Lecture 2: Bayesian Hypothesis Testing

Jim Berger

Duke University

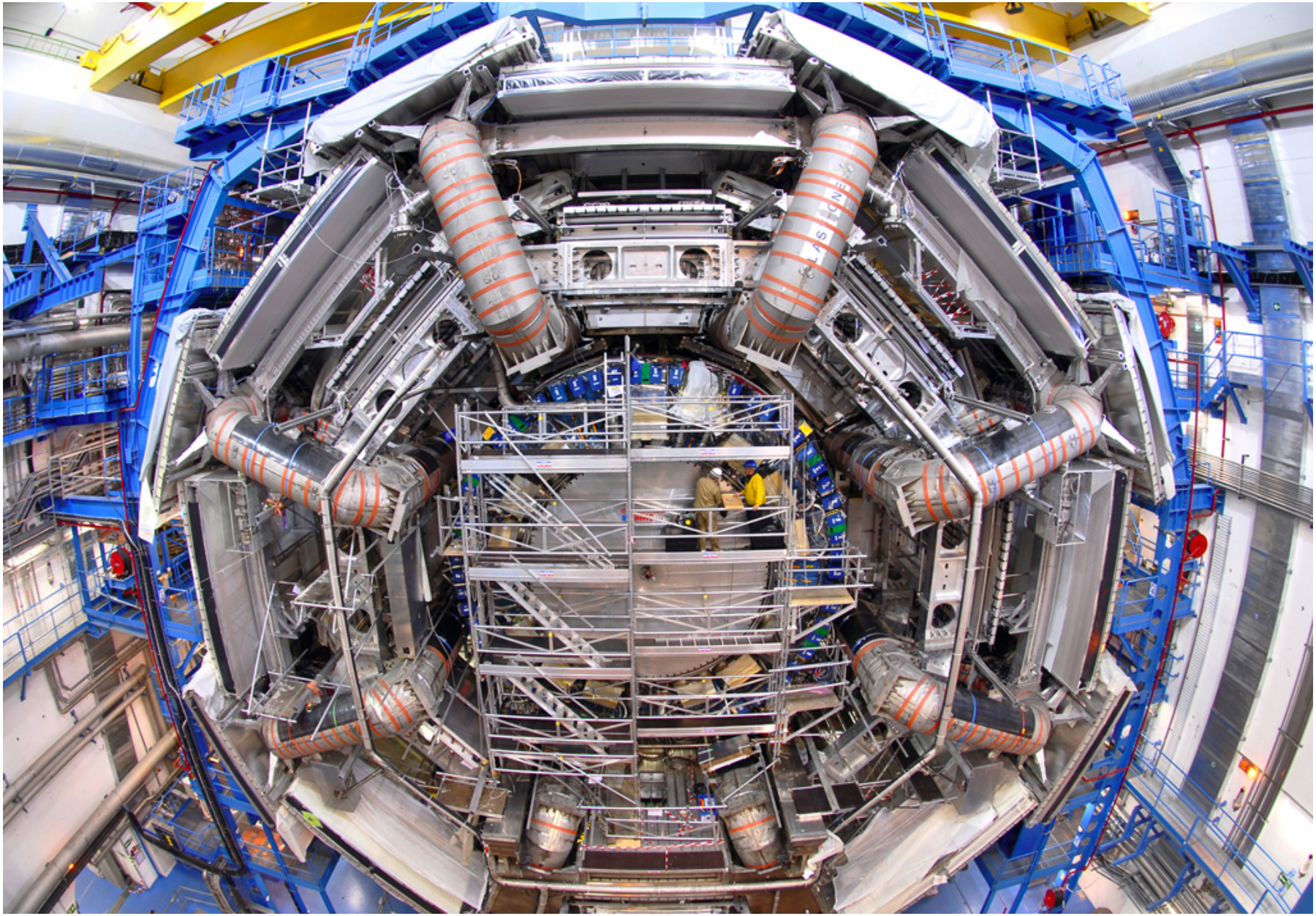
CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Pedagogical introduction to Bayesian testing
- Formal introduction to Bayesian testing
- Precise and imprecise hypotheses
- Choice of prior distributions for testing
- Paradoxes
- Robust Bayesian testing
- Multiple hypotheses and sequential testing
- HIV vaccine example
- Psychokinesis example
- More on p-values and their calibration

I. Pedagogical Introduction to Bayesian Testing

A pedagogical example from high-energy physics: A major goal of the *Large Hadron Collider* at CERN is to determine if the Higgs boson particle actually exists.



Atlas Detector at the Large Hadron Collider (CERN)

Data: $X = \#$ events observed in time T that are characteristic of Higgs boson production in LHC particle collisions.

Statistical Model: X has density

$$\text{Poisson}(x \mid \theta + b) = \frac{(\theta + b)^x e^{-(\theta+b)}}{x!} ;$$

- θ is the mean rate of production of Higgs events in time T;
- b is the (known) mean rate of production of events with the same characteristics from background sources in time T.

To test: $H_0 : \theta = 0$ vs $H_1 : \theta > 0$. (H_0 corresponds to 'no Higgs.')

P-value: $p = P(X \geq x \mid b, \theta = 0) = \sum_{m=x}^{\infty} \text{Poisson}(m \mid 0 + b)$

Case 1: $p = 0.00025$ if $x = 7$, $b = 1.2$

Case 2: $p = 0.025$ if $x = 6$, $b = 2.2$.

Sequential testing: This is actually a sequential experiment, so p should be further adjusted to account for multiple looks at the data.

Bayes factor of H_0 to H_1: *ratio of likelihood under* H_0 *to average likelihood under* H_1 (or "odds" of H_0 to H_1)

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0+b)}{\int_0^\infty \text{Poisson}(x \mid \theta+b)\pi(\theta)\ d\theta} = \frac{b^x\ e^{-b}}{\int_0^\infty (\theta+b)^x\ e^{-(\theta+b)}\pi(\theta)\ d\theta}\ .$$

Subjective approach: Choose $\pi(\theta)$ subjectively (e.g., using the standard physics model predictions of the mass of the Higgs).

Objective approach: Choose $\pi(\theta)$ to be the 'intrinsic prior' (discussed later) $\pi^I(\theta) = b(\theta+b)^{-2}$. (Note that this prior is proper and has median b.)

Bayes factor: is then given by

$$B_{01} = \frac{b^x\ e^{-b}}{\int_0^\infty (\theta+b)^x\ e^{-(\theta+b)} b(\theta+b)^{-2}\ d\theta} = \frac{b^{(x-1)}\ e^{-b}}{\Gamma(x-1,b)}\ ,$$

where Γ is the incomplete gamma function.

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$)

Case 2: $B_{01} = 0.26$ (recall $p = 0.025$)

Posterior probability of the null hypothesis: The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}} .$$

Case 1: $\Pr(H_0 \mid x) = 0.0075$ (recall $p = 0.00025$)

Case 2: $\Pr(H_0 \mid x) = 0.21$ (recall $p = 0.025$)

Complete posterior distribution: is given by

- $\Pr(H_0 \mid x)$, the posterior probability of null hypothesis
- $\pi(\theta \mid x, H_1)$, the posterior distribution of θ under H_1

A useful summary of the complete posterior is $\Pr(H_0 \mid x)$ and C, a (say) 95% posterior credible set for θ under H_1.

Case 1: $\Pr(H_0 \mid x) = 0.0075$; $C = (1.0, 10.5)$

Case 2: $\Pr(H_0 \mid x) = 0.21$; $C = (0.2, 8.2)$

Note: For testing precise hypotheses, confidence intervals alone are *not* a satisfactory inferential summary.

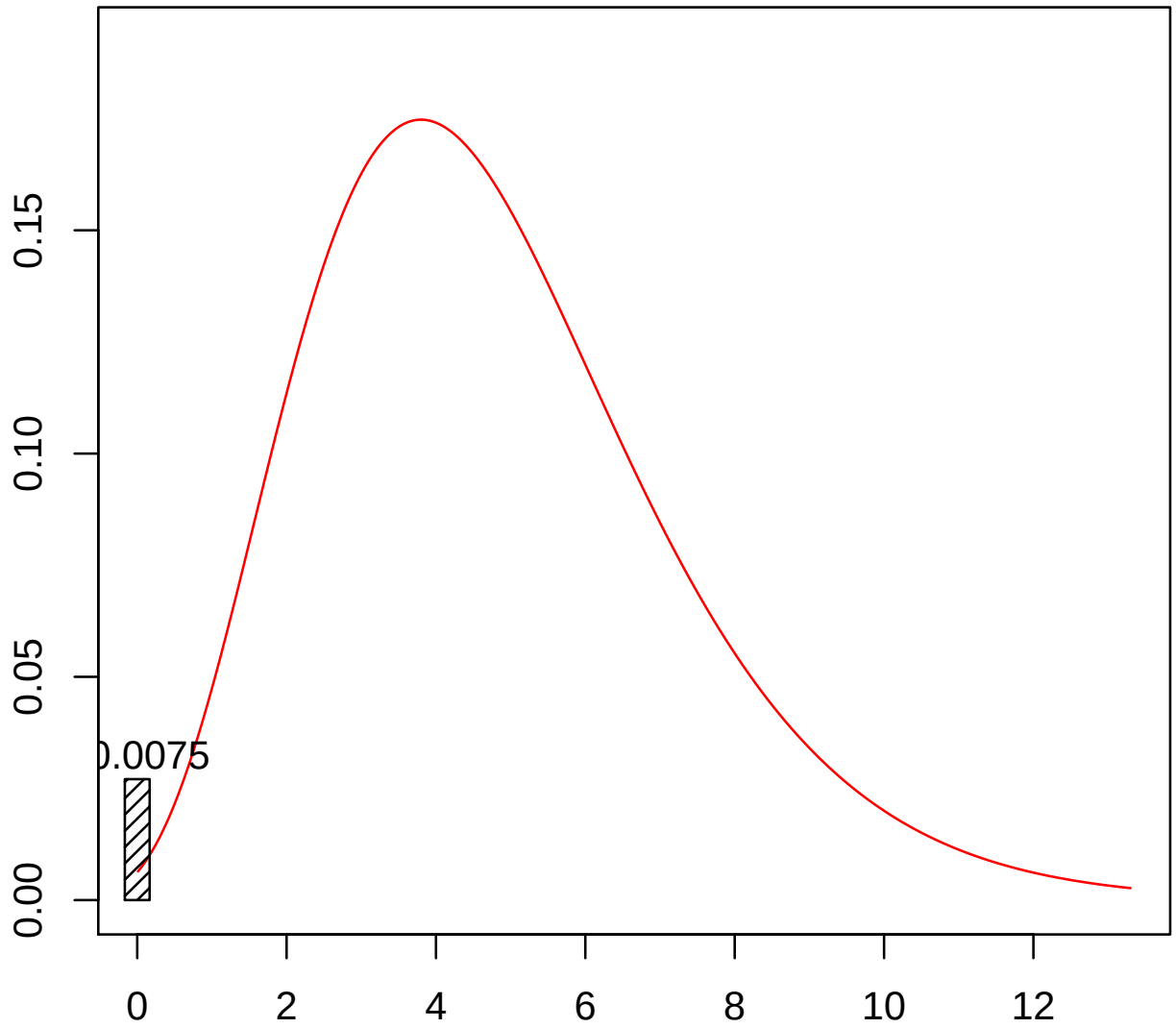


Figure 1: $\Pr(H_0 \mid x)$ (the vertical bar), and the posterior density for θ given x and H_1.

What should a Fisherian or a frequentist think of the discrepancy between the p-value and the objective Bayesian answers in precise hypothesis testing?

Many Fisherians (and arguably Fisher) prefer likelihood ratios to p-values, when they are available (e.g., genetics).

A lower bound on the Bayes factor (or likelihood ratio): choose $\pi(\theta)$ to be a point mass at $\hat{\theta}$, yielding

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0+b)}{\int_0^\infty \text{Poisson}(x \mid \theta+b)\pi(\theta)\, d\theta} \geq \frac{\text{Poisson}(x \mid 0+b)}{\text{Poisson}(x \mid \hat{\theta}+b)} = \min\{1, \left(\frac{b}{x}\right)^x e^{x-b}\}.$$

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$)

Case 2: $B_{01} \geq 0.11$ (recall $p = 0.025$)

Note: Such arguments were first used in
Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," Psychological Review, 70,193-242.

The sources of the difference between p-values and Bayes factors

Consider Case 1, where the p-value $\approx .00025$, but the Bayes factor ≈ 0.0075, differing by a factor of 30:

- A factor of $.0014/.00025 \approx 5.6$ is due to the difference between a tail area $\{X : \ X \geq 7\}$ and the actual observation $X = 7$.
- The remaining factor of roughly 5.4 in favor of the null results from the *Ockham's razor* penalty that Bayesian analysis automatically gives to a more complex model. (A more complex model will virtually always fit the data better than a simple model, and so some penalty is necessary to avoid overfitting.)

Note: A small p-value suggests that something unusual has occurred (unusual at level $[-ep \log p]$) but, if the Bayes factor is not small, the cause of this unusual outcome is not that the *scientific* null hypothesis is wrong. It may be some other cause, such as experimental bias or model misspecification. (One might then argue that the *statistical* null hypothesis is wrong, but this would simply lead to further investigation of the situation, not to a scientific conclusion.)

II. Formal Introduction to Bayesian Testing

Notation

- $\boldsymbol{X} \mid \theta \sim f(\boldsymbol{x} \mid \theta)$.
- To test: $H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$.
- Prior distribution:
  - Prior probabilities $\Pr(H_0)$ and $\Pr(H_1)$ of the hypotheses.
  - Proper prior densities $\pi_0(\theta)$ and $\pi_1(\theta)$ on Θ_0 and Θ_1.
    - $\pi_i(\theta)$ would be a point mass if Θ_i is a point.
- Marginal likelihoods under the hypotheses:

$$m(\boldsymbol{x} \mid H_i) = \int_{\Theta_i} f(\boldsymbol{x} \mid \theta)\pi_i(\theta)\, d\theta, \quad i = 0, 1\,.$$

- Bayes factor of H_0 to H_1:

$$B_{01} = \frac{m(\boldsymbol{x} \mid H_0)}{m(\boldsymbol{x} \mid H_1)}\,.$$

- Posterior distribution $\pi(\theta \mid \boldsymbol{x})$:
  - Posterior probabilities of the hypotheses:

$$\Pr(H_0 \mid \boldsymbol{x}) = \frac{\Pr(H_0) m(\boldsymbol{x} \mid H_0)}{\Pr(H_0) m(\boldsymbol{x} \mid H_0) + \Pr(H_1) m(\boldsymbol{x} \mid H_1)} = 1 - \Pr(H_1 \mid \boldsymbol{x}) \,.$$

  - Posterior densities of the parameters under the hypotheses:

$$\pi_0(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta) 1_{\Theta_0}(\theta)}{m(\boldsymbol{x} \mid H_0)}, \quad \pi_1(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta) 1_{\Theta_1}(\theta)}{m(\boldsymbol{x} \mid H_1)} \,.$$

Two useful expressions:

$$\underbrace{\frac{\Pr(H_0|\boldsymbol{x})}{\Pr(H_1|\boldsymbol{x})}}_{\text{(posterior odds)}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{(prior odds)}} \times \underbrace{B_{01}}_{\text{(Bayes factor)}}$$

$$\Pr(H_0 \mid \boldsymbol{x}) = \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \cdot \frac{1}{B_{01}} \right]^{-1} \,.$$

Conclusions from posterior probabilities or Bayes factors:

- Based on the posterior odds. By default, H_0 accepted if $\Pr(H_0 \mid \boldsymbol{x}) > \Pr(H_1 \mid \boldsymbol{x})$ but often decisions are not reported.
- Alternatively, report Bayes factor B_{01} either because
  - it is to be combined with personal prior odds
  - the 'default' $\Pr(H_0) = \Pr(H_1)$ is used.
  - Jeffreys (1961) suggested the scale

B_{01}	Strength of evidence
1:1 to 3:1	Barely worth mentioning
3:1 to 10:1	Substantial
10:1 to 30:1	Strong
30:1 to 100:1	Very strong
$>$ 100:1	Decisive

Formulation as a decision problem

- Decide between $\begin{cases} a_0, & \text{accept } H_0 \\ a_1, & \text{accept } H_1 \end{cases}$
- With a $0-1$ loss function:

$$L(\theta, a_i) = \begin{cases} 0, & \text{if } \theta \in \Theta_i \\ 1, & \text{if } \theta \in \Theta_j,\ j \neq i \end{cases}$$

- Optimal decision minimizes expected posterior loss, but

$$\begin{aligned} E^{\pi(\theta|x)} L(\theta, a_1) &= \int L(\theta, a_1)\, \pi(\theta \mid \boldsymbol{x})\, d\theta = \Pr(H_0 \mid \boldsymbol{x}) \\ E^{\pi(\theta|x)} L(\theta, a_0) &= \int L(\theta, a_0)\, \pi(\theta \mid \boldsymbol{x})\, d\theta = \Pr(H_1 \mid \boldsymbol{x}), \end{aligned}$$

so that

$$\begin{aligned} a_0 \succ a_1 &\leftrightarrow E^{\pi(\theta|x)} L(\theta, a_0) < E^{\pi(\theta|x)} L(\theta, a_1) \\ &\leftrightarrow \Pr(H_1 \mid \boldsymbol{x}) < \Pr(H_0 \mid \boldsymbol{x}), \end{aligned}$$

so choose the most probable hypothesis.

- More generally, use a $0 - K_i$ loss function:

$$L(\theta, a_i) = \begin{cases} 0, & \text{if } \theta \in \Theta_i \\ K_i, & \text{if } \theta \in \Theta_j,\ j \neq i \end{cases}$$

- Optimal decision a_1 (reject H_0) iif

$$\frac{\Pr(H_0 \mid \boldsymbol{x})}{\Pr(H_1 \mid \boldsymbol{x})} < \frac{K_0}{K_1}\,.$$

- Bayesian rejection regions are usually of same form as classical rejection regions but (fixed) cut-off points are determined by loss functions and prior odds.

III. Precise and Imprecise Hypotheses

(Point Null and One-Sided Hypotheses)

A Key Issue: Is the precise hypothesis being tested plausible?

A *precise hypothesis* is an hypothesis of lower dimension than the alternative (e.g. $H_0 : \mu = 0$ versus $H_0 : \mu \neq 0$).

A precise hypothesis is *plausible* if it has a reasonable prior probability of being true. H_0 : there is no Higgs boson particle, *is* plausible.

Example: Let θ denote the difference in mean treatment effects for cancer treatments A and B, and test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

Scenario 1: Treatment A = standard chemotherapy
Treatment B = standard chemotherapy + steroids

Scenario 2: Treatment A = standard chemotherapy
Treatment B = a new radiation therapy

$H_0 : \theta = 0$ is plausible in Scenario 1, but not in Scenario 2; in the latter case, instead test $H_0 : \theta < 0$ versus $H_1 : \theta > 0$.

Plausible precise null hypotheses:

- H_0 : Males and females of a species have the same characteristic A.
- H_0 : Pollutant A does not affect Species B.
- H_0 : Gene A is not associated with Disease B.
- H_0: There is no psychokinetic effect.
- H_0: Vitamin C has no effect on the common cold.
- H_0: A new HIV vaccine has no effect.
- H_0: Cosmic microwave background radiation is isotropic.

Implausible precise null hypotheses:

- H_0 : Small mammals are as abundant on livestock grazing land as on non-grazing land
- H_0 : Bird abundance does not depend on the type of forest habitat they occupy
- H_0 : Children of different ages react the same to a given stimulus.

An Aside: Hypothesis Testing is Drastically Overused

- Tests are often performed when they are irrelevant.
- Rejection by an irrelevant test is sometimes viewed as "license" to forget statistics in further analysis

A wildlife example:

Habitat Type	Rank	Observed Usage	Hypothesis
A	1	3.8	
B	2	3.6	H_0: "mean usage of habitats by a type
C	3	2.8	of bird is equal for all habitats "
D	4	1.8	Rejected ($p < .025$)
E	5	1.5	
F	6	0.7	

Statistical mistakes in the example

- The hypothesis is not plausible; testing serves no purpose.
- The observed usage levels are given without confidence sets.
- The rankings are based only on observed means, and are given without uncertainties. (For instance, suppose a 95% confidence interval for A is (1.8,5.8), while that for B is (3.5,3.7).)

Approximating a believable precise hypothesis by an exact precise null hypothesis

A precise null, like $H_0 : \theta = \theta_0$, is typically never true *exactly*; rather, it is used as a surrogate for a 'real null'

$$H_0^\epsilon : |\theta - \theta_0| < \epsilon, \quad \epsilon \text{ small}\,.$$

(Even if $\theta = \theta_0$ in nature, the experiment studying θ will typically have a small unknown bias, introducing an ϵ.)

Result (Berger and Delampady, 1987 Statistical Science):
Robust Bayesian theory can be used to show that, under reasonable conditions, if $\epsilon < \frac{1}{4}\,\sigma_{\hat{\theta}}$, where $\sigma_{\hat{\theta}}$ is the standard error of the estimate of θ, then

$$Pr(H_0^\epsilon \mid \mathbf{x}) \approx Pr(H_0 \mid \mathbf{x})\,.$$

Note: Typically, $\sigma_{\hat{\theta}} \approx \frac{c}{\sqrt{n}}$, where n is the sample size, so for large n the above condition can be violated, and using a precise null may not be appropriate, even if the real null is believable.

Posterior probabilities can equal p-values in one-sided testing:

Normal Example:

- $X \mid \theta \sim N(x \mid \theta, \sigma^2)$
- One-sided testing

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

- Choose the usual estimation objective prior $\pi(\theta) = c$, yielding posterior $\theta \mid x \sim N(\theta \mid x, \sigma^2)$.
- Posterior probability of H_0:

$$\begin{aligned} \Pr(H_0 \mid x) = \Pr(\theta \leq \theta_0 \mid x) &= \Phi\left(\frac{\theta_0 - x}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{x - \theta_0}{\sigma}\right) = \Pr(X > x \mid \theta_0) = p\text{-value}. \end{aligned}$$

- The Bayes factor can also be *formally* defined as

$$B_{01} = \frac{\int_{-\infty}^{\theta_0} N(x \mid \theta, \sigma) c \, d\theta}{\int_{\theta_0}^{\infty} N(x \mid \theta, \sigma) c \, d\theta} = \frac{\Phi\left((\theta_0 - x)/\sigma\right)}{1 - \Phi\left((\theta_0 - x)/\sigma\right)} ,$$

  since the unspecified constant c in the Bayes factor cancels.

  – This can be rigorously justified as resulting from a limit of vague proper priors symmetric about θ_0.

- If, instead, one were testing

$$H_0 : 0 \leq \theta \leq \theta_0 \quad vs \quad H_1 : \theta > \theta_0 ,$$

  use of $\pi(\theta) = c$ is much more problematical (in that one appears to be giving infinite prior odds in favor of H_1).

- It has been argued (e.g., Berger and Mortera, JASA99) that objective testing in one-sided cases should not use objective estimation priors.

IV. Choice of Prior Distributions in Testing

A. Choosing priors for "common parameters" in testing:

If pure subjective choice is not possible, here are some guidelines:

- Gold standard: if there are parameters in each hypothesis that have the same group invariance structure, one can use the right-Haar priors for those parameters (even though improper) (Berger, Pericchi and Varshavsky, 1998)
- Silver standard: if there are parameters in each hypothesis that have the same scientific meaning, reasonable default priors (e.g. the constant prior 1) can be used (e.g., in the later vaccine example, where p_1 means the same in H_0 and H_1).
- Bronze standard: try to obtain parameters that have the same scientific meaning (beware the "fallacy of Greek letters"); one strategy often employed is to orthogonalize the parameters, i.e., reparameterize so that the partial Fisher information for those parameters is zero.

An example of testing invariant hypotheses (Dass and Berger, 2003)

- Vehicle emissions data from McDonald et. al. (1995)
- Data consists of 3 types of emissions, hydrocarbon (HC), carbon monoxide (CO) and nitrogen oxides (NO_x) at 4 different mileage states 0, 4000, 24,000(b) and 24,000(a).

Data for 4,000 miles

HC	0.26	0.48	0.40	0.38	0.31	0.49	0.25	0.23
CO	1.16	1.75	1.64	1.54	1.45	2.59	1.39	1.26
NO_x	1.99	1.90	1.89	2.45	1.54	2.01	1.95	2.17
HC	0.39	0.21	0.22	0.45	0.39	0.36	0.44	0.22
CO	2.72	2.23	3.94	1.88	1.49	1.81	2.90	1.16
NO_x	1.93	2.58	2.12	1.80	1.46	1.89	1.85	2.21

Goal: Based on independent data $\mathbf{X} = (X_1, \ldots, X_n)$, test whether the i.i.d. X_i follow the Weibull or the lognormal distribution given, respectively, by

$$\begin{aligned} H_0 : f_W(x \mid \beta, \gamma) &= \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right] \\ H_1 : f_L(x \mid \mu, \sigma^2) &= \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\log x - \mu)^2}{2\sigma^2}\right]. \end{aligned}$$

Both distributions are location-scale distributions in $y = \log x$, i.e., are of the form

$$\frac{1}{\sigma} g\left(\frac{y-\mu}{\sigma}\right)$$

for some density $g(\cdot)$. To see for the Weibull, define $y = \log(x)$, $\mu = \log(\beta)$, and $\sigma = 1/\gamma$; then

$$f_W(y \mid \mu, \sigma) = \frac{1}{\sigma}\, e^{(y-\mu)/\sigma}\, e^{-e^{(y-\mu)/\sigma}}.$$

Berger, Pericchi and Varshavsky (1998) argue that, for two hypotheses (models) with the same invariance structure (here location-scale invariance), one can use the right-Haar prior (usually improper) for both. Here the right-Haar prior is

$$\pi^{RH}(\mu, \sigma) = \frac{1}{\sigma}\, d\sigma d\mu\,.$$

The justification is called *predictive matching* and goes as follows (related to arguments in Jeffreys 1961):

- With only two observations (y_1, y_2), one cannot possibly distinguish between $f_W(y \mid \mu, \sigma)$ and $f_L(y \mid \mu, \sigma)$ so we should have $B_{01}(y_1, y_2) = 1$.
- *Lemma:*

$$\int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma} g\left(\frac{y_1 - \mu}{\sigma}\right) \frac{1}{\sigma} g\left(\frac{y_2 - \mu}{\sigma}\right) \pi^{RH}(\mu, \sigma) d\mu d\sigma = \frac{1}{2|y_1 - y_2|}$$

  for any density $g(\cdot)$, implying $B_{01}(y_1, y_2) = 1$ for two location-scale densities.

Using the right-Haar prior for both models, calculus then yields that Bayes factor of H_0 to H_1 is

$$B(\mathbf{X}) = \frac{\Gamma(n) n^n \pi^{(n-1)/2}}{\Gamma(n-1/2)} \int_0^\infty \left[\frac{v}{n} \sum_{i=1}^n exp\left(\frac{y_i - \bar{y}}{s_y v} \right) \right]^{-n} dv,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$. *Note:* This is also the classical UMP invariant test statistic.

As an example, consider four of the car emission data sets, each giving the carbon monoxide emission at a different mileage level.

For testing, H_0 : Lognormal versus H_1 : Weibull, the results were as follows:

	Data set at Mileage level			
	0	4000	24,000	30,000
B_{01}	0.404	0.110	0.161	0.410

Dass and Berger (2003) showed that this is also a situation where conditional frequentist testing, using the p-value conditioning statistic, can be done resulting in the test:

$$T^B = \begin{cases} \text{if } B(\mathbf{x}) \leq 0.94, & \text{reject } H_0\text{, report Type I CEP} \\ & \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x}))\,; \\ \text{if } B(\mathbf{x}) > 0.94, & \text{accept } H_0\text{, report Type II CEP} \\ & \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})). \end{cases}$$

Note: The CEPs are the same for any value of the parameters under the models (because of invariance) and again equal the objective Bayesian posterior probabilities of the hypotheses.

B. Choosing priors for non-common parameters

If subjective choice is not possible, be aware that

- Vague proper priors are often horrible: for instance, if $X \sim N(x \mid \mu, 1)$ and we test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for θ, the Bayes factor is

$$B_{01}(c) = \frac{f(x \mid 0)}{\int_{-c}^{c} f(x \mid \mu)(2c)^{-1} d\mu} \approx \frac{2c\, f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu) d\mu} = 2c\ f(x \mid 0)$$

  for large c, which depends dramatically on the choice of c.

- Improper priors are problematical, because they are unnormalized; should we use $\pi(\mu) = 1$ or $\pi(\mu) = 2$, yielding

$$B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(1) d\mu} = f(x \mid 0) \quad \text{or} \quad B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(2) d\mu} = \frac{1}{2} f(x \mid 0)\,?$$

- It is curious here that use of vague proper priors is much worse than use of objective improper priors (though neither can be justified).

Various proposed default priors for non-common parameters

- Conventional priors of Jeffreys and generalizations
- Conventional 'robust priors'
- Priors induced from a single prior
- Various data-driven priors (fractional, intrinsic, ...) – discussed in lecture 6.

Jeffreys Conventional prior:

Jeffreys (1961) proposed to deal with the indeterminacy of improper priors in hypothesis testing by:

- Using objective (improper) priors only for 'common' parameters.
- Using 'default' proper priors (but NOT vague proper priors) for parameters that occur in one model but not the other.

Example:

- *Data:* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$
- to choose between

$$M_1 : X_i \sim N(x_i \mid 0, \sigma_1^2)$$
$$M_2 : X_i \sim N(x_i \mid \mu, \sigma_2^2)$$

- Since μ is orthogonal to σ_2^2 (the Fisher information matrix is diagonal), Jeffreys argues that σ_1^2 and σ_2^2 have same meaning $\rightsquigarrow \sigma_1^2 = \sigma_2^2 = \sigma^2$.

- We thus seek $\pi_0(\sigma^2)$ and $\pi_1(\mu, \sigma^2) = \pi_1(\mu \mid \sigma^2)\pi_1(\sigma^2)$.
- The 'common' σ^2 can now be given the same prior, and use of the objective estimation prior $\pi(\sigma^2) = 1/\sigma^2$ is okay, because
  - improper priors are okay for common parameters;
  - answers are very robust to the choice of prior for common parameters. (Kass and Vaidyanathan (1992) also find that Bayes factors are roughly insensitive to choice of a common prior under weaker assumptions than orthogonality; see also Sansó, Pericchi and Moreno, 1996).

  Note: The Berger, Pericchi and Varshavsky (1998) predictive matching argument also applies here, providing a modern argument for this choice.

- $\pi_1(\mu \mid \sigma^2)$ must be proper (and not vague), since μ only occurs in H_1. Jeffreys argued that it
  - should be centered at zero (H_0);
  - should have scale σ (the 'natural' scale of the problem);
  - should be symmetric around zero;
  - should have no moments (more on this later).

  The 'simplest prior' satisfying these is the $Cauchy(\mu \mid 0, \sigma^2)$ prior, resulting in

Jeffreys proposal:

$$\pi_1(\sigma^2) = \frac{1}{\sigma^2} \qquad \pi_2(\mu, \sigma^2) = \frac{1}{\pi\sigma^2} \, \frac{1}{(1 + (\mu/\sigma)^2)} \, .$$

The *robust prior* and Bayesian *t*-test

- Computation of B_{01} for Jeffreys choice of prior requires one-dimensional numerical integration. (Jeffreys gave a not-very-good numerical approximation.)
- An alternative is the 'robust prior' from Berger (1985) (a generalization of the Strawderman (1971) prior), to be discussed in lectures 5 and 7.
  - This prior satisfies all desiderata of Jeffreys;
  - has identical tails and varies little from the Cauchy prior;
  - yields an exact expression for the Bayes factor (Pericchi and Berger)

$$B_{01} = \sqrt{\frac{2}{n+1}} \left(\frac{n-2}{n-1}\right) t^2 \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} \left[1 - \left(1 + \frac{2t^2}{n^2-1}\right)^{-(\frac{n}{2}-1)}\right]^{-1},$$

where $t = \sqrt{n}\bar{x}/\sqrt{\sum(x_i - \bar{x})^2/(n-1)}$ is the usual t-statistic. For $n = 2$, this is to be interpreted as

$$B_{01} = \frac{2\sqrt{2}\, t^2}{\sqrt{3}(1+t^2)\log(1+2t^2/3)}\,. \tag{1}$$

Encompassing approach: inducing priors on hypotheses from a single prior

Sometimes, instead of separately assessing $\Pr(H_0), \Pr(H_1), \pi_0, \pi_1$, it is possible to start with an overall prior $\pi(\theta)$ and *deduce* the $\Pr(H_0), \Pr(H_1), \pi_0, \pi_1$:

$$\Pr(H_0) = \int_{\Theta_0} \pi(\theta)\, d\theta \qquad \text{and} \qquad \Pr(H_1) = \int_{\Theta_1} \pi(\theta)\, d\theta$$

$$\pi_0(\theta) = \frac{1}{\Pr(H_0)}\, \pi(\theta) 1_{\Theta_0}(\theta) \qquad \text{and} \qquad \pi_1(\theta) = \frac{1}{\Pr(H_1)}\, \pi(\theta) 1_{\Theta_1}(\theta)\,.$$

Note: To be sensible, the induced $\Pr(H_0), \Pr(H_1), \pi_0, \pi_1$ must themselves be sensible.

Example: Intelligence testing:

- $X \mid \theta \sim N(x, \mid \theta, 100)$, and we observe $x = 115$.
- To test 'below average' versus 'above average'

$$H_0 : \theta \leq 100 \quad \text{vs} \quad H_1 : \theta > 100.$$

- It is 'known' that $\theta \sim N(\theta \mid 100, 225)$.
- *induced* prior probabilities of hypotheses

$$\Pr(H_0) = \Pr(\theta \leq 100) = \tfrac{1}{2} = \Pr(H_1)$$

- *induced* densities under each hypothesis:

$$\begin{aligned} \pi_0(\theta) &= 2\, N(\theta \mid 100, 225) I_{(-\infty, 100)}(\theta) \\ \pi_1(\theta) &= 2\, N(\theta \mid 100, 225) I_{(100, \infty)}(\theta) \end{aligned}$$

- Of course, we would not have needed to formally derive these.
  - From the original encompassing prior $\pi(\theta)$, we can derive the posterior and $\theta \mid x = 115 \sim N(110.39, 69.23)$.
  - Then directly compute the posterior probabilities:

$$\begin{aligned} \Pr(H_0 \mid x = 115) &= \Pr(\theta \leq 100 \mid x = 115) = 0.106 \\ \Pr(H_1 \mid x = 115) &= \Pr(\theta > 100 \mid x = 115) = 0.894 \end{aligned}$$

V. Paradoxes

Normal Example:

- $X_i \mid \theta \overset{i.i.d.}{\sim} N(x_i \mid \theta, \sigma^2)$, σ^2 known.
- Test $H_0 : \theta = \theta_0 \quad$ versus $\quad H_1 : \theta \neq \theta_0$.
- Can reduce to sufficient statistic $\bar{x} \sim N(\bar{x} \mid \theta\ \sigma^2/n)$.
- Prior on H_1: $\pi_1(\theta) = N(\theta \mid \theta_0, v_0^2)$
- Marginal likelihood under H_1: $m_1(\bar{x}) = N(\bar{x} \mid \theta_0, v_0^2 + \sigma^2/n)$.
- posterior probability:

$$\Pr(H_0 \mid \bar{x}) = \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\frac{1}{(2\pi(v_0^2+\sigma^2/n))^{1/2}} \exp\left\{-\frac{1}{2} \frac{1}{v_0^2+\sigma^2/n} (\bar{x}-\theta_0)^2\right\}}{\frac{1}{(2\pi\sigma^2/n)^{1/2}} \exp\left\{-\frac{1}{2} \frac{1}{\sigma^2/n} (\bar{x}-\theta_0)^2\right\}}\right]^{-1}$$

$$= \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\exp\{\frac{1}{2}\ z^2\ [1 + \frac{\sigma^2}{nv_0^2}]^{-1}\}}{\{1 + nv_0^2/\sigma^2\}^{1/2}}\right]^{-1},$$

where $z = \dfrac{|\bar{x} - \theta_0|}{\sigma/\sqrt{n}}$ is the usual (frequentist) test statistic for this problem.

Comparing $\Pr(H_0 \mid \bar{x})$ with classical p-value for various "n"

z	p-value	$n = 5$	$n = 20$	$n = 100$	$\underline{\Pr}(H_0 \mid \bar{x})$
1.645	0.1	0.44	0.56	0.72	0.4121
1.960	0.05	0.33	0.42	0.60	0.32213
2.576	0.01	0.13	0.16	0.27	0.1334

where $\underline{\Pr}(H_0 \mid \bar{x})$ is the smallest $\Pr(H_0 \mid \bar{x})$ can be among *all* normal priors with mean θ_0.

The Jeffreys-Lindley and Barlett "paradoxes"

In the normal testing scenario of testing $H_0 : \theta = \theta_0$ with a normal $N(\theta \mid \theta_0, v_0^2)$ prior on the alternative,

$$\Pr(H_0 \mid \bar{x}) = \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\exp\{\frac{1}{2}\, z^2\, [1 + \frac{\sigma^2}{n v_0^2}]^{-1}\}}{\{1 + n v_0^2/\sigma^2\}^{1/2}}\right]^{-1}.$$

Jeffreys-Lindley paradox: for large n,

$$\Pr(H_0 \mid \bar{x}) \approx \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\sqrt{n}\, v_0}{\sigma} \exp\{-\frac{1}{2}\, z^2\},$$

so that a classical test can strongly reject the null (which happens with large z) and a Bayesian analysis strongly support the null (if $\sqrt{n}\exp\{-\frac{1}{2}\, z^2\}$ is large).

Bartlett paradox: as $v_0^2 \to \infty$ then $\Pr(H_0 \mid \bar{x}) \to \infty$ so that proper priors in testing can not be 'arbitrarily flat.'

VI. Robust Bayesian Testing

Two versions:

A. Find the spread of answers over a plausible range of prior inputs.

B. Find 'lower bounds' on Bayes factors over any reasonable prior.

First, let's look at the spread of answers over a range of priors:

Scenario:

- We observe $X \sim N(x \mid \mu, 1)$, where μ is the unknown mass of the Higgs boson.
- To test: $H_0 : \mu = 0$ (i.e., the particle does not exist) versus $H_0 : \mu > 0$.

Case 1: $\pi(\mu)$ is Uniform$(0, 10)$ (e.g., known upper limit on μ)

- Observe $x = 2$: $p = 0.025$, while $\Pr(H_0 \mid x = 2) = 0.54$
- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 1.3 \times 10^{-3}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 6.0 \times 10^{-8}$

Case 2: $\pi(\mu)$ is Normal$(4, 1)$ (arising from a previous experiment)

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 4.7 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.8 \times 10^{-8}$

Case 3: $\pi(\mu)$ is a point mass at 4 (the prediction of a new theory).

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 3.4 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 1.1 \times 10^{-7}$

Conservative conversion of p to Pr$(H_0 \mid x)$*:* $\Pr(H_0 \mid x) = (1 + (-ep \log p)^{-1})^{-1}$:

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 8.8 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.7 \times 10^{-8}$

Lower bounds for the Normal Example:

- $X_i \mid \theta \overset{i.i.d.}{\sim} N(x_i \mid \theta, \sigma^2)$, σ^2 known.
- Test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- Prior on H_1: $\pi_1(\theta) = N(\theta \mid \theta_0, v_0^2)$
  - Mean of θ_0 is natural (centering on the null), but choice of v_0^2 is arbitrary.
- Bayes factor:

$$B_{01} = \sqrt{1 + \frac{nv_0^2}{\sigma^2}} \exp\{-\frac{1}{2}\, z^2\, [1 + \frac{\sigma^2}{nv_0^2}]^{-1}\}\,.$$

where $z = \dfrac{|\bar{x} - \theta_0|}{\sigma/\sqrt{n}}$ is the usual (frequentist) test statistic for this problem.

- Compute the minimum of B_{01} and $\Pr(H_0 \mid \bar{x})$ over v_0^2:

  For $z > 1$, the minimizing v_0^2 is $\frac{\sigma^2}{n}\,(z^2 - 1)$ (calculus) and

$$\underline{B}_{01} = \sqrt{e}\, z\, \exp\{-\tfrac{1}{2}\, z^2\}\,, \qquad \underline{\Pr}(H_0 \mid \bar{x}) = [1 + \underline{B}_{01}^{-1}]^{-1}.$$

Bounds for ALL priors

Since $m_1(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\, \pi_1(\theta)\, d\theta \leq f(\boldsymbol{x} \mid \hat{\theta})$

we have

$$B_{01} = \frac{f(\boldsymbol{x} \mid \theta_0)}{m_1(\boldsymbol{x})} \geq \frac{f(\boldsymbol{x} \mid \theta_0)}{f(\boldsymbol{x} \mid \hat{\theta})}$$

and the corresponding bound for $\Pr(H_0 \mid \boldsymbol{x})$ is

$$\underline{\Pr}(H_0 \mid \boldsymbol{x}) \geq \left[1 + \frac{\Pr(H_1)}{\Pr(H_1)} \frac{f(\boldsymbol{x} \mid \hat{\theta})}{f(\boldsymbol{x} \mid \theta_0)}\right]^{-1}$$

For the normal problem, $\hat{\theta} = \bar{x}$ and $f(\boldsymbol{x} \mid \hat{\theta}) = (2\,\pi\,\sigma^2/n)^{-1/2}$ so that, with $\frac{\Pr(H_1)}{\Pr(H_0)} = 1$,

$$B_{01} \geq \exp\{-\frac{z^2}{2}\} \equiv B^*_{01} \qquad \Pr(H_0 \mid \bar{x}) \geq \left[1 + \exp\{\frac{z^2}{2}\}\right]^{-1} \equiv \Pr^*(H_0 \mid \bar{x})$$

In our example:

z	p-value	$\underline{\Pr}(H_0 \mid \bar{x})$	$\Pr^*(H_0 \mid \bar{x})$	B^*_{01}
1.645	0.1	0.4121	0.205	1/3.87
1.960	0.05	0.32213	0.127	1/6.83
2.576	0.01	0.1334	0.035	1/27.60

VII. Multiple Hypotheses and Sequential Testing

Two nice features of Bayesian testing:

- Multiple hypotheses can be easily handled.
- In sequential scenarios, there is no need to 'spend α' for looks at the data; posterior probabilities are not affected by the reason for stopping experimentation.

Normal testing example

The X_i are i.i.d from the $N(x_i \mid \theta, 1)$ density, where

θ = mean effect of T1 - mean effect of T2

Standard Testing Formulation:

$H_0 : \theta = 0$ (no difference in treatments)

$H_a : \theta \neq 0$ (a difference exists)

A More Revealing Formulation:

$H_0 : \theta = 0$ (no difference)

$H_1 : \theta < 0$ (Treatment 2 is better)

$H_2 : \theta > 0$ (Treatment 1 is better)

A default Bayesian analysis:

Prior Distribution:

- Assign H_0 and H_a prior probabilities of 1/2 each
- On H_a, assign θ the "default" Normal(0,2) distribution (so that $Pr(H_1) = Pr(H_2) = 1/4$)

Posterior Distribution:

After observing $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, compute the posterior probabilities of the various hypotheses, i.e.

$Pr(H_0 \mid \boldsymbol{x})$, $Pr(H_1 \mid \boldsymbol{x})$, $Pr(H_2 \mid \boldsymbol{x})$,

and $Pr(H_a \mid \boldsymbol{x}) = Pr(H_0 \mid \boldsymbol{x}) + Pr(H_0 \mid \boldsymbol{x})$

Noting that the Bayes factor of H_0 to H_a can be computed to be

$$B_n = \sqrt{1+2n}\; e^{-n\bar{x}_n^2/(2+\frac{1}{n})}\,,$$

these posterior probabilities are given by

$$\begin{aligned} Pr(H_0 \mid \boldsymbol{x}) &= \frac{B_n}{1+B_n} \\ Pr(H_1 \mid \boldsymbol{x}) &= \frac{\Phi(-\sqrt{n}\,\bar{x}_n/\sqrt{1+\frac{1}{2n}})}{1+B_n} \\ Pr(H_2 \mid \boldsymbol{x}) &= 1 - Pr(H_0 \mid \boldsymbol{x}) - Pr(H_1 \mid \boldsymbol{x}) \end{aligned}$$

Data and posterior probabilities:

Pair	dif	mean	Posterior Probabilities of			
n	x_n	$\bar{x}_n$	H_0	H_a	H_1	H_2
1	1.63	1.63	.417	.583	.054	.529
2	1.03	1.33	.352	.648	.030	.618
3	0.19	0.95	.453	.547	.035	.512
4	1.51	1.09	.266	.734	.015	.719
5	-0.21	0.83	.409	.591	.023	.568
6	0.95	0.85	.328	.672	.016	.657
7	0.64	0.82	.301	.699	.013	.686
8	1.22	0.87	.220	.780	.009	.771
9	0.60	0.84	.177	.823	.006	.817
10	1.54	.091	.082	.918	.003	.915

Comments:

(i) Neither multiple hypothesis nor the sequential aspect caused difficulties. There is no penalty (e.g. 'spending α') for looks at the data

(ii) Quantification of the support for $H_0 : \theta = 0$ is direct. At the 3rd observations, $Pr(H_0 \mid \boldsymbol{x}) = .453$, at the end, $Pr(H_0 \mid \boldsymbol{x}) = .082$

(iii) H_1 can be ruled out almost immediately

(iv) For testing $H_0 : \theta = 0$ versus $H_a : \theta \neq 0$, the posterior probabilities are frequentist error probabilities

The Philosophical Puzzle: How can there be no penalty for looks at the data?

- With unconditional frequentist measures, there must be a penalty, or one could 'sample to a foregone conclusion.'
  - But recall that there was not necessarily a penalty with conditional frequentist measures!
- Bayesian analysis is just probability theory and, if the priors arise physically, there should clearly be no penalty.
- {'statistician's client with a grant application' example.}

But it is difficult; as Savage (1961) said "When I first heard the stopping rule principle from Barnard in the early 50's, I thought it was scandalous that anyone in the profession could espouse a principle so obviously wrong, even as today I find it scandalous that anyone could deny a principle so obviously right."

VIII. HIV Vaccine Example

San Jose Mercury News

mercurynews.com WEST VALLEY 102 Friday, September 25, 2009 THE NEWSPAPER OF SILICON VALLEY 75 cents

AIDS MILESTONE

New path for HIV vaccine

Some in study protected from infection, but trial raises more questions

By Karen Kaplan and Thomas H. Maugh II
Los Angeles Times

Hours after HIV researchers announced the achievement of a milestone that had eluded them for a quarter of a century, reality began to set in: Tangible progress could take another decade.

A Thai and American team announced early Thursday in Bangkok that they had found a combination of vaccines providing modest protection against infection with the virus that causes AIDS, unleashing excitement worldwide. The idea of a vaccine to prevent infection with the human immunodeficiency virus, HIV, had long been frustrating and fruitless.

But by Thursday afternoon, initial euphoria gave way to a more sober assessment. There is still a very long way to go before reaching the goal of producing a vaccine that reliably shields people from HIV.

Some researchers questioned whether the apparent 31 percent reduction in infections was a sta-

See **VACCINE**, *Page 14*

A researcher during the Thai phase III HIV Vaccine Trial, also known as RV 144, tests the "prime-boost" combination of two vaccines.

ASSOCIATED PRESS

Hypotheses, Data, and Classical Test:

- Alvac had shown no effect
- Aidsvax had shown no effect

Question: Would Alvac as a primer and Aidsvax as a booster work?

The Study: Conducted in Thailand with 16,395 individuals from the general (not high-risk) population:

- 74 HIV cases reported in the 8198 individuals receiving placebos
- 51 HIV cases reported in the 8197 individuals receiving the treatment

Model: $X_1 \sim \text{Binomial}(x_1 \mid p_1, 8198)$ and $X_2 \sim \text{Binomial}(x_2 \mid p_2, 8197)$, respectively, so that p_1 and p_2 denote the probability of HIV in the placebo and treatment populations, respectively.

Classical test of $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ yielded a p-value of 0.04.

Bayesian Analysis: Reparameterize to p_1 and $V = 100\left(1 - \frac{p_2}{p_1}\right)$, so that we are testing

$H_0 : V = 0$, p_1 arbitrary

$H_1 : V \neq 0$, p_1 arbitrary.

Prior distribution:

- $Pr(H_i)$ = prior probability that H_i is true, $i = 0, 1$,
- Let $\pi_0(p_1) = \pi_1(p_1)$, and choose them to be either
  - uniform on (0,1)
  - subjective (evidence-based?) priors based on knowledge of HIV infection rates

  Note: the answers are essentially the same for either choice.
- For V under H_1, consider the priors
  - uniform on (-20, 60) (apriori subjective – evidence-based – beliefs)
  - uniform on $(-100c/3, 100c)$ for $0 < c < 1$, to study sensitivity

  (constrained also to $V > 100(1 - \frac{1}{p_1})$.

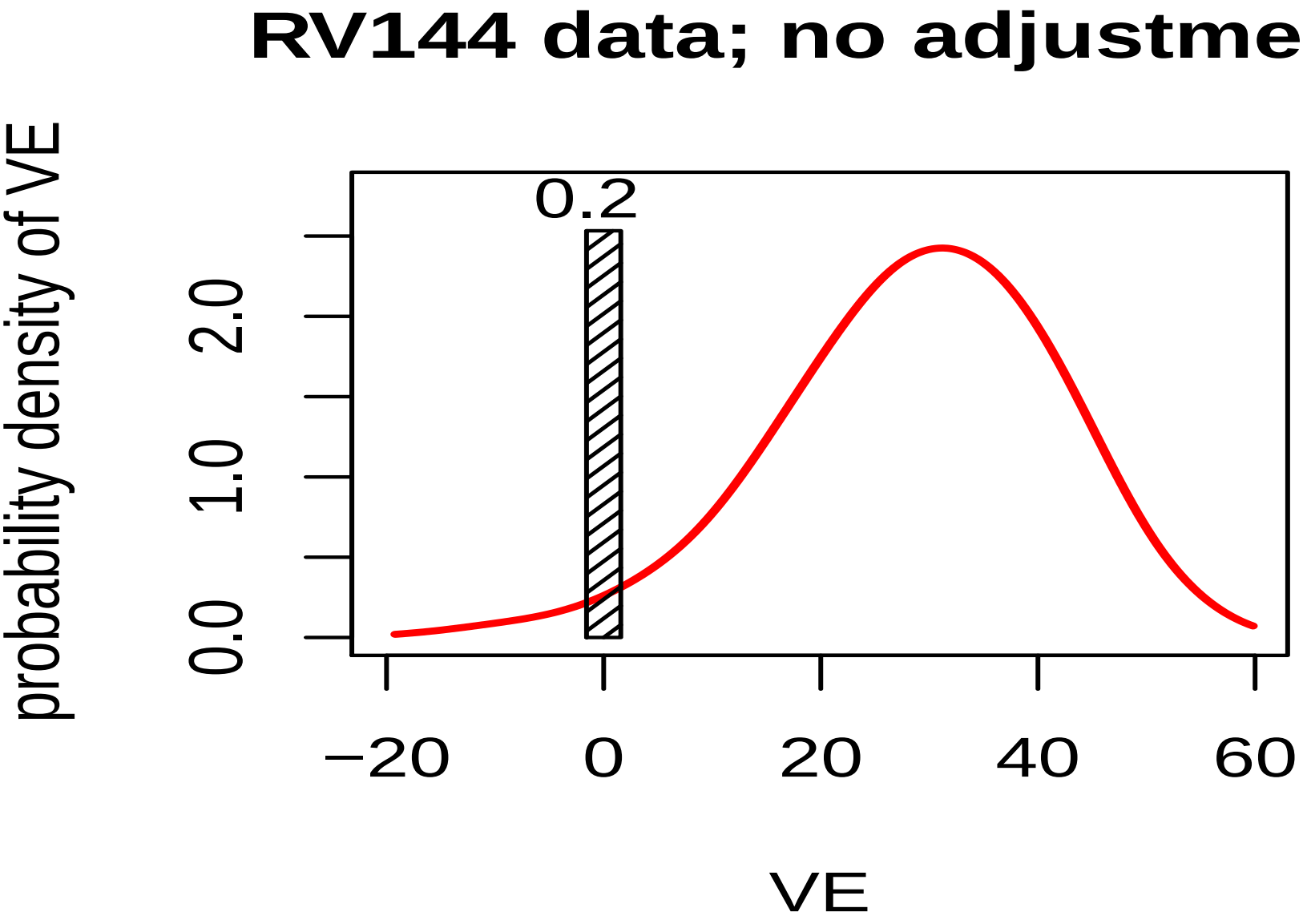
Posterior probability of the null hypothesis:

$$Pr(H_0 \mid \text{data}) = \frac{Pr(H_0)B_{01}}{Pr(H_0)B_{01} + Pr(H_1)} \,,$$

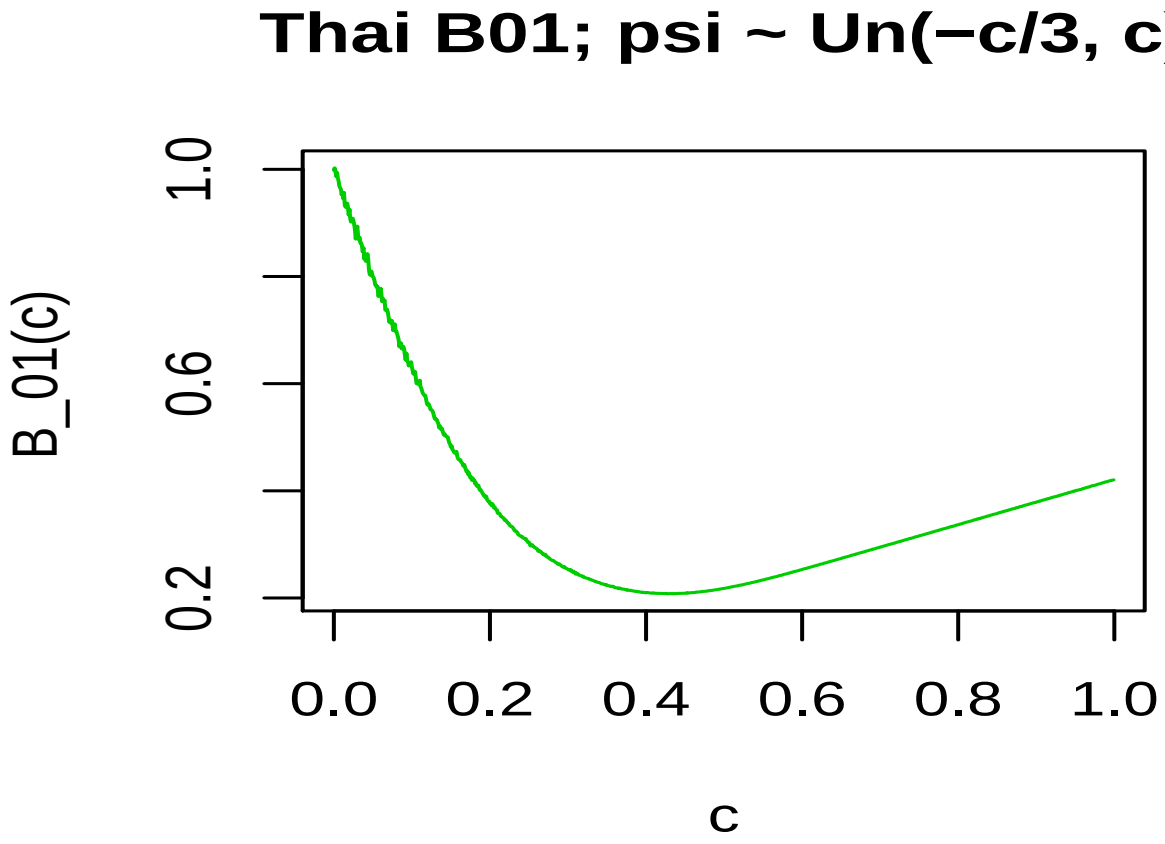
where the Bayes factor of H_0 to H_1 is

$$B_{01} = \frac{\int \text{Binomial}(74 \mid p_1, 8198)\, \text{Binomial}(51 \mid p_1, 8197)\pi_0(p_1)dp_1}{\int \text{Binomial}(74 \mid p_1, 8198)\, \text{Binomial}(51 \mid p_2, 8197)\pi_0(p_1)\pi_1(p_2 \mid p_1)dp_1 dp_2} \,.$$

- For the prior for V that is uniform on (-20, 60),
  $B_{01} \approx 1/4$ (recall, p-value $\approx .04$)
- If the prior probabilitites of the hypotheses are each 1/2, the overall posterior density of V has
  - a point mass of size 0.20 at $V = 0$,
  - a density (having total mass 0.80) on non-zero values of V:


RV144 data; no adjustment
probability density of VE
0.0
1.0
2.0
0.2
−20
0
20
40
60
VE

Robust Bayes: For the prior on V that is uniform on $(-100c/3, 100c)$, the Bayes factor is



Note: There is sensitivity to c, indeed $0.22 < B_{01}(c) < 1$, but why would this cause one to instead report $p = 0.04$, knowing it will be misinterpreted?

Note: Uniform priors are the extreme points of monotonic priors, and so such robustness curves are quite general.

Alternative frequentist perspective:

Let α and $(1-\beta(\theta))$ be the Type I error and power for testing H_0 versus H_1 with, say, rejection region $\mathcal{R} = \{z : z > 1.645\}$. Then

$$\begin{aligned} O &= \text{Odds of } \textit{correct rejection} \text{ to } \textit{incorrect rejection} \\ &= [\text{prior odds of } H_1 \text{ to } H_0] \times \frac{(1-\bar{\beta})}{\alpha}, \end{aligned}$$

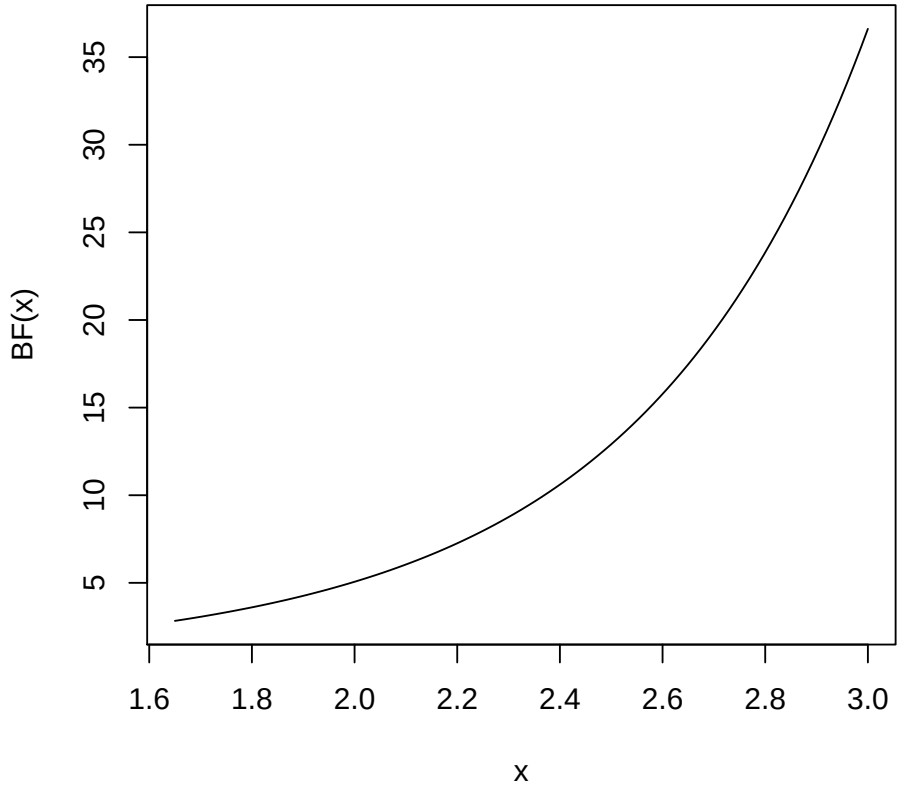
where $(1-\bar{\beta}) = \int (1-\beta(\theta))\pi(\theta)d\theta$ is average power wrt the prior $\pi(\theta)$.

- $\frac{(1-\bar{\beta})}{\alpha} = \frac{\text{average power}}{\text{type 1 error}}$ is the *experimental odds* of correct rejection to incorrect rejection.
- For vaccine example, $(1-\bar{\beta}) = 0.45$ and $\alpha = 0.05$ (the error probability corresponding to $\mathcal{R}$), so $\frac{(1-\bar{\beta})}{\alpha} = 9$.

average power	0.05	0.25	0.50	0.75	1.0	0.01	0.25	0.50	0.75	1.0
type I error	0.05	0.05	0.05	0.05	0.05	0.01	0.01	0.01	0.01	0.01
correct/incorrect	1	5	10	15	20	1	25	50	75	100

But that is pre-experimental; better is to report the actual data-based odds of correct rejection to incorrect rejection, namely the Bayes factor $B_{10}(z)$.

- For vaccine example, here is $B_{10}(z)$ (recall $\frac{(1-\bar{\beta})}{\alpha} = 9$):



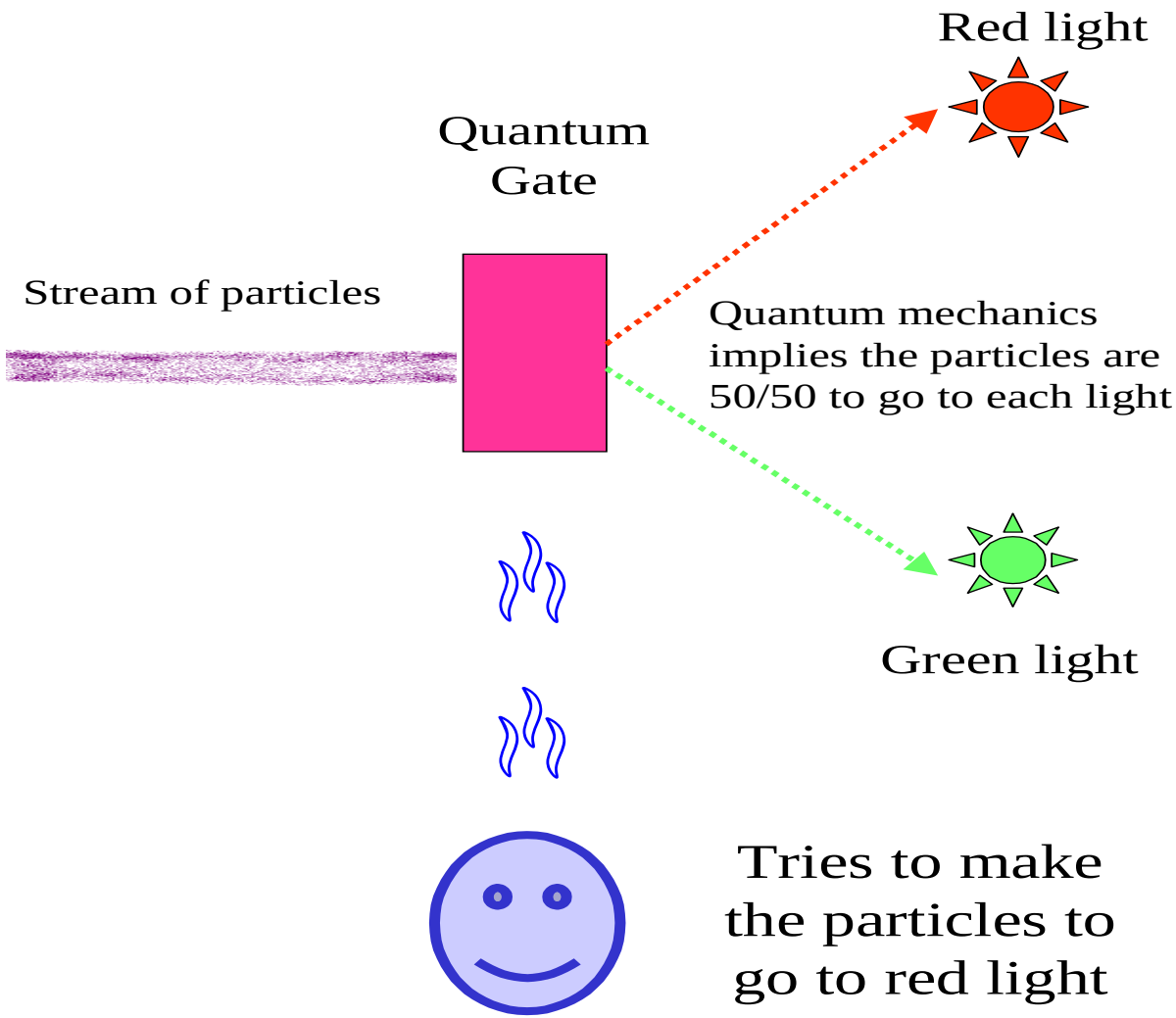
- For simple nulls (or nulls that are simple for the test statistic) $E[B_{10}(Z) \mid H_0, \mathcal{R}] = \frac{(1-\bar{\beta})}{\alpha}$, so reporting $B_{10}(z)$ is a valid conditional frequentist procedure. (Kiefer, 1977 JASA; Brown, 1978 AOS)

IX. Psychokinesis Example

Do people have the ability to perform *psychokinesis*, affecting objects with thoughts?

The experiment:

Schmidt, Jahn and Radin (1987) used electronic and quantum-mechanical random event generators with visual feedback; the subject with alleged psychokinetic ability tries to "influence" the generator.


Red light
Quantum
Gate
Stream of particles
Quantum mechanics
implies the particles are
50/50 to go to each light
Green light
Tries to make
the particles to
go to red light

Data and model:

- Each "particle" is a Bernoulli trial (red = 1, green = 0)
  θ = probability of "1"
  $n = 104,490,000$ trials
  X = # "successes" (# of 1's), $\quad X \sim \text{Binomial}(n, \theta)$
  $x = 52,263,470$ is the actual observation

  $$\begin{aligned} \text{To test } H_0 : \theta &= \tfrac{1}{2} \text{ (subject has no influence)} \\ \text{versus } H_1 : \theta &\neq \tfrac{1}{2} \text{ (subject has influence)} \end{aligned}$$

- P-value $= P_{\theta=\frac{1}{2}}(|X - \frac{n}{2}| \geq |x - \frac{n}{2}|) \approx .0003.$
  Is there strong evidence against H_0 (i.e., strong evidence that the subject influences the particles) ?

Bayesian Analysis: (Jefferys, 1990)

Prior distribution:

$Pr(H_i)$ = prior probability that H_i is true, $i = 0, 1$;

On $H_1 : \theta \neq \frac{1}{2}$, let $\pi(\theta)$ be the prior density for θ.

Subjective Bayes: choose the $Pr(H_i)$ and $\pi(\theta)$ based on personal beliefs

Objective (or default) Bayes: choose

$Pr(H_0) = Pr(H_1) = \frac{1}{2}$

$\pi(\theta) = 1$ (on $0 < \theta < 1$)

Posterior probability of hypotheses:

$$Pr(H_0|x) = \frac{f(x \,|\, \theta = \frac{1}{2})\, Pr(H_0)}{Pr(H_0)\, f(x \,|\, \theta = \frac{1}{2}) + Pr(H_1) \int f(x \,|\, \theta)\pi(\theta) d\theta}$$

For the objective prior,

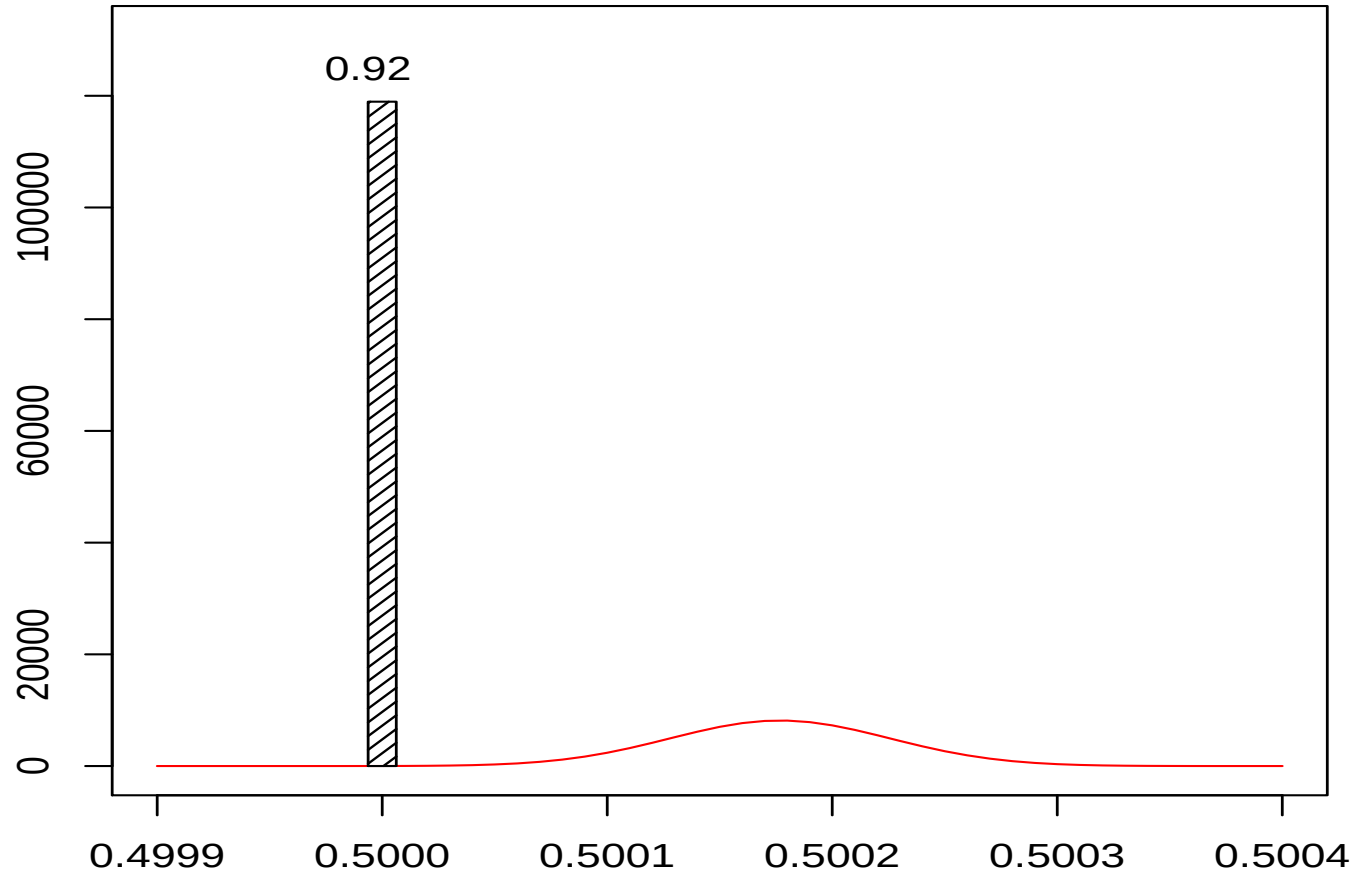
$Pr(H_0 \mid x = 52,263,470) \approx 0.92$

(recall, p-value $\approx .0003$)

Posterior density on $H_1 : \theta \neq \frac{1}{2}$ is

$$\pi(\theta|x, H_1) \propto \pi(\theta) f(x \mid \theta) \propto 1 \times \theta^x (1-\theta)^{n-x},$$

the $Be(\theta \mid 52,263,471\ ,\ 52,226,531)$ density.


0.92
0
20000
60000
100000
0.4999
0.5000
0.5001
0.5002
0.5003
0.5004

Bayes Factor:

$$B_{01} = \frac{\text{likelihood of observed data under } H_0}{\text{'average' likelihood of observed data under } H_1}$$
$$= \frac{f(x \mid \theta{=}\frac{1}{2})}{\int_0^1 f(x \mid \theta)\pi(\theta)d\theta} \approx 12$$

Crash of frequentist and Bayesian conclussions is dramatic.

- $\alpha_0 \leq 0.5$ requires $\pi_1 \geq 0.92$,
- alternatively $\rightsquigarrow$ require a $\pi(\theta)$ under H_1 *extremely* concentrated around $H_0 : \theta = 0.5$ (that is, both hypothesis would then be very precise)

Choice of the prior density or weight function, π, on $\{ \theta : \theta \neq \frac{1}{2} \}$

Consider $\pi_r(\theta) = U(\theta \mid \frac{1}{2} - r, \frac{1}{2} + r)$ the uniform density on $(\frac{1}{2} - r, \frac{1}{2} + r)$

Subjective interpretation: r is the largest chance in success probability that you would expect, given that ESP exists. And you give equal probability to all θ in the interval $(\frac{1}{2} - r, \frac{1}{2} + r)$.

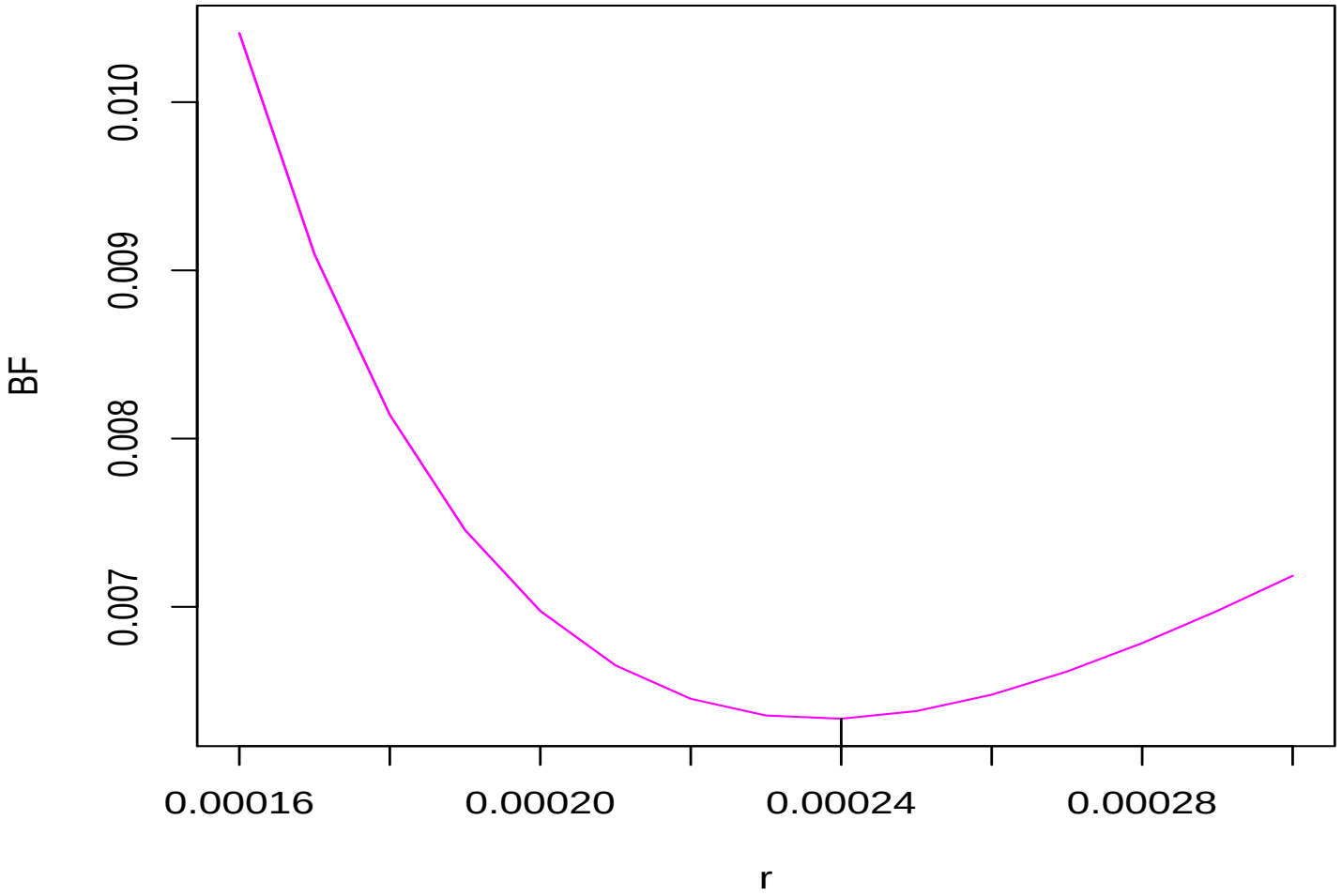
Resulting Bayes factor (letting $FBe(\cdot \mid a, b)$ denote the CDF of the Beta(a, b) distribution)

$$B(r) = \frac{f(x \mid 1/2)}{\int_0^1 f(x \mid \theta)\pi_r(\theta)\, d\theta} = \binom{n}{x} \frac{(n+1)\, r}{2^{n-1}} \left[FB_2 - FB_1\right]^{-1}$$

where

$FB_2 = FBe(\frac{1}{2} + r \mid x + 1, n - x + 1)$ and
$FB_1 = FBe(\frac{1}{2} - r \mid x + 1, n - x + 1)$

For example, $B(0.25) \approx 6$


BF
0.007
0.008
0.009
0.010
0.00016
0.00020
0.00024
0.00028
r

r = largest increase in success probability that would be expected, given ESP exists.

the minimum value of $B(r)$ is $\frac{1}{158}$, attained at the minimizing value of $r = .00024$

Conclusion: Although the p-value is small (.0003), for typical prior beliefs the data would provide evidence *for* the simpler model H_0 : no ESP. Only if one believed a priori that $|\theta - \frac{1}{2}| \leq .0021$, would the evidence for H_1 be at least 20 to 1.

X. More on p-values and their Calibration

Background concerns with p-values

- Concerns with use of p-values trace back to at least Berkson (1937).
- Concerns are recurring in many scientific literatures:

  Environmental sciences: http://www.indiana.edu/~stigtsts/
  Social sciences: http://acs.tamu.edu/~bbt6147/
  Wildlife science: http://www.npwrc.usgs.gov/perm/hypotest/
  http://www.cnr.colostate.edu/~anderson/null.html
- Numerous works specifically focus on comparing the Fisher and N-P approaches (e.g., Lehmann, 1993 JASA: The Fisher, N-P Theories of Testing Hypotheses: One Theory or Two?)
- Even the popular press has become involved: The Sunday Telegraph, September 13, 1998:


Review
My China girl
Josephine, six months after adoption, p4
Vicki Woods
Football, big business and little girls, p3
The great health hoax
Swedish studies pinpoint power line cancer link

Commonly aired issues with p-values

- Inappropriate fixation with $p = 0.05$.
- p-values do not measure the magnitude or importance of the effect being investigated.
- p-values are commonly misinterpreted as
  - the probability that H_0 is true, given the data;
  - the probability an error is made in rejecting H_0;
  - the probability that a 'replicating' experiment would reach the same conclusion.

Cohen (1994): The statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

- Many (most?) null hypotheses are obviously false and will surely be rejected if one simply collects enough data.
  Thompson (1992): "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical significance test to evaluate whether or not there were a lot of subjects, which the researchers already know because . . . they're tired."
- Misuse of p-values is not likely to naturally go away.
  Campbell (1982): "p-values are like mosquitoes [that apparently] have an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them."

Possible solutions?

- Teach p-values for diagnostic purposes, but *not* for inference or decision.
  Hogben (1957): "We can already detect signs of such deterioration in the growing volume of published papers . . . recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration."
  Rozeboom (1997): "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students."

- Only present p-values for precise hypotheses with an appropriate calibration; this will eliminate the worst excesses.

I. J. Good's efforts to calibrate p-values

The difference between p-values and Bayes factors fascinated Good, and he would return to study it throughout his career. As part of his Bayes/non-Bayes compromise, he wanted a simple formula relating the two. Here is a series of suggestions from Good.

- $B_{01} = 3$ or 4 times p: (0.08 or 0.12 in the vaccine example.)
- $10p/3 < B_{01} < 30p$: ($0.133 < B_{01} < 1.2$ in the vaccine example.)
- $B_{01} \approx p\sqrt{n}$: ($B_{01} = 2.5$ in the vaccine example.)
- $p\sqrt{2\pi n}/6 < B_{01} < 6p\sqrt{2\pi n}$: ($1.07 < B_{01} < 38.5$ in the vaccine example.)

Note that his later efforts at calibration all involved the sample size n.

Calibration of p-values from Vovk (1993 JRSSB) and Sellke, Bayarri and Berger (2001 Am. Stat.)

- A *proper* p-value satisfies $H_0 : p(X) \sim \text{Uniform}(0, 1)$.
- Test versus $H_1 : p(X) \sim \text{Beta}(\xi, 1)$, $0 < \xi < 1$. Then, when $p < e^{-1}$,

$$B_{01}(p) = \frac{1}{\xi p^{(\xi-1)}} \geq -e\, p \log(p)\,.$$

  This bound also holds for any alternative $f(p)$, where $Y = -\log(p)$ has a decreasing failure rate (natural non-parametric alternatives).

- The corresponding bound on the conditional Type I frequentist error is

$$\alpha \geq (1 + [-e\, p \log(p)]^{-1})^{-1}\,.$$

p	.2	.1	.05	.01	.005	.001	.0001	.00001
$-ep\log(p)$	.879	.629	.409	.123	.072	.0189	.0025	.00031
$\alpha(p)$	.465	.385	.289	.111	.067	.0184	.0025	.00031

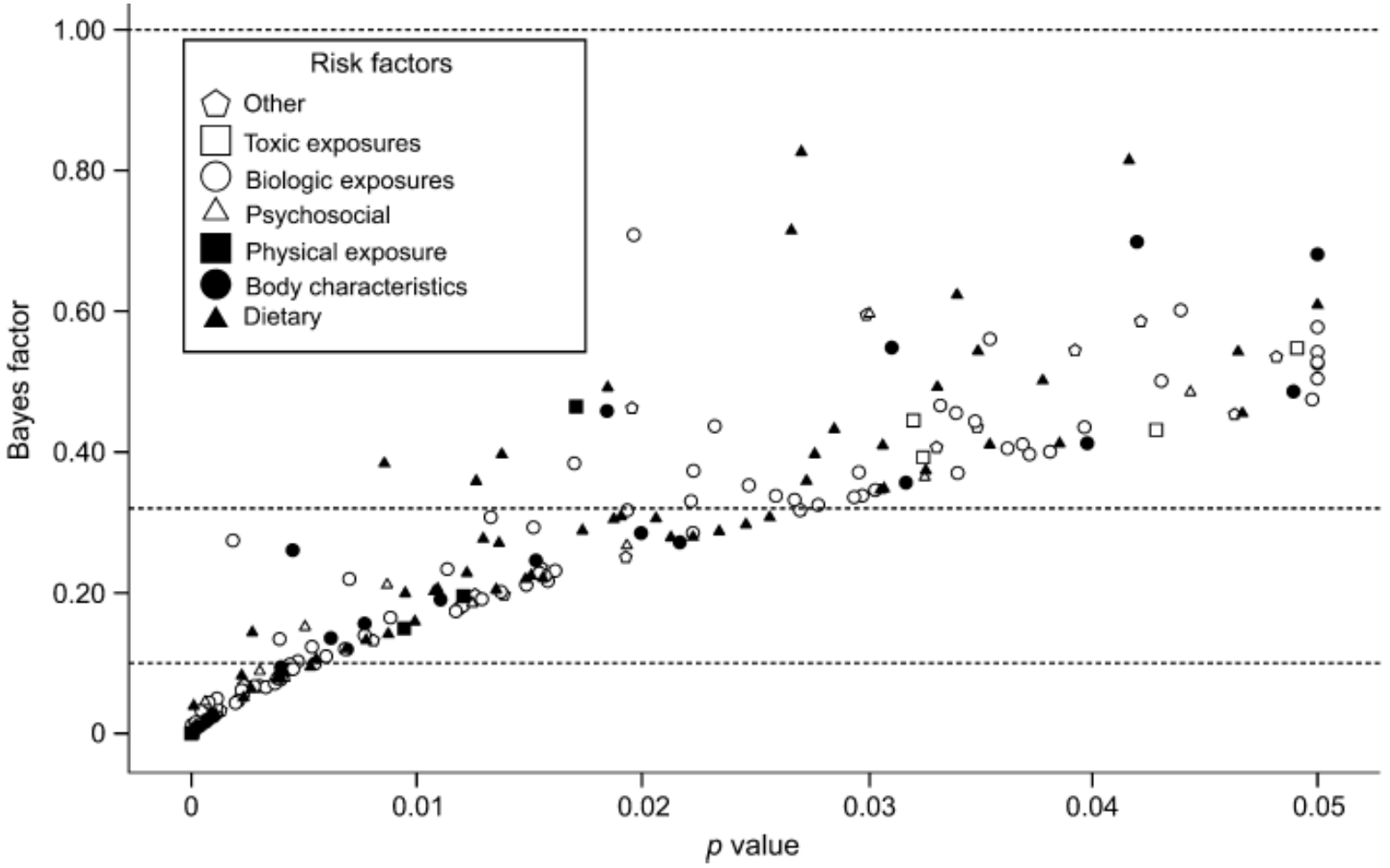


FIGURE 1. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

Figure 2: J.P. Ioannides: Am J Epidemiol 2008;168:374–383

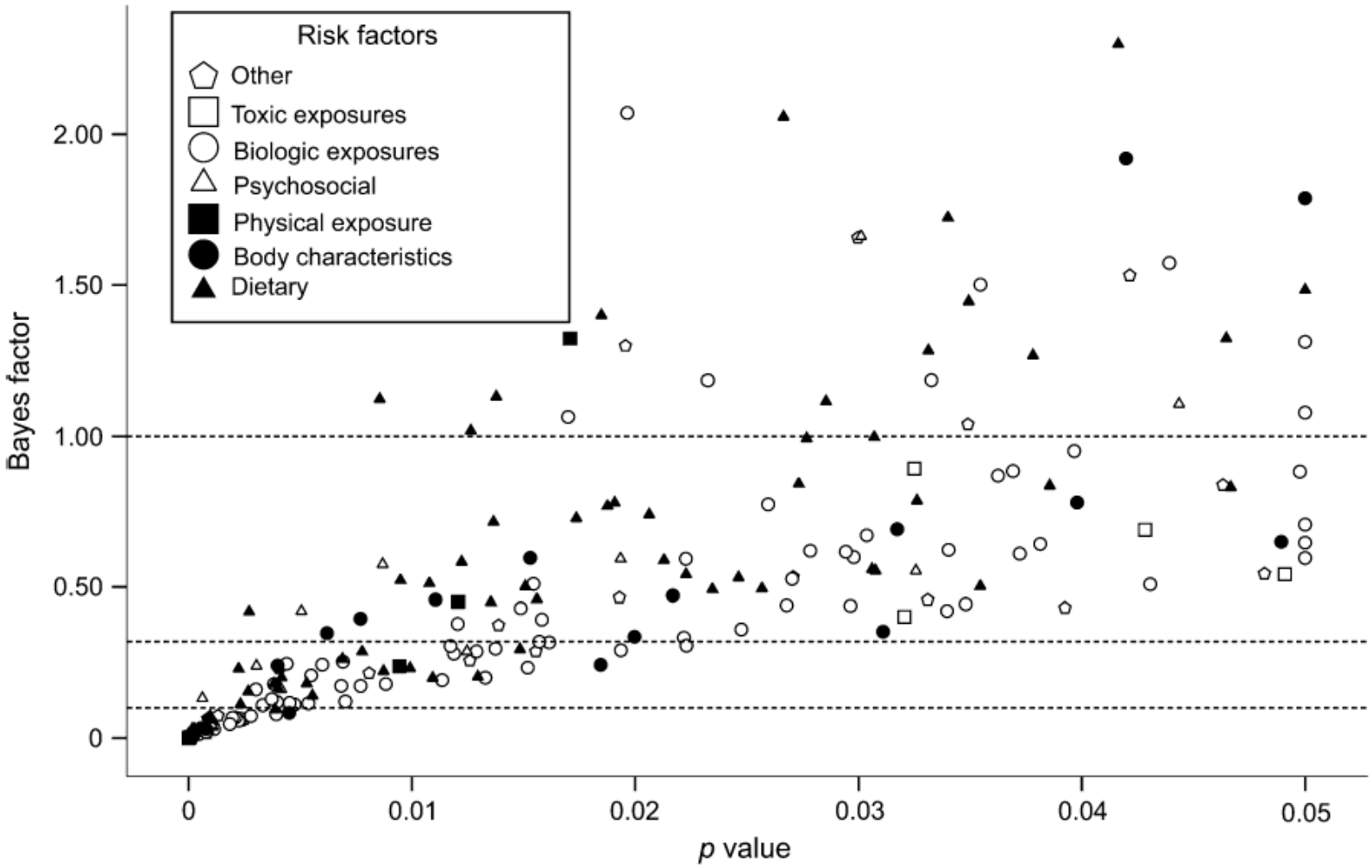


FIGURE 2. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed p value in each study. Shown are calculations assuming θ_A of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

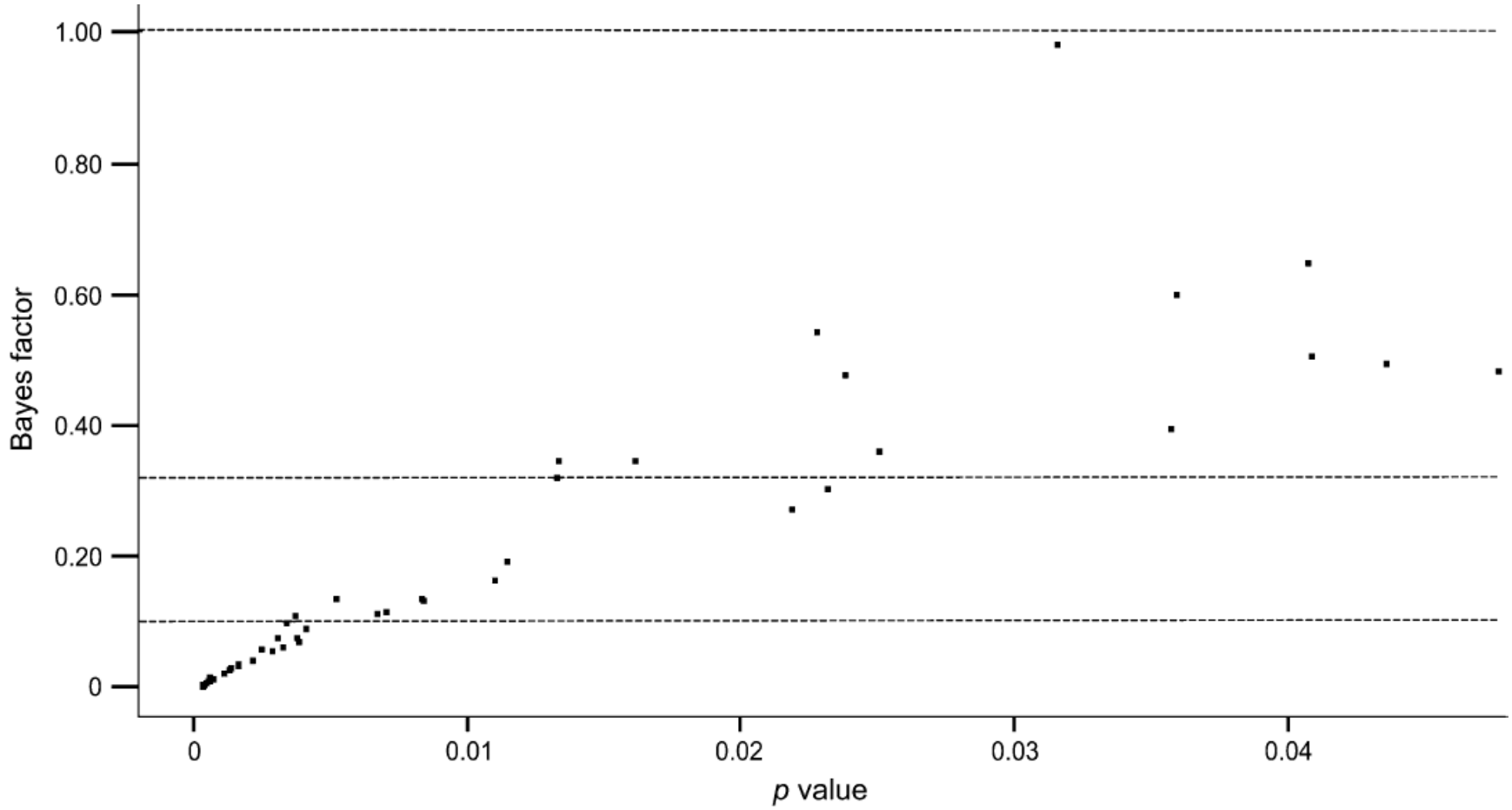


FIGURE 3. Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each meta-analysis. Calculations assume θ_A equal to the median relative risk observed in the 50 genetic associations (relative risk = 1.44). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

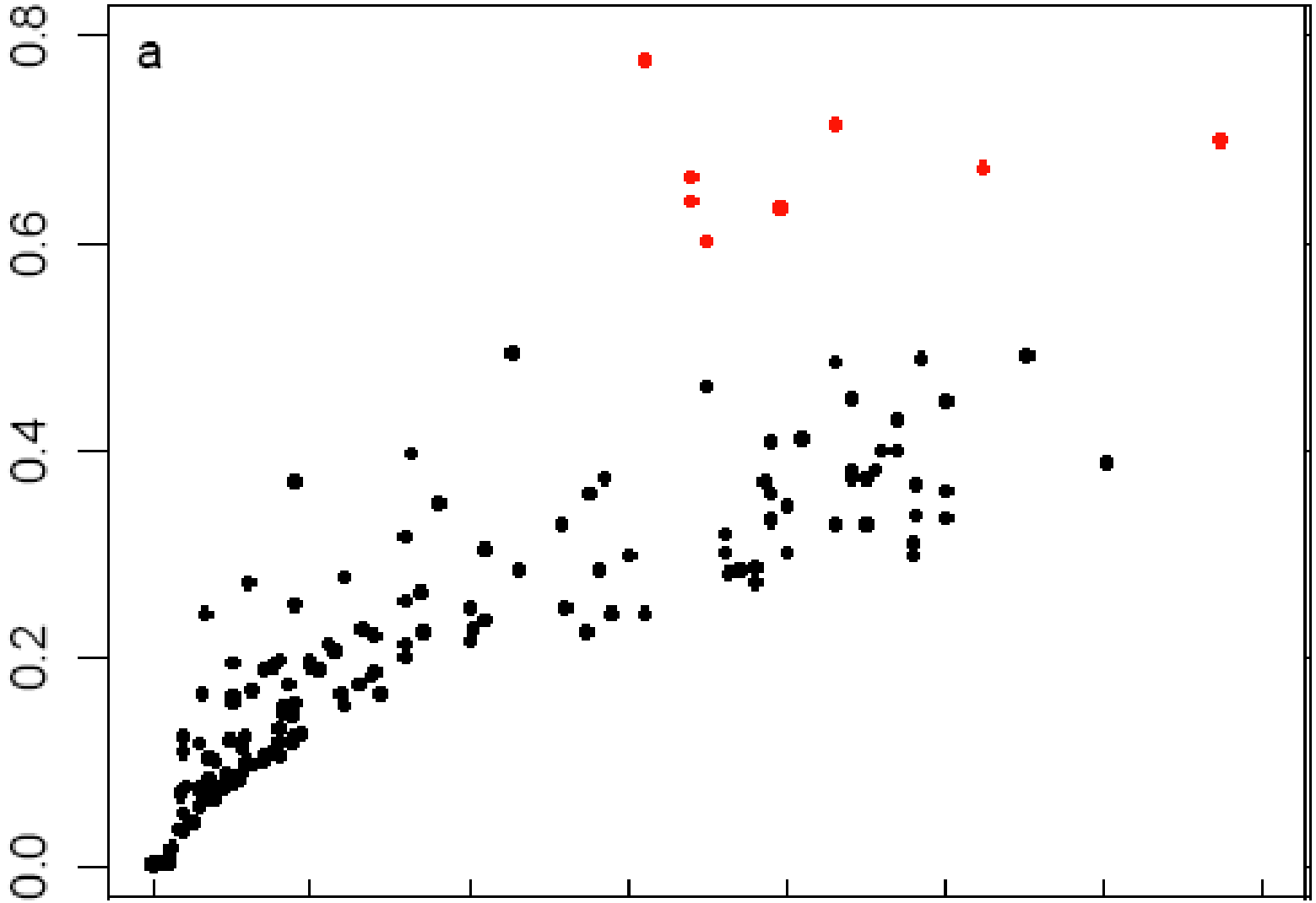


Figure 3: Elgersma and Green (2011): $\alpha(p)$ versus observed p-values for 314 articles in Ecology in 2009.

Two legitimate uses of p-values

- As an indication that something unusual has happened and one should investigate further.
  - Very useful at initial stages of the analysis:
    - If p is not small, don't proceed.
    - If it is small, perform the Bayesian analysis. (I would argue always; physics a year ago had an 8σ result that didn't replicate.)
  - It would be better to measure 'unusual' by $-ep \log p$.
- As a statistic to measure 'strength of evidence', for use in conditional frequentist testing; indeed, it yields conditional frequentist tests that are exactly the same as objective Bayesian tests. (This was the story about how use of Fisher's p-value and conditioning ideas, when combined with Neyman's frequentist testing formulation, results in identical answers as Jeffreys objective Bayesian testing.)