Uniformly Most Powerful Bayesian Tests and Standards for Statistical Evidence

Valen E. Johnson
Texas A&M University

February 27, 2014

STATISTICS
TEXAS A&M UNIVERSITY

Motivation

Multi-agent, Multi-cohort End-Stage Melanoma trial, Standard-of-care survival times:

Biomarker Group	Tmt 1	Tmt 2	Tmt 3
a	4	4	4
b	4	4	4
c	4	4	4
d	6	6	6
e	6	6	6

STATISTICS
TEXAS A&M UNIVERSITY

Assumptions:

- Suppose survival time for patient i in biomarker group b under treatment t is $Y_i \sim Exp(\mu_{bt})$
- You wish to test a null hypothesis $H_0 : \mu_{bt} = \mu_0$ versus $H_1 : \mu_{bt} = \mu_1 > \mu_0$

STATISTICS
TEXAS A&M UNIVERSITY

Alternative hypotheses:

Suppose you know the true mean survival time is μ_t.

- If you want to maximize expected weight of evidence, you take $H_1 : \mu = \mu_t$, because

$$\int_0^1 m_t(y) \log \left[\frac{m_t(y)}{m_0(y)} \right] dy \quad - \quad \int_0^1 m_t(y) \log \left[\frac{m_1(y)}{m_0(y)} \right] dy$$
$$= \quad \int_0^1 m_t(y) \log \left[\frac{m_1(y)}{m_1(y)} \right] dy > 0$$

- This choice of $\mu = \mu_t$ makes all posterior inferences exactly correct, even in a repeated sampling sense

Problems for subjective Bayesian analysis

- μ_t is generally not known.
- There is not a unique prior density for survival times of patients (i.e., drug sponsors, physicians, medical centers, patients, regulatory agencies)
- Similarly for decision theoretic analysis; there is no unique loss function
- Decision to proceed to next trial phase not based on Bayes factor, but whether Bayes factor (or significance level) for particular treatment combination exceeds a threshold.

STATISTICS
TEXAS A&M UNIVERSITY

Probability of exceeding threshold

- In practice, we usually reject H_0 if the Bayes factor exceeds a threshold, say γ. In adaptive trial, we may also reject H_1 if $BF_{10} < 1/\gamma$
- If we believe null is false, then we really want to maximize

$$\mathbf{P}_{\mu_t}[BF_{10}(y) > \gamma].$$

- For exponential data and a point alternative hypothesis, the log of the Bayes factor is

$$\log[BF_{10}(\mathbf{y})] = -n\left[\log(\mu_1) - \log(\mu_0)\right] - \left(\frac{1}{\mu_1} - \frac{1}{\mu_0}\right)\sum_{i=1}^{n} y_i$$

STATISTICS TEXAS A&M UNIVERSITY

Probability of exceeding threshold

- Probability that $\log(BF_{10})$ exceeds $\log(\gamma)$ can be written

$$P_{\mu_t}\left[\sum_{i=1}^{n} y_i > \frac{\log(\gamma) + n\left[\log(\mu_1) - \log(\mu_0)\right]}{\frac{1}{\mu_0} - \frac{1}{\mu_1}}\right]$$

STATISTICS
TEXAS A&M UNIVERSITY

Probability of exceeding threshold

- Probability that $\log(BF_{10})$ exceeds $\log(\gamma)$ can be written

$$P_{\mu_t}\left[\sum_{i=1}^{n} y_i > \frac{\log(\gamma) + n\left[\log(\mu_1) - \log(\mu_0)\right]}{\frac{1}{\mu_0} - \frac{1}{\mu_1}}\right]$$

Minimizing the RHS maximizes $\mathbf{P}_{\mu_t}[BF_{10}(y) > \gamma]$, regardless of the value of μ_t

STATISTICS
TEXAS A&M UNIVERSITY

Notation and assumptions

- H_0, H_1 denote models/hypotheses
- $f(\mathbf{y} \mid \boldsymbol{\theta})$ denotes the sampling density under all models
- $m_i(\mathbf{y})$ denotes the marginal density of data under model i
- Θ denotes parameter space
- $\pi_i(\boldsymbol{\theta})$ denotes the prior density for $\boldsymbol{\theta} \in \Theta$ under model i
- $BF_{10}(\mathbf{y})$ denotes the Bayes factor between H_1 and H_0

Definition

A **uniformly most powerful Bayesian test** for a given evidence threshold γ, in favor of an alternative hypothesis H_1 against a fixed null hypothesis H_0 is a Bayesian hypothesis test in which the Bayes factor for the test satisfies the following inequality

$$\mathbf{P}_{\boldsymbol{\theta}_t}\left[BF_{10}(\mathbf{y}) > \gamma\right] \geq \mathbf{P}_{\boldsymbol{\theta}_t}\left[BF_{20}(\mathbf{y}) > \gamma\right] \tag{1}$$

for any $\boldsymbol{\theta}_t \in \Theta$ and for all alternative hypotheses $H_2 : \boldsymbol{\theta} \sim \pi_2(\boldsymbol{\theta})$:

One parameter exponential family models

- Suppose $\mathbf{x} = \{x_1, \ldots, x_n\}$ are iid with joint density function

$$f(\mathbf{x}) = \exp\left[-\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\theta)\right]\prod_{i=1}^{n} h(x_i),$$

  where $\eta(\theta)$ is strictly monotonic
- Consider a one-sided test of a point null hypothesis that $H_0 : \theta = \theta_0$ against an arbitrary alternative hypothesis.

UMPBT(γ) for one parameter exponential family models

Theorem

Define

$$g_\gamma(\theta, \theta_0) = \frac{\log(\gamma) + n[A(\theta) - A(\theta_0)]}{\eta(\theta) - \eta(\theta_0)},$$

and define $u = \pm 1$ according to whether $\eta(\theta)$ is monotonically increasing or decreasing, and define $v = \pm 1$ according to whether the alternative hypothesis requires θ to be greater than or less than θ_0, respectively.

Then a UMPBT(γ) can be obtained by restricting the support of $\pi_1(\theta)$ to values of θ that belong to the set

$$\arg\min_\theta uv\, g_\gamma(\theta, \theta_0).$$

Implications

- Like classical uniformly most powerful tests, UMPBTs exist for all common 1PEFs
- Unique UMPBTs are often defined by simple alternative hypotheses; exceptions occur when several values of parameter define the same rejection region
- Rejection regions for UMPBTs in exponential family models can generally be matched to rejection regions of UMPTs by appropriate choice of γ and Type I error

STATISTICS
TEXAS A&M UNIVERSITY

Implications

- Like classical uniformly most powerful tests, UMPBTs exist for all common 1PEFs
- Unique UMPBTs are often defined by simple alternative hypotheses; exceptions occur when several values of parameter define the same rejection region
- Rejection regions for UMPBTs in exponential family models can generally be matched to rejection regions of UMPTs by appropriate choice of γ and Type I error

  $\Rightarrow$ This property establishes a connection between BFs and p-values

Asymptotic Properties

Theorem

For a one parameter natural exponential family density, suppose that $A(\theta)$ has three bounded derivatives in a neighborhood of θ_0, and let θ^* denote a value of θ that defines a UMPBT(γ) test and satisfies

$$\frac{dg_\gamma(\theta^*, \theta_0)}{d\theta} = 0. \tag{2}$$

Asymptotic Properties

Theorem

Then the following statements are true.

1. *For some $t \in (\theta_0, \theta^*)$,*

$$|\theta^* - \theta_0| = \sqrt{\frac{2\log(\gamma)}{nA''(t)}}. \quad (3)$$

2. *Under the null hypothesis,*

$$\log(BF_{10}) \to N\left(-\log(\gamma), 2\log(\gamma)\right) \quad \textit{as} \quad n \to \infty. \quad (4)$$

Asymptotics

- As $n \to \infty$, UMPBT(γ) alternative converges to null hypothesis, *for fixed* γ.
- In practice, very large samples are collected for hypothesis tests when either
  1. A very small effect size is being tested, or
  2. Very strong evidence against H_0 is required
- Asymptotic properties of UMPBTs seem consistent with actual statistical practice
- Evidence in favor of true null is probabilistically bounded by log(γ)
- Rates at which to increase γ with n are topic for additional research

STATISTICS
TEXAS A&M UNIVERSITY

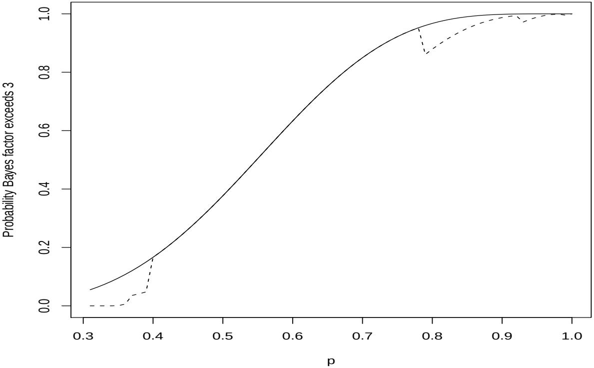
Examples

Binomial data

- Suppose $y \sim Binom(n, \pi)$
- $H_0 : \pi = 0.3,\ n = 10,\ \gamma = 3;\ H_1 : \pi > 0.3$
- UMPBT(γ) value of π_1 satisfies

$$\begin{aligned} \pi_1 &= \arg\min_{\pi} \frac{\log(\gamma) - n[\log(1-\pi) - \log(1-\pi_0)]}{\log[\pi/(1-\pi)] - \log[\pi_0/(1-\pi_0)]} \\ &= 0.525 \end{aligned}$$

$\mathbf{P}[BF_{10} > 3]$ vs data-generating parameter



Normal data

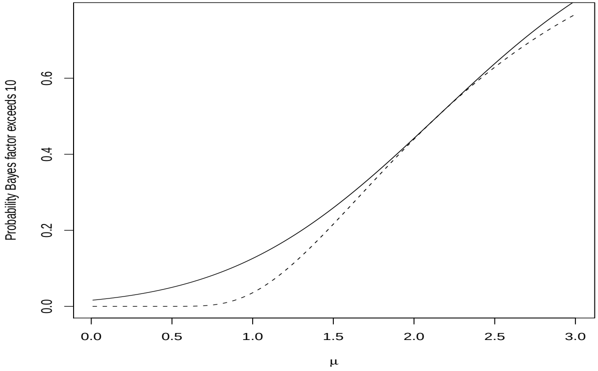
- Suppose $x_1, \ldots, x_n$ iid $N(\mu, \sigma^2)$, σ^2 known
- UMPBT(γ) test of $H_0 : \mu = \mu_0$ is given by

$$\mu_1 = \mu_0 \pm \sigma \sqrt{\frac{2 \log \gamma}{n}},$$

depending on whether $\mu_1 > \mu_0$ or $\mu_1 < \mu_0$.

STATISTICS
TEXAS A&M UNIVERSITY

$\mathbf{P}[BF_{10} > 10]$ vs data-generating parameter for $\sigma^2, n = 1$



Comparison to classical UMPT of normal mean

- Classical one-sided test's rejection region is

$$\bar{x} \geq \mu_0 \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Equating the rejection regions for the UMPBT(γ) test and the UMPT of size α leads to

$$\gamma = \exp\left(z_\alpha^2/2\right)$$

- UMPBT places μ_1 on boundary of classical UMPT rejection region

Comparison to classical UMPT of normal mean

- Classical one-sided test's rejection region is

$$\bar{x} \geq \mu_0 \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Equating the rejection regions for the UMPBT(γ) test and the UMPT of size α leads to

$$\gamma = \exp\left(z_\alpha^2/2\right)$$

- UMPBT places μ_1 on boundary of classical UMPT rejection region
- UMPBT the most "subjective" of objective Bayesian hypothesis tests?

Other Exact UMPBTs

UMPBTs exist for

- Simple tests regarding coefficients in linear models with known observational variances
- Chi-squared tests on one degree of freedom when $H_0 : \lambda = 0$ and $H_1 : \lambda > 0$, λ the non-centrality parameter
- Two-sided tests in 1PEF, under constraint of symmetric alternative.

Approximate UMPBTs

- Approximate UMPBTs can be obtained in normal model hypothesis tests with unknown variances (require data dependent alternative hypotheses).
  1. T-tests (one-sample, paired, two-sample)
  2. Simple tests of linear regression coefficients with unknown observational variance

T-tests

- For one-sample t-test, $P(BF_{10} > \gamma)$ can be expressed as

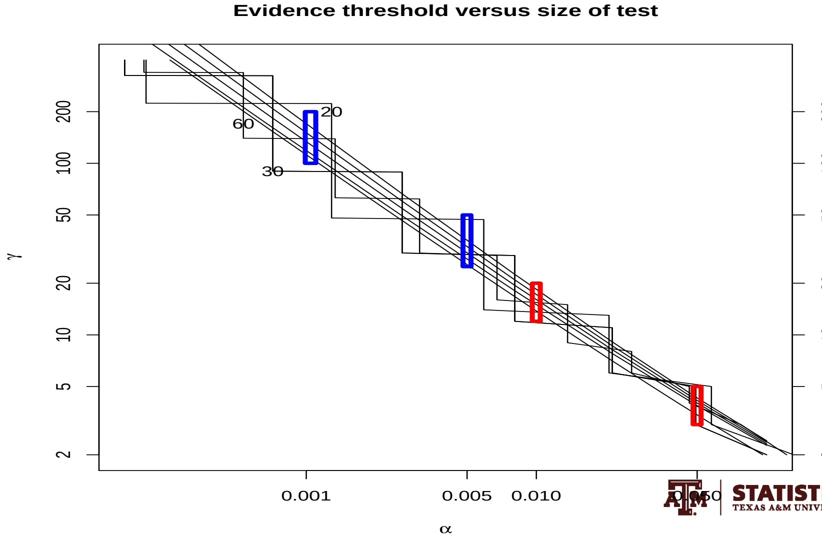
$$P_{\mu_t}\left[a < \bar{x} < b\right]$$

- For two-sample t-test, $P(BF_{10} > \gamma)$ can be expressed as

$$P_{\mu_1-\mu_0}\left[c < \bar{x} < d\right]$$

- The parameters (a, b, c, d) depend on n, γ, and s^2.
- Upper bounds $b, d \to \infty$ with n
- Ignoring upper bound, data-dependent (s^2) approximate UMPBT can be obtained by minimizing a or c.

Bayes evidence thresholds versus test size



Bayes evidence thresholds versus test size

Under assumption of equipoise (i.e., $P(H_0) = P(H_1)$),

- $p = 0.05 \Rightarrow \gamma \in (3, 5) \Rightarrow P(H_0 \,|\, \mathbf{x}) \in (.17, .25)$
- $p = 0.01 \Rightarrow \gamma \in (12, 20) \Rightarrow P(H_0 \,|\, \mathbf{x}) \in (.05, .08)$

- $p = 0.005 \Rightarrow \gamma \in (25, 50) \Rightarrow P(H_0 \,|\, \mathbf{x}) \in (.02, .04)$
- $p = 0.001 \Rightarrow \gamma \in (100, 200) \Rightarrow P(H_0 \,|\, \mathbf{x}) \in (.005, .001)$

STATISTICS
TEXAS A&M UNIVERSITY

Bayes evidence thresholds versus test size

- Standard definitions of "significant" and "highly significant" results correspond to only weak evidence against null hypotheses.
- Definition of "significant" or "highly significant" should require evidence of $> 25 : 1$ or $> 100 : 1$ against the null $\Rightarrow$ p-values of 0.005 or 0.001

Ongoing research:

Scott Goddard, graduate student at Texas A&M, is currently developing "restricted most power Bayesian tests"

- Suppose

$$\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I), \qquad H_0 : \boldsymbol{\beta} = 0, \qquad H_1 : \boldsymbol{\beta} \sim N\left[0, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]$$

- WIth non-informative prior on σ^2, value of g that maximizes probability that $BF_{10} > \gamma$ is

$$argmin \frac{(g+1)}{g}\left[1-(g+1)^{-p/n}\gamma^{-2/n}\right]$$

STATISTICS
TEXAS A&M UNIVERSITY

Restricted most powerful Bayesian tests have applications in

- ANOVA, where they provide correspondence to F tests
- Bayesian variable selection, where γ can be set according to p and n
- Goddard has developed analytic expressions for optimal g and found expressions to set g to control Type 1 error in ANOVA and Bayesian variable selection contexts.

STATISTICS
TEXAS A&M UNIVERSITY

Summary

- UMPBTs provide default objective Bayes factors for the most common of statistical hypothesis tests
- Large sample behavior is reasonable
- Approximately mimic the subjective alternative hypothesis implicit to classical tests (for matched γ and Type I error)
- Correspondence between UMPBTs and UMPTs provide guidance on appropriate definition of significant and highly significant findings, and insight into the non-reproducibility of scientific studies
- Restricted most powerful Bayesian tests can provide default settings for hyperparameters for parametric alternative hypotheses

STATISTICS
TEXAS A&M UNIVERSITY

The End