BAYESIAN STATISTICS 8, pp. 571–576.
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2007

Comparing Normal Means: New Methods for an Old Problem*

José M. Bernardo
Universitat de València, Spain
jose.m.bernardo@uv.es

Sergio Pérez
Colegio de Postgraduados, Mexico
sergiop@colpos.mx

Summary

Comparing the means of two normal populations is a very old problem in mathematical statistics, but there is still no consensus about its most appropriate solution. In this paper we treat the problem of comparing two normal means as a Bayesian decision problem with only two alternatives: either to accept the hypothesis that the two means are equal, or to conclude that the observed data are, under the assumed model, incompatible with that hypothesis. The combined use of an information-theory based loss function, the *intrinsic discrepancy* (Bernardo and Rueda, 2002), and an objective prior function, the *reference prior* (Bernardo, 1979; Berger and Bernardo, 1992), produces a new solution to this old problem which, for the first time, has the invariance properties one should presumably require.

Keywords and Phrases: Bayes factor, BRC, comparison of normal means, intrinsic discrepancy, Kullback–Leibler divergence, precise hypothesis testing, reference prior.

1. STRUCTURE OF THE DECISION PROBLEM

Precise hypothesis testing as a decision problem. Assume that available data $\boldsymbol{z}$ have been generated from an unknown element of the family of probability distributions for $\boldsymbol{z} \in \mathcal{Z}$, $\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}, \boldsymbol{\omega}),\ \boldsymbol{\phi} \in \Phi,\ \boldsymbol{\omega} \in \Omega\}$, and suppose that it is desired to check whether or not these data may be judged to be compatible with the (null) hypothesis $H_0 \equiv \{\boldsymbol{\phi} = \boldsymbol{\phi}_0\}$. This may be treated as a decision problem with only two alternatives; a_0: to accept H_0 (and work *as if* $\boldsymbol{\phi} = \boldsymbol{\phi}_0$) or a_1: to claim that the observed data are incompatible with H_0. Notice that, with this formulation, H_0 is generally a composite hypothesis, described by the family of probability distributions

* This is a synopsis. For the full version, see *Bayesian Analysis* **2** (2007), 571–576.

José M. Bernardo is Professor of Statistics at the Facultad de Matemáticas, Universidad de Valencia, 46100-Burjassot, Valencia, Spain. Dr. Sergio Pérez works at the Colegio de Postgraduados, Montecillo, 56230-Texcoco, Mexico. Work by Professor Bernardo has been partially funded with Grant MTM2006-07801 of the MEC, Spain.

$\mathcal{M}_0 = \{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0),\ \boldsymbol{\omega}_0 \in \Omega\}$. Simple nulls are included as a particular case where there are no nuisance parameters.

Foundations dictate (see *e.g.*, Bernardo and Smith, 1994, and references therein) that, to solve this decision problem, one must specify utility functions $u\{a_i, (\boldsymbol{\phi}, \boldsymbol{\omega})\}$ for the two alternatives a_0 and a_1, and a joint prior distribution $\pi(\boldsymbol{\phi}, \boldsymbol{\omega})$ for the unknown parameters $(\boldsymbol{\phi}, \boldsymbol{\omega})$; then, H_0 should be rejected if, and only if,

$$\int_{\Phi} \int_{\Omega} [u\{a_1, (\boldsymbol{\phi}, \boldsymbol{\omega})\} - u\{a_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\}]\, \pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{z})\, d\boldsymbol{\phi}\, d\boldsymbol{\omega} > 0,$$

where, using Bayes' theorem, $\pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{z}) \propto p(\boldsymbol{z} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\, \pi(\boldsymbol{\phi}, \boldsymbol{\omega})$ is the joint posterior which corresponds to the prior $\pi(\boldsymbol{\phi}, \boldsymbol{\omega})$. Thus, only the utilities difference must be specified, and this may usefully be written as

$$u\{a_1, (\boldsymbol{\phi}, \boldsymbol{\omega})\} - u\{a_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\} = \ell\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\} - u_0,$$

where $\ell\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\}$ is the non-negative terminal loss suffered by accepting $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ given $(\boldsymbol{\phi}, \boldsymbol{\omega})$, and $u_0 > 0$ is the utility of accepting H_0 when it is true. Hence, H_0 should be rejected if, and only if,

$$t(\boldsymbol{\phi}_0 \,|\, \boldsymbol{z}) = \int_{\Theta} \int_{\Omega} \ell\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\}\, \pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{z})\, d\boldsymbol{\phi}\, d\boldsymbol{\omega} > u_0,$$

that is, if the posterior expected loss, the *test statistic* $t(\boldsymbol{\phi}_0 \,|\, \boldsymbol{z})$ is large enough.

The intrinsic discrepancy loss. As one would expect, the optimal decision depends heavily on the particular loss function $\ell\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\}$ used. Specific problems may require specific loss functions, but conventional loss functions may be used to proceed when one does not have any particular application in mind.

A common class of conventional loss function are the *step* loss functions. These *forces* the use of a *non-regular* 'spiked' proper prior which places a lump of probability at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ and leads to rejecting H_0 if, and only if, its posterior probability is too small or, equivalently, if, and only if, the *Bayes factor* against H_0, is sufficiently large. This will be appropriate wherever preferences are well described by a step loss function, and prior information is available to justify a (*highly informative*), spiked prior. It may be argued that many scientific applications of precise hypothesis testing fail to meet one or both of these conditions.

Another example of a conventional loss function is the ubiquitous *quadratic* loss function. This leads to rejecting the null if, and only if, the posterior expected Euclidean distance of $\boldsymbol{\phi}_0$ from the true value $\boldsymbol{\phi}$ is too large, and may safely be used with (typically improper) 'noninformative' priors. However, as most conventional continuous loss functions, the quadratic loss depends dramatically on the particular parametrization used. But, since the model parametrization is arbitrary, the conditions to reject $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ should be *precisely the same*, for any one-to-one function $\boldsymbol{\psi}(\boldsymbol{\phi})$, as the conditions to reject $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\phi}_0)$. This requires the use of a loss function which is invariant under one-to-one reparametrizations.

It may be argued (Bernardo and Rueda, 2002; Bernardo, 2005b) that a measure of the disparity between two probability distributions which may be appropriate for general use in probability and statistics is the *intrinsic discrepancy*,

$$\delta_{\boldsymbol{z}}\{p_{\boldsymbol{z}}(\cdot), q_{\boldsymbol{z}}(\cdot)\} \equiv \min\left\{\int_{\mathcal{Z}} p_{\boldsymbol{z}}(\boldsymbol{z}) \log\frac{p_{\boldsymbol{z}}(\boldsymbol{z})}{q_{\boldsymbol{z}}(\boldsymbol{z})}\, d\boldsymbol{z},\ \int_{\mathcal{Z}} p_{\boldsymbol{z}}(\boldsymbol{z}) \log\frac{p_{\boldsymbol{z}}(\boldsymbol{z})}{q_{\boldsymbol{z}}(\boldsymbol{z})}\, d\boldsymbol{z}\right\}, \quad (1)$$

defined as the *minimum* (Kullback-Keibler) logarithmic divergence between them. This is *symmetric*, *non-negative*, and it is zero if, and only if, $p_{\boldsymbol{z}}(\boldsymbol{z}) = q_{\boldsymbol{z}}(\boldsymbol{z})$ a.e. Besides, it is *invariant* under one-to-one transformations of $\boldsymbol{z}$, and it is *additive* under independent observations; thus if $\boldsymbol{z} = \{x_1, \ldots, x_n\}$, $p_{\boldsymbol{z}}(\boldsymbol{z}) = \prod_{i=1}^n p_x(x_i)$, and $q_{\boldsymbol{z}}(\boldsymbol{z}) = \prod_{i=1}^n q_x(x_i)$, then $\delta_{\boldsymbol{z}}\{p_{\boldsymbol{z}}(\cdot), q_{\boldsymbol{z}}(\cdot)\} = n\, \delta_x\{p_x(\cdot), q_x(\cdot)\}$.

Within a parametric probability model, say $\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, the intrinsic discrepancy induces a loss function $\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \delta_{\boldsymbol{z}}\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_0), p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta})\}$, in which the loss to be suffered if $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}_0$ is not measured terms of the disparity between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, but in terms of the disparity between the *models* labelled by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. This provides a loss function which is *invariant* under reparametrization: for any one-to-one function $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$, $\delta\{\boldsymbol{\psi}_0, \boldsymbol{\psi}\} = \delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}$. Moreover, one may equivalently work with sufficient statistics: if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ is a sufficient statistic for $p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta})$, then $\delta_{\boldsymbol{z}}\{\boldsymbol{\theta}_0, \boldsymbol{\theta})\} = \delta_{\boldsymbol{t}}\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}$. The intrinsic loss may be safely be used with improper priors. In the context of hypothesis testing within the parametric model $p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})$, the intrinsic loss to be suffered by replacing $\boldsymbol{\phi}$ by $\boldsymbol{\phi}_0$ becomes

$$\delta_{\boldsymbol{z}}\{H_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\} \equiv \inf_{\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}} \delta_{\boldsymbol{z}}\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0),\, p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\}, \tag{2}$$

that is, the intrinsic discrepancy between the distribution $p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})$ which has generated the data, and the family of distributions $\mathcal{F}_0 \equiv \{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0), \boldsymbol{\omega}_0 \in \Omega\}$ which corresponds to the hypothesis $H_0 \equiv \{\boldsymbol{\phi} = \boldsymbol{\phi}_0\}$ to be tested. If, as it is usually the case, the parameter space $\boldsymbol{\Phi} \times \boldsymbol{\Omega}$ is convex, then the two minimization procedures in (2) and (1) may be interchanged (Juárez, 2005).

As it is apparent from its definition, the intrinsic loss (2) is the *minimum conditional expected log-likelihood ratio* (under repeated sampling) against H_0, what provides a direct *calibration* for its numerical values; thus, intrinsic loss values of about log(100) would indicate rather strong evidence against H_0.

The Bayesian Reference Criterion (BRC). Any statistical procedure depends on the accepted assumptions, and those typically include many subjective judgements. If has become standard, however, to term 'objective' any statistical procedure whose results only depend on the quantity of interest, the model assumed and the data obtained. The *reference prior* (Bernardo, 1979; Berger and Bernardo, 1992; Bernardo, 2005a), loosely defined as that prior which maximizes the missing information about the quantity of interest, provides a general solution to the problem of specifying an objective prior. See Berger (2006) for a recent analysis of this issue.

The Bayesian reference criterion (Bernardo and Rueda, 2002) is the normative Bayes solution to the decision problem of hypothesis testing described above which corresponds to the use of the *intrinsic loss* and the *reference prior.* Given a parametric model $\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\phi}, \boldsymbol{\omega}),\ \boldsymbol{\phi} \in \Phi,\ \boldsymbol{\omega} \in \Omega\}$, this prescribes to reject the hypothesis $H_0 \equiv \{\boldsymbol{\phi} = \boldsymbol{\phi}_0\}$ if, and only if,

$$d(H_0 \,|\, \boldsymbol{z}) = \int_0^\infty \delta\, \pi(\delta \,|\, \boldsymbol{z})\, d\delta > \delta_0, \tag{3}$$

where $d(H_0 \,|\, \boldsymbol{z})$, termed the *intrinsic (test) statistic*, is the reference posterior expectation of the intrinsic loss $\delta_{\boldsymbol{z}}\{H_0, (\boldsymbol{\phi}, \boldsymbol{\omega})\}$ defined by (2), and where δ_0 is a context dependent positive utility constant, *the largest acceptable average log-likelihood ratio against* H_0 *under repeated sampling.* For scientific communication, δ_0 could conventionally be set to $\log(10) \approx 2.3$ to indicate some evidence against H_0, and to $\log(100) \approx 4.6$ to indicate strong evidence against H_0.

2. NORMAL MEANS COMPARISON

Problem statement in the common variance case. Let available data $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{y}\}$, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $\boldsymbol{y} = \{y_1, \ldots, y_m\}$, consist of two random samples of possibly different sizes n and m, respectively drawn from $\mathrm{N}(x \,|\, \mu_x, \sigma)$ and $\mathrm{N}(y \,|\, \mu_y, \sigma)$, so that the assumed model is $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma) = \prod_{i=1}^{n} \mathrm{N}(x_i \,|\, \mu_x, \sigma) \prod_{j=1}^{m} \mathrm{N}(y_j \,|\, \mu_y, \sigma)$. It is desired to test $H_0 \equiv \{\mu_x = \mu_y\}$, that is, whether or not these data could have been drawn from some element of the family $\mathcal{F}_0 \equiv \{p(\boldsymbol{z} \,|\, \mu_0, \mu_0, \sigma_0), \mu_0 \in \Re,\ \sigma_0 > 0\}$. To implement the BRC criterion described above one should: (i) compute the *intrinsic discrepancy* $\delta\{H_0, (\mu_x, \mu_y, \sigma)\}$ between the family $\mathcal{F}_0$ which defines the hypothesis H_0 and the assumed model $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)$; (ii) determine the *reference* joint prior $\pi_\delta(\mu_x, \mu_y, \sigma)$ of the three unknown parameters when δ is the quantity of interest; and (iii) derive the relevant *intrinsic statistic*, that is the reference posterior expectation $d(H_0 \,|\, \boldsymbol{z}) = \int_0^\infty \delta\, \pi_\delta(\delta \,|\, \boldsymbol{z})\, d\delta$ of the intrinsic discrepancy $\delta\{H_0, (\mu_x, \mu_y, \sigma)\}$.

The intrinsic loss. The (Kullback–Leibler) logarithmic divergence of a normal distribution $\mathrm{N}(x \,|\, \mu_2, \sigma_2)$ from another normal distribution $\mathrm{N}(x \,|\, \mu_1, \sigma_1)$ is given by

$$\begin{aligned}\kappa\{\mu_2, \sigma_2 \,|\, \mu_1, \sigma_1\} &\equiv \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu_1, \sigma_1) \log \frac{\mathrm{N}(x \,|\, \mu_1, \sigma_1)}{\mathrm{N}(x \,|\, \mu_2, \sigma_2)}\, dx \\ &= \frac{1}{2}\left(\frac{\mu_2 - \mu_1}{\sigma_2}\right)^2 + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log \frac{\sigma_1^2}{\sigma_2^2}\right). \qquad (4)\end{aligned}$$

Using its additive property, the (KL) logarithmic divergence of $p(\boldsymbol{z} \,|\, \mu_0, \mu_0, \sigma_0)$ from $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)$ is $n\,\kappa\{\mu_0, \sigma_0 \,|\, \mu_x, \sigma\} + m\,\kappa\{\mu_0, \sigma_0 \,|\, \mu_y, \sigma\}$, which is minimized when $\mu_0 = (n\mu_x + m\mu_y)/(n + m)$ and $\sigma_0 = \sigma$. Substitution yields $h(n, m)\,\theta^2/4$, where $h(n, m) = 2nm/(m + n)$ is the harmonic mean of the two sample sizes, and $\theta = (\mu_x - \mu_y)/\sigma$ is the standardized distance between the two means. Similarly, the logarithmic divergence of $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)$ from $p(\boldsymbol{z} \,|\, \mu_0, \mu_0, \sigma_0)$ is given by $n\,\kappa\{\mu_x, \sigma \,|\, \mu_0, \sigma_0\} + m\,\kappa\{\mu_y, \sigma \,|\, \mu_0, \sigma_0\}$, minimized when $\mu_0 = (n\mu_x + m\mu_y)/(n + m)$ and $\sigma_0^2 = \sigma^2 + (\mu_x - \mu_y)^2\,(mn)/(m + n)^2$. Substitution now yields a minimum divergence $(n + m)/2 \log[1 + h(n, m)/(2(n + m))\,\theta^2$, which is always smaller than the minimum divergence $h(n, m)\,\theta^2/4$ obtained above. Therefore, the required intrinsic loss function is

$$\delta_{\boldsymbol{z}}\{H_0, (\mu_x, \mu_y, \sigma)\} = \frac{n + m}{2}\ \log\left[1 + \frac{h(n, m)}{2(n + m)}\,\theta^2\right], \qquad (5)$$

a logarithmic transformation of the standardized distance $\theta = (\mu_x - \mu_y)/\sigma$ between the two means. The intrinsic loss (5) increases linearly with the total sample size $n + m$, and it is essentially quadratic in θ in a neighbourhood of zero, but it becomes concave for $|\,\theta\,| > (k + 1)/\sqrt{k}$, where $k = n/m$ is the ratio of the two sample sizes, an eminently reasonable behaviour which conventional loss functions do not have. For equal sample sizes, $m = n$, this reduces to $n\ \log[1 + \theta^2/4]$ a linear function of the sample size n, which behaves as $\theta^2/4$ in a neighbourhood of the origin, but becomes concave for $|\,\theta\,| > 2$.

Reference analysis. The intrinsic loss (5) is a piecewise invertible function of θ, the standardized difference of the means. Consequently, the required objective prior is the joint reference prior function $\pi_\theta(\mu_x, \mu_y, \sigma)$ when the standardized difference of the means, $\theta = (\mu_x - \mu_y)/\sigma$, is the quantity of interest. This may easily be obtained using the orthogonal parametrization $\{\theta, \omega_1, \omega_2\}$, with $\omega_1 = \sigma\sqrt{2(m+n)^2 + m\,n\,\theta^2}$, and $\omega_2 = \mu_y + n\,\sigma\,\theta/(n+m)$. In the original parametrization the required reference prior is found to be

$$\pi_\theta(\mu_x, \mu_y, \sigma) = \frac{1}{\sigma^2}\left(1 + \frac{h(n,m)}{4(m+n)}\left(\frac{\mu_x - \mu_y}{\sigma}\right)^2\right)^{-1/2}. \tag{6}$$

By Bayes theorem, the posterior is $\pi_\theta(\mu_x, \mu_y, \sigma \,|\, \boldsymbol{z}) \propto p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)\, \pi_\theta(\mu_x, \mu_y, \sigma)$. Changing variables to $\{\theta, \mu_y, \sigma\}$, and integrating out μ_y and σ, produces the (marginal) reference posterior density of the quantity of interest

$$\begin{aligned}\pi(\theta \,|\, \boldsymbol{z}) &= \pi(\theta \,|\, t, m, n)\\ &\propto \left(1 + \frac{h(n,m)}{4(m+n)}\,\theta^2\right)^{-1/2} \mathrm{NcSt}\left(t \,\middle|\, \sqrt{\frac{h(n,m)}{2}}\,\theta,\; n+m-2\right)\end{aligned} \tag{7}$$

where

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{2/h(n,m)}}\,, \quad s^2 = \frac{n\,s_x^2 + m\,s_y^2}{n+m-2}\,,$$

and $\mathrm{NcSt}(\cdot \,|\, \lambda, \nu)$ is the density of a noncentral Student distribution with noncentrality parameter λ and ν degrees of freedom. The reference posterior (7) is proper provided $n \geq 1$, $m \geq 1$, and $n + m \geq 3$. For further details, see Pérez (2005).

The reference posterior (7) has the form $\pi(\theta \,|\, t, n, m) \propto \pi(\theta)\; p(t \,|\, \theta, n, m)$, where $p(t \,|\, \mu_x, \mu_y, \sigma, m, m) = p(t \,|\, \theta, n, m)$ is the sampling distribution of t. Thus, the reference prior is consistent under marginalization (cf. Dawid, Stone and Zidek, 1973).

The intrinsic statistic. The reference posterior for θ may now be used to obtain the required intrinsic test statistic. Indeed, substituting into (5) yields

$$d(H_0 \,|\, \boldsymbol{z}) = d(H_0 \,|\, t, m, n) = \int_0^\infty \frac{n+m}{2}\,\log\left[1 + \frac{h(n,m)}{2(m+n)}\,\theta^2\right] \pi(\theta \,|\, t, m, n)\, d\theta, \tag{8}$$

where $\pi(\theta \,|\, t, m, n)$ is given by (7). This has no simple analytical expression but may easily be obtained by one-dimensional numerical integration.

Example. The derivation of the appropriate reference prior allows us to draw precise conclusions even when data are extremely scarce. As an illustration, consider a (minimal) sample of three observations with $\boldsymbol{x} = \{4, 6\}$ and $\boldsymbol{y} = \{0\}$, so that $n = 2$, $m = 1$, $\bar{x} = 5$, $\bar{y} = 0$, $s = \sqrt{2}$, $h(n,m) = 4/3$ and $t = 5/\sqrt{3}$. If may be verified numerically that the reference posterior probability that $\theta < 0$ is

$$\Pr[\theta < 0 \,|\, t, h, n] = \int_{-\infty}^{0} \pi(\theta \,|\, t, m, n)\, d\theta = 0.0438,$$

directly suggesting some (mild) evidence against $\theta = 0$ and, hence, against $\mu_x = \mu_y$. On the other hand, using the formal procedure described above, the numerical value of intrinsic statistic to test $H_0 \equiv \{\mu_x = \mu_y\}$ is

$$d(H_0 \,|\, t, m, n) = \int_0^\infty \frac{3}{2}\,\log\left[1 + \frac{2}{9}\,\theta^2\right] \pi(\theta \,|\, t, m, n)\, d\theta = 1.193 = \log[6.776].$$

Thus, given the available data, the expected value of the average (under repeated sampling) of the log-likelihood ratio against H_0 is 1.193 (so that likelihood ratios may be expected to be about 6.8 against H_0), which provides a precise measure of the available evidence against the hypothesis $H_0 \equiv \{\mu_x = \mu_y\}$.

This (moderate) evidence against H_0 is *not* captured by the conventional frequentist analysis of this problem. Indeed, since the sampling distribution of t under H_0 is a standard Student distribution with $n+m-2$ degrees of freedom, the p-value which corresponds to the two-sided test for H_0 is $2(1 - T_{m+n-2}\{|t|\})$, where T_ν is the cumulative distribution function of an Student distribution with ν degrees of freedom (see, *e.g.*, DeGroot and Schervish, 2002, Section 8.6). In this case, this produces a p-value of 0.21 which, contrary to the preceding analysis, suggests lack of sufficient evidence in the data against H_0.

Further results. The full version of this paper (Bernardo and Pérez, 2007) contains analytic asymptotic approximations to the intrinsic test statistic (8), analyzes the behaviour of the proposed procedure under repeated sampling (both when H_0 is true and when it is false), and discusses its generalization to the case of possibly different variances.

REFERENCES

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402 and 457–464 (with discussion).

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 61–77 (with discussion).

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**,113–147.

Bernardo, J. M. (2005a). Reference analysis. *Handbook of Statistics* **25**, D. K. Dey and C. R. Rao, (eds.) Amsterdam: Elsevier17–90.

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **14**, 317–384.

Bernardo, J. M. and Pérez, S. (2007). Comparing normal means: New methods for an old problem. *Bayesian Analysis* **2**, 45–48.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley, (2nd edition to appear in 2008).

DeGroot, M. H. and Schervish, M. J. (2002). *Probability and Statistics*, 3rd ed. Reading, MA: Addison-Wesley.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference *J. Roy. Statist. Soc. B* **35**,189–233 (with discussion).

Juárez, M. A. (2005). Normal correlation: An objective Bayesian approach. *Tech. Rep.*, CRiSM 05-15, University of Warwick, UK.

Pérez, S. (2005). *Objective Bayesian Methods for Mean Comparison.* Ph.D. Thesis, Universidad de Valencia, Spain.