On the Interpretation of P-Values

By:

Thomas M Sellke

Technical Report #17-01

Department of Statistics
Purdue University

July 28, 2017

On the Interpretation of P-Values

Thomas M Sellke
Department of Statistics,
Purdue University,
West Lafayette, IN 47907, USA

October 29, 2012

Abstract

Suppose that the p-value for an hypothesis test has a Uniform [0,1] distribution when the null hypothesis is true. This paper proposes a ”rough and ready” rule for the interpretation of the evidence corresponding to such p-values. The rule is to use $\bar{B}^*(p) = 1/\{epln(1/p)\}$ as an upper bound on the Bayes factor *against* the null hypothesis for $p \leq 1/e = 0.368$. The rule is found to work well for two-sided z-tests, one-sided z-tests, and two-sided t-tests with degrees of freedom at least in the teens. The rule is plausible for Chi-squared tests. If the prior distribution under the alternative is chosen so that the median p-value is between .05 and .01, then $(3/4)\bar{B}^*(p)$ is found to be a good ballpark estimate of the Bayes factor in these situations when the p-value is between .05 and .001.

Keywords: Bayes factors, z tests, t tests, Chi-squared test

1 Introduction

Consider some large family of point-null hypotheses whose plausibilities are to be assessed. The family might come from some particular area of science, and the family might be further restricted, for example to one-sided z-tests or to two-sided t-tests with degrees of freedom between 10 and 20. Suppose that, in this family, some of the null hypotheses are false and some are true, with r being the ratio of false hypotheses to true hypotheses. We assume that p-values for true null hypotheses behave like independent Uniform[0,1] random variables (meaning that, for $0 \leq a \leq b \leq 1$, the fraction of true-null p-values that fall in $[a, b]$ is $b - a$). Suppose that p-values for false null hypotheses behave collectively like independent random variables with continuous density $f(p)$ (meaning that, for $0 \leq a \leq b \leq 1$, the fraction of false-null p-values that fall in $[a, b]$ is $\int_a^b f(p)\, dp)$. Then the ratio of false to true nulls among p-values near p_0 will be approximately $rf(p_0)$, since $f(p_0) = f(p_0)/1$ is the Bayes factor against the null hypothesis when the p-value equals p_0.

One can view the hypothesis-testing scenario of the previous paragraph as a p-value sorting process. If we divide the unit interval [0,1] into many narrow sub-interval "bins", perhaps like (0.04, 0.05] and (0.009, 0.01] and (0.004, 0.005] for example, then the ratio of false to true null hypotheses within a p-value bin centered at p_0 will be approximately $rf(p_0)$ (assuming that the family of tested null hypotheses is large enough for the law of large numbers to hold for that bin, and that the average of $f(p)$ over the bin is approximately $f(p_0)$).

The discussion so far has been "frequentist", in that r and $f(p)$ are described in terms of frequencies in a large population of null hypothesis tests. The value $f(p_0)$ of course also has a Bayesian interpretation. If one's prior probabilities imply a density $f(p)$ when the null is false, then $f(p_0)$ will equal the ratio of posterior odds to prior odds (false versus true) if the only data is that the p-value is p_0. In either case, it is reasonable to regard $f(p_0)$ as a measure of the evidence against the null hypothesis corresponding to a p-value of p_0.

Given enough data from a scientific field, one could try to estimate both the false-to-true odds r and features of the false-null density $f(p)$ for that field, perhaps also conditioning on the type of test statistic. For example, Sterne and Smith (2001) claim that $r = 1/9$ "is consistent with the epidemiological literature" and estimate that the average power for false nulls is about 1/2. Such estimation of the $f(p)$ density, while very worthwhile, has to contend with difficulties like hard-to-model publication bias, changes over time, and hypotheses whose truth is never definitively resolved. However, given the key role of $f(p)$ in the "p-value sorting process" described above, a generally applicable "rule of thumb" describing the typical behavior of $f(p)$ might be useful in interpreting particular p-values and, more generally, in understanding problems involving the reproducibility of science.

This paper will examine the use of $\bar{B}^*(p) = 1/\{epln(1/p)\}$ as an upper bound on $f(p)$, following up on the considerations in Sellke, Bayarri, and Berger(2001). The formula for $\bar{B}^*(p)$ is initially motivated by plausible qualitative requirements on $f(p)$, and then the issue of whether $\bar{B}^*(p)$ is an upper bound on $f(p)$ in various standard situations is examined. Table 1 gives $\bar{B}^*(p)$ for several reference p-values.

p	.1	.05	.01	.005	.001
$1/ep\ln(1/p)$	1.60	2.46	7.99	13.89	53.25

Table 1: Values of $1/\{epln(1/p)\}$ for various reference p-values.

2 Derivation of $\bar{B}^*(p) = 1/epln(1/p)$

If the conditional probability of the null hypothesis, given the p-value, decreases as p decreases and decreases to zero (or at least to a very small value) as p decreases to zero, then $f(p)$ will be decreasing in p and will increase to infinity (or at least up to some very large value) as p decreases to zero. One would also expect $f(p)$ to be rather smooth, both because of the typical smoothness of parametric densities and also because of smoothness in the H_1 prior on background parameters. A family of densities with the right qualitative behavior is the Beta densities

$f(p|\zeta) = \zeta p^{\zeta-1}$, $0 < \zeta \leq 1$.

While the actual $f(p)$ may not closely resemble any particular member of this beta family, it is plausible that, for each fixed p in [0,1] (or at least for a range of p-values of primary interest, like $.001 \leq p \leq .05$) the value of

$f(p)$ is less than $f(p|\zeta)$ for *some* ζ.

Calculus shows that, for $p \leq e^{-1}$, the ζ which maximizes $f(p|\zeta)$ is $\zeta = 1/ln(1/p)$, and

$f(p|\zeta = 1/ln(1/p)) = \bar{B}^{*}(p) = 1/epln(1/p)$.

For $p \geq e^{-1}$, the ζ in (0,1] which maximizes $f(p|\zeta)$ is $\zeta = 1$, for which $f(p|\zeta = 1)$ is the Uniform [0,1] density.

So, the upper bound $\bar{B}^{*}(p) = 1/epln(1/p)$ for $p \leq e^{-1}$ amounts to assuming that the density $f(p)$ is bounded above for $p \leq e^{-1}$ by the "super-density" $1/epln(1/p)$, whose integral over $e^{-1} \leq p < 1$ is infinity.

Let $Y = ln(1/p)$. Then $f(p|\zeta) = \zeta p^{\zeta - 1}$ implies $P(Y \geq y) = P(p \leq e^{-y}) = e^{-\zeta y}$ for $y \geq 0$, so that Y has an Exponential distribution with hazard rate ζ when p has density $f(p|\zeta)$. The null distribution of Y is standard Exponential, with $\zeta = 1$. Let $f_1(y)$ be the density of Y corresponding to the actual H_1 density $f(p)$ for p. If $f_1(y)$ has decreasing hazard rate $h_1(y)$, then

$f_1(y) = h_1(y)exp(-\int_0^y h_1(z))\,dz) \leq h_1(y)e^{-yh_1(y)}$

and the Bayes factor against the null is

$B(y) = f_1(y)/e^{-y} \leq h_1(y)e^{y-yh_1(y)} = h_1(y)p^{h_1(y)-1} = f(p|\zeta = h_1(y)) \leq \bar{B}^{*}(p) = 1/epln(1/p)$ if $p \leq e^{-1}$.

Hence, the upper bound $\bar{B}^{*}(p) = 1/\{epln(1/p)\}$ on the Bayes factor against the null holds for $p \leq e^{-1}$ if $Y = ln(1/p)$ has decreasing hazard rate under the alternative.

The author will forgive the reader if the reader does not find the above derivation of $\bar{B}^{*}(p)$ completely convincing. Let us examine whether the bound $\bar{B}^*(p)$ holds in a variety of standard testing situations.

2.1 Two-sided z tests

Suppose that $X \sim \mathrm{N}(\theta, \sigma = 1)$ and that we test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ based on test statistic $T = |X|$. The usual p-value when $T = t$ is of course $2(1 - \Phi(t))$. Edwards, Lindman, and Savage(1963) showed that, if θ has a Normal distribution with mean 0 under H_1, then the Bayes factor $B(t)$ against the null satisfies

$B(t) \leq exp(t^2/2)/(t\sqrt{e}) = \bar{B}_{NOR}(t)$ for $t \geq 1$.

Figure 1 graphs the ratio $\bar{B}_{NOR}(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 5$. The graph suggests that $\bar{B}^*(p(t))$ is an upper bound on $\bar{B}_{NOR}(t)$ for all $t \geq 1$. In fact, it follows from the log-convexity of Mills' ratio (Theorem 2.5(a) in Baricz (2008)) that $Y = ln(1/p)$ has decreasing hazard rate when θ has a Normal prior with mean 0, which implies that $\bar{B}_{NOR}(t) \leq \bar{B}^*(p(t))$ when $t \geq 1$ for the Edwards-Lindman-Savage bound $\bar{B}_{NOR}(t)$. Using Mills' ratio to approximate the p-value shows that this ratio $\bar{B}_{NOR}(t)/\bar{B}^*(p(t))$ converges to $\sqrt{e/(2\pi)} = 0.6577$ as t goes to ∞.

$\bar{B}_{NOR}(1.96) = 2.01$, and $\bar{B}_{NOR}(2.576) = 6.50$, so these Edwards-Lindman-Savage bounds on the Bayes factor against the null are smaller than is consistent with the conventional interpretation of evidence corresponding to the p-values .05 and .01. A naive explanation of this discrepancy might be to attribute it to special features of Normal priors, such as thin tails. Berger and Sellke (1987) showed that assuming only that the alternative density is

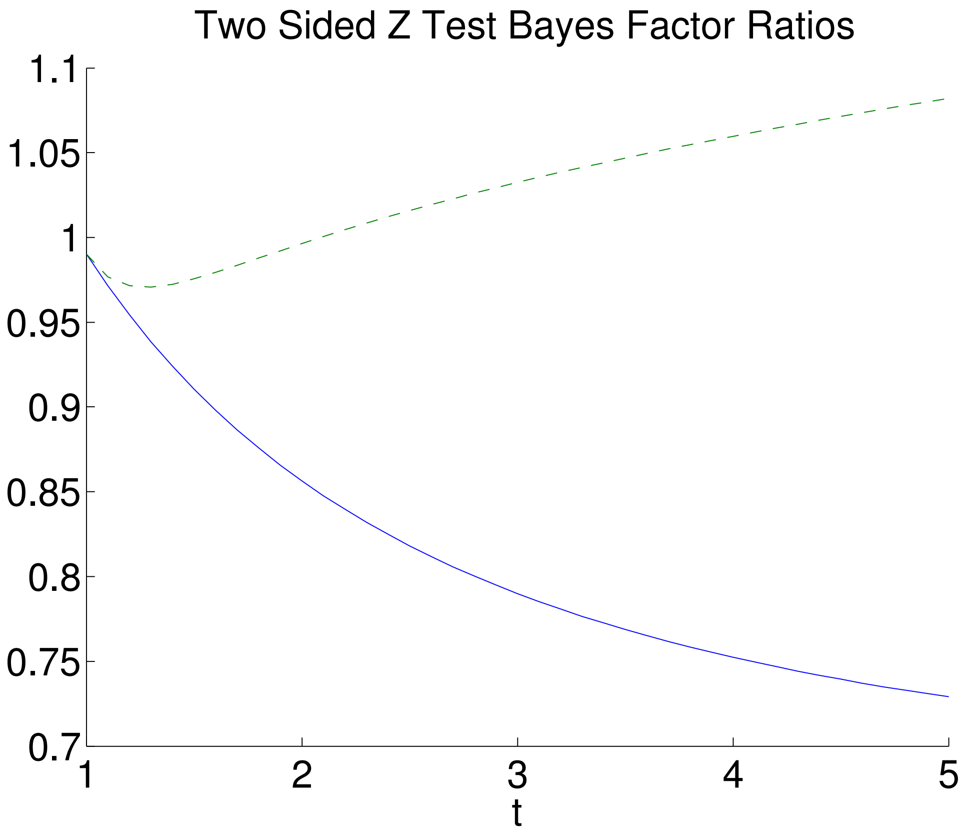

Figure 1: $\bar{B}_{NOR}(t)/\bar{B}^*(p(t))$ is solid curve. $\bar{B}_{US}(t)/\bar{B}^*(p(t))$ is dashed curve.

symmetric and unimodal about 0 gave qualitatively similar upper bounds $\bar{B}_{US}(1.96) = 2.45$ and $\bar{B}_{US}(2.576) = 8.17$. (The symmetry is in fact irrelevant if the analysis is based on T, with the sign of X ignored. Note also that the maximum Bayes factor $\bar{B}_{US}(t)$will always be attained by a symmetric Uniform distribution.) Figure 1 also shows the ratio $\bar{B}_{US}(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 5$. Using Mills' ratio to approximate the p-value shows that this ratio $\bar{B}_{US}(t)/\bar{B}^*(p(t))$ converges to $e/2 = 1.36$ as t goes to ∞.

Let $B_{NOR(2.73)}(t)$ be the Bayes factor against H_0 when θ has a $N(0, \sigma =$

2.73) distribution under H_1. Under this H_1 prior, the average power of a .05 level test is 1/2. Let $B_{NOR(3.68)}(t)$ be the Bayes factor against H_0 when θ has a $N(0, \sigma = 3.68)$ distribution under H_1. Under this H_1 prior, the average power of a .01 level test is 1/2. Figure 2 shows the ratios $B_{NOR(2.73)}(t)/\bar{B}^*(p(t))$ and $B_{NOR(3.68)}(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 5$. Figure 2 also shows the ratio $B_{\pm 4.85}(t)/\bar{B}^*(p(t))$ for the Bayes factor corresponding to the Efron-Gous(2001) Uniform$[-4.85, 4.85]$ "break-even" H_1 prior, for which the Bayes factors here correspond to the Efron-Gous interpretation of R.A.Fisher's strength-of-evidence scale for p-values. For p between .05 and .001 (meaning t between 1.96 and 3.291), $B_{NOR(2.73)}(t)$ and $B_{NOR(3.68)}(t)$ are between $(.6)\bar{B}^*(p(t))$ and $(.8)\bar{B}^*(p(t))$, while $B_{\pm 4.85}(t)$ is between $(.7)\bar{B}^*(p(t))$ and $(1.1)\bar{B}^*(p(t))$ over this range.

2.2 One-sided z tests

Suppose we test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ based on test statistic $T \sim N(\theta, 1)$. The standard p-value when $T = t$ is $1 - \Phi(t))$.

Let $\bar{B}_{ABSNOR}(t)$ be the maximum Bayes factor against H_0 when θ has the distribution of the absolute value of a mean 0 Normal random variable under H_1. Let $\bar{B}_{EXP}(t)$ be the maximum Bayes factor against H_0 when θ has an Exponential distribution under H_1. Let $\bar{B}_{UNI}(t)$ be the maximum Bayes factor against H_0 when θ has a unimodal distribution on $[0, \infty)$ with mode 0. (The maximum will always be attained by a Uniform$[0, b(t)]$ distribution.) Figure 3 plots the ratios $\bar{B}_{ABSNOR}(t)/\bar{B}^*(p(t))$,$\bar{B}_{EXP}(t)/\bar{B}^*(p(t))$,

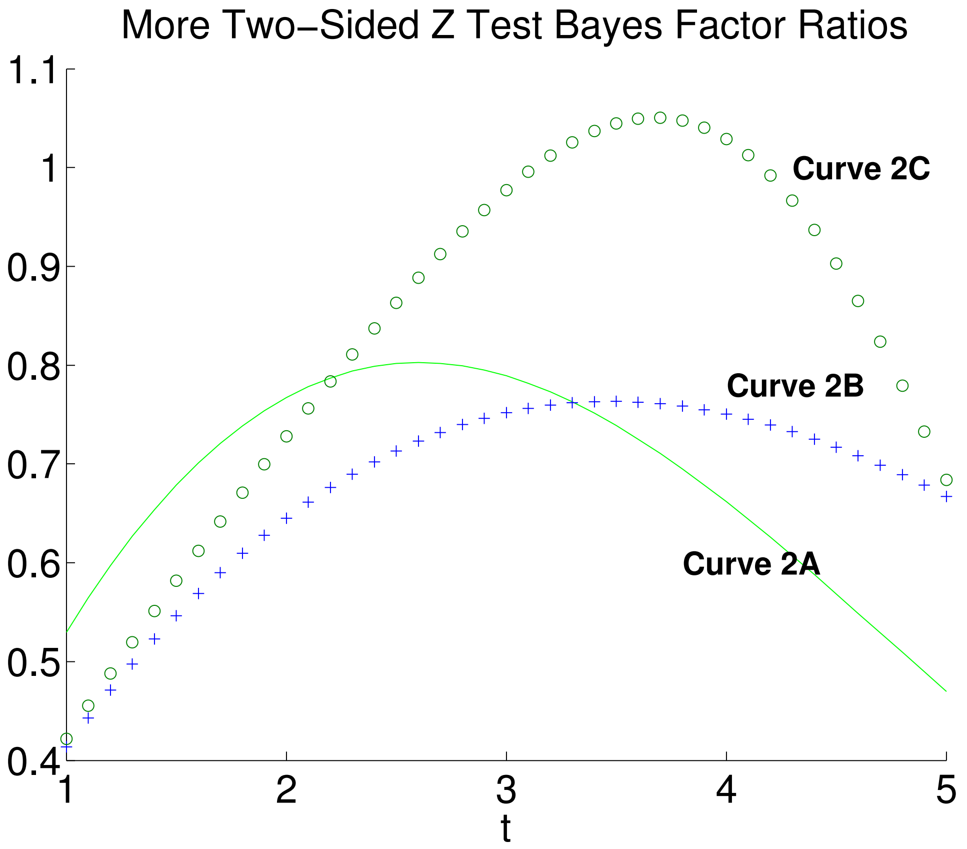

Figure 2: Curve 2A is $B_{NOR(2.73)}(t)/\bar{B}^*(p(t))$, where $B_{NOR(2.73)}(t)$ is the Bayes factor for the mean 0 normal H_1 prior with median p-value .05. Curve 2B is $B_{NOR(3.68)}(t)/\bar{B}^*(p(t))$, where $B_{NOR(3.68)}(t)$ s the Bayes factor for the mean 0 normal H_1 prior with median p-value .01. Curve 2C is $B_{\pm 4.85}(t)/B(p(t)$, where $B_{\pm 4.85}(t)$ is the Bayes factor for the U[-4.85, 4.85] "break-even" H_1 prior.

and $\bar{B}_{UNI}(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 5$. These ratios respectively converge to $\sqrt{e/(2\pi)} = 0.6577$, to 1/2, and to $e/2 = 1.36$ as t goes to ∞.

Let $B_{ABSNOR(2.27)}(t)$ be the Bayes factor when θ has the distribution of the absolute value of a $N(0, \sigma = 2.27)$ random variable under H_1. Under this H_1 prior, the average power of a .05 level test is 1/2. Let $B_{ABSNOR(3.307)}(t)$ be the Bayes factor when θ has the distribution of the absolute value of a $N(0, \sigma = 3.307)$ random variable under H_1. Under this H_1 prior, the average power of a .01 level test is 1/2. Figure 4 graphs the ratios $B_{ABSNOR(2.27)}(t)/\bar{B}^*(p(t))$ and $B_{ABSNOR(3.307)}(t)/\bar{B}^*(p(t))$, along with the previously graphed $\bar{B}_{ABSNOR}(t)/\bar{B}^*(p(t))$. For p between .05 and .001 (meaning t between 1.645 and 3.09), $B_{ABSNOR(2.27)}(t)$ and $B_{ABSNOR(3.307)}(t)$ are between $(.7)\bar{B}^*(p(t))$ and $\bar{B}^*(p(t))$.

Let $B_{EXP(.476)}(t)$ be the Bayes factor when θ has an Exponential distribution with $\lambda = .476$ under H_1. Under this H_1 prior, the average power of a .05 level test is 1/2. Let $B_{EXP(.319)}(t)$ be the Bayes factor when θ has an Exponential distribution with $\lambda = .319$ under H_1. Under this H_1 prior, the average power of a .01 level test is 1/2. Let $B_{[0,4.85]}(t)$ be the Bayes factor when θ has a Uniform[0,4.85] distribution under H_1. According to Efron and Gous(2001), the Bayes factors here for this Uniform[0,4.85] prior are in good agreement with the Efron-Gous interpretation of R.A.Fisher's strength-of-evidence scale for p-values. Figure 5 plots the ratios of these Bayes factors against $\bar{B}^*(p)$ for $1 \leq t \leq 5$, along with the previously graphed ratios $\bar{B}_{EXP}(t)/\bar{B}^*(p(t))$ and $\bar{B}_{UNI}(t)/\bar{B}^*(p(t))$. For p between .05 and .001 (meaning t between 1.645 and 3.09), $B_{EXP(.476)}(t)$ and $B_{EXP(.319)}$ are between$(.6)\bar{B}^*(p(t))$ and

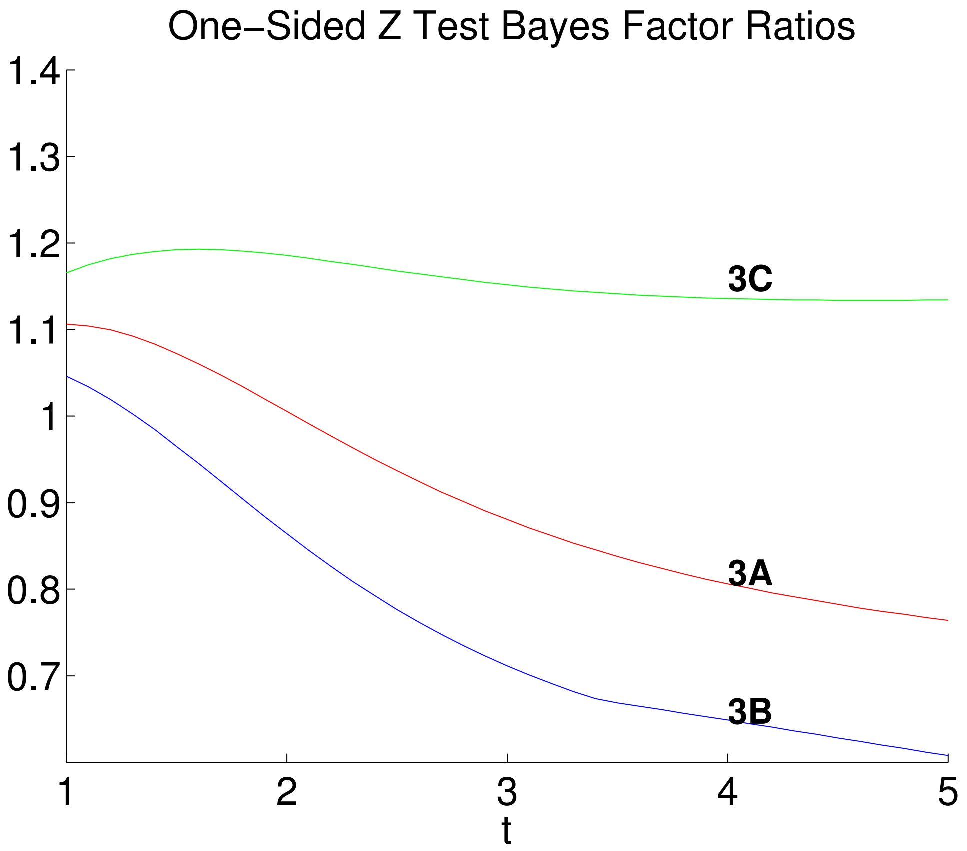

Figure 3: Curve 3A is $\bar{B}_{ABSNOR}(t)/\bar{B}^*(p(t))$, curve 3B is $\bar{B}_{EXP}(t)/\bar{B}^*(p(t))$, and curve 3C is $\bar{B}_{UNI}(t)/\bar{B}^*(p(t))$

$(.85)\bar{B}^*(p(t))$, while $\bar{B}_{UNI}(t)$ is between $(.7)\bar{B}^*(p(t))$ and $(1.15)\bar{B}^*(p(t))$ over this range.

2.3 Two-sided t tests

Suppose we see $X_1, ..., X_{k+1}$ which are iid Normal$(\theta, SD = \sigma)$ random variables. We test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ based on the t-statistic $T = \sqrt{(k+1)k}\bar{X}_{k+1}/\sqrt{(\sum(X_i - \bar{X}_{k+1})^2)}$, whose distribution under H_0 is

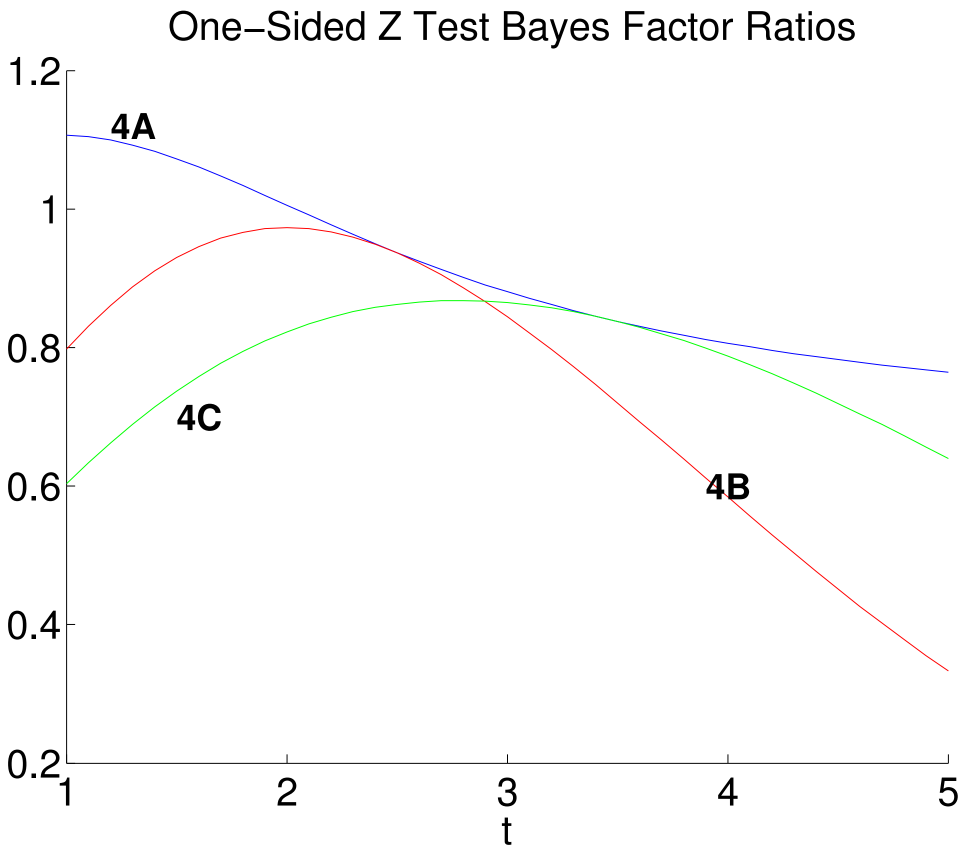

Figure 4: Curve 4A is $\bar{B}_{ABSNOR}(t)/\bar{B}^*(p(t))$, curve 4B is $B_{ABSNOR(2.27)}(t)/\bar{B}^*(p(t))$, and curve 4C is $B_{ABSNOR(3.307)}(t)/\bar{B}^*(p(t))$.

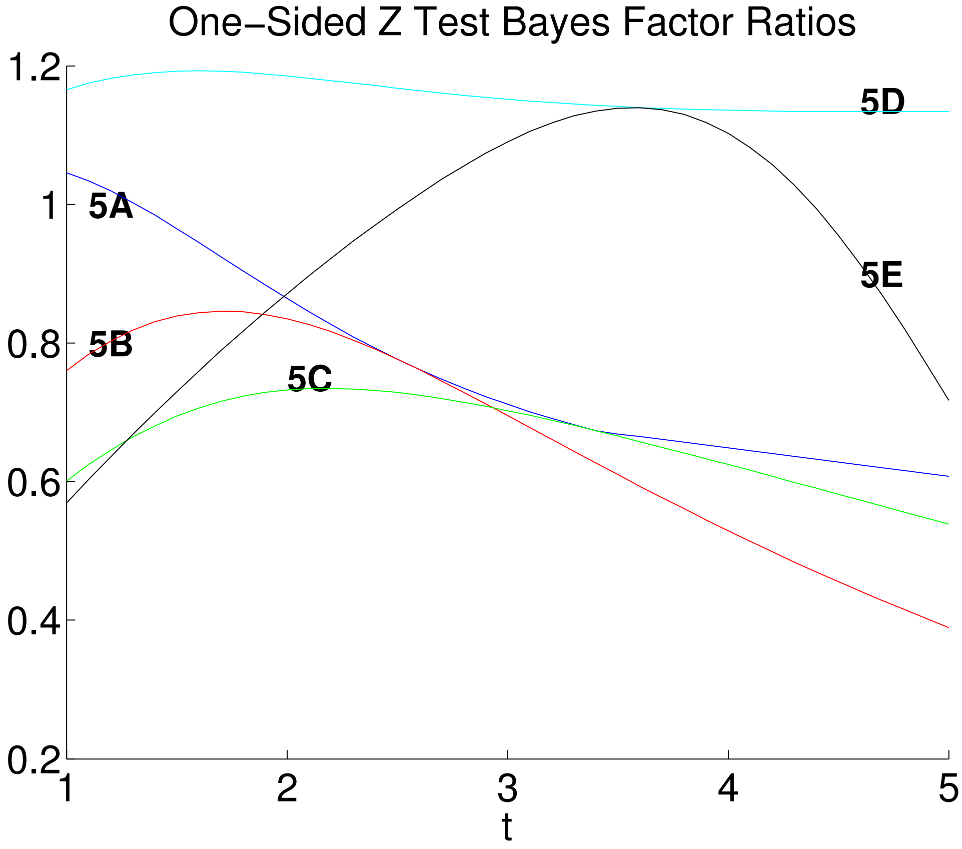

Figure 5: Curve 5A is $\bar{B}_{EXP}(t)/\bar{B}^*(p(t))$. Curve 5B is $B_{EXP(.476)}(t)/\bar{B}^*(p(t))$. Curve 5C is $B_{EXP(.319)}(t)/\bar{B}^*(p(t))$. Curve 5D is $\bar{B}_{UNI}(t)/\bar{B}^*(p(t))$. Curve 5E is $B_{[0,4.85]}(t))/\bar{B}^*(p(t))$

the t-distribution with k degrees of freedom. The behavior of T is determined by the ratio θ/σ. Suppose that θ/σ has a Normal$(0, SD = \tau)$ distribution under H_1. Then under H_1, $T/\sqrt{1 + (k+1)\tau^2}$ will have a t-distribution with k degrees of freedom. Calculus shows that the maximum (with respect to τ) Bayes factor against H_0 when $|T| = t \geq 1$ is

$\bar{B}_k(t) = [((k + t^2)/(k+1))^{(k+1)/2}]/t.$

As one would expect, $\bar{B}_k(t)$ converges to $\bar{B}_{NOR}(t)$ for fixed $t \geq 1$ as k goes to ∞, and the p-value for any $t > 0$ converges as k goes to ∞ to $2(1 - \Phi(t))$, so $\bar{B}_k(t) \leq \bar{B}^*(p(t))$ for k sufficiently large, for any fixed $t \geq 1$. Table 2 shows the minimum values of k for which we have $\bar{B}_k(t) \leq \bar{B}^*(p(t))$, for various standard p-values. For each p, the inequality $\bar{B}_k(t) \leq \bar{B}^*(p)$ holds for $k \geq$ these minimum values. For 10 degrees of freedom and $.05 \geq p \geq .001$, $\bar{B}_{10}(t)$ can exceed $\bar{B}^*(p(t))$, but not by that much. Figure 6 graphs the ratio $\bar{B}_{10}(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 6$, with points corresponding to various p-values indicated. For $k = 10$ and $.05 \geq p \geq .001$, $\bar{B}_{10}(t)$ is very slightly below $\bar{B}^*(p(t))$ near $p = .05$ and otherwise between $\bar{B}^*(p(t))$ and $1.25)\bar{B}^*(p(t))$

p	.05	.01	.005	.001
minimum k	10	14	16	19

Table 2: Minimun degrees of freedom k to guarantee $\bar{B}_k(t) \leq \bar{B}^*(p(t))$ for t test Bayes factor upper bound $\bar{B}_k(t)$, for various p-values.

When k is small and θ/σ is Normal$(0, SD = \tau)$, the maximum Bayes factor (maximizing over τ) can be much larger than $\bar{B}^*(p)$ for $.05 \leq p \leq .001$. Figure 7 graphs the ratio $\bar{B}_3(t)/\bar{B}^*(p(t))$ for $1 \leq t \leq 6$, with points

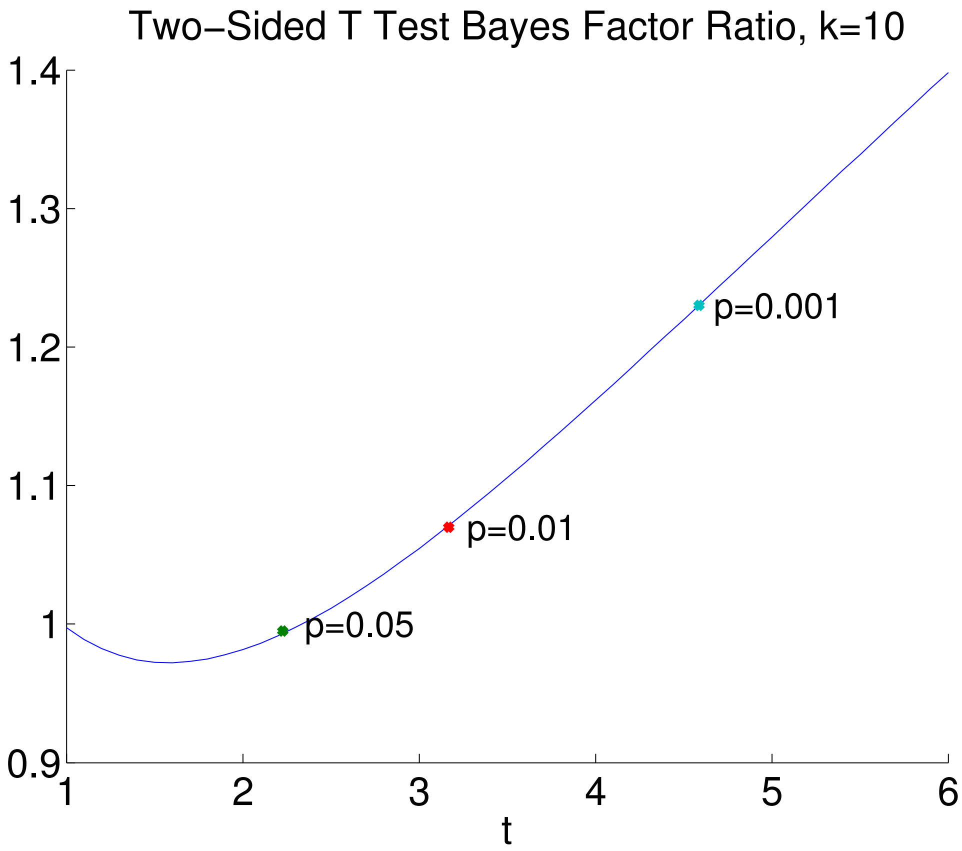

Figure 6: Graph of $\bar{B}_{10}(t)/\bar{B}^*(p(t))$. For $k = 10$ degrees of freedom and $.05 \geq p \geq .001$, $\bar{B}_{10}(t)$ is between $(.99)\bar{B}^*(p(t))$ and $(1.25)\bar{B}^*(p(t))$

corresponding to various p-values indicated.

For other types of t-tests (one-sided and/or two-sample, etc.), the situation will likewise converge to z-tests as the degrees of freedom increase to ∞. Numerical calculations like those above would show how fast this convergence occurs.

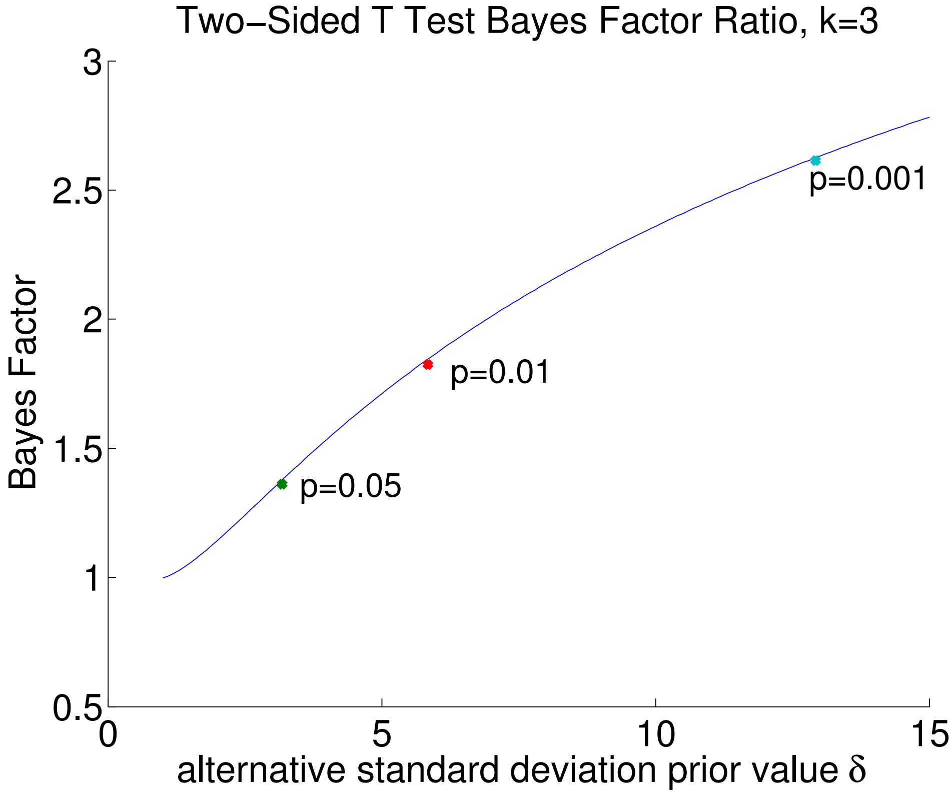

Figure 7: Graph of $\bar{B}_3(t)/\bar{B}^*(p(t))$. For $k = 3$ degrees of freedom and $p = .001$, $\bar{B}_3(t)$ is $> (2.5)\bar{B}^*(p(t))$

2.4 Chi-Squared Tests

Suppose that we see $\mathbf{X} = (X_1, ..., X_k)$ which is multivariate normal $N_k(\boldsymbol{\theta}, \mathbf{I})$ with mean vector $\boldsymbol{\theta}$ and identity covariance matrix. The usual p-value for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \neq \mathbf{0}$ when $T = \sum X_i^2 = t$ is $P(T_0 \geq t)$, where T_0 has a Chi-squared distribution with k degrees of freedom.

Suppose that $\boldsymbol{\theta}$ is distributed according to the scale mixture of multivariate normals given by

$v^2 \sim Exponential(\lambda)$

$\boldsymbol{\theta}|v^2 \sim N_k(\mathbf{0}, v^2\mathbf{I})$

Let $\bar{B}_{\chi^2_k}(t)$ be the maximum Bayes factor for such H_1 priors when $T = t$. Figures 8, 9, 10, and 11 plot the ratio $\bar{B}_{\chi^2_k}(t)/\bar{B}^*(p(t))$ versus t for $k = 1, 3, 6, 15, 30$, respectively, with points corresponding to $p = .10, .05, .01, .001$ indicated.

As k goes to infinity here, the Chi-squared distributions look more and Normal, and the distribution of T for any $\boldsymbol{\theta}$ not too far from $\mathbf{0}$ looks more and more like the null distribution shifted to the right. The situation converges to the one-sided z test situation as k goes to infinity, and the hierarchical prior on v^2 corresponds to an Exponential prior for a one-sided z test.

3 Discussion

We have seen that, in a variety of standard testing situations, the maximum Bayes factor against the null hypothesis , maximized over a large class of

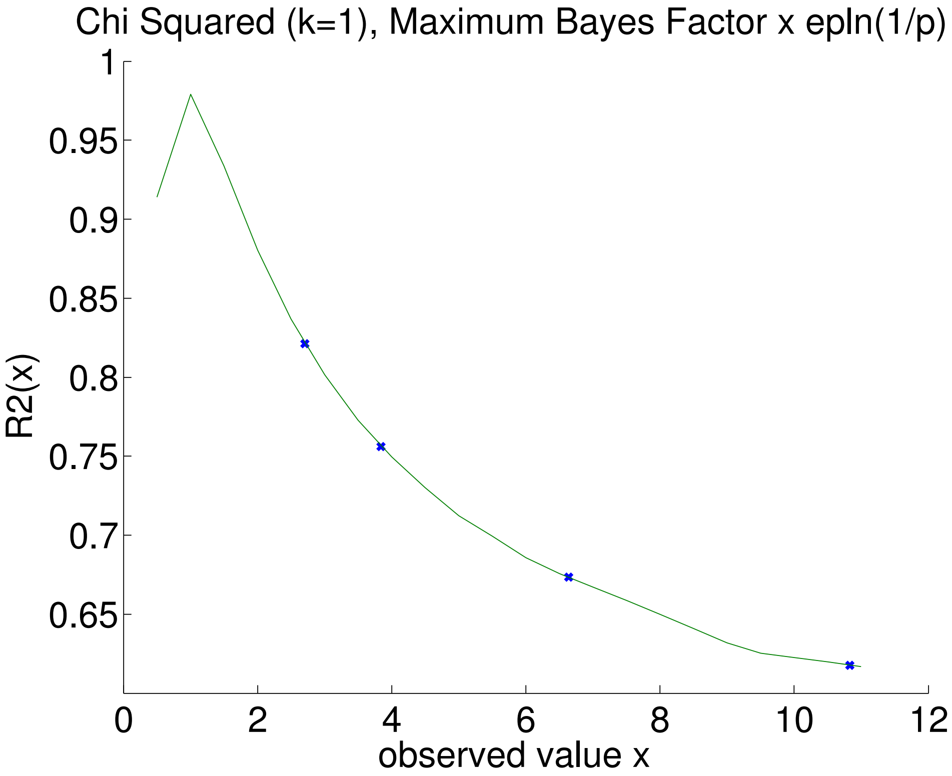

Figure 8: $\bar{B}_{\chi_1^2}(t)/\bar{B}^*(p(t))$

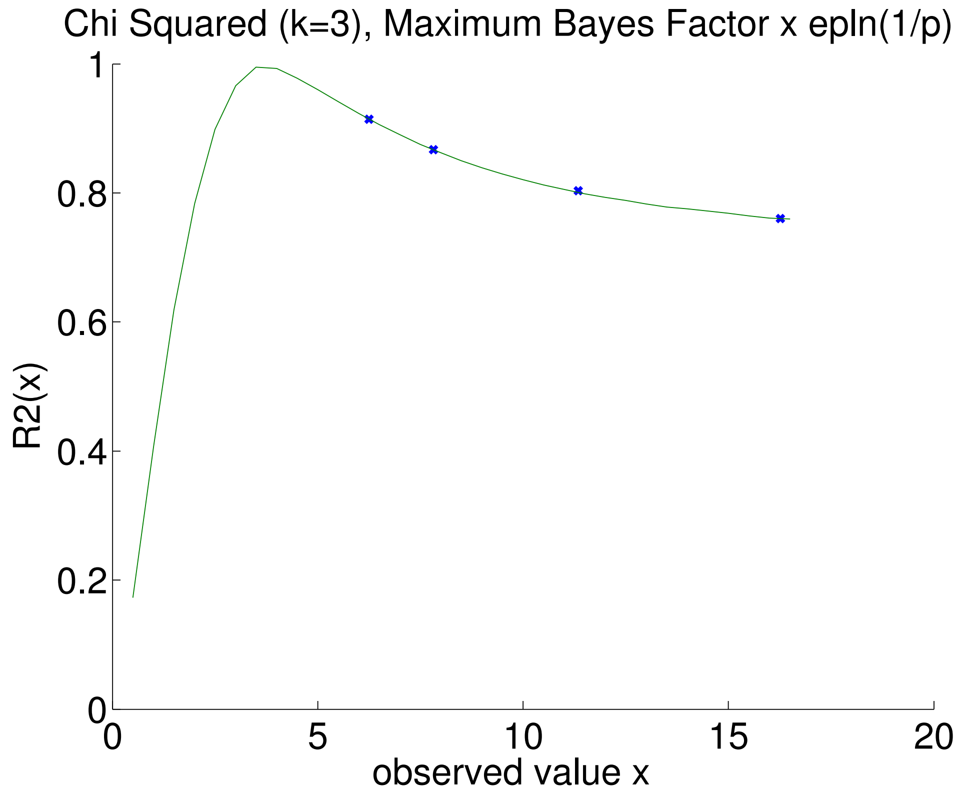

Figure 9: $\bar{B}_{\chi^2_3}(t)/\bar{B}^*(p(t))$

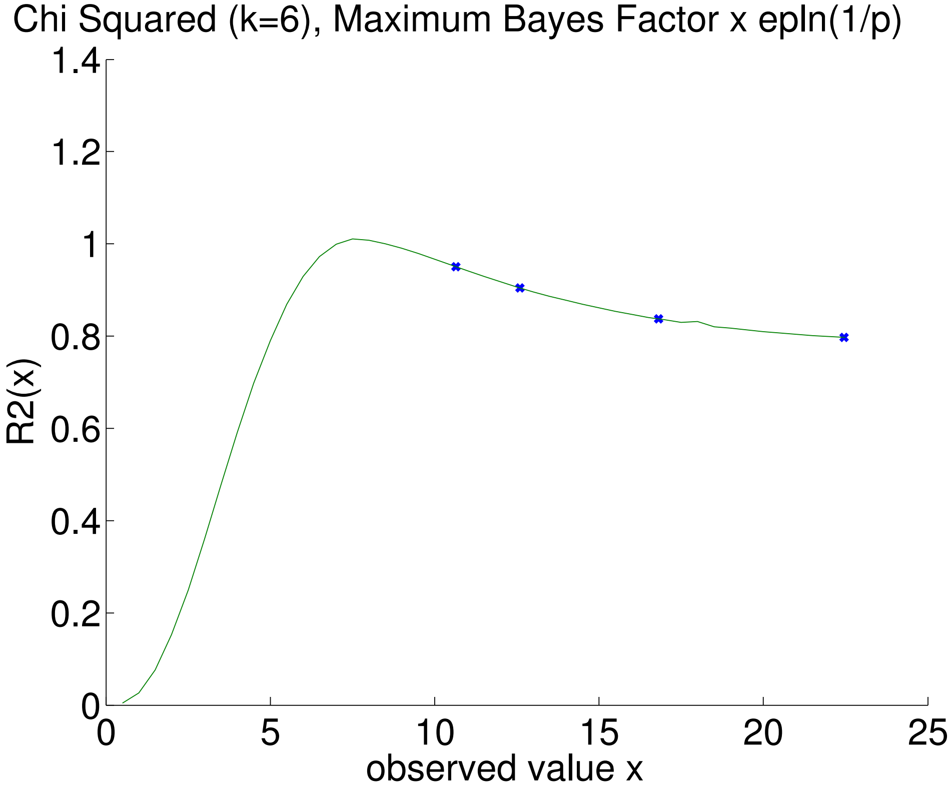

Figure 10: $\bar{B}_{\chi^2_6}(t)/\bar{B}^* p(t))$

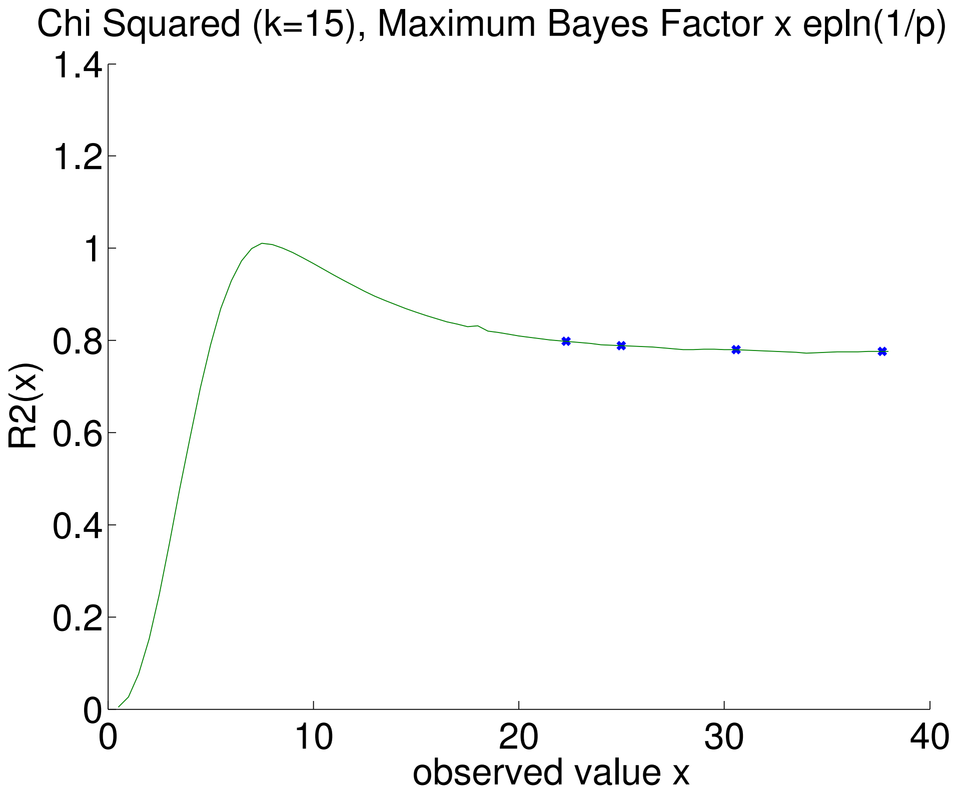

Figure 11: $\bar{B}_{\chi^2_{15}}(t)/\bar{B}^*(p(t))$

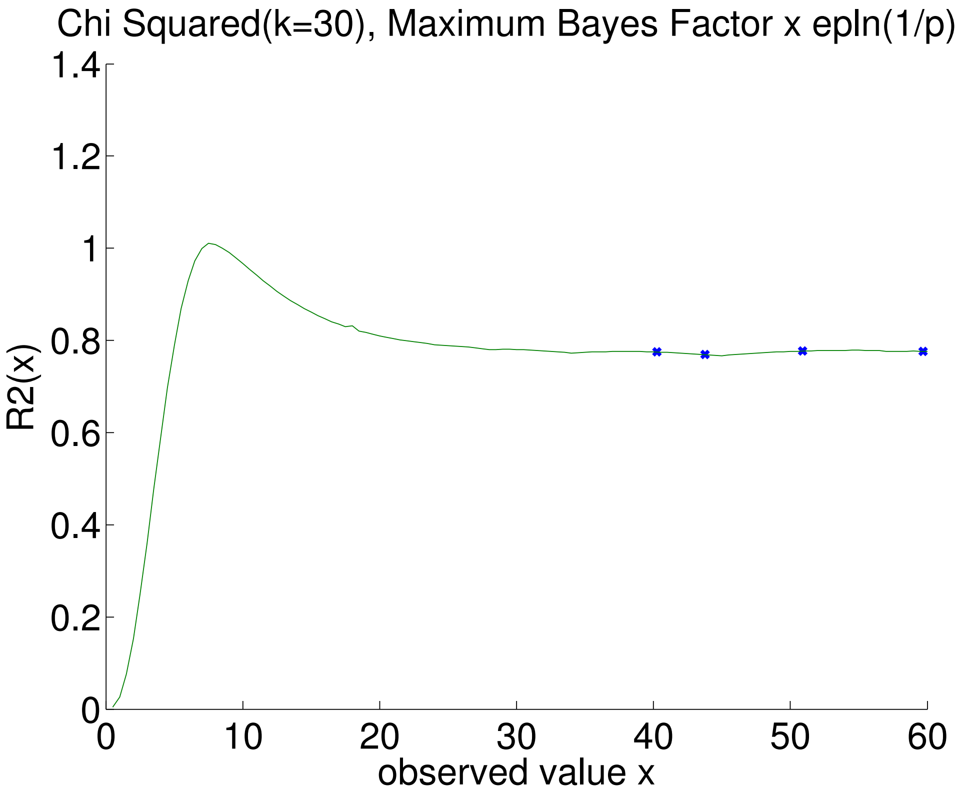

Figure 12: $\bar{B}_{\chi^2_{30}}(t)/\bar{B}^*(p(t))$

"reasonable" alternative prior distributions, is at most a bit bigger than $\bar{B}^*(p) = 1/epln(1/p)$ for $p \leq e^{-1}$. For H_1 priors for which the median p-value is between .05 and .01, we have seen that $(3/4)\bar{B}^*(p)$ is a reasonable ballpark estimate of the Bayes factor for $.05 \geq p \geq .001$. However, for two-sided t-tests with small degree of freedom, the Bayes factor against the null can be significantly larger that $\bar{B}^*(p)$ for some "reasonable" priors on the alternative. These results can be viewed as emphasizing and extending the implications of the Edwards-Lindman-Savage bound $\bar{B}_{NOR}(t)$ above.

Numerous previous authors, in addition to those mentioned above, have considered the calibration of the evidence corresponding to p-values. Bayesian p-value calibrations, as in Good (1982), and objective Bayesian model-selection procedures like the BIC (Bayesian Information Criterion) of Schwarz (1978) typically depend on sample size. However, as discussed in Efron and Gous (2001) and in references given there, the value of the sample size is sometimes ambiguous. The approximate upper bound $\bar{B}^*(p) = 1/epln(1/p)$ here obviously does not depend on sample size.

As can be seen from Curve 2C in Figure 2 (and as has been pointed out personally to the author by Brad Efron), $\bar{B}^*(p)$ is not so different from the Efron-Gous interpretation of Fisher's scale of evidence in terms of Bayes factors. However, the common opinion concerning the strength of evidence against the null inherent in p-values of .05 and .01, to the extent that such common opinion can be expressed numerically as a Bayes factor, seems to be that the evidence is stronger than would correspond to a Bayes factor of about 2 (for p=.05) and about 6 or 8 (for p=.01). And indeed, Efron and Gous (2001) write that the large-n Jeffreys interpretation of p=.01 as “barely worth mentioning” would be “a shocking assertion to the medical scientist for whom a .01 significance level settles the issue”. A Bayes factor of 6 or 8 is certainly worth mentioning, but it does not settle the issue, either.

It should be noted that a formula equivalent to $\bar{B}^*(p) = 1/epln(1/p)$ appears already in Vovk(1993).

The $N(0, SD = \tau)$ alternative priors on θ/σ considered for two-sided t-tests have the very unrealistic feature that the conditional variance of θ, given σ, is a constant multiple of σ^2. It might be more reasonable (whether in the frequentist scenario or in the Bayesian scenario) to assume that the distribution of θ/σ is some mixture of mean 0 Normal distributions. If, for example, the conditional distribution of θ, given σ, were assumed $N(0, SD = \tau(\sigma))$ for some function $\tau(\sigma)$, (with $\tau(\sigma)$ perhaps even constant, making θ and σ independent), then the marginal distribution of θ/σ would be a mixture of mean 0 Normals. If the Bayes factor for a certain p-value p_0 is less than $\bar{B}^*(p_0)$ for *any* $N(0, SD = \tau)$ prior on θ/σ when the degrees of freedom are at least $k(p_0)$, then $\geq k(p_0)$ degrees of freedom would guarantee that the Bayes factor for that p-value p_0 will be less than $\bar{B}^*(p_0)$ when the the prior on θ/σ is a mixture of mean 0 Normal priors. And indeed, an H_1 prior on θ/σ which is a genuine mixture of different mean 0 Normal distributions would generally need fewer than $k(p_0)$ degrees of freedom for $\bar{B}^*(p_0)$ to be an upper bound on the Bayes factor when the p-value is p_0.

To expand on the point of the previous paragraph, there is no reason to think that H_1 distributions considered above (whether frequentist or Bayesian) should actually be Normal, or Absolute-Normal, or Exponential, or the hierarchical priors for the Chi-squared case.. However, our bounds over such classes of priors apply to priors that are mixtures of priors in the corresponding classes. If our (approximate) bound $\bar{B}^*(p)$ on $f(p)$ does not hold for a large family of tests of a certain type (e.g., for a large collection of one-sided z-tests), then it must be that the actual distribution of parameter values under the alternative is not well approximated by mixtures of the distributions studied here.

Here is a possible feature of hypothesis testing that might cause $\bar{B}^*(p)$ to not be an upper bound on $f(p)$ in a frequentist sense, for example in the case of one-sided z tests. Suppose that, for many of the null hypotheses tested, there was preliminary data suggesting that the null hypothesis might be false by about two standard deviations on the scale of measurement in the "official" experiment. Here, "official experiment" means an experiment whose result might be published, at least if the result is interesting (= small-enough p-value?), typically without the preliminary background data entering into the analysis. Null hypotheses without sufficiently promising preliminary data might tend to be winnowed out before progressing to an "official" test. Such winnowing would affect both the prior odds r and the density $f(p)$. The post-winnowing H_1 density of the normal mean might have a mode a standard deviation or two to the right of the null value. Just to illustrate the difficulty, suppose all false null values of the normal mean are between 1.5 and 2 standard deviations from the null value. Then the Bayes factor against the null for a z-score of 1.75 (one-sided p-value p= .04) would be at least $exp(1.75^2 - 0.25^2)/2 = exp(1.5) = 4.48$, while $\bar{B}^*(.04) = 2.86$.

Another relevant issue is stopping rules for collecting data. If the p-value is a bit greater than some cut-off value like .05 or .01, then the experimenter might collect more data in hopes that additional data will result in a decrease in the p-value to below the cut-off value. Such a sampling procedure would cause the final p-value to *not* have a Uniform[0,1] distribution when the null is true. However, suppose a Bayesian has an H_1 prior distribution for which $\bar{B}^*(p)$ is an upper bound on the Bayes factor against the null for any sample size. (An example would be a two-sided z-test of $H_0{:}\theta = 0$ versus $H_1{:}\theta \neq 0$, with a normal mean=0 H_1 prior.) Then because non-informative stopping rules do not affect the posterior distribution, the Bayes factor against H_0 would still be $\leq \bar{B}^*(p)$ for the final p-value p, providing that the stopping rule is non-informative. Likewise in the frequentist scenario, if the parameter distribution under H_1 is such that the ratio of posterior odds to prior odds is bounded above by $\bar{B}^*(p)$ for any fixed sample size, then the posterior to prior odds ratio will still be bounded above by $\bar{B}^*(p)$ (for the final p-value p) if the stopping rule for sampling, given the data, is not affected by the value of the parameter.

P-values, as commonly used in hypothesis testing, have the advantage that they require no consideration of alternative behavior beyond that implicit in the choice of the test statistic, as in, "If the null is false, then it seems like the test statistic T should be stochastically larger than under the null, so let's calculate the p-value based on T." P-values, as commonly interpreted by less sophisticated users, also have the "advantage" that they are thought to give a a universal measure of the posterior plausibility of the null which does not need to be adjusted according to the type of test used or perhaps even according to the prior plausibility of the null. P-values have the disadvantage that the "advantage" in the previous sentence has no logical basis. Obviously, if all tested nulls are true, then all tested nulls with a specified p-value will also be true, and likewise if all tested nulls are false. Also, the example of testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ based on the p-value $P(|T| \geq |t|)$ when $T - \theta$ has a density proportional to $exp(-x^{10^{500}})$ (which is a slightly smoothed Uniform[-1,1]distribution) shows that an observation with an infinitesimal p-value can correspond to essentially no evidence against the null hypothesis in terms of likelihood ratios. Perhaps some sophisticated users of p-values have learned how to interpret them sensibly in their own fields, but there do seem to be serious systemic problems in science with the interpretation of p-values. Ioannidis (2005) describes such systemic problems with p-values in medicine. Ioannidis gives various reasons for why one should not be surprised to have most medical null hypotheses rejected at the .05 level to actually be true nulls. The $\bar{B}^*(p) = 1/\{epln(1/p)\}$ bound may be viewed as complementing Ioannidis (2005) in terms of the sensible assessment of medical research.

The use of $\bar{B}^*(p)$ as an upper bound on $f(p)$ (and the use of $(3/4)\bar{B}^*(p)$ as an estimate of the Bayes factor under the assumption that the median p-value for false nulls is between .05 and .01) has a good justification for many standard situations, such as z tests and two-sided t tests with moderately large degrees of freedom, at least if mixtures of the sorts of priors considered above in those situations are appropriate. The behavior of parameters in higher dimensional problems like Chi-square tests (either in the sense of subjective prior distributions or in the sense of frequentist behavior in a large class of tests) is less easy to model convincingly, so the justification of $\bar{B}^*(p)$ in these situations is more tentative. However, the situations where the bound $\bar{B}^*(p)$ has a good justification constitute a substantial fraction of real-world hypothesis tests.

As noted above, a major "advantage" of p-values is that they require minimal input and a major disadvantage is they are hard, perhaps impossible, to interpret directly in any rational way, for example in a way that has some decision-theoretic justification. Calculating $\bar{B}^*(p)$ requires no more input, aside from a couple of keystrokes on a calculator, than the p-value itself, yet it does have a clear interpretation in terms of what happens when null hypotheses are sorted according to p-values. For example, the fact that $\bar{B}^*(.01) = 7.99$ implies (assuming that $\bar{B}^*(.01)$ really is an upper bound on $f(.01)$) that the ratio of false to true nulls among tests with p close to .01 is at most 8 times the ratio of false to true nulls in the overall population of tests. Even for those who take a dim view of subjective probability, this bound is an easily understandable statement about the “p-value sorting process” which is much more useful with respect to sensible decision making than saying that “only one percent of true nulls have a p-value this small or smaller”. In neither the above frequentist interpretation of $\bar{B}^*(p)$ nor (of course) in the Bayesian interpretation of $\bar{B}^*(p)$ is the role of the prior odds r obscured, as it is with p-values themselves.

For a Bayesian with unlimited time and energy who wants to assess his/her posterior probability of the null, the “correct” approach is of course to specify priors distributions, or perhaps a range of plausible prior distributions, and to go through the usual Bayesian calculations. For a Bayesian who would like a quick and (only slightly?) dirty first pass at the problem, the $\bar{B}^*(p)$ “upper-bound rule-of-thumb” is extremely appealing.

And so herewith a modest proposal. Let’s tell our students and clients that, in a large family of z tests or of t tests with degrees of freedom at least in the teens, the false-to-true odds among tests with p-value p is unlikely to be much more than $\bar{B}^*(p) = 1/\{epln(1/p)\}$ times the false-to-true odds in the whole family of tests, and that often, perhaps typically, the odds will be about $(3/4)\bar{B}^*(p)$ times the false-to-true odds in the whole family of tests. We should, of course, try to explain the assumptions behind the $\bar{B}^*(p) = 1/\{epln(1/p)\}$ upper bound and the $(3/4)\bar{B}^*(p)$ ballpark estimate. Our better students might even have some awareness and appreciation of these assumptions if our explanations are accompanied by homework problems (and, better yet, test problems) involving the calculation of Bayes factors for various H_1 prior distributions. However, even for those students and clients with whom this "assumptions" business does not register, there are significant benefits of this modest proposal. As things stand now, the p-value itself is usually the only number floating around which supposedly gives a measure of the evidence against H_0, and the "one-in-twenty" 05 or the "one-in-a-hundred" .01 give the impression of rendering H_0 implausible or very implausible. Even if the $\bar{B}^*(.05) = 2.46$ and $\bar{B}^*(.01) = 7.99$ values are dubious in some cases, they are still likely to give a better assessment of the evidence against H_0 than the typical "transpose the conditional" interpretation of .05 and .01. And indeed, as mentioned above, these Bayes factors are in rough agreement with the Efron-Gous(2001) interpretation of Fisher's scale of evidence for p-values. Furthermore, a bound on and/or estimate of the Bayes factor might cause people to more sensibly incorporate an assessment of the prior odds into their considerations. If H_0 is a fake null which one has no reason to think true, then even a large p-value does not render it probable. Conversely, if the falsity of H_0 is farfetched (e.g., prior odds of only .01 against H_0), then the posterior odds against H_0 may still be small even with a traditionally small p-value like .01. Clients and perspicacious students will naturally ask about Bayes factors in other situations. The proper answer to that question, I believe, is to say that the Bayes factor will often be more variable over reasonable classes of priors in such other situations, but that $\bar{B}^*(p) = 1/\{epln(1/p)\}$ is typically an upper bound over *some* reasonable classes of priors, and that $\bar{B}^*(p) = 1/\{epln(1/p)\}$ will generally give a more sensible interpretation of the evidence against H_0 than traditional naive interpretations of p-values.

Finally, lets consider where the burden of proof should fall in situations where the Bayes factor against H_0 has a wide range of plausible values. To return to our frequentist scenario, suppose we have a large class of null hypothesis tests, with r being the ratio of false to true hypotheses within the class. The philosophy of hypothesis testing, as typically presented, is that H_0 is innocent until proven guilty, i.e., that H_0 is to be rejected only if the data render H_0 implausible. If the ratio of false to true nulls among tests in our large class having a p-value near .05 can plausibly be anywhere from $2r$ to $7r$, then treating H_0 as innocent until proven guilty should cause one to focus on the factor of 2 rather than on the factor of 7. And so, if for *some* reasonable class of H_1 priors we have $\bar{B}^*(p) = 1/\{epln(1/p)\}$ as an (approximate) upper bound on the Bayes factor against H_0, then we should be hesitant to say that the evidence against H_0 corresponding to a p-value of .05 in this class changes the odds against H_0 by more than a factor of 2.46. Perhaps there are areas of science in which certain p-values correspond to stronger evidence against the null hypothesis than is consistent with the bound $\bar{B}^*(p) = 1/\{epln(1/p)\}$. Identifying such areas and explaining how it comes to be that the bound $\bar{B}^*(p) = 1/\{epln(1/p)\}$ is violated could be very interesting. Absent a justification for doing otherwise, however, it might be reasonable to adopt $\bar{B}^*(p) = 1/\{epln(1/p)\}$ as a default upper bound on the evidence against the null hypothesis.

4 Acknowledgements

The author thanks Jim Berger for his very helpful comments and Sarah Sellke for help with the numerical calculations, done in MATLAB.

References

A. Baricz. Mills ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340:13621370, 2008.

J. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilibility of p-values and evidence (with discussion). *Journal of the American Statistical Association*, 82:112–133, 135–139, 1987.

W. Edwards, H. Lindman, and L. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193242, 1963.

B. Efron and A. Gous. Scales of evidence for model selection: Fisher versus jeffreys. *IMS Lecture Notes-Monograph Series*, 38:193242, 2001.

I. Good. Standardized tail-area probabilities. *Journal of Computation and Simulation*, 16:65–66, 1982.

J. Ioannidis. Why most published research findings are false. *PLoS Med*, 2: e124, 2005.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

T. Sellke, M. Bayarri, and J. Berger. Calibration of p-values for testing precise null hypotheses. *The American Statistician*, 55:62–71, 2001.

J. Sterne and G. D. Smith. Sifting the evidence: what's wrong with significance tests? *British Medical Journal*, 322:226–231, 2001.

V. Vovk. Mellin transforms and asymptotics: harmonic sums. *Journal of the Royal Statistical Society, Series B*, 55:317351, 1993.