Unified Bayesian and Conditional Frequentist Testing of Composite Hypotheses

Sarat C. Dass — University of Michigan
James O. Berger — Duke University

January 2, 1999

Abstract

Testing of a composite null hypothesis versus a composite alternative is considered when both have a related invariance structure. The goal is to develop conditional frequentist tests that allow the reporting of data-dependent error probabilities, error probabilities that have a strict frequentist interpretation and that reflect the actual amount of evidence in the data. The resulting tests are also seen to be Bayesian tests, in the strong sense that the reported frequentist error probabilities are also the posterior probabilities of the hypotheses under default choices of the prior distribution. The new procedures are illustrated in a variety of applications to model selection and multivariate hypothesis testing.

Key words and phrases. Conditional error probabilities; Bayes factors; Posterior probabilities; Default prior distributions; Nested hypotheses; Group invariance.

1 Introduction

This paper considers conditional frequentist and Bayesian unification in testing of composite hypotheses that have suitable invariance structures. A simple example is that in which the observations, $X_1, X_2, \ldots, X_n$, are i.i.d. f and it is desired to test $H_0 : f \; is$ Weibull (β, γ) versus $H_1 : f \; is$ Lognormal (μ, σ^2), where all parameters are unknown.

The classical frequentist approach to testing constructs acceptance and rejection regions and reports associated error probabilities of Type I and Type II. These error probabilities are *unconditional*, in the sense that they depend only on whether the data is in the rejection or acceptance region, and not on the evidentiary strength of the observed data. (Thus, for a normal test at level $\alpha = 0.05$, one reports the error probability of 0.05 whether $z = 2$ or $z = 10$.) The common 'solution' to this perceived shortcoming is to use p-values as data-dependent measures of the strength of evidence against H_0. However, a p-value is not a true error probability in the frequentist sense. Other shortcomings of p-values are discussed in Edwards, Lindman and Savage (1963), Berger and Sellke (1987), and Berger and Delampady (1987) .

To obtain data-dependent error probabilities having a proper frequentist interpretation it is natural to turn to the conditional frequentist approach, formalized by Kiefer (1975,1976,1977) and Brownie and Kiefer (1977). The idea behind this approach is to find a statistic measuring 'strength of evidence' in the data, for or against H_0, and report Type I and Type II error probabilities conditional on this statistic. The main difficulty in the approach is that of finding an appropriate choice for the conditioning statistic. In Kiefer (1977) and Brown (1978), admissibility considerations were employed to find "optimal" conditioning statistics, but this proved successful in identifying a conditioning statistic only in the simplest of testing problems, that of symmetric simple versus simple testing. Likewise, suitable ancillary statistics are rarely available for testing problems, so the conditional frequentist approach to testing has languished.

In the Bayesian approach to testing, reported error probabilities vary with the observed data and automatically reflect its evidentiary strength. However, it was long believed that Bayesian error probabilities are incompatible with frequentist error probabilities, leading to a sharp divide between Bayesians and frequentists over the issue of testing. (Note that Bayesian error probabilities can be close to p-values in one-sided testing - see Casella and Berger, 1987 - but, again, p-values do not have a direct interpretation as frequentist error probabilities; furthermore, we will concentrate on testing of null hypotheses that are more precise than the alternatives, in which case conditional frequentist or Bayesian error probabilities differ dramatically from p-values.)

For testing simple hypotheses, Berger, Brown and Wolpert (1994) and Wolpert (1996) found a conditioning statistic, $S(X)$, that reflects the evidentiary strength in the data and leads to a conditional frequentist test that is very easy to implement. Surprisingly, they observed that the ensuing conditional frequentist Type I and Type II error probabilities coincide exactly with the Bayesian posterior probabilities of H_0 and H_1, respectively. Therefore, a frequentist and a Bayesian using this test will not only reach the same decision (rejecting or accepting the null) after observing the data, but will also report the same values for the error probabilities. In this sense, the proposed test represents a unified testing procedure. Berger, Boukai and Wang (1997)generalized this to testing a simple null hypothesis versus a composite alternative. The sequential version of this problem was considered in Berger, Boukai and Wang (1998).

This paper considers the case in which both hypotheses are composite, the situation that arises most frequently in practice. Attention is restricted to problems in which the conditional frequentist Type I error probability will exactly equal the Bayesian posterior probability of H_0, so that the 'unification' obtains in a very strong sense. The situations in which this happens are situations in which there is a suitable invariance structure to the problem, so that the frequentist Type I error probability is constant over the null parameter space. This covers the majority of classical testing problems, however.

It should be mentioned that, while the original motivation for this work was to obtain unification between Bayesian and frequentist testing, the new conditional frequentist test has a number of compelling advantages, even when viewed solely from the frequentist perspective. One advantage has already been mentioned, namely that the error probabilities will vary with the evidentiary strength in the data. Another advantage is that the new test is simpler than unconditional tests, in several ways. First, its use in practice is considerably more straightforward, as is indicated by the application in Section 2. Second, there is, in a sense, only one general testing procedure that covers all testing problems (having suitable invariance structure). In particular, the procedure applies equally well to sequential problems, so that sequential testing is no harder than fixed sample size testing; among the many benefits is the elimination of complications such as 'spending alpha' in sequential clinical trials. (Indeed, the new conditional tests essentially follow the Stopping Rule Principle, eliminating another major perceived division between the Bayesian and frequentist schools; see Berger, Boukai, and Wang, 1997, for discussion.) Finally, from a pedagogical perspective, one no longer needs to be greatly concerned that a naive user might misinterpret a frequentist error probability as a posterior probability; here, they are the same.

Section 2 of the paper illustrates the new conditional test in the situation of testing a Weibull model versus a Lognormal model. Section 3 reviews the conditional frequentist and

Bayesian approaches for simple versus simple hypothesis testing, primarily to set notation. Section 4 discusses the general methodology of composite hypothesis testing in the presence of group structures. Several classical multivariate testing scenarios are considered in Section 5, in part to indicate the scope of applicability of the conditional testing method and, in part, to begin the process of redoing classical testing from the conditional perspective. As will become clear, this will be a major project involving objective Bayesians and conditional frequentists.

Section 6 considers the design problem of choosing an optimal sample size. At first sight, it might seem that design evaluations would be the same for conditional and unconditional testing; after all, before obtaining the data one can only perform an unconditional average over all possible data. What can change, however, is the design criterion itself. For instance, one might well desire to choose a sample size so that the reported conditional error probabilities are small, with specified certainty; if it is the conditional error probability that will be reported, ensuring that it is likely to be small is the natural goal. Designs based on such conditional criteria will, in general, be different than designs based on unconditional criteria.

2 An Application

2.1 The new conditional frequentist test for testing Weibull versus Lognormal models

The bulk of the paper uses group-theoretic language to describe the results, but the actual resulting methodology is easy to use. To illustrate this, we begin with an application to testing between the two-parameter Weibull and Lognormal distributions.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. f, where f is one of

$$H_0: \quad f_W(x;\beta,\gamma) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right], \quad x>0,\, \beta>0,\, \gamma>0 \tag{1}$$

$$H_1: \quad f_L(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp[-\frac{(\log x-\mu)^2}{2\sigma^2}], \quad x>0,\, \sigma>0. \tag{2}$$

Thus, (1) is the family of Weibull distributions with *unknown* parameters $\beta > 0$ and $\gamma > 0$, whereas (2) is the family of Lognormal distributions with *unknown* parameters $\mu \in R$ and $\sigma > 0$.

A (conventional) Bayesian approach to testing Weibull vs. Lognormal is described in

Section 4.1, leading to the test statistic

$$B_n = \frac{\Gamma(n)}{\Gamma((n-1)/2)} \sqrt{n} (n\pi)^{(n-1)/2} \int_0^\infty \left[\frac{\sigma}{n} \sum_{i=1}^{n} \exp\left(\frac{z_i - \bar{z}}{s_z \sigma} \right) \right]^{-n} d\sigma \, , \tag{3}$$

where $z_i = \log(x_i)$, $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ and $s_z^2 = \frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})^2$. The test statistic B_n is the **Bayes factor** of H_0 to H_1, based on the data $X_1, X_2, \ldots, X_n$ and standard non-informative priors. It is often regarded by Bayesians as the odds of H_0 to H_1 arising from the data, although this interpretation is not necessary for what follows. Indeed, it suffices here to simply view B_n as a specified test statistic.

Define the conditional test, T^*, based on B_n as follows:

$$T^* = \begin{cases} \text{if } B_n \leq r, & \text{reject } H_0 \text{ and report conditional error} \\ & \text{probability (CEP) } \alpha^*(B_n) = B_n/(1+B_n) \\ \text{if } r < B_n < a, & \text{make no decision} \\ \text{if } B_n \geq a, & \text{accept } H_0 \text{ and report conditional error} \\ & \text{probability (CEP) } \beta^*(B_n) = 1/(1+B_n), \end{cases} \tag{4}$$

where a and r are defined as in (20) and (21). It should be noted that the *no-decision* region, $r < B_n < a$, arises as an artifact of the analysis, rather than as a planned feature of the methodology; this region will be further discussed later.

It will be seen that T^* is an actual frequentist test; the reported CEPs, $\alpha^*(B_n)$ and $\beta^*(B_n)$, are conditional frequentist Type I and Type II error probabilities, conditional on a certain statistic measuring strength of evidence in the data. Furthermore, $\alpha^*(B_n)$ and $\beta^*(B_n)$ will be seen to have the Bayesian interpretation of being (objective) posterior probabilities of H_0 and H_1, respectively. Thus T^* is simultaneously a conditional frequentist and a Bayesian test.

Note that this new conditional test is easy to use. The conditional error probabilities are simple functions of B_n. The only potential computational difficulty is determining r and a but, for the examples we present in this paper, they were quite easily calculated via simulation. Note also that, while conditioning on the strength of evidence in the data is the underlying principle behind the procedure, the conditioning statistic does not directly appear in (4). In other words, this statistic can be viewed as part of the theoretical background of the procedure, but need not be described as part of the methodology in application.

Because T^* is also a Bayesian test, it inherits many of the positive features of Bayesian tests. Several of these features (e.g., not being affected by the stopping rule) were discussed in the introduction. Among other features, one that is perhaps of particular interest to frequentists is consistency: as the sample size grows, the test will eventually pick the right model, assuming the data comes from either a Lognormal or a Weibull distribution. (If the data actually arises from a third model, the test would pick the hypothesis which is closest in Kullback-Leibler divergence to the true model : Berk, 1966, Dmochowski, 1995.)

2.2 Application to car emission data

We illustrate application of the conditional test defined in (4) with data from McDonald et. al. (1995). The results obtained using T^* will also be compared with the results from classical tests.

McDonald et. al. (1995) studies three types of emissions from vehicles at four different mileage states. The types of emissions are hydrocarbon (HC), carbon monoxide (CO) and nitrogen oxide (NO_x) at mileage states 0, 4,000, 24,000 before maintenance and 24,000 after maintenance. There were 16 vehicles measured at each mileage state.

The results of testing using T^* are given in Table 1. The values of a and r for T^* were found to be 1.00 and 0.90, respectively. Therefore, the no-decision region is the interval [0.90,1.00]. Out of the 12 cases considered, only one value of B_n, 0.962, fell in the no-decision region. This corresponds to the emission of NO_x at 24,000 miles before maintenance. Note that a Bayesian using B_n in this case would reject H_0, but not with much confidence since the Bayes factor is very close to 1. Indeed, we have generally found that the no-decision region rarely arises and, when it does, the evidence is typically very weak for or against H_0.

McDonald et. al. (1995) discusses several classical tests to distinguish between the Weibull and the Lognormal distributions, including the Pearson chi-square goodness of fit test, the Kolmogorov-Smirnov test, the Srinivasan test, the Shapiro-Wilks test, the Ratio of Maximum Likelihood (RML) test, and the Mann-Scheuer-Fertig test. Critical values and powers of the tests were computed (via extensive simulation) at various levels of α, and are summarized in Table 2. Clearly the RML test is the most powerful among the classical tests considered. The results of the various tests for the emissions data are given in Table 3.

In comparing T^* with the classical tests, the most important feature to note is simply that the CEPs reflect the "strength of evidence" of the observed data. As an illustration, compare T^* and the Mann-Schuer test for the emissions of CO and NO_x at 4,000 miles. In both cases, the Mann-Schuer test rejects the Weibull model at fixed level 0.05. In contrast, the conditional test, T^*, reports a CEP of 0.099 for the emission of CO at 4,000 miles, but

Table 1: Decisions and conditional error probabilities (CEP) of T^*

	Mileage			
	0	4,000	24,000 (before maint.)	24,000 (after maint.)
	HC			
B_n	1.437	1.429	0.512	0.339
Decision	Weibull	Weibull	Lognormal	Lognormal
CEP	0.410	0.412	0.339	0.253
	CO			
B_n	0.406	0.111	0.161	0.410
Decision	Lognormal	Lognormal	Lognormal	Lognormal
CEP	0.288	0.099	0.139	0.291
	NO_x			
B_n	5.184	0.418	0.962	0.532
Decision	Weibull	Lognormal	No-decision	Lognormal
CEP	0.162	0.295	-	0.347

Table 2: Critical values and power of classical tests

	$\alpha = .20$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
	Critical values			
Chi-square	4.0	5.5	6.5	9.5
Kolmogorov-Smirnov	.159	.175	.191	.220
Srinivasan	.156	.174	.188	.218
Shapiro-Wilk	.937	.922	.901	.862
RML	1.006	1.038	1.067	1.140
	Power			
Chi-square	.282	.166	.117	.041
Kolmogorov-Smirnov	.448	.313	.212	.092
Srinivasan	.471	.330	.224	.096
Shapiro-Wilk	.568	.433	.306	.150
RML	.731	.605	.479	.238

a much larger CEP of 0.295 for NO_x at 4,000 miles. Obtaining such data-dependent error probabilities was our pimary goal.

The value of conditional error probabilities would have been even more apparent had we not 'cheated' in Table 3, by reporting rejection at different levels. The strict frequentist paradigm requires specification of the rejection region in advance, and does not allow reporting of any type of attained level of significance. It would have been permissible to, in advance, choose different levels for different tests, but one cannot do so after the fact and maintain a strict frequentist interpretation. Thus, in this example, someone using the CEP from T^* can claim that it is a strict frequentist error probability, whereas someone using attained levels from Table 3 cannot claim that they are frequentist error probabilities. This is not a pedantic distinction: true frequentist error probabilities (conditional or not) are typically much larger than attained levels, so that the common use of attained significance levels often produces a misleading sense of accuracy.

Another interesting distinction between T^* and the classical tests concerns the effect of the choice of the null model. This choice has no effect on T^*, which operates symmetrically between the hypotheses, but it can have a pronounced effect on classical tests. For example, consider the case of testing NO_x at 24,000 miles before maintenance. When the test is posed

Table 3: Conclusions of the classical tests

	Mileage			
	0	4,000	24,000 (before maint.)	24,000 (after maint.)
	HC			
	H_0 : **lognormal vs.** H_1**: weibull**			
Chi-square	R^1	A	R^2	A
K-S	R^1	R^1	A	A
Srinivasan	R^1	R^1	A	A
Shapiro-Wilk	R^1	R^2	A	A
RML	R^1	R^1	A	A
	H_0**: weibull vs.** H_1**: lognormal**			
RML	A	A	R^1	R^2
Mann-Schuer	A	A	R^1	R^1
	CO			
	H_0 : **lognormal vs.** H_1**: weibull**			
Chi-square	A	A	A	A
K-S	R^1	A	R^1	A
Srinivasan	R^1	A	R^1	A
Shapiro-Wilk	A	A	A	A
RML	A	A	A	A
	H_0**: weibull vs.** H_1**: lognormal**			
RML	R^1	R^{**}	R^*	R^1
Mann-Schuer	A	R^*	R^{**}	A

Table 3: Conclusions of the classical tests (cont.)

	Mileage			
	0	4,000	24,000 (before maint.)	24,000 (after maint.)
			NO_x	
	H_0 : **lognormal vs.** H_1**: weibull**			
Chi-square	A	A	A	A
K-S	A	A	A	A
Srinivasan	A	A	A	A
Shapiro-Wilk	R^2	A	A	A
RML	R^*	A	A	A
	H_0**: weibull vs.** H_1**: lognormal**			
RML	A	R^1	A	R^1
Mann-Schuer	A	R^*	A	A

A Accept null hypothesis

R^1 Reject null hypothesis at $\alpha = .20$ level

R^2 Reject null hypothesis at $\alpha = .10$ level

R^* Reject null hypothesis at $\alpha = .05$ level

R^{**} Reject null hypothesis at $\alpha = .01$ level as Lognormal vs. Weibull, the RML test accepts the null. When the test is posed as Weibull vs. Lognormal, the RML test again accepts the null. Of course, in classical testing, two 'acceptances' of the null simply means that the evidence for any one hypothesis is weak. This can be confusing to nonstatisticians, however. Such inconsistencies never arise with T^*; with weak evidence, T^* will either simply give a large CEP or will end up concluding no-decision, but this will not depend on which model is labeled the null hypothesis.

3 Conditional frequentist and Bayesian testing for simple hypotheses.

The primary purpose of this section is to introduce the notation needed for conditional frequentist and Bayesian testing. With little loss of efficiency in exposition, this can be done while reviewing the simple versus simple testing scenario, since the later developments will be based on reduction (in part through invariance) to this situation.

3.1 The conditional frequentist approach

Let X be a random variable, representing all the data, from the observation space $\mathcal{X}$ and consider testing the *simple* hypotheses

$$H_0 : X \sim m_0(x) \quad \text{versus} \quad H_1 : X \sim m_1(x), \tag{5}$$

where m_0 and m_1 are two specified probability density functions. Then

$$B(x) = \frac{m_0(x)}{m_1(x)} \tag{6}$$

is the *likelihood ratio* of H_0 to H_1 (or equivalently, the *Bayes factor in favor of* H_0). Let F_0 and F_1 be the c.d.f's of $B(X)$ under H_0 and H_1. The decision to accept or reject the null will be based on $B(x)$, where small values of $B(x)$ correspond to rejection of H_0.

The most powerful (unconditional) test of (5) is defined by a critical value c where

$$\begin{aligned} &\text{if } B(x) \le c, \quad \text{reject } H_0, \\ &\text{if } B(x) > c, \quad \text{accept } H_0. \end{aligned} \tag{7}$$

The unconditional frequentist Type I and Type II error probabilities are $\alpha = P_0(B(X) \le c) \equiv F_0(c)$ and $\beta = P_1(B(X) > c) \equiv 1 - F_1(c)$.

The conditional frequentist approach considers a statistic, $S(X)$, that represents evidentiary strength (for or against H_0), and then reports error probabilities conditional on

$S(X) = s$, where s denotes the observed value of $S(X)$. The resulting conditional error probabilities are

$$\begin{aligned} \alpha(s) &= \Pr(\text{Type I error} \mid S(X) = s) = P_0(B(X) \leq c \mid S(X) = s) \\ \beta(s) &= \Pr(\text{Type II error} \mid S(X) = s) = P_1(B(X) > c \mid S(X) = s). \end{aligned} \tag{8}$$

Thus (7) becomes

$$\begin{aligned} &\text{if } B(x) \leq c, \quad \text{reject } H_0 \text{ and report conditional error probability } \alpha(s) \\ &\text{if } B(x) > c, \quad \text{accept } H_0 \text{ and report conditional error probability } \beta(s). \end{aligned} \tag{9}$$

3.2 The Bayesian approach

For Bayesian testing of (5), one specifies prior probabilities for H_0 and H_1; here we will take them to each be 1/2. (See Berger, Brown, and Wolpert, 1994, for discussion of conditional frequentist testing corresponding to more general choices of these prior probabilities.) Then the posterior probability (given the data x) of H_0 is

$$P(H_0|x) = \alpha^*(B(x)) \equiv \frac{B(x)}{1 + B(x)} \tag{10}$$

and the posterior probability of H_1 is

$$P(H_1|x) = \beta^*(B(x)) \equiv \frac{1}{1 + B(x)}. \tag{11}$$

The standard Bayesian test for this situation can be written as

$$T: \begin{cases} \text{if } B(x) \leq 1, & \text{reject } H_0 \text{ and report} \\ & \quad \text{posterior error probability } \alpha^*(B(x)). \\ \text{if } B(x) > 1, & \text{accept } H_0 \text{ and report} \\ & \quad \text{posterior error probability } \beta^*(B(x)). \end{cases} \tag{12}$$

3.3 The new conditional test

Berger, Brown and Wolpert (1994) and Wolpert (1996) chose, as the conditioning statistic to measure strength of evidence in the data,

$$S(X) = \min\{B(X), \psi^{-1}(B(X))\}, \tag{13}$$

where ψ is defined in (19) over the domain $\mathcal{X}_d = \{X : 0 \leq S(X) \leq r\}$ and r is defined in (20) and (21). The intuition behind this statistic being a measure of the strength of evidence in the data is discussed in Berger, Boukai, and Wang (1997). Note that this conditioning statistic is only defined on $\mathcal{X}_d$, so that the complement of this region was termed the *no-decision* region. The resulting conditional frequentist test, T^B, was shown to be

$$T^B : \begin{cases} \text{if } B(x) \leq r & \text{reject } H_0 \text{ and report CEP } \alpha^*(B(x)), \\ \text{if } r < B(x) < a & \text{make no decision,} \\ \text{if } B(x) \geq a & \text{accept } H_0 \text{ and report CEP } \beta^*(B(x)), \end{cases} \tag{14}$$

where a is defined in (20) and (21).

The surprise here is that this conditional frequentist test is the same as the Bayes test in (12), except in the no-decision region. As mentioned earlier, however, data in the no-decision region rarely occurs and, when it does, the evidence for either hypothesis is typically very weak.

4 General Methodology

4.1 Composite hypotheses testing

This section generalizes the unified testing theory to the testing of two composite hypotheses that have the same group structure. Section 4.2 will consider the case when the alternative hypothesis contains additional parameters. For definitions and explanations of group-theoretic terms in bold type, see the Appendix.

Suppose $X = (X_1, X_2, \ldots X_n)$, where the X_i are i.i.d. f and we are interested in testing the composite hypotheses

$$H_0 : f = f_0(\cdot|\theta_0),\, \theta_0 \in \Theta_0 \quad \text{versus} \quad H_1 : f = f_1(\cdot|\theta_1),\, \theta_1 \in \Theta_1, \tag{15}$$

where θ_0 and θ_1 are the (unknown) nuisance parameters, and f_0 and f_1 are group invariant densities. Furthermore, we require that the group acting on H_0 and H_1, G say, be the same so that each family, f_0 and f_1, is **G-invariant**. The action of G on the observation space induces groups $\bar{G}_0$ and $\bar{G}_1$ acting on Θ_0 and Θ_1, respectively. It will be assumed that $\bar{G}_0$ and $\bar{G}_1$ are **transitive** on their respective domains. If, in addition, $\bar{G}_0$ and $\bar{G}_1$ have trivial **isotropy subgroups**, it follows that there is an **isomorphism** that maps Θ_0 to Θ_1. The family of densities in H_0 can, therefore, be parameterized by $\theta_1 \in \Theta_1$ instead of $\theta_0 \in \Theta_0$, and the action of the group $\bar{G}_0$ on Θ_0 can be replaced by the action of $\bar{G}_1$ on Θ_1. Subsequently, under this reparameterization, it can be assumed that the family of densities in (15) have the same parameter space Θ, with a common group $\bar{G}$ (arising from the action of G) acting on them.

For a set $A \in \bar{G}$ and $\bar{g}, \bar{h} \in \bar{G}$, the set $A \cdot \bar{g}$ denotes the right translate of A and the set $\bar{g} \cdot A$ denotes the left translate of A.

Definition 1 *A measure μ on $\bar{G}$ is said to be **relatively invariant** with **left multiplier** α_l and **right multiplier** α_r if $\mu(A \cdot \bar{g}) = \alpha_r(\bar{g}) \cdot \mu(A)$ and $\mu(\bar{g} \cdot A) = \alpha_l(\bar{g}) \cdot \mu(A)$.*

Note that $\alpha_r = 1$ and $\alpha_l = \Delta_l$ (the **left-hand moduli** of $\bar{G}$) corresponds to the right-Haar measure on $\bar{G}$. We will denote the right-Haar measure by ν. Similarly, $\alpha_r = \Delta_r$ (the **right-hand moduli** of $\bar{G}$) and $\alpha_l = 1$ corresponds to the left-Haar measure on $\bar{G}$. We denote the left-Haar measure on $\bar{G}$ by μ_L.

Define a function $\phi : \bar{G} \longrightarrow \Theta$ by $\phi(\bar{g}) = \bar{g} \circ e$, where e is the identity element of Θ. Since $\bar{G}$ is assumed to be transitive, the function ϕ is onto. Thus a prior μ on $\bar{G}$ induces a prior $\mu_\phi \equiv \mu\phi^{-1}$ on Θ. We will say that μ_ϕ is relatively invariant if μ is relatively invariant. As a special case, ν_ϕ will denote the prior induced by the right-Haar measure ν on $\bar{G}$.

For any relatively invariant prior μ_ϕ, the Bayes factor of H_0 to H_1 is

$$B(x) = \frac{\displaystyle\int_\Theta f_0(x|\theta)\, d\mu_\phi(\theta)}{\displaystyle\int_\Theta f_1(x|\theta)\, d\mu_\phi(\theta)}. \tag{16}$$

As a special case of (16), the Bayes factor corresponding to the right-Haar prior is given by

$$B^\nu(x) = \frac{\displaystyle\int_\Theta f_0(x|\theta)\, d\nu_\phi(\theta)}{\displaystyle\int_\Theta f_1(x|\theta)\, d\nu_\phi(\theta)}. \tag{17}$$

The following theorem, whose proof is standard and hence omitted, explains our interest in relatively invariant priors. The reader is referred to Eaton (1989) for a proof.

Theorem 1 *For a relatively invariant prior μ_ϕ, the distribution of B(X) in (16) does not depend on the nuisance parameter θ under H_0 or H_1.*

Denote by F_0^* and F_1^* the distribution function of $B(X)$ under H_0 and H_1, respectively, with densities f_0^* and f_1^*. By Theorem 1, F_0^* and F_1^* do not depend on θ_0 and θ_1, i.e., are completely known. Hence, if we choose to base the test on the statistic $B(X)$, the hypotheses in (15) reduce to the simple versus simple test of

$$H_0 : B(X) \sim f_0^* \qquad \text{versus} \qquad H_1 : B(X) \sim f_1^*. \tag{18}$$

The theory of Berger, Brown, and Wolpert (1994) can then be applied, as follows, to yield a unified conditional frequentist and Bayesian test.

Define the conditioning statistic as in (13), where

$$\psi(s) = F_0^{*\,-1}(1 - F_1^*(s)), \tag{19}$$

and let

$$r = 1,\ a = \psi(1) \quad \text{if} \quad F_0^*(1) < 1 - F_1^*(1) \tag{20}$$

$$r = \psi^{-1}(1),\ a = 1 \quad \text{if} \quad F_0^*(1) > 1 - F_1^*(1). \tag{21}$$

The conditional test in (14) then becomes

$$T^B = \begin{cases} \text{if } B(x) \le r & \text{reject } H_0 \text{ and report CEP} \\ & \alpha^*(B(x)) = f_0^*(B(x))/[f_1^*(B(x)) + f_0^*(B(x))], \\ \text{if } r < B(x) < a & \text{make no decision,} \\ \text{if } B(x) \ge a & \text{accept } H_0 \text{ and report CEP} \\ & \beta^*(B(x)) = f_1^*(B(x))/[f_1^*(B(x)) + f_0^*(B(x))]\ . \end{cases} \tag{22}$$

It can be concluded from Berger, Brown, and Wolpert (1994) that this is simultaneously a conditional frequentist test and a Bayesian test for (18).

There is a subtlety here on the Bayesian side, namely that the reduced problem in (18) need not be equivalent to the original testing problem in (15). The issue is that the posterior probability of H_0 given $B(x)$, namely $P(H_0|B(x)) = f_0^*(B(x))/[f_1^*(B(x)) + f_0^*(B(x))]$, is not necessarily the same as the original $P(H_0|x) = B(x)/(1 + B(x))$. In other words, the reduction to (18) may not be valid from a Bayesian perspective. The following theorem shows that all is well if the right-Haar prior distribution is used.

Theorem 2 *For the Bayes factor, B^ν, derived from the right-Haar prior,* $P(H_0|B^\nu(x)) = P(H_0|x)$ *and* $P(H_1|B^\nu(x)) = P(H_1|x)$.

Proof: See the Appendix.

Finally, note that $B^\nu(X)$ is the statistic used to construct the most powerful invariant test in classical testing. Thus reduction to (18), based on $B^\nu(X)$, is also natural from a frequentist perspective. Thus the recommended unified test, $T^{B^\nu} \equiv T^*$, is given by

$$T^* = \begin{cases} \text{if } B^\nu(x) \le r & \text{reject } H_0 \text{ and report CEP} \\ & \alpha^*(B^\nu(x)) = B^\nu(x)/(1 + B^\nu(x)) \\ \text{if } r < B^\nu(x) < a & \text{make no decision} \\ \text{if } B^\nu(x) \ge a & \text{accept } H_0 \text{ and report CEP} \\ & \beta^*(B^\nu(x)) = 1/(1 + B^\nu(x))\ . \end{cases} \tag{23}$$

This has the property that the reported CEPs are simultaneously (i) Type I and Type II error probabilities, conditional on $S(x)$, and (ii) Bayesian posterior probabilities of H_0 and H_1, respectively. Note that, implicit in this result is the fact that the conditional error probabilities are constant over the parameter spaces.

4.2 Testing when H_0 is nested in H_1

Now suppose that the alternative hypothesis contains additional parameters, i.e., the test is of

$$H_0 : f = f_0(\cdot|\theta_0) \quad \text{versus} \quad H_1 : f = f_1(\cdot|\theta_1, \xi), \tag{24}$$

where $\theta_0 \in \Theta_0, \theta_1 \in \Theta_1$ and $\xi \in \Omega$ are unknown parameters. As before, we assume that the group G acting on the families of densities is the same, and that Θ_0 has been reparameterized, if necessary, so that a common group $\bar{G}$, arising from the action of G, acts on the parameter spaces. Finally, assume that the parameterization of (θ_1, ξ) is such that ξ is not affected by the group action, i.e., that, for any $\bar{g}_1 \in \bar{G}_1$ and $(\theta_1, \xi) \in (\Theta_1, \Omega)$, $\bar{g}_1 \circ (\theta_1, \xi) = (\bar{g}_1 \circ \theta_1, \xi)$. It will often be necessary to reparameterize the problem to achieve this, as will be seen in the examples.

The prior density that will be considered for (θ_1, ξ) will be of the form $\pi(\theta_1, \xi) = \mu_\phi(\theta_1)\pi(\xi)$, where μ_ϕ is a relatively invariant prior on Θ_1 (and will also be the prior on Θ_0) and π is a proper prior on Ω. Typically, π will be chosen to be a conventional proper prior used for nested Bayesian testing (following Jeffreys, 1961). As in Berger, Boukai, and Wang (1997), one can integrate over ξ to form the 'marginal alternative' model

$$m_1(x|\theta_1) = \int_\Omega f_1(x|\theta_1, \xi)\, \pi(\xi)\, d\xi, \tag{25}$$

and then consider the test of

$$H_0 : f = f_0(x|\theta_0) \quad \text{versus} \quad H_1 : f = m_1(x|\theta_1). \tag{26}$$

Because of the choice of parameterization, it is easy to see that m_1 retains the original group action on Θ_1, so this reduced testing situation is of exactly the form discussed in Section 4.1. Hence the corresponding unified conditional frequentist tests, T^B and T^*, can be constructed, based on $B(x)$ defined as in (16) and (17), respectively, with $f_1(x|\theta)$ replaced by $m_1(x|\theta)$. The following theorems define the properties of these conditional tests in the original testing problem (24). The proof of Theorem 3 essentially follows the lines of the corresponding result in Berger, Boukai and Wang (1997), while that of Theorem 4 is essentially the same as the proof of Theorem 2.

Theorem 3 *For T^B as in (22), but defined for the testing problem in (26),*

$$P(H_0|B(x)) = P_{H_0}(\textit{Reject } H_0|S(x)) \tag{27}$$

$$P(H_1|B(x)) = E^{\pi(\xi|S(x))}\, P_\xi(\textit{Accept } H_0|S(x)), \tag{28}$$

where $P_\xi(\textit{Accept } H_0|S(x))$ is the conditional Type II error probability under (θ,ξ) in H_1 (which will depend only on ξ), and $\pi(\xi|S(x))$ is the conditional posterior distribution of ξ given $S(x)$.

Theorem 4 *For T^* as in (23), but defined for the testing problem in (26), $P(H_0|B^\nu(x)) = P(H_0|x)$ and $P(H_1|B^\nu(x)) = P(H_1|x)$.*

Note that the conditional Type I error probability for T^B is still exactly equal to the posterior probability of H_0 given $B(x)$. (Implicit, again, is the fact the conditional Type I error probability is constant over θ_0.) This equality is, in a sense, the main goal, since Type I error is often perceived to be of primary importance in classical statistics. Under the alternative, however, the conditional Type II error probability is no longer constant (it varies with ξ), so that it cannot equal the posterior probability of H_1 given $B(x)$. Interestingly, (28) shows that the posterior probability of H_1 given $B(x)$ is the *average* of the conditional Type II error probabilities, the averaging being done with respect to the posterior distribution of ξ conditional on $S(x)$. As the latter posterior can be thought of as describing where ξ is likely to be under the alternative, this *average power* is a very reasonable quantity to consider for a frequentist. See Berger, Boukai, and Wang (1997) for further discussion.

Finally, as in Section 4.1, we actually recommend use of T^*, since the posterior probabilities of H_i given $B(x)$ are then equal to the posterior probabilities of H_i given x.

Before proceeding to the examples, it should be mentioned that the above development of conditional frequentist tests is clearly dependent on the existence of suitable default or conventional Bayesian tests. Indeed, we have chosen examples for which such default Bayesian procedures exist. The development of default Bayesian procedures for important testing situations thus defines a joint research agenda for Bayesians and conditional frequentists.

5 Examples

5.1 Testing Weibull versus Lognormal

This is the testing problem discussed in Section 2.1. The group G, acting on the observation space, $\mathcal{X}$, that leaves both H_0 and H_1 invariant is

$$G = \{g_{b,c} : g_{b,c}(x) = b\, x^c,\ \ b > 0, c > 0\}. \tag{29}$$

Then, $\bar{G}_0$ and $\bar{G}_1$ are, respectively,

$$\bar{G}_0 = \{\bar{g}_{0,(b,c)} : \bar{g}_{0,(b,c)}(\beta, \gamma) = (\, b \cdot \beta^c, \gamma / c\,)\} \tag{30}$$

and

$$\bar{G}_1 = \{\bar{g}_{1,(b,c)} : \bar{g}_{1,(b,c)}(\mu, \sigma) = (\, c\mu + \log(b),\ c \cdot \sigma\,)\}. \tag{31}$$

The reparameterization of (β, γ) to (μ, σ), given by $\mu = \log(\beta)$ and $\sigma = 1/\gamma$, yields the same group action on each parameter space. The right-Haar prior induced by this group action on (μ, σ) is $\nu_\phi(\mu, \sigma) = 1/\sigma$, which is the well-known right-Haar prior for location-scale problems. Using (17), the Bayes factor for this problem reduces to B_n in (3). Hence, T^* in (4) is the unified Bayesian-conditional frequentist test with $\alpha^*(B_n)$ and $\beta^*(B_n)$, respectively, being (i) posterior probabilities of H_0 and H_1, and (ii) Type I and Type II error probabilities conditional on $S(x)$.

The no-decision regions corresponding to various sample sizes are given in Table 4 and are quite innocuous, usually arising only when the Bayes factor is in a small region near one, which would indicate weak evidence in any case. These were computed by simulation of the distribution of $B(X)$ under both H_0 and H_1. The results here were based on 1000 generated values of the random variable.

5.2 Testing Exponential versus Lognormal

Let $X_1, X_2, \ldots, X_n$ be i.i.d. from f and consider testing

$$H_0 : f \textit{ is } \text{Exponential}(\theta) \qquad \text{versus} \qquad H_1 : f \textit{ is } \text{Lognormal}(\mu, \sigma^2).$$

This is an example in which H_1 contains an additional parameter and the analysis of Section 4.2 must be employed. The group acting on the observation space, $\mathcal{X}$, is the multiplicative group given by $G = \{\, g_c : g_c(x) = c\, x,\ c > 0\}$. Using the notation of Section 4.2, we can

Table 4: The no-decision region for testing Weibull versus Lognormal.

Sample size, n	**No-decision region**
5	(0.91,1.00)
10	(0.93,1.00)
20	(0.83,1.00)
30	(0.68,1.00)
40	(0.75,1.00)
50	(0.82,1.00)

define $\Theta_0 = \{\theta : \theta > 0\}$, $\Theta_1 = \{\mu : \mu \in R\}$ and $\Omega = \{\sigma : \sigma > 0\}$, and the groups $\bar{G}_0$ and $\bar{G}_1$ are, respectively, $\bar{G}_0 = \{\bar{g}_c : \bar{g}_c(\theta) = c\theta\}$ and $\bar{G}_1 = \{\bar{g}_c : \bar{g}_c(\mu) = \log(c) + \mu\}$. Observe that σ is invariant under this group action, as was required of the parameterization. Furthermore, the transformation $\mu = \log(\theta)$ results in a common group action on $\bar{G}_0$ and $\bar{G}_1$, leading to the usual right-Haar prior $\nu_\phi(\theta) = 1/\theta$.

To complete the analysis, it is necessary to choose a proper prior, $\pi(\sigma)$, for the computation of (25) and the resulting Bayes factor. A proper conventional prior that has been suggested for this testing problem is the *intrinsic prior* of Berger and Pericchi (1996), given by

$$\pi(\sigma) = \sqrt{2}E\left(\frac{\mid Z \mid}{(1+\cosh(\sqrt{2}\sigma Z))}\right), \tag{32}$$

where the expectation is with respect to the standard normal random variable Z. The resulting Bayes factor of H_0 to H_1 can be shown to be

$$B(x) = \frac{\Gamma(n)\sqrt{2n}\pi^{n/2}\prod_{i=1}^n x_i}{\Gamma(n/2)(\sum_{i=1}^n x_i)^n} \cdot \left(\int_0^\infty \frac{1}{(1+\cosh(\sqrt{2v}))(v+\sum_{i=1}^n (y_i-\bar{y})^2)^{n/2}} dv\right)^{-1},$$

where $y_i = \log(x_i)$. Then (23) defines the unified conditional frequentist and Bayesian test.

Table 5 gives the no-decision region of this test for various values of n. These were computed by simulation from the distributions of $B(X)$ under H_0 and H_1, using 5000 generated values of the random variable. Again, the no-decision regions correspond to Bayes factors that would be considered very weak evidence.

Table 5: The no-decision region for testing Exponential versus Lognormal.

Sample size, n	**No-decision region**
10	(1.00,2.20)
20	(1.00,2.09)
30	(1.00,1.91)
40	(1.00,1.91)
50	(1.00,1.82)
60	(1.00,1.57)

5.3 Testing correlations in multivariate normal populations

Let $X_1, X_2, \ldots, X_N$ be i.i.d. p-vector observations from the normal population $N_p(\mu, \Sigma)$, p$\geq$2. Of interest is $\rho = \rho_{12\cdot 34\ldots p}$, the partial correlation of the first and second components given the others; that is, we want to know if there is an association between the first and second components after the linear dependence of the others has been eliminated. The test of no association versus association can be formulated as

$$H_0 : \ \rho = 0 \qquad \text{versus} \qquad H_1 : \ \rho \neq 0. \tag{33}$$

This testing problem is invariant under G, the group of affine transformations $X \longrightarrow AX + b$, where $b \in R^p$ and A is a matrix of the form

$$A_{p\times p} \quad = \quad \begin{pmatrix} D & U \\ 0 & C \end{pmatrix},$$

where $D_{2\times 2}$ is diagonal with positive entries d_1 and d_2, $U_{2\times(p-2)}$ is arbitary and $C_{(p-2)\times(p-2)}$ is non-singular. G induces a group of transformations $\bar{G}$ on the parameter space (μ, Σ) by

$$\bar{G} = \{(A, b) : (A, b) \circ (\mu, \Sigma) = (A\mu + b, A\Sigma A^{'})\}. \tag{34}$$

The right-Haar density on $\bar{G}$ is

$$\nu(b, D, U, C) \quad = \quad \frac{1}{|\det(D)|\,|\det(C)|^p}\, dD\, dC\, dU\, db.$$

The elements dD,dC, dU and db represent elements of Lebesgue measure on the respective spaces. To better understand the action of $\bar{G}$ on the parameter space, let

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{35}$$

where Σ_{11} is a 2×2 matrix, Σ_{12} is a $2 \times (p-2)$ matrix and Σ_{22} is a $(p-2) \times (p-2)$ matrix. Define $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}$. Then, $\Sigma_{11\cdot 2}$ can be represented by the entries

$$\Sigma_{11\cdot 2} = \begin{pmatrix} \sigma^2_{11\cdot 34\ldots p} & \rho\, \sigma_{11\cdot 34\ldots p}\, \sigma_{22\cdot 34\ldots p} \\ \rho\, \sigma_{11\cdot 34\ldots p}\, \sigma_{22\cdot 34\ldots p} & \sigma^2_{22\cdot 34\ldots p} \end{pmatrix}.$$

Reparameterize (μ, Σ) by $(\mu, \sigma_{11\cdot 34\ldots p}, \sigma_{22\cdot 34\ldots p}, \Sigma_{12}, \Sigma_{22}, \rho)$. The action of $\bar{G}$ in the new parameterization is

$$\begin{aligned} &(A, b) \circ (\, \mu,\, \sigma_{11\cdot 34\ldots p},\, \sigma_{22\cdot 34\ldots p},\, \Sigma_{12},\, \Sigma_{22},\, \rho\,) \\ &= (A\mu + b,\, d_1\sigma_{11\cdot 34\ldots p},\, d_2\sigma_{22\cdot 34\ldots p},\, D\Sigma_{12}C' + U\Sigma_{22}C',\, \rho\,). \end{aligned}$$

Using the notation in Section 4.2,

$$\Theta_0 = \Theta_1 = (\mu, \sigma_{11\cdot 34\ldots p}, \sigma_{22\cdot 34\ldots p}, \Sigma_{12}, \Sigma_{22}).$$

Under the null, $\rho = 0$ and $\Omega = \{\rho \neq 0\}$. The right-Haar prior ν_ϕ, induced by ν on $\bar{G}$, is

$$\nu_\phi(\sigma_{11\cdot 34\ldots p}, \sigma_{22\cdot 34\ldots p}, \Sigma_{12}, \Sigma_{22}) = \frac{1}{\sigma_{11\cdot 34\ldots p}} \cdot \frac{1}{\sigma_{22\cdot 34\ldots p}}\, d\sigma_{11\cdot 34\ldots p}\, d\sigma_{22\cdot 34\ldots p}\, d\Sigma_{12}\, \frac{1}{|\Sigma_{22}|^{(p+3)/2}}\, d\Sigma_{22}.$$

The prior on the 'extra parameter' ρ under H_1 will be chosen to be uniform, as suggested by Jeffreys (1961). This is proper, since the range of ρ is compact.

The Bayes factor, $B(x)$, is then given by

$$B(x) \;=\; \frac{2^{N-p+1}\, \Gamma^2((N-p+1)/2)}{\int_{-1}^{1} (1-\rho^2)^{-(N-p+1)/2} f(\rho, \hat{\rho})\, d\rho},$$

where $f(\rho, \hat{\rho})$ is given by

$$f(\rho, \hat{\rho}) \;=\; \int_0^\infty \int_0^\infty (yz)^{(N-p-1)/2} \exp\{-\frac{1}{2(1-\rho^2)}(\, y - 2\sqrt{yz}\, \hat{\rho}\, \rho + z)\}\, dy\, dz.$$

As in Jeffreys (1961), the substitution $yz = \alpha^2, \alpha > 0$, and $y/z = e^{2\beta}, \beta \in R$, allows this to be expressed as

$$B(x) = \frac{2^n \Gamma^2(n/2)}{\int_{-1}^{1} \int_0^\infty (1-\rho^2)^{n/2} (\cosh(\beta) - \rho\hat{\rho})^{-n}\, d\beta\, d\rho}, \tag{36}$$

where $n = N - p + 1$.

The test T^* based on $B(x)$ is given by (23), with a and r defined as in (20) and (21). The Bayes factor in (36) is a function only of $|\hat{\rho}|$, and decreases as $|\hat{\rho}|$ increases. Writing the distribution of $|\hat{\rho}|$ under H_0 and H_1 as $F_{0,|\hat{\rho}|}$ and $F_{1,|\hat{\rho}|}$, respectively, the test T^* can, alternatively, be given in terms of $\hat{\rho}$ as

$$T^* = \begin{cases} \text{if } |\hat{\rho}| \geq \rho_r & \text{reject } H_0 \text{ and report} \\ & \text{CEP } \alpha^*(B(\hat{\rho})) = B(\hat{\rho})/(1+B(\hat{\rho})) \\ \text{if } \rho_a < |\hat{\rho}| < \rho_r & \text{make no decision} \\ \text{if } |\hat{\rho}| \leq \rho_a & \text{accept } H_0 \text{ and report} \\ & \text{CEP } \beta^*(B(\hat{\rho})) = 1/(1+B(\hat{\rho}))\ , \end{cases}$$

where ρ_r and ρ_a satisfy, with $B(\rho^*) = 1$,

$$\int_0^{\rho_a \wedge \rho^*} F_{0,|\hat{\rho}|}(t)\, dt = \int_{\rho_r \vee \rho^*}^1 F_{1,|\hat{\rho}|}(t)\, dt.$$

5.4 Testing for equality of means of several multivariate populations

Suppose X_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, are independent observations from $N(\mu_i, \Sigma)$, where the μ_i's and Σ are unknown. We are interested in testing the following hypotheses:

$$H_0: \text{ all } \mu_i = \mu_0 \qquad \text{versus} \qquad H_1: \text{not } H_0.$$

Both H_0 and H_1 are invariant under G, the group of affine transformations $X \longrightarrow AX + b$, where A is non-singular and $b \in R^p$.

The conventional hierarchical prior that is recommended for this testing problem is specified as follows. Given μ_0 and Σ, let

$$\mu_i \overset{\text{iid}}{\sim} N_p(\mu_0, \tau\Sigma), \text{ for } i = 1, 2, \ldots, k,$$

$$\tau \sim g(\tau) = \tfrac{1}{\sqrt{2\pi}}\, \tau^{-3/2} \exp(-\tfrac{1}{2\tau}).$$

See Berger, Boukai and Wang (1997) for discussion of this prior when the null hypothesis is that all the means $\mu_i, i = 1, 2, \ldots, k$, are equal to zero. It is natural to also use this when, conditional on μ_0 and Σ, the means all equal μ_0.

$\bar{G}$ is the group of transformations on (μ_0, Σ) given by

$$\bar{G} = \{(A, b) : (A, b) \circ (\mu_0, \Sigma) = (A\mu_0 + b, A\Sigma A^{'})\}.$$

The right-Haar measure on $\bar{G}$ is

$$\nu(A, b) = \frac{1}{|A|^p}\, dA\, db.$$

The induced prior, ν_ϕ, on (μ_0, Σ) is

$$\nu_\phi(\mu_0, \Sigma) = \frac{1}{|\Sigma|^{(p+1)/2}}\, d\Sigma\, d\mu_0 .$$

Note that $\tau = 0$ corresponds to H_0 while $\tau > 0$ corresponds to H_1.

From (17), the Bayes factor is

$$B(X) = \left(\int_0^\infty \frac{|\det(S_0)|^{(n-1)/2}}{|\det(S_\tau)|^{(n-1)/2}} \cdot \frac{1}{(1+n\tau)^{p(k-1)/2}} \cdot g(\tau)\, d\tau \right)^{-1},$$

where

$$\begin{aligned} S_0 &= \sum_{i=1}^{k}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' + n \sum_{i=1}^{k} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \\ S_\tau &= \sum_{i=1}^{k}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' + \frac{n}{1+n\tau} \sum_{i=1}^{k} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})'. \end{aligned}$$

The unified test, T^*, is then given in (23), with a and r calculated using formulas (20) and (21).

6 Design Aspects

6.1 Motivation

In designing an experiment for which a conditional test is to be used, it is natural to incorporate the conditional nature of the inference into the design criterion. In this section we illustrate this notion on the simplest design problem, that of choosing the sample size n.

Denote, by CEP_i, the reported conditional error probability when H_i is true, $i = 0, 1$. Thus, under H_0, $CEP_0 = 1/(1 + B(x))$ and, under H_1, $CEP_1 = B(x)/(1 + B(x))$. For specified values of $0 < p_i < 1$ and $\alpha_i > 0$, it is attractive to choose the smallest sample size such that

$$P_{H_0}\{\, CEP_0 < \alpha_0 \,\} \geq p_0 \tag{37}$$

and

$$P_{H_1}\{\, CEP_1 < \alpha_1 \,\} \geq p_1. \tag{38}$$

If (37) holds then, under the null hypothesis and with a confidence of p_0, we pre-experimentally expect to report CEP_0 less than α_0. Likewise, (38) implies that we are confident, under H_1, that the reported CEP_1 will be less than α_1. Typically, p_0 and p_1 will be close to 1

Table 6: The smallest n achieving the design goal, for testing Weibull versus Lognormal.

	Values of p				
B	0.80	0.85	0.90	0.95	0.99
5	58	69	84	113	170
10	72	83	96	122	190
100	110	120	139	170	236

and α_0, α_1 small. This pre-experimental assurance, that the reported error will be small, is precisely the type of assurance that practitioners are likely to find appealing.

Allowing p_0 and p_1 to differ and α_0 and α_1 to differ enables differing treatments of Type I and Type II errors. In the following examples, however, we simply choose $\alpha_0 = \alpha_1 = \alpha$ and $p_0 = p_1 = p$. Furthermore, instead of choosing

various values of α, we choose values of $B = \alpha^{-1} - 1$, which are the corresponding desired evidence levels in terms of Bayes factors or odds.

Testing Weibull versus Lognormal: This is the example considered in Section 2 and Section 5.1. Table 6 gives the smallest values of n for which the indicated design goals would be achieved. Thus, if one desired to ensure that, with pre-experimental probability 0.90, the conclusion of the study would report odds of at least 10 to 1 in favor of one the models, it would be necessary to choose a sample size of at least $n = 96$. The computations in Table 6 were performed by simulation, using 1000 generated values of $B(X)$.

Testing Exponential versus Lognormal: For the testing problem of Section 5.2, the sample size needed to achieve the design goal is given in Table 7. These numbers were computed using 5000 generated values of $B(X)$.

Acknowledgements

This work formed part of the first author's Ph.D. thesis at Purdue University, and was supported by the National Science Foundation, Grants DMS-9303556 and DMS-9802261, and by a Purdue Research Foundation grant.

Table 7: The smallest n achieving the design goal, for testing Exponential versus Lognormal.

	Values of p				
B	0.80	0.85	0.90	0.95	0.99
5	27	34	45	68	122
10	40	45	58	81	142
100	73	84	100	125	181

Appendix: Group-theoretic definitions and proof of Theorem 3

Definition 2 *Let G be a group of measurable one-to-one transformations of $\mathcal{X}$ onto itself, $g : x \longrightarrow g \circ x$, such that the family $\{P_\theta : \theta \in \Theta\}$ is closed with respect to this transformation, i.e., for $x \sim P_\theta, g \circ x \sim P_{\theta'}$ for some $\theta^{'} \in \Theta$, defined by $\theta^{'} \equiv \bar{g} \circ \theta$. In this case, the family $\{P_\theta\}$ is **G-invariant**.*

The action of the group G on $\mathcal{X}$ induces another group $\bar{G}$ on the parameter space Θ. Actually it can be shown that G and $\bar{G}$ are isomorphic to each other.

Definition 3 *For a group of measurable transformations G acting on a space $\mathcal{X}$, G is said to be transitive on $\mathcal{X}$ if for any $x, x^{'} \in \mathcal{X}$, there is a g in G such that $x^{'} = g \circ x$.*

Definition 4 *The isotropy subgroup of G at x is the subgroup*

$$G_x = \{g \in G : g \circ x = x\}. \tag{39}$$

The isotropy subgroup of G is said to be trivial if $G_x = e, \forall x \in \mathcal{X}$. Define a function $\phi : \bar{G} \longrightarrow \mathcal{X}$ by $\phi(\bar{g}) = \bar{g} \circ e$, where e is the identity element of $\mathcal{X}$. Then, transitivity of $\bar{G}$ on $\mathcal{X}$ implies that ϕ is onto. If, furthermore, G has a trivial isotropy subgroup, then the function ϕ is also one-to-one and, in that case, it is an automorphism of G and $\mathcal{X}$.

Definition 5 *Let G be a group of measurable transformations of $\mathcal{X}$ onto itself. A **maximal invariant**, $\tau(x)$, is a function on $\mathcal{X}$ satisfying*

- *$\tau(x)$ is invariant under G, i.e., $\tau(g \circ x) = \tau(x)$ for $g \in G$ and $x \in \mathcal{X}$*

- *$\tau(x)$ takes different values on different orbits of G, i.e. $\tau(x_1) = \tau(x_2) \Rightarrow x_1 = gx_2$ for some $g \in G$.*

Definition 6 *The **left-hand moduli** of G, Δ_l, is such that, for any right-Haar measure ν, $\nu(g \cdot A) = \Delta_l(g)\nu(A) \;\; \forall g \in G$. Similarly, the **right-hand moduli** of G, Δ_r, is such that, for any left-Haar measure μ_L, $\mu_L(A \cdot g) = \Delta_r(g)\mu_L(A) \;\; \forall g \in G$.*

Δ_l and Δ_r are special examples of multipliers. A continuous function $\alpha(\cdot)$, $\alpha : G \to R^+$ is said to be a **multiplier** if, $\forall\, g, h \in G$, $\alpha(g \cdot h) = \alpha(g) \cdot \alpha(h)$.

Assume the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ admits a family of densities $\{p(\cdot|\theta) : \theta \in \Theta\}$ with respect to a σ-finite dominating measure λ. Let $G, \bar{G}$ act on $\mathcal{X}, \Theta$, respectively, and λ be relatively invariant under G with multiplier $\chi(\cdot)$. If the densities satisfy

$$p(x|\theta) = p(gx|\bar{g}\theta)\chi(g), \tag{40}$$

then the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is G-invariant. Furthermore, $gP_\theta = P_{\bar{g}\theta}$.

Definition 7 *A measure μ is said to be relatively invariant with left-multiplier α_l and right-multiplier α_r if, for any subset A of G and any $g \in G$, $\mu(g \cdot A) = \alpha_l(g)\mu(A)$ and $\mu(A \cdot g) = \alpha_r(g)\mu(A)$.*

Proof of Theorem 2: We first state a few basic results:

Theorem 5 *[Wijsman] For $i = 0, 1$, let P_i be a distribution on $\mathcal{X}$ with density p_i, with respect to a χ-relatively invariant measure λ. Let $\tau(x)$ be a maximal invariant with distributions P_0^τ and P_1^τ, respectively. Then P_0^τ and P_1^τ are absolutely continuous with respect to a dominating measure μ^τ and, for any $x \in \mathcal{X}$,*

$$\frac{dP_0^\tau}{dP_1^\tau}(\tau(x)) = \frac{\displaystyle\int p_0(gx)\chi(g)d\mu_L(g)}{\displaystyle\int p_1(gx)\chi(g)d\mu_L(g)} \;, \tag{41}$$

where μ_L is a left-invariant measure on G.

Proof: *See Wijsman (1990).*

Theorem 6 *For $i = 0, 1$, let $\mathcal{P}_i$ denote the family of distributions with densities $\{f_i(\cdot|\theta) : \theta \in \Theta\}$ with respect to a χ-relatively invariant measure λ. Also, assume that the class of* *densities* $\{f_i(\cdot|\theta) : \theta \in \Theta\}$*, for* $i = 0, 1$*, are* G*-invariant. Then,*

$$\frac{dP_0^\tau}{dP_1^\tau} = \frac{\int f_0(x|\theta) d\nu_\phi(\theta)}{\int f_1(x|\theta) d\nu_\phi(\theta)} \equiv B^\nu(x). \tag{42}$$

Proof:

$$\begin{aligned}
B^\nu(x) &\equiv \frac{\int f_0(x|\theta) d\nu_\phi(\theta)}{\int f_1(x|\theta) d\nu_\phi(\theta)} \\
&= \frac{\int f_0(x_2|\bar{g} \circ e) d\nu(\bar{g})}{\int f_1(x_2|\bar{g} \circ e) d\nu(\bar{g})} \\
&= \frac{\int f_0(g^{-1}x_2|e)\chi(g^{-1}) d\nu(\bar{g})}{\int f_1(g^{-1}x_2|e)\chi(g^{-1}) d\nu(\bar{g})} \\
&= \frac{\int f_0(gx_2|e)\chi(g) d\mu_L(g)}{\int f_1(gx_2|e)\chi(g) d\mu_L(g)} \\
&= \frac{dP_0^\tau}{dP_1^\tau}(\tau(x)) \quad (\text{ by Theorem 5 }),
\end{aligned}$$

where $d\nu(\bar{g}^{-1}) = d\mu_L(g)$. QED.

Theorem 7 *Let* F_0^* *and* F_1^* *be the c.d.f.s of* $B^\nu(X)$*, and denote their densities with respect to Lebesgue measure by* f_0^* *and* f_1^**. Then,*

$$\frac{f_0^*(b)}{f_1^*(b)} = b \quad \forall b > 0. \tag{43}$$

Proof:

$$\begin{aligned}
F_0^*(b) = \int_0^b f_0^*(t)dt &= \int_{\{x : B^\nu(x) \le b\}} dP_0^\tau(\tau(x)) \\
&= \int_{\{x : B^\nu(x) \le b\}} \frac{P_0^\tau(\tau(x))}{P_1^\tau(\tau(x))} \cdot dP_1^\tau(\tau(x)) \\
&= \int_{\{x : B^\nu(x) \le b\}} B^\nu(x) \cdot dP_1^\tau(\tau(x)) \qquad (\text{see Theorem 6}) \\
&= \int_0^b t \cdot f_1^*(t)dt.
\end{aligned}$$

Differentiating both sides w.r.t. b yields the result. QED.

To complete the proof of Theorem 2, note that

$$P(H_0|B^\nu(x)) = \frac{f_0^*(B^\nu(x))}{f_0^*(B^\nu(x)) + f_1^*(B^\nu(x))} = \frac{B^\nu(x)}{1 + B^\nu(x)} \tag{44}$$

and

$$P(H_1|B^\nu(X)) = \frac{f_1^*(B^\nu(x))}{f_0^*(B^\nu(x)) + f_1^*(B^\nu(x))} = \frac{1}{1 + B^\nu(x)}, \tag{45}$$

the final equalities in (44) and (45) following from (43).

References

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis (Second Edition).* Wiley.

Berger, J. O., Boukai, B. and Wang, W. (1997) Unified Frequentist and Bayesian testing of a precise hypothesis. *Statistical Sciences*, **12,** no. 3, 133–160.

Berger, J. O., Boukai, B. and Wang, Y. (1997) Properties of Unified Bayesian-Frequentist Tests. In *Advances in Statistical Decision Theory and Applications*, pp. 207–223. Birkhouser, Boston.

Berger, James O., Boukai, B. and Wang, Y. (1998) Simulteneous Bayesian-Frequentist Sequential Testing of Nested Hypotheses. *Biometrika.*

Berger, James O., Brown, Lawrence D. and Wolpert, Robert L. (1994) A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing. *The Annals of Statistics*, **22,** 1787–1807.

Berger, James O. and Delampady, Mohan (1987) Testing Precise Hypotheses (c/r: P335-352). *Statistical Science*, **2,** 317–335.

Berger, James O. and Pericchi, L. R. (1995) The Intrinsic Bayes Factor for linear models. In *Bayesian Statistics 5* (ed. et. al. J. M. Bernardo), pp. 23–42. Clarendon Press, Oxford.

Berger, James O. and Sellke, Thomas (1987) Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence (c/r: P123-133, 135-139, 1201-1201). *Journal of the American Statistical Association*, **82,** 112–122.

Berk, Robert H. (1966) Limiting Behavior of Posterior Distributions when the Model is Incorrect (corr: V37 p745). *The Annals of Mathematical Statistics*, **37,** 51–58.

Brown, Lawrence D. (1978) A Contribution to Kiefer's Theory of Conditional Confidence Procedures. *The Annals of Statistics*, **6,** 59–71.

Dmochowski, J. (1995) Properties of Intrinsic Bayes Factors. Ph.D. Thesis. Purdue University, W. Lafayette.

Eaton, Morris L. (1983) *Multivariate Statistics: A Vector Space Approach.* Wiley, New York.

Eaton, Morris L. (1989) *Group Invariance Applications in Statistics.* Institute of Mathematical Statistics, Hayward, California.

Jeffreys, H. (1961) *Theory of Probability (3rd ed.).* Clarendon Press, Oxford.

Kass, Robert E. and Raftery, Adrian E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90,** 773–795.

Kass, Robert E. and Wasserman, Larry (1995) A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90,** 928–934.

Kiefer, J. (1977) Conditional Confidence Statements and Confidence Estimators (c/r: P808-827). *Journal of the American Statistical Association*, **72,** 789–807.

McDonald, Gary C., Vance, L. C. and Gibbons, D. I. (1995) Some Tests for Discriminating Between Lognormal and Weibull distributions - An Application to Emissions Data. In *Recent Advances in Life-Testing and Reliability - A Volume in honor of Alonzo Clifford Cohen, Jr.* (ed. N. Balakrishnan). CRC Press, Inc.

Wijsman, R. A. (1990) *Invariant Measures on Groups and their Uses in Statistics.* Institute of Mathematical Statistics, Hayward, California.

Wolpert, Robert L. (1996) Testing Simple Hypotheses. In *Data Analysis and Information Systems. Proceedings of the 19th Annual Conference of the Gesellschaft für Klassifikation e.V.*, pp. 289– 297. Springer, Berlin.