

Principled Bayesian Workflow: Practicing Safe Bayes.

Keith O'Rourke, Health Canada

August 2, 2019

So now you can Bayes - need to do it safely (especially in a regulatory environment).

- ▶ Motivated by courses I gave in Health Canada in 2010/11 - “resulted” in less than critical review of submissions.
- ▶ Today is largely to provide motivation to read and work through material/programs developed by Michael Betancourt on Principled Bayesian Workflow.
- ▶ GitHub for Rstan LINK
- ▶ GitHub for Pystan LINK
- ▶ I'll do a largely conceptual introduction and work through a “toyed down” example. Files on Github LINK.

My relevant background.

- ▶ Provided statistical support and mentor-ship to research fellows at the University of Toronto and Toronto Hospital (1985-1998).
- ▶ Had no degree in statistics but still taught one course and gave multiple tutorials.
- ▶ Was recruited to the Ottawa Hospital (1998-2001).
- ▶ Did a DPhil in Statistics at Oxford (2001-2007).
- ▶ Visited the Statistical Sciences Department at Duke University, *the world's leading center for Bayesian statistics* (2007-2008).

My relevant background (cont.)

- ▶ Joined Health Canada 2009, gave Bayesian Courses in 2010/11.
- ▶ Currently at Pest Management Regulatory Agency.
- ▶ Gave webinars on Bayes for the American Statistical Society and CSEB in 2012 (first reviewed positively, not the second).
- ▶ Currently an author on Statistical Modeling, Causal Inference, and Social Science LINK a mostly Bayesian perspective blog.

Some of my early experience in Bayesian work.

- ▶ Used someone's published Bayesian model - obtained positive posterior probabilities for negative proportions.
- ▶ The priors they had specified were independent even though proportions were dependent.
- ▶ Contacted the author - they were unconcerned.
- ▶ Claimed it was a result of my not having adequate data and so no need to worry.
- ▶ Sometimes it is hard to know when you have adequate data but the prior is obviously wrong and *should* be fixed.
- ▶ (This sort of thing was be done a lot in early 2000s given limitations).

Conceptual review of Bayes.

Seeing all of Bayes all at once.

- ▶ Need to go back and focus on basics - not always popular.
- ▶ Once told that they hide the best techniques in martial arts in the beginners form.
- ▶ Two stage simulation (described later) is not a cheap trick but rather a way to *really, really* understand (I first came up with it in 2004). Told I must be wrong.
- ▶ Later learned Don Rubin (Harvard) came up with it 1984.

Conceptual review of Bayes (cont.)

Seeing all of Bayes all at once.

- ▶ Andrew Gelman blogged numerous times - in Bayes you write down the model and simply watch it work (crickets. . .).
- ▶ From a two stage sampling perspective it is obvious.
- ▶ But so much more!
- ▶ Or as Michael Betancourt put it: Work it, Make it, Do it, Makes Us Harder, Better, Faster, Stronger.

Conceptual review of Bayes (cont.)

- ▶ Numerous (wild?) claims about Bayes.
- ▶ For instance in an American Statistical Association webinar (2010/11).
- ▶ “No need to worry if you are unsure about the prior - lots of noninformative priors that work just great.”
- ▶ “... the term “noninformative” appears to be a Whorfian trick of language. Just as an “empty” gasoline drum is more dangerous than a “full” one (Whorf 1941), a noninformative prior may distort estimates more than would an informative prior.” (McElreath and Koster, 2014)
- ▶ **Take no one's word for it!**

Need to worry if you are unsure about the prior.

- ▶ One of the first I heard about was from Peter F. Thall at M.D. Anderson Cancer Center.
- ▶ Not obvious at the time perhaps given the lack of awareness of the dangers of (silly?) default flat (noninformative) priors.
- ▶ In 2011 I emailed him and got back *It seems that we are both in the How to tell if you have a prior with undesirable properties business. I got into this years ago when I was doing CRM dose finding with the model $\text{logit}\{\text{Prob}(\text{toxicity} \mid \text{dose})\} = a + b \log(\text{dose})$ and just assumed noninformative normal priors on a and b , with variance = 100. The resulting CRM design does strange things with the first 3 to 6 patients. It drove me and my programmer crazy for a month.*
- ▶ See Stan Prior Choice Recommendations wiki [LINK](#).

Conceptual review of statistics.

Seeing all of statistics all at once.

- ▶ In both Bayes and Frequentist statistics the holy grail is clearly seeing what repeatedly would happen when trying to learn from observations like the ones you have in hand.
- ▶ A helpful? metaphor of discerning what cast the shadows.
- ▶ Think of learning about an object just from the shadows it casts while being unable to look directly at the object.
- ▶ We see those shadows but really are only interested in what is casting them. They may look very scary, but the object casting them maybe mice. (All too common to take noisy observations as the reality that generated them.)

Strategy: Generate realistic Fake Universes to better grasp reality.

- ▶ Make something where you can easily “see” what would repeatedly happen.
- ▶ In analytical chemistry you can repeatedly spike test tubes with know trace amounts of X.
- ▶ In statistics we can't repeatedly spike humans with say know faint rates.
- ▶ So we have to represent faint rates abstractly - make a fake universe were you set the rate.
- ▶ Most statisticians prefer to use probability models (and math) to do that.

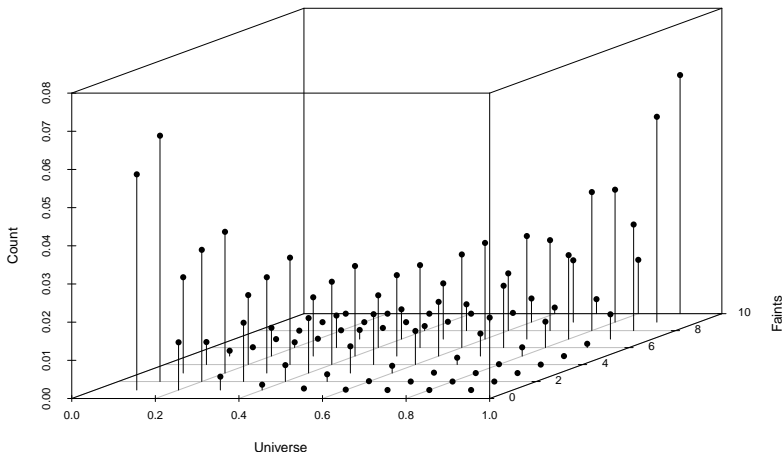
Strategy: Generate realistic Fake Universes to better grasp reality (cont.)

- ▶ Today almost anyone can use probability models (and simulation!!!) to make fake universes.
- ▶ Math: The exact study of ideal states of things (a fake universe).
- ▶ Simulation: The approximate study of the same.
- ▶ Two stage simulation is the just the Bayesian version of generating realistic enough? fake universes!
- ▶ My toy example today will be annual faint rates in business leaders “rounds” at Toronto Hospital where 3 out of 9 fainted the one year I heard about it.

Lets plot the set of fake universes.

Two stage simulation, simulate the faint rate and then number of faints given that rate.

Set of Fake Universes for Fainting.



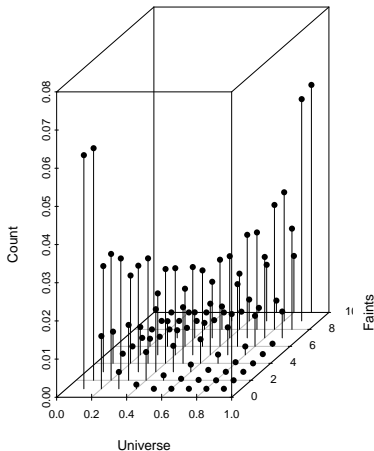
Why make fake universes of faint rates and occurrences?

- ▶ Once the joint model is set - the prior and data generating model - Bayes is deductive.
- ▶ The joint model is the first premise, the data is the second premise and the posterior is the conclusion.
- ▶ Assuming adequate sampling from the posterior - the quality of what follows totally depends the premises.
- ▶ *Can be very bad if the premises are deficient (either joint model or data).*

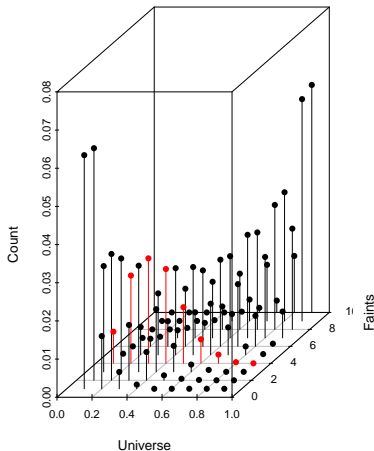
Picture Bayes as deductive.

First premise on the left - do they adequately represent our grasp of reality (as in an artist's pencil sketch)? Note the conclusion in red on the right is clearly contained in the premises (QED).

Set of Fake Universes for Fainting.



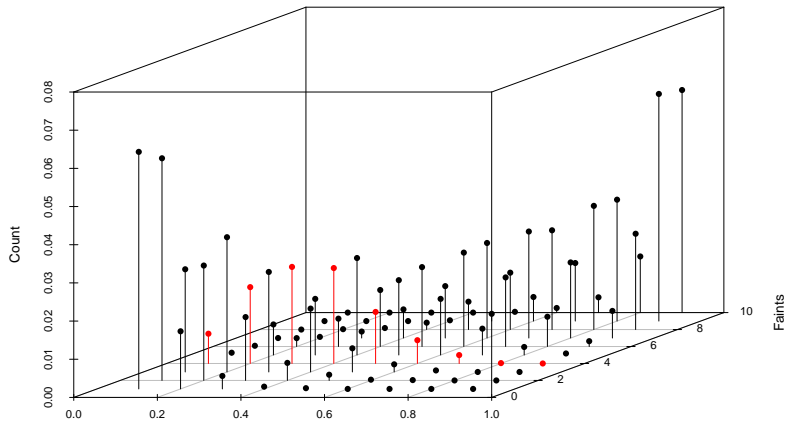
Note the posterior it implies, given 3 out of 9



Consistent with domain expertise?

Note the 9 fake universes on the x axis and 10 possible faint totals on the y axis with how often on the z axis. A proportional set of fake universes and faint totals that we judge could be encountered in topics *like this* (Bayes/Frequency reference sets).

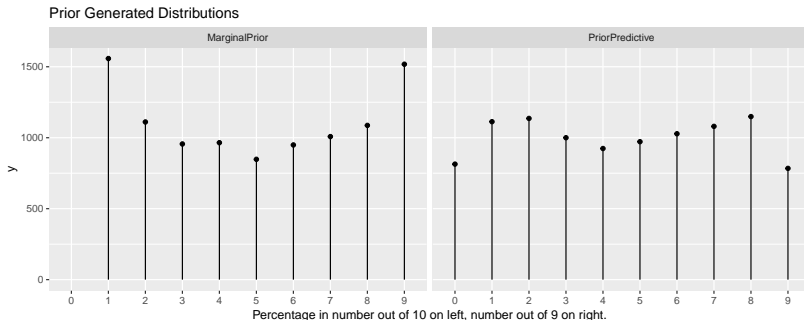
The proportional subset of universes most compatible with our universe shown in red



Things to consider on the last slides.

- ▶ Are these fake universes and repeatedly happening faint totals roughly like our universe?
- ▶ Biggest fail so far may be air density greater than concrete in air quality modeling.
- ▶ *If not too wrong, provides an ample laboratory to evaluate the performance of the given joint model and data!*
- ▶ Often a good idea to use different joint model to create the fake universes to evaluate the Bayesian model actually being used in the analysis.

Marginal views and disagreements on views.



- ▶ Prior predictive distribution is on the right - seem realistic?
- ▶ Marginal prior distribution over all universes is on the left - do these parameters seem realistic.
- ▶ Some argue only look at the second (which has advantages especially with many parameters) while I would look at both.

But so much more to make of the set of fake universes generated.

##	TrueProp1	3 Faints	TrueProp2	Not 3 Faints
## 1	0.5	3	0.9	8
## 2	0.2	3	0.2	0

- ▶ Samples on the left are valid posterior samples - right?
- ▶ What are the ones on the right a valid posterior sample from?
- ▶ But you know how to get many posterior samples using MCMC.
- ▶ If your MCMC is valid the ranks of two stage sample with the MCMC sampleS (say 500) should be uniform.
- ▶ Now we can check the MCMC sampling for every point in the fake universes generated (true parameter, valid sample).

Three tasks in principled Bayesian workflow to be safer.

- ▶ Three tasks in Bayesian Workflow - assess adequacy of joint model, posterior sampling and data.
- ▶ Assessing the adequacy of joint model has only become practical in the past 5 or so years.
- ▶ Around 2012 even experts were refusing (even saying not kosher in Bayes).
- ▶ Still some resistance to assessing the adequacy of the data?
- ▶ Lots can and did go wrong and it remains the most challenging aspect (and an incentive to just ignore it?).

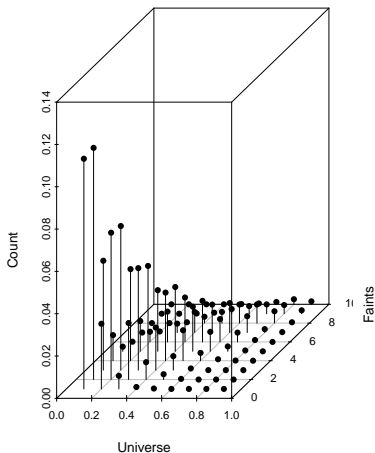
What two stage fake Bayesian universe generation enables.

Makes safe Bayes practical (was not until recently!).

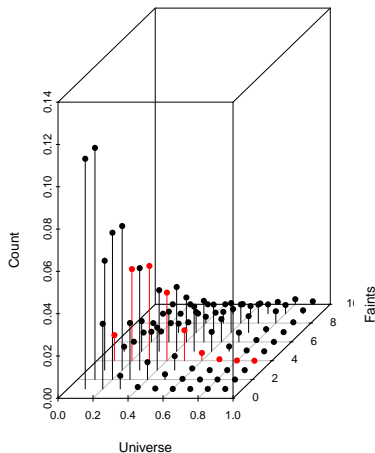
- ▶ Domain Expertise Consistency: Is our model consistent with our domain expertise?
- ▶ Computational Faithfulness: Are our computational tools sufficient to accurately fit the model?
- ▶ Model Sensitivity: How do we expect our inferences to perform over the possible samples in fake universes?
- ▶ (Model Adequacy: Is our model rich enough to capture the relevant structure of our universe?)

A more sensible set of fake universes?

Set of Fake Universes for Fainting.

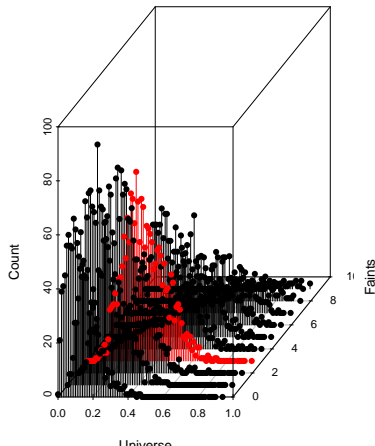
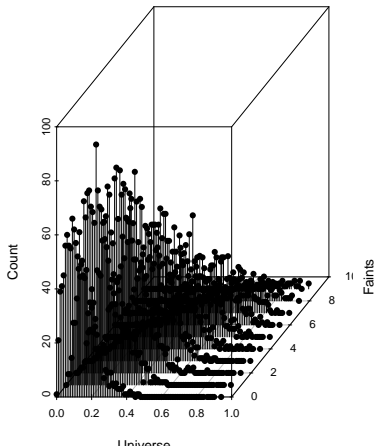


Note the posterior it implies, given 3 out of 9

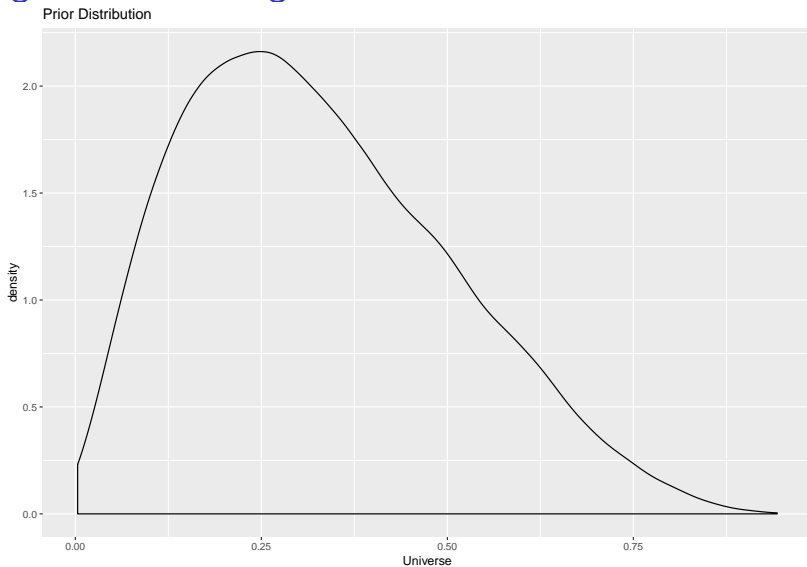


Domain Expertise Consistency: Is our model consistent with our domain expertise?

Now use Rstan which has prior predictive functions (used to have to leave out data ~ statement in model block).

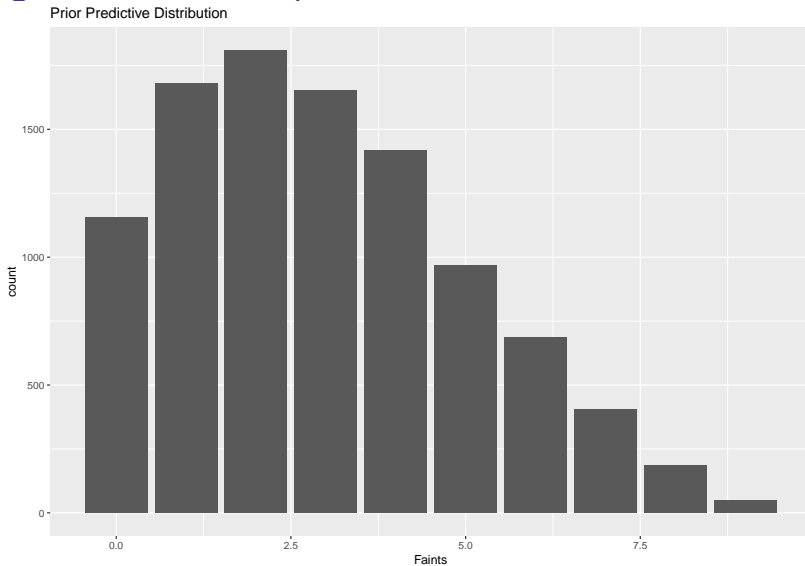


Marginal views: Marginal Prior



- Do these parameters seem realistic?

Marginal views: Prior predictive distribution.



- Prior predictive distribution - these observations seem realistic?

Computational Faithfulness: Are our computational tools sufficient to accurately fit the model?

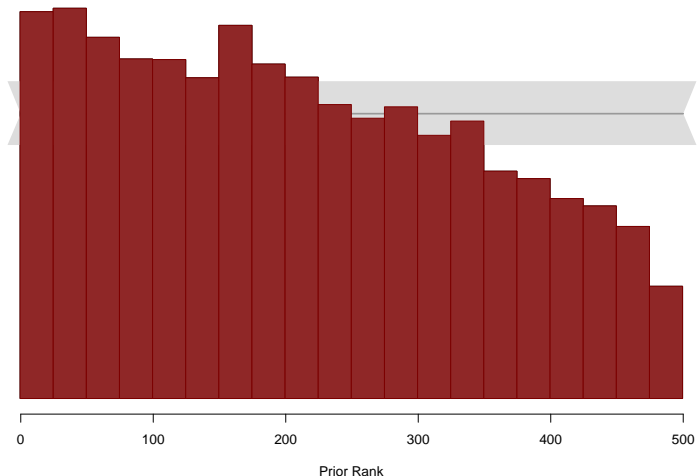
Lets look at some samples of the two stage simulation

##	Universe	Faints	Universe	Faints
## 1	0.2418439	3	0.09257722	0
## 2	0.4519268	3	0.19122446	0
## 14	0.3327996	3	0.25878004	4

- ▶ If the MCMC is valid the ranks of two stage sample with the MCMC sampleS should be uniform.
- ▶ Now we can check the MCMC sampling for every point in the fake universes generated (true parameter, valid sample).

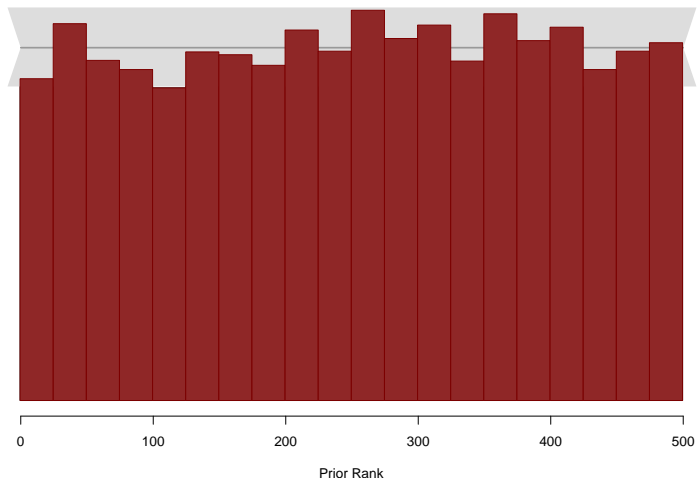
Computational Faithfulness - are ranks uniform?

Lets obtain 500 MCMC samples for each two stage sample in Rstan.
What went wrong? My priors did not match.



Now, lets do that correctly in Stan

Lets obtain 500 MCMC samples for each two stage sample in Rstan using same priors.



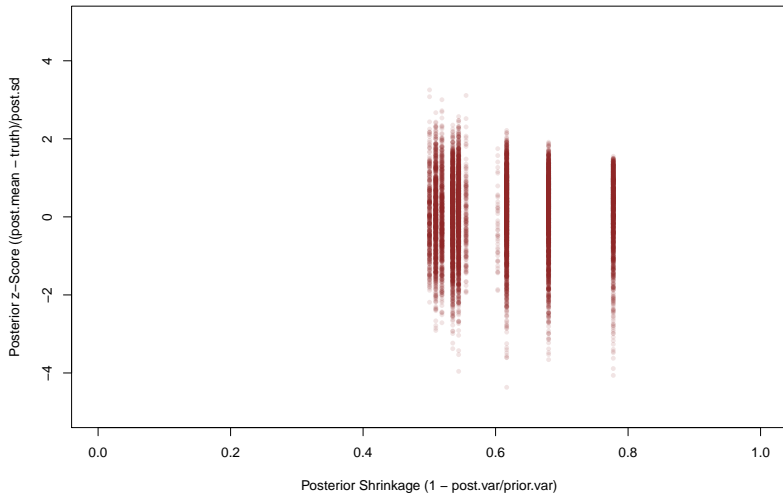
Now actually you should first run a set of utility diagnostic checks.

```
warning_code <- ensemble_output[1,]  
if (sum(warning_code) != 0) {  
  print ("Some simulated posterior fits in the joint ensemble")  
  for (r in 1:R) {  
    if (warning_code[r] != 0) {  
      print(sprintf('Replication %s of %s', r, R))  
      util$parse_warning_code(warning_code[r])  
      print(sprintf('Simulated theta = %s', simu_thetas[r]))  
      print(" ")  
    }  
  }  
} else {  
  print ("No posterior fits in the joint ensemble encountered")  
}
```

```
## [1] "No posterior fits in the joint ensemble encountered"
```

Model Sensitivity.

Generic Posterior Shrinkage and z-Scores.



Model Sensitivity: More focused assessments.

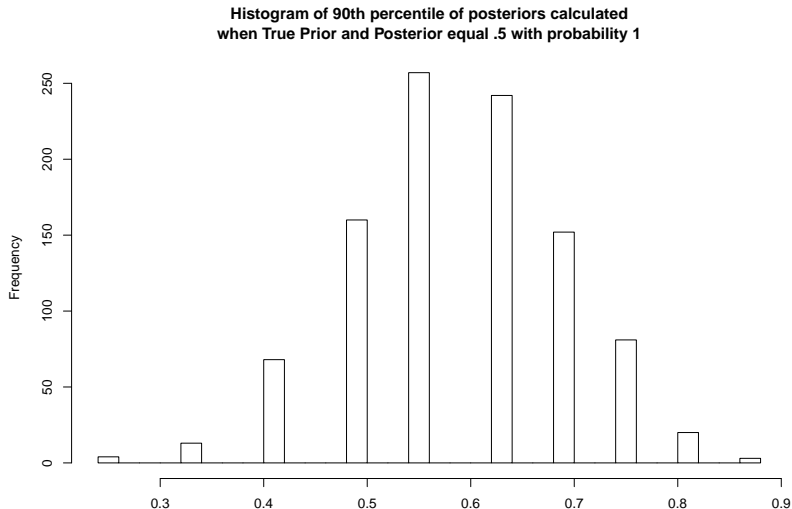
- ▶ FDA 2010 guidance [LINK](#)
- ▶ “FDA usually recommends you provide simulations of your trial at the planning (or IDE) stage. This will facilitate FDA’s assessment of the operating characteristics of the Bayesian trial; specifically, the type I and type II error rates.”
- ▶ Actually carried out a feasibility study of calculating the type I and type II error rates with a major pharma consulting firm and Yongmin Yu at Health Canada in 2012.

Pick one universe and repeatedly sample and see what happens.

- ▶ Think of needing to know if faint rate not $\geq 50\%$.
- ▶ Set the probability of fainting to .5 and simulate data just for that fake universe.
- ▶ Type 1 error? See how often $((\text{post_mean_theta} - .5) / \text{post_sd_theta}) < -1.65$.
- ▶ Calculate the 90th percentile of posterior probability.
- ▶ See how often that is claimed $< 50\%$.

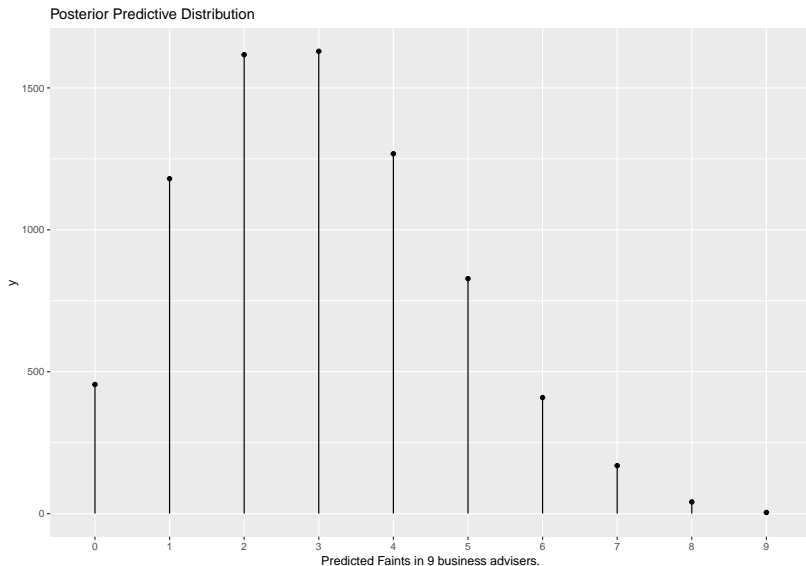
What repeatedly happens in this fake universe of 50% fainting.

```
## [1] "Type 1 Error = 0.085"
```



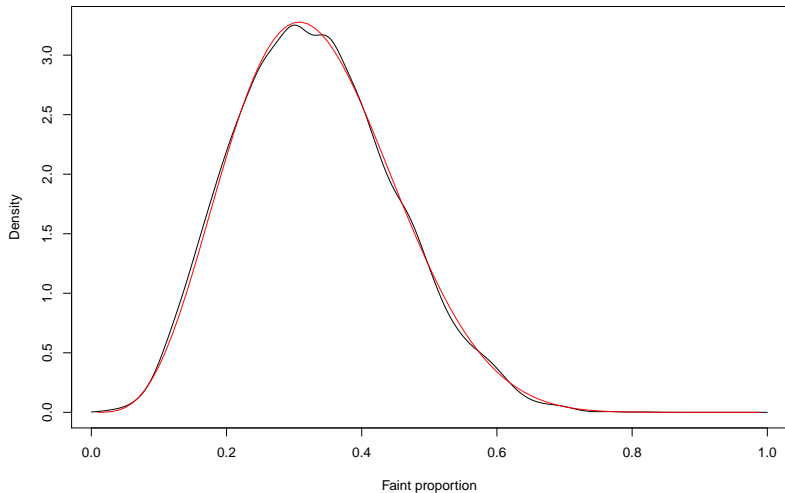
Model Adequacy: Checking the posterior itself.

Not much to see with this toy example.



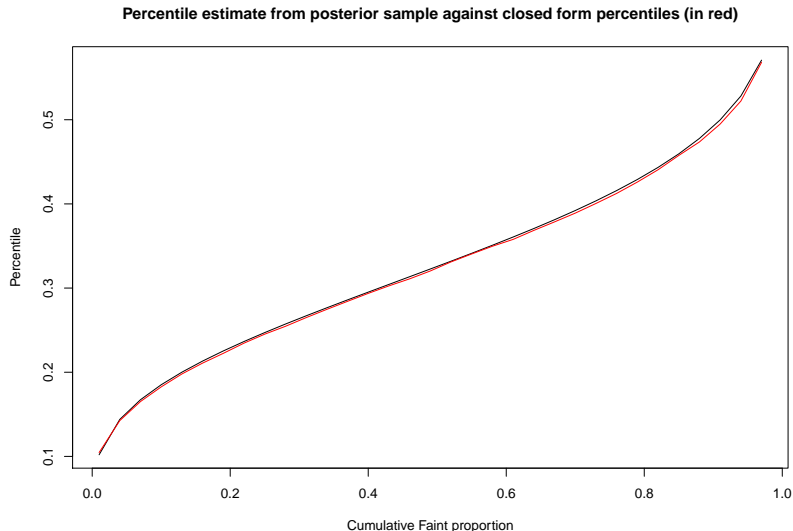
Posterior plot.

Density estimate from posterior sample against closed form posterior (in red)



Its a developing field - checking accuracy of posterior quantiles.

Check Stan website if a concern.



Summary: Bayes enables powerful analyses.

With great power comes great responsibility.

- ▶ Safe Bayes requires adequate assessment of:
- ▶ The joint model specified.
- ▶ The posterior sampling carried out (usually MCMC).
- ▶ The adequacy of the study design/data.
- ▶ This is a developing field - one needs to stay current.

Summary (cont.).

- ▶ The joint model defines a set of of fake universes and what repeatedly happens in them.
- ▶ That set should (!must) include mostly *close enough to our universe* fake universes.
- ▶ If not the posterior might (!will) be misleading for our universe.
- ▶ Two stage sampling provides a way to assess if fake universes look like ours.

Summary (cont.).

- ▶ The posterior needs to be adequately sampled from.
- ▶ Two stage sampling offers an additional direct assessment tool of whether sampling from the posterior was adequate.
- ▶ Extraction of information cannot exceed what is actually in the data (if it does, its error not signal).
- ▶ Two stage sampling offers a way to assess the distribution of information in the study design/data.

Summary (cont.).

- ▶ How often the Bayesian analysis will be misleading is important.
- ▶ Smart people don't like being repeatedly wrong (Don Rubin).
- ▶ Study designs (especially with large sample sizes) can mitigate a poor set of fake universes.
- ▶ Inadequate posterior sampling and poor data likely fatal.
- ▶ But don't assume or take anyone's word for it - check!