

If only Problem 1 is relevant, the power function reduces to $1 - \beta(\alpha, \mu_1)$. In this situation of a completely specified, so-called "simple" alternative hypothesis, one can also calculate a p value under the alternative:

$$p_1 \;=\; \int_{x_{obs}}^{+\infty} f(x \,|\, \mu_1)\, dx. \tag{1.5}$$

Small values of p_1 are evidence against H_1 in the direction of H_0. Equation (1.4) shows that β is generally a function of both α and μ; it is then easy to see that $p_1 = \beta(p_0, \mu_1)$, and that the rejection criterion $p_0 \leq \alpha$ is equivalent to $p_1 \geq \beta(\alpha, \mu_1)$. The relationships between p_0, p_1, α, and β are illustrated in Fig. 1.

2 Choice of the null hypothesis

Since we are dealing with two hypotheses, the standard model vs. the exotic model, setting up a hypothesis test requires that we first decide which hypothesis is to be the "null", and which one the "alternative". Even though this testing situation may appear symmetric between the two hypotheses, there are in fact two important asymmetries that need to be coordinated:

1. Explanatory power asymmetry
   Setting aside the question of the charge of the heavy quarks observed at the Tevatron, the standard and exotic models both explain precision electroweak data with equal success. However, the exotic model requires more parameters to achieve this (more quarks, more Higgs bosons,...), and has therefore less explanatory power than the more parsimonious standard model. If we subscribe to Ockham's razor, the regulative principle according to which unproved assumptions should not be unnecessarily multiplied, then the standard model is favored a priori, before inspecting the top charge data.

2. Error control asymmetry
   As explained in section 1.1, in any test of a null hypothesis versus an alternative, one can consider two kinds of error: Type-I, whereby the null hypothesis is incorrectly rejected, and Type-II, whereby the alternative is incorrectly rejected. However, the probability of only one of these errors can be directly controlled by the experimenter. Once this is done, the probability of the other error is constrained by the resolution of the measurement and cannot be reduced at will. The Neyman-Pearson approach is to fix the Type-I error probability α at some level and then adjust the critical region so as to minimize the Type-II error probability β.

It follows from these two asymmetries that the choice of the null hypothesis should be based on a consideration of the consequences of the two types of error. From the first asymmetry it appears that incorrectly rejecting the standard model is a worse error than incorrectly rejecting the exotic model. Assuming that one would want full control of the probability of the worse error, the second asymmetry then implies that the standard model should be taken as null hypothesis.

In this approach, if we fail to reject the null hypothesis, we will be left with the better model, until such time as new and more convincing data-based evidence forces us to adopt an improved model. Such a testing strategy is particularly safe in situations where the measurement resolution is low. Indeed, it would be rather embarrassing to test the exotic model and then fail to reject it because of a lack of measurement resolution.

2.1 The minimax strategy

A strategy that bypasses the need to select one hypothesis as the null is to calculate two p values, p_0 under the standard model and p_1 under the exotic model, and then to reject the model with the smaller p value. Data that are consistent with the standard model will tend to have large p_0 and small p_1, and vice-versa for data that are consistent with the exotic model.

It is easy to see that this strategy implies equal Type-I and Type-II errors for continuous test statistics. Indeed, suppose this were not true, and that $\alpha < \beta$ for example. If we observe $X = x_\alpha$, we will find that $p_0 = \alpha < \beta = p_1$. By continuity, if we observe a slightly larger value of X, we will then have $\alpha < p_0 < p_1$. The first of these inequalities implies that we must accept H_0, whereas the second one implies that we must reject it. Since this is a contradiction, our premise that $\alpha < \beta$ must be wrong. A similar argument shows that α cannot be strictly larger than β. Thus we must have $\alpha = \beta$.

A test with $\alpha = \beta$ is sometimes called equal-tailed. It can also be derived from a minimax criterion, namely by minimizing the maximum probability of error, $\max\{\alpha, \beta\}$. Indeed, decreasing α tends to increase β, and vice-versa, so that the maximum of these two error rates is minimized at the equilibrium value for which $\beta = \alpha$.

Advantages of minimax tests are that they automatically take care of the choice of α, and that they yield a unique probability of error: whether we accept or reject H_0,

the probability of a mistake is the same, and it minimizes the maximum probability of error among all frequentist tests based on the same test statistic. The formulation in terms of p_0 and p_1 has the further advantage that it shows us directly if the data agree with *neither* hypothesis (p_0 and p_1 both small), or with *both* (p_0 and p_1 both large). The downside is that we must be willing to set both hypotheses on the same footing, and to relinquish control of the error rates. The latter will tend to be small if the experiment has good resolution, and large otherwise. If the CDF and DØ experiments were to adopt a minimax strategy in the top charge analysis, one would be able to compare their sensitivities by examining their α values.

The following example illustrates that if the resolution improves with the sample size n, then the minimax test will be consistent, meaning that the probability of selecting the correct hypothesis will go to 1 as $n \to \infty$. Contrast this with fixed-size testing, where the probability of selecting the correct hypothesis is $1-\alpha$ or $1-\beta$ regardless of sample size.

Example 1 (Gaussian approximation to the top charge analysis)
A useful approximation to the top charge analysis is to treat the distribution of x as Gaussian with mean μ and known width σ:

$$f(x\,|\,\mu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma}. \tag{2.1}$$

The standard model p value can now be calculated explicitly, using equation (1.1); this yields:

$$p_0(x_{obs}) \;=\; \frac{1}{2}\Big[1+\mathrm{erf}\Big(\frac{x_{obs}-\mu_0}{\sqrt{2}\,\sigma}\Big)\Big], \quad \text{where} \quad \mathrm{erf}(x) \equiv \frac{2}{\pi}\int_0^x e^{-t^2}\,dt, \tag{2.2}$$

where we explicitly indicated the dependence of p_0 on the observed value of X. To calculate the exotic model p value, we note that *large* values of X are evidence *against* H_1 in the direction of H_0. Therefore:

$$p_1(x_{obs}) \;\equiv\; \int_{x_{obs}}^{+\infty} f(x\,|\,\mu_1)\,dx \;=\; \frac{1}{2}\Big[1+\mathrm{erf}\Big(\frac{\mu_1-x_{obs}}{\sqrt{2}\,\sigma}\Big)\Big]. \tag{2.3}$$

The minimax critical region, $p_0(X) \le p_1(X)$, is equivalent to $X \le \bar{\mu}$, where

$$\bar{\mu} \;=\; \frac{\mu_0+\mu_1}{2}. \tag{2.4}$$

The frequentist error rate of the test is then:

$$\alpha \;=\; \mathbb{P}\Big[X \le \bar{\mu}\,\Big|\,H_0\Big] \;=\; \frac{1}{2}\Big[1-\mathrm{erf}\Big(\frac{\Delta\mu}{2\sqrt{2}\,\sigma}\Big)\Big] \;=\; \mathbb{P}\Big[X > \bar{\mu}\,\Big|\,H_1\Big] \;=\; \beta, \tag{2.5}$$

where $\Delta\mu \equiv \mu_0 - \mu_1$. An experiment with good resolution has $\sigma \ll \Delta\mu$ and therefore *small* α. In the top charge analysis $\Delta\mu = 1$ and $\sigma \approx 0.38$, yielding $\alpha \approx 9.4\%$, which is still a rather substantial probability of error. One way to reduce it would be to take

more data, if this reduces the value of the parameter σ. Suppose for example that σ decreases at the rate of $1/\sqrt{n}$, where n is the sample size. If we replace σ by $\sigma/\sqrt{n}$ in the expression for the p.d.f. (2.1), then the probability to correctly accept H_0 becomes:

$$\mathbb{P}\Big[p_0 > p_1 \,\Big|\, H_0\Big] \;=\; \frac{1}{2}\Big[1 + \mathrm{erf}\Big(\frac{\sqrt{n}\,\Delta\mu}{2\sqrt{2}\,\sigma}\Big)\Big], \tag{2.6}$$

which goes to 1 as n goes to infinity. ■

3 Unconditional confidence levels

As explained in section 1.1, the standard frequentist procedure for testing a null hypothesis H_0 versus an alternative H_1 is to choose a rejection threshold α, calculate a p value p_0 under H_0, and reject the latter if $p_0 \leq \alpha$. The probability of incorrectly rejecting H_0 is then α, whereas that of incorrectly rejecting H_1 is labeled β, and in general $\beta \neq \alpha$. A common question is the following. When the test leads us to accept H_0, what is our confidence that H_0 is true? Similarly, when the test rejects H_0, what is our confidence that H_0 is false? Standard frequentist theory answers each of these questions by providing two confidence levels, both of which associate the notion of confidence with the long-run probability of "being right". The argument goes as follows. Suppose first that the result of the test is to accept H_0. We know that if H_0 is true, the probability of making the right decision is $1-\alpha$. Therefore our confidence in our decision to accept H_0 is $1-\alpha$, and we will call this our *acceptance confidence level*. On the other hand, if H_1 is true we know that the probability of accepting H_0 is β. Since this would be the wrong decision, our confidence in our rejection of H_1 is $1-\beta$, which we will refer to as our *exclusion confidence level*. A similar argument can be made when we accept H_1 instead of H_0. The overall situation is summarized in Table 1.

Acceptance Confidence Levels:
I. *When H_0 is accepted, our confidence that H_0 is true equals $1-\alpha$.*
II. *When H_1 is accepted, our confidence that H_1 is true equals $1-\beta$.*
Exclusion Confidence Levels:
II. *When H_0 is accepted, our confidence that H_1 is false equals $1-\beta$.*
I. *When H_1 is accepted, our confidence that H_0 is false equals $1-\alpha$.*

Table 1: Confidence levels for a frequentist test of H_0 versus H_1. The roman numerals in the first column refer to the frequentist error types from which the confidence levels are derived.

A somewhat unsatisfactory aspect of frequentist confidence levels is that there are

two of them for each outcome of the test. When accepting H_0 for example, it would seem more natural if our confidence that H_0 is true were always equal to our confidence that H_1 is false, since H_0 and H_1 are the only contending alternatives. Unfortunately, frequentist confidence levels are not conditioned on the action taken, which is known and unique, but on the true state of nature, which is unknown and bivalent. From an evidential point of view however, it is possible to argue that one of the confidence levels is more relevant than the other. If the data lead us to accept H_0 for example, then this only constitutes positive evidence in favor of H_0 if there is a substantial probability of rejecting H_0 when H_1 is true. [10] It is this probability, $1-\beta$, that constitutes the evidentially relevant confidence level when accepting H_0. Similarly, $1-\alpha$ is the evidentially relevant confidence level when accepting H_1. These are the exclusion confidence levels of the test and we take a closer look at their meaning in the next two subsections.

3.1 Coverage interpretation of exclusion confidence levels

Suppose that we do a test of H_0 versus H_1' (Problem 2 above), and find H_0 to be accepted. Our confidence that H_1' is false is then given by the exclusion confidence level function $1-\beta(\alpha,\mu)$. Interestingly, it is possible to associate $1-\beta(\alpha,\mu)$ with the coverage of an interval. Indeed, the alternative hypothesis in Problem 2 specifies a range of values for μ rather than a single particular value. This makes it meaningful to calculate confidence intervals for μ. For example, a γ-confidence level lower limit μ_L on μ satisfies the equation:

$$\int_{x_{obs}}^{\infty} f(x\,|\,\mu_L)\,dx \;=\; 1-\gamma. \tag{3.1}$$

Now, if our test led us to accept $H_0 : \mu = 1$, we may be interested in determining the confidence level of the largest interval that contains $\mu = 1$ but not some μ' in the alternative hypothesis ($\mu' < 1$). Since any value of μ under H_1' is lower than the value of μ under H_0, this largest interval must have the form of a lower-limit μ'. According to the above equation, its coverage is:

$$\gamma \;=\; 1 \;-\; \int_{x_{obs}}^{\infty} f(x\,|\,\mu')\,dx \;>\; 1 \;-\; \int_{x_\alpha}^{\infty} f(x\,|\,\mu')\,dx \;=\; 1 \;-\; \beta(\alpha,\mu'), \tag{3.2}$$

where the inequality follows from $x_{obs} > x_\alpha$, which must be true since we accepted H_0. This result leads to the following coverage interpretation of $1-\beta(\alpha,\mu')$:

Coverage Interpretation 1 (when accepting H_0)
In the event that $H_0 : \mu = 1$ is accepted, the largest confidence interval containing $\mu = 1$ but not $\mu = \mu' < 1$ has coverage at least $1-\beta(\alpha,\mu')$.

This interpretation has an interesting application to sensitivity calculations in searches for new physics. Whenever one fails to reject the standard model hypothesis H_0 in such a search, there is interest in determining what region of parameter space in the new physics model can be excluded at some confidence level γ or higher. How does

one characterize the sensitivity of this procedure prior to the measurement? One way is to report the set S_γ of all μ values for which $1-\beta(\alpha,\mu) \geq \gamma$. Then it follows from the definition of $\beta(\alpha,\mu)$ that if the true value of μ is in S_γ, the probability of making a discovery at the α significance level is at least γ, and it follows from the above coverage interpretation that if one fails to make a discovery, any value of μ inside S_γ will be excluded with a confidence level of at least γ. The set S_γ can therefore be called the sensitivity set of the measurement. Further details are provided in Ref. [11].

A similar coverage interpretation can be found whenever the test rejects H_0. In this case our confidence that H_0 is false equals $1-\alpha$. In the context of Problem 2 we can construct confidence intervals on μ, and since we rejected H_0, we will be particularly interested in intervals that exclude $\mu = 1$, i.e. upper limits. A γ-confidence level upper limit μ_U on μ satisfies:

$$\int_{-\infty}^{x_{obs}} f(x\,|\,\mu_U)\,dx \;=\; 1-\gamma. \tag{3.3}$$

Conversely, the confidence level of a given upper limit μ' is given by:

$$\gamma \;=\; 1 \;-\; \int_{-\infty}^{x_{obs}} f(x\,|\,\mu')\,dx \;\geq\; 1 \;-\; \int_{-\infty}^{x_\alpha} f(x\,|\,\mu')\,dx \;=\; 1 \;-\; \beta(\alpha,\mu'), \tag{3.4}$$

where this time we used the fact that $x_{obs} \leq x_\alpha$ to derive a lower bound. In the limit $\mu' \to 1$ this yields $\gamma \geq 1-\alpha$ (cfr. equation 1.2), and we have:

Coverage Interpretation 2 (when rejecting H_0)
In the event that $H_0 : \mu = 1$ is rejected, the largest confidence interval containing μ values less than 1, but not $\mu = 1$, has coverage at least $1-\alpha$.

Note the difference between this interpretation and the previous one. In Interpretation 1 the rejected hypothesis is composite ($\mu < 1$), and one can construct many "largest" confidence intervals, one for each value of μ under the alternative hypothesis. In Interpretation 2 the rejected hypothesis is simple ($\mu = 1$), and therefore there is only one largest confidence interval.

One may object to the above coverage interpretations that, when the parameter space is bounded, they seem to constitute uninteresting statements about trivial confidence limits. Indeed, if we look at Interpretation 2 for example, it states that when H_0 is rejected we are interested in the interval $0 \leq \mu < 1$, and that this interval has coverage at least $1-\alpha$. However, since μ is bounded between 0 and 1, we already know that the interval $0 \leq \mu \leq 1$ has 100% coverage. The only difference between these two intervals is that the former does not contain the point $\mu = 1$, a fact that is irrelevant in the standard Neyman-Pearson construction of upper limits [12, section II.B]. This suggests that the confidence limits must be constructed more carefully if one is to obtain meaningful coverage interpretations. With a bounded parameter space, a simple and general way to proceed is described in Ref.[13]. Instead of basing the construction of confidence limits directly on the data x, one uses an estimator $\hat{\mu}$ that respects the physical boundaries:

$$\hat{\mu} \;=\; \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases} \tag{3.5}$$

The probability distribution of $\hat{\mu}$ has finite probability masses at $\hat{\mu} = 0$ and $\hat{\mu} = 1$ and is continuous between these boundaries. If we now apply the Neyman-Pearson construction to a plot of μ versus $\hat{\mu}$, we find that the parameter values $\mu = 0$ and $\mu = 1$ each acquire a "lump" of confidence, in such a way that the two intervals $0 < \mu \leq 1$ and $0 \leq \mu < 1$ have coverage strictly less than 1. Thus, with this approach to parameter boundaries the coverage interpretations become non-trivial again.

3.2 Bayesian interpretation of exclusion confidence levels

Confidence levels are introduced by frequentists to characterize the reliability of decisions made when testing hypotheses, and care should be taken not to confuse them with the Bayesian posterior probabilities of the hypotheses. Nevertheless it is interesting to ask whether there are any conditions under which such confusion would be excusable. Supposing that a Bayesian has assigned prior probabilities π_0 and $\pi_1 \equiv 1 - \pi_0$ to H_0 and H_1 respectively, how should she choose α in order to match the posterior probability of H_0 with a frequentist confidence level? First we need to figure out which confidence level to match, the acceptance one or the exclusion one? To answer this, remember that even though a frequentist anticipates two types of error *before* a test, he can only have committed one type of error *after* the test. For example, if H_0 was accepted, then only the possibility of error resulting from H_1 being true remains, and the corresponding frequentist probability of error is β. In the Bayesian approach, if we accept H_0, then the posterior probability of error equals the posterior probability of H_1. Hence, from a posterior point of view the only sensible matching that can be sought is between the posterior probability of H_1 and β, or between the posterior probability of H_0 and the exclusion confidence level $1 - \beta$. Similarly, when H_1 is accepted one would like to match the posterior probability of H_1 with the exclusion confidence level $1 - \alpha$.

There is one more assumption we need to make in order to match the Bayesian and frequentist descriptions of the test, and that is that the measurement result is binary, only telling us whether H_0 was accepted or rejected. Any additional information about the "strength of evidence" in favor of H_0 or H_1 must be suppressed, otherwise the Bayesian statistician will have an unsurpassable advantage over the frequentist. With this assumption, Bayes' theorem yields, for the posterior probability of H_i given that H_i was accepted:

$$\begin{aligned}
\mathbb{P}\Big[H_0\Big|H_0 \text{ accepted}\Big] &= \frac{\mathbb{P}\Big[H_0 \text{ accepted}\Big|H_0\Big]\,\mathbb{P}\Big[H_0\Big]}{\mathbb{P}\Big[H_0 \text{ accepted}\Big|H_0\Big]\,\mathbb{P}\Big[H_0\Big] \;+\; \mathbb{P}\Big[H_0 \text{ accepted}\Big|H_1\Big]\,\mathbb{P}\Big[H_1\Big]} \\
&= \frac{(1-\alpha)\,\pi_0}{(1-\alpha)\,\pi_0 \;+\; \beta\,\pi_1}, \qquad (3.6) \\
\mathbb{P}\Big[H_1\Big|H_1 \text{ accepted}\Big] &= \frac{(1-\beta)\,\pi_1}{(1-\beta)\,\pi_1 \;+\; \alpha\,\pi_0}. \qquad (3.7)
\end{aligned}$$

The condition that yields the desired matching of Bayesian and frequentist confidence levels turns out to be "maximin": α must be such that it maximizes the minimum prior

probability of a successful test outcome. Here, "successful outcome" means correctly accepting H_0 or H_1, and depends on the acceptance probability as well as on the prior probability of the relevant hypothesis. Thus we need to maximize:

$$\min\{(1-\alpha)\pi_0,\, (1-\beta)\pi_1\}.$$

In general, a decrease in α results in an increase in β, and vice-versa. Maximizing the minimum therefore leads to:

$$(1-\alpha)\,\pi_0 \;=\; (1-\beta)\,\pi_1. \tag{3.8}$$

Substituting this result in equations (3.6) and (3.7) yields

$$\mathbb{P}\Big[H_0\Big|H_0 \text{ accepted}\Big] \;=\; 1-\beta \quad \text{and} \quad \mathbb{P}\Big[H_1\Big|H_1 \text{ accepted}\Big] \;=\; 1-\alpha,$$

as desired. We emphasize that this agreement between Bayesian and unconditional frequentist confidence levels can only be achieved for measurements that reveal nothing more than whether the data lies in the critical region. Information about the degree of "extremeness" of the data with respect to one or the other hypothesis is presumed unavailable. It should also be pointed out that the maximin matching condition is not the usual procedure a Bayesian would follow to choose α. Indeed, a more standard Bayesian approach is to minimize the risk of an incorrect decision, as will be described in section 4. Nevertheless, the maximin condition will reappear in section 5.3, where it will help reconcile Bayesian and *conditional* frequentist confidence levels.

4 Choice of rejection threshold for unconditional testing

A rather arbitrary aspect of frequentist tests is the choice of α. Values commonly found in the statistics literature include 1% and 5%, whereas standards in high energy physics are typically much more stringent ($\alpha = 1.3 \times 10^{-3}$ for evidence and $\alpha = 2.8 \times 10^{-7}$ for discovery). For low resolution measurements such as the top quark charge analysis, small values of α result in low power $1-\beta$, and therefore in a low confidence level of the maximum exclusion interval(s) when the standard model is accepted (Coverage Interpretation 1).

There is another reason for being careful in choosing α. Suppose for simplicity that the distribution of x under H_0 is symmetric with mean $\mu_0 = 1$, and that its distribution under H_1 is identical in shape but has mean $\mu_1 = 0$. Suppose also that we choose α in such a way that $x_\alpha < 1/2$, and that we subsequently observe $x = x_{obs}$ with $x_\alpha < x_{obs} < 1/2$. In this case we will be accepting $H_0 : \mu = 1$ even though x_{obs} is closer to 0 than to 1. This is clearly inconsistent from an evidential point of view, independently of the fact that the frequentist Type-I error rate is still properly characterized by α. For a different choice of α it is similarly possible to have $1/2 < x_{obs} \leq x_\alpha$, forcing one to reject H_0 even though x_{obs} is farther from 0 than from 1.

To describe this conflict between evidence and error probability more generally, consider the likelihood ratio in favor of H_0:

$$B_{01} \;\equiv\; \frac{f(x_{obs}\,|\,\mu=\mu_0)}{f(x_{obs}\,|\,\mu=\mu_1)}. \tag{4.1}$$

The notation B_{01} indicates that for problems without systematic uncertainties the likelihood ratio coincides with the Bayes factor, which will be defined in all generality in section 6. It is now easy to see that the above conflict arises whenever

$$(p_0 \le \alpha \text{ and } B_{01} > 1) \quad \text{or} \quad (p_0 > \alpha \text{ and } B_{01} \le 1), \tag{4.2}$$

p_0 being the p value under H_0. One way to avoid this conflict is to use B_{01} as test statistic and set $\alpha = \alpha^\star$, where $\alpha^\star$ is the probability that $B_{01} \le 1$ under H_0.

An interesting special case occurs when the testing problem satisfies the condition of likelihood ratio symmetry (LRS). Under this condition, the distribution of B_{01} under H_0 equals that of $B_{10} \equiv 1/B_{01}$ under H_1. Therefore:

$$\alpha^\star \;\equiv\; \mathbb{P}(B_{01} \le 1\,|\,H_0) \;=\; \mathbb{P}(B_{10} \le 1\,|\,H_1) \;=\; \mathbb{P}(B_{01} \ge 1\,|\,H_1) \;\equiv\; \beta^\star. \tag{4.3}$$

Thus, LRS tests that avoid conflict (4.2) are minimax.

The proposed solution to conflict (4.2) completely eliminates the freedom of choosing α. However, this freedom can be restored if one is willing to entertain evidential concepts from Bayesian statistics. A Bayesian will start by assigning prior probabilities π_i to the hypotheses H_i, and will then reject H_0 whenever the posterior odds in favor of that hypothesis are less than 1:

$$B_{01}\,\frac{\pi_0}{\pi_1} \;<\; 1. \tag{4.4}$$

A conflict between error probability and evidence now occurs whenever:

$$(p_0 \le \alpha \text{ and } B_{01}\frac{\pi_0}{\pi_1} > 1) \quad \text{or} \quad (p_0 > \alpha \text{ and } B_{01}\frac{\pi_0}{\pi_1} < 1), \tag{4.5}$$

and can be avoided by using B_{01} as test statistic and setting $\alpha = \alpha^{\star\star}$, the probability that $B_{01} \le \pi_1/\pi_0$ under H_0. This way, any value of α can be obtained by a suitable choice of π_0. In particular, $\alpha^{\star\star} = \alpha^\star$ corresponds to $\pi_0 = \pi_1 = 1/2$.

The argument for setting $\alpha = \alpha^{\star\star}$ can also be derived from Bayesian risk considerations. First, define $G_i(y)$ to be the cumulative probability distribution of $y \equiv B_{01}$ under H_i, so that:

$$\alpha \;=\; G_0(c) \quad \text{and} \quad \beta \;=\; 1 - G_1(c), \tag{4.6}$$

where c is the critical value of B_{01} for the test. Then, if the cost of incorrectly rejecting H_i is estimated to be ℓ_i, the Bayesian risk of an incorrect decision is:

$$R(c) \;\equiv\; \ell_0\,\alpha\,\pi_0 \;+\; \ell_1\,\beta\,\pi_1 \;=\; \ell_0\,G_0(c)\,\pi_0 \;+\; \ell_1\,\Big[1 - G_1(c)\Big]\,\pi_1. \tag{4.7}$$

A natural criterion is to choose c so as to minimize this risk. If $g_i(y)$ is the probability density function corresponding to $G_i(y)$, we have:

$$\frac{dR}{dc} \;=\; \ell_0\,g_0(c)\,\pi_0 \;-\; \ell_1\,g_1(c)\,\pi_1 \;=\; \Big(\ell_0\,\pi_0\,c - \ell_1\,\pi_1\Big)\,g_1(c), \tag{4.8}$$

where we used the property that $g_0(y) = y\, g_1(y)$, as proved in Appendix C.1, see equation (C.5). Equation (4.8) shows that the risk is minimized for

$$c \;=\; \frac{\ell_1\, \pi_1}{\ell_0\, \pi_0}. \tag{4.9}$$

Hence for equal costs, $\ell_0 = \ell_1$, we recover the $\alpha = \alpha^{\star\star}$ rule.

Example 2 (Gaussian approximation to the top charge analysis, continued) Using equation (4.1) with the p.d.f. of equation (2.1) yields, for the Bayes factor in favor of H_0:

$$B_{01} \;=\; e^{\frac{\Delta\mu}{\sigma^2}\,(x-\bar{\mu})}, \tag{4.10}$$

where $\Delta\mu \equiv \mu_0 - \mu_1$ and $\bar{\mu} \equiv (\mu_0 + \mu_1)/2$. The distributions of $y \equiv B_{01}$ under H_i, $i = 0, 1$, are log-normal:

$$g_i(y) \;=\; \frac{1}{\sqrt{2\pi}\, y\, \Delta\mu/\sigma}\, \exp\left[-\frac{1}{2}\left(\frac{\ln y \;+\; (2i-1)\Delta\mu^2/2\sigma^2}{\Delta\mu/\sigma}\right)^2\right], \tag{4.11}$$

$$G_i(y) \;\equiv\; \int_0^y g_i(t)\, dt \;=\; \frac{1}{2}\left[1 \;+\; \mathrm{erf}\left(\frac{\ln y \;+\; (2i-1)\Delta\mu^2/2\sigma^2}{\sqrt{2}\,\Delta\mu/\sigma}\right)\right]. \tag{4.12}$$

The median and mean expected Bayes factors are given by:

$$\mathrm{Median}(B_{01}\,|\,H_i) \;=\; e^{\frac{1}{2}(1-2i)\left(\frac{\Delta\mu}{\sigma}\right)^2}, \quad \mathrm{Mean}(B_{01}\,|\,H_i) \;=\; e^{(1-2i)\left(\frac{\Delta\mu}{\sigma}\right)^2}. \tag{4.13}$$

Using the approximate top charge analysis values $\Delta\mu = 1$ and $\sigma \approx 0.38$ yields a median expected B_{01} of 31.9 and a mean expected B_{01} of 1017.6, both under H_0. If σ is reduced by a factor of $\sqrt{2}$, the median becomes 952.1 and the mean 906568. As the measurement resolution improves, the skewness of the distribution increases, making it more likely that a large B_{01} will be observed if H_0 is true.

The value of α that avoids conflict (4.2) is:

$$\alpha^\star \;=\; G_0(1) \;=\; \frac{1}{2}\left[1 \;-\; \mathrm{erf}\left(\frac{\Delta\mu}{2\sqrt{2}\sigma}\right)\right]. \tag{4.14}$$

Note that the LRS condition is satisfied in this problem, so that $\alpha^\star$ is minimax and $\beta^\star = \alpha^\star$. The critical boundary in x space that corresponds to $\alpha^\star$ is obtained by setting $B_{01} = 1$ in eq. (4.10) and solving for x:

$$x_{\alpha^\star} \;=\; \bar{\mu}. \tag{4.15}$$

For a given choice of the prior probability π_0 of H_0, the value of α that avoids conflict (4.5) is equal to the probability of $B_{01} \le \pi_1/\pi_0$ under H_0:

$$\alpha^{\star\star} \;=\; G_0\left(\frac{\pi_1}{\pi_0}\right) \;=\; \frac{1}{2}\left[1 \;-\; \mathrm{erf}\left(\frac{\Delta\mu}{2\sqrt{2}\,\sigma} \;+\; \frac{\sigma}{\sqrt{2}\,\Delta\mu}\,\ln\frac{\pi_0}{\pi_1}\right)\right], \tag{4.16}$$

π_0/π_1	Minimum Bayes Risk α	β	Bayes/Freq. Matching α	β
1	0.0941	0.0941	0.0941	0.0941
10	1.42×10^{-2}	0.3297	0.9000	4.55×10^{-5}
100	1.09×10^{-3}	0.6679	0.9900	3.56×10^{-7}
1000	4.06×10^{-5}	0.9048	0.9990	5.27×10^{-9}
10000	7.33×10^{-7}	0.9855	0.9999	1.08×10^{-10}

Table 2: Gaussian approximation to the top charge analysis ($\Delta\mu = 1$ and $\sigma = 0.38$): frequentist error rates α and β corresponding to given prior odds π_0/π_1 for the criterion of minimum Bayes risk and that of matching Bayes and frequentist confidence levels. Note that under inversion of the prior odds, the α and β values are simply interchanged. For $\pi_0/\pi_1 = 0.1$ for example, the criterion of minimum Bayes risk yields $\alpha = 0.3297$ and $\beta = 1.42 \times 10^{-2}$.

and corresponds to the critical boundary in x:

$$x_{\alpha^{\star\star}} \;=\; \bar{\mu} \;-\; \frac{\sigma^2}{\Delta\mu} \ln\left(\frac{\pi_0}{\pi_1}\right). \tag{4.17}$$

Thus, if our prior belief in H_0 is very strong ($\pi_0 \gg \pi_1$), it will be quite possible to accept H_0 even though x_{obs} is closer to μ_1 than to μ_0. Using the same values for $\Delta\mu$ and σ as before, Table 2 lists the α and β values corresponding to given prior odds π_0/π_1. Two criteria are considered: minimum Bayes risk (equation 4.16) and matching of frequentist and Bayesian confidence levels (equation 3.8). Not surprisingly, these two criteria give very different results. Someone with a strong prior belief in H_0 and wanting to minimize the risk of an incorrect decision will require strong evidence before rejecting H_0; this person will therefore want to choose a small α, but must be willing to incur a rather large β. On the other hand, if the criterion is to match Bayesian and frequentist confidence levels, strong prior belief in H_0 that accords with the data will yield a large Bayesian posterior confidence level. Such data will lead to acceptance of H_0, in which case the frequentist exclusion confidence level is $1 - \beta$, which is large if α is large. Thus, large π_0 is now associated with *large* α. The table clearly indicates the opposite effects of the two criteria as π_0/π_1 increases. Only for $\pi_0 = \pi_1$ is there full agreement between them. ■

Discovery claims in high energy physics are typically based on single, significant measurements. This indicates predominant interest in an evidential interpretation of the hypothesis tests we perform, rather than in an error-rate interpretation. As the above discussion illustrates, the freedom of choosing the discovery threshold α in this context requires the elicitation of prior odds on the tested hypotheses. Further discussion of this aspect of hypothesis testing can be found in Ref. [14].

5 Unified Bayes/conditional frequentist testing

The two confidence levels discussed in section 3 are unconditional in the sense that they are known before looking at the data: they measure the performance of the test over the long run, but do not quantify what the experimenter has learned from the observed data after a single measurement. For example, whether the observations fall right on top of the critical boundary or deep into the critical region, we will reject the null hypothesis and report the same unconditional confidence levels, even though the strength of evidence may be very different in the two cases. Within the frequentist paradigm, this deficiency can be addressed by conditioning on an appropriate aspect of the data.

To simplify the argument, we only consider Problem 1 from section 3 and assume for the moment that there are no nuisance parameters. In this case the Bayes factor B_{01} defined in equation (4.1) is actually a likelihood ratio and is in one-to-one correspondence with the variable x (the measured fraction of charge +2/3 quarks). The advantage of working with Bayes factors is that this simplifies generalization to the more complex problem that will be considered in section 6. Let $g_i(y)$ be the probability density function of $y \equiv B_{01}$ under H_i, and $G_i(y)$ the corresponding cumulative distribution function. If we observe $y = y_{obs}$, we can calculate two p values in this problem:

$$\begin{aligned} p_0 &= G_0(y_{obs}), \\ p_1 &= 1 \; - \; G_1(y_{obs}). \end{aligned} \tag{5.1}$$

Given the fact that small values of B_{01} disfavor H_0, the above definitions imply that a small p_0 is evidence against H_0 and a small p_1 is evidence against H_1. Let c be the critical value of the test in Bayes factor space. The unconditional Type-I and Type-II error probabilities are then:

$$\begin{aligned} \alpha &= G_0(c), \\ \beta &= 1 \; - \; G_1(c), \end{aligned} \tag{5.2}$$

and we reject H_0 if $y_{obs} \leq c$, i.e. if $p_0 \leq \alpha$. The first step in the construction of a conditional test is to find a statistic that represents the strength of evidence exhibited by the data against H_0 or H_1. By conditioning the test on this statistic it will then be possible to derive frequentist error rates that reflect the evidential content of the data.

5.1 Definition of the conditioning statistic

From a frequentist point of view it seems quite natural that the required conditioning statistic should involve a combination of p_0, p_1, α, and β. When $\alpha = \beta$ a plausible choice of statistic is $\max\{p_0, p_1\}$, since this is neutral with respect to both hypotheses when they are considered on the same footing. However, when the error rates are unequal, the criterion for rejecting H_0 will be different from that for rejecting H_1, and the p values will need to be adjusted in order to represent the same evidence against H_0 and H_1. Some simple possibilities are to replace p_0 and p_1 by p_0/α and p_1/β, or by $(1-p_0)/(1-\alpha)$ and $(1-p_1)/(1-\beta)$. By basing the conditioning statistic on the latter

option we will be able to reconcile Bayesian and frequentist inference in section 5.3; we therefore define:

$$Q \;\equiv\; \min\left\{\frac{1-p_0}{1-\alpha},\, \frac{1-p_1}{1-\beta}\right\} \;=\; \min\left\{\frac{1-G_0(y_{obs})}{1-\alpha},\, \frac{G_1(y_{obs})}{1-\beta}\right\}. \tag{5.3}$$

This definition is illustrated in Figure 2. The idea of Ref. [15] is to calculate frequentist error probabilities (or confidence levels) that are conditional on Q and therefore take the observed evidence into account. In general the result of conditioning does not change when a monotonic transformation is applied to the conditioning statistic, since subsets of data points that are associated to fixed values of the statistic are unchanged by such a transformation. For convenience we will perform the transformation $\mathcal{T}:\; Q \to G_1^{-1}[(1-\beta)\,Q]$; this transformation maps $(1-p_1)/(1-\beta)$ into the Bayes factor $y = B_{01}$, and $(1-p_0)/(1-\alpha)$ into $\psi^{-1}(y)$, where ψ^{-1} is the inverse of the function [3]

$$\psi(y) \;=\; G_0^{-1}\Big[1 - G_1(y)/\rho\Big], \quad \text{with} \quad \rho \;\equiv\; \frac{1-\beta}{1-\alpha}. \tag{5.4}$$

Since $\mathcal{T}$ is monotonically *increasing*, minima remain minima, and Q is mapped into

$$S(y) \;=\; \min\Big\{\psi^{-1}(y),\, y\Big\}. \tag{5.5}$$

As shown in Figure 3, for every $s < 1$ the equation $s = S(y)$ has two solutions in y, one smaller than c and the other larger. If y is the smaller solution, the other one can be written as $\psi(y)$ [4] and we have:

$$\frac{\mathbb{P}(B_{01} \le y \,|\, H_1)}{\mathbb{P}(B_{01} > \psi(y) \,|\, H_0)} \;=\; \frac{G_1(y)}{1 - G_0(\psi(y))} \;=\; \rho, \tag{5.6}$$

where the last equality follows directly from the definition of $\psi(y)$. This shows that the two solutions of $s = S(y)$ have proportional tail probabilities, and the proportionality constant is independent of s. For this reason the statistic S is sometimes called the proportional tail statistic. [16] In the space of Bayes factors $S(y)$ induces a partition whose elements can be written as:

$$\mathcal{Y}_s \;\equiv\; \Big\{y:\; S(y) = s\Big\} \;=\; \Big\{y:\; y = s \text{ or } y = \psi(s)\Big\}, \quad \text{where} \quad s \in [0, c]. \tag{5.7}$$

The statistic S takes a particularly simple form in minimax tests that are likelihood ratio symmetric (LRS). The LRS condition was defined in section 4. It implies that, for all $y > 0$:

$$\begin{aligned} G_0(y) \;&=\; \mathbb{P}\Big[B_{01} \le y \,\Big|\, H_0\Big] \;=\; \mathbb{P}\Big[B_{10} \le y \,\Big|\, H_1\Big] \\ &=\; 1 \;-\; \mathbb{P}\Big[B_{01} \le 1/y \,\Big|\, H_1\Big] \;=\; 1 \;-\; G_1(1/y). \end{aligned} \tag{5.8}$$

[3] The meaning of the function ψ is that if y is less than c and constitutes evidence of magnitude Q against H_0, then $\psi(y)$ is greater than c and constitutes evidence of the same magnitude Q against H_1.

[4] To see this, consider that $S(y) = y$ implies by definition that $y < \psi^{-1}(y)$. Applying the function ψ to both sides of this inequality, and using the fact that $\psi(y)$ decreases with y, we obtain $\psi(y) > y$. Therefore $S(\psi(y)) = \min\{y, \psi(y)\} = y$, showing that $\psi(y)$ is the other solution and the larger one.

Thus, in situations where minimaxity and LRS are simultaneously satisfied, equation (5.4) becomes $\psi(y) = 1/y$ and S becomes the smaller of the Bayes factor in favor of H_0 and the Bayes factor in favor of H_1. The partition elements (5.7) become:

$$\mathcal{Y}_s \;=\; \Big\{y:\; y = s \text{ or } y = 1/s\Big\}, \quad \text{where} \quad s \in [0, c] \quad \text{(if minimax and LRS)}. \tag{5.9}$$

5.2 Calculation of the conditional error rates

Having identified a suitable conditioning statistic S that reflects the evidential content of the data, our next task is to calculate frequentist error rates that are conditioned on the observed value of S. The meaning of these error rates is as follows. Suppose we take data and observe a Bayes factor in favor of H_0 equal to y_{obs}. The conditioning statistic then has the value $s_{obs} = S(y_{obs})$. Our observation automatically selects the partition element $\mathcal{Y}_{s_{obs}}$ that contains both s_{obs} and $\psi(s_{obs})$. Although in terms of Bayes factors $\mathcal{Y}_{s_{obs}}$ only contains two distinct elements, in terms of datasets the cardinality of $\mathcal{Y}_{s_{obs}}$ may be much larger, since the mapping from dataset to Bayes factor is many-to-one. Consider then the subensemble of datasets for which the observed Bayes factor belongs to $\mathcal{Y}_{s_{obs}}$. By construction this subensemble contains the dataset actually observed in our measurement, plus all the datasets, not actually observed, that have the same evidential power as our data, as measured by the statistic S. The questions we now wish to answer are: if we repeatedly perform our test on the datasets *of this subensemble*, and H_0 is true, what is the frequentist probability $\tilde{\alpha}(s_{obs})$ that we will incorrectly reject H_0; and if H_1 is true, what is the frequentist probability $\tilde{\beta}(s_{obs})$ that we will incorrectly reject H_1? Note that these error probabilities depend on s_{obs}, in contrast with the unconditional error rates α and β introduced in section 1.1.

Appendix C.1 describes the calculation of conditional frequentist error rates. Dropping the subscript $_{obs}$ to simplify the notation, the result is:

$$\begin{aligned}
\tilde{\alpha}(s) &\equiv \mathbb{P}\Big[\text{Rejecting } H_0 \,\Big|\, H_0 \text{ true and } B_{01} \in \mathcal{Y}_s\Big] \;=\; \frac{\rho\, s}{\rho\, s + 1} \;=\; \frac{\rho\, B_{01}}{\rho\, B_{01} + 1}, \\
\tilde{\beta}(s) &\equiv \mathbb{P}\Big[\text{Accepting } H_0 \,\Big|\, H_1 \text{ true and } B_{01} \in \mathcal{Y}_s\Big] \;=\; \frac{1}{\rho\, \psi(s) + 1} \;=\; \frac{1}{\rho\, B_{01} + 1}
\end{aligned} \tag{5.10}$$

(remember that $B_{01} = s$ if $B_{01} \le c$, and $B_{01} = \psi(s)$ if $B_{01} > c$). The conditional frequentist test is then:

$$\begin{cases} \text{If } p_0 \le \alpha, \text{ reject } H_0 \text{ and report conditional error probability } \tilde{\alpha}(s); \\ \\ \text{If } p_0 > \alpha, \text{ accept } H_0 \text{ and report conditional error probability } \tilde{\beta}(s). \end{cases} \tag{5.11}$$

Figure 4 compares the conditional error rates to the unconditional ones for our simplified Gaussian model of the top charge analysis. As expected, the conditional error rates decrease on each side of the critical value, reflecting the change in strength of evidence as the Bayes factor y increasingly favors one hypothesis over the other. Note also that the conditional rates exceed the unconditional ones near the critical value. This is consistent with the fact that, under a given hypothesis, the *expectation* of the conditional error rate equals the unconditional one [15]. In other words, the conditional test divides up the unconditional error probabilities among the various partitions.

5.3 Bayesian interpretation

A remarkable property of the conditional frequentist test (5.11) is that it can be given a Bayesian interpretation [15]. All that is needed is a "matching" condition between the prior probabilities π_i of H_i and the unconditional frequentist error rates (α, β). It turns out that the required condition is the very same maximin condition that was found to work for *un*conditional testing in section 3.2, namely equation (3.8). In terms of the parameter ρ this is:

$$\rho \;\equiv\; \frac{1-\beta}{1-\alpha} \;=\; \frac{\pi_0}{\pi_1}. \tag{5.12}$$

Substituting this result in the expressions for the conditional frequentist error probabilities yields:

$$\begin{aligned} \tilde{\alpha}(s) &= \frac{\pi_0\, B_{01}}{\pi_0\, B_{01} \;+\; \pi_1} \;=\; \frac{\pi_0\, f(x_{obs}\,|\,\mu_0)}{\pi_0\, f(x_{obs}\,|\,\mu_0) \;+\; \pi_1\, f(x_{obs}\,|\,\mu_1)}, \\ \tilde{\beta}(s) &= \frac{\pi_1}{\pi_0\, B_{01} \;+\; \pi_1} \;=\; \frac{\pi_1\, f(x_{obs}\,|\,\mu_1)}{\pi_0\, f(x_{obs}\,|\,\mu_0) \;+\; \pi_1\, f(x_{obs}\,|\,\mu_1)}, \end{aligned} \tag{5.13}$$

where we used equation (4.1) to express the Bayes factor B_{01} in terms of the original p.d.f. $f(x\,|\,\mu)$. The right-hand sides of the above equations are the Bayesian posterior probabilities of H_0 and H_1. As a result, if the test rejects H_0 for example, the Bayesian's posterior belief in H_0 will exactly match the conditional frequentist Type-I error probability. The insidious but common shift in meaning between "the probability that H_0 was incorrectly rejected" and "the probability that H_0 is true" no longer needs to be a source of concern.

An important comment is that conditional frequentist tests can be constructed with conditioning statistics other than the $S(y)$ defined in (5.5). However, only $S(y)$, or a one-to-one transformation of it, leads to conditional error rates that have the proper structure to allow a Bayesian interpretation[5].

5.4 Choice of rejection threshold for conditional testing

So far we have achieved one of our main goals, which was to provide confidence levels that take the evidential strength of the data into account. Furthermore, these confidence levels are trivially easy to compute from the Bayes factor: there is no need to calculate the conditioning statistic itself nor its distribution. In fact, modern experimenters will often calculate a Bayes factor to complement the perspective of a frequentist analysis; once this is done, conditional frequentist error rates are only one step away with the help of equation (5.10). Finally, the equivalence between these frequentist error rates and Bayesian posterior probabilities provides an additional, reassuring layer of interpretation.

[5] Statisticians generally recommend to condition on a statistic that is ancillary, i.e. whose distribution is independent of the parameter of interest. Indeed, inferences obtained by conditioning on a non-ancillary statistic may ignore some relevant information contained in the data. Although $S(y)$ is ancillary for LRS problems with $\alpha = \beta$, it is not universally so, and conditioning on it in general is only justified by the fact that it leads to a frequentist test that is simultaneously Bayesian.

As formulated, the conditional frequentist test (5.11) still depends on the choice of a rejection threshold α. This is illustrated in Fig. 5, which shows conditional frequentist error rates for three choices of α. A closer examination of this plot reveals a somewhat unappealing feature. Suppose for example that we set $\alpha = 0.01$, and observe a Bayes factor larger than the corresponding critical value but smaller than 1. In this case we will accept the null hypothesis, even though this will force us to report a conditional error rate that is much larger than the one we would have reported had we chosen $\alpha = 0.094$ and correspondingly rejected the null. Given the choice between accepting and rejecting a hypothesis, it seems unsatisfactory to make the decision that has the larger probability of being wrong. In the next subsections we examine some possible solutions to this problem.

5.4.1 Adding a loss structure

The first solution is to introduce an asymmetric cost function and emphasize the *risk* of a decision instead of its probability of being wrong. For example, if the cost of incorrectly rejecting H_0 is considered higher than that of incorrectly accepting H_0, one will naturally tolerate a higher Type-II error rate. To formalize this idea, define ℓ_i, a positive number, to be the loss incurred when incorrectly rejecting H_i, and assume that the loss is zero when a correct decision is made. The frequentist risk is then the expected loss when a given hypothesis is true: it is $\ell_0\alpha$ when H_0 is true and $\ell_1\beta$ when H_1 is true, where α and β are conditional or unconditional error rates, depending on the context. Next, let the critical boundary in Bayes factor space be the solution of

$$c \;=\; \frac{\ell_1}{\ell_0}\,\frac{1 - G_0(c)}{G_1(c)} \tag{5.14}$$

with respect to c. The modified test is then:

$$\begin{cases} \text{If } B_{01} \le c, \text{ reject } H_0 \text{ and report risk } \dfrac{\ell_0\,\rho\,B_{01}}{\rho\,B_{01}+1}; \\[2ex] \text{If } B_{01} > c, \text{ accept } H_0 \text{ and report risk } \dfrac{\ell_1}{\rho\,B_{01}+1}; \end{cases} \tag{5.15}$$

where "risk" now has simultaneous meaning as "conditional frequentist risk" and as "Bayes posterior risk". We leave it to the reader to verify that this test procedure always selects the decision with the lower risk (note that the equation for c is equivalent to $c = \ell_1/(\rho\ell_0)$).

5.4.2 Adding a no-decision region

Another possible solution is to add a "no-decision" region to the test, that would cover all the values of the test statistic B_{01} for which the reported conditional error probability is larger than 50%. For example, according to the test (5.11), when $B_{01} \le c$ we reject H_0 and report $\rho B_{01}/(\rho B_{01}+1)$ as conditional error probability. This probability will be larger than 50% if $B_{01} > 1/\rho$, so that the interval $]1/\rho,\, c[$ should be included in the

no-decision region (NDR) whenever $1/\rho < c$. In addition, we must make sure that the NDR respects the partition structure induced by the statistic S. Suppose for example that y and y' both belong to the same partition element, with y being in the rejection region and y' in the NDR. Observing y would then cause us to reject H_0, but our reported conditional error probability would be incorrect since it would be based on the wrong assumption that we would have accepted H_0 if we had observed y' instead. The solution is to arrange for the NDR to contain only *complete* partition elements, and this requires the NDR to be an interval of the form $]r, \psi(r)[$ for some r. In the above situation where $1/\rho < c$, one has $\psi(1/\rho) > \psi(c) = c$. Hence it is sufficient to extend the NDR from $]1/\rho,\, c[$ to $]1/\rho,\, \psi(1/\rho)[$. For a general formulation, define:

$$r \;\equiv\; \begin{cases} 1/\rho & \text{if } 1/\rho \le c, \\ \psi^{-1}(1/\rho) & \text{if } 1/\rho \ge c, \end{cases} \qquad \text{and} \qquad a \;\equiv\; \psi(r). \tag{5.16}$$

The modified test then replaces (5.11) with

$$\begin{cases} \text{If } B_{01} \le r, & \text{reject } H_0 \text{ and report conditional error probability } \tilde{\alpha}(s); \\ \text{If } r < B_{01} < a, & \text{make no decision;} \\ \text{If } B_{01} \ge a, & \text{accept } H_0 \text{ and report conditional error probability } \tilde{\beta}(s). \end{cases} \tag{5.17}$$

or, in terms of the p value under H_0:

$$\begin{cases} \text{If } p_0 \le G_0(r), & \text{reject } H_0 \text{ and report conditional error probability } \tilde{\alpha}(s); \\ \text{If } G_0(r) < p_0 < G_0(a), & \text{make no decision;} \\ \text{If } p_0 \ge G_0(a), & \text{accept } H_0 \text{ and report conditional error probability } \tilde{\beta}(s). \end{cases} \tag{5.18}$$

5.4.3 Choosing an empty no-decision region

The form of test 5.18 suggests a third way to avoid reporting error probabilities that are larger than 50%, namely by eliminating the no-decision region by selecting the test for which $a = r$. According to equation (5.16) this amounts to requiring $\psi(r) = r$, which is solved by $r = c$; the definition of r then implies that $c = 1/\rho$, or, using (5.2):

$$c \;=\; \frac{1 - G_0(c)}{G_1(c)}. \tag{5.19}$$

This condition can also be derived from equation (5.14) by equalizing the losses ℓ_0 and ℓ_1.

Tests with likelihood ratio symmetry satisfy $G_0(y) = 1 - G_1(1/y)$ for all $y > 0$ (see equation 5.8); setting $y = 1$ shows that $G_0(1) = 1 - G_1(1)$, so that condition (5.19) is satisfied for $c = 1$ in such tests. Equation (5.19) together with the choice $c = 1$ yields the minimax test ($\alpha = \beta$). Since the Gaussian approximation to the top charge

analysis is likelihood ratio symmetric, it is worth noting that in this case the statistic Q of equation (5.3) is equivalent to $\max\{p_0, p_1\}$ for conditioning purposes, and that the conditional frequentist test can be reformulated as follows:

$$\begin{cases} \text{If } p_0 \le p_1, \quad \text{reject } H_0 \text{ and report conditional error prob. } \tilde{\alpha}(s) = \dfrac{B_{01}}{B_{01}+1}, \\ \text{If } p_0 > p_1, \quad \text{accept } H_0 \text{ and report conditional error prob. } \tilde{\beta}(s) = \dfrac{1}{B_{01}+1}, \end{cases} \tag{5.20}$$

where we used the fact that $\rho = 1$ for minimax tests.

5.4.4 Two examples

In this subsection we present two examples to illustrate some of the ideas previously described.

Example 3 (Gaussian approximation to the top charge analysis, continued) Table 3 illustrates the conditional frequentist test for three choices of the unconditional Type-I error rate α in the Gaussian approximation to the top charge analysis. For very small values of α (such as the 5σ threshold in high energy physics), poor experimental resolution can lead to no-decision regions that are quite large compared to typical values of the Bayes factor B_{01}. As the resolution improves however, the no-decision region will tend to shrink. ■

The second example is that of a test that does not enjoy likelihood ratio symmetry.

Example 4 (Measurement of a lifetime) Assume that we make a single measurement from the exponential distribution:

$$f(t\,|\,\tau) \;=\; \frac{e^{-t/\tau}}{\tau} \qquad (t > 0), \tag{5.21}$$

and wish to test

$$H_0 : \tau \;=\; \tau_0 \quad \text{versus} \quad H_1 : \tau \;=\; \tau_1, \qquad \text{with } \tau_0 \;<\; \tau_1. \tag{5.22}$$

The Bayes factor in favor of H_0 is

$$B_{01} \;=\; \frac{1}{\gamma}\, e^{-t\,(1-\gamma)/\tau_0}, \qquad \text{where } \gamma \;\equiv\; \frac{\tau_0}{\tau_1} \;<\; 1. \tag{5.23}$$

We have that $0 < B_{01} < 1/\gamma$, and the distribution of $y \equiv B_{01}$ under H_i is given by

$$g_0(y) \;=\; \frac{\gamma^{\frac{1}{1-\gamma}}}{1-\gamma}\, y^{\frac{\gamma}{1-\gamma}}, \qquad G_0(y) \;\equiv\; \int_0^y g_0(t)\,dt \;=\; (\gamma\, y)^{\frac{1}{1-\gamma}}, \tag{5.24}$$

$$g_1(y) \;=\; \frac{\gamma^{\frac{1}{1-\gamma}}}{1-\gamma}\, y^{\frac{\gamma}{1-\gamma}-1}, \qquad G_1(y) \;\equiv\; \int_0^y g_1(t)\,dt \;=\; (\gamma\, y)^{\frac{\gamma}{1-\gamma}}. \tag{5.25}$$

α	c	β	NDR	P_0(NDR)	P_1(NDR)	CEP
$\Delta\mu = 1,\ \sigma = 0.38$						
2.8×10^{-7}	6.1×10^{-5}	0.99	$[2.1\times10^{-5}, 113.33]$	0.685	0.996	(0.780)
0.01	0.07	0.38	$[0.04, 1.60]$	0.122	0.386	0.0477
0.094	1.00	0.094	$\emptyset$	0	0	0.0304
$\Delta\mu = 1,\ \sigma = 0.27 \approx 0.38/\sqrt{2}$						
2.8×10^{-7}	8.5×10^{-6}	0.90	$[6.7\times10^{-6}, 10.35]$	0.111	0.908	0.0107
0.01	0.17	0.084	$[0.11, 1.08]$	0.027	0.075	0.00113
0.032	1.00	0.032	$\emptyset$	0	0	0.00105

Table 3: Characteristics of some conditional frequentist tests for the Gaussian approximation to the top charge analysis. For two values of the resolution σ and three choices of the unconditional Type-I error rate α, the table gives the critical value c of the Bayes factor B_{01}, the Type-II error rate β, the no-decision region NDR, the probabilities of the NDR under H_0 and H_1, and finally the conditional error probability when the median expected Bayes factor under H_0 is observed (cfr. Example 2). For $\sigma = 0.38$ this Bayes factor is 31.9 and corresponds to $p_0 = 0.5$ and $p_1 = 0.0042$. For $\sigma = 0.27$ it is 952.1 and corresponds to $p_0 = 0.5$ and $p_1 = 0.00011$. The median expected Bayes factor leads to acceptance of H_0 in all cases except the first ($\sigma = 0.38$, $\alpha = 2.8\times10^{-7}$), where B_{01} falls inside the no-decision region and the conditional probability of error is higher than 50%. Note that the last line in each sub-table corresponds to an equal-tailed (minimax) test.

Note that $1-G_1(1/y) = 1-(\gamma/y)^{\gamma/(1-\gamma)} \neq G_0(y)$, so that the likelihood ratio symmetry condition is not satisfied. As a consequence, when choosing the critical value c of B_{01}, three criteria that are equivalent under LRS will now yield three different results. The first criterion assumes equal prior probabilities for H_0 and H_1 and equal costs for incorrectly rejecting these hypotheses. Equation (4.9) then gives $c = 1$. The second criterion is to choose an equal-tailed test. For $\gamma = 1/2$ this will be achieved for $c \approx 1.236$, yielding $\alpha = \beta \approx 0.382$. The third criterion is to choose the value of c for which the conditional error probability is never larger than 50%. This value solves equation (5.19) and is here given by:

$$c = \frac{1}{\gamma^\gamma\,(1+\gamma)^{1-\gamma}}. \tag{5.26}$$

If $\gamma = 1/2$ we obtain $c \approx 1.155$. This is illustrated in Figure 6. ■

6 Systematic uncertainties

In Bayesian statistics the treatment of systematic uncertainties presents no particular problem as long as these uncertainties can be modeled by nuisance parameters ν with proper prior distributions $\varphi_i(\nu)$ under each hypothesis H_i, $i = 0, 1$. As this is usually the case in high energy physics, we will restrict the discussion accordingly. The first step is to average the probability density $f_i(x \,|\, \nu)$ of the data x under hypothesis H_i over the nuisance prior $\varphi_i(\nu)$ to obtain the marginal distribution:

$$f_i^\dagger(x) \;=\; \int_{H_i} f_i(x \,|\, \nu)\; \varphi_i(\nu)\; d\nu. \tag{6.1}$$

Here, the integral is over the nuisance parameter region that applies when H_i is true. Note that ν can be a *vector* of nuisance parameters, and that neither the number of its components nor their physical meaning needs to be the same under H_0 and H_1. The ν dependence of the original hypothesis test:

$$H_0 :\; x \;\sim\; f_0(x \,|\, \nu) \qquad \text{versus} \qquad H_1 :\; x \;\sim\; f_1(x \,|\, \nu), \tag{6.2}$$

can now be eliminated by considering the modified test:

$$H_0^\dagger :\; x \;\sim\; f_0^\dagger(x) \qquad \text{versus} \qquad H_1^\dagger :\; x \;\sim\; f_1^\dagger(x). \tag{6.3}$$

An important quantity is the Bayes factor in favor of H_0, which generalizes the likelihood ratio (4.1):

$$B_{01} \;=\; \frac{f_0^\dagger(x)}{f_1^\dagger(x)}. \tag{6.4}$$

Note that the Bayes factor for H_0 versus H_1 is the likelihood ratio for $H_0^\dagger$ versus $H_1^\dagger$. The Bayesian testing procedure is then:

$$\begin{cases} \text{If } B_{01} \le \dfrac{\pi_1}{\pi_0}, \text{ reject } H_0^\dagger \text{ and report } \dfrac{\pi_0 B_{01}}{\pi_0 B_{01} + \pi_1} \text{ as posterior probability of error;} \\[2ex] \text{If } B_{01} > \dfrac{\pi_1}{\pi_0}, \text{ accept } H_0^\dagger \text{ and report } \dfrac{\pi_1}{\pi_0 B_{01} + \pi_1} \text{ as posterior probability of error.} \end{cases} \tag{6.5}$$

An interesting question is whether these posterior probabilities of error can be given a conditional frequentist interpretation. An obvious starting point is to try to use the same conditioning statistic S (equation 5.5) as for the case without nuisance parameters. This can be done provided we replace the function ψ with a version $\psi^\dagger$ that does not depend on ν:

$$\psi^\dagger(y) \;=\; G_0^{\dagger\,-1}\Big[1 - G_1^\dagger(y)/\rho\Big], \tag{6.6}$$

where $G_i^\dagger(y)$ is the c.d.f. of $y \equiv B_{01}$ under $H_i^\dagger$. It is shown in Appendix C.2, equation (C.10), that the corresponding p.d.f of B_{01} is given by:

$$g_i^\dagger(y) \;=\; \int_{H_i} g_i(y \,|\, \nu)\; \varphi_i(\nu)\; d\nu, \tag{6.7}$$

where $g_i(y \,|\, \nu)$ is the p.d.f. of B_{01} under H_i. Although we have used Bayesian integrations over ν to construct a ν-independent version of S, the latter can be viewed as nothing more than a known function of the data, i.e. as a frequentist statistic. On the other hand, the error rates derived by conditioning on S will now be functions of ν, say $\tilde{\alpha}(\nu \,|\, s)$ and $\tilde{\beta}(\nu \,|\, s)$, and are therefore unknown. However, it turns out that the posterior probabilities of error in equation (6.5) can be interpreted as *average* conditional frequentist error rates, where the average is taken with respect to an appropriate posterior distribution. Let $p_i(\nu \,|\, s)$ be the posterior distribution of the nuisance parameters ν under H_i, conditional on the observed value s of the proportional tail statistic S. Appendix C.2 shows that:

$$\begin{aligned} \tilde{\alpha}^\dagger(s) \;&\equiv\; \mathbb{E}^{p_0(\nu \,|\, s)}\Big[\tilde{\alpha}(\nu \,|\, s)\Big] \;\equiv\; \int_{H_0} \tilde{\alpha}(\nu \,|\, s)\; p_0(\nu \,|\, s)\; d\nu \;=\; \frac{\pi_0 B_{01}}{\pi_0 B_{01} + \pi_1}, \\ \tilde{\beta}^\dagger(s) \;&\equiv\; \mathbb{E}^{p_1(\nu \,|\, s)}\Big[\tilde{\beta}(\nu \,|\, s)\Big] \;\equiv\; \int_{H_1} \tilde{\beta}(\nu \,|\, s)\; p_1(\nu \,|\, s)\; d\nu \;=\; \frac{\pi_1}{\pi_0 B_{01} + \pi_1}. \end{aligned} \tag{6.8}$$

In order to reexpress the test (6.5) in terms of p values we begin by observing that the frequentist probability of $B_{01} \leq \pi_1/\pi_0$ depends on ν:

$$\alpha(\nu) \;\equiv\; \mathbb{P}\Big[B_{01} \leq \frac{\pi_1}{\pi_0} \,\Big|\, H_0\Big] \;=\; \int_0^{\pi_1/\pi_0} g_0(y \,|\, \nu)\; dy, \tag{6.9}$$

where $g_0(y \,|\, \nu)$ is the distribution of B_{01} under H_0. To eliminate ν we average $\alpha(\nu)$ over the ν prior under H_0 (averaging over a posterior would destroy α's status as a pre-experimental quantity):

$$\begin{aligned} \alpha^\dagger \;\equiv\; \mathbb{E}^{\varphi_0(\nu)}\Big[\alpha(\nu)\Big] \;&=\; \int_{H_0} \alpha(\nu)\; \varphi_0(\nu)\; d\nu \;=\; \int_{H_0}\int_0^{\pi_1/\pi_0} g_0(y \,|\, \nu)\; \varphi_0(\nu)\; dy\; d\nu \\ &=\; \int_0^{\pi_1/\pi_0} g_0^\dagger(y)\; dy. \end{aligned} \tag{6.10}$$

Suppose next that we observe $B_{01} = y_{obs}$. The p value under $H_0^\dagger$ is then:

$$p_0 \;=\; \int_0^{y_{obs}} g_0^\dagger(y)\; dy, \tag{6.11}$$

and is known as a prior-predictive p value. The test (6.5) is now easily seen to be equivalent to:

$$\begin{cases} \text{If } p_0 \leq \alpha^\dagger, \text{ reject } H_0^\dagger \text{ and report average conditional error probability } \tilde{\alpha}^\dagger(s); \\ \text{If } p_0 > \alpha^\dagger, \text{ accept } H_0^\dagger \text{ and report average conditional error probability } \tilde{\beta}^\dagger(s), \end{cases} \tag{6.12}$$

where p_0, $\alpha^\dagger$, $\tilde{\alpha}^\dagger(s)$, and $\tilde{\beta}^\dagger(s)$ are defined by equations (6.11), (6.10), and (6.8), respectively. Strict frequentists will object to procedure (6.12) on the grounds that it tests $H_0^\dagger$ instead of H_0, and uses a prior-predictive p value to do so. Often however,

nuisance parameters model systematic uncertainties that do not have a strict frequentist character, in which case (6.12) still offers a reasonable approach. In other cases it can be argued that the use of prior-predictive p values to test H_0 is an approximation that is often conservative and therefore acceptable. [17]

7 Summary

The main purpose of this note was to motivate and construct frequentist hypothesis tests whose results can be judged on the basis of a measure of confidence that takes into account the evidence contained in the observed data. Several test procedures were discussed along the way; we summarize them here:

1. Standard frequentist

   - Procedure: choose α, then reject H_0 if $p_0 \leq \alpha$ and accept H_0 otherwise.
   - Pros: freedom of choosing α allows full control of the probability of incorrectly rejecting H_0 (this is of crucial importance when H_0 is the standard model).
   - Cons: confidence levels $1 - \alpha$ and $1 - \beta$ are unconditional and do not reflect the strength of evidence contained in the data; very small α leads to small power $1 - \beta$ if the measurement resolution is poor; the probability of selecting the correct hypothesis is always the same, regardless of sample size (inconsistency).

2. Minimax frequentist

   - Procedure: reject H_0 if $p_0 \leq p_1$ and accept H_0 otherwise; use the error rate α to evaluate the reliability of the decision: α will decrease as the experimental resolution improves.
   - Pros: procedure only has one error rate since $\alpha = \beta$, and this error rate depends directly on the measurement resolution; the probability of selecting the correct hypothesis approaches 1 in large samples.
   - Cons: confidence level is unconditional and does not reflect the strength of evidence contained in the data; the user does not get to choose α; both hypotheses are treated on an equal basis, even if one hypothesis is a priori far more credible than the other.

3. Bayes

   - Procedure: select prior probabilities for the null and alternative hypotheses, calculate their posterior probabilities, and reject the hypothesis with the smaller posterior probability.
   - Pros: posterior probabilities are a direct measure of the evidence contained in the data, taking measurement resolution into account.

- Cons: requires the choice of prior probabilities for the hypotheses (this is similar to the choice of α in the standard frequentist test).

4. Unified Bayes/conditional frequentist, with no-decision region

   - Procedure: choose α and compute the no-decision region $]r, a[$ (equation 5.16); compute the Bayes factor B_{01}, then reject H_0 if $B_{01} \leq r$, accept H_0 if $B_{01} \geq a$, and make no decision otherwise; if a decision was made, use B_{01} to calculate the corresponding conditional error probability (equation 5.10).
   - Pros: this is essentially the standard frequentist procedure, supplemented with a data-based statement of error probability; the error probability can also be interpreted as a Bayesian posterior hypothesis probability.
   - Cons: the no-decision region adds a minor complication to the test. In general however, that region does appear sensible and tends to shrink with increasing measurement resolution.

5. Unified Bayes/conditional frequentist, with loss structure

   - Procedure: for each hypothesis, estimate the loss that would result from incorrectly rejecting it; calculate the corresponding value of α (equation 5.14), then reject H_0 if $p_0 \leq \alpha$ and accept H_0 otherwise; report the conditional frequentist risk associated with the decision.
   - Pros: the reported risk is conditional on the observed data and has a unified frequentist/Bayes interpretation; there is no non-decision region.
   - Cons: it may not be clear how to estimate losses (although one possibility would be to set the loss ratio ℓ_1/ℓ_0 at whatever value gives $\alpha = 2.7 \times 10^{-7}$); the concept of risk is probably unfamiliar to most high energy physicists.

6. Unified Bayes/conditional frequentist, with empty no-decision region

   - Procedure: set α equal to the value that makes the no-decision region empty (equation 5.19), then follow the standard frequentist test procedure; report the conditional frequentist error probability associated with the decision taken.
   - Pros: all the advantages of a unified Bayes/frequentist test; no non-decision region.
   - Cons: the value of α cannot be chosen by the user (however, the reported CEP gives a data-based measure of the confidence level of one's decision); both hypotheses are treated on (approximately) the same basis.

In choosing a procedure, the user will need to take several issues into consideration. The first one is whether the two hypotheses should be treated on the same basis, i.e. whether the cost of an incorrect rejection is the same for both. Even when it is difficult to quantify this cost, knowing that it is different for the two hypotheses helps determine which one should be the null. Assuming that a frequentist test is desired, the second

issue is the choice of rejection threshold α. If H_0 and H_1 are treated on the same basis, a minimax test may be preferred (option 2). Otherwise, the choice of α will depend on prior belief and cost considerations. Finally, if a confidence level is desired that takes observed evidence into account, it can be calculated directly from the Bayes factor B_{01}. In this case, options 4, 5, or 6 would be appropriate.

In general there seems to be little reason *not* to choose a unified test, due to the richness of interpretation this method adds, with a minimum of effort, to the standard frequentist and Bayesian tests. For the top charge analysis, the lack of symmetry in explanatory power between the standard and exotic models disfavors option 6. On the other hand, the loss structure involved in option 5 is somewhat overly abstract. Thus one is left with recommending option 4 for this analysis.

Acknowledgements

The motivation for this work was provided by several discussions with authors of the CDF top charge analysis: Jaroslav Antoš, Andy Beretvas, Veronique Boisvert, Yen-Chu Chen, and Veronica Sorin. We thank them for raising some of the issues studied in this note. We also thank Louis Lyons for helpful comments on a previous version of this note.

Appendix

A Summary of CDF's top charge analysis

The CDF analysis is based on a sample of 193 lepton+jets and 44 dilepton $t\bar{t}$ candidates, corresponding to an integrated luminosity of 1.5 fb^{-1}. In these 237 events CDF identifies a total of 225 top or antitop candidates for which the b-jet charge can be estimated. A maximum likelihood method is then used to estimate the fraction f_+ of candidates with a standard model top charge; the result is $f_+ = 0.87$. To test the standard model hypothesis that the true value of f_+ is 1, CDF chooses a rejection threshold α that minimizes the probability of incorrectly rejecting the standard model while keeping the power of the test at a reasonable level, given the limited resolution of the measurement. A satisfactory setting is found for $\alpha = 1\%$, where the power is 87%, i.e. $\beta = 13\%$. Using f_+ as test statistic, CDF calculates the p value under the standard model and finds $p_{sm} = 0.31$. Since $p_{sm} > \alpha$, the standard model is accepted and the exotic model is excluded with 87% confidence (see Table 1).

CDF also calculates the Bayes factor in favor of the standard model and finds $B_{01} \approx 403$. According to a conventional interpretation of the evidence provided by Bayes factors [18], this result *very strongly favors* the standard model over the exotic one. Although this is not part of the official CDF interpretation, one can use equation (5.10) to convert the Bayes factor into a conditional frequentist error probability CEP. With $\rho \equiv (1-\beta)/(1-\alpha) \approx 0.88$, one obtains CEP$= 1/(\rho B_{01} + 1) \approx 0.28\%$.

Finally CDF calculates Feldman-Cousins intervals on f_+, finding $f_+ > 0.4$ (0.6) at the 95% (68%) confidence level. That these intervals are one-sided is partly due to the limited resolution of the measurement. With better resolution, Feldman-Cousins intervals could have turned out two-sided, leading to the potentially confusing report of an interval that excludes the standard model whereas the hypothesis test accepts it. The only assured way to avoid this problem is to use a lower limit ordering rule with an estimator of f_+ that respects the physical boundaries, as in equation (3.5). Such a method is described in Ref. [13].

The issue of what information to report after a test has resulted in the acceptance of a null hypothesis H_0 deserves further comment. Clearly, claiming consistency of the data with H_0 is meaningless if the test is not sensitive. One way to avoid such claims is to examine the power of the test, $1-\beta$. However, β is calculated with respect to a predesignated value of α without regard for the specific data one has observed. Ref. [10] argues that a more relevant quantity is the probability $\zeta(\delta)$ of observing a larger discrepancy with $H_0 : \mu = \mu_0$ if the true value of μ is $\mu_0 + \delta$:

$$\zeta(\delta) \;=\; \mathbb{P}\Big[p_0(T) \le p_0(t_{obs}) \,\Big|\, \mu = \mu_0 + \delta\Big], \tag{A.1}$$

where $p_0(T)$ is the p value under H_0, evaluated at the test statistic T, and t_{obs} is the observed value of T. Thus, a large $\zeta(\delta)$ indicates that the data can definitely distinguish H_0 from a hypothesis that is a distance δ away from H_0. Rather than reporting $\zeta(\delta)$, one can construct the set I_γ of μ values for which $\zeta(\mu - \mu_0) \le \gamma$, where γ is a probability near 1. This is a γ-confidence level interval that contains μ_0 (provided

$p_0(t_{obs}) \leq \gamma$), plus all the μ values that the data can hardly distinguish from μ_0. Due to this interpretation, the set I_γ provides a more insightful complement to the result of the test than a Feldman-Cousins interval.

B Summary of DØ's top charge analysis

The DØ collaboration has published a top charge analysis based on 370 pb^{-1} of data [4]. Using 32 measurements of the top quark charge in a sample of 16 lepton+jets events, they construct a likelihood ratio for the standard model versus the exotic model. By comparing the observed likelihood ratio to the distribution expected under the exotic model, DØ obtains a p value of 0.078. It is important to keep in mind that, in contrast with CDF, DØ's p value is calculated *under the exotic hypothesis*. To emphasize this difference in the choice of null hypothesis, we will use the notation p_{xm} for DØ's p value.

The next step in DØ's interpretation is to claim that $1 - p_{xm}$ is the confidence level with which they can exclude that their data set is solely composed of exotic quarks with charge $4e/3$. Strictly speaking, this is a misuse of standard frequentist terminology. As shown in Fig. 4, p values consistently *under*estimate both conditional and unconditional frequentist error probabilities. Therefore, $1 - p_{xm}$ consistently *over*estimates the corresponding confidence levels.

It is nevertheless possible to argue that $1 - p_{xm}$ is a *hypothetical* confidence level. Indeed, anyone with an a priori rejection threshold α at least as large as DØ's observed p_{xm} will reject the exotic model. In other words, p_{xm} can be interpreted as the smallest Type-I error probability α for which the exotic model would be rejected. Correspondingly, $1 - p_{xm}$ can be interpreted as the largest Type-I confidence level [5]. A conservative observer might not find this interpretation very compelling however, since it would clearly be more reassuring to know the *largest* probability that one committed an error, or equivalently, the *lowest* confidence that one may claim in the action taken. This caveat is one motivation for labeling such p value based confidence levels "hypothetical". Another motivation is that the value of a hypothetical confidence level is unknown before the measurement. Thus, it cannot be used to verify an actual frequentist error rate in an ensemble of experiments containing the experiment actually performed.

Even if we accept the notion of hypothetical confidence levels, say for the purpose of quantifying evidence rather than establishing error rates [19], DØ's analysis summary is still unsatisfactory because it fails to report *both* confidence levels of the test. Indeed, DØ *rejects* their null hypothesis (the exotic model), but they only report the hypothetical exclusion confidence level, which in their case is the largest probability for *accepting* the exotic model if it is true. As noted in section 3, when rejecting the null hypothesis, the other confidence level of interest is the probability of making this decision when the null is false, i.e. the power of the test. Unfortunately this important information is missing from DØ's summary.

Assuming that we are testing simple hypotheses (all heavy quarks in the sample have charge $2e/3$ or all have charge $4e/3$, but there is no mixing), the power $1 - \beta$ is a

function of the rejection threshold α only. Thus, when rejecting the exotic model, DØ could either report the entire function $1 - \beta(\alpha)$, or just the hypothetical acceptance confidence level $1 - \beta(\alpha = p_{xm})$.

C Derivation of conditional frequentist error rates

Here we derive results (5.10) and (6.8) given in the text. The cases with and without systematic uncertainties are treated separately.

C.1 Without systematic uncertainties

In the notation of section 5, the elements of the conditioning partition are labeled by the real number $s \in [0, c]$, where c is the critical value. If a Bayes factor $y \leq c$ is observed, the null hypothesis is rejected and the conditional probability of error is:

$$\tilde{\alpha}(s) \;=\; \mathbb{P}_0\Big[y = s \;\Big|\; y = s \text{ or } y = \psi(s)\Big] \;=\; \frac{g_0(s)}{g_0(s) \;+\; g_0(\psi(s)) \left|\dfrac{d\psi}{ds}\right|}, \tag{C.1}$$

where $g_i(y)$ is the p.d.f. of the Bayes factor y under H_i and $\psi(s)$ is given by equation (5.4). Differentiating $\psi(s)$ yields:

$$\frac{d\psi}{ds} \;=\; -\frac{g_1(s)}{\rho\, g_0\left[\psi(s)\right]}, \tag{C.2}$$

so that:

$$\tilde{\alpha}(s) \;=\; \frac{\rho\, g_0(s)}{\rho\, g_0(s) \;+\; g_1(s)}. \tag{C.3}$$

To continue, we write $f_i(x) \equiv f(x \,|\, \mu_i)$ for the p.d.f. of the data x under H_i and note the following:

$$\begin{aligned}\int_0^s g_0(y)\, dy \;&=\; \int_{\{x: B_{01}(x) \leq s\}} f_0(x)\, dx \;=\; \int_{\{x:\, B_{01}(x) \leq s\}} \frac{f_0(x)}{f_1(x)}\, f_1(x)\, dx \\ &=\; \int_{\{x:\, B_{01}(x) \leq s\}} B_{01}(x)\, f_1(x)\, dx \;=\; \int_0^s y\, g_1(y)\, dy, \end{aligned} \tag{C.4}$$

where we twice changed variable according to $g_i(y)dy = f_i(x)dx$, and used the definition of the Bayes factor in favor of H_0, $B_{01}(x) \equiv f_0(x)/f_1(x)$. Differentiating the first and last members of the above sequence of equalities with respect to s yields:

$$g_0(s) \;=\; s\, g_1(s). \tag{C.5}$$

The expression for $\tilde{\alpha}(s)$ becomes then:

$$\tilde{\alpha}(s) \;=\; \frac{\rho\, s}{\rho\, s \;+\; 1} \;=\; \frac{\rho\, B_{01}}{\rho\, B_{01} \;+\; 1}, \tag{C.6}$$

where the second equality results from the fact that when we reject H_0 it is because $B_{01} \leq c$ and therefore $B_{01} = s$. A similar calculation for $\tilde{\beta}(s)$ yields:

$$\tilde{\beta}(s) \;=\; \frac{1}{\rho\,\psi(s) \;+\; 1}. \tag{C.7}$$

In this case however, we report $\tilde{\beta}(s)$ as the conditional error rate when we *accept* H_0, i.e. when $B_{01} > c$. Since $s \in [0, c]$ this means that $B_{01} = \psi(s)$, so that:

$$\tilde{\beta}(s) \;=\; \frac{1}{\rho\,B_{01} \;+\; 1}. \tag{C.8}$$

This concludes the proof of equations (5.10). [15]

C.2 With systematic uncertainties

Suppose that there are systematic uncertainties (nuisance parameters ν) under both H_0 and H_1. The hypothesis test then has the form (6.2) and the Bayes factor $B_{01}(x)$ is given by equation (6.4). Let $g_i(y\,|\,\nu)$ be the conditional p.d.f. of $B_{01}(x)$ given ν under H_i, and let $g_i^\dagger(y)$ be the p.d.f. of $B_{01}(x)$ under $H_i^\dagger$ (defined in equation 6.3). This means that under the change of variables $x \leftrightarrow y$, where $y = B_{01}(x)$, we have $g_i^\dagger(y)\,dy = f_i^\dagger(x)\,dx$ and $g_i(y\,|\,\nu)\,dy = f_i(x\,|\,\nu)\,dx$. Therefore:

$$\begin{aligned}
\int_0^s g_i^\dagger(y)\,dy \;&=\; \int_{\{x:\,B_{01}(x)\leq s\}} f_i^\dagger(x)\,dx \;=\; \int_{\{x:\,B_{01}(x)\leq s\}} \int_{H_i} f_i(x\,|\,\nu)\,\varphi_i(\nu)\,d\nu\,dx \\
&=\; \int_{H_i} \int_{\{x:\,B_{01}(x)\leq s\}} f_i(x\,|\,\nu)\,\varphi_i(\nu)\,dx\,d\nu \;=\; \int_{H_i} \int_0^s g_i(y\,|\,\nu)\,\varphi_i(\nu)\,dy\,d\nu \\
&=\; \int_0^s \int_{H_i} g_i(y\,|\,\nu)\,\varphi_i(\nu)\,d\nu\,dy.
\end{aligned} \tag{C.9}$$

Differentiating with respect to s the first and last expressions in this string of equalities yields:

$$g_i^\dagger(s) \;=\; \int_{H_i} g_i(s\,|\,\nu)\,\varphi_i(\nu)\,d\nu \tag{C.10}$$

We will need two results from Appendix C.1; the first one is that the argument leading to eq. (C.5) can be recycled here to obtain:

$$g_0^\dagger(s) \;=\; s\,g_1^\dagger(s), \tag{C.11}$$

and the second one is a daggered version of equation (C.2):

$$\frac{d\psi^\dagger}{ds} \;=\; -\,\frac{g_1^\dagger(s)}{\rho\,g_0^\dagger\,[\psi^\dagger(s)]}, \tag{C.12}$$

where $\psi^\dagger(s)$ is defined in equation (6.6).

When $B_{01} \leq c$ we reject $H_0^\dagger$; the conditional Type-I error rate is then:

$$\tilde{\alpha}(\nu \,|\, s) \;=\; \mathbb{P}\Big[B_{01} \leq c \,\Big|\, H_0 \text{ and } S = s\Big] \;=\; \frac{g_0(s \,|\, \nu)}{g_0(s \,|\, \nu) \;+\; g_0(\psi^\dagger(s) \,|\, \nu) \left|\dfrac{d\psi^\dagger}{ds}\right|}. \tag{C.13}$$

Under H_0, the conditional posterior p.d.f. of ν given s is:

$$p_0(\nu \,|\, s) \;=\; \frac{\left[g_0(s \,|\, \nu) + g_0(\psi^\dagger(s) \,|\, \nu) \left|\dfrac{d\psi^\dagger}{ds}\right|\right] \varphi_0(\nu)}{m_0(s)}, \tag{C.14}$$

where, using eq. (C.10):

$$m_0(s) \;\equiv\; \int_{H_0} \left[g_0(s \,|\, \nu) + g_0(\psi^\dagger(s) \,|\, \nu) \left|\frac{d\psi^\dagger}{ds}\right|\right] \varphi_0(\nu)\, d\nu \;=\; g_0^\dagger(s) + g_0^\dagger(\psi^\dagger(s)) \left|\frac{d\psi^\dagger}{ds}\right|. \tag{C.15}$$

The posterior expected conditional error rate is then:

$$\begin{aligned}
\mathbb{E}^{p_0(\nu \,|\, s)}\Big[\tilde{\alpha}(\nu \,|\, s)\Big] \;&=\; \int_{H_0} \tilde{\alpha}(\nu \,|\, s)\; p_0(\nu \,|\, s)\; d\nu \\
&=\; \int_{H_0} \frac{g_0(s \,|\, \nu)\, \varphi_0(\nu)}{m_0(s)}\; d\nu && \text{by (C.13) and (C.14)} \\
&=\; \frac{g_0^\dagger(s)}{g_0^\dagger(s) + g_0^\dagger(\psi^\dagger(s)) \left|\dfrac{d\psi^\dagger}{ds}\right|} && \text{by (C.10) and (C.15)} \\
&=\; \frac{g_0^\dagger(s)}{g_0^\dagger(s) + g_1^\dagger(s)/\rho} && \text{by (C.12)} \\
&=\; \frac{s}{s + 1/\rho} && \text{by (C.11)} \\
&=\; \frac{\rho B_{01}}{\rho B_{01} + 1}, && \text{(C.16)}
\end{aligned}$$

where the last equality follows from the fact that $B_{01} = s$ on the set $\{B_{01} \leq c \text{ and } S = s\}$. This result shows that the posterior probability of H_0 equals the average of the conditional Type-I error probability with respect to the posterior distribution of ν given s, under H_0.

When $B_{01} > c$, a similar calculation [20] leads to:

$$\mathbb{E}^{p_1(\nu \,|\, s)}\Big[\tilde{\beta}(\nu \,|\, s)\Big] \;=\; \frac{1}{\rho B_{01} + 1}. \tag{C.17}$$

Using the matching condition (5.12), equations (C.16) and (C.17) establish the result (6.8).

Figures

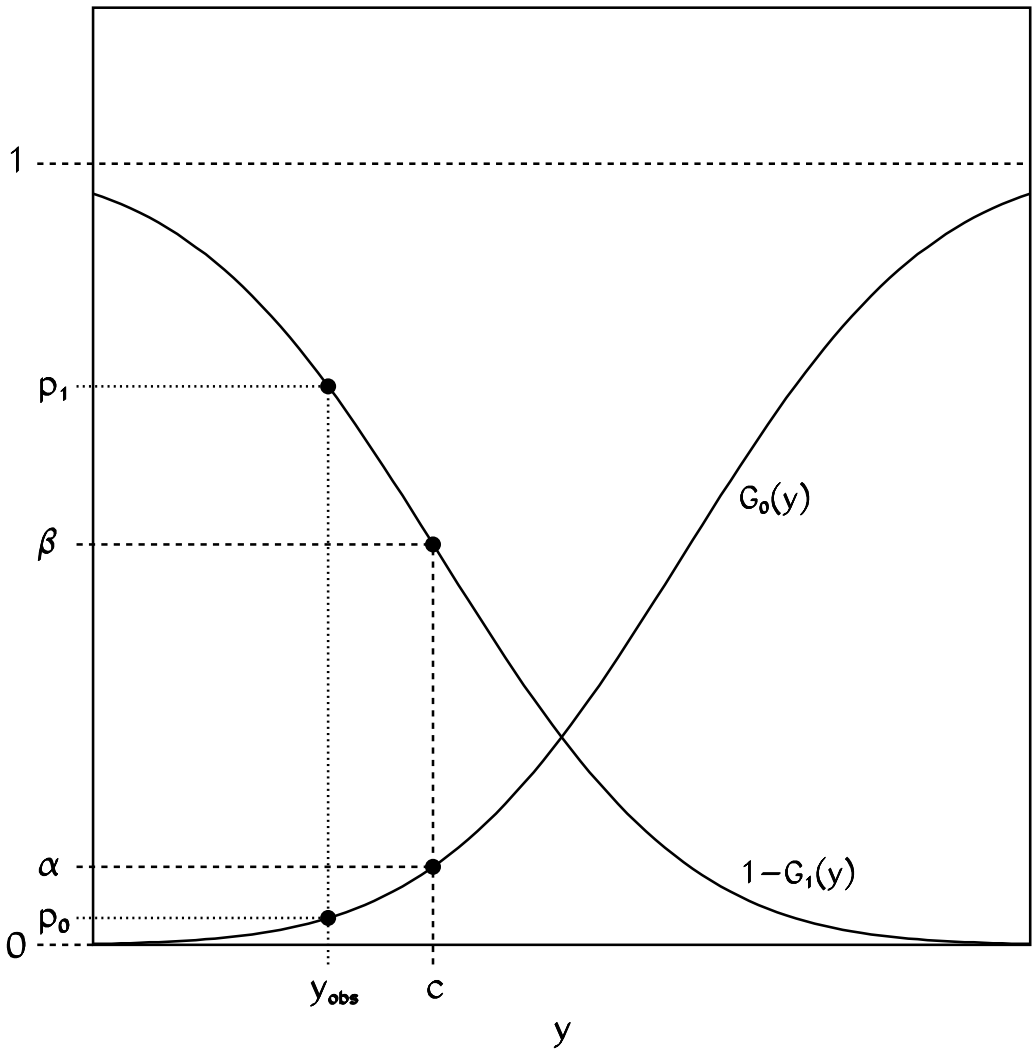

Figure 1: Graphic demonstrating the relationships between error rates and p values in a hypothesis test of H_0 versus H_1. The test is based on the observation of a statistic y whose cumulative probability distribution under hypothesis H_i is $G_i(y)$, where $i = 0$ or 1. The critical region consists of all y values lower than or equal to c, and the corresponding error probabilities are $\alpha = G_0(c)$ for Type-I, and $\beta = 1-G_1(c)$ for Type-II. When a value y_{obs} of y is observed, one can calculate two p values: $p_0 = G_0(y_{obs})$ under H_0, and $p_1 = 1 - G_1(y_{obs})$ under H_1. The null hypothesis H_0 will be rejected if $y_{obs} \leq c$. The graph shows that this is equivalent to $p_0 \leq \alpha$, as well as to $p_1 \geq \beta$. As c moves to the right, α and β approach each other until they become equal at the crossing point of the two curves. That point corresponds to the so-called minimax test, for which $p_0 \leq \alpha$ if and only if $p_0 \leq p_1$.

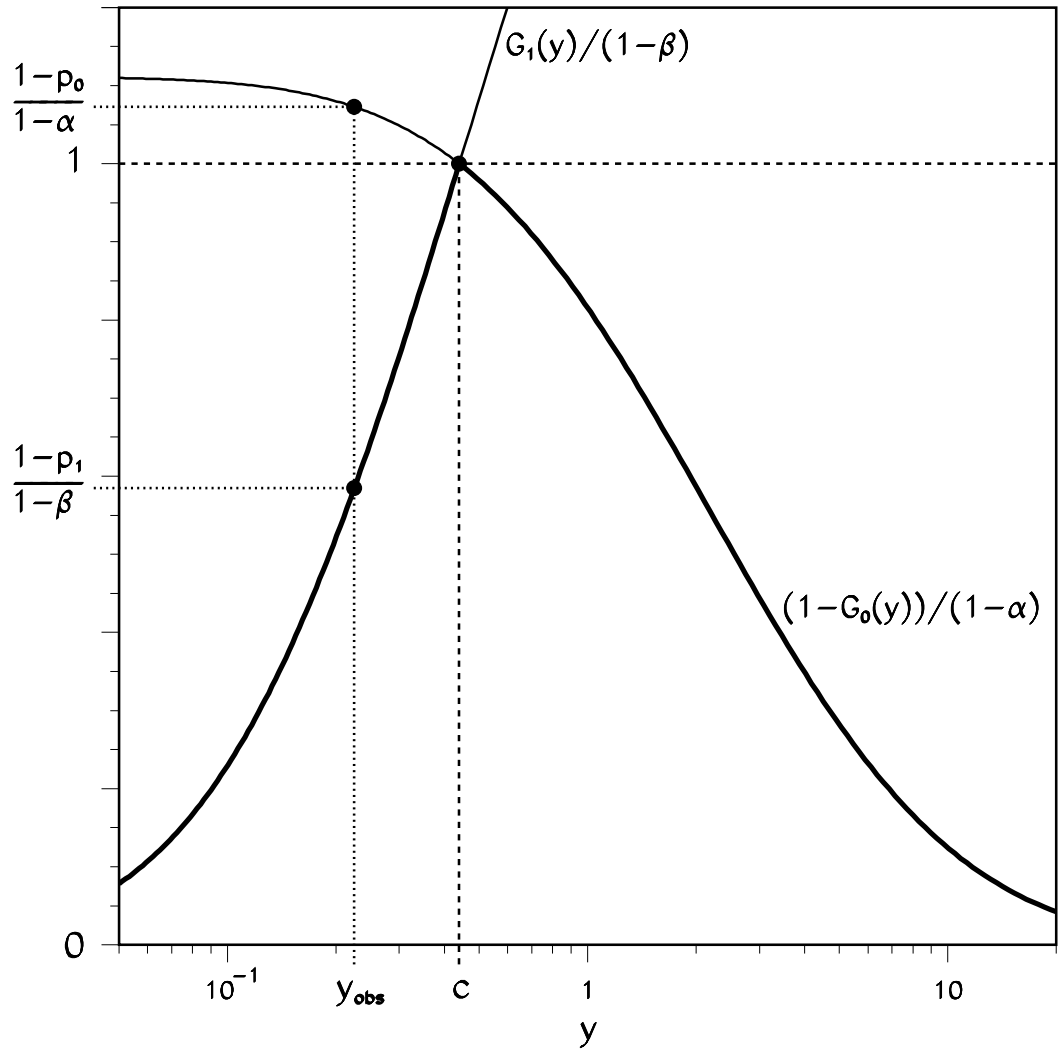

Figure 2: Graphic illustrating the conditioning statistic $Q \equiv \min\{(1-p_0)/(1-\alpha), (1-p_1)/(1-\beta)\}$ of section 5.1 (thick solid line) for the case of Example 2, with $\Delta\mu = 1$ and $\sigma = 0.8$. In addition, α is set to 0.10, which corresponds to $c \approx 0.44$ and $\beta \approx 0.51$; the observation y_{obs} equals 0.225, yielding $p_0 \approx 0.035$ and $p_1 \approx 0.72$, so that $Q_{obs} = (1-p_1)/(1-\beta) = 0.58$. Note that we always have $0 \leq Q \leq 1$.

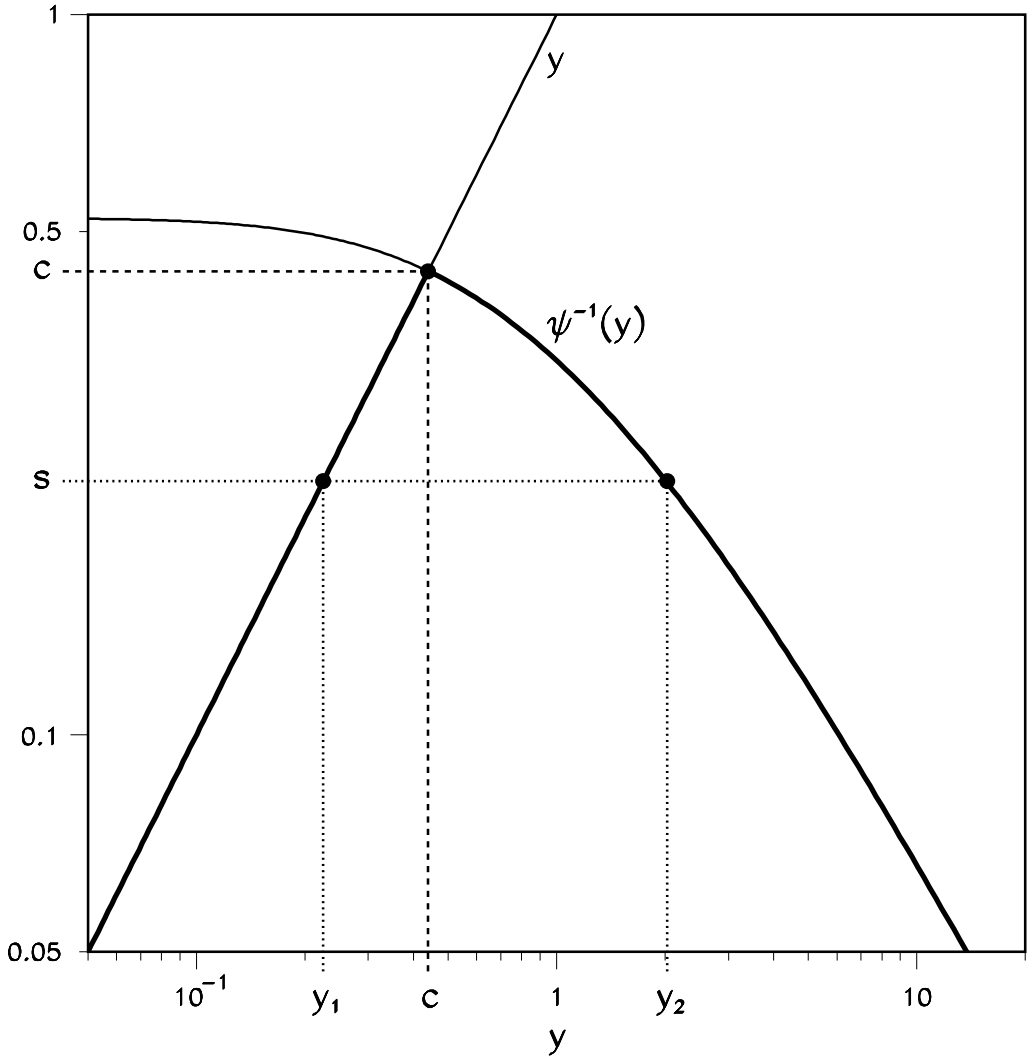

Figure 3: Graphic illustrating the partition induced by the conditioning statistic $S(y) \equiv \min\{\psi^{-1}(y),\, y\}$ in the space of Bayes factors y, where $\psi^{-1}(y) \equiv G_1^{-1}[\rho(1 - G_0(y))]$ and $\rho \equiv (1-\beta)/(1-\alpha)$. For every $s < c$, the equation $s = S(y)$ has two solutions in y, one on each side of the critical boundary c. Following the dotted lines on the graph shows that $s = y_1 = \psi^{-1}(y_2)$; therefore $y_2 = \psi(y_1)$, where y_1 is the smaller of the two solutions (see footnote 4 on pg. 17). By construction, y_2 represents the same evidence against H_1 as y_1 does against H_0.

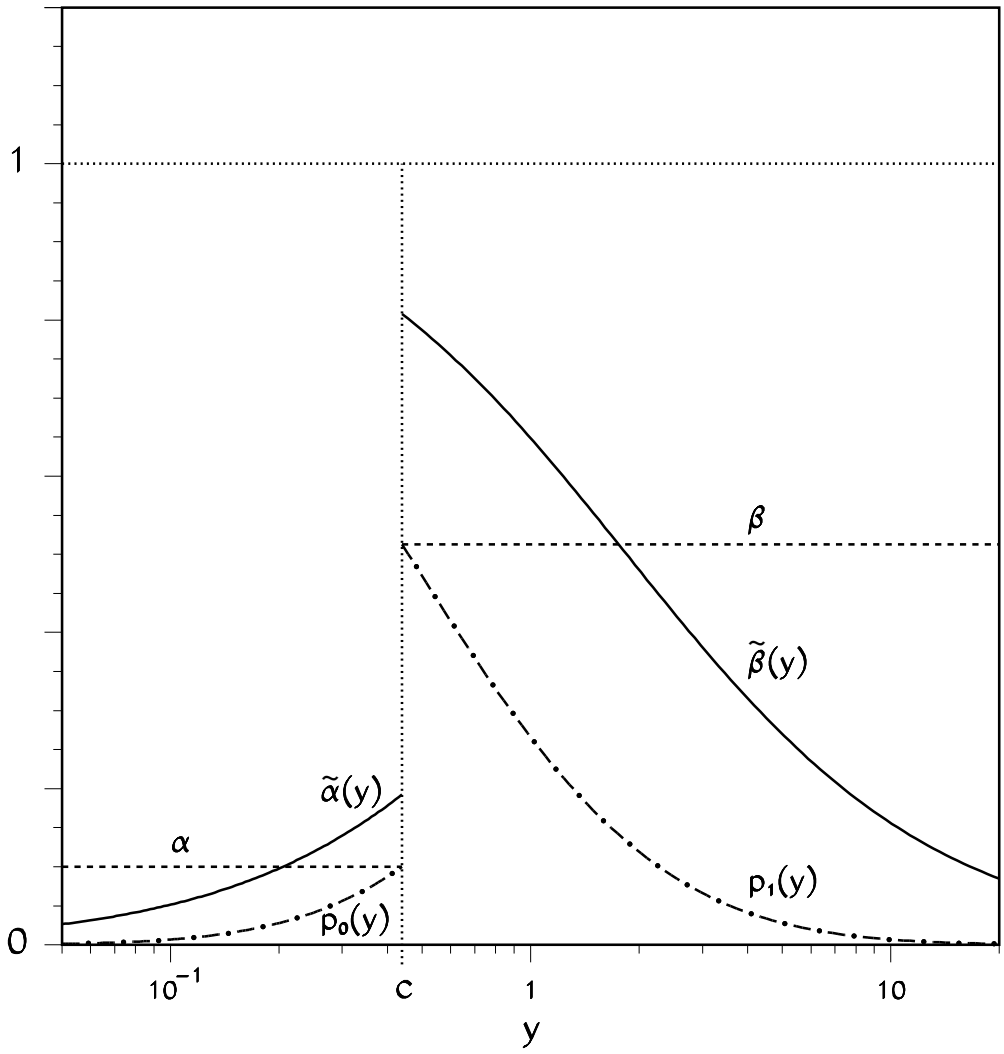

Figure 4: Unconditional frequentist error rates (horizontal dashed lines), unified conditional frequentist and Bayesian error rates (solid lines), and p values (dot-dashed lines), for the same test setup as in Fig.2. Note that the p values consistently *under*estimate the frequentist and Bayesian error rates.

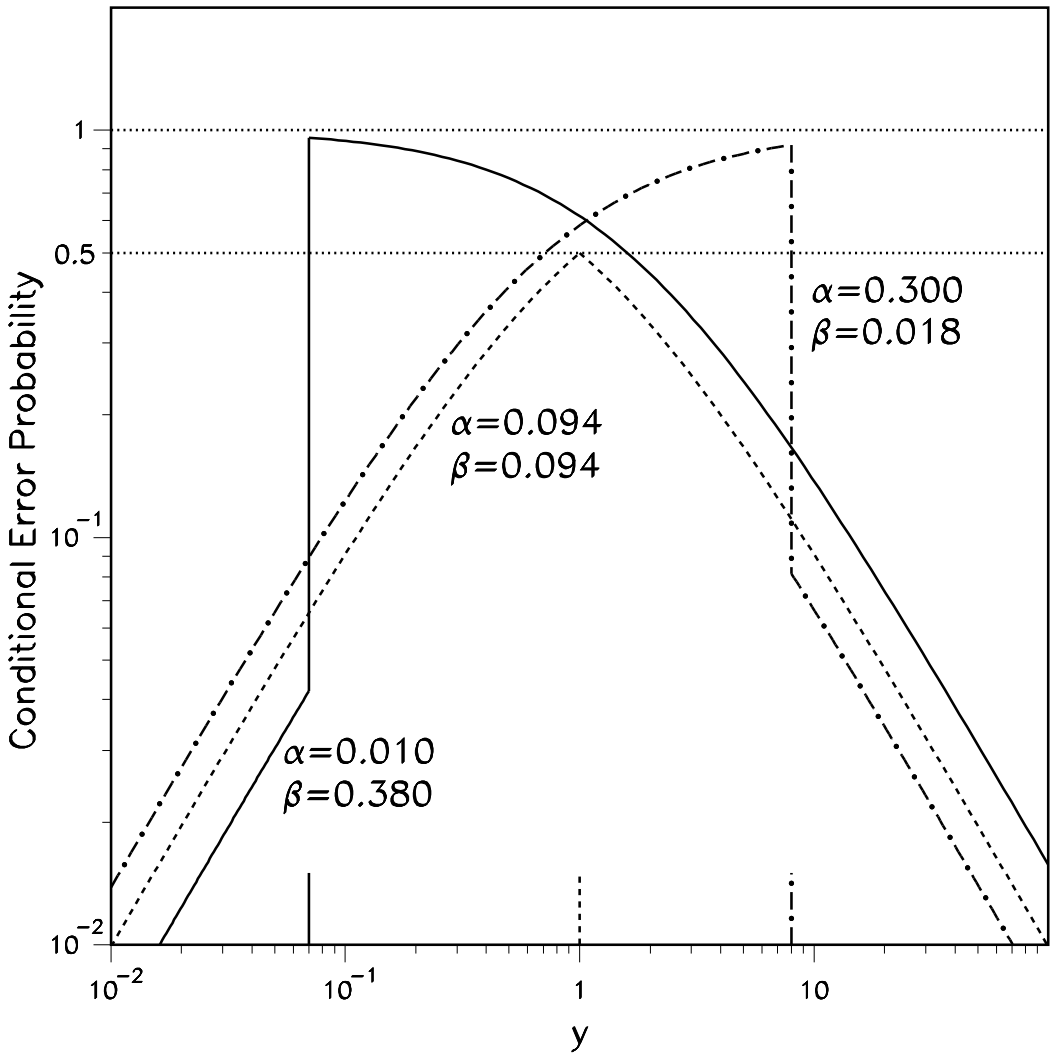

Figure 5: Conditional error probabilities for three values of the unconditional Type-I error rate α in the Gaussian approximation to the top charge analysis (Example 2 in the text, with $\Delta\mu = 1$ and $\sigma = 0.38$): $\alpha = 0.01$ (solid), $\alpha = 0.094$ (dashes), and $\alpha = 0.30$ (dot-dashes). In each case, a line segment of the same type indicates the critical value of the corresponding test along the abscissae. For example, if one chooses $\alpha = 0.01$ and obtains a Bayes factor to the right of the solid line segment, one will accept the null hypothesis and report the conditional error probability given by the solid curve. The value $\alpha = 0.094$ corresponds to the minimax test, for which $\beta = \alpha$ unconditionally. The conditional error probabilities for this test are always below 50%.

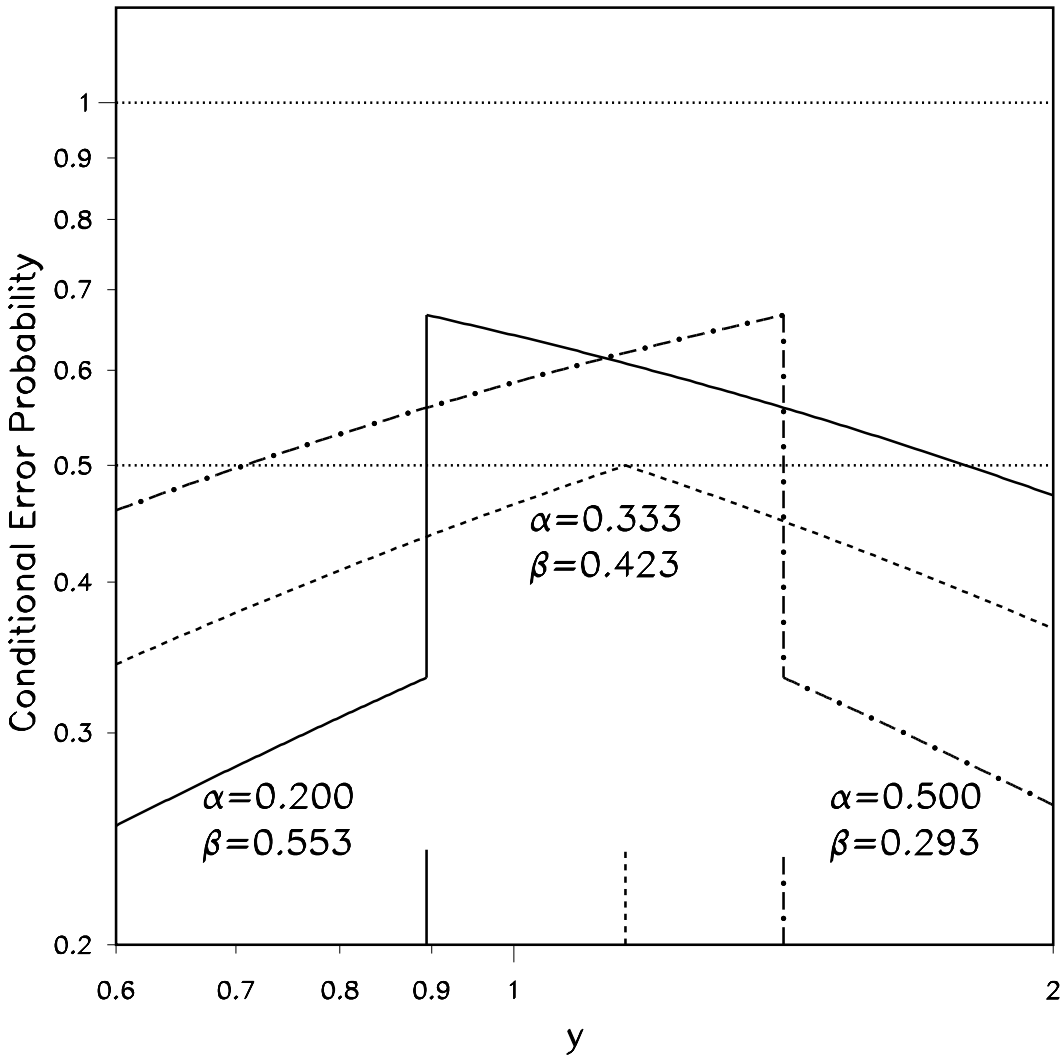

Figure 6: Same as Figure 5, but for a test based on a lifetime measurement (Example 4 in the text). Conditional error rates are shown for three values of the unconditional Type-I error rate α: 0.200 (solid), 0.333 (dashes), and 0.500 (dot-dashes). In each case, a line segment of the same type indicates the critical value of the corresponding test along the abscissae. Because the measurement does not enjoy likelihood ratio symmetry, the test that yields conditional error rates never larger than 50% is *not* minimax.

References

[1] D. Chang, W.-F. Chang, and E. Ma, "Alternative interpretation of the Fermilab Tevatron top events," Phys. Rev. D **59**, 091503 (1999). 3

[2] D. Chang, W.-F. Chang, and E. Ma, "Fitting precision electroweak data with exotic heavy quarks," Phys. Rev. D **61**, 037301 (2000). 3

[3] D. Choudhury, T. M. P. Tait, and C. E. M. Wagner, "Beautiful mirrors and precision electroweak data," Phys. Rev. D **65**, 053002 (2002). 3

[4] V. M. Abazov *et al.* (DØ Collaboration), "Experimental discrimination between charge $2e/3$ top quark and charge $4e/3$ exotic quark production scenarios," Phys. Rev. Lett. **98**, 041801 (2007). 3, 30

[5] U. Heintz, E. Shabalina, and M. Weber, "Clarification on the DØ measurement of the top quark charge," http://www-d0.fnal.gov/Run2Physics/WWW/results/final/TOP/T06D/extra/topQ.htm (April 2007). 3, 30

[6] A. Beretvas *et al.*, "Finding the charge of the top quark in the dilepton channel," CDF note 8367, FERMILAB-CONF-06-251-E (July 2006). 3

[7] J. Antoš, Y.-C. Chen, and A. Beretvas, "Top charge reconstruction," CDF internal note 8638 (February 2007). 3

[8] V. Boisvert *et al.*, "Measuring the sign of the top charge using the top decay products," CDF internal note 8654 (February 2007). 3

[9] J. Antoš, Y.-C. Chen, and A. Beretvas, "Analysis of top charge reconstruction," CDF internal note 8713 (July 2007). 3

[10] D. G. Mayo and D. R. Cox, "Frequentist statistics as a theory of inductive inference," IMS Lecture Notes — Monograph Series: 2nd Lehmann Symposium — Optimality, Vol. 49, pg.77-97 (2006); arXiv:math/0610846v1 [math.ST] 27 Oct 2006. 9, 29

[11] G. Punzi, "Sensitivity of searches for new signals and its optimization," in *Proceedings of the conference on statistical problems in particle physics, astrophysics and cosmology (PhyStat2003)*, ed. L. Lyons, R. Mount, and R. Reitmeyer, SLAC report eConf: C030908, SLAC-R-703. 10

[12] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D **57**, 3873 (1998). 10

[13] M. Mandelkern and J. Schultz, "The statistical analysis of Gaussian and Poisson signals near physical boundaries," J. Math. Phys. **41**, 5701 (2000). 10, 29

[14] L. V. Manderscheid, "Significance levels — 0.05, 0.01, or?," J. Farm Economics **47**, 1381 (1965). 15

[15] J. O. Berger, L. D. Brown, and R. L. Wolpert, "A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing," Ann. Statist. **22**, 1787 (1994); see also http://www.stat.duke.edu/~berger/papers/brown.html. 17, 18, 19, 32

[16] R. L. Wolpert, "Testing simple hypotheses," in *Studies in classification, data analysis, and knowledge organization* **7** (H.H. Bock and W. Polasek, eds.), Springer-Verlag, Berlin (1996), pg. 289; see also http://ftp.stat.duke.edu/WorkingPapers/96-32.html. 17

[17] L. Demortier, "P values and nuisance parameters," in *Proceedings of the PhyStat LHC Workshop on Statistical Issues for LHC Physics*, ed. H.B. Prosper, L. Lyons, and A. De Roeck, CERN Yellow Report CERN-2008-001 (7 March 2008). 26

[18] R. E. Kass and A. E. Raftery, "Bayes factors," J. Amer. Statist. Assoc. **90**, 773 (1995). 29

[19] D. R. Cox, "Statistical significance tests," Br. J. Clin. Pharmac. **14**, 325 (1982). 30

[20] J. O. Berger, B. Boukai, and Y. Wang, "Unified frequentist and Bayesian testing of a precise hypothesis [with discussion]," Statist. Sci. **12**, 133 (1997); see also http://www.stat.duke.edu/~berger/papers/statsci.html. 33