Changing Statistical Significance with the Amount of Information: The Adaptive α Significance Level ☆

María-Eglée Pérez, Luis Raúl Pericchi

Department of Mathematics, University of Puerto Rico, Rio Piedras Campus, Box 70377, San Juan, PR 00936-8377, Puerto Rico

Abstract

We put forward an "adaptive α" which changes with the amount of sample information. Equivalently, we put forward a calibration for p-values which may be interpreted as a Bayes/non-Bayes compromise, and which leads to statistical consistency. This adaptive α can be used to test hypothesis trough probability intervals, with a potential radical simplification of Bayesian Testing, using computer packages that only produce intervals. Our procedures adapt the size of the intervals taking into consideration the amount of observed information.

Keywords: Significance Principle, Posterior Probability Principle, Bayes-non Bayes compromise, p-value calibration, adaptive confidence level

1. Motivation

The amount of information varies wildly from problem to problem. Still the most traditional statistics is anchored in fixed significance levels, and constant tables of evidence to judge p-values.

Significance Principle: *The observed significance (tail probabilities) against a hypothesis has comparable interpretation for different sample sizes and amounts of (Fisher) information.*

There is no clear justification to the use of fixed significance, except by tradition. On the contrary, there is a vast literature implicitly critical of it, most of it publish outside statistical journals, see for example Siegfried (2010) and Ionnaidis (2005) and references within.

An alternative principle is used in Bayesian analysis.

Posterior Probability Principle: *The probability of a hypothesis and a model has comparable interpretations for different sample sizes and amounts of (Fisher) information.*

☆This work was sponsored in part by NIH Grant: P20-RR016470. M.E. Pérez research was also sponsored by NSF Grant: HRD 0734826.

Email addresses: maria.perez34@upr.edu (María-Eglée Pérez), luis.pericchi@upr.edu (Luis Raúl Pericchi)

We argue that this principle is far more sensible than *Significance Principle*, which assumes significance levels (and p-values) based on wildly different information. Insistence in the Significance Principle has resulted in the strong difference among statistical and practical significance, and makes hypothesis testing inconsistent by design (since no matter how informative our experiment is, the type I error does not go to zero). In this article, we put forward rules for adapting the significance to sample sizes. The calibration of the adaptive significance levels (in short, adaptive α's) proposed is consistent with posterior probabilities of hypothesis, and it is achieved in two ways: using the concept of a "canonical experiment" , or alternatively by direct approximation of Bayes factors.

Basic derivation of the adjustment is presented in section 2, section 3 focuses on the special case of testing the mean of a normal distribution, calibration strategies are discussed in section 4 and an example is presented in section 5. In section 6 an adjustment for confidence intervals is proposed. Finally, in section 7 some further issues are addressed.

2. Basic Derivation

2.1. Basic Derivation of Formulae

Lets start by assuming we are comparing the models:

$$M_i : \mathbf{X} \text{ has density } f_i(\mathbf{x}|\theta_i),\ i = 0, 1$$

where $\theta_i, i = 0, 1$, are the unknown parameters with dimensions $q_i, i = 0, 1$.

We cannot compare the two likelihoods directly, since models depend on unknown parameters. The natural probabilistic idea is then to *integrate out* the unknown parameters,

$$B_{01}(\mathbf{x}) = \frac{\int f_0(\mathbf{x}|\theta_0)\pi(\theta_0)d\theta_0}{\int f_1(\mathbf{x}|\theta_1)\pi(\theta_1)d\theta_1},$$

under the conditional priors for each model $\pi(\theta_i), i = 0, 1$, forming in this way the Bayes factor of M_0 vs. M_1. Posterior model probabilities are then readily obtained, using Bayes Theorem, as

$$P(M_0|\mathbf{x}) = [1 + B_{10}(\mathbf{x}) \cdot B_{10}]^{-1}, \tag{1}$$

where $B_{10} = P(M_1)/P(M_0)$, the ratio of prior model probabilities.

The methods proposed in this paper are based on these ideas, but they do not require the assessment of prior distributions by the user. Instead we will use well established statistical practices to directly construct summaries of evidence.

In implementing the *Posterior Probability Principle* the following facts will be used repeatedly:

1. *Approximation to Bayes factors under regularity conditions:* Laplace's asymptotic method, under regularity conditions, gives the following neat approximation (see for example Berger and Pericchi 2001):

$$B_{0,1}^{L} = \frac{f_0(\mathbf{x}|\hat{\theta}_0)|\hat{I}_0|^{-1/2}}{f_1(\mathbf{x}|\hat{\theta}_1)|\hat{I}_1|^{-1/2}} \cdot \frac{(2\pi)^{q_0/2}\pi_0(\hat{\theta}_0)}{(2\pi)^{q_1/2}\pi_1(\hat{\theta}_1)}, \tag{2}$$

where $\hat{\theta}_i$ and $\hat{I}_i$ are the MLE and the observed information matrix respectively under M_i with q_i adjustable parameters. Notice that the first factor typically goes to ∞ or to 0 as the sample size accumulates, but the second factor stays bounded; thus, it is useful to rewrite it as:

$$-2\log(B_{01}) = -2\log\left(\frac{f_0(y|\hat{\theta}_0)}{f_1(x|\hat{\theta}_1)}\right) - 2\log\left(\frac{|\hat{I}_1|^{1/2}}{|\hat{I}_0|^{1/2}}\right) + C \tag{3}$$

2. Assume H_0 nested in H_1. *Under* H_0, the distribution of minus twice the log-likelihood ratio is asymptotically central chi-squared, that is

$$-2\log\left(\frac{f_0(\mathbf{x}|\hat{\theta}_0)}{f_1(\mathbf{x}|\hat{\theta}_1)}\right) \sim \chi^2(q) \tag{4}$$

where $q = q_1 - q_0$ and the hypothesis H_j has q_j adjustable parameters, $j = 0, 1$, see for example Casella and Berger (2002), Theorem 10.3.3.

3. The log-Bayes factor (2) can be approximated as:

$$-2log(B_{01}) = -2\log\left(\frac{f_0(\mathbf{x}|\hat{\theta}_0)}{f_1(\mathbf{x}|\hat{\theta}_1)}\right) - qlog(n^*) + C^*, \tag{5}$$

see Schwarz (1978), Kass and Wasserman (1995), Pauler (1998), Berger and Pericchi (2001). Here n^* is the *effective sample size* (Pauler 1998; Berger, Bayarri and Pericchi 2012). This effective sample size is not, in general, equal to the sample size, a point often overlooked when the Schwarz's approximation is used. The constant C^* depends on the prior assumptions and does not goes to zero as the effective sample size grows.

4. ***Key Point:*** *Condition for stabilization of the Bayes factor*
Using (4) and (5), instead of letting the quantile fixed (as under the significance principle) we re-define significance as a quantity depending on n^* according to the following rule:

$$\chi^2_{\alpha_{n^*}}(q) = \chi^2_{\alpha}(q) + q\log(n^*). \tag{6}$$

Then the Bayes factor will converge to a constant (and $\chi^2_{\alpha_{n^*}}(q)$ replaces the fixed quantile). We are ready to get the converse of equation (6) to get a calibration of p-values.

5. *Asymptotic expansion for large* $\chi^2(q)$ for the upper tail:

$$1 - F(\chi^2(q)) = 1 - Pr(\chi^2(q)) \sim \frac{(\chi^2(q))^{\frac{q}{2}-1} \exp(-\frac{\chi^2(q)}{2})}{2^{\frac{q}{2}-1}\Gamma(\frac{q}{2})}, \tag{7}$$

see for example equation 26.4.12 in Abramowitz and Stegun (1972).

Now we equate the adaptive levels: $\alpha(n, q)$ to the approximate upper tail:

$$\alpha \approx \frac{(\chi^2_\alpha)^{\frac{q}{2}-1} \exp(-\frac{\chi^2_\alpha}{2})}{2^{\frac{q}{2}-1}\Gamma(\frac{q}{2})}, \tag{8}$$

Key Point 2: If we replace the fixed quantile $\chi^2_\alpha(q)$ by $\chi^2_{\alpha_{n^*}}(q)$ as in (6), the following result is obtained:

$$\alpha_{n^*}(q) = \frac{[\chi^2_\alpha(q) + q\log(n^*)]^{\frac{q}{2}-1}}{2^{\frac{q}{2}-1} n^{*\frac{q}{2}} \Gamma(\frac{q}{2})} \times C_\alpha \tag{9}$$

which is the simple (approximate) calibration we have been looking for, and which defines the adaptive α_{n^*} levels and also the corresponding adaptive quantiles, suitable for testing via adaptive intervals for any q. Later we will give guidelines for determining C_α

2.2. The Square Root n times log(n) formula

A very important case arises when $q = 1$. For this situation, (9) simplifies to:

$$\alpha(n) = \frac{C_\alpha}{\sqrt{n^* \times (\log(n^*) + \chi^2_\alpha(1))}}, \tag{10}$$

We will call this expression the **square root** $n^* \times \log(n^*)$ **formula**. This elegant formula has antecedents, some of which we will revisit in next section.

3. The Normal mean example

In this section, the problem of comparing models $M_1 : X \sim N(0, \sigma^2)$ and $M_2 : X \sim N(\mu, \sigma^2)$ where σ^2 is known will be considered.

3.1. Antecedents

Some prior attempts of adjusting α levels (or p-values) for this problem can be found in literature. Two of them are mentioned below:

- In Good (1992) the author presents the $\sqrt{N}$ **rule**

  $$P_{stan} = min[0.5, Pval \times \sqrt{N}/10],$$

  This expression is calibrated for $N = 100$, that is, it is based on the assumption that the p-value is appropriate for $N = 100$. Note that the factor $\sqrt{log(N) + \chi^2_\alpha(1)}$ in (10) is neglected.

- In Cox and Hinkley (1979), page 397, the following expression is proposed for α's that would make hypothesis testing approximately equivalent to Bayes factors:
$$\alpha_n = const_{CH} \cdot n^{-1/2} \cdot \log(n),$$
notice the typo on this important expression, comparing it with (10).

3.2. Direct derivation for the Normal Case:

Although equation (4) is quite general, for normal samples and $q = 1$ we may have a more direct derivation.

Taking square root in equation (6), $\sqrt{k^2_{\alpha/2}} \sim \sqrt{log(n^*) + \chi^2_\alpha(1)}$, and using Mill's Ratio formula

$$\alpha/2 = 1 - \Phi(\sqrt{log(n^*) + \chi^2_\alpha(1)}) \sim \frac{\phi(\sqrt{\log(n^*) + \chi^2_\alpha(1)})}{\sqrt{\log(n^*) + \chi^2_\alpha(1)}},$$

we find again the square root $n^* \times \log(n^*)$ formula,

$$\alpha^*_n = cons_{PP} \cdot \sqrt{2/\pi} \cdot \frac{1}{\sqrt{n \times (\log(n) + \chi^2_\alpha(1))}}$$

4. Strategies to find the calibration constant

We now introduce two strategies for choosing the constant C_α. These strategies are simple enough for fast implementation in practice, even though there could be other strategies providing better approximations.

1. **The strategy of a reference experiment:** Following the form of presentation of I.J. Good, the calibration will assume that we believe α level is adequate for a specific sample size. The adaptive α is defined to be (**square root** $n \times \log(n)$ **formula**),
$$\alpha_{ref}(n) = \frac{\alpha * \sqrt{n_0 \times (\log(n_0) + \chi^2_\alpha(1))}}{\sqrt{n^* \times (\log(n^*) + \chi^2_\alpha(1))}}. \quad (11)$$
Here n_0 is the sample size of a ***reference experiment*** resulting from a experimental design for which the experimenter has specified a Type I error α and a Type II error β for a specific point of statistical importance. For n_0, $\alpha(n) = \alpha$, but decreases (increases) as the sample size grows (decreases).

2. **The strategy of a simple approximation:** The simplest approximation in (2), which is implicit in the BIC approximation, comes from assuming priors $\pi(\theta_i)$ to be $N(\theta_i|\hat{\theta}_i, I_i^{*(-1)})$, where $I_i^{*(-1)} = \hat{I}_i/n^*$, with $\hat{I}_i$ being the observed Fisher Information matrix. This leads to to a $C^*_\alpha = 1$ in (5) and then a $C_\alpha = \exp(-\chi^2_\alpha(q)/2)$ in (9).

In the following example we will see how both strategies perform.

5. Example: Extra-Sensorial Perception Experiment

The question of interest is if a "gifted" individual can change the proportion of 0's and 1's that are emitted with "perfectly" balanced proportions. So we want to test the null hypothesis $H_0 : p = 1/2$ vs the alternative $H_1 : p \neq 1/2$.

What is n_0? Assume that you design an experiment, with $\alpha = 0.01$ with type II error $\beta = 0.05$ to detect a difference of $\Delta = 0.01$. This design yields, by the usual methods, $n_0 = 44,529$.

The actual data, (Good 1992) consists in a huge sample with $n = 104,490,000$ and $S = 52,263,471$ successes, with a ratio of success of $S/n = 0.5001768$. The p-value against the null is minute, $p = 0.0003$, leading to a compelling rejection of H_0 if no adjustment in α is done.

However, adapting α according to equation (11),

$$\alpha_{ref}(n) = \frac{0.01 * \sqrt{44,529 \times (\log(44,529) + \chi^2_{0.01}(1))}}{\sqrt{104,490,000 \times (\log(104,490,000) + \chi^2_{0.01}(1))}} = 0.00017.$$

With this value of $\alpha < p$, it is not possible anymore to reject H_0. This corresponds with sensible intuition, that the data rather supports the null hypothesis, instead of giving reason to reject it.

When the simple approximation strategy is used, the adjusted value of α is $\alpha_s = 7.077512 \times 10^{-7}$. This non calibrated adjustment will produce smaller values for the size of the test.

6. Adaptive Testing Intervals

Consider again equation (6), which gives a condition on $\chi^2_{\alpha(n,q)}$ such that the Bayes factor converges to a constant C when $n \to \infty$; this, we suggest, is a reasonable condition of consistency for inference based on the Posterior Probability Principle. In such a case, confidence intervals could be defined using the adjusted quantile which can be used for testing hypothesis, by checking if the value of the parameter corresponding to the null hypothesis H_0 is inside the interval. On the other hand, the actual value of the C (which defines a level curve for the Bayes factor as a function of n) should be calibrated so that the length of the confidence interval is adjusted with n. This can be done using a similar idea to the one presented in previous section for the adjustment of the level α of the test: Assign C using a designed "canonical" experiment. Assuming $q = 1$, the adjusted quantile for $\chi^2_{\alpha(n,1)}$ is then,

$$\chi^2_{\alpha(n,1)} \approx \chi^2_\alpha + \log(\frac{n}{n_0})$$

Using the last equation, the adjustment of a normal quantile becomes

$$z_{\alpha(n,1)/2} = \sqrt{\chi^2_{\alpha(n,1)}} \approx \sqrt{\chi^2_\alpha + \log(\frac{n}{n_0})} \tag{12}$$

Note that when $n = n_0$, no adjustment is necessary.

These types of adjustments open the possibility of an spectacular simplification: Bayesian Hypothesis Testing by Intervals with adjusted sizes, that can be routinely performed, including MCMC packages, using the familiar rule: "Reject the null if it is out of the probability interval but now using (14) instead of $z_{\alpha/2}$".

Example: Extra-Sensorial Perception Experiment revisited

The 99% unadjusted confidence interval for the probability of changing a digit in the sequence, p, is $(0.5001762, 0.5001774)$, which clearly does not contain the value 0.5.

If we adjust the quantile using equation (12) and $n_0 = 44,529$ (the sample size corresponding to a designed canonical experiment, calculated in previous section), the adjusted confidence interval is $(0.4999912, 0.5003624)$. This interval now contains the value 0.5, so the null hypothesis $H_0 : p = 0.5$ is not rejected.

Note that using a level of confidence of $100 \times (1 - \alpha(n))\%$, where $\alpha(n)$ is the adjusted value of α calculated in previous section, we obtain as confidence interval $(0.4999930, 0.5003606)$, which is very close to the interval calculated using the quantile adjusted using equation (12). Using a level of confidence based on the simple approximation, the interval obtained is $(0.4999342, 0.5004194)$, which is again similar.

7. Final Comments, Questions and Some Answers

1. Is this procedure for adjusting the significance level α equivalent to the use of a specific prior for calculating the Bayes factor? The simple approximation is based on an actual Bayesian procedure based on normal priors. Nevertheless, when the reference experiment strategy is selected, there cannot be an equivalent prior to get a Bayes factor since according to the Lower Bound ($LB(pval)$), given by Sellke, Bayarri and Berger(2001),

$$BF_{01} \geq LB(pval) = -e \cdot pval \cdot log(pval), \tag{13}$$

   which is 0.125 for $\alpha = 0.01$. Our procedure directly yields an asymptotic behavior similar to a Bayes factor, but not the actual value **if** $P(H_0) = 1/2$. If the prior probability of the null is let to be smaller than a half, according to the rule: $P(H_0) = \left[1 + B_{0,1}\frac{1-\alpha}{\alpha}\right]^{-1}$, then the posterior probability of H_0 becomes very similar to the adjusted p-values.
2. Note that approximate formula (9) is valid for $q > 1$, and as such is useful for more than one parameter intervals. For example for two-dimensional intervals, $q = 2$ the approximate formula is remarkably simple, namely: $\alpha_{n^*}(q = 2) = \frac{C_\alpha}{n^*}$.
3. This procedure does not keep a constant ratio of type I error over type II error evaluated at a distinguished point θ_1. Rather, an average type 2 error on the whole of the alternative H_1 is kept in control. The situation is discussed at length in Pericchi and Pereira (2012).
4. What is the bridge between Frequentism and Bayesianism here? We are avoiding the specification of a prior, but we are introducing in the analysis

assessments that a frequentist statistician will be prepared to acknowledge through the design of a "canonical" experiment or a strategy equivalent to BIC.

5. What are the advantages of the adaptive α's: i) it yields a consistent procedure; ii) it alleviates the problem of the divergence between practical and statistical significance. iii) it makes possible to perform Bayesian testing by computing intervals with the calibrated α-levels.

References

Abramowitz, M., and Stegun, I. (1972), *Handbook of Mathematical Functions* (9th ed.),New York, Dover Publications.

Berger, J.O., and Pericchi, L.R. (2001), "Objective Bayesian Model Selection. Introduction and Comparisons", in *"Model Selection"*, ed. P. Lahiri , Lecture Notes–Monograph Series, Volume 38, Beachwood, OH: Institute of Mathematical Statistics, pp. 135-207.

Berger, J.O., Bayarri, M.J., and Pericchi, L.R. (2012), "The Effective Sample Size". *Econometrics Review* (to appear)

Casella, G., and Berger, R.L. (2002), *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Cox, D.R., and Hinkley, D.V. (1979),*Theoretical Statistics*. Boca Raton, FL: Chapman and Hall.

Good, I.J. (1992), "The Bayes/Non-Bayes Compromise. A Brief Review". *Journal of the American Statistical Association*, 87, 597–606.

Ioannidis, J. P. A. (2005), "Why most published research findings are false". *PLoS Medicine*, 2: e124. doi:10.1371/journal.pmed.0020124

Kass, R.E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypothesis and its Relationship to the Schwarz Criterion". *Journal of the American Statistical Association*, 90, 928–934.

Pauler, D.K. (1998), "The Schwarz Criterion and Related Methods for Normal Linear Models". *Biometrika*, 85, 13–27

Pericchi, L.R. and Pereira, C.A.B. (2012), "Changing the paradigm of fixed significance levels: Testing Hypothesis by Minimizing Sum of Errors Type I and Type II". (Submitted)

Schwarz, G. (1978), "Estimating the Dimension of a Model". *The Annals of Statistics*, 6, 461–464.

Sellke, T., Bayarri, M.J., and Berger, J.O. (2001), "Calibration of p values for testing precise null hypotheses". *American Statistician* 55, 62–71.

Siegfried, T. (2010), "Odds Are, It's Wrong: Science fails to face the shortcomings of statistics". *ScienceNews*, March 27th, 2010; Vol.177 #7 (p. 26)