

# A Bayesian's Perspective on Model Checking, Improvement, and Selection

With Illustrations from Source Detection in Astronomy

David A. van Dyk

Statistics Section, Imperial College London

CERN, February 2013

# Outline

- 1 Bayes Factors
- 2 p-values
- 3 ppp-values
- 4 Other Methods
- 5 More Bayes Factors
- 6 The Bottom Line
- 7 Bayes for HEP

# Topics for discussion

- Are two  $3\sigma$  events more evidence than one?
  - ★ Profile versus marginalization.
  - ★ LEE?
- Does LEE-correction depend on whether you bin data?
  - ★ In a sense the resolution increases with  $\sigma$ .
- Background models in other science
  - ★ Astronomy: spectral analysis, image analysis.
  - ★ Signal processing: (often?) static background
  - ★ I guess no one does it more carefully.
- Changing  $H_A$  to match  $p_0$  with  $\alpha$ .

# Bayes Factors and Posterior Probabilities

Bayesian methods have no trouble with unknown parameters

- The prior predictive distribution:

$$p_i(y) = \int p_i(y|\theta)p_i(\theta)d\theta$$

- How likely is  $y$  under model  $i$  (likelihood + prior dist'n).
- Compare models with **Relative Probability**: (avoids tail prob)

$$\text{Bayes Factor} = \frac{p_0(y)}{p_A(y)}.$$

or the *posterior probability of  $H_0$* :

$$\Pr(H_0|y) = \frac{p_0(y)\pi_0}{p_0(y)\pi_0 + p_A(y)(1 - \pi_0)}.$$

*What do these probabilities mean if neither model holds??*

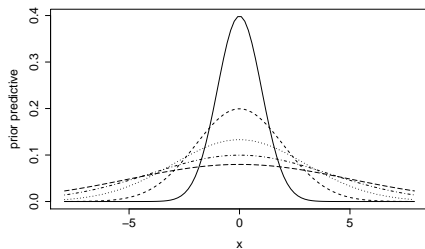
# The Choice of Prior Dist'n Matters!

## Example:

Likelihood:  $y \sim N(\mu, 1)$ .

Prior Dist'n:  $\mu \sim N(0, \tau^2)$ .

Prior Pred.:  $y \sim N(0, 1 + \tau^2)$ .



*Value of  $p_A(y)$  depends on  $\tau^2$ !*

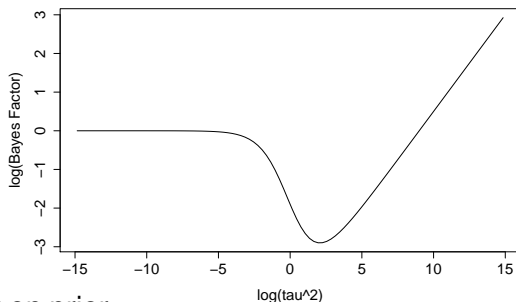
*Must think hard about choice of prior and report!*

# The Choice of Prior Dist'n Matters!

## Bayes Factor:

$$H_0 : y \sim N(0, 1).$$

$$H_A : y \sim N(0, 1 + \tau^2).$$



**Cons:** Depend heavily on prior.

Decision-based method: neither model may hold.

**Pros:** Probability based principled method,  
no problem with nuisance parameters,  
models needn't be nested.

# How to Choose the Prior Dist'n.

- Prior Predictive Distribution is improper w/ improper prior!
  - ★ Prior must be proper, at least for parameters which differ
- Default prior distributions are much harder to come by.
- Subjective prior distributions:
  - ★ Especially illusive in this setting.
  - ★ What are likely parameters values in hypothesized models?
- Problem is even more complicated when:
  - ★ Parameter space is large.
  - ★  $H_0$  and  $H_A$  have different (non-nested) parameters.
- Possible Solutions (see below)
  - ★ Evaluate Bayes Factor over a class of priors.
  - ★ Methods that aim to generate reasonable priors.

# P-values

In the model selection framework, we make a decision between

$H_0$  There is no source.

$H_A$  There is a source.

To quantify the degree of evidence, *p-value* is often reported:

$$\text{p-value} = \Pr(T > T^{\text{obs}} | \text{no source}).$$

(Ignoring the problem of nuisance parameters.)

*A Tail Probability, rather than a Relative Probability.*



# The Abuse of P-values

- Fisher would not approve of using p-values in the model selection (Neyman-Pearson) framework.
- He proposed the p-value as an index to
  - ★ help researchers understand if their model was sufficient
  - ★ guide them in designing future experiments.
  - ★ measure the *strength* of evidence.

*“In an acceptance procedure...acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid out in advance; no thought is given to the particular case, and the tester’s state of mind, or his capacity for learning... By contract, the conclusions drawn by a scientific worker from a test of significance are provisional, and involve an intelligent attempt to understand the experimental situation.”<sup>1</sup>*

---

<sup>1</sup>Fisher, *Statistical Methods and Scientific Induction* (1955)

# Dangers of Using P-values for Model Selection

Although the use of p-values in model selection is endemic, they are not easily interpreted (for a precise  $H_0^2$ ):

- 1 When compared to Bayes Factors or  $\Pr(H_0|\text{data})$ , p-values *vastly overstate the evidence for  $H_1$* .
  - ★ Even using the prior least favorable to  $H_1$  (in a large class).
- 2 Computed given data as extreme or more extreme than  $y$ .
  - ★ This is a **tail probability**.
  - ★ This is *much stronger evidence* for  $H_1$  than  $y$ .
  - ★ Agree with Bayes measures given “as/more extreme”.
- 3 P-values cannot be calibrated to match Bayes Measures
  - ★ Depends on sample size, model, and precision of  $H_0$ .

*P-values bias inference in the direction of **false** discovery.*

---

<sup>2</sup>Berger & Delampady, *Testing Precise Hypotheses*, Stat. Sci., 1987

# Why are p-values so popular?

Use of p-values for model selection

- aims to quantify the level of evidence in a significance test,
- but p-values are not suited to this task.

## Maybe it is just a bad habit....

Assessment of P-values

**Cons:** Biased toward discovery and uninterpretable.

**Pros:** Everyone is doing it...

# Posterior Predictive P-values

How do we compute p-value with unknown param's under  $H_0$ ?

$$\text{p-value} = \Pr(T > T^{\text{obs}} | H_0).$$

- 1 Careful choice of  $T$ , dist'n may not depend on unknowns.
- 2 Use estimates of unknowns under  $H_0$ .
- 3 Average over the posterior dist'n of unknowns under  $H_0$ :

$$\text{ppp-value} = \int \Pr(T > T^{\text{obs}} | H_0) p(\theta | y) d\theta.$$

*ppp-values may be very weak with poor choice of  $T$ . Use LRT!*

Otherwise use partial predictive dist'n:  $p(\theta | y \setminus t) \propto p(y | t, \theta) \pi(\theta) = \frac{p(y|\theta)\pi(\theta)}{p(t|\theta)}$ .

*Designed and Used for Model Checking.*

# The Likelihood Ratio Test Statistic

Neyman-Pearson Testing and P-values require a Test Statistic.

- Often derived on a case-by-case basis.
- An important general Test Statistic: (A Relative Probability)

$$\text{Likelihood Ratio} = \frac{\sup_{\theta \in \Theta_0} p(y|\theta)}{\sup_{\theta \in \Theta_A} p(y|\theta)} \equiv \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_A)}.$$

- Typically calibrated with a tail probability. (Is this necessary?)
- Under conditions  $-2 \log(\text{likelihood ratio}) \overset{\text{asy}}{\sim} \chi^2$ , eg:
  - 1  $\Theta_0$  must be in the interior of  $\Theta_A$ .
  - 2 If  $\Theta_0$  is on the boundary, but all parameters are identified under  $H_0$ ,  $-2 \log(\text{likelihood ratio}) \overset{\text{asy}}{\sim}$  mixture of  $\chi^2$ .

# An Example

## Spectral Analysis in High-Energy Astrophysics<sup>3</sup>

- We fit a power-law continuum and test for an added emission line of (a) known or (b) unknown location.

Model 0: There is no emission line.

Model 1: There is an emission line with fixed location.

Model 2: There is an emission line with unknown location.

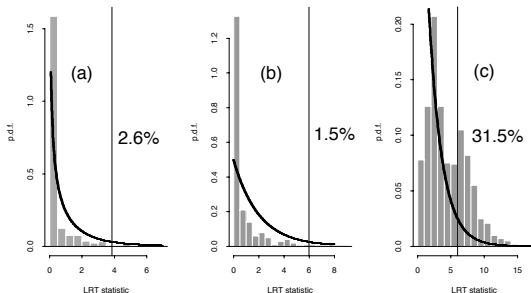
- With known location all parameters are identified under  $H_0$  but  $H_0$  is on the boundary.
- With unknown location *not* all parameters are identified under  $H_0$  and  $H_0$  is on the boundary.
- We also tested for (c) a two-parameter absorption feature.

---

<sup>3</sup>Protossov, et al., 2002, ApJ

# Results

Asymptotic theory may be conservative or anti-conservative.



*By fitting the line location, we adjust for the LEE.*

Assessment of ppp-values:

**Pro:** Can handle nuisance parameters.

**Con:** As p-values, they are not well-suited to *model selection*.

**Pro:** They can be used for *model checking*:

*Is the data consistent with the model?*

# Other Methods

There are *Many* other methods....

- 1 Posterior Likelihood Ratio (\*\*)
- 2 “Posterior Bayes Factors” (\*\*)
- 3 Conditional Error Probabilities (\*\*)
- 4 Expected-posterior prior distributions (\*\*)
- 5 Bayesian Model Averaging
- 6 Decision Theory
- 7 Information Criteria (e.g., AIC, BIC, etc.)
- 8 “Default Bayes Factors”



# Posterior Likelihood Ratio

Dempster (1974) suggested computing the posterior distribution of the LRT (with a sharp null):

$$\text{Likelihood Ratio} = \frac{p(y|\theta_0)}{p(y|\theta)} \equiv \frac{L(\theta_0)}{L(\theta)}.$$

We can compute

$$\Pr\left(\frac{L(\theta_0)}{L(\theta)} < k \mid y\right) \quad \text{e.g., for } k = 0.3, 0.1, \text{ or } 0.05$$

Compared with

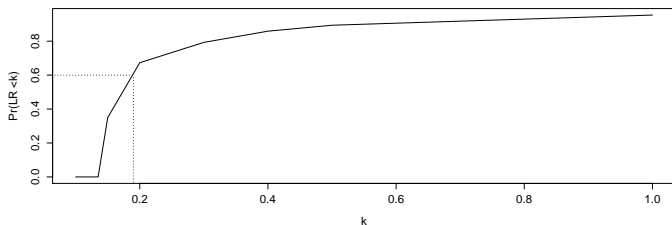
- 1 Bayes Factor:  $L(\theta_0) / \int L(\theta)p(\theta)d\theta$  (high prior dependence)
- 2 Posterior Bayes Factor:  $L(\theta_0) / \int L(\theta)p(\theta|y)d\theta$  (use data twice)

*Posterior LR uses full distribution, rather than a mean.*

# Posterior Likelihood Ratio: Example

Suppose<sup>4</sup>

- $y_i \sim N(\mu, 1)$  for  $i = 1, \dots, 25$  and  $\bar{y} = 0.4$ ,
- Test:  $H_0 : \mu = 0$  versus  $H_1 : \mu \neq 0$ ,
- $Z = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma} = 2$  and p-value = 0.0456,
- Posterior Bayes Factor is 0.19,  $\Pr(LR < PBF) = 0.6$



<sup>4</sup>Following Aitkin (Stat & Comput, 1997)

# Posterior Likelihood Assessment

- Pros:**
- Involves relative likelihood rather than tail probabilities.
  - Bayesian method that isn't sensitive to choice of prior.
  - Clean interpretation as a Bayesian posterior probability.
- Cons:**
- Must interpret the probability that threshold is obtained.
  - Susceptible to the charge that it uses the data twice.
  - Statistical properties are not well studied.

# Conditional Error Probabilities

## Ancillary Statistics:

- 1 Suppose that data  $y$  can be transformed into  $(s, t)$  so that

$$p(y|\theta) \propto p(s|t, \theta)f(t).$$

- 2 Because its dist'n does not depend on  $\theta$ ,  $t$  is ancillary.
- 3 Conventional statistical wisdom:

*Condition on ancillary statistics when possible.*

- ★ Inference for  $\theta$  based on  $p(s|t, \theta)$  tends to be more precise.
  - ★ Subsets defined by fixing  $t$  are the “*relevant subsets*”.
- 4 Intervals, tests, and p-values computed conditional on  $t$ .
    - ★ But  $t$  is not typically unique, so methods are not unique.

# Conditional Error Probabilities

- 1 With simple hypotheses, define (Berger, Brown & Wolpert, AoS, '94)

$p_0$ : The p-value as we have defined it.

$p_1$ : The p-value with  $H_0$  and  $H_A$  interchanged.

$S$  The maximum of  $p_0$  and  $p_1$ .

- 2 Reject  $H_0$  if  $p_0 < p_1$  accept  $H_0$  and accept otherwise.

- 3 Report the conditional error probabilities:

$\alpha(s)$ : Probability of Type 1 error given  $S = s$ .

$\beta(s)$ : Probability of Type 2 error given  $S = s$ .

- 4 With  $\pi_0 = 0.5$ ,

$$\alpha(s) = \Pr(H_0|y) = \frac{\text{Bayes Factor}(y)}{1 + \text{Bayes Factor}(y)}$$

$$\beta(s) = \Pr(H_A|y) = \frac{1}{1 + \text{Bayes Factor}(y)}.$$

# Conditional Error Probabilities

*Example of the use of conditioning to improve the frequency properties of statistical procedures.*

## Assessment of Conditional Error Probabilities

- Pros:**
- Error probs depend on the data (unlike Neyman-Pearson)
  - Conditioning can be technically challenging. Here conditional error probs are computed via Bayes Factor.
  - Seems to unify Bayesian and Frequency methods.
- Cons:**
- Frequency conditional prob's make eyes glaze over.
  - Requires Bayes Factors for composite hypotheses.  
(Prior determines conditioning statistic: Influential!)
  - Bayesian Wolf in a Frequentist Sheep's clothing?  
Test statistic is Bayes Factor for composite hypothesis.

# Other Methods

There are *Many* other methods....

## 1 Bayesian Model Averaging

**Pros:** Bayesian, but less dependent on the choice of prior.

**Cons:** More appropriate for prediction than model selection.

## 2 Decision Theory

**Pros:** Derives rules tailored to specific scientific goals.

**Cons:** Sensitive to choice of Loss Function and Prior.

## 3 Information Criteria (e.g., AIC, BIC, etc.)

**Pros:** Simple to compute with an intuitive form!

**Cons:** Ad hoc—with questionable statistical properties.

## 4 “Default Bayes Factors”

**Pros:** Derive a proper prior dist'n based on training sample.

**Cons:** Result depends on the choice of training sample.

# Mitigating sensitivity of Bayes Factors to choice of prior

- 1 Intrinsic or Default Bayes Factors<sup>5</sup>.
  - ★ Use the minimal data subset resulting in a proper posterior.
  - ★ Use this as the prior when computing the Bayes Factor.
  - ★ Average over all possible minimal subsets.
- 2 Expected-posterior prior distributions<sup>6</sup>.
  - ★ Devise a single non-informative predictive dist'n  $p^*(y)$ .
  - ★ Compute the corresponding self-consistent proper priors

$$p(\theta) \propto \int p^*(\theta|y^*)p^*(y^*)dy^* \quad \text{with} \quad p^*(\theta|y^*) \propto p(y^*|\theta)p_0(\theta)$$

Think of  $y^*$  as training sample to obtain proper prior, but average over  $y^*$ .

- 3 In my experience astronomers will use subjective priors... as long as they don't have to be too precise about them.

---

<sup>5</sup>Berger and Pericchi, 1996, JASA

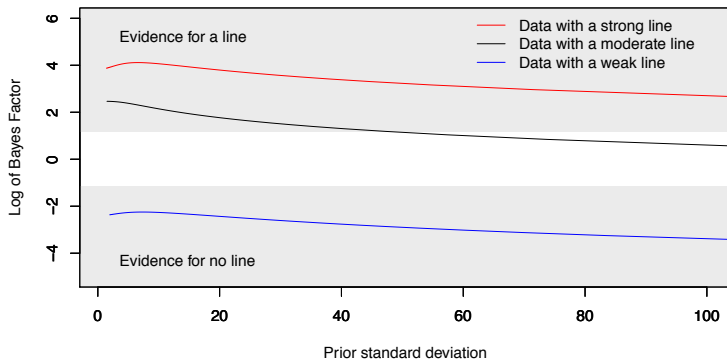
<sup>6</sup>Perez and Berger, 2002, Biometrika



## Spectral Analysis in High-Energy Astrophysics:

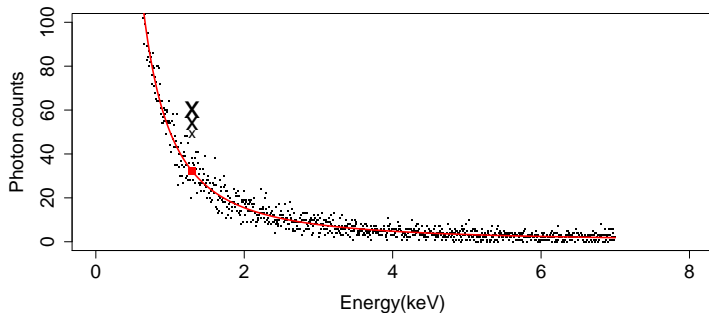
$H_0$ :  $Y_i \sim \text{POISSON}(\alpha E_i^{-\beta})$  with  $(\alpha, \beta)$  fixed.

$H_A$ :  $Y_i \sim \text{POISSON}(\alpha E_i^{-\beta} + \omega I_{\{i=\mu\}})$  with  $(\alpha, \beta, \mu)$  fixed.



*Decision may be the same over a range of reasonable priors.*

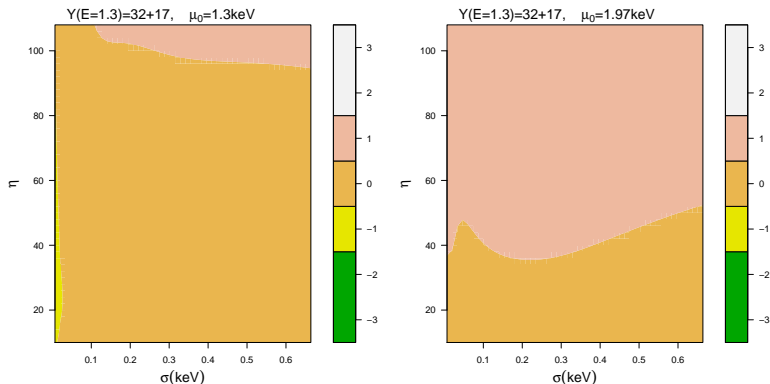
- Fix  $\alpha$  and  $\beta$  throughout
- The “true” emission line is at  $\mu = \mathbf{1.3}$  keV.
- The intensity from the continuum in this bin is **32**.
- We control the strength of support for  $H_A$  by altering the observed counts at 1.3keV.



# Results: A weak spectral line

*Prior on spectral line:*  $\omega \sim U(0, \eta)$  and  $\mu \sim N(\mu_0, \sigma^2)$ .

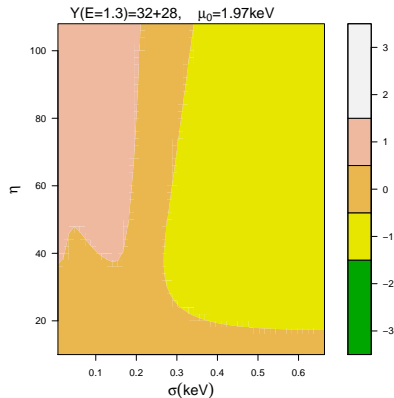
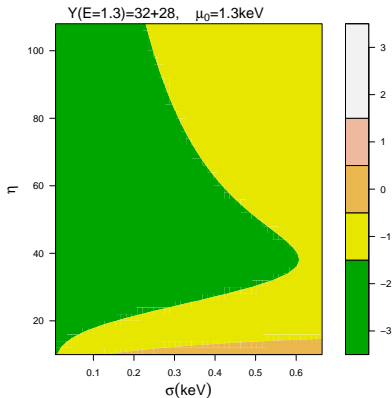
- We vary all three hyper parameters.



*Diffuse or misplaced priors weaken evidence*

# Results: A strong spectral line

Prior on spectral line:  $\omega \sim U(0, \eta)$  and  $\mu \sim N(\mu_0, \sigma^2)$ .



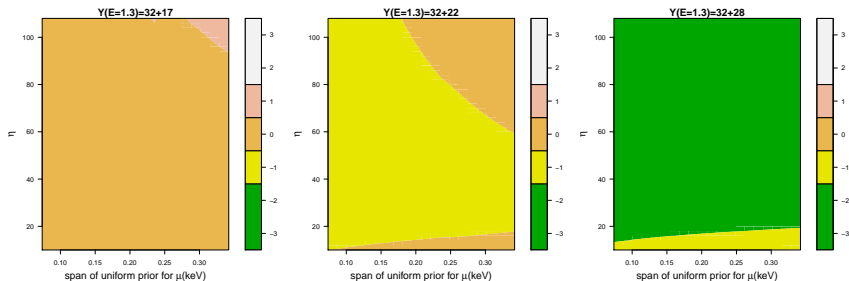
Under equal prior odds:

Yellow to Green transition  $\Pr(H_0|y) = 0.032$

transition to yellow:  $\Pr(H_0|y) = 0.250$

# Results: Using Astronomer's Priors

*Prior on spectral line:  $\omega \sim U(0, \eta)$  and  $\mu \sim U(1.3 \pm \kappa)$ .*

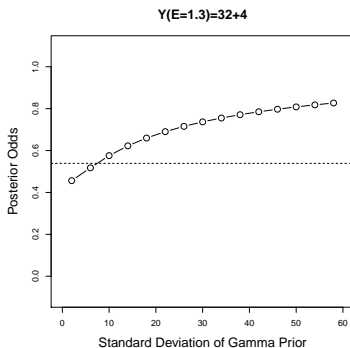
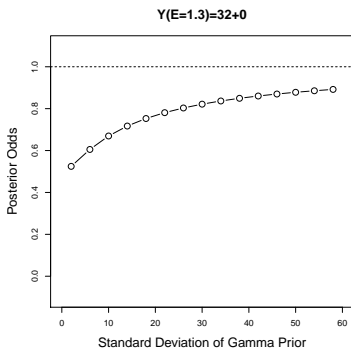


ppp-values (based on 1000 MC samples)

$Y(E = 1.3)$	32 + 17	32 + 22	32 + 28
$H_A$ : known line location	0.008	0.002	0.000
$H_A$ : fitted line location (0.5 – 7.0keV)	0.539	0.184	0.006

# Results: Comparing Bayes Factors with P-values

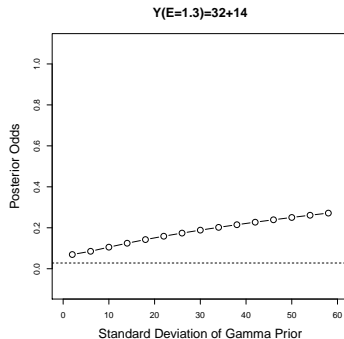
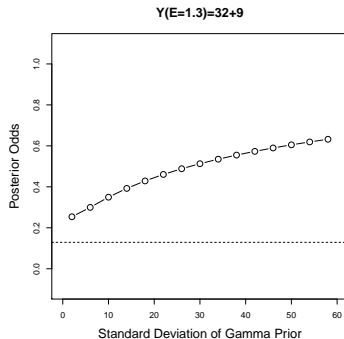
*Prior on spectral line:*  $\omega \sim \text{GAMMA}(\text{correct mode, sd}), \mu$  fixed  
*Prior odds: 50-50.* *ppp-value: dotted line*



*In left plot Bayes Factor  $\geq 1$ .*

# Results: Comparing Bayes Factors with P-values

*Prior on spectral line:*  $\omega \sim \text{GAMMA}(\text{correct mode, sd}), \mu$  fixed  
*Prior odds: 50-50.* *ppp-value: dotted line*



*Evidence decreases with more diffuse priors.  
 Bayes Factors are more conservative.*

# Results: Comparing Bayes Factors with P-values

Prior on spectral line:  $\omega \sim U(0, \eta)$  and  $\mu \sim U(1.3 \pm \kappa)$ .

$H_A$ : known line location

- ppp-value= 0.002

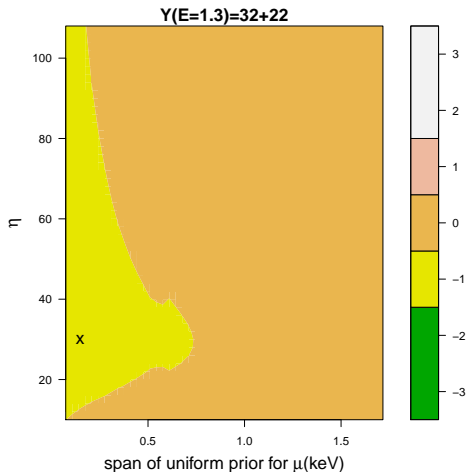
$H_A$ : UNknown line location

- ppp-value=0.184

minimum Bayes Factor = 0.044  
(span = 0.07,  $\eta = 30$ )

Prior on  $\mu$

- *let's us decide where to look,*
- *penalizes us for looking too many places.*





# The Bottom Line

- 1 It is hard to justify p-values for model selection.

*We feel that the correct interpretation of a P-value, although perhaps objective, is nearly meaningless, and that the actual meaning usually ascribed to a P-value by practitioners contains hidden and extreme bias.  
— J. Berger and M. Delampady (Stat Sci., 1987).*

- 2 Bayes Factors are *highly* dependent on choice of prior.

*Bayesians address the question everyone is interested in by using assumptions no one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone. — L. Lyons.*

# The Bottom Line

- 1 At least the Bayesian can clearly identify the assumptions.
- 2 So... I prefer Bayes Factors—but with:
  - 1 Careful choice of prior distribution.
  - 2 Clearly identified prior distribution.
  - 3 Comprehensive analysis of sensitivity to prior.
- 3 If no informative prior is available, identify classes of prior distribution that lead to one choice or the other.

*As Always: Try several methods and  
compare results!!!*

# Bayes for High Energy Physics

## Bayes Factors for Detection:

$$\text{Bayes Factor} = \frac{p_0(y)}{p_A(y)} = \frac{\int p(y|\theta, \mu = 0)p(\theta)d\theta}{\int p(y|\theta, \mu, m)p(\theta, \mu, m)d\theta d\mu dm}.$$

- Looking for lines at “all” masses simultaneously.
- This accounts for the LEE automatically.
- Doesn't address exclusion or Higgs mass, just detection.
- Priors will be important

$\theta$ : Reference (improper?) priors, hierarchical?

$\mu$ : “Looking” in the range zero to one.

$m$ : Summarize where we look. (*Sorry Bob!!*)

# Bayes for High Energy Physics

## Using Priors to Quantify Search

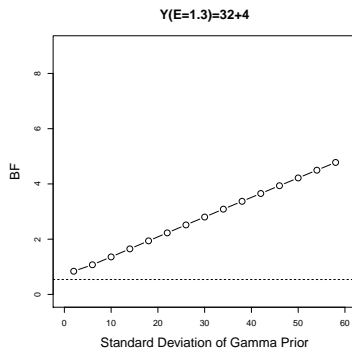
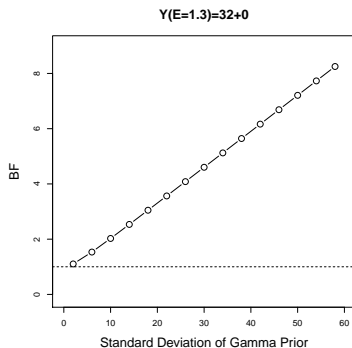
- The prior for mass acts qualitatively like a LEE correction.
- Can we quantify this?
  - ★ Analytic in simple problem? (Have analytic LEE correction.)
  - ★ Numerical results: Compare effect of prior on Bayes factor with choice of  $H_A$  on p-value and analytic LEE correction.

## Estimating the Mass

- Given detection, estimate  $m$  in a secondary analysis.
- In a mass-by-mass analysis we can use HPD for exclusion and mass estimation.
- A mass-by-mass analysis does not seem very Bayesian.
- In a full parameter space fit we can estimate the mass.
- Can we avoid penalizing masses with low sensitivity?

# Results: Comparing Bayes Factors with P-values

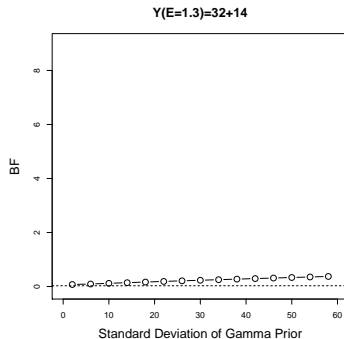
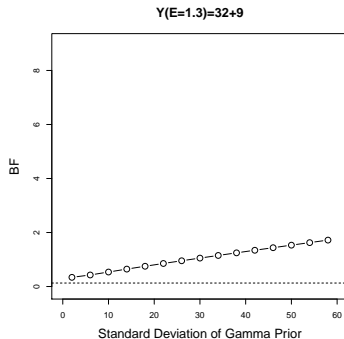
*Prior on spectral line:*  $\omega \sim \text{GAMMA}(\text{correct mode, sd}), \mu$  fixed  
*Prior odds: 50-50.* *ppp-value: dotted line*



*In left plot Bayes Factor  $\geq 1$ .*

# Results: Comparing Bayes Factors with P-values

*Prior on spectral line:  $\omega \sim \text{GAMMA}(\text{correct mode, sd}), \mu$  fixed*  
*Prior odds: 50-50. ppp-value: dotted line*



*Evidence decreases with more diffuse priors.*  
*Bayes Factors are more conservative.*