Article

Significance Tests for Binomial Experiments: Ordering the Sample Space by Bayes Factors and Using Adaptive Significance Levels for Decisions.

Carlos A. de B. Pereira[1,*], Adriano Polpo[2] and Eduardo Y. Nakano[3]

[1] Institute of Mathematics and Statistics, University of São Paulo, São Paulo 05508-090, Brazil; cadebp@gmail.com
[2] Department of Statistics, Federal University of São Carlos, São Carlos 13565-905, Brazil; polpo@ufscar.br
[3] Department of Statistics, University of Brasília, Brasília 70910-900, Brazil; nakano@unb.br

* Correspondence: cpereira@ime.usp.br; Tel.: +55 11 99115 3033

Abstract: The main objective of this paper is to find a close link between the adaptive level of significance, presented here, and the sample size. We, statisticians, know of the inconsistency, or paradox, in the current classical tests of significance that are based on *p*-value statistics that is compared to the canonical significance levels (10%, 5% and 1%): "Raise the sample to reject the null hypothesis" is the recommendation of some ill-advised scientists! This paper will show that it is possible to eliminate this problem of significance tests. The Bayesian Lindley's paradox – "increase the sample to accept the hypothesis" – also disappears. Obviously, we present here only the beginning of a possible prominent research. The intention is to extend its use to more complex applications such as survival analysis, reliability tests and other areas. The main tools used here are the Bayes Factor and the extended Neyman-Pearson Lemma.

Keywords: Significance level, sample size, Bayes ratio, likelihood function, optimal decision, significance test.

1. Introduction

Recently, the use of p-values in tests of significance has been criticized. The question posed by [1] and discussed by [2-4] concerns the misuse of canonical values of significance level (0.10, *, 0.05, **, 0.01, ***, and 0.001, ****). More recently, a publication by the American Statistical Association [5] makes recommendations for scientists to be concerned with choosing the appropriate level of significance. Pericchi and Pereira [6] considers the calculation of adaptive levels of significance in an apparently successful solution for the correction of the significance level choices. This suggestion eliminates the risk of a breach of the principle of likelihood. However, that article deals only with simple null hypotheses, although the alternative may be compounded. Another constraint was the dependence of the parametric space dimension; it was only about one-dimensional spaces. More recent is the article of [7] commented on Nature Human Behavior [8]. In a genuinely Bayesian context, [9] introduced the value *e* (*e*-value, *e* for evidence) as an alternative to the classic *p*-value. A correction to make the null hypothesis invariant under transformations was presented by [10], and a more theoretical review can be seeming in [11,12]. The *e*-value was the basis of the solution of an astrophysical problem described by [13]. The relationship between *p*-values and *e*-values is discussed by [14]. However, while the e-value works independently of dimensions, setting its significance level is not an easy task. This has made us look for a way to obtain a modified *p*-value that allows us to better understand how to obtain the optimal significance level of a problem of any finite dimension. This work is based on three of our papers [15-17]. It has taken a long time to see the possibility of using them in combination and with reasonable adjustments: Bayes Factor takes the place of the Likelihood Ratio and the average value of the Likelihood function replaces its maximum value. The mean of the likelihood function under the null hypothesis will be the density used in the calculation of the new value p, the P-value. The basis of all our work is the extended Neyman-Pearson lemma in its Bayesian form, see [18] sections Optimal Tests (Theorem 1) and Bayes Test Procedures (pp. 451-452).

This paper will show that it is possible to eliminate problems with the current significance tests. Lindley's paradox [19] – "increase the sample to accept the hypothesis" – also disappears.

2. Blending Bayesian and classical concepts

2.1. Statistical model

As usual, let x and θ be random vectors (could be scalars) such that: $x \in \boldsymbol{X} \subset \Re^n$, $\boldsymbol{X}$ being the sample space; and $\theta \in \Theta \subset \Re^n$, Θ being the parametric space. To state the relation between the two random vectors, the statistician considers the following: a family of probability density functions indexed by the conditioning parameter θ, $\{f(x|\theta); \theta \in \Theta\}$; a prior probability density function $g(\theta)$; and the posterior density function $g(\theta|x)$. In order to be appropriate, indexed by x, the family of likelihood functions $\{f(x|\theta); x \in \boldsymbol{X}\}$ must be measurable in the prior σ-algebra.

With the defined statistical model, a partition of the parametric space is defined by the consideration of a null hypothesis that should be confronted with its alternative**:**

$$\mathbf{H}: \theta \in \Theta_{\mathbf{H}} \text{ and } \mathbf{A}: \theta \in \Theta_{\mathbf{A}} \text{ where } \Theta_{\mathbf{H}} \cup \Theta_{\mathbf{A}} = \Theta \text{ with } \Theta_{\mathbf{H}} \cap \Theta_{\mathbf{A}} = \emptyset. \tag{1}$$

In the case of composite hypotheses with the partition elements having the same dimension, the model would be complete. These cases would not be involved with partitions for which there are components with zero Lebesgue measure. In case of precise hypotheses - the partition components have different dimensions - we must add other elements:

i. Positive probabilities of the hypotheses, $\pi(\mathbf{H}) > 0$ e $\pi(\mathbf{A}) = 1 - \pi(\mathbf{H}) > 0$; and
ii. A density on the subset that has the smaller dimension. The choice of this density should be coherent with the original prior density over the global parameter space.

Consider the common case for which the null hypothesis is the one defined by the subset of smallest dimension. In this case we use the surface integral to normalize the values of the prior density in the null set so that the sum or volume of these values is equal to the unit. Figure 1 illustrates how this procedure is taken. Recall that an a priori density can be looked at as a preference system in the parametric space and the preference systems must be kept even within the null hypothesis: coherence in access to a priori distributions is crucial. Further details on this procedure can be found in [20] and [16].

2.2. Significance index

By significance index we mean a real function over the sample space that is used for decision-making with respect to accept/reject the null hypothesis, **H**. We begin this section by stating the extended Neyman-Pearson Lemma presented by De Groot [18].

Let $f_{\mathbf{H}}(x)$ and $f_{\mathbf{A}}(x)$ be probability density functions over the sample space, $\boldsymbol{X}$. The decision problem is to choose one of these densities as being the true generator of the observed data. Consider now a binary function $\delta(x)$ used to define the decision procedure. Defining a partition of the sample space as $\boldsymbol{X}_{\mathbf{H}} \cup \boldsymbol{X}_{\mathbf{A}} = \boldsymbol{X}$ with $\boldsymbol{X}_{\mathbf{H}} \cap \boldsymbol{X}_{\mathbf{A}} = \emptyset$**,** the test function is

$$\delta(x) = \begin{cases} 0, & if\ x \in \boldsymbol{X}_{\mathbf{H}} \\ 1, & if\ x \in \boldsymbol{X}_{\mathbf{A}} \end{cases}. \tag{2}$$

To define the relevance of a hypothesis in relation to its alternative, one should choose two positive real numbers, say A and B : $A > B, A = B$ and $A < B$, meaning preference for the null hypothesis, indifference, and preference for the alternative. The decision rule is reject the null hypothesis, **H,** whenever the function equal one and do not reject otherwise. The optimal test is obtained by the following theorem for which the probabilities of the two types of errors – type I and type II – are

$$\alpha(\delta) = \Pr\{rejecting\ \mathbf{H}|\mathbf{H}\ is\ true\} = \Pr\{\delta(x) = 0|f_{\mathbf{H}}\}$$

and (3)

$$\beta(\delta) = \Pr\{not\ rejecting\ \mathbf{H}|\mathbf{H}\ is\ false\} = \Pr\{\delta(x) = 1|f_{\mathbf{A}}\}.$$

Neyman-Pearson-DeGroot Theorem: Let δ^* be a test that reject **H** favoring **A** if $Af_{\mathbf{H}}(x) < Bf_{\mathbf{A}}(x)$, do not reject **H** if $Af_{\mathbf{H}}(x) > Bf_{\mathbf{A}}(x)$, and being indifferent if $Af_{\mathbf{H}}(x) = Bf_{\mathbf{A}}(x)$. Then, for any other test δ,

$$A\alpha(\delta) + B\beta(\delta) \geq A\alpha(\delta^*) + B\beta(\delta^*). \tag{4}$$

To obtain the Bayesian version of the theorem consider a loss function that is zero if the decision is correct, w_{A} (w_{H}) if the decision favors **A** (**H**) when **H** (**A**) is the true state of nature. In addition, if π is the prior probability of **H** and using δ as the test function, the risk function would be

$$\boldsymbol{r}(\delta) = w_{\mathbf{A}}\pi\alpha(\delta) + w_{\mathbf{H}}(1-\pi)\beta(\delta). \tag{5}$$

Consequently, to obtain the Bayesian version of the theorem it is enough to replacing (πw_{A}) and $(1-\pi)w_{\mathrm{H}}$ for A and B, respectively. Both the classical and the Bayesian versions of the theorem are enunciated comparing in fact the ratio $\frac{f_{\mathbf{H}}}{f_{\mathbf{A}}}$ with the constant K, for which

$$K = \frac{B}{A} = \frac{(1-\pi)w_{\mathbf{H}}}{\pi w_{\mathbf{A}}}. \tag{6}$$

Important is to remember that this general version of Neyman-Pearson's theorem, from the classical point of view, will only apply to simple versus simple hypotheses. It is not common to consider a density function under a composite hypothesis. However, it is true that some classical methods use optimization by considering the maximum of the likelihood function both under **H** and under **A**: recall that the likelihood function can be represented as $\mathfrak{I}_x = \{L(\theta|x) = f(x|\theta); \forall\theta \in \Theta\}$.

Also under the Bayesian paradigm, the likelihood function L (L for likelihood) plays an important role, as it could not be otherwise, since it is the only considered objective function that shows association between the sample x and the parameter θ. However, instead of optimization, integration is the Bayesian tool. With the *a priori* densities, the following conditional expectations are calculated:

$$f_{\mathbf{H}}(x) = E\{L(\theta|x)|x, \theta \in \Theta_{\mathbf{H}}\} \text{ and } f_{\mathbf{A}}(x) = E\{L(\theta|x)|x, \theta \in \Theta_{\mathbf{A}}\}. \tag{7}$$

These functions are the Bayesian predictive densities under the respective hypotheses. Both are probability density functions over the sample space $\boldsymbol{X}$. The ratio between the two functions is known as the Bayes factor or Bayes ratio,

$$BF(x) = \frac{f_{\mathbf{H}}(x)}{f_{\mathbf{A}}(x)}. \tag{8}$$

To define a confidence index, alternative to the usual *p*-value, it is necessary to stablish an order over the sample space. Montoya-Delgado et al [17] suggests the use of the Bayes factor values of all sample points to induce the necessary order. The steps to perform a significance test are as follows:

1. Access a prior density for the parameter of interest, $g(\theta)$;
2. Clearly define the alternative hypotheses **H** and **A**;
3. Obtain the predictive functions under the two alternative hypotheses. In the case for which the parametric subspaces defined by the hypotheses are of different dimensions, the definition of a priori density under the subset of smaller dimension, say **H**, is obtained as follows:

$$g(\theta|\mathbf{H}) = \begin{cases} 0 & if\ \theta \notin \Theta_{\mathbf{A}} \\ \dfrac{g(\theta)}{\oint_{\Theta_{\mathbf{H}}} g(y)dy} & if\ \theta \in \Theta_{\mathbf{H}} \end{cases} \tag{9}$$

The denominator is the surface integral over the subspace $\Theta_{\mathbf{H}}$. In addition to this density and only in the case distinct dimensions of $\Theta_{\mathbf{H}}$ and $\Theta_{\mathbf{A}}$, consider a positive probability π of **H** be the true hypothesis. Figure 1 well illustrates how should be the choice of $g(\theta|\mathbf{H})$.

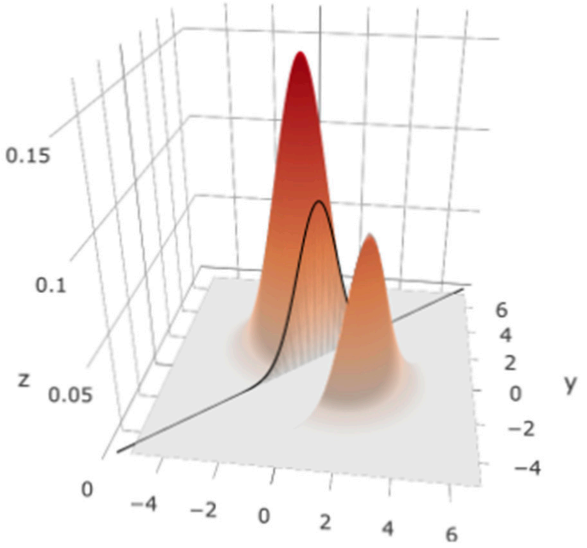

Figure 1. Bivariate Normal density cut in the subspace of equal marginal means to show the prior density in that subspace.

4. Define the loss function considering mainly the differences of importance - social, for example - between the hypotheses;
5. Use the Bayes factor to order the sample space: $\{BF(x): x \in \boldsymbol{X}\} \subset \mathfrak{R}$ establishes the order of each $x \in \boldsymbol{X}$. This ordering can be used independently of the dimensions of the spaces $\boldsymbol{X}$ and $\boldsymbol{\Theta}$.
6. Using the above theorem, compute the optimal errors and use the value of $\alpha(\delta^x)$ as the adaptive level of significance, which will depend on the loss function, the probability densities, the a priori probability π, and especially on the sample size.
7. Calculate the significance index, the *P*-value, which will take the following form: being x_0 the observed value of the statistic and $C_0 = \{x; x < x_0\}$ the observed tail, the *P*-value will take the expression $P_0 = \int_{C_0} f_{\mathbf{H}}(x)dx$. Clearly, this may be either single or a multiple integral.
8. Compare the value P_0 with the value of $\alpha(\delta^*)$. Reject (do not reject) **H** if $P_0 \underset{(>)}{<} \alpha(\delta^*)$. In the case of equality, take either decision without prejudice to optimization.
9. Finally, if $\alpha(\delta^*)$ is fixed a priori, calculate the sample size needed to make this fixed value optimal according to the Neyman-Pearson-DeGroot theorem.

3. Illustrative examples

This section introduces four simple examples to illustrate the appropriateness of the new *P*-value and how this adaptive level of significance relates with sample sizes.

3.1. Example 1 – comparing two proportions

A doctor wants to show that the incorporation of a new technology in a treatment can produces better results than a conventional one. He planned a clinical trial with two arms, case/control, each with eight patients. The cases arm used the new treatment and the arm of the controls was for the conventional one. For instance, details of an alike clinical trial are shown by [21]. The observed results were that only one of the patients in the controls arm responded positively although in the cases arm the positive respondents were four.

The most common classical significance tests result in the following *p-values*: the Pearson χ^2 *p-value* was 0.106 that changed to 0.281 when the Yates continuity correction was applied and the Fisher's exact *p-value* was 0.282. Traditional analysts would conclude that there were no statistically significant differences between the two treatments, whenever they would use anyone of the canonical significance levels. Note that these procedures were for testing a sharp hypothesis against a composed one: $\mathbf{H}: \theta_0 = \theta_1$ and $\mathbf{A}: \theta_0 \neq \theta_1$, comparing the proportion of success of the two treatments. In the sequel, we calculate the proposed $P-value$ and use the optimal significance level $\alpha(\delta^*)$ to making the decision of choosing one of the hypotheses.

To be fair in our comparisons we consider independent uniform (noninformative) prior distributions for θ_0 and θ_1. With these suppositions and the likelihoods being binomials with sample sizes $n = 8$, the predictive probability functions under the two hypotheses are

$$f_{\mathbf{H}}(x,y) = \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}} \quad \text{and} \quad f_{\mathbf{A}}(x,y) = \frac{1}{81} \;\forall\; (x,y) \in \{0,1,\dots,8\} \times \{0,1,\dots,8\}. \tag{10}$$

The variables x and y represent the possible observed values of the number of positively respondents of the two arms. Table 1 and Figure 2 present, for all possible results, the Bayes Factor values (8).

Table 1. Bayes Ratio of all possible results in a clinical trial with arms size of *n*=8.

x	*y*									Sum
	0	1	2	3	4	5	6	7	8	
0	4.765	2.382	1.112	0.476	0.183	0.061	0.017	0.003	4.E-04	9
1	2.382	2.541	1.906	1.173	***0.611***	0.267	0.093	0.024	0.003	9
2	1.112	1.906	2.052	1.710	1.166	0.653	0.290	0.093	0.017	9
3	0.476	1.173	1.710	1.866	1.633	1.161	0.653	0.267	0.061	9
4	0.183	***0.611***	1.166	1.633	1.814	1.633	1.166	0.611	0.183	9
5	0.061	0.267	0.653	1.161	1.633	1.866	1.710	1.173	0.476	9
6	0.017	0.093	0.290	0.653	1.166	1.710	2.052	1.906	1.112	9
7	0.003	0.024	0.093	0.267	0.611	1.173	1.906	2.541	2.382	9
8	4E-04	0.003	0.017	0.061	0.183	0.476	1.112	2.382	4.765	9
Sum	9	9	9	9	9	9	9	9	9	**81**

Note: Cells with red numbers form the region Ψ^* and bold-italic cells form the region Ψ_{obs}.

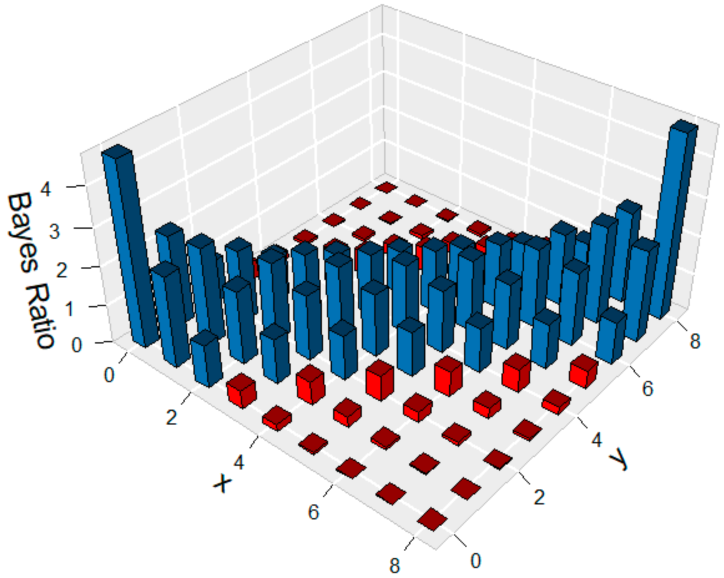

Figure 2. Bayes Factor of all possible results in a clinical trial with arms size of *n*=8 each.

To obtain the proposed $P-value$, define the set $\Psi_{obs}$of (*x*,*y*) for which its Bayes factors are smaller than the Bayes factor of the observed sample point; i.e.,

$$\Psi_{obs} = \{(x,y) \in \{0,1,\ldots,8\} \times \{0,1,\ldots,8\}: BR < BR_{obs}\}.$$

Hence, the significance index, $P-value$, is the sum of all predictive probabilities (under **H**) in Ψ_{obs}:

$$P-value = \sum_{(x,y)\in\Psi_{obs}} f_{\mathbf{H}}(x,y) = \sum_{(x,y)\in\Psi_{obs}} \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}}. \qquad (11)$$

Recalling the observed result of the clinical trial, $(x,y)=(1,4)$, the observed Bayes factor is $BR_{obs} = 0.661$. The italic-bold cells in Table 1 identify the set Ψ_{obs}. Thus, according (11), the $P-value$ is $P = 0.0923$.

To obtain the optimal solution we minimize the sum of the errors probability, $\alpha(\delta)+\beta(\delta)$. This optimal solution is the result of comparing the Bayes ratio with constant K (6) to make the choice according to the Neyman-Pearson-DeGroot theorem. Defining the set of (*x*,*y*) which Bayes Ratio is less than *K*, i.e., $\Psi^* = \{(x,y)\in\{0,1,\ldots,8\}\times\{0,1,\ldots,8\}: BR < K\}$, the optimal type I and type II errors are given by:

$$\alpha^*(\delta) = \sum_{(x,y)\in\Psi^*} f_{\mathbf{H}}(x,y) = \sum_{(x,y)\in\Psi^*} \frac{\binom{8}{x}\binom{8}{y}}{17\binom{16}{x+y}},$$

and (12)

$$\beta^*(\delta) = \sum_{(x,y)\notin\Psi^*} f_{\mathbf{A}}(x,y) = \sum_{(x,y)\notin\Psi^*} \frac{1}{81}.$$

In this example, we consider that the two hypotheses are of equal importance, $\pi = 0.5$ and $w_{\mathbf{H}} = w_{\mathbf{A}} = 1$, resulting in $K=1$. The set Ψ^* was identified by red cells in Table 1. From (12), we obtain the optimal adaptive level of significance $\alpha(\delta^*) = 0.1245$ and probability of the second kind of error $\beta(\delta^*) = 0.4815$. The high value of the probability of the second kind of error is expected whenever the sample sizes are small. Contrary to the classical results, the conclusion now is the most intuitive one; the null hypothesis is rejected since $P < \alpha(\delta^*)$.

The physician, owner of the data in Example 1, looking at our analysis, asked about the sample size needed to obtain at most 10% of a level of significance of our procedure. The answer could be obtained by the next example that shows the case of two arms with 20 patients each.

3.2. Example 2 – two proportions varying sample sizes

Consider now a Clinical Trial as in Example 1 but with arms size of n=20. Now, the observed result is $(x, y) = (4,10)$. We leave to the reader the simple exercise of repeating the calculus of Example 1 with different samples. Considering independent uniform (noninformative) prior distributions for θ_0 and θ_1 and that the two hypotheses are of equal importance, $\pi = 0.5$ and $w_{\mathbf{H}} = w_{\mathbf{A}} = 1$. The predictive probability functions under hypotheses $\boldsymbol{H}: \theta_0 = \theta_1$ and $\mathbf{A}: \theta_0 \neq \theta_1$ are

$$f_{\mathbf{H}}(x,y) = \frac{\binom{20}{x}\binom{20}{y}}{41\binom{40}{x+y}} \quad \text{and} \quad f_{\mathbf{A}}(x,y) = \frac{1}{441} \;\forall\; (x,y) \in \{0,1,\dots,20\} \times \{0,1,\dots,20\}, \tag{13}$$

and the observed Bayes Ratio is $BR_{obs} = 0.415$, which leads to the following results: significance index $P = 0.02901$; optimal adaptive level of significance $\alpha(\delta^*) = 0.0995$; and second kind of error $\beta(\delta^*) = 0.3651$. The classical χ^2 *p-value* is $p = 0.0467$ that indicates the rejection of the null hypothesis considering the canonical 5% level of significance. This agrees with our decision of also rejecting the null hypothesis since again $P < \alpha(\delta^*)$. It is interesting to see the relative distance between the index and the level of significance. For the χ^2 test we have $1-\frac{0.0467}{0.1}=0.53$ and the adaptive case obtains $1-\frac{0.029}{0.0995}=0.71$.

Figure 3 presents the optimal adaptive level of significance and the type II error according to sample size. As expected, the probabilities of both errors decrease when the sample size increases.

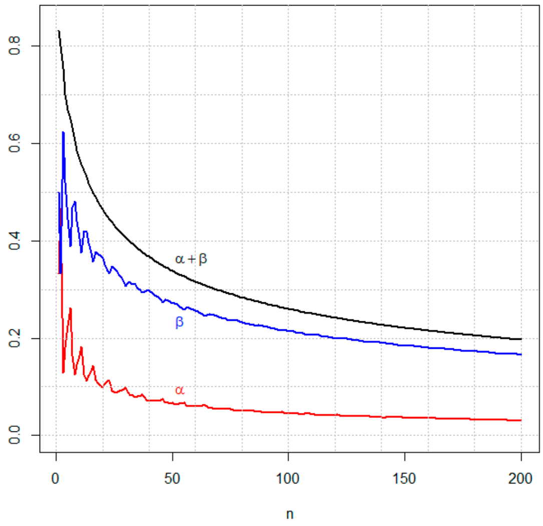

Figure 3. Probability of errors according to sample size n in each arm.

The response to the question about the sample size needed to obtain a significance level of at most 10% the answer is n = 20 in each arm. For a level of at most 5%, we need a sample size of n=90 in each arm.

We calculated the optimal adaptive level of significance and the second kind of error for different arm sizes, n_1 and n_2. The results are presented in Table 2. Once we fixed the total sample size, an unbalanced sample has larger (both type I and II) errors when compared to a balanced sample. The greater the imbalance of the sample, the greater the error. For example, the errors of an unbalanced sample with n_1= 60 and n_2 = 10 is larger than a balanced sample with n_1 = n_2 = 20 (Table 2).

Pericchi and Pereira [6] presented a closed asymptotic formula that relates sample size and level of significance in the simple case of testing $\mathbf{H}: \theta = \theta_0$ vs $\mathbf{A}: \theta \neq \theta_0$, in a binomial with parameters θ and n. The natural future project is to find this type of relation in other complex statistical problems such as the one presented in the above examples.

The following example is an attempt to show that our P-value should not violate the principle of verisimilitude. Recall that violation of this principle produced the main criticisms of the Bayesian community about classical p-values.

Table 2. Optimal levels of significance (α) and probabilities of type II error (β) for two proportions: Two independent binomial likelihoods and various sample sizes.

n_1	n_2	α	β	n_1	n_2	α	β	n_1	n_2	α	β	n_1	n_2	α	β
10	10	0.1639	0.4050	50	50	0.0667	0.2718	80	10	0.1130	0.3648	90	70	0.0529	0.2323
20	10	0.1318	0.3939	60	10	0.1097	0.3741	80	20	0.0834	0.3122	90	80	0.0493	0.2281
20	20	0.0995	0.3651	60	20	0.0860	0.3193	80	30	0.0704	0.2847	90	90	0.0468	0.2240
30	10	0.1159	0.3900	60	30	0.0765	0.2903	80	40	0.0634	0.2671	100	10	0.1111	0.3627
30	20	0.1045	0.3333	60	40	0.0689	0.2747	80	50	0.0603	0.2530	100	20	0.0818	0.3079
30	30	0.0997	0.3070	60	50	0.0626	0.2652	80	60	0.0553	0.2455	100	30	0.0684	0.2795
40	10	0.1250	0.3703	60	60	0.0591	0.2572	80	70	0.0531	0.2380	100	40	0.0617	0.2601
40	20	0.0868	0.3357	70	10	0.1130	0.3675	80	80	0.0508	0.2327	100	50	0.0559	0.2479
40	30	0.0850	0.3029	70	20	0.0865	0.3132	90	10	0.1131	0.3626	100	60	0.0538	0.2368
40	40	0.0706	0.2968	70	30	0.0727	0.2876	90	20	0.0810	0.3114	100	70	0.0512	0.2291
50	10	0.1126	0.3761	70	40	0.0645	0.2717	90	30	0.0707	0.2804	100	80	0.0483	0.2238
50	20	0.0883	0.3240	70	50	0.0603	0.2593	90	40	0.0648	0.2608	100	90	0.0467	0.2188
50	30	0.0767	0.2992	70	60	0.0575	0.2501	90	50	0.0575	0.2506	100	100	0.0449	0.2150
50	40	0.0718	0.2817	70	70	0.0539	0.2446	90	60	0.0550	0.2401				

3.3. Example 3: Test for one proportion and the likelihood principle

The main example for the violation of the likelihood principle is the case of positive binomials in comparison with negative binomials. For the same values of x, the number of successes in n independent Bernoulli trials, the two distributions produce different p-values that can lead to different decisions if compared with the same level of significance. The present example shows that the method introduced here will produce the same decisions if the observed sample size and the number of successes are the same. The reasons are that, although different, the $P-values$ are compared with different levels of significances: the decisions about the null hypothesis are going to be the same and there will be no violation of the Likelihood Principle. Changing the notation let the sample vector be composed by the number of success and the number of failures, (x, y), and the corresponding vector of probabilities be $(\theta_0.\theta_1)$ with $\theta_0 = 1-\theta_1$. Consider that $\mathbf{H}: \theta_1 = 0.5$ vs $\mathbf{A}: \theta_1 \neq 0.5$ are the hypotheses to be confronted. Considering uniform (noninformative) prior distribution for θ_1 and that the two hypotheses are of equal importance, $\pi = 0.5$ with $w_{\mathbf{H}} = w_{\mathbf{A}} = 1$, the predictive densities to build the significance tests are as follows:

1. For positive binomial

$$f_{\mathbf{H}}(x) = \binom{x+y}{x}\left(\frac{1}{2}\right)^{x+y} \quad \text{and} \quad f_{\mathbf{A}}(\mathrm{x}) = (\mathrm{x}+\mathrm{y}+1)^{-1};$$

$$(14)$$

2. For negative binomial

$$f_{\mathbf{H}}(x) = \binom{x+y-1}{x}\left(\frac{1}{2}\right)^{x+y} \quad \text{and} \quad f_{\mathbf{A}}(x) = y[(x+y)(x+y+1)]^{-1}.$$

Clearly, the Bayes factors (8) are equal for the two models and since from the theorem they will be compared with the same constant, the decisions about the null hypothesis shall be the same. On the other hand, both the p-values and the significance level are different for the two models. For instance, if we consider the observations $(x, y) = (3,10)$ and $(x, y) = (10,3)$ for positive binomial we obtain the same results for both samples; $\alpha = 0.09$, $\beta = 0.43$ and $P = 0.02$. For the negative binomial, the two observed points will produce different significance levels and both error probabilities. For the first (second) sample, one stops observing whenever the number of successes reach 3 (10). For the first result, we have $\alpha = 0.18$, $\beta = 0.48$ and $P = 0.01$, and for the second $\alpha = 0.12$, $\beta =$

0.33 and $P = 0.01$. Then, the decisions based on the positive binomials are equal to the ones based on negative binomials for the same (x, y).

Table 3 presents the predictive densities under several kinds of hypotheses for one proportion. For all kinds of hypotheses, positive and negative binomial models, for the same (x, y), produce equal Bayes factors.

Table 3. Predictive densities under several hypotheses for one proportion.

Hypotheses	**Predictive densities under H**[(1)]
H: $\theta=\theta_0$	$C(x,y)\theta_0^x(1-\theta_0)^y$
H: $\theta\neq\theta_0$	$C(x,y)\dfrac{B(U,V)}{B(u,v)}$
H: $\theta\leq\theta_0$	$C(x,y)\dfrac{B(\theta_0;U,V)}{B(\theta_0;u,v)}$
H: $\theta>\theta_0$	$C(x,y)\dfrac{B(U,V)-B(\theta_0;U,V)}{B(u,v)-B(\theta_0;u,v)}$
H: $\theta_1\leq\theta\leq\theta_2$	$C(x,y)\dfrac{B(\theta_2;U,V)-B(\theta_1;U,V)}{B(\theta_2;u,v)-B(\theta_1;u,v)}$
H: $(\theta<\theta_1)\cup(\theta>\theta_2)$	$C(x,y)\dfrac{B(U,V)-B(\theta_2;U,V)+B(\theta_1;U,V)}{B(u,v)-B(\theta_2;u,v)+B(\theta_1;u,v)}$
H: $(\theta_1\leq\theta\leq\theta_2)\cup(\theta_3\leq\theta\leq\theta_4)$	$C(x,y)\dfrac{B(\theta_2;U,V)-B(\theta_1;U,V)+B(\theta_4;U,V)-B(\theta_3;U,V)}{B(\theta_2;u,v)-B(\theta_1;u,v)+B(\theta_4;u,v)-B(\theta_3;u,v)}$
H: $(\theta<\theta_1)\cup(\theta_2<\theta<\theta_3)\cup(\theta>\theta_4)$	$C(x,y)\dfrac{B(U,V)-B(\theta_2;U,V)+B(\theta_1;U,V)-B(\theta_4;U,V)+B(\theta_3;U,V)}{B(u,v)-B(\theta_2;u,v)+B(\theta_1;u,v)-B(\theta_4;u,v)+B(\theta_3;u,v)}$

[(1)] prior distribution for θ: $\theta\sim Beta(u,v)$; $U = u + x$; $V = v + y$; $C(x,y) = \binom{x+y}{x}$ for positive binomial or $C(x,y) = \binom{x+y-1}{x}$ for negative binomial; $B(r,s) = \int_0^1 z^{r-1}(1-z)^{s-1}dz$ is the beta functions; and $B(p;r,s) = \int_0^p z^{r-1}(1-z)^{s-1}dz$ is the incomplete beta function.

3.4. Example 4

This is an example of Pereira and Wechsler [15], showing that the critical region is not always the tails of the null distribution; it can be a union of disjoint intervals.

Let x be a normal random variable with zero mean and unknown variance σ^2. The interest was to test **H**: $\sigma^2 = 2$ vs **A**: $\sigma^2 \neq 2$. A χ_1^2 (qui-squared distribution with one degree of freedom) is taken as a prior density for σ^2. After some integration exercise, we can establish the predictive densities for our significance test as

$$f_{\mathbf{A}}(x) = \{\pi(1+x^2)\}^{-1} \quad \text{and} \quad f_{\mathbf{H}}(x) = \left(2\sqrt{\pi}\right)^{-1}\exp\left(-\frac{x^2}{4}\right). \tag{15}$$

Respectively, the Cauchy density and a normal density with zero mean and variance equal two. Figure 4 shows the Bayes Ratio for all sample points that is confronted with the constant 1.1 to indicate the decision about the null hypothesis. The sample points that do not favor the null hypothesis are just a center area together with the heavy tails of the Cauchy density. The set that favors **H** does not include the central area:

$$X_H = \{x | x \in (-2.8; -0.6) \cup (0.6; 2.8)\} \tag{16}$$

The critical region under other side includes the interval $(-0.6; 0.6)$, a considerable center area.

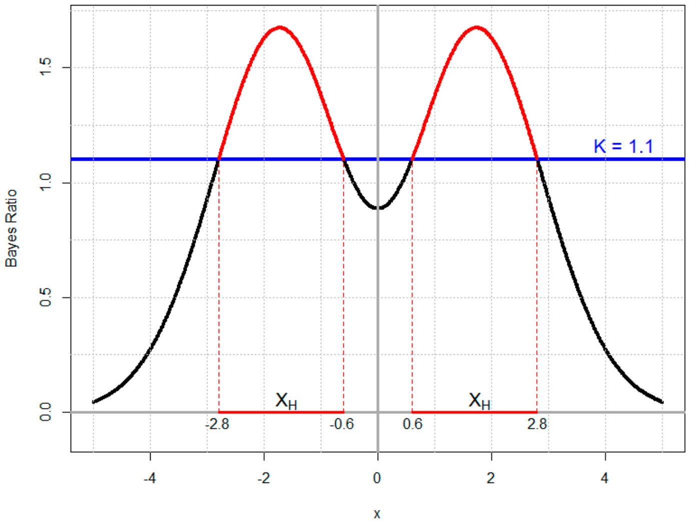

Figure 4. Bayes Ratio for N(0;2) vs Cauchy.

4. Final remarks

Most users of statistics question the logic of using the canonical significance levels for classical testing of hypotheses. We believe that there are no formal reasons for using those stablished numbers. On the other hand, here we use the natural logic of optimization for defining the adaptive significance level. We do not see any complex model that prevents the use of the significance test presented in the present paper. Although we, together with our colleagues, have already see the possibility of some additional work in testing different hypotheses, it still has a lot to do to make our $P-value$ popular. For example, considering the simple cases presented here with small changes in the hypotheses comparing two proportions can bring difficulties: $\pi \leq \theta$ against $\pi > \theta$ does give us more work than expected. Imagine now working in large sample problems in general contingency tables. It in fact remains a lot of work to be done, mainly in multivariate problems. There will have problems that give no space for improper priors to work. Hope this is the starting point of a new statistical significance testing area.

Acknowledgments: The first and second authors are grateful to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the financial support. CABP grant number 308776/2014-3; AP grant number 304025/2013-5. Our research group, GIS – group of inductive statistics, contributed to this work by discussing and given suggestions, mainly Fernando Corrêa Filho and Mark Gannon. We are very grateful for all the collaboration from these colleagues.

Author Contributions: The first author presented the problems discussed here and motivated the co-authors for the development of the work. With the third author, he defined the project of the article. The second and third author were responsible for the entire computational apparatus and the formatting of the article. The three authors wrote the article together.

Conflicts of Interest: The three authors declare no conflict of interest.

Peer-reviewed version available at *Entropy* **2017**, *19*, 696; doi:10.3390/e19120696

References

1. Johnson,V.E. Revised standards for statistical evidence, ***PNAS***, 2013, 110(48): p.19313–17.
2. Gaudart, J.; Huiartb,L.; Milligan, P.J.; Thiebautd, R.; Giorgi, R. Reproducibility issues in science, is P value really the only answer? ***PNAS***, 2014, 111(19): E1934.
3. Gelman, A.; Robert, C.P. Revised evidence for statistical standards, ***PNAS***, 2014, 111(19): E1933.
4. Pericchi, L.; Pereira, C.A.B.; Pérez, M.E. Adaptive revised evidence for statistical standards, ***PNAS***, 2014, 111(19): E1935.
5. Wasserstein, R.L; Lazar, N.A. The ASA's statement on p-values: Context, process, and purpose. ***TAS*,** 2016, 70(2), p.129–33.
6. Pericchi, L.R.; Pereira, C.A.B. Adaptive significance levels using optimal decision rules: Balancing by weighting the error probabilities. ***BJPS***, 2016, 30(1), p.70-90.
7. Benjamin, D.; Berger, J.; Johannesson, M.; et al. Redefine statistical significance. ***PsyArxiv Preprints***, 2017, Retrieved from osf.io/preprints/psyarxiv/mky9j.
8. Nature News. Big names in statistics want to shake up much-maligned P value, Available on line: https://www.nature.com/articles/d41586-017-02190-5?WT.mc_id=TWT_NatureNews&sf101140733=1, July 2017 (accessed on 28th august 2017).
9. Pereira, C.A.B.; Stern, J.M. Evidence and credibility: a full Bayesian test of precise hypothesis. ***Entropy***, 1999, 1, p.104–15.
10. Madruga, M.R.; Pereira, C.A.B.; Stern, J.M. Bayesian evidence test for precise hypotheses. ***J Statistical Planning & Inference***, 2002, 117, p.185-98.
11. Pereira, C.A.B.; Stern, J.M.; Wechsler, S. Can a significance test be genuinely Bayesian? ***Bayesian Analysis***, 2008**,** 3(1), p.79-100.
12. Stern, J.M.; Pereira, C.A.B. Bayesian epistemic values: focus on surprise, measure probability! Logic Journal of the IGPL, 2013, 22, p.236-54.
13. Chakrabarty, D. A New Bayesian Test to Test for the Intractability-Countering Hypothesis. ***JASA***, 2017, 112(518), p. 561-77.
14. Diniz, M.A.; Pereira, C.A.B.; Polpo, A.; Stern, J.M.; Wechsler, S. Relationship between Bayesian and frequentist significance indices. ***Int. J for Uncertainty Quantification***, 2012, 2(2), p.161–72.
15. Pereira, C.A.B.; Wechsler, S. On the concept of p-value. ***BJPS***, 1993, 7, p.159–77.
16. Irony, T.Z.; Pereira, C.A.B. Bayesian hypothesis test: using surface integrals to distribute prior information among the hypotheses, ***Resenhas***, 1995, 2(1), p.27-46
17. Montoya-Delgado, L.E.; Irony, T.Z.; Pereira, C.A.B.; Whittle, M.R. An unconditional exact test for the Hardy-Weimberg equilibrium law: Sample space ordering using the Bayes factor. ***Genetics***, 2001, 158(2), p.875–83.
18. DeGroot, M.H. ***Probability and Statistics***, Addison-Wesley, 1986.
19. Lindley, D.V. A Statistical Paradox", ***Biometrika***, 1957, 44 (1–2), p.187–92.
20. Pereira, C.A.B. ***Testing hypotheses of different dimensions: Bayesian view and classical interpretation*** (in Portuguese). Professor thesis, Inst Math & Statistics, USP, 1985.
21. Lopes, A.C.; Greenberg, B.D.; Canteras, M.M.; Batistuzzo, M.C.; Hoexter, M.Q.; Gentil, A.F.; Pereira, C.A.B.; Joaquim, M.A.; de Mathis, M.E.; D'Alcante, C.C.; Taub, A.; de Castro, D.G.; Tokeshi, L.; Sampaio, L.A.; Leite, C.C.; Shavitt, R.G.; Diniz, J.B.; Busatto, G.; Norén, G.; Rasmussen, S.A.; Miguel, E.F. Gamma Ventral Capsulotomy for Obsessive-Compulsive Disorder: A Randomized Clinical Trial. ***JAMA Psychiatry***, 2014, 71(9), p.1066-76.