Making peace with p's: Bayesian tests with straightforward frequentist properties

Ken Rice, Department of Biostatistics

April 6, 2011

Biowhat?

Biostatistics is the application of statistics to topics in biomedical science. UW Biostat is part of the School of Public Health.

- We interpret ‘biomedical’ broadly; *I* work in cardiovascular genetics, my colleagues are experts in clinical trials, environmental health, infectious diseases, health services...
- We are *consistently* ranked the #1 Biostatistics department in the US*
- *Many* outstanding statisticians; NAS members, IoM advisors, an FRSNZ, one (Dutch) knight, an army of ASA fellows

Today’s topic is more ‘stat’ than ‘bio’ – but matters, for high-volume studies of small effects.

* We *may* also be the US department most aware of the *shortcomings* of rank-based analysis

Overview

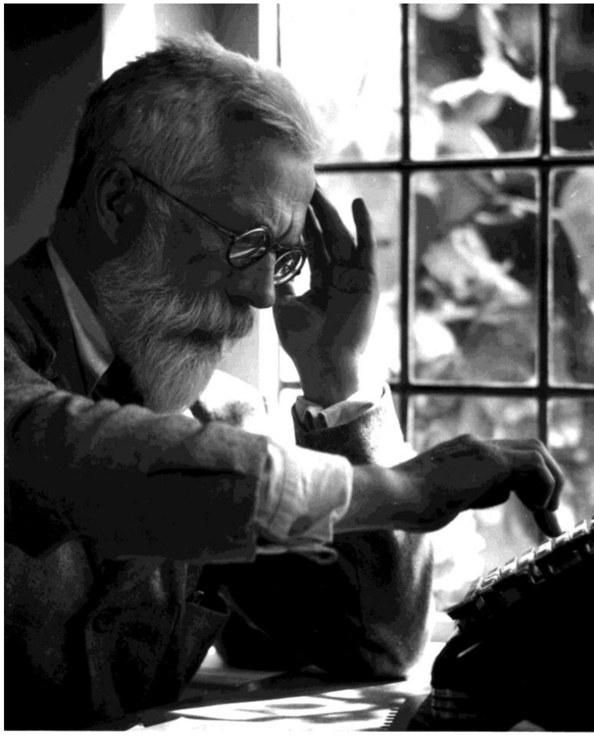
Biostatistics... "with the p's and the t's" ?

Today I will discuss;

- Testing, as Fisher saw it
- Bayes – making decisions
- Bayes – making testing decisions
- Some extensions

All of this is (surprisingly) contentious – but perhaps it doesn't need to be.

What is a Fisherian test?



Ronald Fisher
(1890–1962)

44 Storey's Way
(1943–1957)

Every experiment may be said to exist
only in order to give the facts a chance
of disproving the null hypothesis

The Design of Experiments, pg 18

What is a Fisherian test?

Fisher developed tests that choose between;

- h=1: Reject the null hypothesis
- h=0: Conclude nothing

This is **different** to Neyman-Pearson style tests;*

- h=1: Reject the null hypothesis
- h=0: Accept the null hypothesis

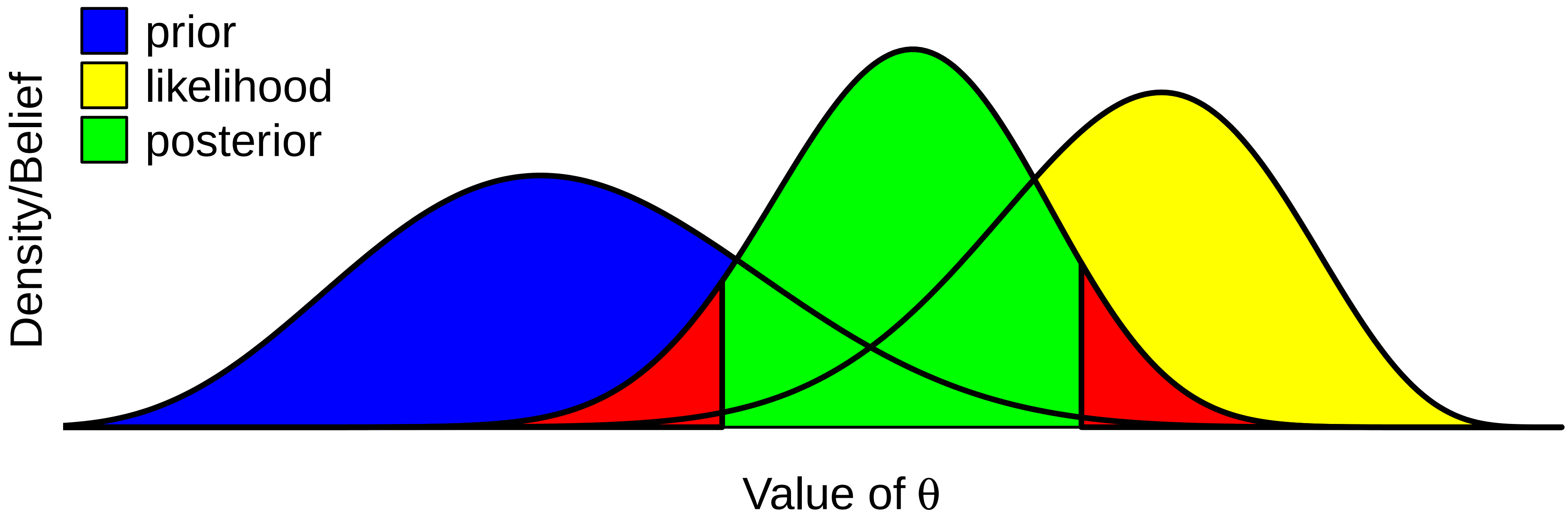
Type I errors can occur in both forms; any test that sets h=1 when $p < \alpha$ fixes the Type I error rate (frequentist)

Type II errors **do not occur** in the Fisherian approach.

* For fun, see Hurlbert & Lombardi (2009) *Ann Zool Fennici* Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian

Bayesian decisions

Bayes' theorem: posterior $\propto$ prior $\times$ likelihood...



Common sense reduced to calculus

Laplace

Bayesian: One who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule

Stephen Senn

Bayesian decisions

Bayes' theorem: posterior $\propto$ prior $\times$ likelihood...

Common sense reduced to calculus

Laplace

Bayesian: One who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule

Stephen Senn

Bayesian decisions

Based on deep results, Bayesian decision theory says we should make decisions that minimize loss averaged over the posterior. This decision is the **Bayes rule**.

The **loss function** specifies how *bad* it is, if our decision is d but the true state of nature is θ. For $\theta \in \mathbb{R}$;

- $L = (\theta - d)^2$: quadratic loss; decide $d = \mathbb{E}[\,\theta|Y\,]$, the posterior mean
- $L = |\theta - d|$: absolute loss; decide d = posterior median
- $L = h\mathbf{1}_{\theta=\theta_0} + (1-h)\mathbf{1}_{\theta\neq\theta_0}$: classic Bayesian testing;

$$h = \begin{cases} 0, & \mathbb{P}[\,\theta = \theta_0\,] > 0.5 \\ 1, & \mathbb{P}[\,\theta = \theta_0\,] < 0.5 \end{cases}$$

Classic Bayesian tests offer NP-style choices; θ_0 or θ_0^C

Bayesian decisions

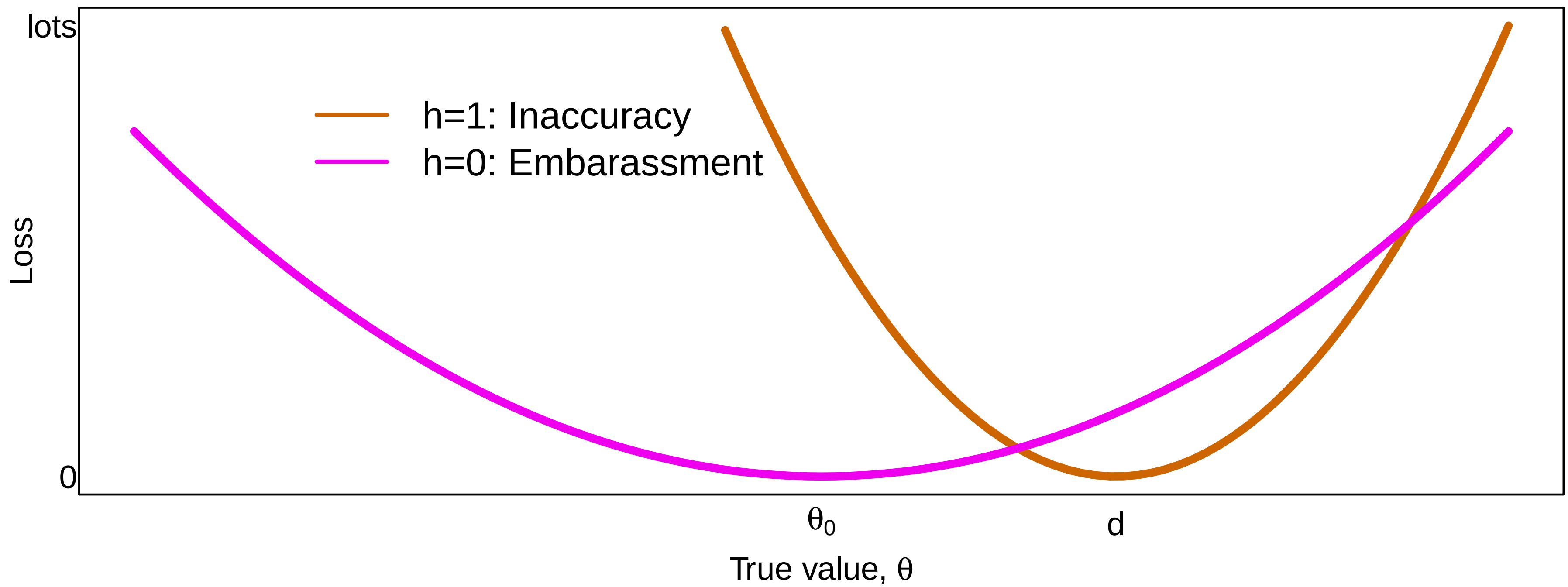
But how might a Bayesian be Fisherian – rejecting the null, or concluding nothing? One way is to decide between;

- **Inaccuracy**
  - make an estimate, which may be badly 'off'
  - $(\theta - d)^2$
- **Embarrassment**
  - 'conclude nothing', which is bad if you miss an exciting signal
  - $(\theta - \theta_0)^2$

$$L_\gamma = \underbrace{(1-h) \times \gamma^{1/2}(\theta - \theta_0)^2}_{\propto \text{ embarrassment}} + \underbrace{h \times \gamma^{-1/2}(\theta - d)^2}_{\propto \text{ inaccuracy}}$$

Bayesian decisions

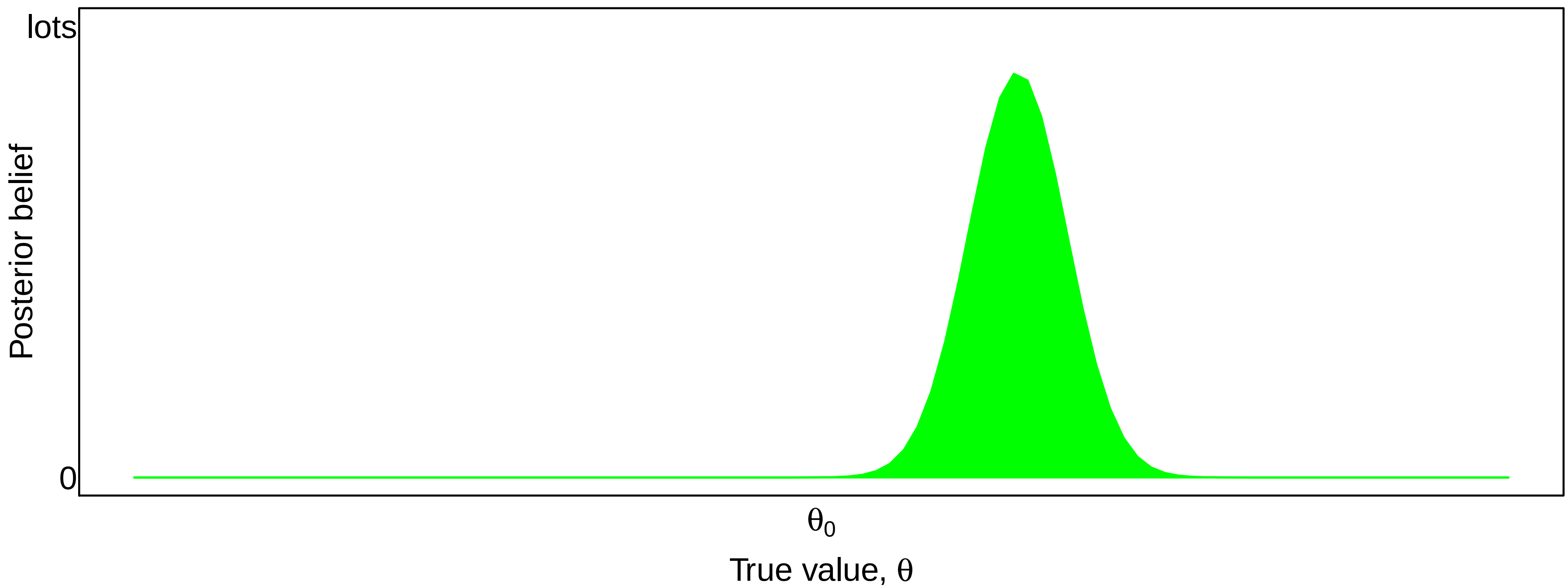
As a function of θ:



Inaccuracy is worse than embarrassment, so scale embarrassment by $0 \leq \gamma \leq 1$. Embarrassment is γ times cheaper than inaccuracy

Bayesian decisions

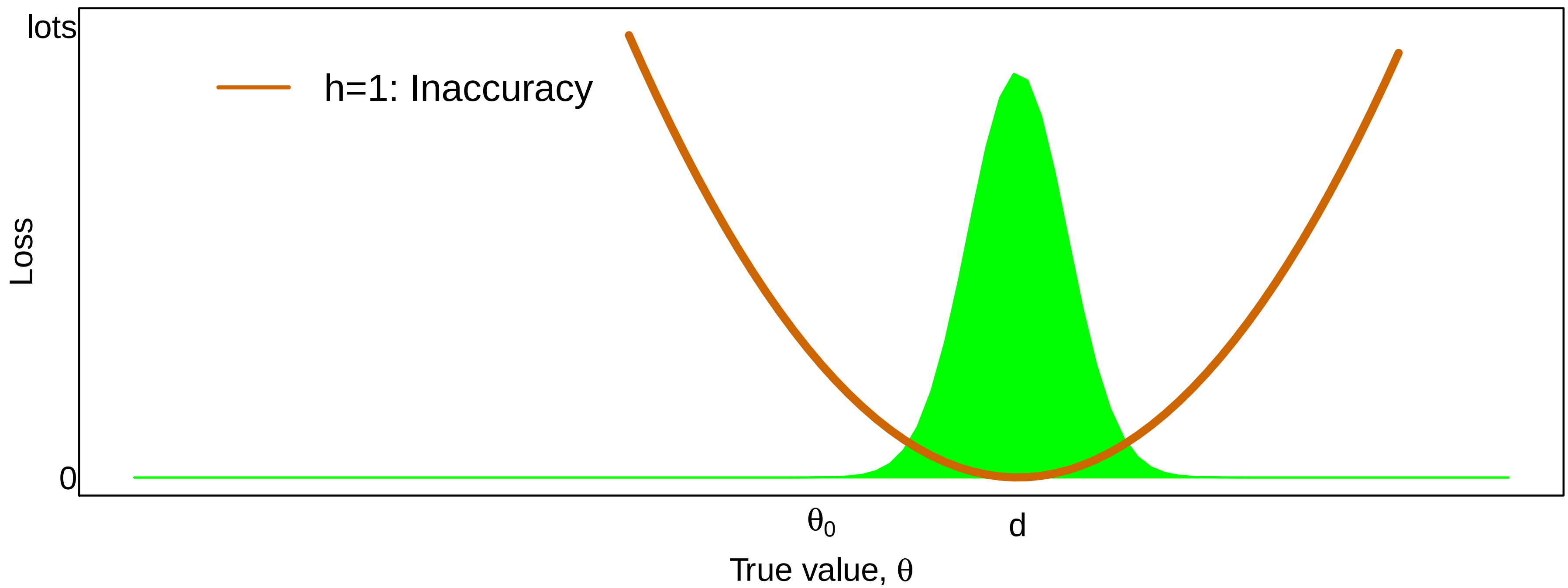
Let's try it, for a revolting green posterior distribution;



Beliefs are centered near θ_0, but also have some uncertainty

Bayesian decisions

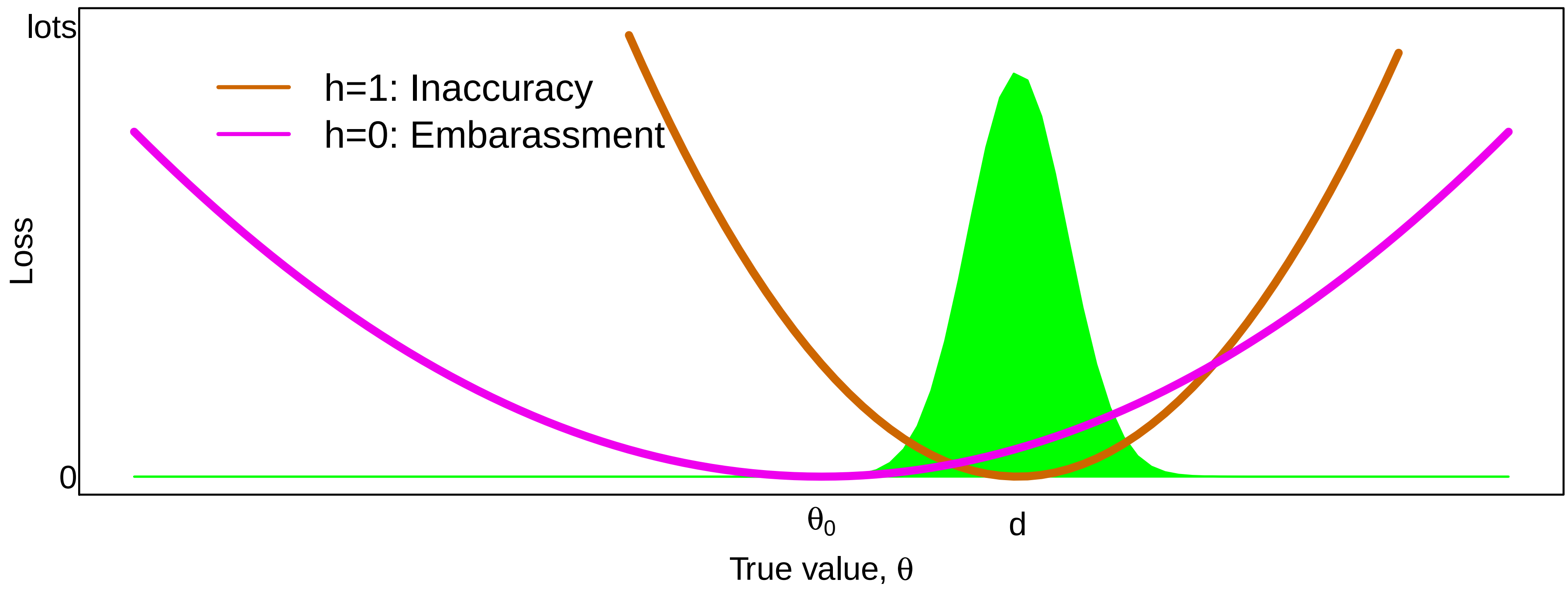
Let's try it, for a revolting green posterior distribution;



Choosing $h = 1$, we'd select the posterior mean, for d

Bayesian decisions

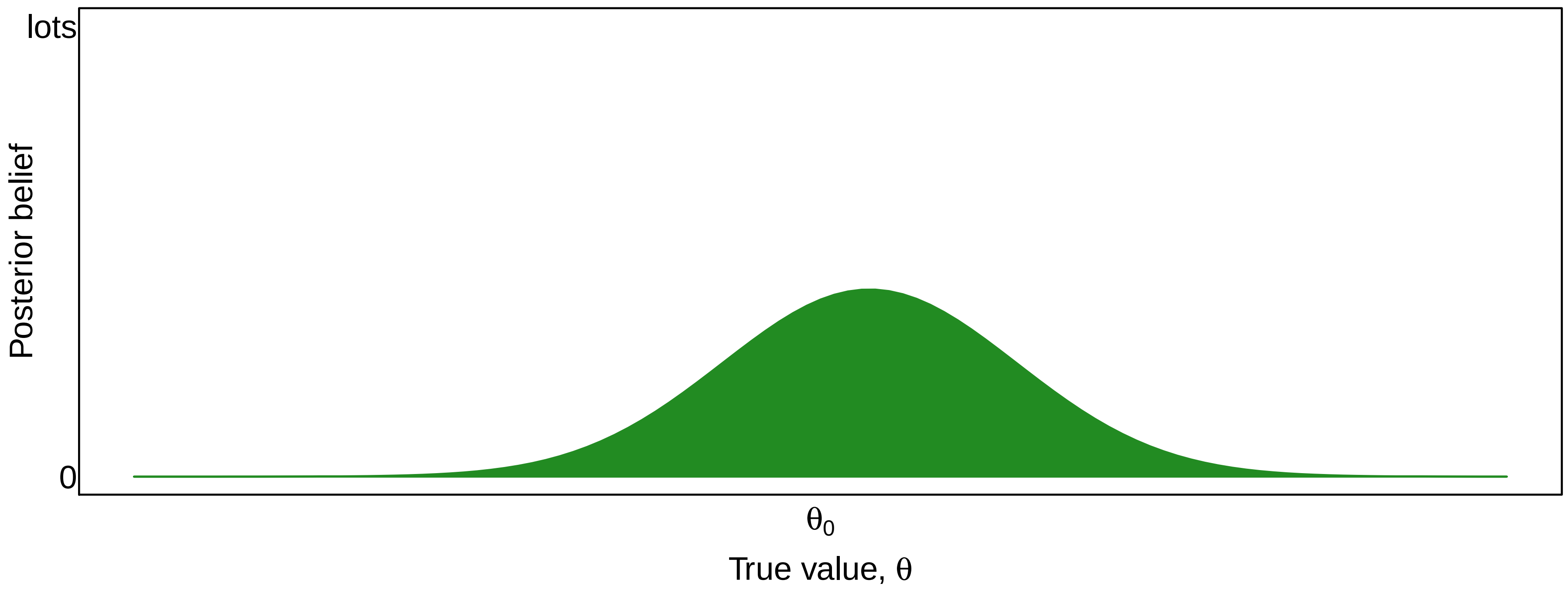
Let's try it, for a revolting green posterior distribution;



Looks better to choose $h = 1$, here

Bayesian decisions

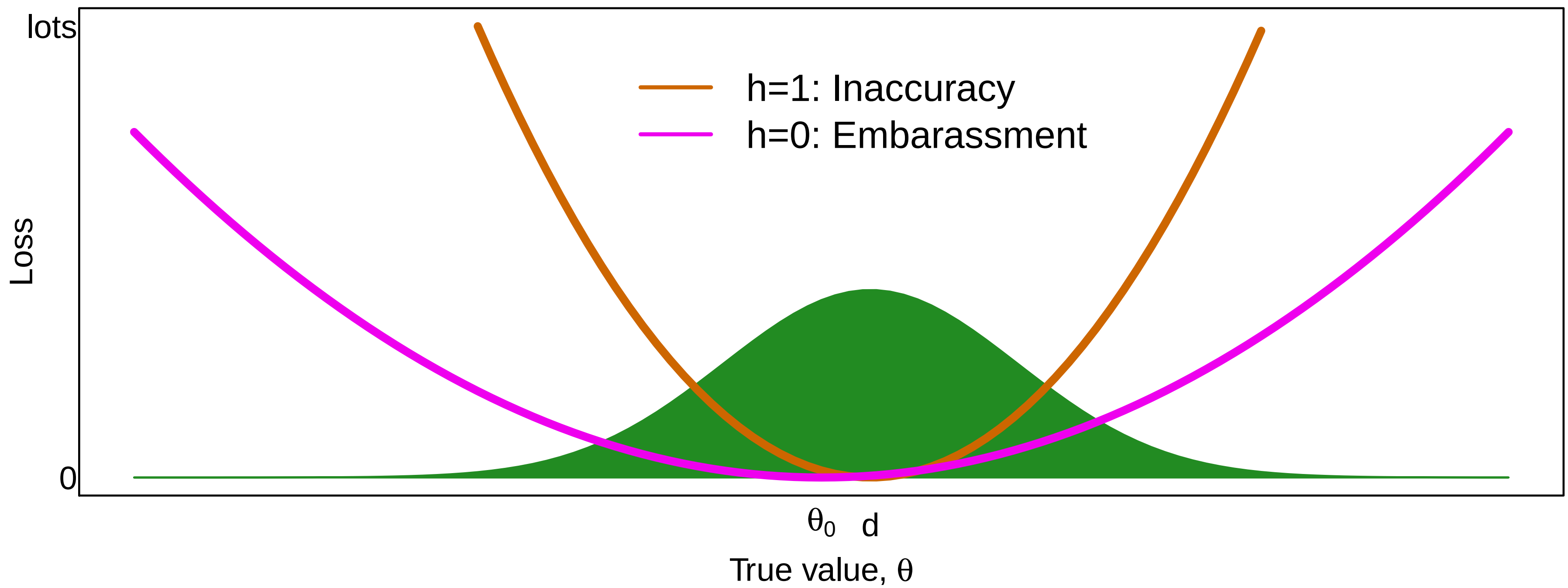
Another example;



This posterior is diffuse, with mean closer to θ_0

Bayesian decisions

Another example;



Here, we do better choosing $h = 0$

Bayesian decisions

We get the Bayes rule formally by minimizing a quadratic; we decide $h = 1$ (inaccuracy) iff

$$\boxed{\frac{\mathbb{E}[\theta - \theta_0 | Y]^2}{\text{Var}[\theta | Y]} \geq \frac{1-\gamma}{\gamma}}$$

- If $h = 1$, d is the posterior mean, $\mathbb{E}[\theta | Y]$ (may be inaccurate)
- If $h = 0$, any d is equally good/bad; we make **no conclusion** (embarrassing!)

Note that a non-committal decision is $\neq$ a non-committal prior/likelihood/posterior

On sanity

Scientifically, this loss is sane. Embarrassment and inaccuracy are measured on the same **scientifically relevant** scale

On sanity

Trading $h = 0, 1$ vs $(\theta - \theta_0)^2$? Apples *vs* oranges;

A PARADOX IN DECISION-THEORETIC INTERVAL ESTIMATION

George Casella, J. T. Gene Hwang and Christian Robert

Cornell University and Université Paris VI

Abstract: Decision-theoretic interval estimation usually employs a loss function that is a linear combination of volume and coverage probability. Such loss functions, however, may result in paradoxical behavior of Bayes rules. We investigate this paradox in the case of Student's t, and suggest ways of avoiding it using a different loss function. Some properties of the resulting Bayes rules are also examined. This alternative approach may also be generalized.

Connections

Moreover, this sane test shouldn't upset frequentists;

Bayes rule	Wald test
$$\frac{\mathbb{E}[\,\theta - \theta_0 \vert Y\,]^2}{\text{Var}[\,\theta \vert Y\,]} \geq \frac{1-\gamma}{\gamma}$$	$$\frac{(\hat{\theta} - \theta_0)^2}{\widehat{\text{Var}}\hat{\theta}} \geq \chi^2_{1,\,1-\alpha}$$

- Interpreting γ in terms of α is straightforward
- Justify your choice of γ! (but $\gamma = 0.21 \approx \alpha = 0.05$, if you *must*... $\gamma = 0.03$ for $\alpha = 5 \times 10^{-8}$)
- For 'nice' situations, by Bernstein-von Mises as $n \to \infty$ the posterior is essentially a Normal likelihood, and everyone agrees
- Classic Bayes Tests can give **opposite results** from Wald tests (the 'Jeffreys/Lindley paradox') – particularly for small θ and large n. With the 'new' tests, this **does not happen**

Example

An old genetics problem – testing Hardy-Weinberg Equilibrium;

Genotype	AA	Aa	aa	Total
Count	n_{AA}	n_{Aa}	n_{aa}	n
Proportion	p_{AA}	p_{Aa}	p_{aa}	1

Under *exact* HWE, for *some* p_A the proportions are

$$\{p_{AA}, p_{Aa}, p_{aa}\} = \{p_A^2, 2p_A(1-p_A), (1-p_A)^2\}$$

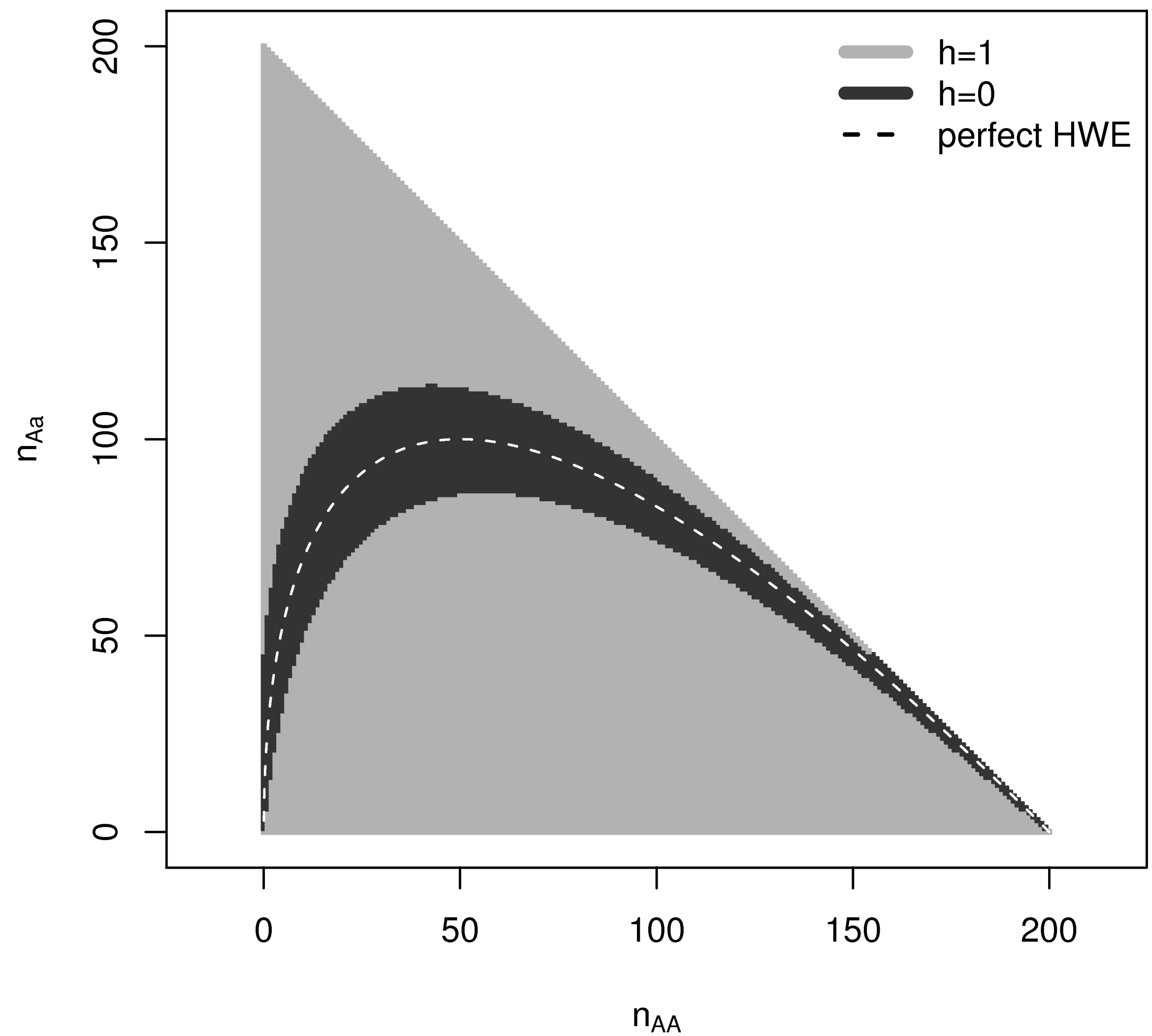
Deviations from HWE can measured by;

$$\theta = \frac{2(p_{aa}+p_{AA}) - 1 - (p_{aa}-p_{AA})^2}{1-(p_{aa}-p_{AA})^2}.$$

Under *exact* HWE, we get $\theta = \theta_0 = 0$. Using a flat prior on $\{p_{AA}, p_{Aa}, p_{aa}\}$, $\gamma = 0.21$, let's use the Bayesian test...

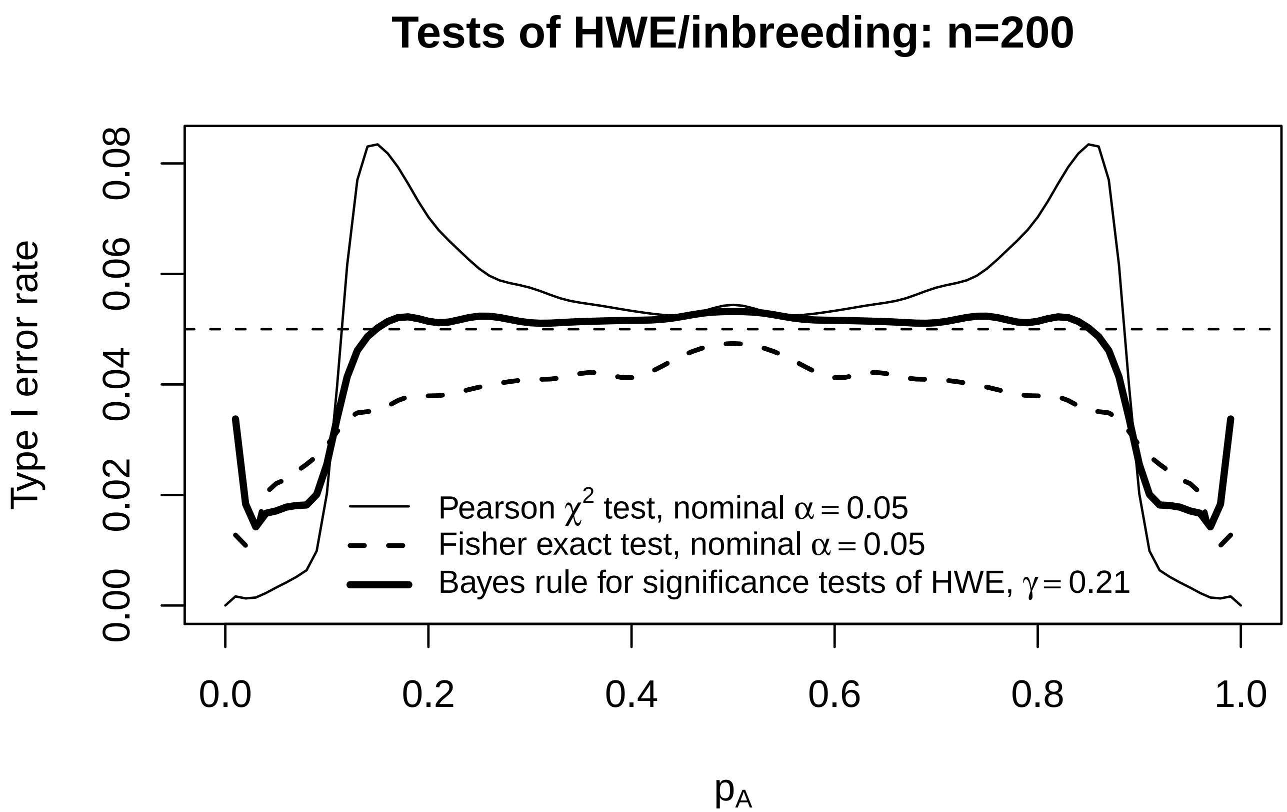
Example

All possible Bayesian answers, for n=200;



Example

Any Bayes test has frequentist properties – ours has *good* ones!

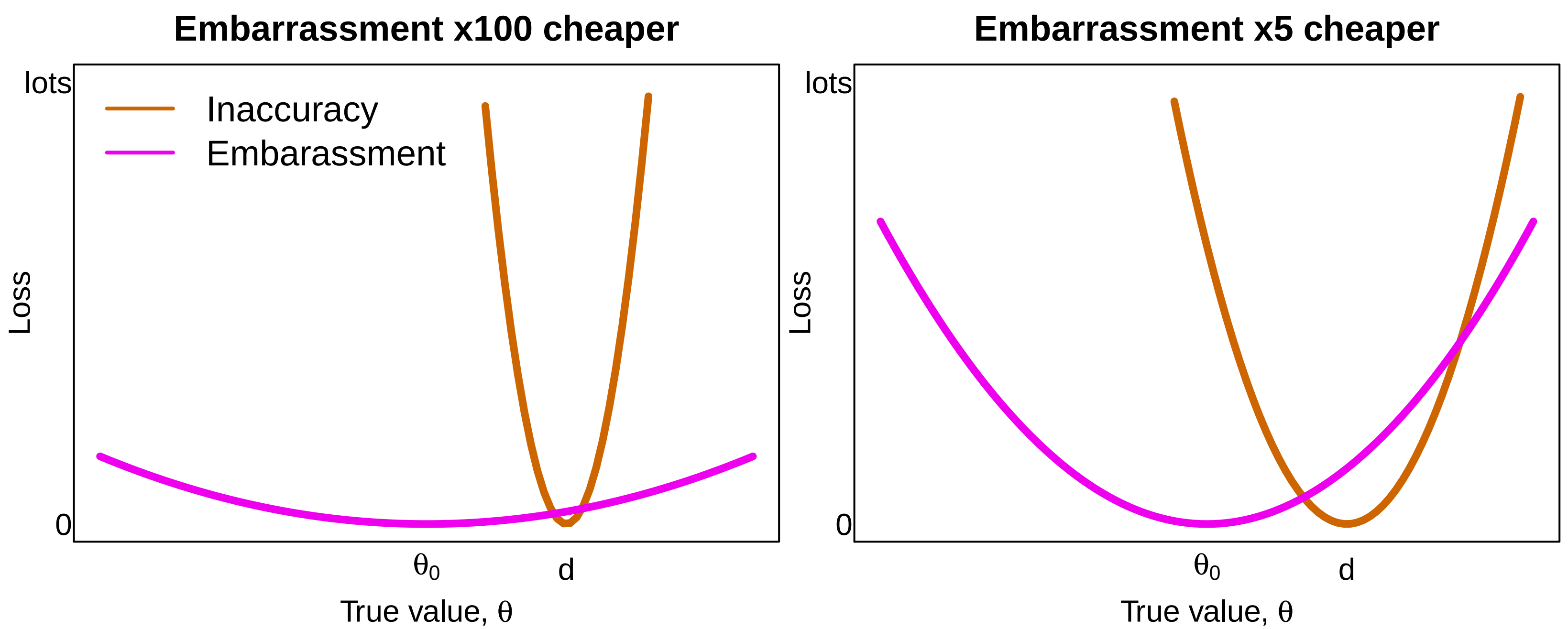


The other tests are;

- A simple Pearson χ^2 test, based on $(O - E)^2$
- Fisher's test (!), which is exact but conservative

A dual problem

A related problem; if you had to suffer **both** embarrassment and inaccuracy – which tradeoff would you choose?



This 'dual' decision problem has loss function

$$L = \frac{1}{\sqrt{1+w}}(\theta - \theta_0)^2 + \sqrt{1+w}\,(d-\theta)^2,$$

for positive decision w, which parameterizes the tradeoff.

A dual problem

The Bayes rule looks familiar;

$$w = \frac{\mathbb{E}[\theta - \theta_0|Y]^2}{\text{Var}[\theta|Y]} \quad \approx \quad \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\text{Var}}\hat{\theta}}.$$

- The Bayes rule **is** the Wald statistic, modulo the prior's influence
- Two-sided p-values are essentially (sane) Bayesian decisions
- Making decision $\{d, w\}$ lets *others* make testing $\{h, d\}$ decision, for *any* tradeoff γ – a *complementary* problem
- Viewed as Bayesian or frequentist, p **does not** measure evidence in favor of $H_0 : \theta = \theta_0$;
  - Neither p nor w represents $\mathbb{P}[\theta = \theta_0]$ – we can give zero support to $\theta = \theta_0$ and still decide $h = 0$.
  - It's *known* that p *alone* behaves **un**like any sane measure of evidence (Schervish 1996)

Interim conclusions

Big points so far;

- Two-sided p values are **not evil, or unBayesian**
- Bayesian analysis can be Fisherian, without difficulty

Also;

- Getting $p < \alpha$ is not 'proof' of anything. Fisherian approaches make this obvious
- The (abstract) concept of repeated sampling is unhelpfully confusing. Embarrassment and inaccuracy make sense with regard to one dataset
- Calibration of anything is hard. Expressing loss in units of θ connects 'the statistics' with 'the science'

Interim conclusions

There are several extensions to this work;

- Multivariate $\boldsymbol{\theta}$
- Shrinkage
- Model-robust inference, ‘sandwich’ approaches
- Set-valued decisions
- Point masses at $\theta = \theta_0$
- Simpler measures of embarrassment and inaccuracy
  – using only sign($\theta - \theta_0$)

Other extensions include multiple testing (Bonferroni, FDR)

Final Conclusions

- *If* you want to do tests, this framework is attractive. But **not doing tests at all** is also reasonable, if your loss looks nothing like those seen here

- Many of the results we teach as ps and ts are **better** justified as Bayesian procedures. The Bayesian version is [I think] easier to motivate and understand – and criticize, when it's used inappropriately

- If methods are chosen because they are 'cookbook', justification as Bayes and/or frequentist doesn't matter. But this choice *shouldn't* be cookbook

Final Conclusions

Thanks to;

- Dane for the invite
- Adam Szpiro
- Thomas Lumley (Auckland)
- Jim Berger and SAMSI (initial work)

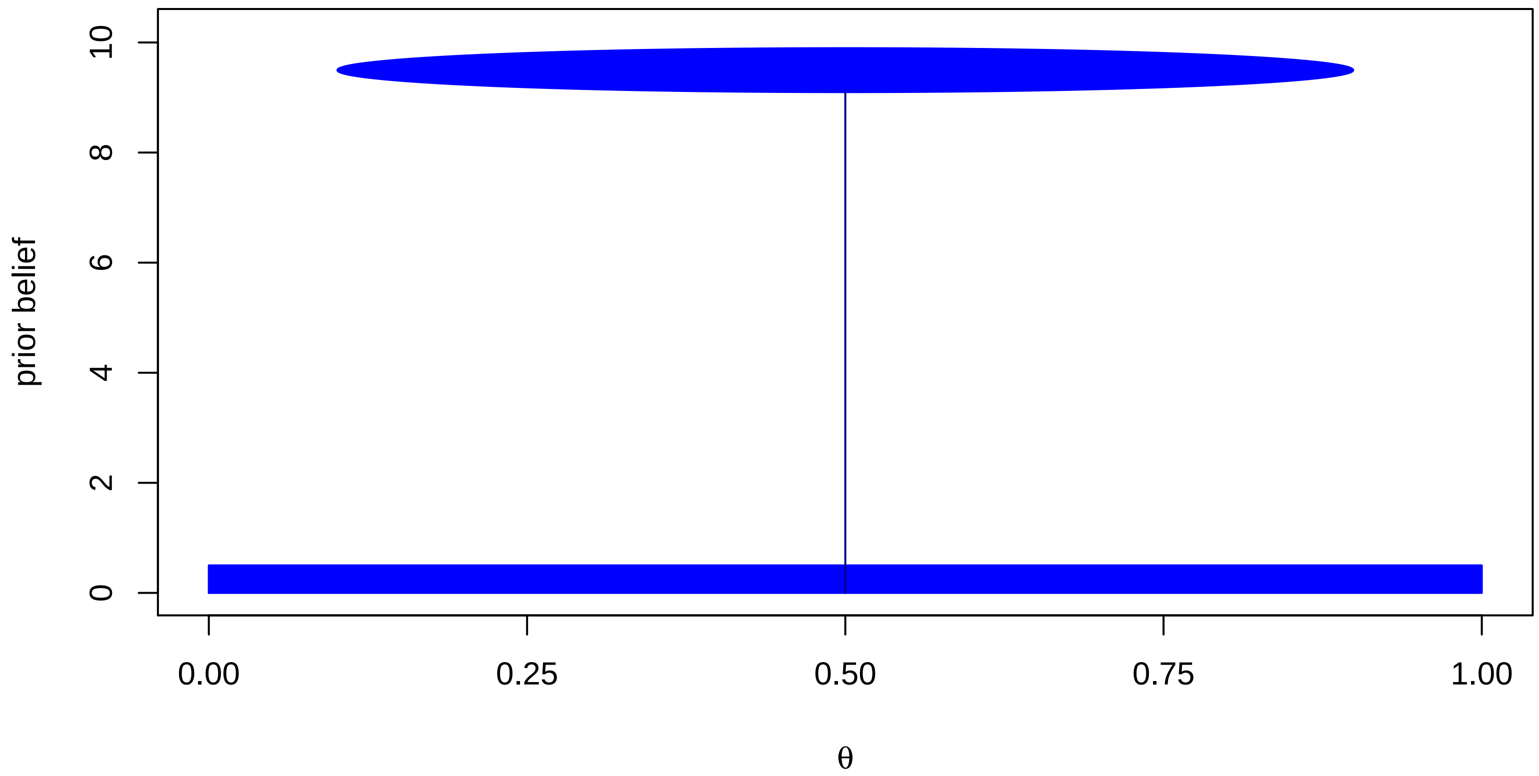
References:

- Rice (2010) A Decision-Theoretic Formulation of Fisher's Approach to Testing, *American Statistician*
- Szpiro, Rice, and Lumley (2011) Model-Robust Regression and a Bayesian 'Sandwich' Estimator *Annals of Applied Statistics*

faculty.washington.edu/kenrice

Bonus Tracks: Lindley's what?

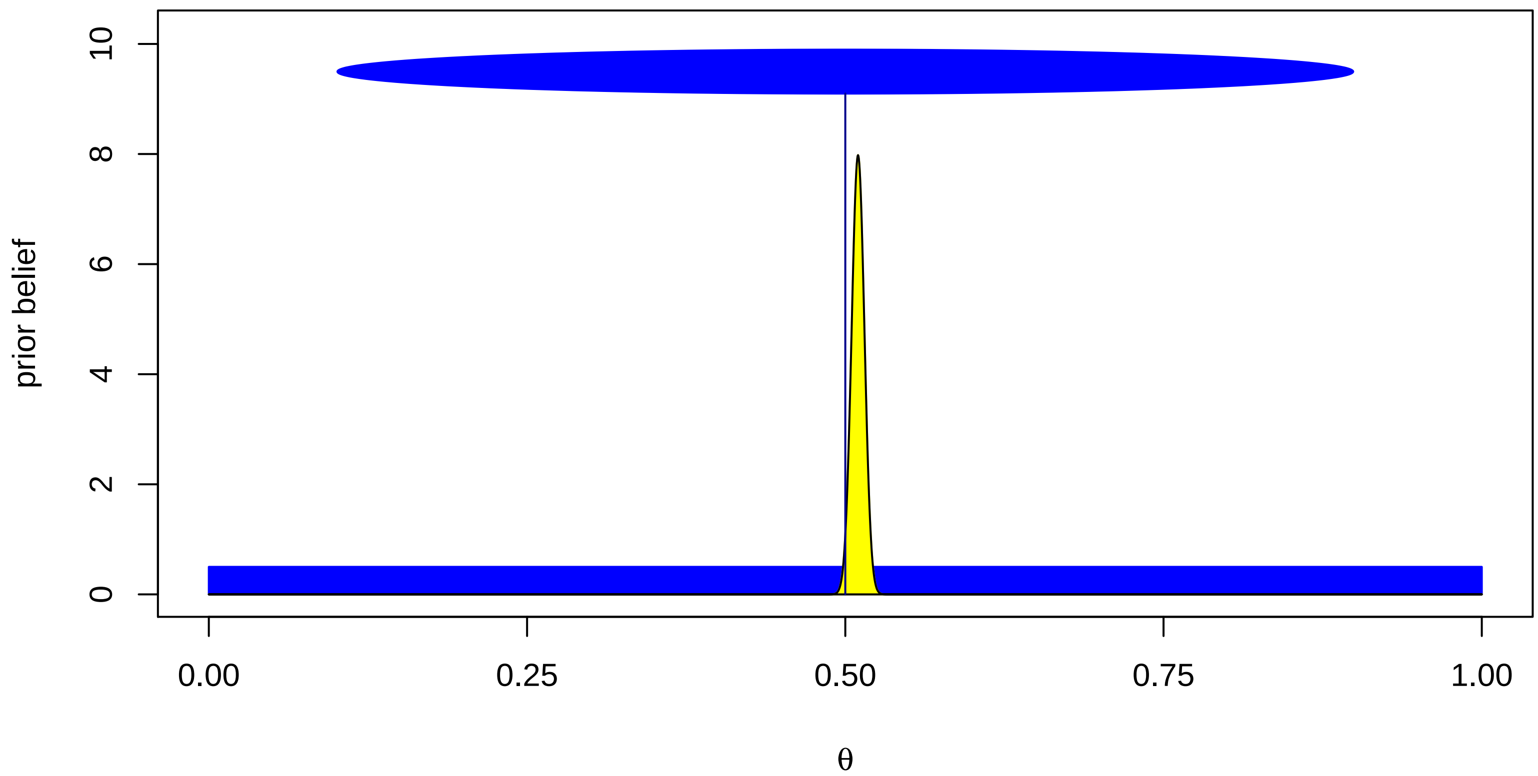
Some Bayesians *hate* p-values – they often have priors like this;



Blue ellipse 'concentrates' at *exactly* $\theta = 1/2$; otherwise diffuse

Bonus Tracks: Lindley's what?

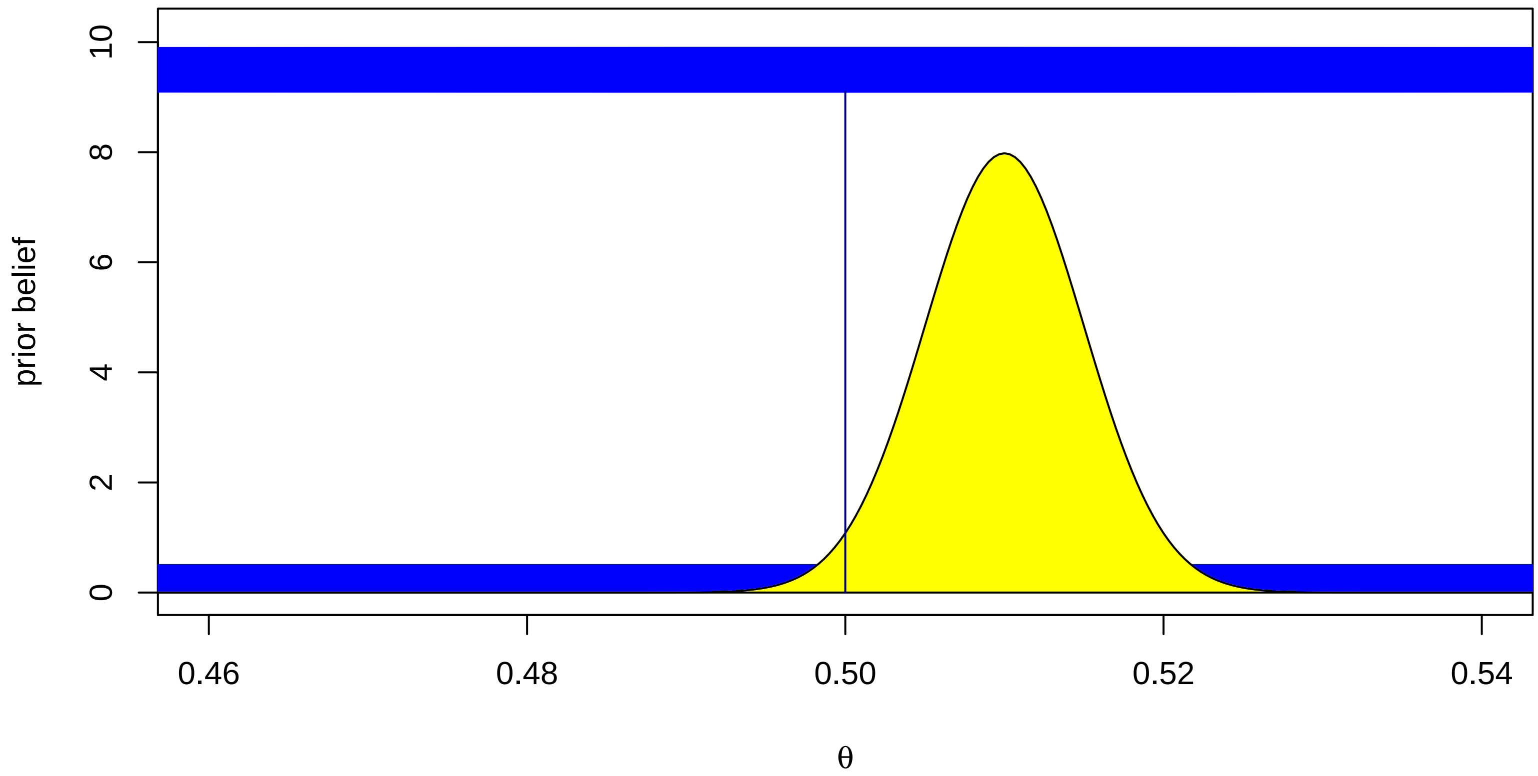
You do a massive study, and get e.g. 51% heads in 10,000 tries;



51% is hard to see, plotted on this scale - let's zoom in;

Bonus Tracks: Lindley's what?

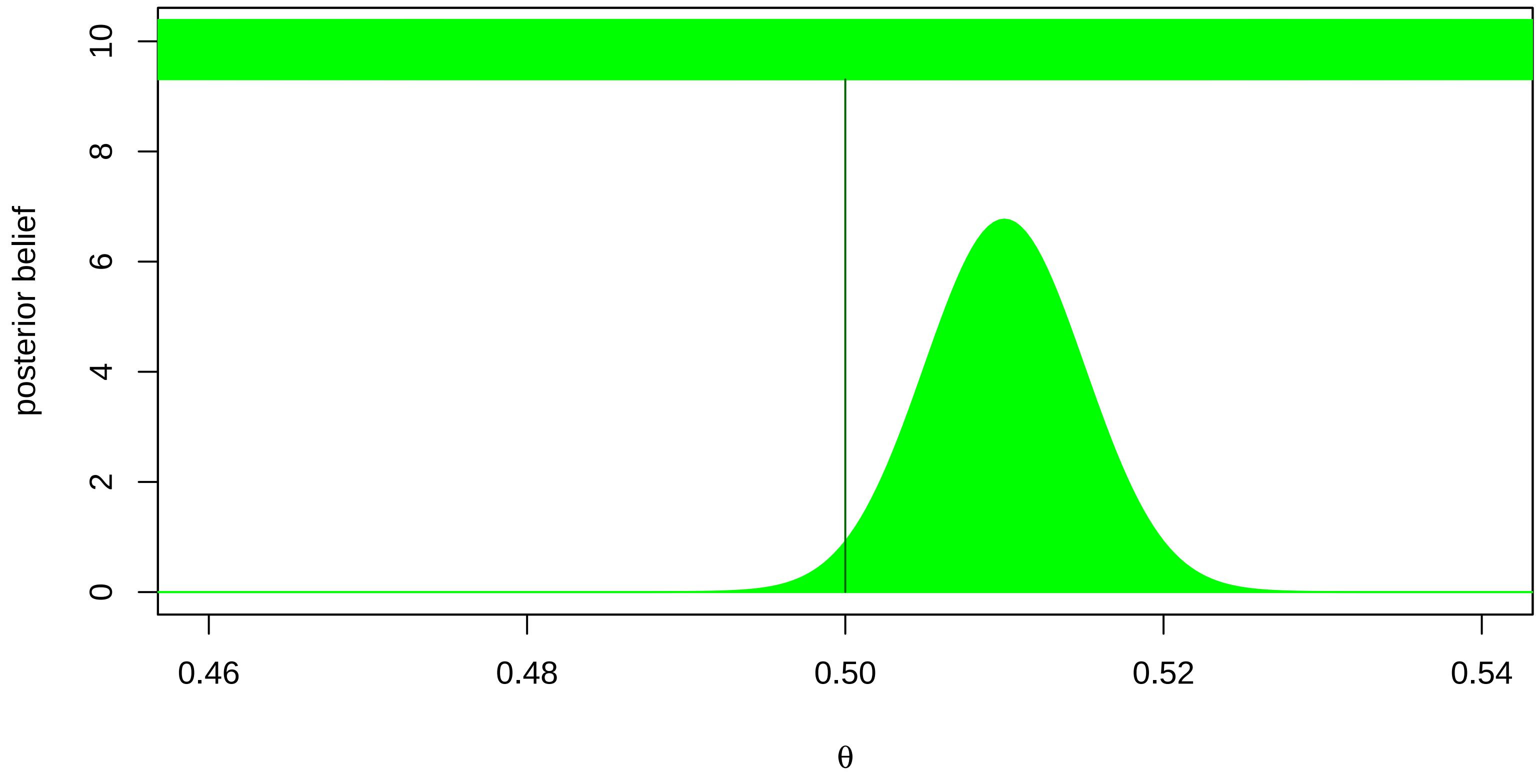
You do a massive study, and get e.g. 51% heads in 10,000 tries;



Wald test rejects ($p < 0.05$, no prior) but small effect estimate

Bonus Tracks: Lindley's what?

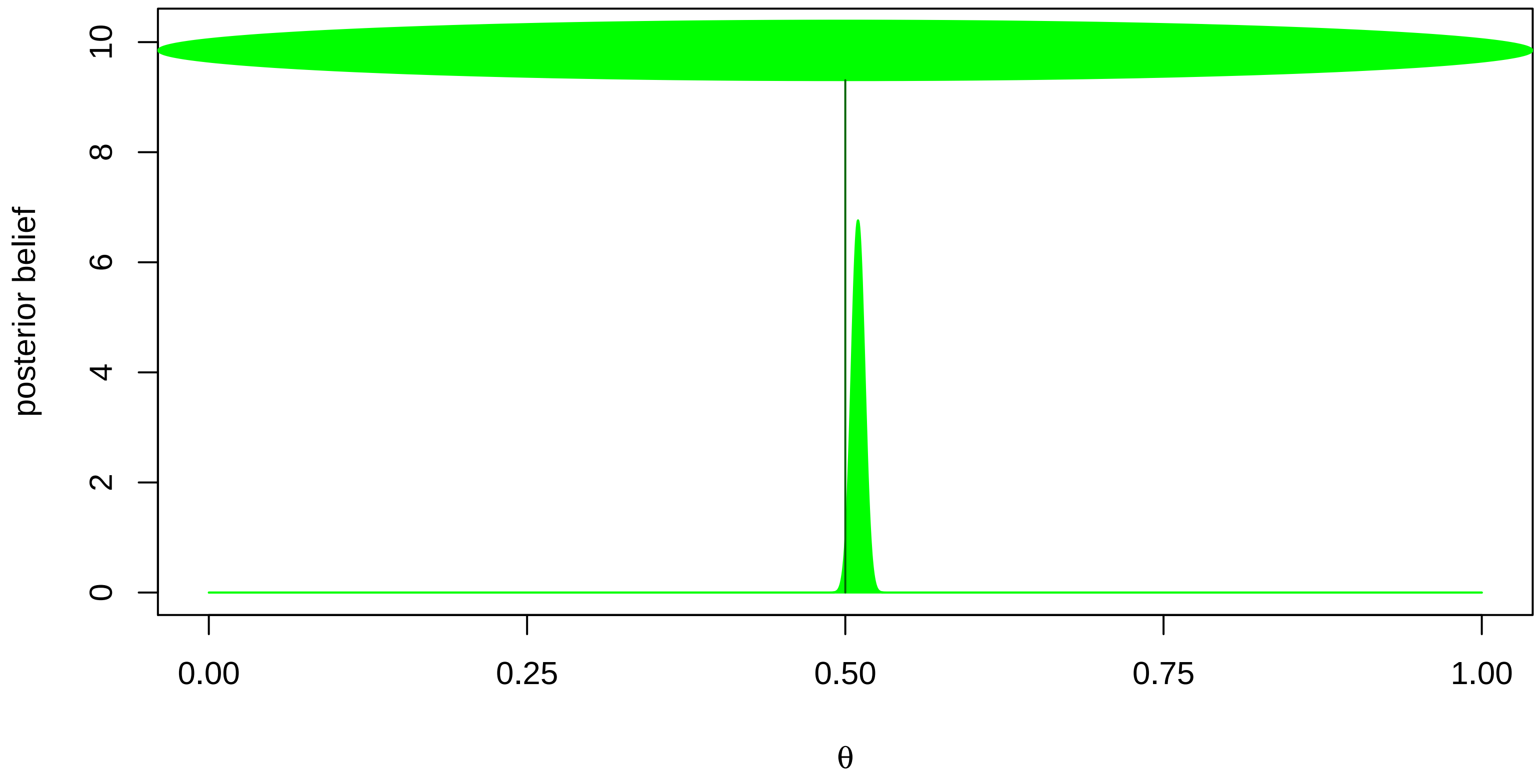
Zoomed-in revolting green posterior; (prior $\times$ likelihood)



Now let's zoom out, for the big picture...

Bonus Tracks: Lindley's what?

Bigger ellipse $\Rightarrow$ Bayesian Taleban believe $\theta = 0.5$ **more strongly**



But the Wald test **rejects** $\theta = 0.5$ (?) – *for unpointy priors*

Bonus Tracks: Lindley's what?

This phenomenon is called the **Jeffreys/Lindley paradox**

- Jeffreys spotted it, Lindley made it famous
- Our prior had 50:50 support for null, alternative – but this doesn't matter; classic Bayes tests use how much *more* we believe the null (a.k.a. the Bayes factor)
- With point null priors, we can still trade embarassment for inaccuracy, but the 'balance' in the prior *does* matter (seems sensible to me!)
- In my experience, a lot else can go wrong with 'pointy' priors like this, and they are not 'real'. But some Bayesians really like them.