Sociedad de Estadística e Investigación Operativa
Test (1998) *Vol.* **7**, *No.* 1, pp. 207–216

Reconciling Bayesian and frequentist evidence in the point null testing problem

Miguel A. Gómez–Villegas and Luis Sanz
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas, Universidad Complutense de Madrid
28040 Madrid, Spain

Abstract

For the point null hypothesis testing problem it is shown that, in some situations, the classical evidence against H_0, expressed in terms of the p–value, is in the range of Bayesian measures of evidence. In these situations, it is therefore possible to reconcile measures of evidence between Bayesian and frequentist approaches. More specifically, for the class of unimodal, symmetric and nonincreasing prior distributions, it is shown that the infimum of the posterior probability of H_0 is numerically equal to the p value. The discrepancy which appears in the literature dedicated to this subject until now, is due to the form of the mixed distribution and not due to its use as a prior.

Key Words: Posterior probability, p–value, point null hypothesis.

AMS subject classification: 62F15; 62A15

1 Introduction

In testing a point null hypothesis Berger and Sellke (1987) and Berger and Delampady (1987) calculated the discrepancy between the classical approach, expressed in terms of the p value, and the Bayesian approach, expressed through the Bayes factor and the posterior probability.

Alternatively, Cassella and Berger (1987) notice that there is no discrepancy in the one-sided testing problem, and they argue that the difference in the point null case is because a mixed type of prior distribution is used. That is to say, the distribution assigns mass π_0 to the point $\theta = \theta_0$, and spreads out the remainder, $1 - \pi_0$, over $\theta \neq \theta_0$ according to a density $\pi(\theta)$.

The inconvenience of using p–values is well known, some important papers about this topic are Lindley (1957), Berger and Delampady (1987) and Hwang et al. (1992).

Received: July 1996; Accepted: January 1997

The different behavior of the p–value in the one–sided and two–sided testing problem is one of the reasons for the present paper. We will prove that the discrepancy observed until now between the classical and the Bayesian approach in the two–sided (point null) problem, can be tempered if one uses a mixed prior distribution with a convenient value of π_0, the mass assigned to the null hypothesis. In this case, the posterior probability of the point null hypothesis and the p value match.

This does not imply that the mixed distribution must be modified in this way, in our opinion, we must resolve the problem with the Bayesian approach. If you want to know what would have happened with the classical approach, you can use the value of π_0 as given in Theorem 1. In either case this paper shows the effect, in the posterior probability of the null hypothesis, of the choice of value π_0.

There is a substantial literature on reconciliation between p–values and posterior probabilities. Some important references are Edwards el al. (1963), Pratt (1965), Dickey and Lientz (1970), Cox and Hinkley (1974), DeGroot (1977), Bernardo (1980), Rubin (1984), Casella and Berger (1987), Ghosh and Mukerjee (1992) and Mukhopadhyay and DasGupta (1997).

In Section 2 we present some preliminaries. Section 3 contains the relationship between classical and Bayesian evidence, and Section 4 contains some conclusions and comments.

2 Preliminaries

We consider the point null testing problem

$$H_0^* : \theta = \theta_0 \quad \text{versus} \quad H_1^* : \theta \neq \theta_0, \tag{2.1}$$

based on observing a random variable, X, with density $f(x|\theta)$ continuous at θ_0. We suppose, moreover, that the probability of $\theta = \theta_0$ is $\pi_0 > 0$, in such a way that the prior information is given by a mixed distribution assigning mass π_0 to the point $\theta = \theta_0$, and spreading the remainder, $1 - \pi_0$, according to a density $\pi(\theta)$ over $\theta \neq \theta_0$.

Following Berger and Sellke (1987) we seek to minimize $\Pr(H_0|x)$ over a reasonable class of prior distributions, defined by

$G_{US} = \{$ All distributions which are unimodal, symmetric about θ_0,

and nonincreasing on $|\theta - \theta_0|$ $\}$.

If one wants to compare the p-value with the posterior probability of the null hypothesis, it seems reasonable to work with a class of priors instead of a single prior distribution, since the p–value is based on the objective frequentist model and doesn't use prior information. Thus an extensive class of prior distributions which represent our prior belief is used, and the posterior distribution within this class is computed.

We consider the infimum of the posterior probability of the null hypothesis, as our Bayesian measure of evidence, where the infimum is taken over G_{US}. We take the infimum rather than the supremum or any other bound, because when the infimum is small the null hypothesis must be rejected according to the interpretation of the p–value. Another reason can be found in Berger and Sellke (1987), Comment 3. This development is similar to that of Casella and Berger (1987) who reconcile Bayesian and frequentist evidence in the one–sided testing problem and, as we have said above, we are interested in clarifying the reason for the discrepancy between both approaches in the point null testing problem.

As in Berger and Delampady (1987), we propose that a precise hypothesis can be represented as

$$H_0 : |\theta - \theta_0| \leq \varepsilon \quad \text{versus} \quad H_1 : |\theta - \theta_0| > \varepsilon, \tag{2.2}$$

where ε is "small", and we replace the point null hypothesis by this interval hypothesis.

An interesting discussion about the difference between (2.1) and (2.2) can be found in Lindley (1988) and the discussion contained therein.

The advantage of replacing (2.1) by (2.2) is twofold:

(i) we do not need a prior distribution of mixed type in (2.2)

(ii) if we take (2.1), given $\pi(\theta)$,then we can fix a value of ε and compute

$$\pi_0 = \int_{|\theta-\theta_0|\leq\varepsilon} \pi(\theta) d\theta. \tag{2.3}$$

The choice of ε is more intuitive than just selecting an arbitrary value for π_0 – in the literature it is usually $\frac{1}{2}$.

One can argue that a hypothesis with prior probability as in (2.3) is not believable, particularly when ε is small, but we are talking about a scenario where the protagonist is $\pi(\theta)$; and we can choose ε to make π_0

equal to 1/2, as is usually done in the literature. If π_0 is small, this is the prior distribution which, in some situations, produces the same effect as the p–value, as we will show.

We illustrate a real world statistical motivation for this approximation with two examples.

The first one can be seen in Kass and Raftery (1995). They suppose that progress within the educational system is determined largely by educationally relevant attributes and not by other educationally irrelevant attributes, such as social class. If we reduce this problem to test if social class is relevant to educational attainment, we should use a prior distribution with π_0 (and ε) rather small.

The second example is contained in Casella and Berger (1987). In a regression problem we may be interested in testing $H_0^* : \beta = 0$ where β is a regression coefficient. It would indicate that the independent variable has no effect on the response variable and, credibly, the researcher would not place a high prior probability on H_0^* since, in this case, the independent variable would not be included in the experiment.

3 Main results

In this section we observe that if we consider the class of prior distributions G_{US}, the value of ε in (2.2) can be chosen such that $\underline{Pr}(H_0^*|x)$, the infimum of the posterior probability of the null hypothesis and the p value match.

Theorem 3.1. *For the hypothesis in (2.1), if we define π_0 as in (2.3) and*

$$\int_{\mathcal{R}} f(x|\theta)\, d\theta < \infty, \tag{3.1}$$

then

$$\inf_{\pi \in G_{US}} Pr(H_0^*|x) = \left(1 + \frac{1}{2\varepsilon} \int_{-\infty}^{+\infty} \frac{f(x|\theta)}{f(x|\theta_0)}\, d\theta \right)^{-1}. \tag{3.2}$$

Proof. Computing the infimum of $\Pr(H_0^*|x)$ over the class G_{US} is the same as computing it over the class, G_U, of uniform distributions $U(\theta_0 - k, \theta_0 + k)$, k varying in $\Re$, see Casella and Berger (1987), Lemma 3.1, so

$$\inf_{\pi \in G_{US}} Pr(H_0^*|x) = \inf_{\pi \in G_U} \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + (1 - \pi_0) \int_{\theta \neq \theta_0} \pi(\theta) f(x|\theta)\, d\theta},$$

it should be noted that π_0, according to (2.3), depends on k. By replacing π_0 by (2.3) for G_U, this expression becomes

$$\inf_k \frac{\frac{\varepsilon}{k} f(x|\theta_0)}{\frac{\varepsilon}{k} f(x|\theta_0) + \left(1 - \frac{\varepsilon}{k}\right) \int_{|\theta-\theta_0|\leq k} \frac{1}{2k} f(x|\theta)\, d\theta},$$

then

$$Pr(H_0^*|x) = \frac{2f(x|\theta_0)}{2f(x|\theta_0) + \left(\frac{1}{\varepsilon} - \frac{1}{k}\right) \int_{|\theta-\theta_0|\leq k} f(x|\theta)\, d\theta},$$

and the variation of $Pr(H_0^*|x)$ in relation to k is given by

$$\frac{\partial}{\partial k} Pr(H_0^*|x) =$$

$$\frac{-\,2f(x|\theta_0)\left[\frac{1}{k^2} \int_{|\theta-\theta_0|\leq k} f(x|\theta)\, d\theta \;+\; \left(\frac{1}{\varepsilon} - \frac{1}{k}\right) [f(x|k) + f(x|-k)]\right]}{\left(2f(x|\theta_0) + \left(\frac{1}{\varepsilon} - \frac{1}{k}\right) \int_{|\theta-\theta_0|\leq k} f(x|\theta)\, d\theta\right)^2}.$$

We observe that

$$\frac{\partial}{\partial k} Pr(H_0^*|x) < 0,$$

therefore, $\Pr(H_0^*|x)$ is decreasing in k and the minimum is attained on the boundary, i.e k goes to infinity. Then

$$\inf_{\pi \in G_{US}} Pr(H_0^*|x) = \lim_{k\to\infty} \frac{2f(x|\theta_0)}{2f(x|\theta_0) + \left(\frac{1}{\varepsilon} - \frac{1}{k}\right) \int_{|\theta-\theta_0|\leq k} f(x|\theta)\, d\theta},$$

and hence we obtain (3.2), because

$$\frac{1}{k} \int_{|\theta-\theta_0|\leq k} f(x|\theta)\, d\theta = \frac{1}{k} \int_{\Re} I_{(\theta_0-k,\theta_0+k)}(\theta) f(x|\theta)\, d\theta \leq \frac{1}{k} \int_{\Re} f(x|\theta)\, d\theta.$$

which proves the theorem. □

For a fixed ε, expression (3.2) gives us the infimum of the posterior probability of the point null hypothesis.

If an appropriate statistic, $T(X)$, exists, the p value – or observed significance level – of the observed data, x, for the point null testing problem is

$$p(x) = Pr_{\theta=\theta_0}(|T(X)| \geq |T(x)|). \tag{3.3}$$

A more general definition can be seen in Lehmann (1986), pg 170. For the interval hypothesis, the p–value is

$$p_\varepsilon(x) = \sup_{\theta:\, |\theta-\theta_0|\le\varepsilon} Pr_\theta(|T(X)| \ge |T(x)|).$$

Berger and Delampady (1987) seek conditions under which $p_\varepsilon(x) \approx p(x)$.

Moreover, a value of ε exists, say ε^*, so that the p–value and the infimum of the posterior probability (3.2) are equal. In order to compute ε^*, it is sufficient to equal (3.2) and (3.3), and then

$$\varepsilon^* = \frac{1}{2}\frac{p(x)}{1-p(x)}\int_{-\infty}^{+\infty}\frac{f(x|\theta)}{f(x|\theta_0)}\,d\theta. \tag{3.4}$$

We observe that we cannot choose ε^* directly, since (3.4) depends on x. We choose a value of ε near to ε^*, since $\underline{Pr}(H_0^*|x)$ is a continuous function of ε, we will obtain that $\underline{Pr}(H_0^*|x,\varepsilon)$ is approximately equal to $\underline{Pr}(H_0^*|x,\varepsilon^*)$ and then (3.2) is equal to the p–value.

Example 3.1. Let us suppose $X = (X_1, X_2, \ldots, X_n)$, where X_i are i.i.d. $N(\theta,\sigma^2)$ random variables, with σ^2 known. Then

$$\underline{Pr}(H_0^*|x) = \left[1 + \frac{1}{2\varepsilon}\frac{\sqrt{2\pi}\sigma}{\sqrt{n}}e^{\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2}\right]^{-1}.$$

Table 1 shows, for $\sigma^2 = 1$ and $n = 10$, the infimum of the posterior probability of the null hypothesis, for some specific, important values of t, and some values of ε.

In Table 1 it can be observed that $\underline{Pr}(H_0^*|x,\varepsilon)$ is relatively close to the p–value. This is also illustrated in Figure 1, where $\underline{Pr}(H_0^*|x,\varepsilon)$ is represented for $\varepsilon = 0.1$, 0.2, 0.3, and 0.4, together with the p–value and $\underline{Pr}(H_0^*|x)$ when $\pi_0 = \frac{1}{2}$. We can see that this last curve is far from the p–value and $\underline{Pr}(H_0^*|x,\varepsilon)$.

Whereas the value of ε^* given by (3.4) is

$$\varepsilon^* = \frac{1}{2}\frac{p(x)}{1-p(x)}\frac{\sqrt{2\pi}\sigma}{\sqrt{n}}e^{\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2},$$

	t			
ε	1.645	1.960	2.576	3.291
0.1	0.0612	0.0356	0.0091	0.0011
0.3	0.1636	0.0998	0.0267	0.0034
0.5	0.2449	0.1560	0.0437	0.0056

Table 1: Infimum of the posterior probabilities of the null hypothesis over G_{US}, for X ∼Normal

Figure 1: P–value, $\underline{Pr}(H_0^*|x,\varepsilon)$ and $\underline{Pr}(H_0^*|x,\pi_0=\frac{1}{2})$

if we let $t = \dfrac{|\overline{x}-\theta_0|}{\sigma}\sqrt{n}$, then

$$\varepsilon^* = \frac{1}{2}\frac{p(t)}{1-p(t)}\frac{\sqrt{2\pi}\sigma}{\sqrt{n}}e^{\frac{t^2}{2}}.$$

Furthermore, in order to show the equivalence between (2.1) and (2.2) we can compute

$$\underline{Pr}(H_0|x) = \inf_{\pi\in G_{US}} Pr(H_0|x)$$

which is given by

$$\begin{aligned}\underline{Pr}(H_0)|x) &= \lim_{k\to\infty} Pr(H_0|x)\\ &= \lim_{k\to\infty}\frac{\Phi(t+\varepsilon\frac{\sqrt{n}}{\sigma})-\Phi(t-\varepsilon\frac{\sqrt{n}}{\sigma})}{\Phi(t+k\frac{\sqrt{n}}{\sigma})-\Phi(t-k\frac{\sqrt{n}}{\sigma})}\\ &= \Phi\left(t+\varepsilon\frac{\sqrt{n}}{\sigma}\right)-\Phi\left(t-\varepsilon\frac{\sqrt{n}}{\sigma}\right).\end{aligned}$$

Table 2 shows, for the same setup as Table 1, for some specific values of t, the values of ε^* such that $\underline{Pr}(H_0^*|x)$ equals the p value, it also shows $\underline{Pr}(H_0|x)$ and the infimum of the posterior probability of H_0^* when $\pi_0 = \frac{1}{2}$. We observe that the second and fourth columns are close if we choose ε properly. On the other hand, the discrepancy between the second and fifth columns is bigger, i.e. when we set $\pi_0 = \frac{1}{2}$ in the mixed distribution.

t	p value = $\underline{Pr}(H_0^*\|x)$	ε^*	$\underline{Pr}(H_0\|x)$	$\underline{Pr}(H_0^*\|x, \pi_0 = \frac{1}{2})$
1.645	0.1	0.170	0.1198	0.390
1.960	0.05	0.142	0.0575	0.290
2.576	0.01	0.111	0.0121	0.109
3.291	0.001	0.089	0.0011	0.018

Table 2: Values of ε^* such that the p value equals $\underline{Pr}(H_0^*|x)$, $\underline{Pr}(H_0|x)$, for an appropriate ε and $\underline{Pr}(H_0^*|x, \pi_0 = \frac{1}{2})$. All infimums are taken over G_{US}

4 Discussions and conclusions

For the problem of testing a point null hypothesis we believe that taking a density, $\pi(\theta)$, and fixing a suitable value of ε in accordance with the intuition of the decision maker, is an appropriate method. If the decision is to choose a value of ε close to (3.4), the infimum of $\Pr(H_0|x)$ is approximately equal to the infimum of $\Pr(H_0^*|x)$. Thus, from the Bayesian point of view, we take the same decision in both problems (2.1) and (2.2), whenever the prior distribution used is near to the distribution in which the infimum is reached. Taking π_0 in the mixed distribution to be equal to the probability of the interval, as in (2.3), has its advantages. First it allows us to fix the value of the prior probability of the null hypothesis, π_0, for those cases where we do not know how to do it. Secondly, a mixed prior distribution is not used for testing a precise hypothesis like (2.2), so the discrepancy observed by Casella and Berger (1987) is not due to using a mixed prior distribution as they asserted. Thirdly, by Theorem (3.1), the infimum is achieved when the prior is the improper prior distribution which seems a natural form of impartiality. In this situation we have shown that the p–value, the posterior probability of the point null hypothesis and the posterior probability of the interval are close.

Finally, taking the infimum over the class G_{US}, we proved that this was close to the p–value, and it is was shown that the p–value is in the range of the Bayesian measure of evidence, whenever the infimum has been reached. On the other hand, it is clear that if we perform a subjective Bayesian analysis, the posterior probability of the null is larger than the p–value.

Further research must be carried out with our construction in the point null testing problem using other classes of prior distributions and other classical measures.

Acknowledgements

The authors wish to thank the associate editor and two unknown referees for many helpful comments and valuable suggestions.

References

Berger, J.O. and Delampady, M. (1987). Testing precise hyphoteses. *Statistical Science*, **2**, 317–352.

Berger, J.O. and Sellke, T. (1987). Testing a point null hyphoteses: The irreconciliability of p–values and evidence (with discussion). *Journal of the American Statistical Association*, **82**, 112–122.

Bernardo, J.M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds) Valencia: University Press, 605–647 (with discussion).

Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one–sided testing problem (with discussion). *Journal of the American Statistical Association*, **82**, 106-111.

Casella,G. and Berger, R.L. (1987). Comment, in testing precise hyphoteses (Berger, J.O. and Delampady, M. (1987)). *Statistical Science*, **2**, 317–352.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Dickey, J.M. (1977). Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, **72**, 138–142.

Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242. Reprinted in *Robustness of Bayesian Analysis* (J.B.Kadane, ed.). Amsterdam: North–Holland, 1984, 1–62.

Ghosh,J.K. and Mukerjee,R. (1992). Non–informative priors. *Bayesian Statistics 4* (J.M.Bernardo, J.O.Berger, A.P.Dawid and A.F.M.Smith, eds.). Oxford: University Press, 195–210 (with discussion).

Gómez–Villegas, M.A. and Gómez Sánchez–Manzano, E. (1992). Bayes factor in testing precise hyphoteses. *Communications in Statistics. Theory and Methods*, **21**, 1707-1715.

Hwang, J.T., Casella, G., Robert, Ch., Wells, M.T. and Farrell, R.H. (1992). Estimation of accuracy in testing. *The Annals of Statistics*, **20**, 1, 409–50.

Kass, R.E. and Raftery, A.E. (1995). Review. Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* New York: Wiley.

Lindley, D.V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

Lindley, D.V. (1988). Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Statistcs 3* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.). Oxford: University Press, 307-326.

Mukhopadhyay, S. and DasGupta, A. (1997). Uniform approximation of Bayes solutions and posteriors: frequentistly valid Bayes inference. *Statistics and Decisions*, **15**, 51–7.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151–1172.