Unified Hypothesis Testing for Bayesian and Frequentist Approaches

Ivair Silva
Department of Statistic, Federal University of Ouro Preto, Ouro Preto, MG, Brasil
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Healph Care Institute, Boston, MA, USA - ivairest@gmail.com

Abstract

Hypothesis testing plays a fundamental role in the modern practice of statistical analysis. Essentially, there are two approaches for performing hypothesis testing, the Bayesian and the frequentist. These approaches not always agree about which of two hypotheses is to be taken as true. Currently, the efforts to accommodate both approaches under the same decision rule are not applicable for the general case of any hypothesis testing problem. This paper offers an unified approach that enables to place Bayesian and frequentist tests under the same decision rule. The post-market vaccine safety surveillance problem is used to illustrate how to construct a test based on the proposed unified approach.

Keywords: Decision set; evidence measures; Bayes Factor; p-value; performance measures.

1. Introduction In theoretical statistical inference, a hypothesis is a statement about a population parameter, $\theta \in \Re$. The objective of a statistical hypothesis test is to decide which of two hypotheses is true. The general format of the two hypotheses is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_i, $i = 0, 1$, is a subset of the parameter space, Θ, $\Theta_0 \cap \Theta_1 = \emptyset$, and they not necessarily form a partition of Θ. H_0 and H_1 are called the null hypothesis and the alternative hypothesis, respectively. The decision of taking H_i as true is based on a random sample, $X_1, \cdots, X_n$, from the population. In practice, an arbitrary real-valued function of the sample, the 'test statistic', is the measure used to define the decision rule. A transformation of the test statistic can be applied in order to construct a 'measure of evidence in favor of H_i', $i = 0, 1$. Small observed values for the measure of evidence in favor of H_i suggest rejection of H_i. Therefore, the use of evidence measures to draw a decision has both, intuitive and practical appeals. However, the reasoning used to construct an evidence measure is the main point of divergence between the Bayesian and the frequentist approches.

In the frequentist approach, the well-known evidence measure is the p-value. The p-value is a random variable, with support in the $(0, 1)$ interval, that enables the use of the same critical value, α, for any problem, where α is the so called significance level, an arbitrary bound for the Type I error probability. The use of p-values is severely criticized by the Bayesian practitioners. According to them, the decision rule should consider not only the Type I error probability, but also other performance measures like the expected loss incurred by the possible wrong decisions from a test or the cost of such decisions. The Bayesian approach consists on using the observed sample from the population to update the analyst's uncertainty, apart the data, about the plausibility of H_i. With the Bayes rule, the analyst can then obtain the so called posterior uncertainty (posterior distribution) about the veracity of each hypothesis. This reasoning has an elegant and intuitive appeal, but it is also criticised by frequentists. Under the frequentist point of view, because θ is an unknown but fixed constant, the probability of H_i come to be true, given the data or not, is 0 or 1 (CASELLA; BERGER, 2000, p. 379), then such probability should not be confounded with the posterior uncertainty of the analyst.

The discussion concerning the correct reasoning to construct an evidence measure, and the correct post-experimental interpretation of it, has generated an enormous amount of scientific papers and books. But this paper is not dedicated to contribute to such discussion. Instead, here the goal is to reconcile the goods of each approach in an unified, and simple, methodology. Many authors have presented proposals to unify

the Bayesian and the frequentist decision rules for specific forms of hypotheses and for a limited class of data distributions. Berger et al. (1997) introduced an unified test for a precise null hypothesis versus a composite alternative hypothesis. Sarat (2001) presented an interesting solution for discrete distributions. Casella & Berger (1998) reconciled frequentist and Bayesian tests for the one-sided testing problem. But, a general theory is still an open problem. Here this challenge is faced by introducing a non-conventional, but still frequentist, class of p-values. Elements of this class are defined as functions of the Bayesian measure of evidence. Consequently, the decision rule becomes concordant between the two approaches, but all the well-known performance properties from both approaches are preserved. Presenting an overview of the main definitions concerning the hypothesis test theory is, apparently, a natural path for constructing an unified theory, thus, this is the subject of the next section. Section 3 introduces the unified proposal, and Section 4 illustrates the use of the unified approach in the construction of a test for the post-market vaccine safety problem. Section 5 lists the main conclusions.

2. Overview of Hypothesis Test Approaches The decision rule used for taking H_i as true is the neuralgic point for the objectives of this work.

Definition 1.(***hypothesis test***) A hypothesis test is a rule that establishes for which values of the sample the hypothesis H_0 is taken as true and for which values the hypothesis H_1 is taken as true.

Some authors have encouraged the use of an alternative form for the decision rule allowing for a third option, the no-decision option. When a no-decision region is incorporated for testing precise hypotheses, the frequentist and the Bayesian methods can be made equivalent (BERGER et al., 1994). Such approach is particularly important for the field of sequential analysis (JENNISON; TURNBULL, 2000). However, as shall be shown in Section 3, an unified approach is possible without the requirement of a no-decision region.

Let $F_X(x|\theta)$ denote the probability distribution of the random variable X parameterized by θ, a fixed and unknown constant. If $F_X(x|\theta)$ is a continuous distribution, let $f_X(x|\theta)$ be the corresponding probability density function. Without loss of generality, the reasoning and definitions are described for the continuous case, but they can be easily extended to the discrete case by analogy. Let $W(\tilde{\mathbf{X}})$ be a real-valued function of the random sample $\tilde{\mathbf{X}} = (X_1, \cdots, X_n)$, called 'test statistic', and let $\aleph$ be the sample space of $\tilde{\mathbf{X}}$. The decision rule for choosing H_i as true is usually based on a subset of the real line.

Definition 2. (***decision set***) For a given test statistic $W(\tilde{\mathbf{X}})$, and an observed sample $\tilde{\mathbf{x}}$, H_1 is taken as true if $W(\tilde{\mathbf{x}}) \in \mathbf{R}(c)$, and H_0 is taken as true if $W(\tilde{\mathbf{x}}) \notin \mathbf{R}(c)$. $\mathbf{R}(c)$ is called 'decision set' and given by $\mathbf{R}(c) = \{w \in \Re : w \geq c\}$, where c is an arbitrary constant named 'critical value'. Naturally, $\mathbf{R}(c)$ can also be defined for the reverted inequality.

2.1. Bayesian Approach Before looking to a sequence of realizations of X, an analyst may have some idea of what would be the most plausible values of θ and, conversely, what are the less plausible possibilities of it. This uncertainty about which values are more, or less, plausibles for θ can be subjectively transmitted through a kind of probability measure. This probability measuring of the analyst's uncertainty apart the data is called prior distribution, and its respective probability density function is denoted here by $\pi_\theta(y)$, $y \in \Re$. For a lighter terminology, here $\pi_\theta(y)$ shall be called just by the prior distribution instead of the density associated to the prior distribution. If $\pi_\theta(y)$ is equal to (or very close to) zero for a fixed value of y, it means that, under the analyst's view, y is almost surely not the value of θ. If $\pi_\theta(y)$ is elevated for a certain y, then the analyst believes that θ may be very close to y.

Using the empirical information, $\tilde{\mathbf{x}}$, the analyst can update his uncertainty about θ by using the Bayes' rule. This updated conditional probability measure, denoted with $\pi_\theta(y|\tilde{\mathbf{x}})$, is called 'posterior distribution'. Thus:

$$\pi_\theta(y|\tilde{\mathbf{x}}) = f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}|y)\pi_\theta(y)/m(x), \tag{1}$$

where $f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}|y)$ is the likelihood function evaluated at $\theta := y$, and $m(\tilde{\mathbf{x}}) = \int_{-\infty}^{\infty} f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}|y)\pi_\theta(y)dy$.

The evidence measure used for constructing the decision rule is then based on $\pi_\theta(y|\tilde{\mathbf{x}})$. The two main measures of evidence are the posterior distribution of H_0 given the data, and the Bayes Factor (BERGER et al., 1997). These two measures produce equivalent decision regions if the hypotheses are precise, i.e., of the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, with $\theta_0 \neq \theta_1$ constants. Other two famous Bayesian evidence measures are the e-value, proposed by Pereira (1999), and the posterior risk (BRIGHENTI, 2007). Because the unified approach introduced in Secion 3 is applicable for any Bayesian evidence measure, only the Bayes Factor is described here for illustrating the Bayesian reasoning. The Bayes Factor measure in favor of H_0, $BF(\tilde{\mathbf{X}})$, is the ratio of corresponding posterior to prior odds,

$$BF(\tilde{\mathbf{X}}|\pi_\theta) = \frac{Pr\left[\theta \in \Theta_0|\tilde{\mathbf{X}}, \pi_\theta\right] Pr\left[\theta \in \Theta_1|\pi_\theta\right]}{Pr\left[\theta \in \Theta_1|\tilde{\mathbf{X}}, \pi_\theta\right] Pr\left[\theta \in \Theta_0|\pi_\theta\right]}. \tag{2}$$

It is worth noting that the probabilities above are with respect to the prior distribution $\pi_\theta(y)$. Also, it is remarkable that, for a realized value $\tilde{\mathbf{x}}$, and if the hypotheses are precise with the prior given as the two point distribution $\pi_\theta(\theta_0) = a$ and $\pi_\theta(\theta_1) = 1 - a$, $a \in (0, 1)$, then the Bayes Factor in favor of H_0 becomes the likelihood ratio $f_X(\tilde{\mathbf{x}}|\theta_0)/f_X(\tilde{\mathbf{x}}|\theta_1)$.

In the general case, before having a realized value of $\tilde{\mathbf{X}}$, the random variable $BF(\tilde{\mathbf{X}}|\pi_\theta)$ is actually a test statistic, and then a decision set of the form given in Definition 2 can be used to make a decision.The choice of the critical value c can be guided by the scale suggested by Jeffrey (1961). Basically, the scale favors a subjective interpretation of $BF(\tilde{\mathbf{x}}|\pi_\theta)$ in the logarithm scale. If $0 < -\log BF(\tilde{\mathbf{x}}|\pi_\theta) \leq 0.5$, the evidence against H_0 is poor. $0.5 < -\log BF(\tilde{\mathbf{x}}|\pi_\theta) \leq 1$ means substantial evidence against H_0. For $1 < -\log BF(\tilde{\mathbf{x}}|\pi_\theta) \leq 2$, the evidence against H_0 is strong. Finally, if $-\log BF(\tilde{\mathbf{x}}|\pi_\theta) > 2$, the evidence against H_0 is decisive.

Alternatively, c can be fixed in a way to comply with requirements over some performance measure. For example, consider hypotheses of the form $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, where θ_0 is a fixed constant. If $\theta \geq \theta_0$ and H_1 is taken as true, a mistake has been made, i.e., the error of the Type I has occurred. If $\theta < \theta_0$ and H_0 is taken as true, a mistake has been made, i.e., the error of the Type II has occurred. The other possibilities represent the correct actions. The decision theoretic analysis offers the concept of 'risk function", also suggestively named 'expected loss'. The risk function is a performance measure for a decision rule in a test. It is intuitive that some wrong decisions are more serious than others. Considering the hypotheses of the form above again, if the Type I error occurs but θ is slightly bigger than θ_0, then the mistake may not be very serious. But the mistake is possibly serious if θ is much bigger than θ_0. The intensity of how 'serious' a mistake is, for each possible value of θ, can be transmitted through a real-valued function of θ, the 'loss function' $l_j(\theta)$, $j = 1, 2$. For $j = 1$, the loss function represents the loss associated to the Type I error, and $j = 2$ is for the loss associated to the Type II error. Let $L(\theta)$ denote the loss function as a random variable. Returning to general case, the risk function, $r(\theta)$, for a fixed θ, is:

$$r(\theta) = E[L(\theta)] = l_1(\theta) \times Pr[BF(\tilde{\mathbf{X}}|\pi_\theta) \in \mathbf{R(c)}|\theta] \times I_{\{\theta \in \Theta_0\}}(\theta) + l_2(\theta) \times Pr[BF(\tilde{\mathbf{X}}|\pi_\theta) \notin \mathbf{R(c)}|\theta] \times I_{\{\theta \in \Theta_1\}}(\theta).$$

Therefore, a good candidate for c is the one that minimizes $r(\theta)$. Or, if the global minimum does not exist, restricted minimization can be tried by imposing restrictions to other measures, like to the Type II error probability associated to elevated losses, for example.

2.2. Frequentist Approach A particular way of constructing a test is to restrict the choice of $\mathbf{R}(c)$ to the class that promotes a Type I error probability smaller than or equal to a pre-specified constant, α. Tests with this property are said of 'α-level', i.e., $\mathbf{R}(c)$ promotes a α-level test if, for each $\theta^* \in \Theta_0$, holds that $Pr[W(\tilde{\mathbf{X}}) \in \mathbf{R}(c)|\theta = \theta^*] \leq \alpha$. There is no strong concerns for imposing particular shapes for $W(\tilde{\mathbf{X}})$. Usually, what really matters is the control over the Type I error probability. Thus, unlike the Bayesian approach, the target performance measure is the Type I error probability, not the risk function (remind that the risk has the Type I error probability on its calculation). One can also invert the priority and choose c to comply with a certain bound for the Type II error probability, but the reasoning leads, in essence, to the same method because of the trade-off between these measures. Increasing the first will automatically decrease the second.

Anyway, although not common, the risk function can also be the target measure in the construction of a frequentist test.

Conventionally, the evidence measure in favor of H_0 is the so called 'p-value'. A quite general definition of p-value is given by Casella and Berger (2001) as following:

Definition 3. (***p-value***) A p-value $p(\tilde{\mathbf{X}})$ is a test statistic satisfying $0 \leq p(\tilde{\mathbf{x}}) \leq 1$ for each point $\tilde{\mathbf{x}} \in \aleph$. H_1 is evidentiated as being true if $p(\tilde{\mathbf{x}}_0)$ is small for a given observed $\tilde{\mathbf{x}}_0$ of $\tilde{\mathbf{X}}$.

Just as the Bayes Factor measure, the p-value is designed to be an evidence measure that favors the subjective interpretation of how strong is the evidence against H_0. A p-value much close to zero indicates a strong evidence against H_0. Also, because the p-value is a random variable with support in the $(0, 1)$ interval, it can also be interpreted as a test statistic given in a normalized scale.

Important: The general definition of p-value given above does not support the restricted and, unfortunately, extensively spread wrong interpretation of p-value as a data-dependent probability. Consider, for example, the likelihood ratio test statistic $\lambda(\tilde{\mathbf{x}}) := \sup_{\theta^* \in \Theta_0} f_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}|\theta^*) / \sup_{\theta \in \Theta} f_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}|\theta)$. According to Definition 3, $\lambda(\tilde{\mathbf{X}})$ is a p-value because, for each point $\tilde{\mathbf{x}} \in \aleph$, (i) $0 < \lambda(\tilde{\mathbf{x}}) < 1$, and (ii) a small value of $\lambda(\tilde{\mathbf{x}})$ indicates that H_0 is not likely to be true, i.e., H_1 would be $1/\lambda(\tilde{\mathbf{x}})$ times more likely to be true than H_0. This is in complete agreement with the likelyhood principle (CASELLA; BERGER, 2001, p. 291). But, obviously, $\lambda(\tilde{\mathbf{x}})$ is not a post-experimental probability.

Like in Section 2.1, the critical value c has to be fixed before having a realization of $\tilde{\mathbf{X}}$, and this choice depends on each problem and of which is the performance measure of interest. Because the Type I error probability is, usually, the performance measure of interest in frequentist tests, a convenient class of frequentist evidence measures is the class of 'valid p-values'.

Definition 4. (***valid p-value***) A p-value $p(\tilde{\mathbf{X}})$ is valid if, for each $\theta^* \in \Theta_0$, and arbitrary $0 \leq \alpha \leq 1$, holds that $Pr[p(\tilde{\mathbf{X}}) \leq \alpha | \theta = \theta^*] \leq \alpha$.

The advantage of using a valid p-value as test statistic is that the corresponding critical value is an universal constant. It is given by the own level of the test, the constant α. This direct relation of the critical value of the statistic $p(\tilde{\mathbf{X}})$ with the level of the test is a convenient and elegant characteristic of the frequentist approach. Thus, the use of p-values for such problems has not logical flaws as a measure of evidence used solely to make the decision. The bad reputation of the p-value may be explained by the wrong interpretation from some practitioners that the p-value is a post-experimental probability of H_0 being true, or of being rejected, or some other probabilistically erroneous interpretations. Ironically, the wrong viewing of a p-value as a data-dependent probability may had unfortunately been reinforced by the extensive use of a convenient and very famous class of p-values, the p-value of the tail-type.

Definition 5. (***p-value of the tail-type***) Let $W(\tilde{\mathbf{X}})$ be a test statistic. For an observed $\tilde{\mathbf{x}}$, the p-value of the tail-type, $p_t(\tilde{\mathbf{x}})$, is given by: $p_t(\tilde{\mathbf{x}}) = \sup_{\{\theta^* \in \Theta_0\}} Pr[W(\tilde{\mathbf{X}}) \in \mathbf{R}(W(\tilde{\mathbf{x}})) | \theta = \theta^*]$.

The tail-type p-value is always valid, i.e., $Pr[p_t(\tilde{\mathbf{X}}) \leq \alpha | \theta = \theta^*] \leq \alpha$, for each $\theta^* \in \Theta_0$ (CASELLA; BERGER, 2001, p. 397). This important property may have contributed to its extensive use. But, the fact that it is constructed in terms of the tail of the test statistic distribution, evaluated at the supremum in Θ_0, might help to explain why so many people wrongly interpret the tail-type p-value as the probability of H_0 be rejected. Berger and Boos (1994) proposed an alternative class of p-values, the p-values of the confidence set type, which is also valid and can produce tests with improved power.

The conventional strategy is to find a function $W(\tilde{\mathbf{X}})$ such that, for a fixed c, promotes the smallest Type II error probability, for each $\theta \in \Theta_1$, in the class of α-level tests. When it exists, such test is called the most powerfull test. For precise hypotheses, the solution is given by the Neyman-Pearson Lemma (CASELLA;

BERGER, 2001, p.388). An important fact is that the Bayesian test based on the Bayes Factor of minimum risk is equivalent to the most powerful test of Neyman-Pearson Lemma (BRIGHETI, 2007). But this type of unification has not been presented for the general case yet.

3. On an Unified Approach Let $T(\tilde{\mathbf{X}})$ denote a statistic for θ based on the random sample $\tilde{\mathbf{X}}$. A particular but quite usual option is $T(\tilde{\mathbf{X}}) := \tilde{\mathbf{X}}$. Let $B\left(T(\tilde{\mathbf{X}})\right)$, or simply $B(T)$, be some Bayesian measure of evidence in favor of H_0, based on $T(\tilde{\mathbf{X}})$, and such that H_1 is taken as true if $B(T)$ is small. Let k be the critical value for $B(T)$, i.e., H_1 is taken as true if $B(T) \in \mathbf{R}(k) := \{b \in \Re : b \leq k\}$. Thus, the 'unified p-value' is defined as following.

Definition 6. (***unified p-value***) Let $B(T)$ be a Bayesian measure of evidence. If t is an observed value of $T(\tilde{\mathbf{X}})$, the unified p-value, $p_u(\tilde{\mathbf{X}})$, is given by: $p_u(\tilde{\mathbf{X}}) = \sup_{\{\theta^* \in \Theta_0\}} Pr\left[B(T) \in \mathbf{R}\left(B(t)\right) | \theta = \theta^*\right]$.

Thus, the Bayesian and the frequentist approaches are equivalent if H_1 is taken as true for $p_u(\tilde{\mathbf{X}}) \leq \alpha$ and, otherwise, H_0 is taken as true. The level of the test is $\alpha = \sup_{\{\theta^* \in \Theta_0\}} Pr\left[B(T) \in \mathbf{R}\left(B(k)\right) | \theta = \theta^*\right]$. This p-value is of the tail-type, and then it is also valid (see Section 2.2). This is a convenient property because, in applications where the Type I error is of meaningful importance, the critical value is equal to α, the desired level, just as the common practice in a frequentist test. That this approach has all elements of a Bayesian test is obvious. To see that it has all of the frequentist too, observe that, according to the descriptions given in Section 2.2, in this setting $W(\tilde{\mathbf{X}}) := B\left(T(\tilde{\mathbf{X}})\right)$, and a p-value can be used to make the decision with the level of the test previously known. More important: it does not matter if the Bayesian or if the frequentist approach is focused by the user, the decision is the same in both for a fixed data set. These conclusions can be conveniently remarked through a theorem.

Theorem (**Equivalence of Bayesian and Frequentist tests**). *Consider the class of hypothesis tests of the form given in Definition 1. Thus, for each Bayesian test there always exists an equivalent frequentist test. Conversely, for each frequentist test, there always exists an equivalent Bayesian test.*

Proof. Let $\aleph$ be the sample space of $\tilde{\mathbf{X}}$. To prove the first assertion, let $B(T(\mathbf{X}))$ to represent the Bayesian measure of evidence specified for a certain problem with decision region $\mathbf{R}_B(k) = \{b \in \Re : b \leq k\}$. The equivalent frequentist test is simply obtained by using the p-value $p_u(\tilde{\mathbf{X}})$ from Definition 6. Therefore, for each $\tilde{\mathbf{x}} \in \aleph$, $B(T(\tilde{\mathbf{x}})) \in \mathbf{R}_B(k)$ iff $p_u(\tilde{\mathbf{x}}) \leq \alpha$, where $\alpha = \sup_{\{\theta^* \in \Theta_0\}} Pr\left[B(T) \in \mathbf{R}_B\left(B(k)\right) | \theta = \theta^*\right]$.
To prove the converse, let $T(\tilde{\mathbf{X}})$ be a test statistic with support in the real line. The decision set used for the frequentist test in this demonstration is of the form $\mathbf{R}(c) = \{t \in \Re : t \geq c\}$, but the same reasoning is applicable for decision sets of the form $\{t \in \Re : t \leq c\}$. If t_0 is a realized value of $T(\tilde{\mathbf{X}})$, define the event $A_{t_0} = \{t \in \Re : t \geq t_0\}$. An equivalent Bayesian test is obtained by taking H_1 as true if $\pi(A_{t_0}) \in \mathbf{R}_B(k^*) = \{b \in \Re : b \leq k^*\}$, where $\pi(A_{t_0})$ is a Bayesian measure of evidence given by $\pi(A_{t_0}) = Pr(A_{t_0}) \times Pr[\theta \in \Theta_0 | A_{t_0}, \pi_\theta]$, and π_θ is an arbitrary prior distribution. Because $\pi(A_{t_0})$ is non-increasing with t_0, the solution for obtaining the equivalent decision set is given by fixing $k^* := \pi(A_c)$. Therefore, for each $\tilde{\mathbf{x}} \in \aleph$, $T(\tilde{\mathbf{x}}) \in \mathbf{R}(c)$ iff $\pi(A_{t_0}) \in \mathbf{R}_B(k^*)$. □

4. Example Using the Unified Approach - on a test for post-market vaccine safety surveillance In the post-market vaccine safety surveillance context, X_t is the random variable that counts the number of adverse events in a known risk window from 1 to W days after a vaccination that was administrated in a period $[0, t]$ (LIEU et al., 2007). Commonly, under the null hypothesis, X_t is supposed to have a Poisson distribution with mean μ_t, where μ_t is a known function of the population at risk, adjusted for age, gender and any other covariates of interest. Under H_1, X_t is Poisson with mean $\theta\mu_t$, where θ is the unknown increased relative risk due to the vaccine. The form of the hypotheses to be tested are $H_0 : \theta \leq 1$ versus $H_1 : \theta \geq 1$. Usually, the monitoring of adverse events is made by a sequential analysis fashion but, by simplicity, here the test is constructed for a single analysis, then, the index t is neglected, i.e., X_t is ferreded only as X here. Also, in this example μ is supposedly equal to 20.

In this example, set $T(X) = X$. For the prior distribution, consider $\pi_\theta(y) = e^{-y} I_{(0,\infty)}(y)$. For the Bayesian measure of evidence, adopt the Bayes Factor $BF(\tilde{\mathbf{X}}|\pi_\theta)$, expression (2). Inspired on the Jeffrey's (1961) scale described in Section 2.1, consider to use a critical value equal to 2 in the minus log scale, which corresponds to a critical value of $k = e^{-2} = 0.1353$ in the $BF(\tilde{\mathbf{X}}|\pi_\theta)$ scale. Therefore, the decision set is $R(0.1353) = \{b \in \Re : b \leq 0.1353\}$. Denote $Pr[\theta \in \Theta_0|\tilde{\mathbf{X}}]$ simply buy $\pi(\tilde{\mathbf{X}})$ and $\pi_0 = Pr[\theta \in \Theta_0|\pi_\theta] = 1 - e^{-1}$. The probability of taking H_1 as true is:

$$Pr[BF(\tilde{\mathbf{X}}|\pi_\theta) \leq 0.1353|\theta] = Pr[\pi(\tilde{\mathbf{X}}) \leq 0.1353 \times \pi_0(1 - 0.8647\pi_0)^{-1}|\theta] = Pr(X \geq 25|\theta). \quad (3)$$

Concerning the loss function, let $l(\theta) = I_{(0,1]}(\theta)/\theta + \theta I_{(1,\infty)}(\theta)$. With this, the performance measures (Type I and Type II error probabilities, and expected loss) can be easily calculated. If the true θ is equal to 0.5, this tuning setting parameterization leads to an expected loss equal to $Pr(X \geq 25|\theta = 0.5)/0.5 = 0.3135$. If $\theta = 1.5$, the expected loss is $1.5Pr(X < 25|\theta = 1.5) = 1.2648$. The exact size of this tet is $Pr(X \geq 25|\theta = 1) = 0.1568$, which is also the critical value in the p-value scale. At this moment, and before the realization of $\tilde{\mathbf{X}}$, the user has the opportunity to adjust k in order to obtain a smaller size. But, this would also increase the Type II error probability and then the gain in terms of expected loss is not guaranteed. Thus, for this example, the critical value $k = 0.135$ is kept because the current resulting expected loss is of major importance. Observe that k was previously define following the Bayesian reasoning. Conversely, k could be fixed as a result of a previously size α required for the test.

5. Conclusions The unified approach introduced in this paper is quite general in the sense of being applicable for data having any probability distribution and any form for the hypotheses. Prior to the realization of the test, the approach enables the analyst: (i) to use its prior uncertainty about θ in the decision rule when this is desired; (ii) to be aware of the maximum Type I error probability (size) of the test; (iii) to be aware of the expected loss; (iv) to express the Bayesian measure of evidence in the normalized (0,1) scale (p-value).

References

Berger, J. 0., & Boukai, B., & Wang, Y. (1997). Unified Frequentist and Bayesian Testing of a Precise Hypothesis, Statistical Science, Vol. 12, No. 3, p. 133-160.

Berger, R. L., & Boos, D. D. (1994). P-values Maximized over a Confidence Set for the Nuisance Parameter. Journal of the American Statistical Association, vol. 89, p. 1012-1016.

Berger, J. O., & Brown, L. D., & Wolpert, R. L. (1994). A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing, The Annals of Statistics, Vol. 22, No. 4, p. 1787-1807.

Brighenti, C. R. G. (2007). Teste com Erros Frequentistas Condicionais e Testes com Interpretao Bayesiana e Frequentista Condicional. PhD Thesis. Federal University of Lavras, Lavras, MG, Brazil.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem (with discussion). Journal of the American Statistical Association, Vol. 82, 106-111.

Casella, G., & Berger, R. L. (2001). Statistical Inference - Second Edition. Duxbury Resource Center, Thomson Learning.

Jeffrey, H. (1961). The Theory of Probability (3 ed.). Oxford. p. 432

Lieu, T. & Kulldorff, M. & Davis, R. & Lewis, E. & Weintraub, E. & Yih, W. & Yin, R. & Brown, J. & Platt, R. (2007) Real-time vaccine safety surveillance for the early detection of adverse events. Medical Care, Vol. 45(S), 8995.

Pereira, C. A. B., & Stern, J. M. (1999). Evidence and credibility: full bayesian significance test of precise hypothesis. Entropy, 1:99-110.