PMLR: Proceedings of Machine Learning Research, vol. 62, 97-108, 2017 ISIPTA '17

Reconciling Bayesian and Frequentist Tests: the Imprecise Counterpart

Inés Couso COUSO@UNIOVI.ES
Statistics and O.R.
Universidad de Oviedo
Gijón (Spain)

Antonio Álvarez-Caballero ANALCA3@GMAIL.COM
Metrology and Models
Universidad de Oviedo
Gijón (Spain)

Luciano Sánchez LUCIANO@UNIOVI.ES
Computer Sciences and A.I.
Universidad de Oviedo
Gijón (Spain)

Abstract

Imprecise Dirichlet Process-based tests (IDP-tests, for short) have been recently introduced in the literature. They overcome the problem of deciding how to select a single prior in Bayesian hypothesis testing, in the absence of prior information. They make use of a "near-ignorance" model, that behaves a priori as a vacuous model for some basic inferences, but it provides non-vacuous posterior inferences. We perform empirical studies regarding the behavior of IDP-tests for the particular case of Wilcoxon rank sum test. We show that the upper and lower posterior probabilities can be expressed as tail probabilities based on the value of the U statistic. We construct an imprecise frequentist-based test that reproduces the same decision rule as the the IDP test. It considers a neighbourhood around the U-statistic value. If all the values in the neighbourhood belong to the rejection zone (resp. to the acceptance region), the null hypothesis is rejected (resp. accepted). Otherwise, the judgement is suspended. This construction puts a step forward in the reconciliation between frequentist and Bayesian hypothesis testing.

Keywords: Wilcoxon rank sum test; imprecise tests; one-sided test; frequentist test; Bayesian test; IDP test; interval p-values.

1. Introduction

The problem of reconciling Bayesian and frequentist techniques has been extensively treated in the literature and seems to be still open. In the frequentist setting, the level of significance of the outcome against the null hypothesis is determined in terms of the p-value. Notwithstanding the "probability that the null hypothesis is true" has no meaning in this framework, but it has been argued that some practitioners attach such a meaning to the p-value (see Casella and Berger (1987) for further discussion). Alternatively, under the Bayesian approach, evidence takes the form of the posterior probability about the null hypothesis, based on the combination of prior evidence and the evidence provided by the dataset. The relation between the p-value and the posterior probability of the null hypothesis has been examined by different authors (see Berger and Selke (1987); Casella and Berger (1987); DeGroot (1973); Pratt (1965); Shafer (1982); Jeffreys (1939) among many oth-

ers). For two-sided tests, it has been noticed by several of them that the p-value tends to be smaller than the posterior probability of the null hypothesis (see Berger and Selke (1987); Lindley (1957)) for some collections of priors, while for the one-sided testing problem situations can be found where they are approximately equal (see Pratt (1965); Casella and Berger (1987)). In particular, Casella and Berger (1987) prove that for some classes of reasonable and impartial priors, and under some additional requirements about the distribution of X, the p-value coincides with the infimum of the posterior probability of the null hypothesis. With respect to the large discrepancies between the infimum for the posterior probability and the p-value in the two-sided problem observed in Berger and Selke (1987), Casella and Berger (1987) question the impartiality of the priors considered by the authors. The problem of selecting an appropriate prior (specially in those cases where no initial information is available) has been a subject of study of many authors. One solution to this problem has been proposed initially by Ferguson (1973) and afterwards by Rubin (1981) under the name of Bayesian Bootstrap (BB). Notwithstanding, the BB model cannot be regarded as non-informative, since it assigns zero probability to any set that does not include the observations. In order to overcome this issue, Benavoli et al. (2015) introduced a new kind of test, by means of replacing a single prior by a collection of priors based on the imprecise Dirichtlet process (IDP). The combination of this near-to-ignorance prior information with our evidence obtained from the sample leads to a pair of dual upper and lower posterior probabilities. The IDP-based test has the advantage of not deciding when this decision is somehow prior-dependent. In other words, when the action that minimizes the risk (expected loss) is not the same for all the prior probabilities, the IDP suspends its judgment. The authors have exemplified their proposal with an IDP-based version of the well known Wilcoxon rank sum test, also called the Mann-Withney-Wilcoxon test, or simply, the MWW test (Mann and Whitney (1947); M.P. Fay (2010)).

Consider two random variables X and Y whose cdf's satisfy $F_X(x) = F_Y(x + \Delta)$, $\forall\, x \in \mathbb{R}$. The null hypothesis of the traditional MWW test is that $P(X \leq Y) \leq 0.5$ against the alternative hypothesis $P(X \leq Y) > 0.5$. When the distribution of $X - Y$ is continuous, we can interpret a significant Mann-Whitney-Wilcoxon test as showing that the median of the difference is negative (Couso et al. (2015)). The IDP-based procedure will assign a pair of upper and lower probabilities to the null hypothesis, $\overline{P}(H_0|(\vec{x}, \vec{y}))$ and $\underline{P}(H_0|(\vec{x}, \vec{y}))$, that encompass the collection of posterior probabilities associated to the selected collection of priors. The authors propose the following decision rule, for some threshold $\gamma \in (0, 1)$:

- If both the upper and the lower posterior probabilities are on one side of the threshold γ, we will either reject (left side) or accept (right side) the null hypothesis.
- Alternatively, if they satisfy the inequalities $\underline{P}(H_0|(\vec{x}, \vec{y})) < \gamma \leq \overline{P}(H_0|(\vec{x}, \vec{y}))$, then we are in an indeterminate decision, i.e, we suspend our judgement.

After presenting their new proposal, the authors have performed some empirical comparisons with respect to the Bayesian bootstrap-based test as well as with the traditional frequentist MWW test, under different conditions for the shift parameter Δ. They suggest that when the IDP based test is indeterminate, both the frequentist and the Boostrap Bayesian test behave as "random guessers". What they check in fact is that, for some values of Δ, the proportion of rejections under those situations is nearly 50%, which coincides with the proportion of rejections of a randomized test derived from the IDP test (the one called the 50/50 test by the authors) that returns the same response as the IDP test when it is determinate, and a random answer otherwise.

Our paper deepens the study about the relations between this new imprecise test and its precedents. We will first empirically check that the p-value of the traditional MWW test coincides with the posterior probability of the null hypothesis for the BB test. Afterwards, we will show that there is a one-to-one correspondence between the upper (resp. the lower) posterior probability of the IDP test and the p-value of the MWW test. Thus the outcome of the latter is univocally determined by the upper (equivalently, by the lower) posterior probability of the former. In fact, upper and lower posterior probabilities derived from the IDP-based approach can be calculated in terms of the cdf of $U + \epsilon$ and $U - \epsilon$, for some $\epsilon > 0$, where U represents the MWW statistic. On the basis of this relation, we construct an imprecise frequentist-based test whose performance mimics the one of the IDP-based test. These findings help us to better understand the behaviour of the new IDP-based test, and put a step forward in the reconciliation between the frequentist and Bayesian approaches in this imprecise setting. In particular, this kind of imprecisiation over the set of priors seems to produce similar effects on the decision mechanism as an imprecisiation of data around the observations.

2. Preliminaries

The Mann-Withney U test (also called Wilcoxon rank sum or Mann-Whitney-Wilcoxon test) is used to check whether or not it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second one, assuming that both selections are independent from each other.

Consider two independent samples containing n_1 and n_2 elements respectively from each population. The U statistic is calculated as the sum of the ranks of the elements contained in the first sample, with the minimum value $n_1(n_1+1)/2$ subtracted. In other words, it counts the number of items (x_i, y_j) such that x_i is less than or equal to y_j,

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{[X_i,\infty)}(Y_j).$$

Under the assumption $P(X \leq Y) = 0.5$, the expectation and the variance of U are respectively:

$$\mu_0 = \frac{n_1\, n_2}{2} \qquad\qquad \sigma_0^2 = \frac{n_1\, n_2(n_1+n_2)}{12}.$$

Let us consider the one-sided test of $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$, where $\theta = P(X \leq Y)$. The rejection region of the Mann-Whitney U test of size α is defined in terms of U as follows:

$$R_\alpha = \left\{ (\vec{x}, \vec{y}) \,:\, \frac{U(\vec{x}, \vec{y}) - \mu_0}{\sigma_0} > z_\alpha \right\},$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the quantile $1-\alpha$ of the distribution $N(0,1)$. Alternatively, it can be defined as:

$$R_\alpha = \{ (\vec{x}, \vec{y}) \,:\, p(\vec{x}, \vec{y}) < \alpha \},$$

where

$$p(\vec{x}, \vec{y}) = 1 - \Phi\left(\frac{U(\vec{x}, \vec{y}) - \mu_0}{\sigma_0} \right)$$

denotes the p-value of the sample, i.e.,

$$p(\vec{x}, \vec{y}) = \inf\{\alpha \in (0,1) \,:\, (\vec{x}, \vec{y}) \in R_\alpha\}.$$

Under the Bayesian approach, the problem of hypothesis testing is seen as a decision problem where the possible actions are $a = 0$ (accept H_0) and $a = 1$ (reject H_0). We start from a prior distribution over the set parametric space $\Theta = [0, 1]$, determined by a density function $\pi : \Theta \to \mathbb{R}^+$. A loss function $\ell : \Theta \times \{0, 1\} \to \mathbb{R}$ links the action to the unknown value of the parameter $\theta = P(X \leq Y)$: when the true state of nature is $\theta \in \Theta$ and we take the action $a \in \{0, 1\}$ we incur in a loss $\ell(\theta, a)$ determined as follows:

	$a = 0$	$a = 1$
$\theta \leq 0.5$	0	K_0
$\theta \geq 0.5$	K_1	0

The decision rule d that minimizes the posterior expected loss is the one defined as follows:

- $d(\vec{x}, \vec{y}) = 1$ if $P(H_0|(\vec{x}, \vec{y})) < \frac{K_0}{K_0+K_1}$
- $d(\vec{x}, \vec{y}) = 0$ otherwise,

where $P(H_0|(\vec{x}, \vec{y}))$ denotes the posterior probability of the null hypothesis calculated as follows:

$$P(H_0|(\vec{x}, \vec{y})) = \int_{-\infty}^{0.5} L(\vec{x}, \vec{y}; \theta)\pi(\theta)\, d\theta,$$

and $L(\vec{x}, \vec{y}; \theta)$ represents the likelihood function.

The Dirichlet process was proposed by Ferguson (1973) as a second-order probability (in our context, a probability on the space of joint probability distributions for (X, Y)). Since every joint distribution determines a specific value for $\theta = P(X \leq Y)$, a Dirichlet process determines a (prior) probability distribution over the parametric space, Θ. But how do we choose this prior in case of lack of information? Rubin (1981) addressed this problem by means of selecting the so-called Bayesian bootstrap. It is the Bayesian analogue to the Efron's bootstrap Efron (1979). Instead of simulating the sampling distribution of a statistic estimating a parameter, it simulates the posterior distribution of the parameter. This choice nevertheless seems controversial (see Rubin (1981) and Benavoli et al. (2015) for detailed discussions), since it cannot be seen as a representation of a lack of knowledge. In fact, the Bayesian bootstrap assigns probability one to the collection of observations (see Rubin (1981)). To overcome this issue, Benavoli et al. (2015) proposed to use the imprecise Dirichlet process (IDP). It is considered as a prior near-ignorance model. In fact, it corresponds to a set of priors that generates vacuous prior probabilities and therefore, leading to an infimum and a supremum for the (prior) expectations of $\theta = P(X \leq Y)$ respectively equal to 0 and 1. This collection of priors leads to a collection of posterior probabilities for H_0 and H_1, given the dataset, whose bounds we will respectively denoted by $\underline{P}(H_0|(\vec{x}, \vec{y}))$ and $\overline{P}(H_0|(\vec{x}, \vec{y}))$. To perform the hypothesis test $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$, they compare each of these bounds with $\gamma = \frac{K_0}{K_0+K_1}$ and consider the following decision rule:

- $d_I(\vec{x}, \vec{y}) = 1$ if $\overline{P}(H_0 : (\vec{x}, \vec{y})) < \frac{K_0}{K_0+K_1}$
- $d_I(\vec{x}, \vec{y}) = 0$ if $\underline{P}(H_0 : (\vec{x}, \vec{y})) > \frac{K_0}{K_0+K_1}$
- $d_I(\vec{x}, \vec{y}) =?$ otherwise,

where "0", "1" and "?" respectively denote "accept H_0", "reject H_0" and "no decision'.

3. Relations between p-value, posterior probability and upper and lower posterior probabilities

3.1 Formal relations between p-value and Bayesian posterior probability

As mentioned in the Introduction, different authors have studied the relations between the frequentist p-value and the Bayesian posterior probability of the null hypothesis. Casella and Berger (1987) studied this relation for one-sided tests under some additional conditions: In particular, when the underlying distribution is assumed to be symmetric and it satisfies the property of monotone likelihood ratio (MLR), then the p-value coincides with the infimum of the set of posterior probabilities for H_0, for several reasonable collections of priors.

We can prove an additional result relating the p-value and the posterior probability of the null hypothesis derived from any prior. It requires a MLR condition but it does not require any symmetry about the underlying distribution.

Definition 1 *The set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T if we can represent the likelihood ratio as*

$$\frac{L(\vec{x},\vec{y};\theta_1)}{L(\vec{x},\vec{y};\theta_2)} = g_{\theta_1,\theta_2}(T(\vec{x},\vec{y})),$$

where g_{θ_1,θ_2} is strictly increasing for every pair $\theta_1 > \theta_2$.

Now we will prove that, when the family of distributions satisfies the MLR property with respect to some statistic T, the posterior probability associated to a one-sided-test is increasing wrt T:

Lemma 2 *Let us suppose that the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T and let us consider the test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then the posterior probability $P(H_0|\vec{x})$ can be expressed as an increasing function of $T(\vec{x})$, i.e.:*

$$T(\vec{x}) < T(\vec{x}') \Rightarrow P(H_0|\vec{x}) < P(H_0|\vec{x}').$$

The following result is well known in the literature:

Theorem 3 *Assume the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T. Let us consider the one-sided test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then the test $\delta : \mathbb{R}^n \to \{0,1\}$ defined as follows:*

$$\delta(\vec{x}) = \begin{cases} 0 & \text{if } T(\vec{x}) \leq c \\ 1 & \text{if } T(\vec{x}) > c \end{cases}$$

is a uniformly most powerful (UMP) test (among all the tests of size $\alpha_c = P_{\theta_0}(T > c)$).

We deduce the following result:

Theorem 4 *Let us suppose that the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T and that the cdf of T is strictly increasing for some θ_0. Let us consider the test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Let us consider the family of UMP tests associated to the rejection regions $\{R_\alpha : \alpha \in (0,1)\}$, each of them defined as $R_\alpha = \{\vec{x} :$*

$T(\vec{x}) > c_\alpha\}$, *with* $P_{\theta_0}(T > c_\alpha) = \alpha$. *Let us consider the p-value associated to this family of tests as follows:*

$$p(\vec{x}) = \inf\{\alpha : \vec{x} \in R_\alpha\} = P_{\theta_0}(T > T(\vec{x})),\ \forall \vec{x}. \tag{1}$$

Let us consider an arbitrary prior over Θ. *Then there exists a (one-to-one) strictly increasing function* $h : [0,1] \to [0,1]$ *linking the posterior probability of* H_0 *and the p-value as* $P(H_0|x) = g(p(\vec{x}))$, $\forall \vec{x}$, *and therefore*

$$p(\vec{x}) < p(\vec{x}') \Leftrightarrow P(H_0|\vec{x}) < P(H_0|\vec{x}').$$

As a consequence we can state the following corollary:

Corollary 5 *Let us suppose that the set of distributions* $\{P_\theta : \theta \in \Theta\}$ *has a monotone likelihood ratio (MLR) with respect to a statistic* T *and that the cdf of* T *is strictly increasing for some* θ_0. *Let us consider the test* $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$. *Let us consider the family of UMP tests associated to the rejection regions* $\{R_\alpha : \alpha \in (0,1)\}$, *each of them defined as* $R_\alpha = \{\vec{x} : T(\vec{x}) > c_\alpha\}$, *with* $P_{\theta_0}(T > c_\alpha) = \alpha$. *Let us consider an arbitrary prior over* Θ, *an arbitrary pair of loss values* K_0 *and* K_1, *and the Bayesian test associated to it. Then there exists a UMP frequentist test that coincides with it, the size of it being an increasing function of* $\gamma = \frac{K_0}{K_0+K_1}$.

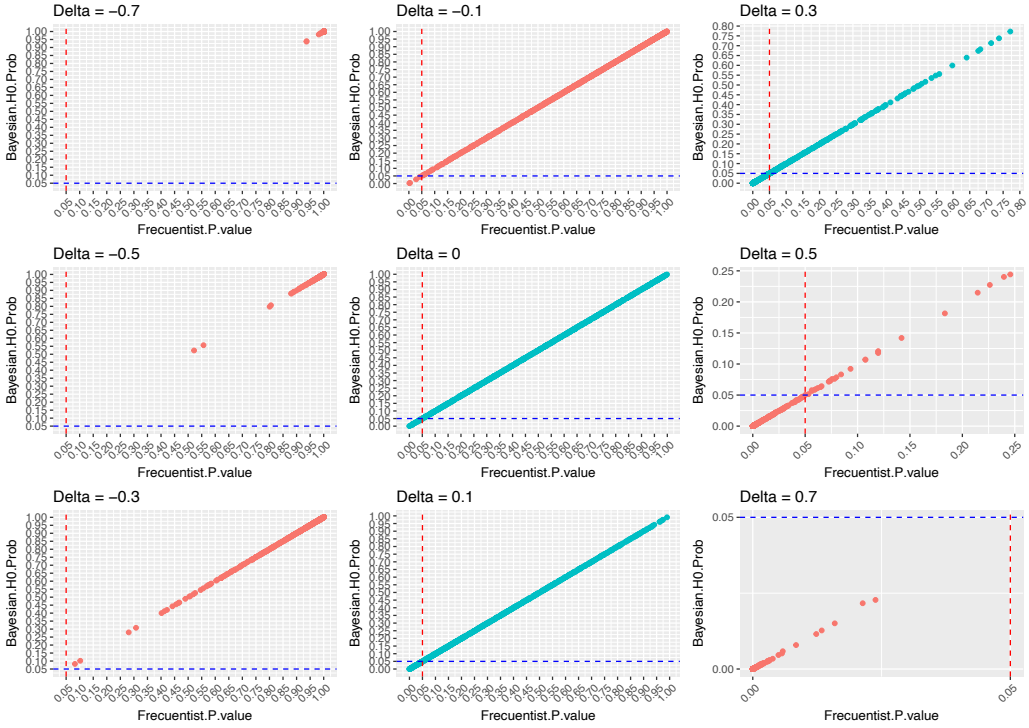
According to the above result, under the condition of MLR, and regardless the prior distribution we select, there exists a one-to-one correspondence between γ and α. This is to say, if we set an arbitrary prior, there exists a bijection $h : [0,1] \to [0,1]$ such that the Bayesian test associated to $\gamma = \frac{K_0}{K_0+K_1}$ coincides with the UMP test of size $\alpha = h(\gamma)$. The next section deals with the particular case of the MWW and its variations considered in Benavoli et al. (2015). In that particular case, this one-to-one correspondence is the identity, i.e., the p-value coincides with the posterior probability of the null hypothesis. Furthermore, we empirically show that the upper and lower posteriors can be also calculated as strictly increasing functions of the p-value.

3.2 Relations between the p-value and the pair of upper and lower posterior probabilities: an empirical study

Benavoli et al. (2015) have developed an empirical study in order to compare their IDP-based test with the MWW frequentist test and the DP-based test obtained as the prior strength goes to zero (called the Bayesian Bootstrap Dirichlet Process test -the BB-DP test, for short-). They have considered a Monte Carlo experiment in which n_1, n_2 observations from X, and Y respectively are generated, where $X \equiv N(0,1)$ and $Y \equiv N(\Delta,1)$, and Δ ranges from -1.5 to 1.5. For each value of Δ, they have performed 20000 Monte Carlo runs. They first compare the performance of the IDP test and the BB-DP test. They consider three different options for the loss quotient $\gamma = 1$, $\gamma = 0.1$ and $\gamma = 0.05$. They conclude that, in all those cases where the first of them is determinate, both of them return the same answer, the difference between them focussing only on those samples for which the first one is indeterminate. In a second round of experiments, they compare the IDP test with the frequentist MWW test. They select the significance level $\alpha = 0.05$ in order to construct the frequentist test, and $\gamma = 0.05$ in order to define the IDP test. Again, the frequentist test returns the same answer when the IDP test is determinate. They also compute the proportion of rejections of the MWW among those samples for which the IDP test is indeterminate. They observe that such

a proportion increases with respect to Δ. As an example, for $n_1 = n_2 = 10$ and $\Delta = 0.9$, the IDP is indeterminate in 30% of the runs, and the MWW test rejects the null hypothesis 50% of them. As it returns the same proportion of rejections as a 50/50 randomized test derived from the IDP test and they conclude that the MWW test "guesses at random" 30% of the times. Let us nevertheless notice that the MWW does not return a random answer from a given sample.

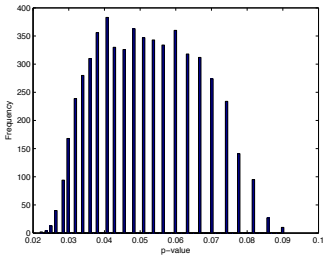
In this section, we deepen this study, with the aim of providing further insight about the behavior of the three tests (BB-DP, IDP and MWW) in practice. On one side, the p-value of the (frequentist) MWW test coincides with the posterior probability of the null hypothesis for the BB-DP, as we empirically show:



The posterior probability of the BB test depends on a bootstrap-based computation, and therefore small differences between the values of the posterior probability may occur if we launch the algorithm repeated times for the same sample $(\vec{x}, \vec{y})$, the average of those posterior probabilities being the p-value. Consequently, the MWW test of size $\alpha = 0.05$ coincides with the BB-DP test for $\gamma = 0.05$.

Let us now examine the relation between the frequentist test and the IDP test. Since the p-value of the MWW coincides with the posterior probability of the BB test, we know that it is always bounded by the upper and lower posterior probabilities associated to the IDP test, $\overline{P}(H_0|(\vec{x}, \vec{y}))$ and $\underline{P}(H_0|(\vec{x}, \vec{y}))$. Therefore, we can write $\overline{P}(H_0|(\vec{x}, \vec{y})) = p(\vec{x}, \vec{y}) + \delta(\vec{x}, \vec{y})$ and $\underline{P}(H_0|(\vec{x}, \vec{y})) = p(\vec{x}, \vec{y}) - \delta'(\vec{x}, \vec{y})$, with $\delta(\vec{x}, \vec{y}) > 0$ and $\delta'(\vec{x}, \vec{y}) > 0$ for every pair of samples $(\vec{x}, \vec{y})$.

Let us now recall an empirical result from Benavoli et al. (2015) about the distribution of the p-values over the collection of samples for which the IDP is indeterminate, i.e., those pairs of samples $(\vec{x}, \vec{y})$ satisfying the inequalities $\underline{P}(H_0|(\vec{x}, \vec{y})) < \gamma < \overline{P}(H_0|(\vec{x}, \vec{y}))$. The figure illustrates the distribution of the p-values for $\Delta = 0.5$ and $n_1 = n_2 = 20$ and $\gamma = 0.05$:



According to the above notation, these are the samples satisfying the following inequalities:

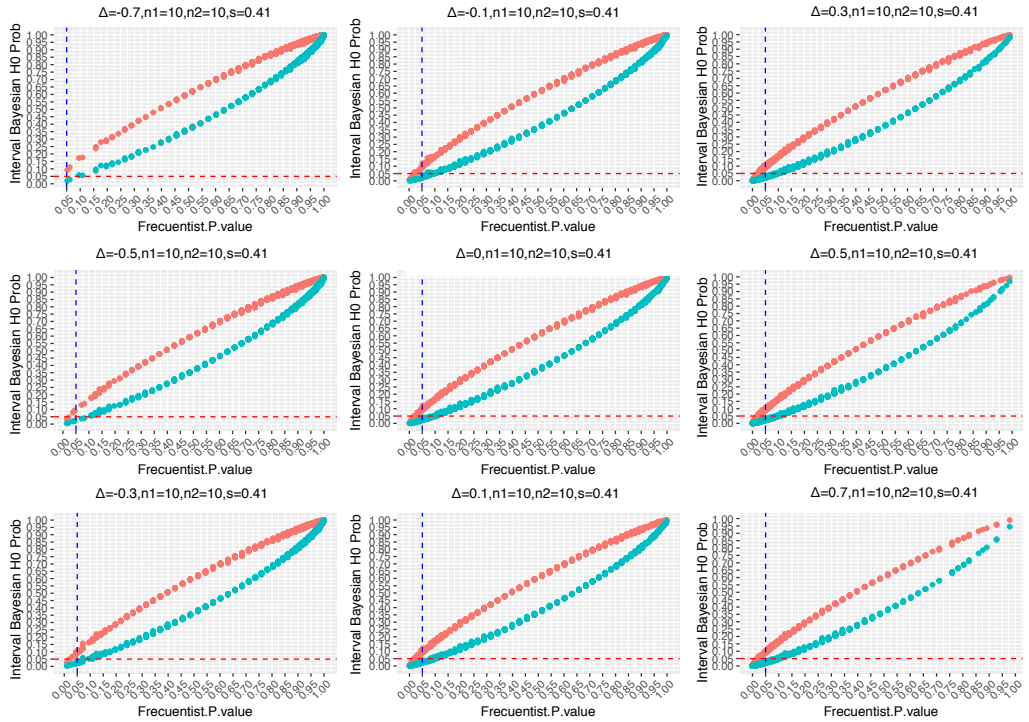
$$p(\vec{x},\vec{y}) - \delta'(\vec{x},\vec{y}) < 0.05 < p(\vec{x},\vec{y}) + \delta(\vec{x},\vec{y})$$

or, equivalently

$$0.05 - \delta(\vec{x},\vec{y}) < p(\vec{x},\vec{y}) < 0.05 + \delta'(\vec{x},\vec{y}).$$

According to the above graph, we observe that the p-values are all of them in a neighbourhood of 0.05, and therefore δ and δ' take small values.

In order to get further information, we have computed and plotted, for every sample $(\vec{x},\vec{y})$, the upper and lower posterior probabilities from the IDP against the corresponding p-value, for different values of Δ and different sample sizes. Due to length restrictions, we just include the graphs for a specific choice of both sample sizes. In particular we have selected $n_1 = n_2 = 10$:



According to the above simulations, $\overline{P}(H_0|(\vec{x},\vec{y}))$ and $\underline{P}(H_0|(\vec{x},\vec{y}))$ can be written as functions of the p-value. For other sample sizes we have observed a similar shape of the graph, although the

difference $\overline{P}(H_0|(\vec{x},\vec{y})) - \underline{P}(H_0|(\vec{x},\vec{y}))$ is well known to decrease with respect to both sample sizes.

Furthermore, these two functions do not depend on the particular choice of Δ. Notwithstanding, it is well known that the p-value follows a uniform distribution over the unit interval when $\Delta = 0$, and as far as we get far away from $\Delta = 0$, the distribution of the p-values tends to concentrate over an extreme of the interval (the left side extreme for positive values of Δ and the right extreme for negative valued of Δ). The pair of upper and lower posterior probabilities also concentrate over the same extremes of the intervals for big values of Δ.

Let us now analyse some features of this functional relation. The p-value (which coincides with the posterior probability of the BB-DP test) is always bounded by $\overline{P}(H_0|(\vec{x},\vec{y}))$ and $\underline{P}(H_0|(\vec{x},\vec{y}))$, but it does not coincide with their half sum in general. On the other hand, when we plot their difference against the p-value, we observe that it increases from 0 to 0.5 and decreases from 0.5 to 1.

According to Equation 1, the p-value of a pair of samples $(\vec{x},\vec{y})$ can be expressed $p(\vec{x},\vec{y}) = G_0(U(\vec{x},\vec{y}))$, with $G_0 = 1 - F_0$, where U denotes the MWW statistic and F_0 denotes the cdf of U under the assumption $\theta = 0.5$. The cdf F_0 corresponds to a unimodal distribution, and symmetric around μ_0. In other words, the density function f_0 is increasing on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞). Let us now consider $g_1(u) = G_0(u-\epsilon) - G_0(u)$ and $g_2(u) = G_0(u) - G_0(u+\epsilon)$. Both functions are increasing on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞). Therefore we easily deduce that:

If either $U(\vec{x},\vec{y}) < U(\vec{x}',\vec{y}') < \mu_0$ or $U(\vec{x},\vec{y}) > U(\vec{x}',\vec{y}') > \mu_0$, then

$$g_1(U(\vec{x},\vec{y})) < g_1(U(\vec{x}',\vec{y}')) \text{ and } g_2(U(\vec{x},\vec{y})) < g_1(U(\vec{x}',\vec{y}')).$$

According to our Monte Carlo simulations, this is exactly what happens with the differences $\overline{P}(H_0|(\vec{x},\vec{y})) - P(H_0|(\vec{x},\vec{y}))$ and $P(H_0|(\vec{x},\vec{y})) - \underline{P}(H_0|(\vec{x},\vec{y}))$, i.e.:

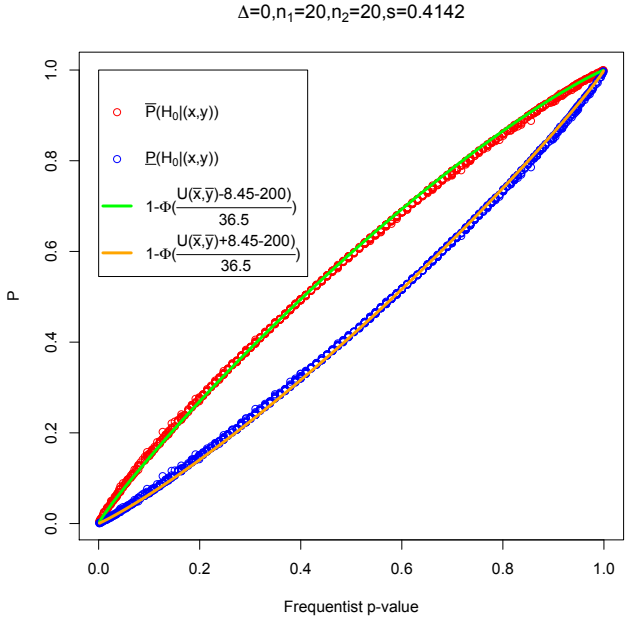
If $U(\vec{x},\vec{y}) < U(\vec{x}',\vec{y}') < \mu_0$ or $U(\vec{x},\vec{y}) > U(\vec{x}',\vec{y}') > \mu_0$, then

$$\overline{P}(H_0|(\vec{x},\vec{y})) - P(H_0|(\vec{x},\vec{y})) < \overline{P}(H_0|(\vec{x}',\vec{y}')) - P(H_0|(\vec{x}',\vec{y}')) \text{ and}$$

$$P(H_0|(\vec{x},\vec{y})) - \underline{P}(H_0|(\vec{x},\vec{y})) < P(H_0|(\vec{x}',\vec{y}')) - \underline{P}(H_0|(\vec{x}',\vec{y}')).$$

Therefore, it seems that the difference $\overline{P}(H_0|(\vec{x},\vec{y})) - P(H_0|(\vec{x},\vec{y}))$ is increasing with respect to $U(\vec{x},\vec{y})$ on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞). Something similar happens with the difference $P(H_0|(\vec{x},\vec{y})) - \underline{P}(H_0|(\vec{x},\vec{y}))$.

Since $P(H_0|(\vec{x},\vec{y}))$ coincides with $p(\vec{x},\vec{y}) = G_0(U(\vec{x},\vec{y}))$ then we can deduce that there is a strictly increasing relation between $\overline{P}(H_0|(\vec{x},\vec{y}))$ and $P(U(\vec{x},\vec{y}) - \epsilon)$, for an arbitrary but fixed ϵ and the same happens with $\underline{P}(H_0|(\vec{x},\vec{y}))$ and $P(U(\vec{x},\vec{y}) + \epsilon)$. We have examined the nature of this strictly increasing (one-to-one) correspondence, and we have observed that it is in fact the identity.



This opens a door to the reconciliation between the Bayesian and the frequentist approaches also in the imprecise framework, following the path of Casella and Berger (1987) for the precise case. On one hand, there is a one-to-one correspondence between the upper (resp. the lower) posterior probability of the IDP test and the p-value of the MWW test. Furthermore, we can easily construct an imprecise test that relies on the MWW U-statistic and that mimics the behavior of the IDP test. Let us take an arbitrary α and let us define the new imprecise test as follows:

$$\delta(\vec{x},\vec{y}) = \begin{cases} 0 & \text{if } U(\vec{x},\vec{y}) \leq c_\alpha - \epsilon \\ 1 & \text{if } U(\vec{x},\vec{y}) > c_\alpha + \epsilon \\ ? & \text{otherwise,} \end{cases} \tag{2}$$

where c_α is such that $P_{\theta_0}(U > c_\alpha) = \alpha$.

According to our simulations, for a specific choice of $\gamma = \frac{K_0}{K_0+K_1}$ and the triple (s, n_1, n_2), there exists $\epsilon = g(s, n_1, n_2)$ such that the above imprecise test for $\alpha = \gamma$ coincides with the IDP test. Furthermore, the upper and lower posterior probabilities of the null hypothesis do respectively coincide with $G_0(U(\vec{x},\vec{y}) + \epsilon)$ and $G_0(U(\vec{x},\vec{y}) - \epsilon)$.

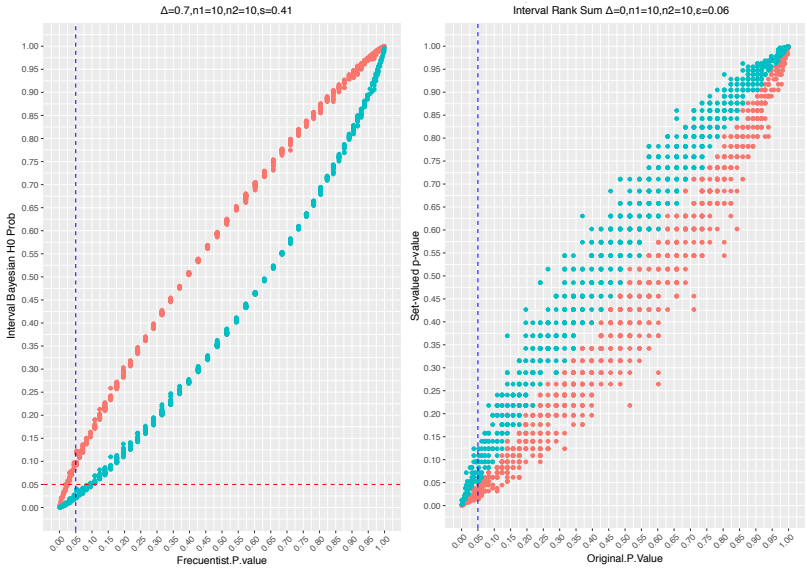
3.3 Conclusions and future directions

We have constructed an imprecise "frequentist" test that mimics the behavior of the so-called IDP test. It basically works as follows: it calculates the interval of values $(U(\vec{x},\vec{y}) - \epsilon, U(\vec{x},\vec{y}) + \epsilon)$ and it inherently considers the collection of samples $(\vec{x}', \vec{y}')$ such that $U(\vec{x},\vec{y}) - \epsilon < U(\vec{x}',\vec{y}') < U(\vec{x},\vec{y}) + \epsilon$. If all of them are either in the rejection or the acceptance zone of the frequentist test, then the decision is clear. Otherwise, the outcome of the test is indeterminate. Thus, we conclude that, at least for the MWW test, the kind of "imprecisiation" over the set of priors considered in the IDP-based test may produce similar effects on the decision mechanism as an imprecisiation around the statistic values.

Let us notice that, in this specific case where the statistic is based on the ranks of the observations, but not on their numerical values, the statistic is not continuous with respect to those numerical values. Thus the following alternative test (see Perolat et al. (2015) for further discussion about it):

$$\begin{cases} 0 & \text{if } p(\vec{x}', \vec{y}') \leq \alpha, \forall\, (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ 1 & \text{if } p(\vec{x}', \vec{y}') > \alpha, \forall\, (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ ? & \text{otherwise.} \end{cases} \tag{3}$$

would produce different outcomes in practice, as we observe below:



Notwithstanding, for other frequentist tests, different from MWW, based on continuous statistics the variation proposed in Equation 3, could report similar results as the one provided in Equation 2, for adequate selections of ϵ and δ. For those cases, it seems that the kind of imprecisiation over the set of priors considered in the IDP-based test may produce similar effects on the decision mechanism as an imprecisiation around the sample values. Let us remind the reader that our empirical comparison in this paper refers to the case where the frequentist test completely coincides with the Bayesian one. But this may be not the case for other tests where the MLR condition is not satisfied.

In those cases we might directly compare the IDP-based test with an imprecise version of the Bayesian test as follows:

$$\begin{cases} 0 & \text{if } P(H_0|\vec{x}', \vec{y}') \leq \gamma, \forall\, (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ 1 & \text{if } P(H_0|\vec{x}', \vec{y}') > \gamma, \forall\, (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ ? & \text{otherwise.} \end{cases} \tag{4}$$

Such a comparison could shed further light about the behaviour of IDP-based tests in practice. We conjecture that they could lead to similar decision rules. If our conjecture is true, this alternative procedure would lead to equivalent but computationally more efficient algorithms. On the other side, it would reflect that the kind of imprecisiation over the priors considered by this almost-ignorance model produces similar effects in the decision procedure as an imprecisiation around the sample values.

Acknowledgments

This paper has been partially supported by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias). We thank three anonymous reviewers for insightful comments on our manuscript.

References

A. Benavoli, F. Mangili, F. Ruggeri, and M. Zaffalon. Imprecise Dirichlet process with application to the hypothesis test on the probability that $X \leq Y$. *Journal of Statistical Theory and Practice*, 9(3):658–684, 2015.

J. Berger and T. Selke. Testing a point null hypothesis: the irreconciability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–120, 1987.

G. Casella and R. L. Berger. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397):106–111, 1987.

I. Couso, S. Moral, and L. Sánchez. The behavioral meaning of the median. *Information Sciences*, 294:127–138, 2015.

M. H. DeGroot. Doing what comes naturally: Interpreting a tail area as a posterior probability or likelihood ratio. *Journal of the American Statistical Association*, 68(344):966–969, 1973.

B. Efron. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1 (2):209–230, 1973.

H. Jeffreys. *Theory of Probability*. Oxford University Press., 1939.

D. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

H. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

M. P. M.P. Fay. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

J. Perolat, I. Couso, K. Loquin, and O. Strauss. Generalizing the Wilcoxon rank-sum test for interval data. *International Journal of Approximate Reasoning*, 56(A):108–121, 2015.

J. Pratt. Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society, Ser. B*, 27(2):169–203, 1965.

D. Rubin. Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

G. Shafer. Lindley's Paradox. *Journal of the American Statistical Association*, 77(378):325–351, 1982.