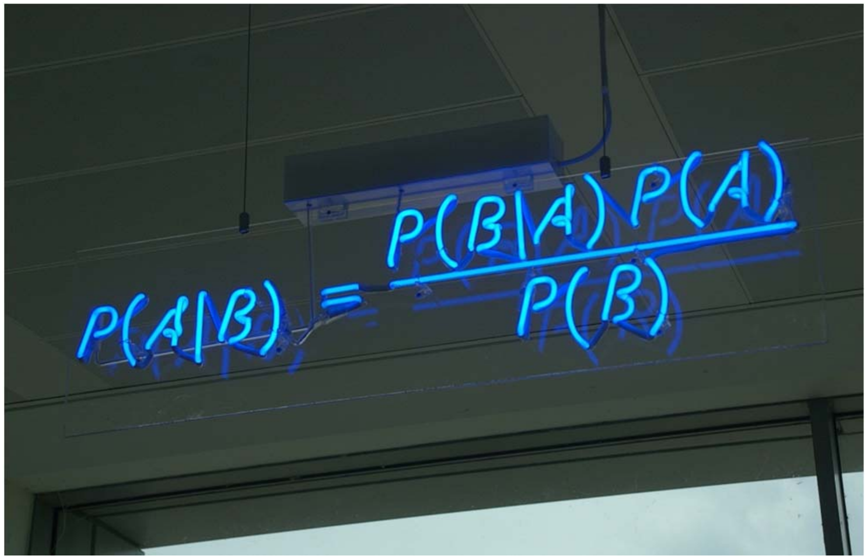


Enhanced Credibility
Bayesian Statistics and Actuarial Modeling

CAS RPM Seminar
Philadelphia

March 20, 2012

Jim Guszcza, FCAS, MAAA

Deloitte Consulting LLP
University of Wisconsin-Madison

Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Agenda

Concepts

Computation

Case Studies

Why Bayes, Why Now

From John Kruschke, Indiana University:

"An open letter to Editors of journals, Chairs of departments, Directors of funding programs, Directors of graduate training, Reviewers of grants and manuscripts, Researchers, Teachers, and Students":

Statistical methods have been evolving rapidly, and many people think it's time to adopt modern Bayesian data analysis as standard procedure in our scientific practice and in our educational curriculum. Three reasons:

1. Scientific disciplines from astronomy to zoology are moving to Bayesian data analysis. **We should be leaders of the move, not followers.**
2. Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive. **Bayesian analyses are readily computed with modern software and hardware.**
3. Null-hypothesis significance testing (NHST), with its reliance on *p* values, has many problems. **There is little reason to persist with NHST now that Bayesian methods are accessible to everyone.**

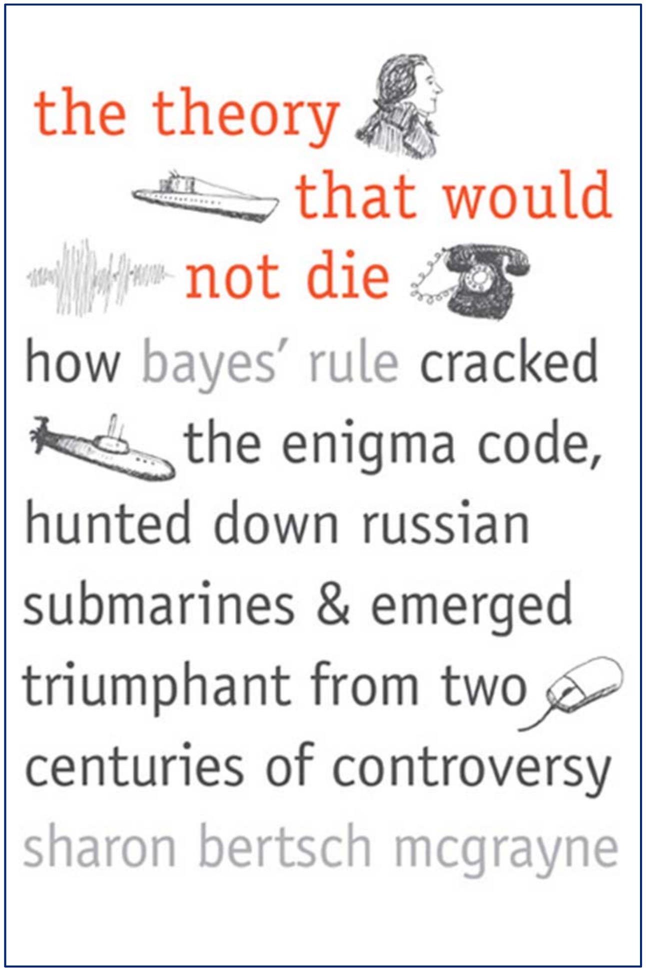
My conclusion from those points is that we should do whatever we can to encourage the move to Bayesian data analysis.

(I couldn't have said it better myself…)

Why Bayes, Why Now

From an Interview with Sharon Bertsch McGrayne in *Chance* Magazine:

"When I started research on [my] book, I could Google the word 'Bayesian' and get 100,000 hits. Recently I got 17 million."



the theory that would not die
how bayes' rule cracked the enigma code, hunted down russian submarines & emerged triumphant from two centuries of controversy
sharon bertsch mcgrayne

Our Profession’s Bayesian Heritage: Early

- Late 18th Century: Thomas Bayes and Pierre-Simon Laplace formulate the principles of “inverse probability”
  - Probabilistic inference from data to model parameters
  - Bayes’ intellectual executor, Richard Price, became perhaps the world’s first consulting actuary (Equitable Life Assurance company, London)
  - Price’s – and perhaps Bayes’ – thinking was influenced by the publication of David Hume’s Treatise on Human Nature (1740)

- 1918: A. W. Whitney “The Theory of Experience Rating”.
  - Advocated combining the claims experience of a single risk with that of a cohort (class, portfolio, …) of similar risks.

$$\boxed{\overline{\mu} = Z \cdot \hat{\mu}_{risk} + (1 - Z) \cdot \hat{\mu}_{class} \quad , \quad Z = \frac{w}{w + k}}$$

  - Estimated pure premium should be a weighted average of the individual risk’s claim experience with that of the cohort… *k* is judgmentally determined.

Our Profession’s Bayesian Heritage: Early-Modern

- 1950: Arthur Bailey publishes “Credibility Procedures: Laplace’s Generalization of Bayes’ Rule and the Combination of Collateral Knowledge with Observed Data”.
  - Anticipates Hans Bühlmann's subsequent work.
  - Quoted Richard Price on making inferences from available data.

“At present, practically all methods of statistical estimation appearing in textbooks… are based on an equivalent to the assumption that any and all collateral information or a priori knowledge is worthless. There have been rare instances of rebellion against this philosophy by practical statisticians who have insisted that they actually had a considerable store of knowledge apart from the specific observations being analyzed… However it appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data.”

Our Profession’s Bayesian Heritage: Mid-Century Modern

- 1967: Bühlmann’s “greatest accuracy” Bayes credibility model.
  - Let X_{ij} denote dollars of loss associated with risk *i* at time *j*.
  - Assume $X_1, \ldots, X_m$ are iid, conditional on a parameter (vector) θ
  - Let m(θ$_i$) denote “risk premium”: $m(\theta_i) \equiv E[X_{ij} | \theta_i]$

- Bühlmann minimizes mean squared errors:

$$E\left[m(\theta_i) - \alpha - \sum_j \beta_j X_{ij}\right]^2$$

- … to arrive at an estimator for m(θ$_i$):

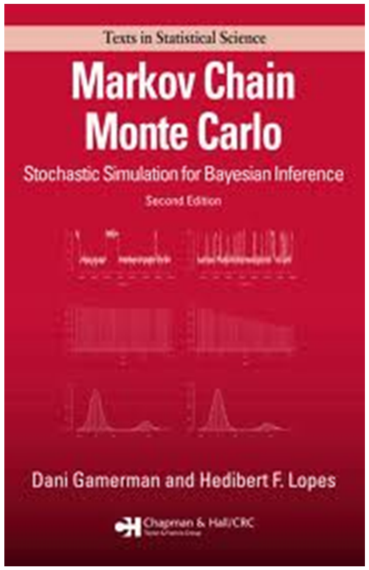
$$z_i \cdot \overline{X}_i + (1 - z_i) \cdot \mu$$

- … where:

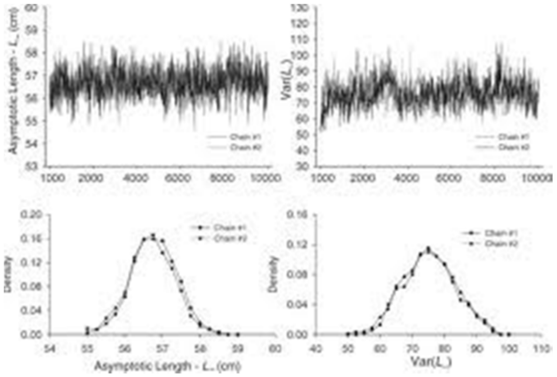
$$z_i = \frac{n_i}{n_i + k} \quad , \quad k = \frac{E\left[Var\left(X_{ij} \mid \theta_i\right)\right]}{Var\left(m(\theta_i)\right)}$$

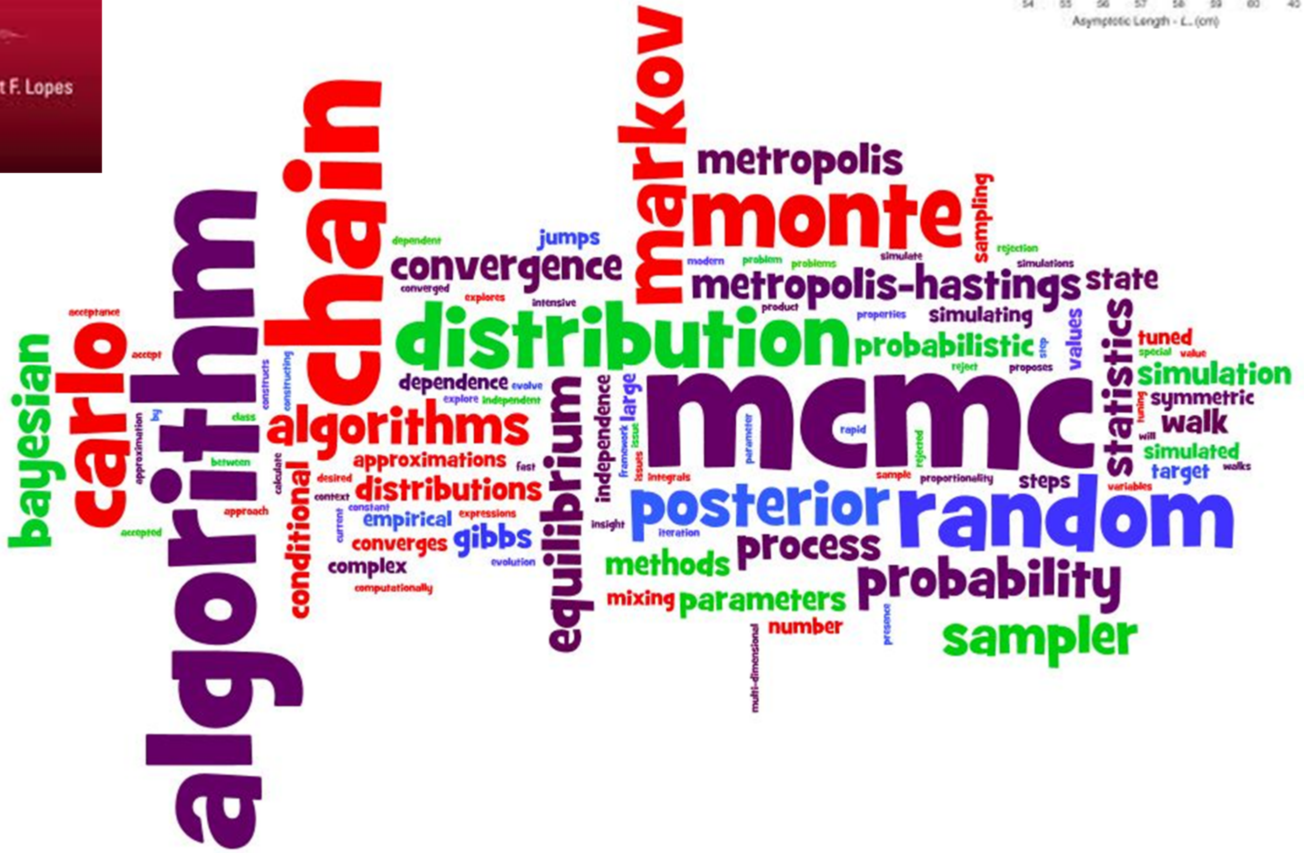
- The within/between variances in *k* are estimated from the data.

Our Profession’s Bayesian Heritage: Modern


Texts in Statistical Science
Markov Chain Monte Carlo
Stochastic Simulation for Bayesian Inference
Second Edition
Dani Gamerman and Hedibert F. Lopes
Chapman & Hall/CRC

?





Bayesian Concepts

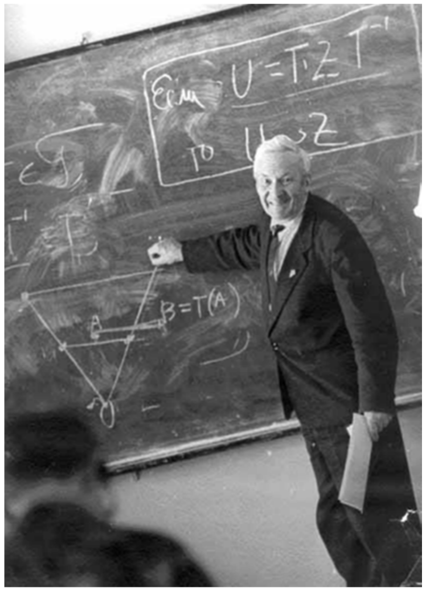
Vocabulary

- Exchangeability
- Credible intervals vs confidence intervals
- Predictive distributions
- Shrinkage / Credibility
- Hierarchical models
- “Borrowing strength”
- Markov Chain Monte Carlo Simulation

Interpreting Probability

It All Starts a Certain Difference of Opinion

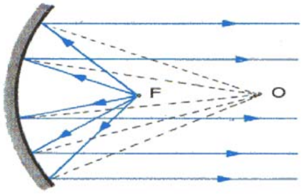
- From a mathematical point of view, probabilities are countably additive, [0,1]-valued functions. Period.
  - For all events E: $\mathrm{Prob}(E) \geq 0$
  - If Ω denotes the sample apace: $\mathrm{Prob}(\Omega) = 1$
  - For pairwise disjoint $\{E_i\}$: $\mathrm{Prob}(E_1 \cup E_2 \cup \ldots) = \sum \mathrm{Prob}(E_i)$



It All Starts a Certain Difference of Opinion

- From a mathematical point of view, probabilities are countably additive, [0,1]-valued functions. Period.
  - For all events E: Prob(E) ≥ 0
  - If Ω denotes the sample apace: Prob(Ω) = 1
  - For pairwise disjoint $\{E_i\}$: Prob($E_1 \cup E_2 \cup \ldots$) = $\sum$Prob(E_i)

- **But whenever mathematics is applied to the world, the relevant concepts must be interpreted.**
  - E.g. in optics a parabola might represent a reflective surface.
  - In biology it might represent some sort of growth.
  - The mathematics is the same either way.
  - But the interpretation is crucial.



Parabolic growth
Self-limiting
Amount
Years
0
2
4
6
8
10
12
0
20
40
60
80
100
120
140
160
180
200
Self-limiting
Reserve growth

- What is the analogous interpretation of probability functions?

Take a Simple Example

- Consider the toss of an ordinary coin.
- **Prob(Heads) = ½** is a mathematical statement.
- But what does this statement mean?

The Frequentist View

- **Probabilities represent frequencies in sequences of repeated events**
  - Emanating from situations involving physical randomization.

- “The probability of heads is ½” means that the coin will come up heads roughly half the time in a sequence of tosses.
  - The more tosses, the closer we this relative frequency approaches 0.50.
  - Prob(H) = ½ means:

$$\boxed{\Pr(H) = \frac{1}{2} \quad \Leftrightarrow \quad \lim_{n\to\infty} \frac{n_H}{n} = \frac{1}{2}}$$

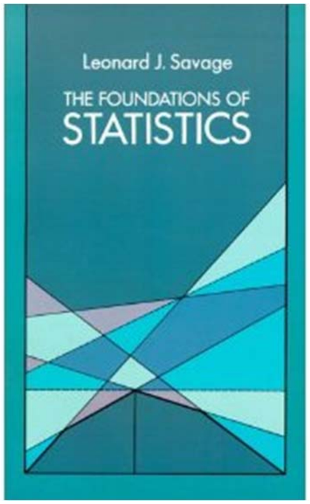
- Many people find this interpretation most acceptable because it is “physical” and “objective” and therefore “scientific”.

The Bayesian View

- **Probabilities represent degrees of certainty or uncertainty.**
- “The probability of heads is ½” means that the speaker believes that the coin is fair.
- Ideally (s)he would be willing to pay $1 for a gamble that pays $2 if the coin lands heads and $0 otherwise.
- People often object to the Bayesian notion because it is “subjective” and therefore presumably not appropriate in scientific investigations.
  - “My belief is that the probability of an earthquake in San Francisco in the next decade is 30%”
  - “Who cares about what you believe?”

Subjective Probability

- “Subjective Probability”
  - Maybe too loaded a term?
  - Historically a lot of confusion and (rather geeky) polemics
  - “PROBABILITY DOES NOT EXIST” – Bruno de Finetti



- “Evidential probability”
  - Maybe a more helpful term?

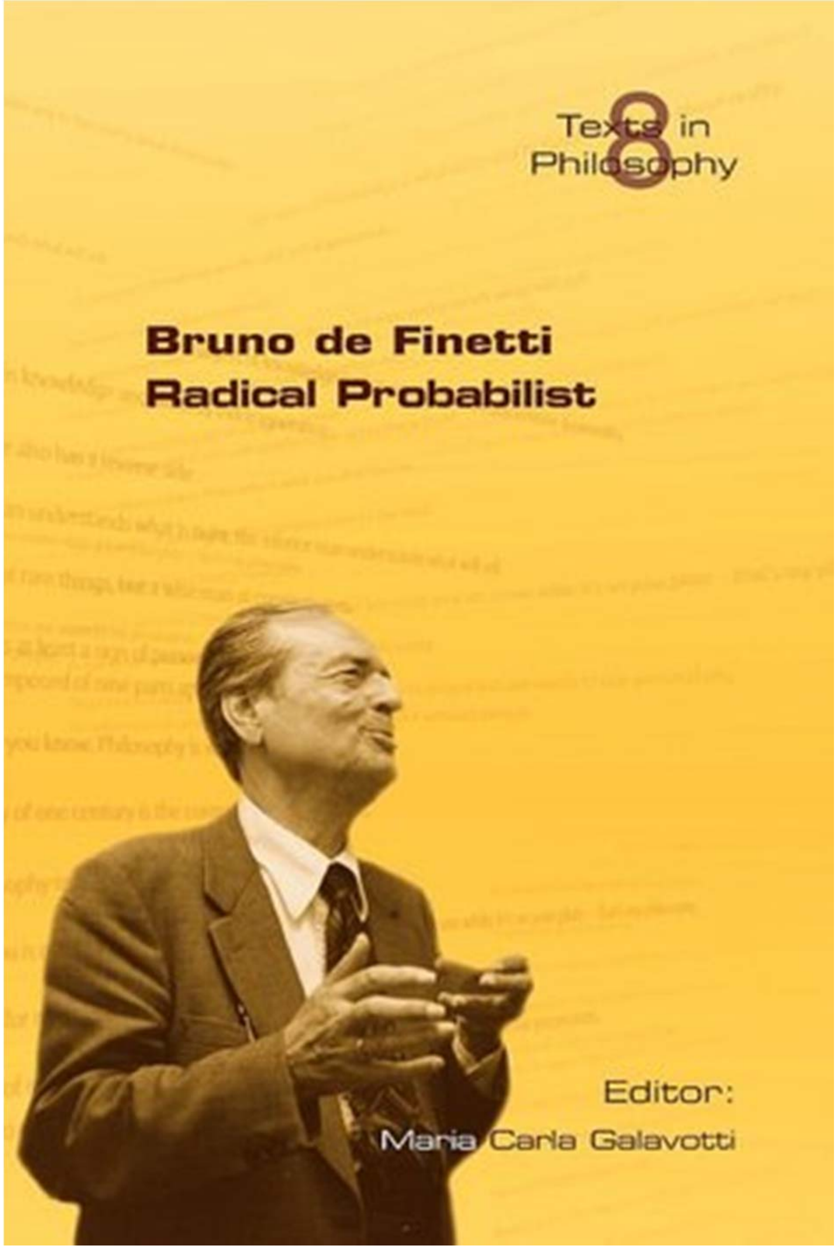
- It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis.”

  – L. J. Savage, *The Foundations of Statistics*

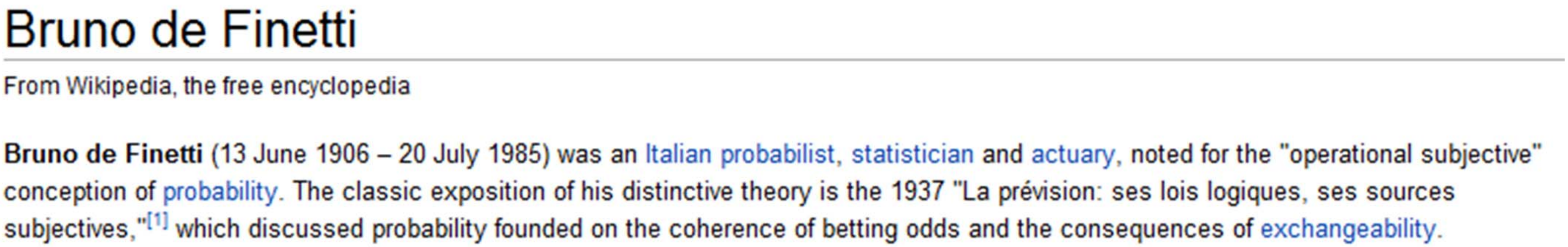
A Farsighted Bayesian

- Bruno de Finetti was one of the most important Bayesian theorists of the 20th century.

- Some interesting history:
  - de Finetti started off as an actuary
  - Independently rediscovered the ideas of the Bloomsbury mathematician/economist Frank Ramsey.
  - Jimmy Savage introduced de Finetti's work to the English-speaking world.
  - Savage and de Finetti both appreciated Arthur Bailey's work in credibility theory in the 1950s.



A Representation of de Finetti



Bruno de Finetti

From Wikipedia, the free encyclopedia

Bruno de Finetti (13 June 1906 – 20 July 1985) was an Italian probabilist, statistician and actuary, noted for the "operational subjective" conception of probability. The classic exposition of his distinctive theory is the 1937 "La prévision: ses lois logiques, ses sources subjectives,"[1] which discussed probability founded on the coherence of betting odds and the consequences of exchangeability.

Single-Case Probabilities

- The interpretation of probabilities in terms of limiting relative frequencies is intuitive at first.

- But often in life and in actuarial science we also find it intuitive to assign probabilities to events that are not part of a sequence of independent random trials.
  - What is the probability Obama will win a 2nd term office?
  - What is the probability of a magnitude 6.7 or greater earthquake in the San Francisco bay area before 2030?
  - What is the probability that the ultimate losses for a cohort of insurance claims incurred in 2012 will fall in the $1M-$1.2M range?
  - What is the probability that the Los Angeles will be the target of a terrorist attach within the coming decade?

Measuring Probabilities

- What is the probability Obama will win a 2nd term office?
- Frequentist answer: I can only answer if Obama’s reelection can be viewed as a repeatable event in which the uncertainty is due to randomness.
  - And the probability is the relative frequency after the event is embedded in this long run of repeated trials.
  - If it can’t be so embedded… no answer.
- Bayesian answer: the uncertainty is due to lack of knowledge.
  - I can quantify my beliefs through betting behavior
  - Suppose I will pay $50 for a lottery ticket that will return $200 if Obama is reelected; nothing otherwise.
  - Then my subjective probability is of Obama being elected is 25%.

Probability as Coherence: Dutch Book Arguments

- Attributed to Frank Ramsey (of the Bloomsbury Group) and Bruno de Finetti.

- If someone’s subjective probabilities do not obey the probability axioms, then they are “incoherent” in the sense that:

- Someone could write a “Dutch Book” against that person.

- A series of bets in which the person would lose money on any outcome.

- In principle, subjective probabilities can be measured through betting behavior.

Learning from Data

Classical and Bayesian Methodology – Learning from Data

- Let's continue thinking about coin tosses.

- Suppose Persi pulls a coin from his pocket and flips it **12** times. **3** of these tosses land heads.

- If Persi were to toss the coin again, what is the probability it would land heads?

- This seems like a silly example but:
  - When thinking about difficult conceptual issues it helps to start with simple examples.
  - And besides, it's not silly. Suppose last year a company sold medical malpractice insurance to 12 heart surgeons in a new zip code, 3 of which had large claims… this year they are thinking about underwriting a 13th heart surgeon in the same state…

How Frequentist and Bayesian Analyses Differ

- The methodological differences between frequentists and Bayesians emanate from the philosophical difference about the interpretation of probability.

- **Frequentists**: the "true probability of heads" is a fact about the world that is manifested in relative frequencies in repeated tosses.
  - The outcome of 3 heads in 12 tosses is one of many possible outcomes of sampling from the "true distribution in the sky".
  - **Probability is assigned to the data… not to model parameters**

- **Bayesians**: the data is a fact in the world. We assign probabilities to quantities we are uncertain about…
  - Probabilities are not assigned to data (although we can incorporate observation errors/sampling mechanisms in a model).
  - Rather, **probabilities are assigned to model parameters** which we do not know with certainty.

A Frequentist Analysis

- To repeat: the data (*h*=3, *n*=12) is viewed as the random outcome of a sampling process that could be repeated ad infinitum.

- From a frequentist POV, what can we infer from the data?

- Let's assume the events {H,T,T,H…} are iid Bernoulli(θ)

- From this assumption it follows that the likelihood function is:

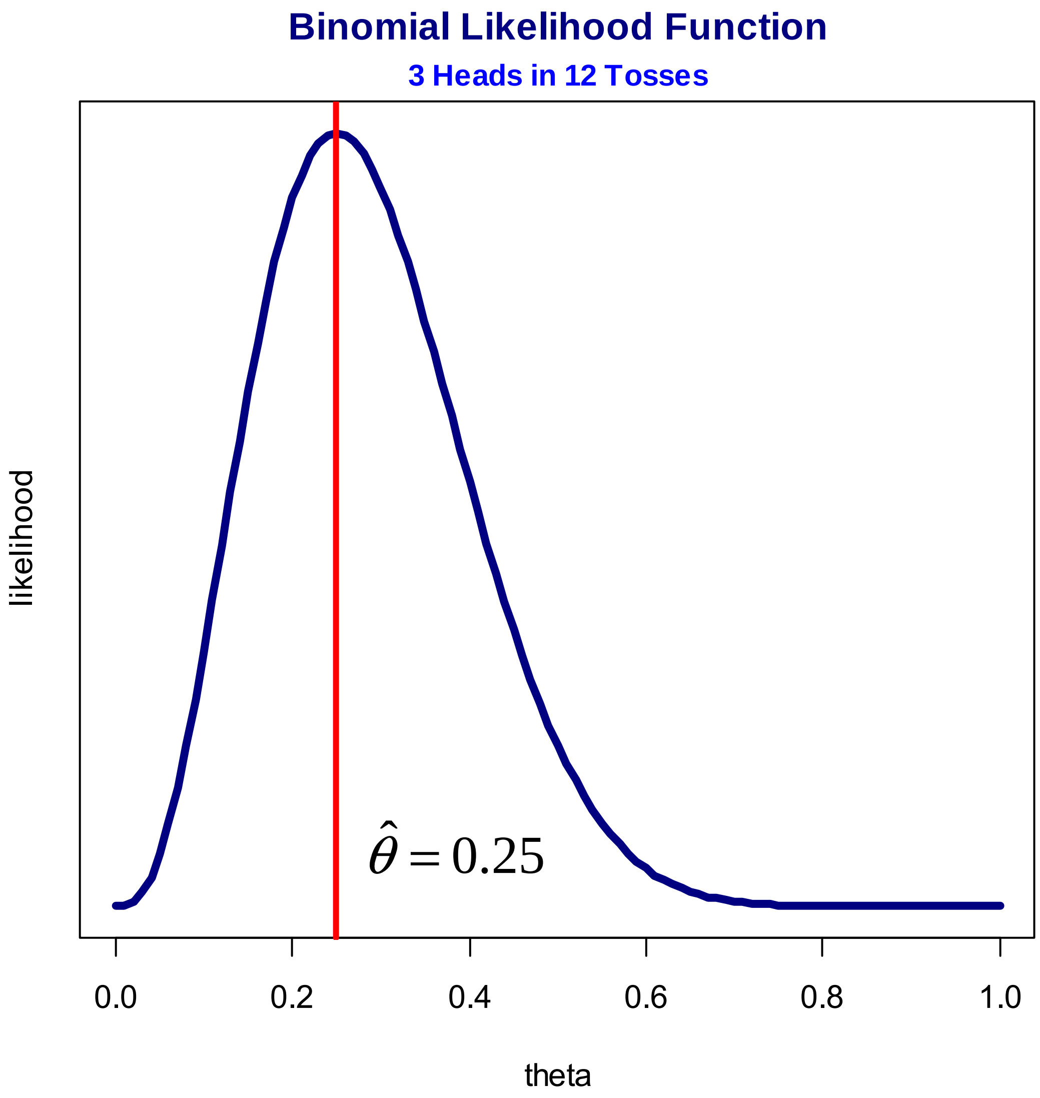
$$\begin{aligned} L(\theta \mid h=3, n=12) &= \prod_{i=1}^{12} \theta^{H_i} (1-\theta)^{T_i} \\ &= \theta^{\Sigma H_i} (1-\theta)^{\Sigma T_i} \\ &= \theta^{3} (1-\theta)^{9} \end{aligned}$$

The Frequentist Estimate

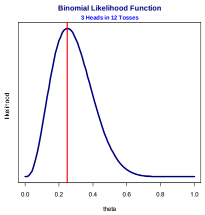
- The Maximum Likelihood Estimate (MLE) is the value of θ that maximizes the likelihood function:

$$L(\theta) = \theta^3 (1-\theta)^9$$

- In this example: MLE = 0.25



What the MLE Means



- Note: the likelihood function $L(\theta)= \theta^3(1-\theta)^9$ is **not a probability function**!

- It is a function of θ, with the data {H,T,T,H,…} regarded as fixed.
  - Remember frequentists don’t assign probabilities to unknown parameters.

- When we maximize likelihood, we select the value θ that results in the model under which the actual observations are most likely to be observed.

- The MLE tells us “what we think” about the coin given the observed data.

- But “how sure are we” about “what we think”?

Measuring "Confidence"

- Twelve tosses aren't that many. How reliable is the maximum likelihood estimate of 0.25?
- To address this question we construct a confidence interval:

$$\boxed{\Pr\left(LB < \theta < UB\right) = 0.95}$$

- **LB and UB are random values** calculated from the data.
- Here, (LB,UB) = (0.0549, 0.5719)
- Does this mean that there is a 95% probability that θ falls in the interval (0.0549, 0.5719)?
- Actually, no.

Measuring “Confidence”

- Confidence interval: $\boxed{\Pr\left(LB < \theta < UB\right) = 0.95}$
- Again we repeat: frequentists only assign probabilities to repeatable, physically random events like {H,T,T,H,…}… … not to parameters like θ.
- θ either is or is not in the interval [0.0549, 0.5719]
- Again… what does the above statement mean?

Frequently Asked Question

- Confidence interval: $\boxed{\Pr\left(LB < \theta < UB\right) = 0.95}$

- What this mean? Answer:

- Suppose we repeated our experiment many times…
  - For each of the next 1000 days, Persi will flips his coin 12 times.
  - Each time he will construct a confidence interval like the one above
  - The resulting interval will differ each day according to how many tosses come up heads on that day.

- But what we can say is that approximately 950 of these intervals will contain the true value of heads!
  - Is this really what people think when they talk about confidence intervals?

Frequently Asked Question

- Confidence interval: $\boxed{\Pr\left(LB < \theta < UB\right) = 0.95}$

- Our 95% “level of confidence” is a measure of the method used to calculate LB and UB…

- … not a measure of our belief that θ lies in the specific interval determined by any particular sample.

- It all goes back to the fundamental principle that probabilities can be assigned only to repeatable random quantities.
    - $\{X_1, X_2, \ldots\}$, LB, UB are such quantities.
    - θ is not.
    - What a tangled web we weave.

The Bayesian Alternative

- For an alternate approach let’s go back to the Bayesian first principle.

- We assign probabilities to quantities that we are uncertain about.

- We are uncertain about whether the coin is fair… what is the “true probability of heads” θ?

- θ can take on values between [0,1].

- So the Beta(α,β) distribution is a good choice.

MahaBeta

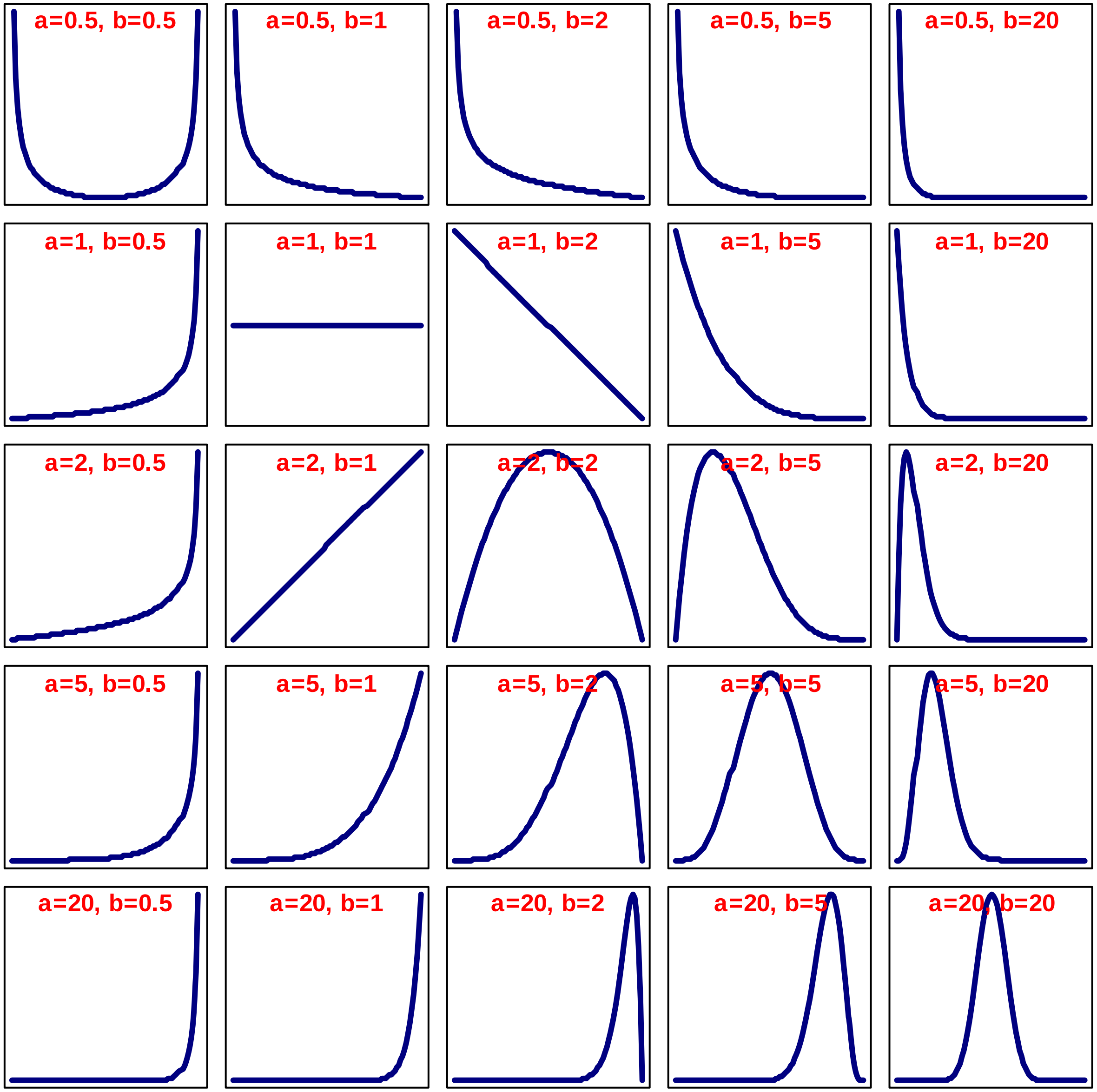
$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The Beta(α,β) family distributions is:

- Defined on [0,1].

- Very flexible.

- In just about any realistic scenario this family will contain a reasonable choice for modeling our (un)certainty about the probability of heads.

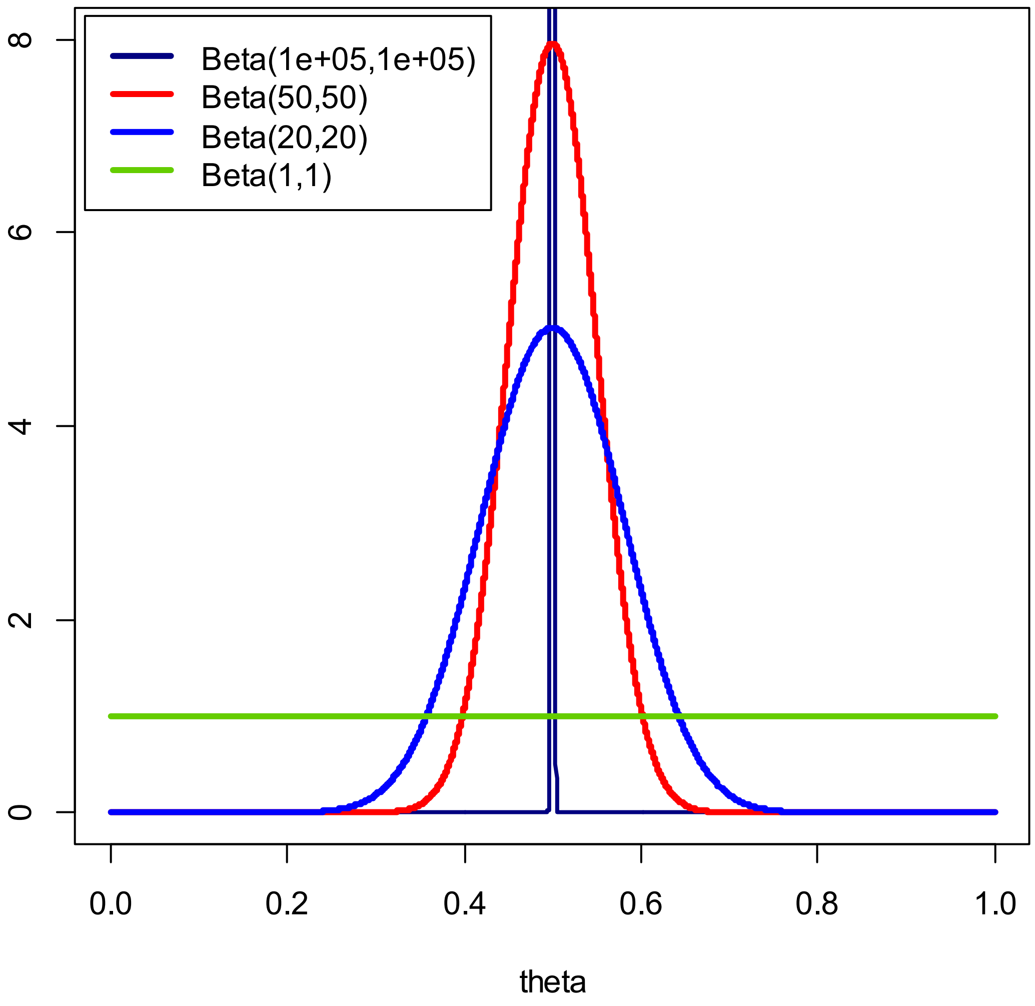


Choice of Priors

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The blessing and the curse of Bayesian statistics:
- We model uncertainty quantities with probabilities.
- So even before we take our data into account we need to select a “prior” probability distribution for θ.

A Few Possible Prior Probability Functions

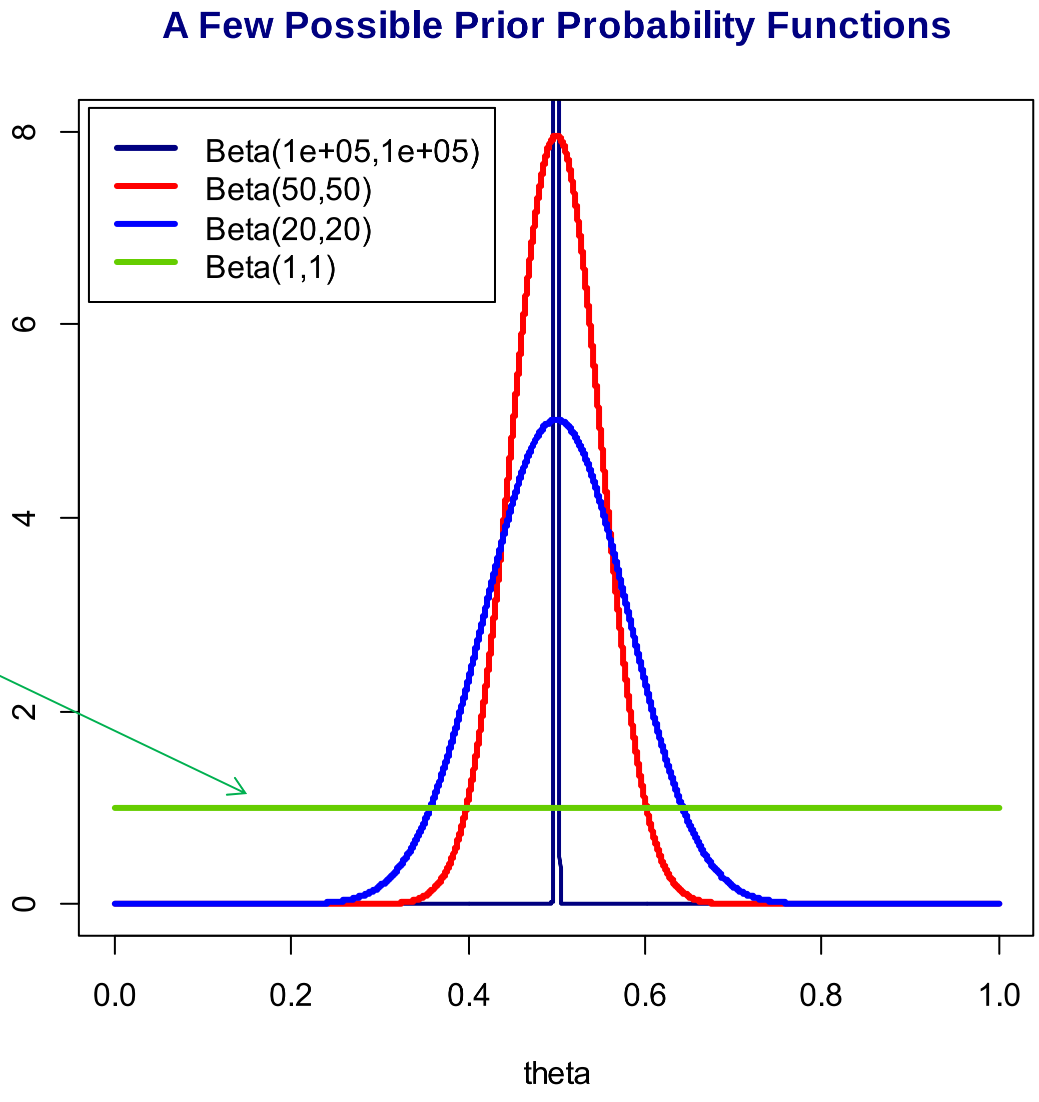


Choice of Priors

$$f(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Here are a few choices that are symmetric with respect to the possibility of the coin being biased towards heads of tails.

- Beta(1,1) – the “flat prior”… we have no idea whether the coin is biased, or how biased it is.

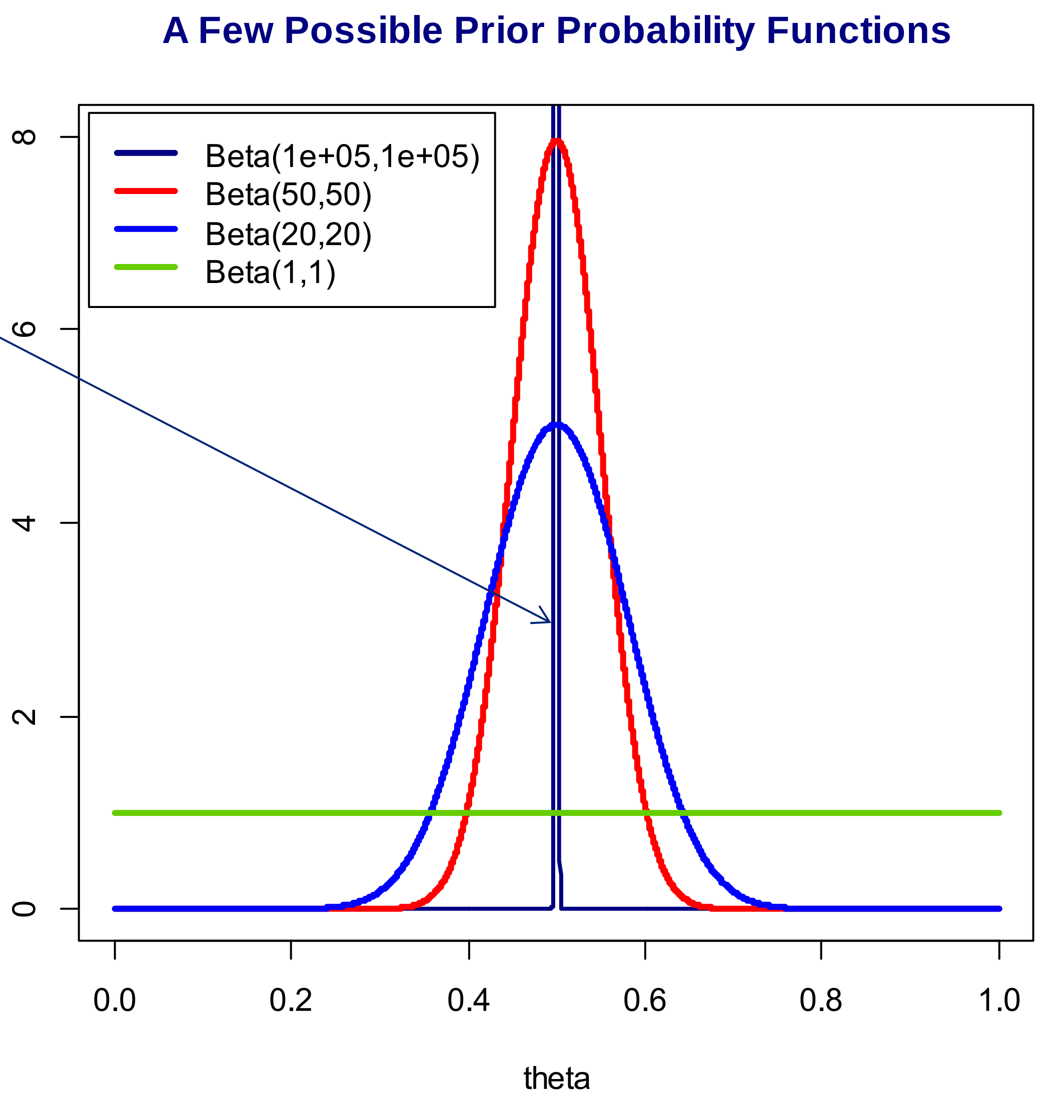


Choice of Priors

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Beta(100000,100000): we have virtual prior certainty that the coin is fair.

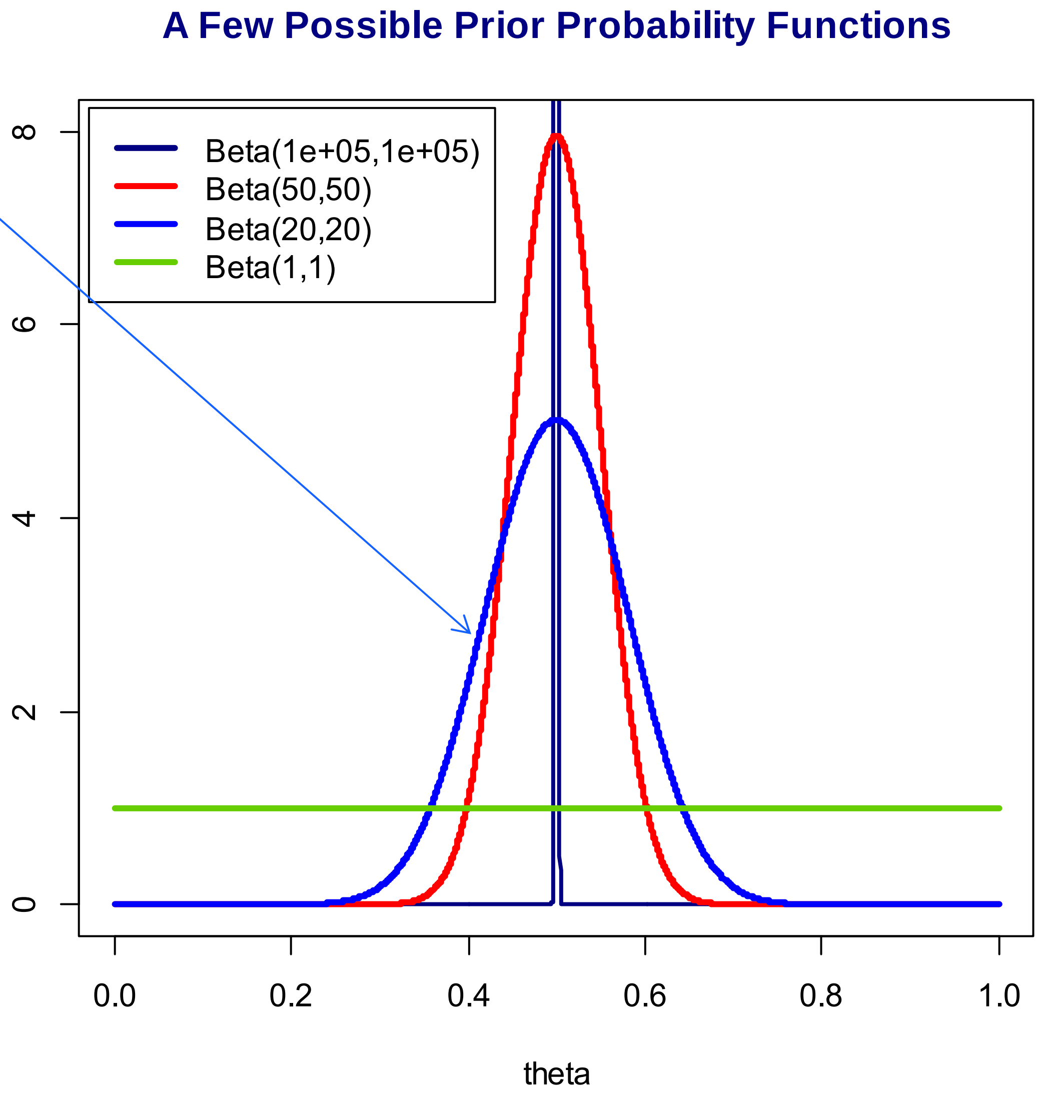
- In the limiting case where we have prior certainty, it means no possible evidence could change our mind.



Choice of Priors

$$f(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Beta(20,20): an intermediate case.
- We have some reason to think that the coin is biased.
  - E.g. maybe Persi is a magician that has been known to flip biased coins in the past.
- But we don’t believe that the coin is more likely to be biased towards heads or tails.



Updating Subjective Probability

- The prior distribution summarizes our beliefs before we have taken the data into account.

- The data (3 heads in 12 tosses) might lead us to change our beliefs about the coin…

- … so our probability function over θ should change accordingly.

- A well known foundational paper on this topic:

Updating Subjective Probability

PERSI DIACONIS and SANDY L. ZABELL*

Updating Subjective Probability

- Bayes' **Theorem** (a mathematical fact):

$$\Pr(H \mid E) = \frac{\Pr(H \wedge E)}{\Pr(E)} = \frac{\Pr(E \mid H)\Pr(H)}{\Pr(E)}$$

- Bayes' **updating rule** (a methodological premise):

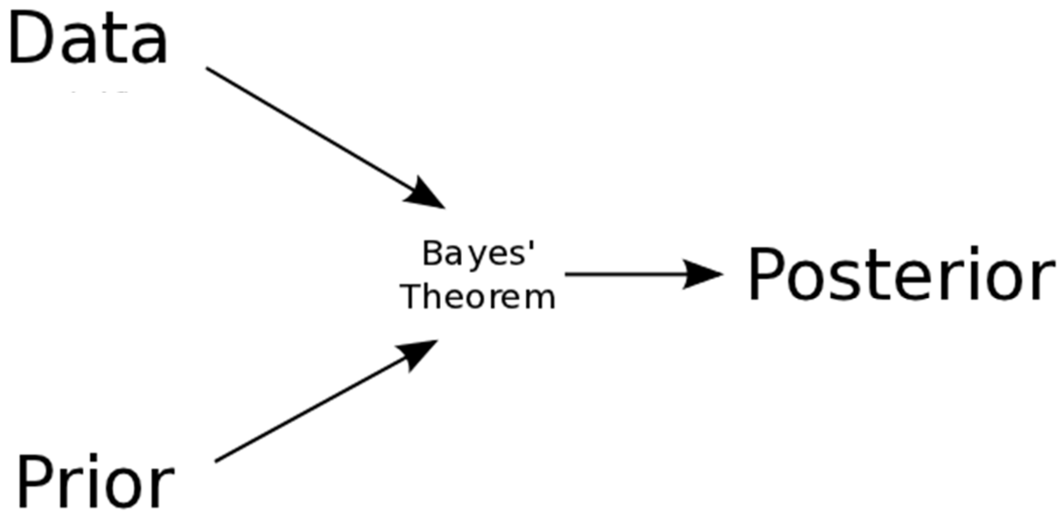
- Let *P*(*H*) represents our belief in hypothesis *H* before receiving evidence *E*.

- Let *P**(*H*) represent our belief about *H* *after* receiving evidence *E*.

- **Bayes Rule: *P**(H) = Pr(*H*|*E*)**

$$\boxed{\Pr(H) \quad \xrightarrow[E]{} \quad \Pr(H \mid E)}$$

Updating Subjective Probability



Data

Bayes'
Theorem

Posterior

Prior

$$\boxed{\Pr(H) \quad \rightarrow \quad \Pr(H \mid E) = \frac{\Pr(H \wedge E)}{\Pr(E)} = \frac{\Pr(E \mid H)\Pr(H)}{\Pr(E)}}$$

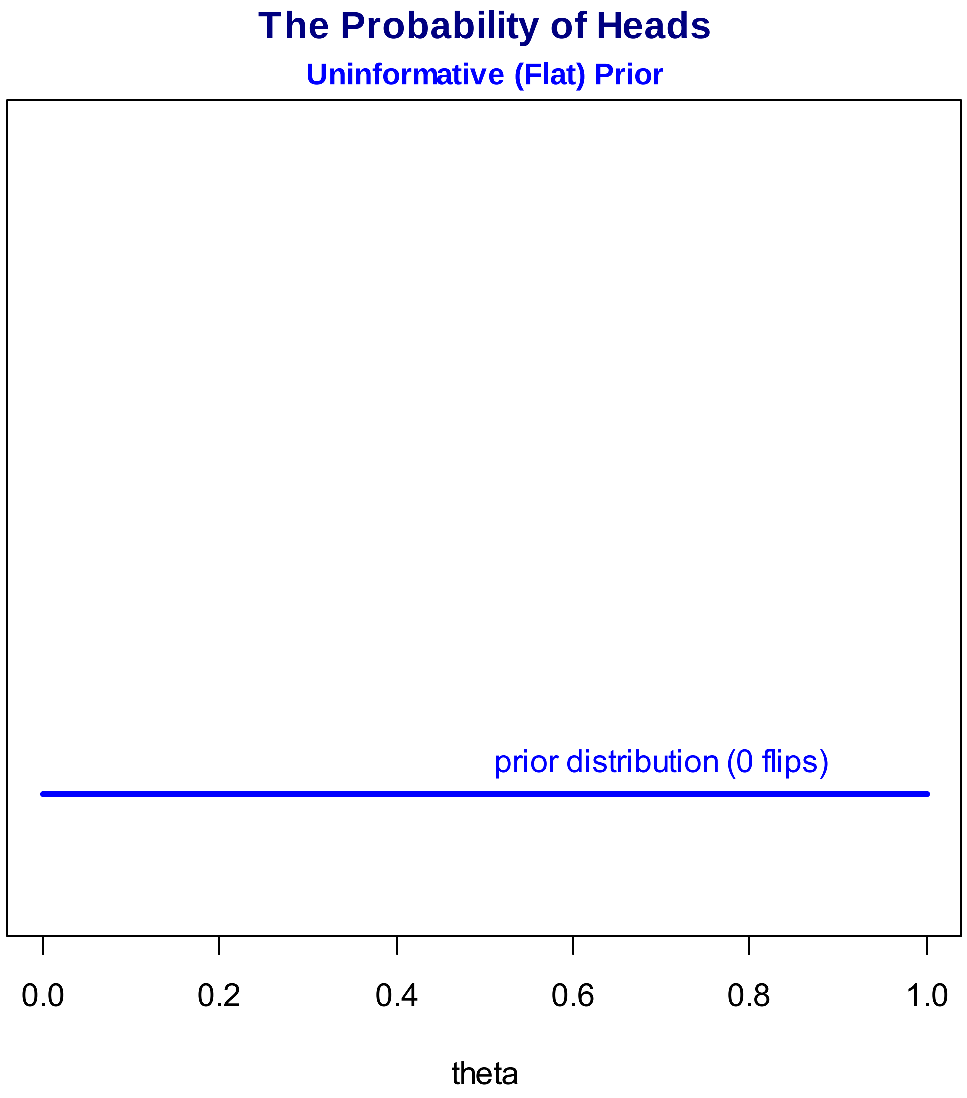
The Beta-Binomial Case

- Bayes' Theorem: $f(\theta | X) = \dfrac{f(X|\theta)\pi(\theta)}{\int f(X|\theta)\pi(\theta)d\theta}$
- Likelihood: $f(X|\theta) = \theta^3(1-\theta)^9$
- Prior: $\pi(\theta) = \dfrac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- So by Bayes Rule:
- Posterior: $f(\theta | X) \propto f(X|\theta)\pi(\theta) = \theta^{\alpha+2}(1-\theta)^{\beta+8}$

➔ updating takes the form: $\boxed{Beta(\alpha,\beta) \underset{3heads}{\rightarrow} Beta(\alpha+3,\beta+9)}$

The Probability of Heads: Prior Uncertainty

- Let's assume absolutely no prior knowledge about whether, or the degree to which, the coin is biased.
  - Maybe Persi drew it at random from a large urn of coins, which have uniformly distributed physical probabilities of heads.
  - Or maybe we just have no idea what tricks Persi might have up his sleeve.
  - Kind of like life.

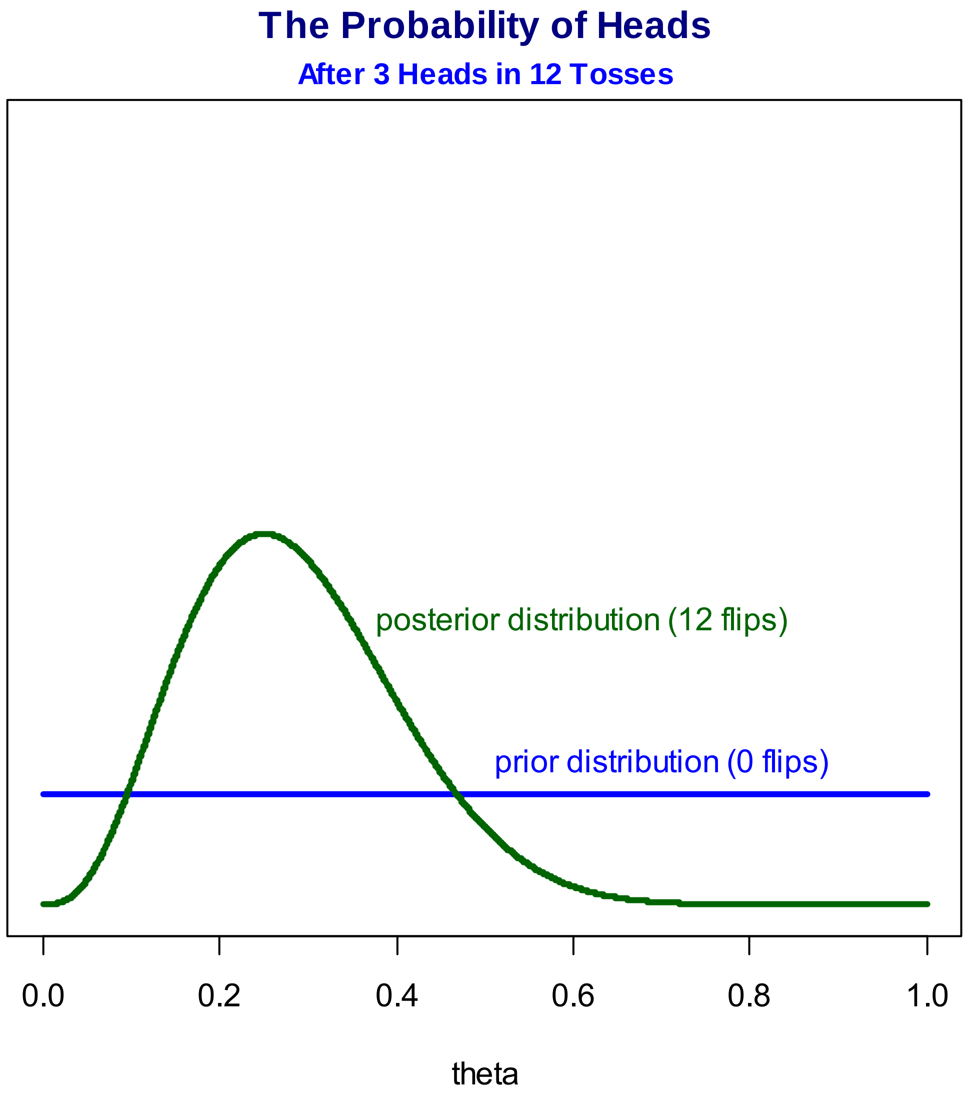


The Probability of Heads: After 12 Flips

- Our data is 3 heads in 12 tosses.

- We use the Bayes updating rule to update our belief (probability) about θ in the light of the data.

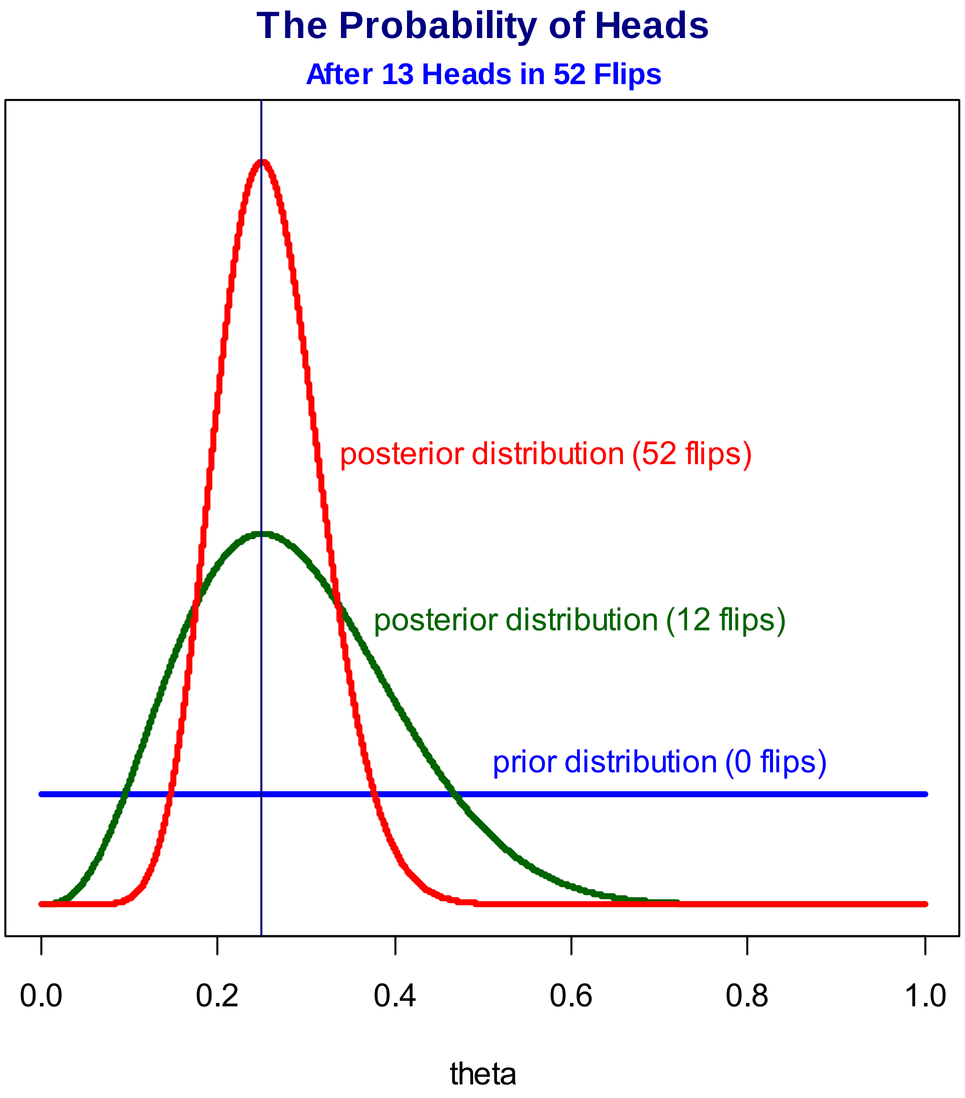
$$Beta(1,1) \quad \rightarrow \quad Beta(4,10)$$



The Probability of Heads: After 52 Flips (Scenario A)

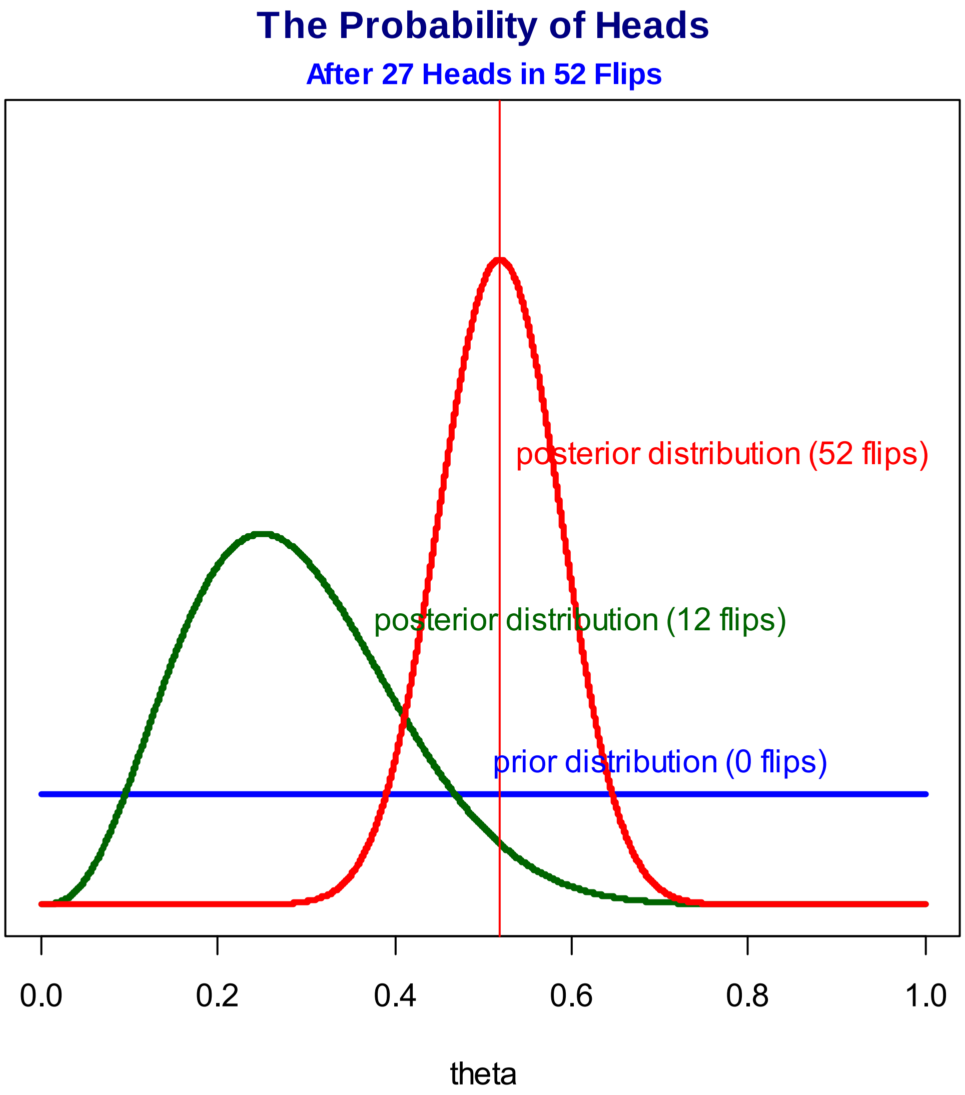
- Suppose Persi flips the coin another 40 times and the total number of heads in all 52 tosses is 13.
- 13/52 = 0.25
- So our posterior distribution is still peaked at the same place, but contains less variability around the mode.

$$Beta(4,10) \quad \rightarrow \quad Beta(14,40)$$



The Probability of Heads: After 52 Flips (Scenario B)

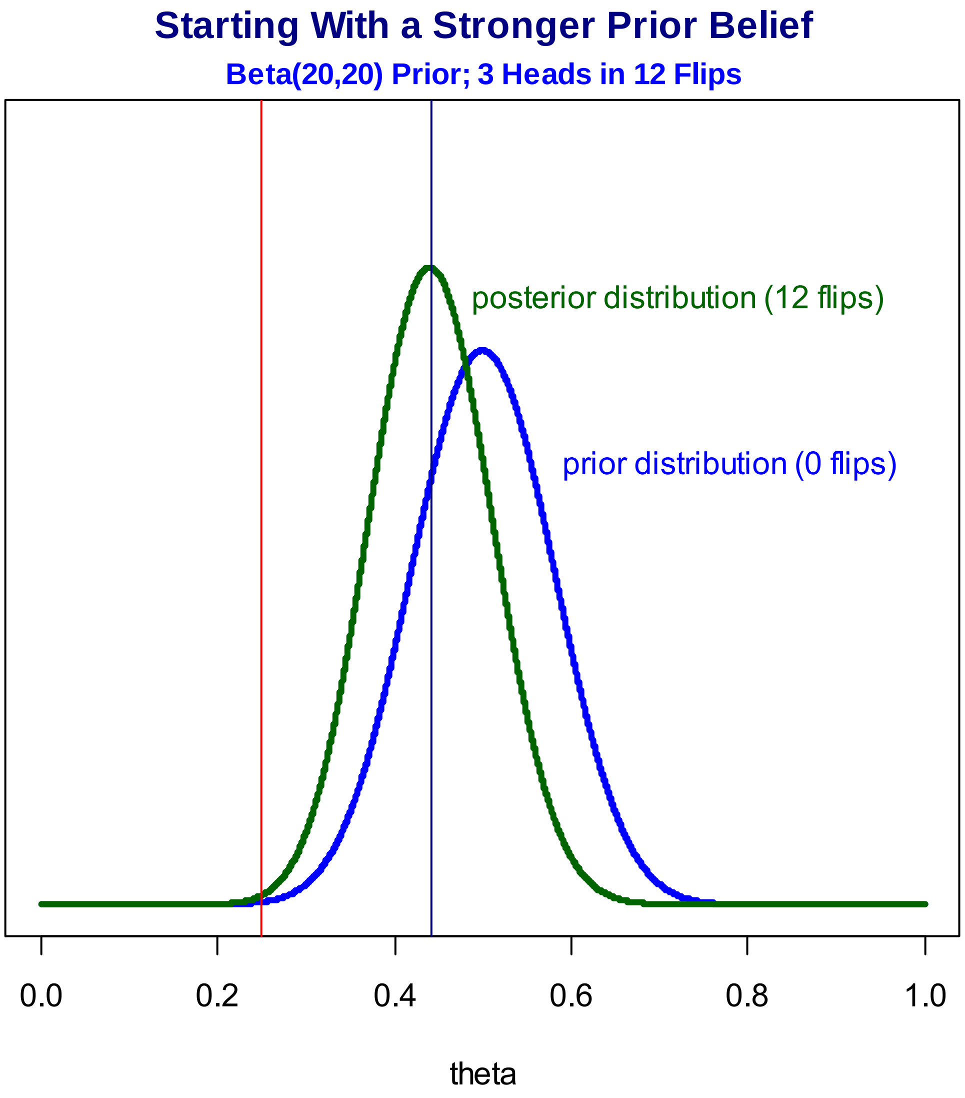
- Of course it could have turned out differently.
- Here's what our posterior would look like if Persi's luck changed and the total number of heads in 52 flips ended up being 27.
- Close to symmetric.



$$Beta(4,10) \quad \rightarrow \quad Beta(28,26)$$

And if We'd Started with an Informative Prior…

- A major weakness of the Bayesian paradigm: the **need** to specify a prior.
- A major strength of the Bayesian paradigm: the **ability** to specify a prior!
  - A rigorous way of incorporating expert judgment and background knowledge in one's analysis.
- The data (3/12) is strongly tempered by our prior belief.
  - A "shrinkage" phenomenon.
  - This is a good thing



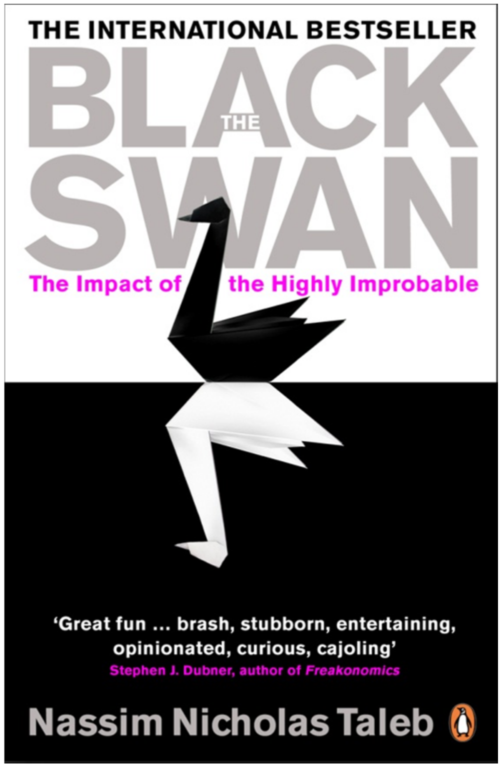
The Frequentists’ Declaration of Independence

- So we’ve got the prior probability distribution covered.
- What about the likelihood function?
- Recall that the frequentist MLE method began by assuming that the coin tosses as iid Bernoulli.
- We assume **independence**.
- This makes sense given the frequentist premise that θ is fixed and the data {H, T, T, H, …} is a random draw from a “sampling distribution in the sky”.
- But does independence make sense from a Bayesian POV?

Taleb’s Question

- But does independence make sense from a Bayesian POV?
- Let’s take a page from Nassim Taleb’s book.

Assume that a coin is fair, i.e., has an equal probability of coming up heads or tails when flipped. I flip it ninety-nine times and get heads each time. What are the odds of my getting tails on my next throw?


THE INTERNATIONAL BESTSELLER
THE BLACK SWAN
The Impact of the Highly Improbable
‘Great fun … brash, stubborn, entertaining, opinionated, curious, cajoling’
Stephen J. Dubner, author of Freakonomics
Nassim Nicholas Taleb

Taleb’s Question

- *…What are the odds of my getting tails on my next throw?*

Dr. John: Trivial question. One half, of course, since you are assuming 50 percent odds for each and independence between draws.

NNT: What do you say Tony?

Fat Tony: I’d say no more than 1 percent, of course.

NNT: Why so? I gave you the initial assumption of a fair coin, meaning that it was 50 percent either way.

Fat Tony: You are either full of #$@& or a pure sucker to buy that “50 pehcent” business. The coin gotta be loaded. It can’t be a fair game.
(Translation: It is far more likely that your assumptions about the fairness are wrong that the coin delivering ninety-nine heads in ninety-nine throws.)

NNT: But Dr. John said 50 percent.

Fat Tony (whispering in my ear): I know these guys with the nerd examples from the bank days. They think way to slow. And they are too commoditized. You can take them for a ride.

Bayesians Aren't So Certain

- Independence implies that:

$$\Pr\left(\theta \mid X_1 = H \; \& \; X_2 = H \; \& \ldots \& \; X_{99} = H\right) = f\left(\theta\right) = \frac{1}{2}$$

- From a Bayesian POV, this implies prior <u>certainty</u> that the coin is fair.

- Prior certainty: **Pr(*H*|*E*) = Pr(*H*)**
  - Our beliefs about proposition *H* will not change.
  - Regardless of how strong the evidence ***E*** is.

- We cannot be certain about model parameters θ
- We must average over the possible values using a prior/posterior probability distribution as a weight.

A Fair Exchange

- Rather than assume **independence**, Bayesians adopt the weaker assumption of **exchangeability**.

- Exchangeability is a kind of symmetry condition that presumably reflects a corresponding symmetry in our beliefs.
  - “the future will resemble the past.”

- Exchangeability: the order of a finite set of random variables does not affect the joint probability. For all *n* and permutations σ:

$$\boxed{\Pr(X_1 = e_1, X_2 = e_2, \ldots, X_n = e_n) = \Pr(X_1 = e_{\sigma(1)}, X_2 = e_{\sigma(2)}, \ldots, X_n = e_{\sigma(n)})}$$

de Finetti's Representation Theorem

- Suppose $\{X_i\}$ is exchangeable:

$$\Pr(X_1 = e_1, X_2 = e_2, \ldots, X_n = e_n) = \Pr(X_1 = e_{\sigma(1)}, X_2 = e_{\sigma(2)}, \ldots, X_n = e_{\sigma(n)}) \qquad \forall n, \sigma$$

- Then the limiting relative frequency $\lim_n \rightarrow ({}^1/_n \sum X_i)$ exists with probability 1 and:

$$\Pr(\sum_{i=1}^{n} X_i = k) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \, d\mu(\theta)$$

- ***An exchangeable sequence of is a mixture of iid sequences.***
  - As the posterior $\mu(\theta)$ becomes sharply peaked, the Bayesian predictive distribution approaches the frequentist model.

The Importance of Exchangeability

- Analogous to the frequentist assumption of independence, some form of exchangeability assumption is implicit in all Bayesian models.
- Often, this will be a “conditional” exchangeability assumption.
- E.g. the observations across times, states, policies, … are exchangeable only once we’ve reflected the relevant information in the model.
  - Inflation across time…
  - Variables capturing aspects of a state’s economy, regulatory environment…
  - Variables capturing aspects of a policy…

Frequentism as a Limiting Case of Bayesianism

- de Finetti's result – and its extensions – shows that a predictive distribution can be represented as a Bayesian mixture of frequentist likelihood models.

- Consider the limiting case where our posterior distribution is sharply peaked around a specific value of θ.
  - Either through prior certainty like Dr John
  - Or a big data set, such as 495 heads in 1000 tosses

- In this case, the frequentist model is a good approximation of the Bayesian predictive distribution.

A Few Words About Hypothesis Tests

The Strength of Evidence

- For a Bayesian:
  - An estimate like the mean of the posterior distribution summarizes what we think in the light of 3 heads in 12 tosses
  - A credible interval summarizes the strength of this belief

- For a frequentist, the story is less simple:
  - The MLE summarizes what the data tells us about the coin
  - Confidence intervals summarize the strength of this evidence

- Another frequentist tool: assessing the “significance” of evidence using *p*-values
  - Is the evidence strong enough to reject a null hypothesis?
  - Ubiquitous in actuarial science and general scientific research
  - But should it be?

Minding Our *p*’s

- *p*-value: the probability of the observed outcome… or a more extreme outcome… assuming the null hypothesis is true.
  - A measure of “surprise”

- In the coin example, a natural null hypothesis is that the coin is fair: θ=½

- The probability of 3 or fewer heads assuming θ=½ is:

$$p - value = \sum_{i=0}^{3} \binom{12}{i} \left(\frac{1}{2}\right)^{i} \left(1 - \frac{1}{2}\right)^{12-i} = 0.073$$

- We “fail to reject at the 5% significance level” the hypothesis that the coin is fair.

p Soup

- One reaction to this logic: do we really go through life either “rejecting” or “failing to reject” things based on what we see?
  - Or do we take actions based on provisional beliefs that are shaped by evidence?

- But there is a deeper issue.

- On the previous slide we tacitly assumed that Persi set out to flip the coin 12 times.
  - 3 is the random quantity

- But what if Persi had set out to keep flipping the coin until the 3rd head appears?
  - 12 is the random quantity

A Likely Story?

- Binomial scenario: Persi flips 12 times
  - We do not reject the hypothesis that the coin is fair.

$$p-value=\sum_{i=0}^{3}\binom{12}{i}\left(\frac{1}{2}\right)^{i}\left(1-\frac{1}{2}\right)^{12-i}=0.073$$

- Negative Binomial scenario: Persi keeps flipping until n_H=3:
  - We do reject the hypothesis that the coin is fair.

$$p-value=1-\sum_{i=0}^{8}\binom{i+2}{2}\left(\frac{1}{2}\right)^{3}\left(1-\frac{1}{2}\right)^{i}=0.0327$$

- **Whether or not we reject depends on what Persi intended to do when he started flipping!**

A Bayesian Update

- Recall that the Bayesian method is to update our prior probability via the likelihood function:

$$\theta^{\alpha}(1-\theta)^{\beta} \quad \rightarrow \quad \kappa\boxed{\theta^{3}(1-\theta)^{8}}\,\theta^{\alpha}(1-\theta)^{\beta}$$

➔ Bayesian updating obeys the "**likelihood principle**"

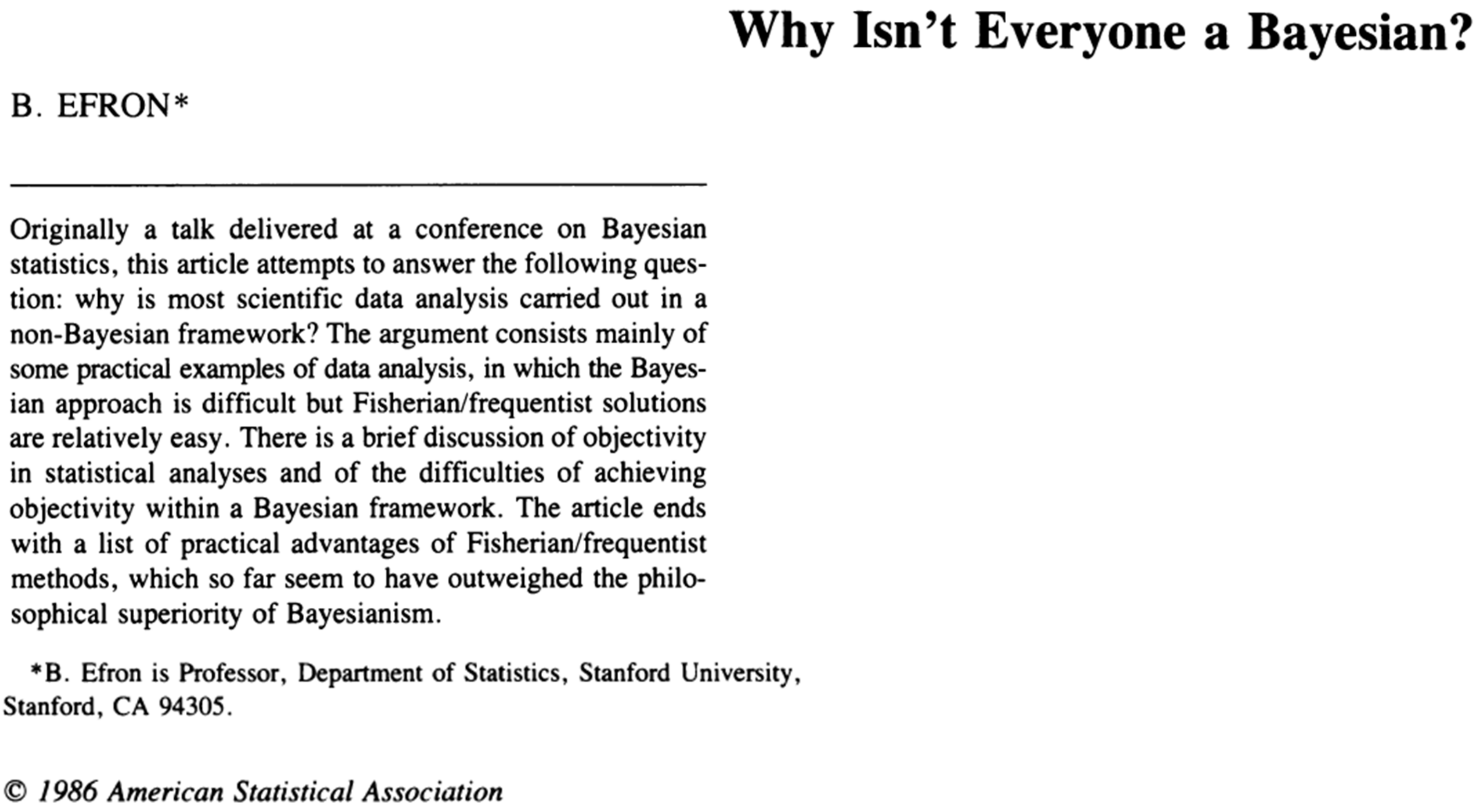
- **All of the information in the data is contained in the likelihood function.**

- This use of *p*-values violates the likelihood principle.
  - Our conclusions depend on results that *could have* happened in different repetitions of he trial.
  - The data isn't enough… we need to know what Persi *intended to do*.
  - … now which looks more "subjective"… frequentist or Bayesian?

  - … and think about the implications of this in the medical/clinical trials domain.

Why Isn’t Everyone a Bayesian?

Why Isn’t Everyone a Bayesian?



Why Isn’t Everyone a Bayesian?

B. EFRON*

Originally a talk delivered at a conference on Bayesian statistics, this article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends with a list of practical advantages of Fisherian/frequentist methods, which so far seem to have outweighed the philosophical superiority of Bayesianism.

*B. Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305.

Why Isn’t Everyone a Bayesian?

- Given that the Bayesian framework is so great, why isn’t it used more in practice?

- **Answer 1:** Actually, it is… things have changed rapidly.

- **Answer 2:** Thoughts on why frequentism has been dominant.
  - (Jim’s speculation): Cognitive biases… failures of probabilistic reasoning
    - E.g. the Monty Hall problem, the prosecutor’s fallacy, Kahneman’s blue taxis
  - Much of classical statistics is “automatic” in ways that can be programmed into canned software packages (PROCs).
  - Argument that Bayesian statistics is “subjective” and science isn’t “subjective”.
  - ***Bayesian computation has traditionally been very difficult.***
    - Pre-1990s: Bayesian practice was largely limited to ad hoc credibility formulas and conjugate prior relationships.

Bayesian Computation is Hard

- Remember Bayes' Theorem:

$$f(\theta \mid X) = \frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta)d\theta}$$

The great virtue of the Bayesian framework:

- It enables us to calculate a **predictive distribution** for future outcomes *Y* given past outcomes *X*: *f*(*Y*|*X*)
  - E.g. in loss reserving, we can get a predictive distribution of future claim payments *Y* given a loss triangle of past payments *X*.

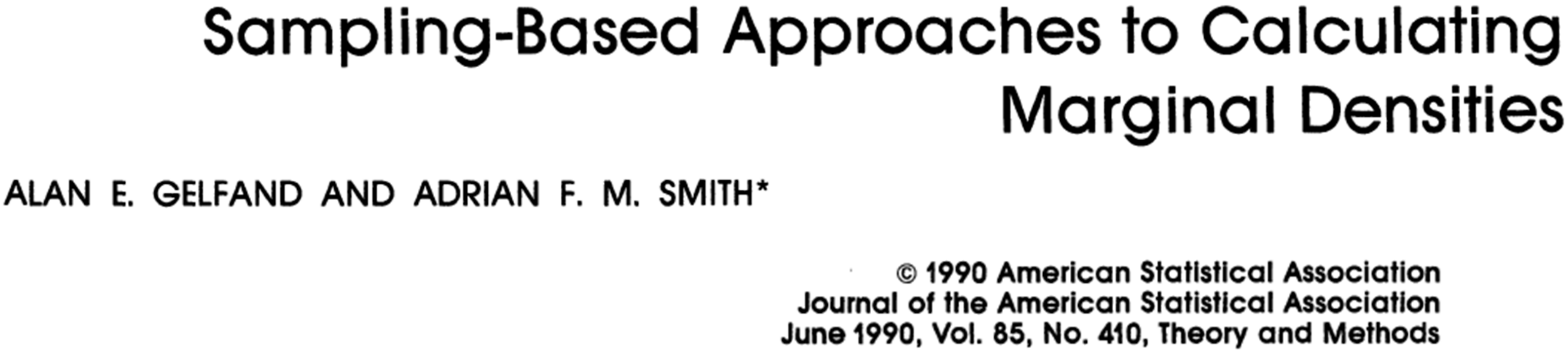
$$\boxed{f(Y \mid X) = \int f(Y \mid \theta) f(\theta \mid X) d\theta = \int f(Y \mid \theta) \left(\frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta)d\theta} \right) d\theta}$$

- But in practice all of this integration is intractable… impasse.

Bayesian Computation

A New World Order

- This impasse came to an end ~1990 when a simulation-based approach to estimating posterior probabilities was introduced.
  - (Circa the fall of the Soviet empire and Francis Fukuyama's "end of history")



Sampling-Based Approaches to Calculating Marginal Densities

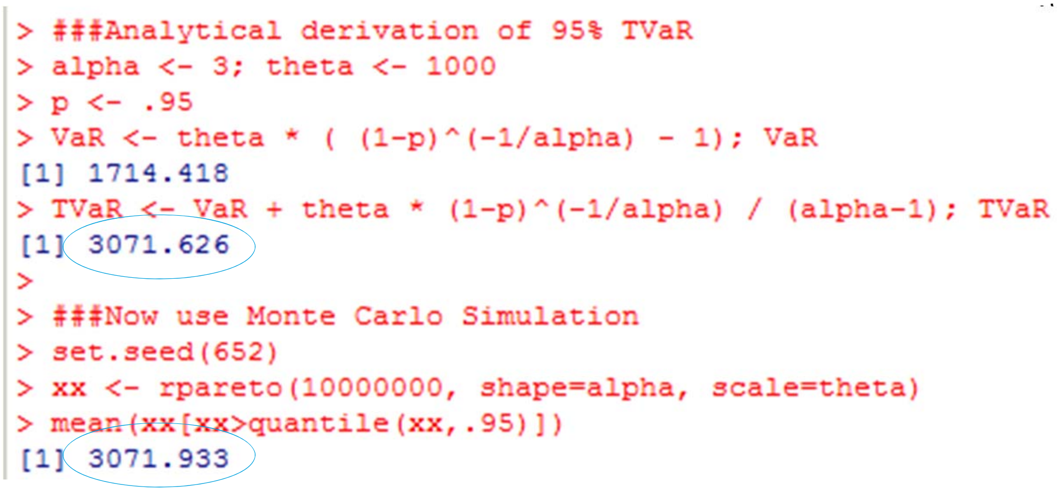
ALAN E. GELFAND AND ADRIAN F. M. SMITH*

© 1990 American Statistical Association
Journal of the American Statistical Association
June 1990, Vol. 85, No. 410, Theory and Methods

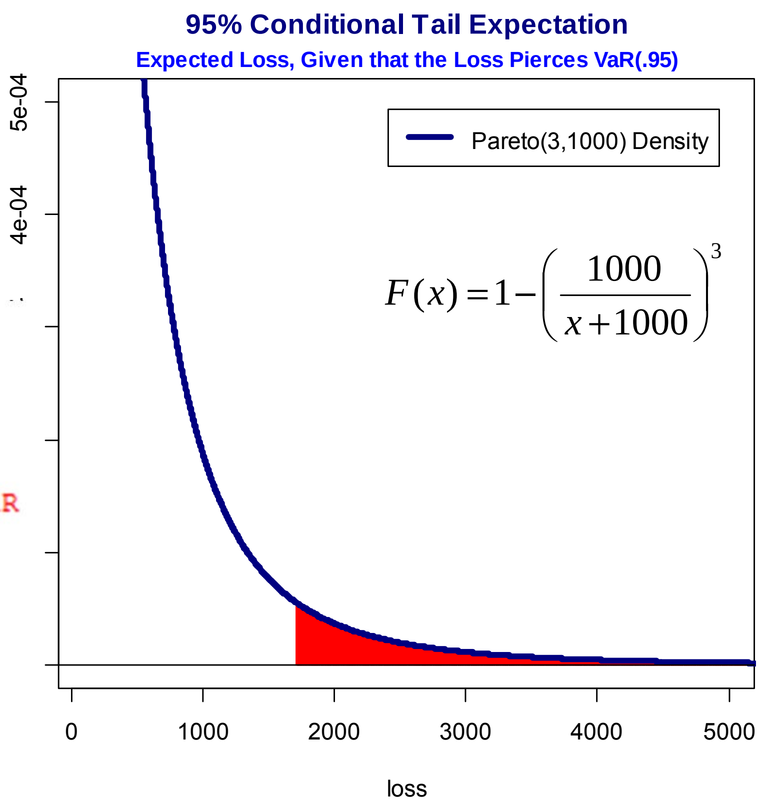
Monte Carlo Simulation – Review

- Recall that Monte Carlo simulation enables us to bypass tough integration problems by taking **independent samples** from the distribution and averaging over the samples.

- Easy example: 95% Conditional Tail Expectation (aka TVaR) for a Pareto(3,1000) distribution.



```
> ###Analytical derivation of 95% TVaR
> alpha <- 3; theta <- 1000
> p <- .95
> VaR <- theta * ( (1-p)^(-1/alpha) - 1); VaR
[1] 1714.418
> TVaR <- VaR + theta * (1-p)^(-1/alpha) / (alpha-1); TVaR
[1] 3071.626
>
> ###Now use Monte Carlo Simulation
> set.seed(652)
> xx <- rpareto(10000000, shape=alpha, scale=theta)
> mean(xx[xx>quantile(xx,.95)])
[1] 3071.933
```



Why Traditional Monte Carlo Isn’t Enough

- Monte Carlo simulation is all well and good when we can write down the probability distribution in a computer program.

- But the problem in Bayesian computation is that **we generally can’t write down an expression for the posterior probability distribution**!

- Specifically: the integral in the denominator gets very nasty very quickly… especially when θ is a vector of parameters…

$$\boxed{f(\theta \mid X) = \frac{f(X \mid \theta)\pi(\theta)}{\int f(X \mid \theta)\pi(\theta)d\theta}}$$

Metropolis-Hastings Sampling

A Random Walk Down Parameter Lane

- OK so we can't do Monte Carlo because in general we can't write down the posterior probability density $f(\theta|X)$.

- But what if we could set up a random walk through our parameter space that… in the limit… passes through each point in the probability space in proportion to the posterior probability density.

- **If we could**, then we could just use the most recent *x*000 steps of that random walk as a good approximation of the posterior density…

- **Yes we can!**

Chains We Can Believe In

- The **Metropolis-Hastings sampler** generates a **Markov chain** $\{\theta_1, \theta_2, \theta_3, \ldots\}$ in the following way:

1. Time t=1: select a random initial position θ_1 in parameter space.
2. Select a **proposal distribution** $p(\theta)$ that we will use to select proposed random steps away from our current position in parameter space.
3. Starting at time t=2: repeat the following until you get convergence:
   a) At step t, generate a proposed $\theta^* \sim p(\theta)$
   b) Also generate $u \sim \text{unif}(0,1)$
   c) If $R > u$ then $\theta_t = \theta^*$. Else, $\theta_t = \theta_{t-1}$.

$$\boxed{R = \frac{f(\theta^* \mid X)}{f(\theta_{t-1} \mid X)} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})}}$$

- Step 3c) implies that at step t, we accept the proposed step θ^* with probability $\min(1,R)$.

Bayesian Computation is Easy?

- At each step we flip a coin with probability of heads min(1,*R*) and accept θ* if the coin lands heads.
  - Otherwise reject θ* and stay put at $\theta_{t\text{-}1}$.

- But why is this any easier? *R* contains the dreaded posterior density *f*(θ|X) that we can't write down.

$$R = \frac{f(\theta^* \mid X)}{f(\theta_{t-1} \mid X)} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{f(\theta^* \mid \theta_{t-1})}$$

- Here's why:

$$\begin{aligned} R &= \frac{f(\theta^* \mid X)\pi(\theta^*) \Big/ \int f(X \mid \vartheta)\pi(\vartheta)d\vartheta}{f(\theta_{t-1} \mid X)\pi(\theta_{t-1}) \Big/ \int f(X \mid \vartheta)\pi(\vartheta)d\vartheta} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})} \\ &= \frac{f(\theta^* \mid X)\pi(\theta^*)}{f(\theta_{t-1} \mid X)\pi(\theta_{t-1})} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})} \end{aligned}$$

Bayesian Computation is Easy

- At each step we flip a coin with probability of heads min(1,*R*) and accept θ* if the coin lands heads.
  - Otherwise reject θ* and stay put at $\theta_{t\text{-}1}$.

- But why is this any easier? *R* contains the dreaded posterior density f(θ|X) that we can't write down.

$$R = \frac{f(\theta^*\,|\,X)}{f(\theta_{t-1}\,|\,X)} \cdot \frac{p(\theta_{t-1}\,|\,\theta^*)}{f(\theta^*\,|\,\theta_{t-1})}$$

- Here's why:

The integrals in the denominator of Bayes theorem cancel out… they are functions only of the data *X*, not the parameters θ.

$$\begin{aligned} R &= \frac{f(\theta^*\,|\,X)\pi(\theta^*) \Big/ \cancel{\int f(X \quad (\vartheta)d\vartheta}}{f(\theta_{t-1}\,|\,X)\pi(\theta_{t-1}) \Big/ \cancel{\int f(\quad \pi(\vartheta)d\vartheta}} \cdot \frac{p(\theta_{t-1}\,|\,\theta^*)}{p(\theta^*\,|\,\theta_{t-1})} \\ &= \frac{f(\theta^*\,|\,X)\pi(\theta^*)}{f(\theta_{t-1}\,|\,X)\pi(\theta_{t-1})} \cdot \frac{p(\theta_{t-1}\,|\,\theta^*)}{p(\theta^*\,|\,\theta_{t-1})} \end{aligned}$$

Now We Can Go to the Metropolis

- So now we have something we can easily program into a computer.

- At each step, give yourself a coin with probability of heads min(1,*R*) and flip it.

$$R = \frac{f(\theta^* \mid X)\pi(\theta^*)}{f(\theta_{t-1} \mid X)\pi(\theta_{t-1})} \cdot \frac{p(\theta_{t-1} \mid \theta^*)}{p(\theta^* \mid \theta_{t-1})}$$

- If the coin lands heads move from $\theta_{t\text{-}1}$ to θ^*

- Otherwise, stay put.

- The result is a Markov chain (step *t* depends only on step *t*-1… not on prior steps). And it converges on the posterior distribution.

Simple Illustration

- Let’s illustrate MH via a simple example.
- “Target” density that we want to simulate: the lognormal.

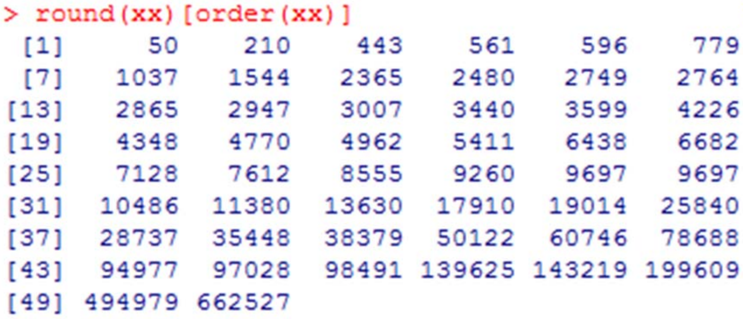
$$\boxed{f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-z^2 / 2\right) \quad , \quad z = \frac{\ln(x) - \mu}{\sigma}}$$

- We take logs so that we add/subtract rather than multiply/divide
- “Target” “density”: $tgt(x, \mu, \sigma) = -\ln(\sigma) - 0.5 * \left(\frac{\log(x) - \mu}{\sigma}\right)^2$
  - As noted before, we can eliminate terms that cancel out
- Proposal densities: $p(\mu^* \mid \mu_{t-1}) = N(\mu_{t-1}, 1) \quad ; \quad p(\sigma^* \mid \sigma_{t-1}) = N(\sigma_{t-1}, 1)$
  - The proposal (μ^*,σ^*) is a standard normal step away from the current location.

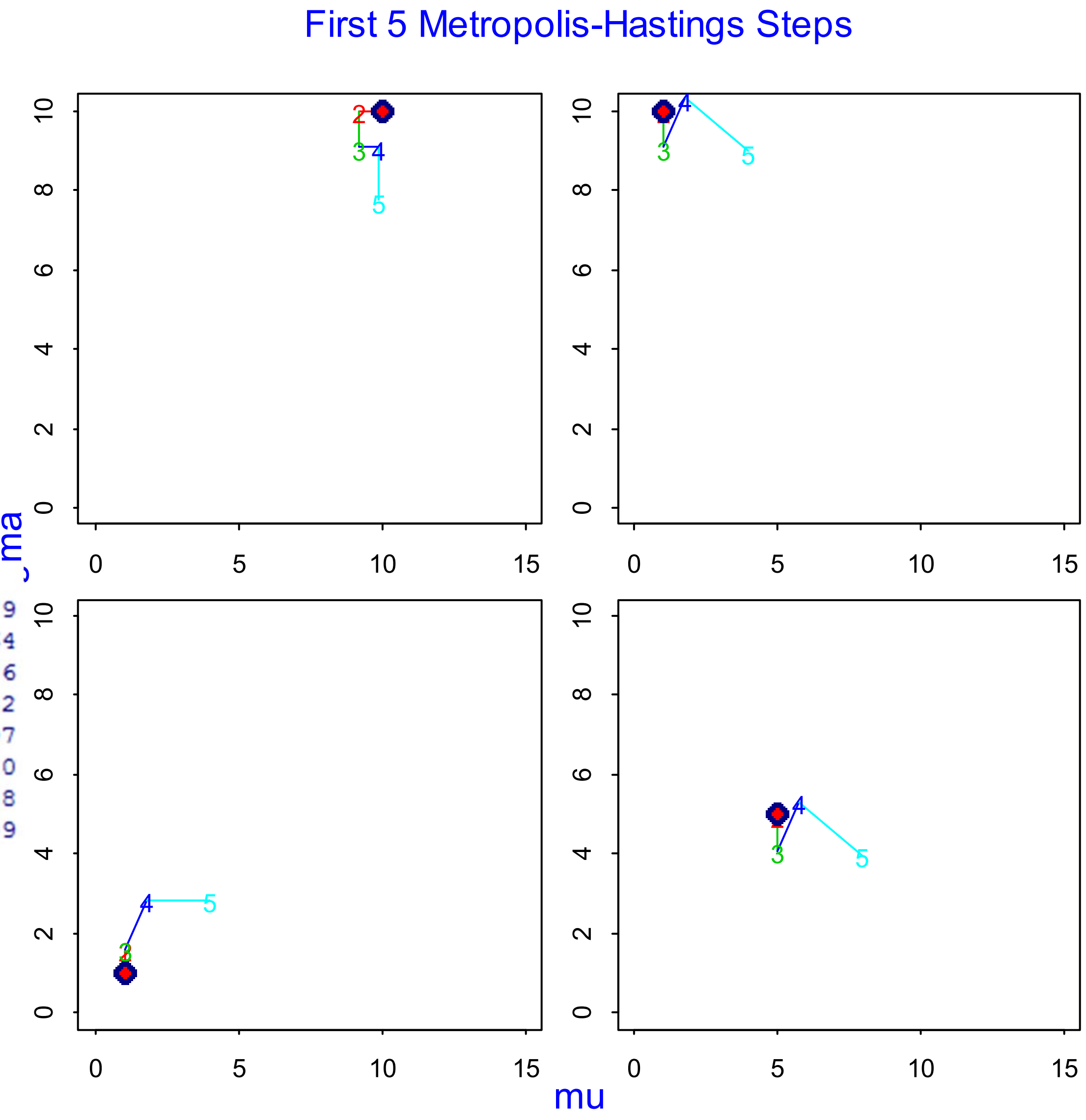
Random Walks with 4 Different Starting Points

- We estimate the lognormal density using 4 separate sets of starting values.

- Data: 50 random draws from lognormal(9,2).

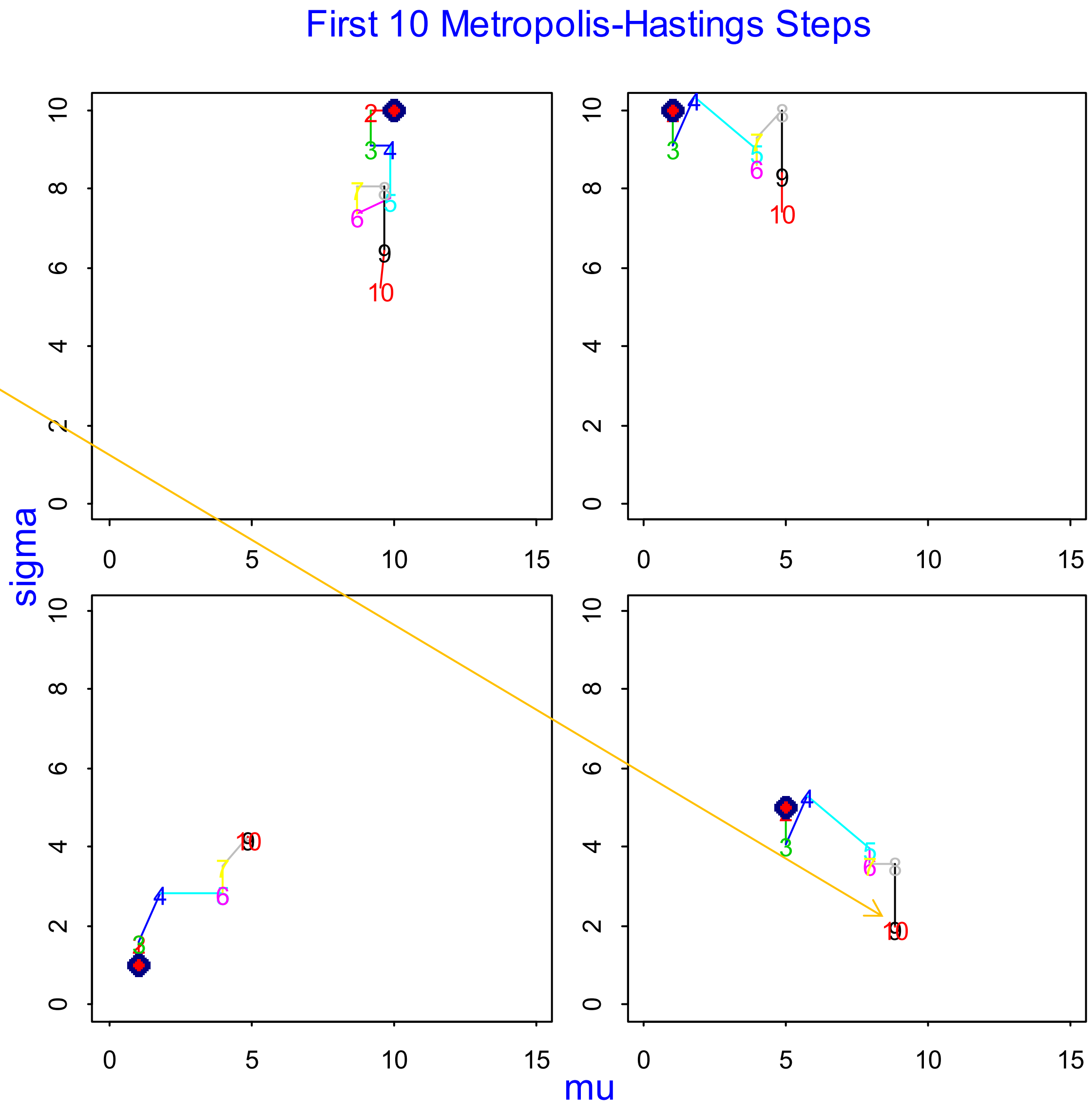


```
> round(xx)[order(xx)]
 [1]     50    210    443    561    596    779
 [7]   1037   1544   2365   2480   2749   2764
[13]   2865   2947   3007   3440   3599   4226
[19]   4348   4770   4962   5411   6438   6682
[25]   7128   7612   8555   9260   9697   9697
[31]  10486  11380  13630  17910  19014  25840
[37]  28737  35448  38379  50122  60746  78688
[43]  94977  97028  98491 139625 143219 199609
[49] 494979 662527
```



Random Walks with 4 Different Starting Points

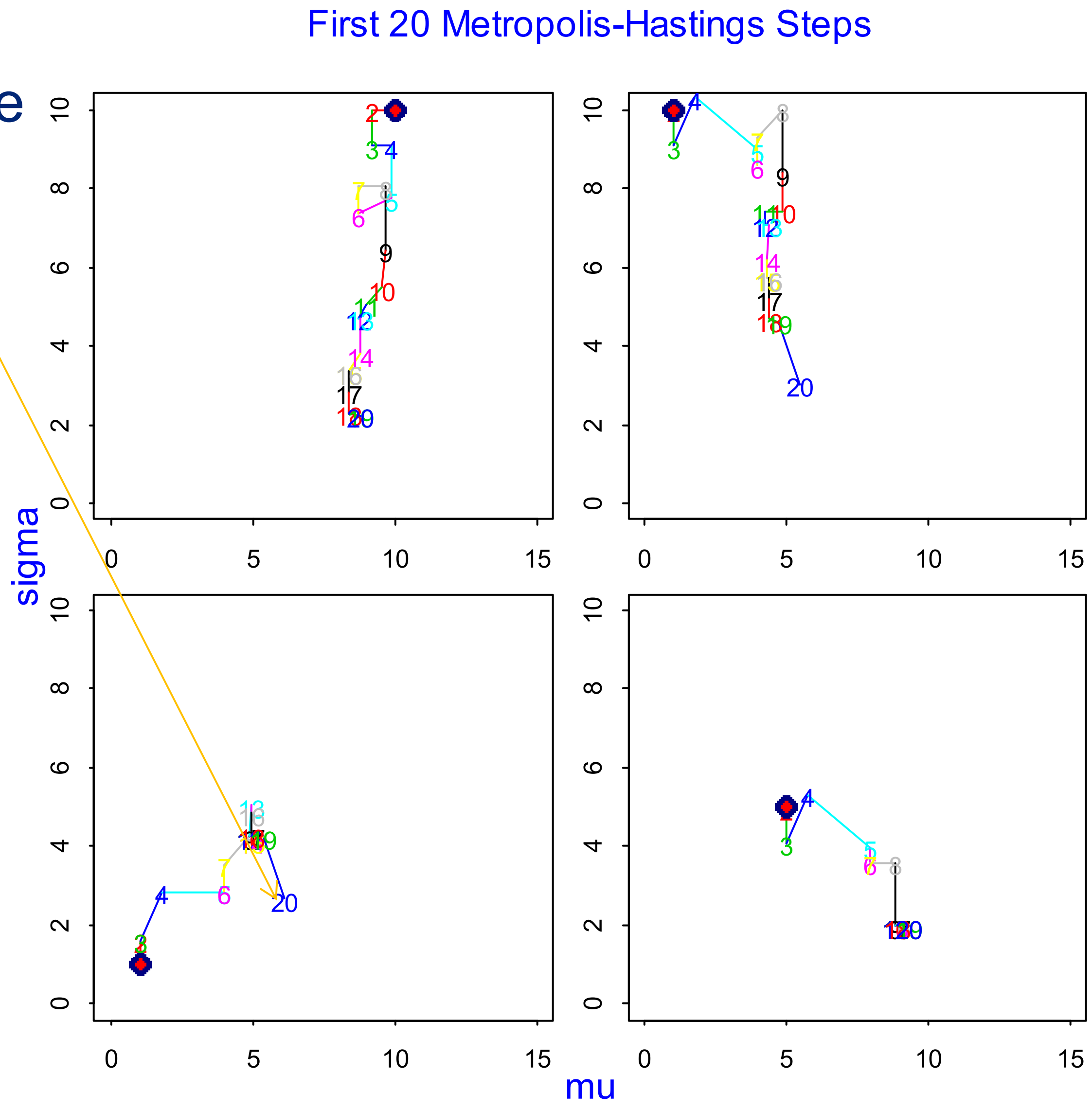
- After 10 iterations, the lower right chain is already in the right neighborhood.



First 10 Metropolis-Hastings Steps

Random Walks with 4 Different Starting Points

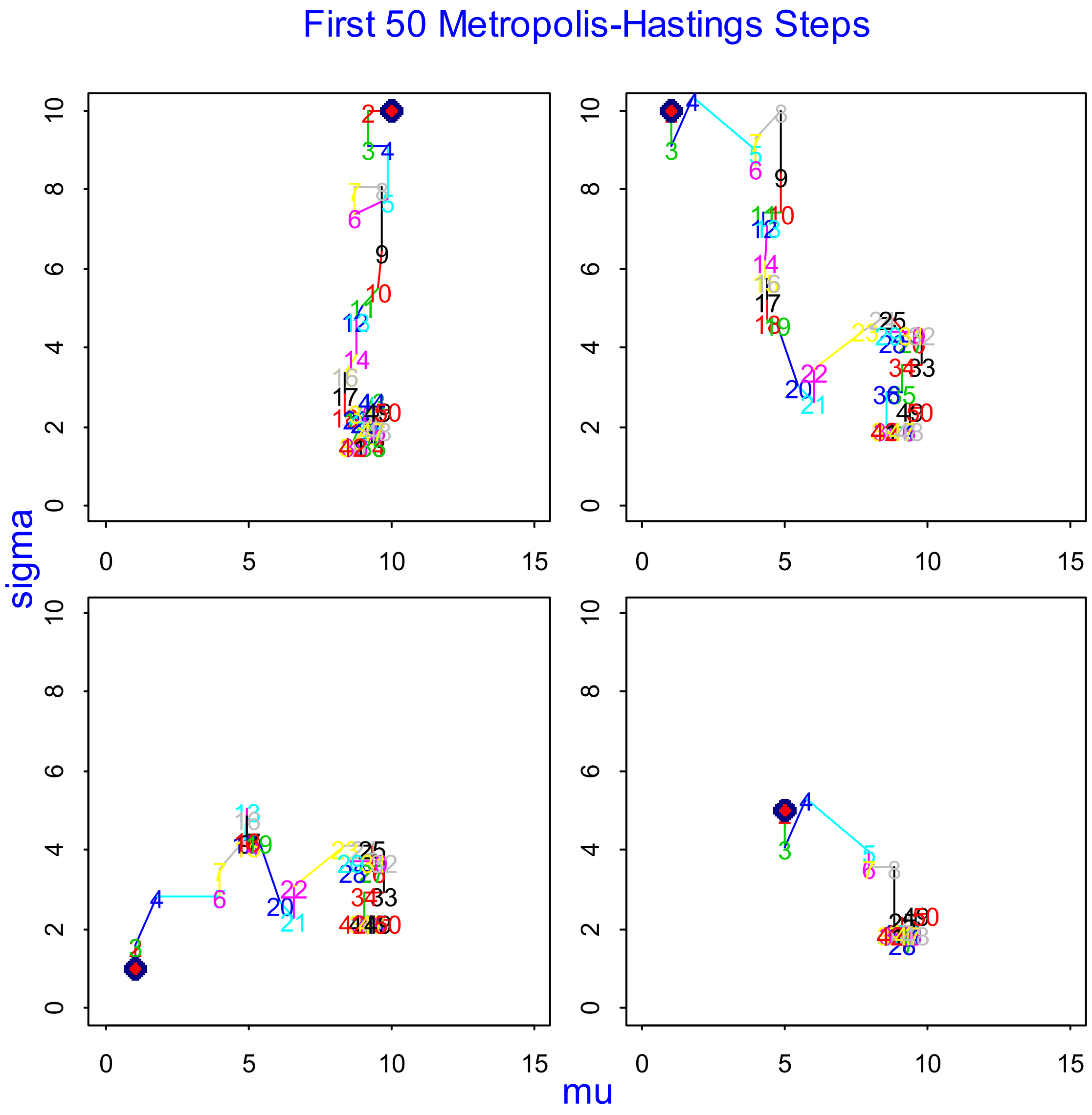
- After 20 iterations, only the 3rd chain is still in the wrong neighborhood.



First 20 Metropolis-Hastings Steps

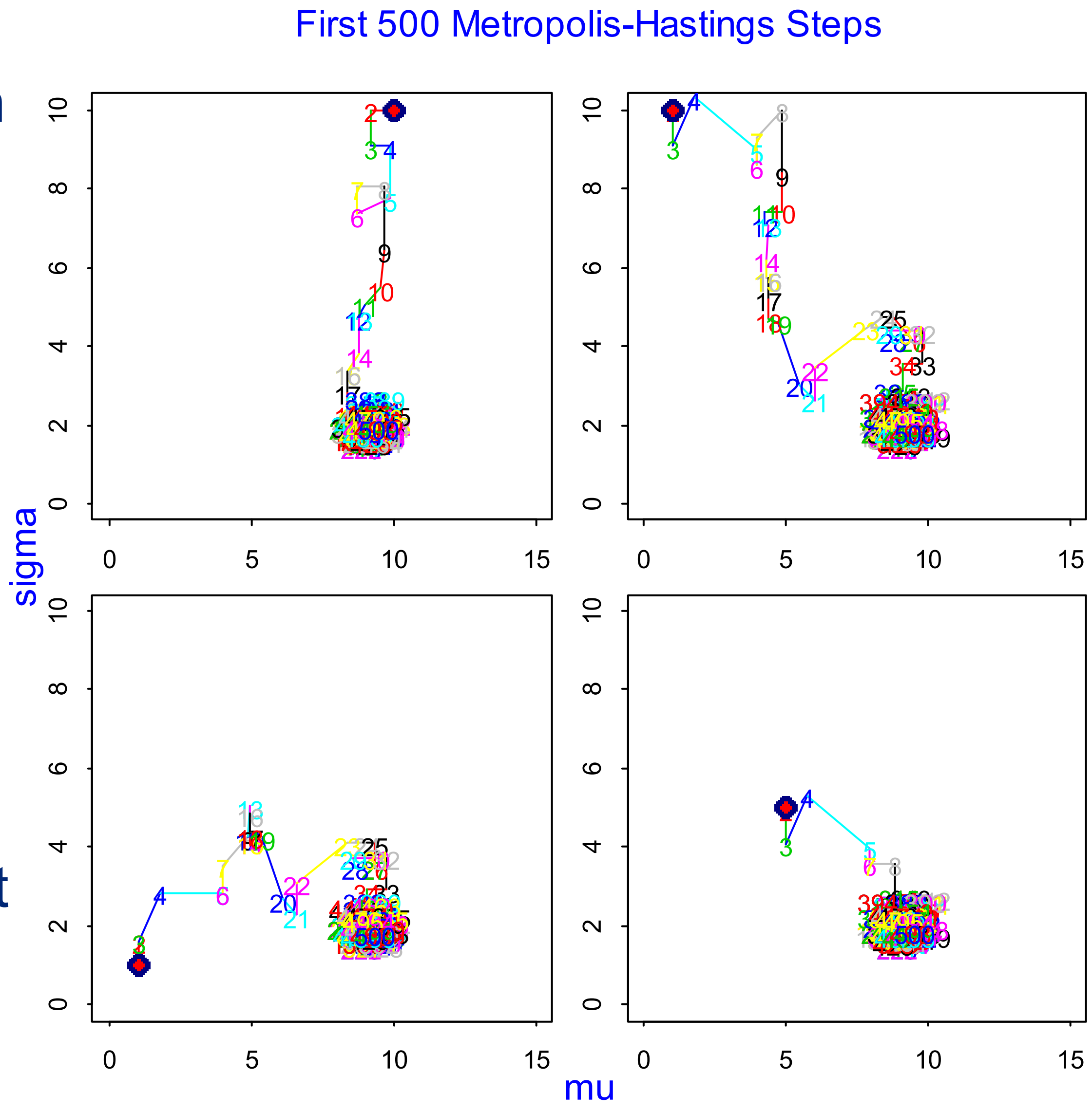
Random Walks with 4 Different Starting Points

- After 50 iterations, all 4 chains have arrived in the right neighborhood.



Random Walks with 4 Different Starting Points

- By 500 chains, it appears that the burn-in has long since been accomplished.
- The chain continues to wander.
- The time the chain spends in a neighborhood approximates the posterior probability that (μ,σ) lies in this nbd.



In 3D

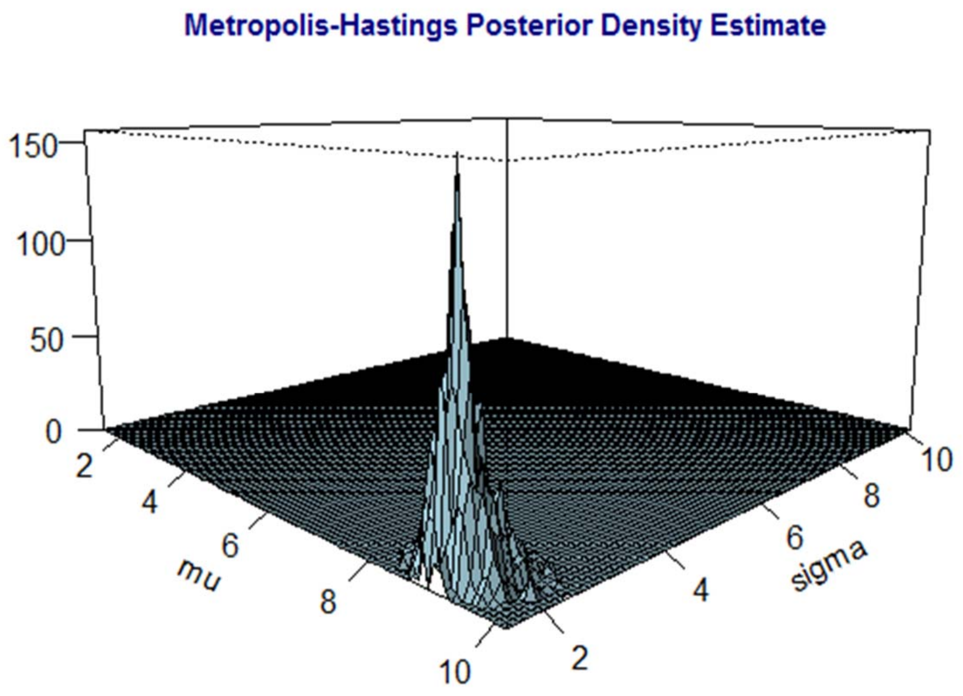
- The true lognormal parameters are:
  μ=9 and σ=2

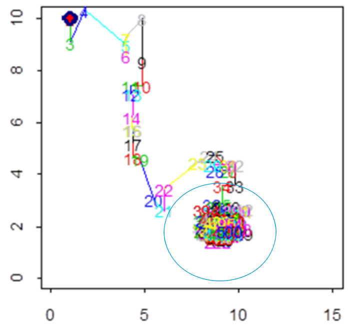
- The MH algorithm yields an estimate of the posterior density:

$$f(\mu, \sigma \mid X_1, X_2, \ldots, X_{50})$$

- This density results from a diffuse prior

- It is based on the informatior available in the data.

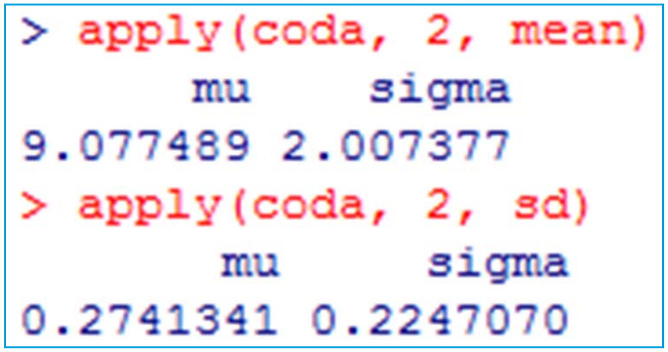

Metropolis-Hastings Posterior Density Estimate



Metropolis-Hastings Results

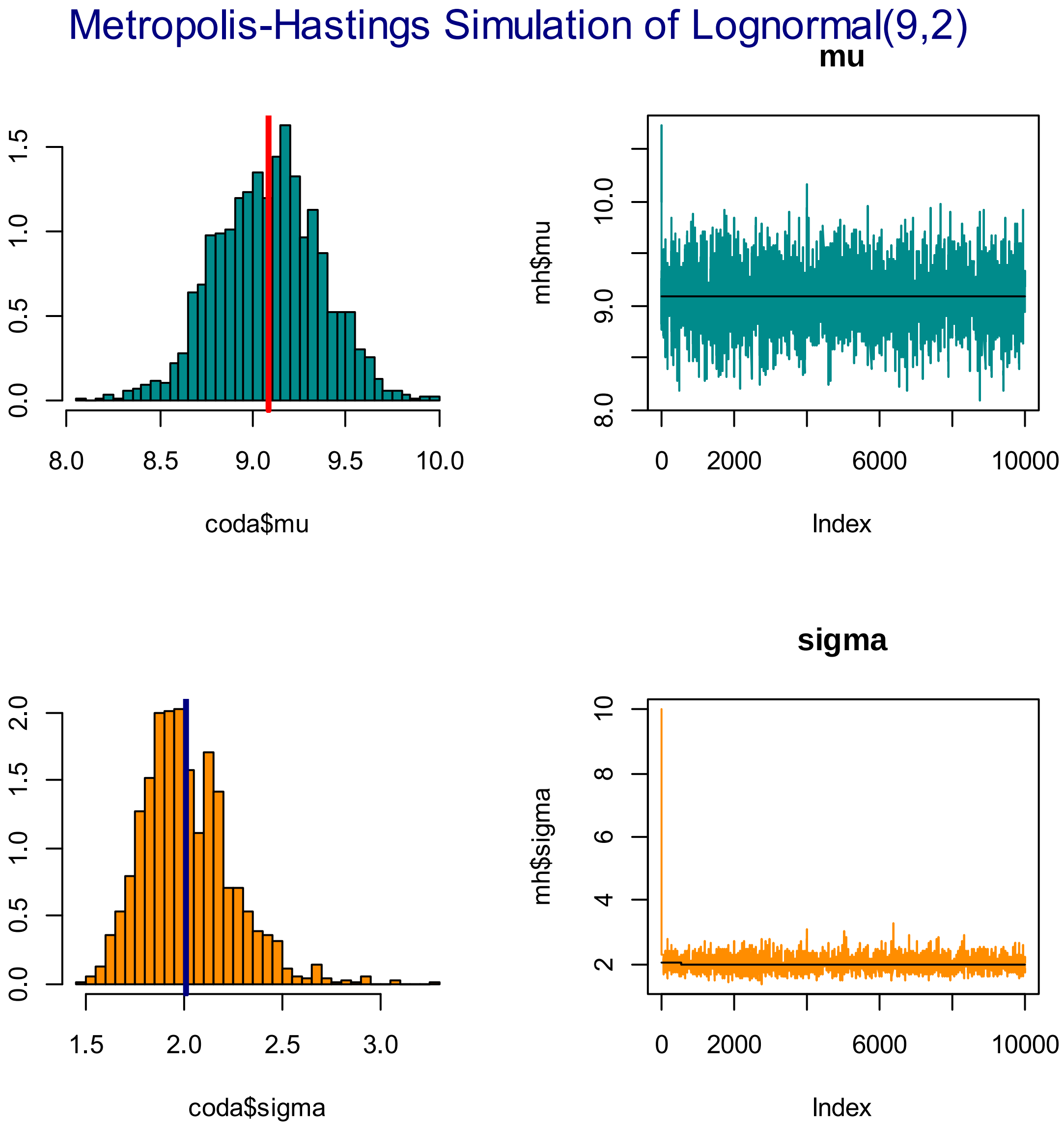
- The true lognormal parameters are:
  μ=9 and σ=2

- The MH simulation is gives consistent results:



```
> apply(coda, 2, mean)
      mu    sigma
9.077489 2.007377
> apply(coda, 2, sd)
       mu     sigma
0.2741341 0.2247070
```

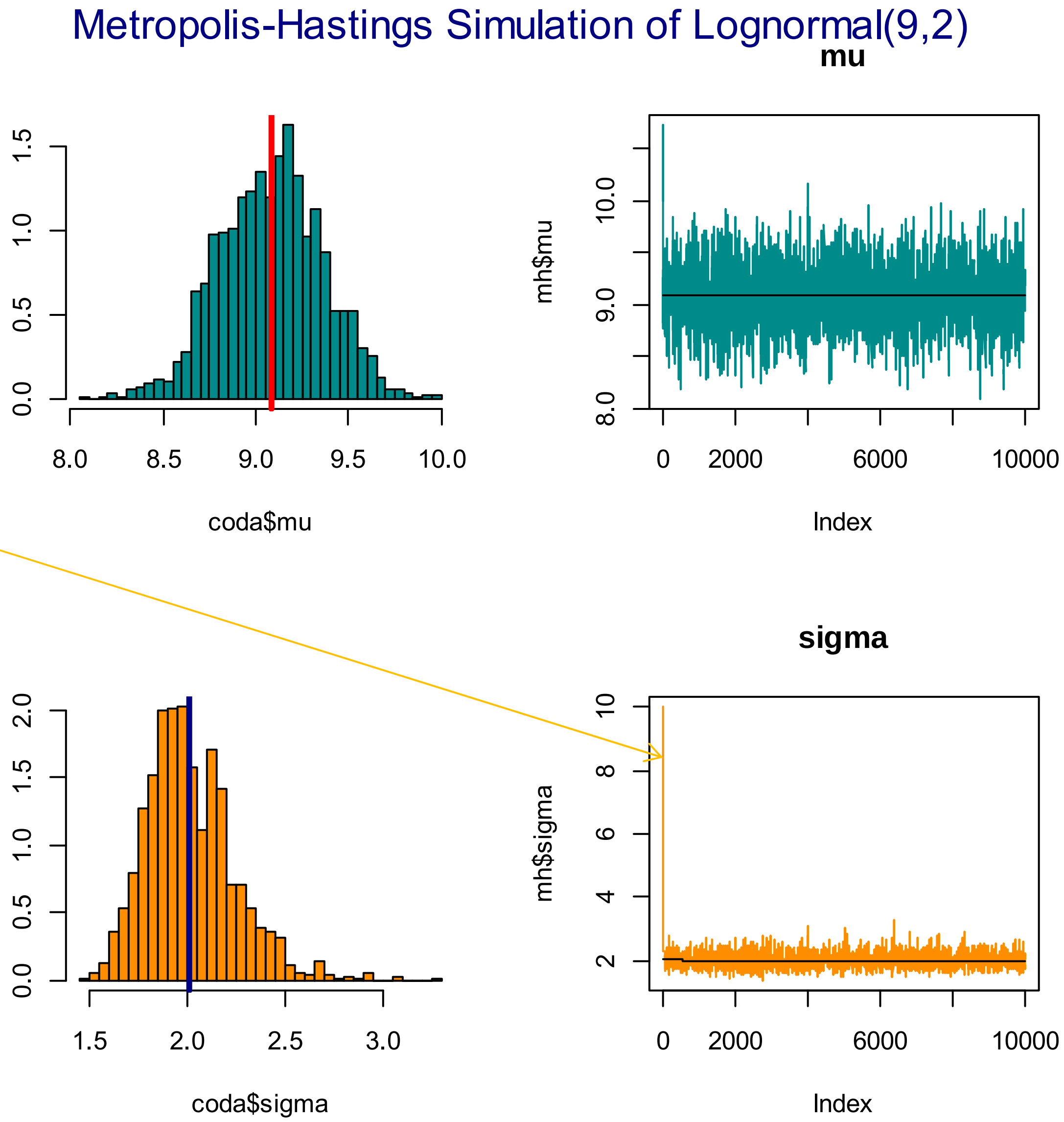
- Only the final 5000 of the 10000 MH iterations were used to estimate μ,σ



Metropolis-Hastings Results

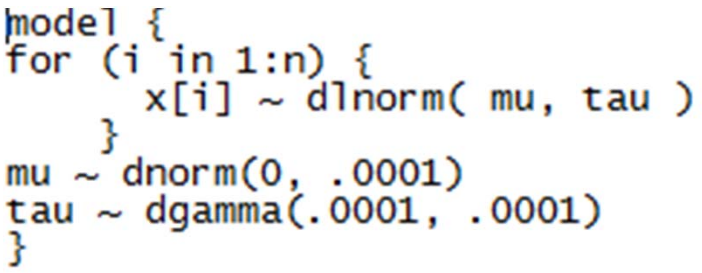
- The true lognormal parameters are:
  μ=9 and σ=2

- Note the very rapid convergence despite unrealistic initial values.

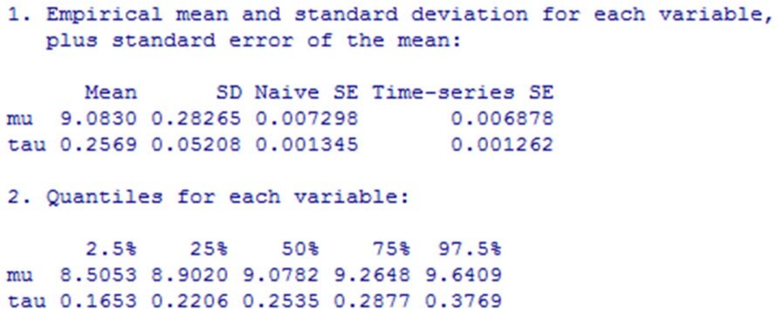


An Easier Way to Get the Same Result

- Call **JAGS** from within R



```
model {
for (i in 1:n) {
        x[i] ~ dlnorm( mu, tau )
      }
mu ~ dnorm(0, .0001)
tau ~ dgamma(.0001, .0001)
}
```

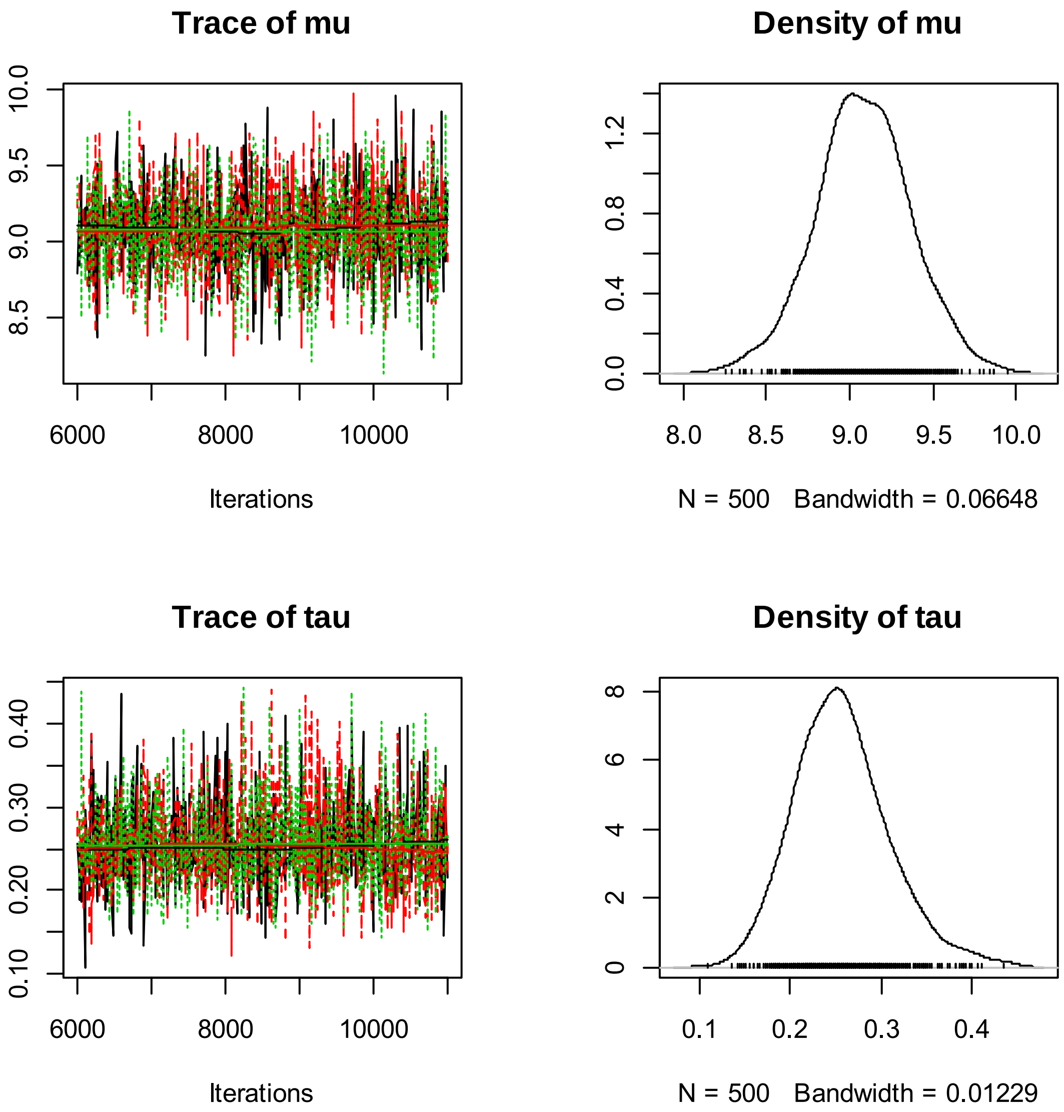


```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean      SD Naive SE Time-series SE
mu  9.0830 0.28265 0.007298       0.006878
tau 0.2569 0.05208 0.001345       0.001262

2. Quantiles for each variable:

      2.5%    25%    50%    75%  97.5%
mu  8.5053 8.9020 9.0782 9.2648 9.6409
tau 0.1653 0.2206 0.2535 0.2877 0.3769
```



Case Studies

Case Study #1

Fitting an Ambiguous Loss Model

JAGS: Just Another Gibbs Sampler

- Gibbs Sampling is a special case of Metropolis-Hastings sampling in which:
  - Each random draw is always accepted (faster convergence)
  - No need to specify a proposal density

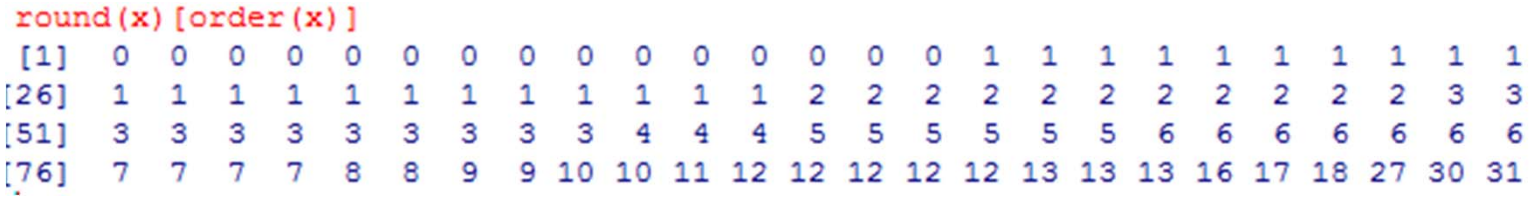
- Sequentially take draws from the conditional distributions. Continue until the chain settles down.

- The open-source packages BUGS and JAGS implement Gibbs sampling.
  - Specify the model in a high-level language
  - Call from within R

JAGS Case Study: Pareto Data

- Suppose we are given data for 100 losses and are told that they represent losses in $1M's for a new line of specialty insurance.

- We multiply the numbers by 10 for convenience:
  - (round the numbers only for display purposes… not in the analysis)



```
 round(x)[order(x)]
  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  1  1
 [26]  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  3  3
 [51]  3  3  3  3  3  3  3  3  3  4  4  4  5  5  5  5  5  5  6  6  6  6  6  6  6
 [76]  7  7  7  7  8  8  9  9 10 10 11 12 12 12 12 12 13 13 13 16 17 18 27 30 31
```

- We are asked to estimate the 99th percentile Value at Risk (VaR).

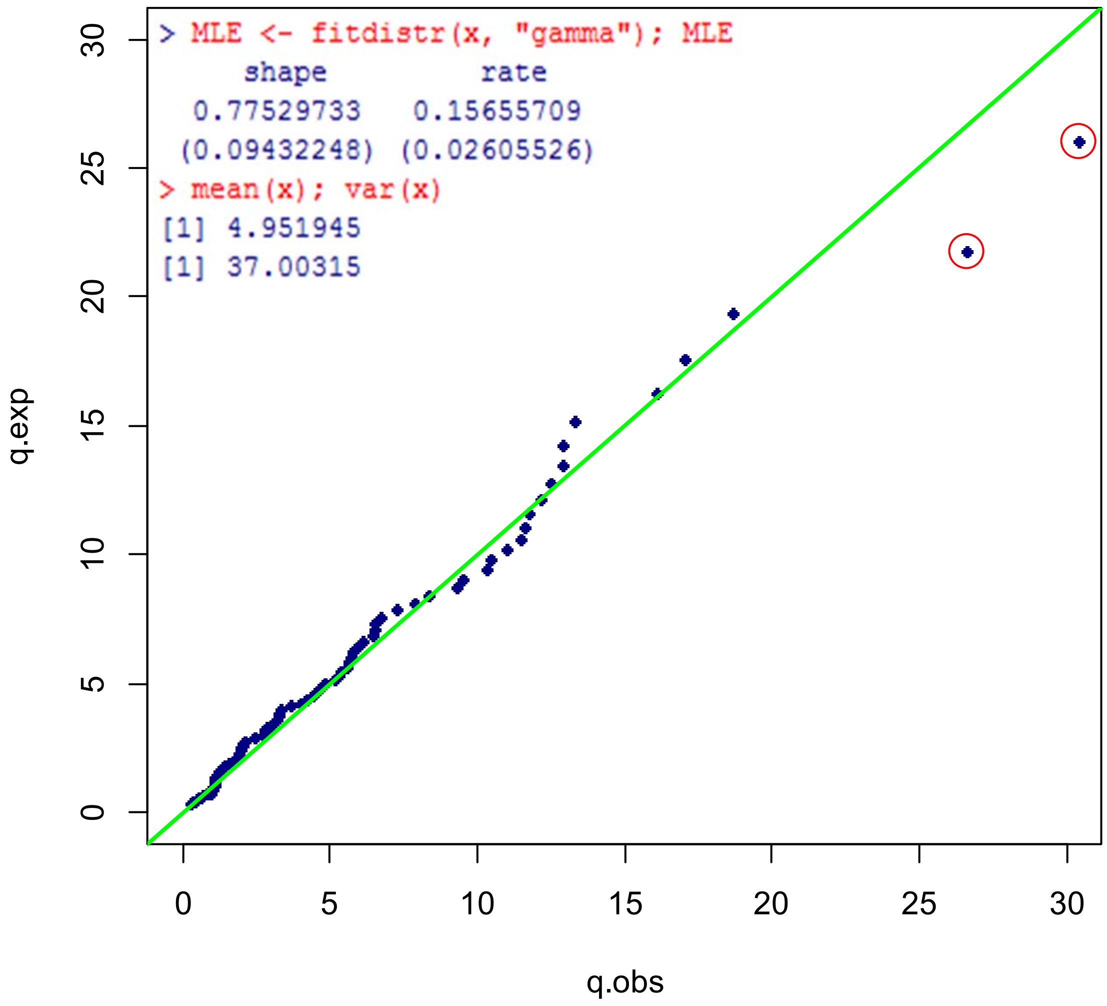
Exploratory Data Analysis

- Just to help visualize the data:
  - Perform gamma MLE fit
  - Create a QQ plot.

- Data doesn't look terribly inconsistent with a gamma…

- … but is this like concluding that the coin is biased after 12 tosses?

QQ Plot of Data Against MLE Gamma



```
> MLE <- fitdistr(x, "gamma"); MLE
     shape          rate
  0.77529733   0.15655709
 (0.09432248) (0.02605526)
> mean(x); var(x)
[1] 4.951945
[1] 37.00315
```

Thinking More About the Problem

- The scale parameter λ of our gamma(δ,λ) model is proportional to the $e^{\alpha+\beta 1X1+\beta 2X2+\ldots}$ from a gamma GLM.

- We're not given any covariates, but that doesn't mean that different risks don't have different expected loss amounts.

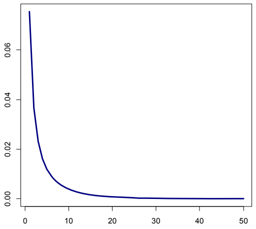
- So maybe we should let λ vary randomly: $\lambda \sim$ gamma(α,θ)

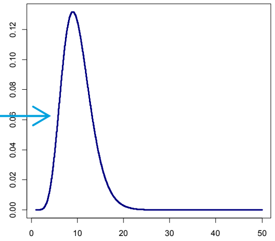
- And since we are uncertain about the values of δ,α,θ, we should specify prior distributions for them.

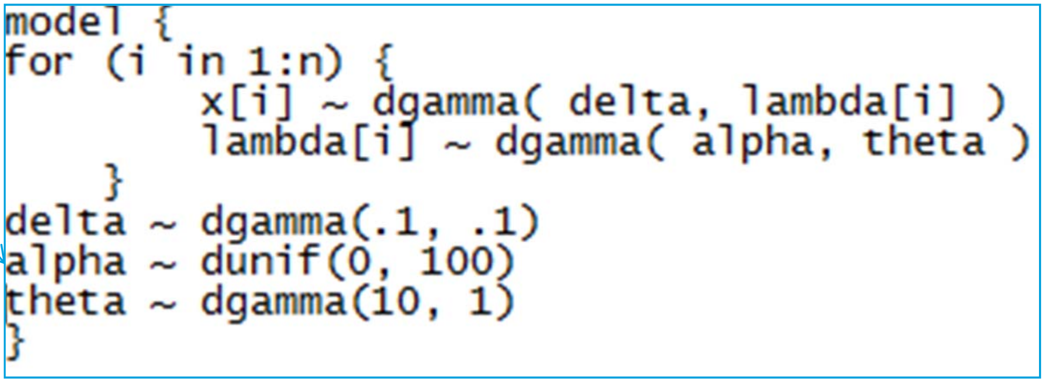
The Model

- We let λ vary randomly
  - This is assuming that losses are generated from a mixture of processes, each with a different innate expected size of loss.
  - Analogous to putting covariates in a Gamma GLM

- Other assumptions:
  - If δ=1 ➔ gamma mixture of exponentials ➔ Pareto(α,θ)
    - But rather than assume this, we put a diffuse distribution on δ.
  - Informative prior on θ reflects overall scale of the data.
  - Diffuse prior on α



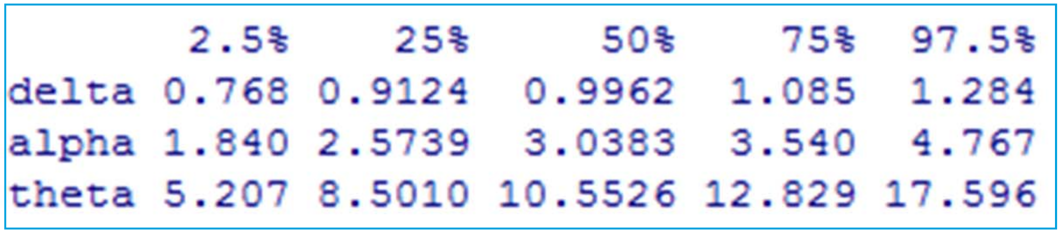




```
model {
for (i in 1:n) {
        x[i] ~ dgamma( delta, lambda[i] )
        lambda[i] ~ dgamma( alpha, theta )
    }
delta ~ dgamma(.1, .1)
alpha ~ dunif(0, 100)
theta ~ dgamma(10, 1)
}
```

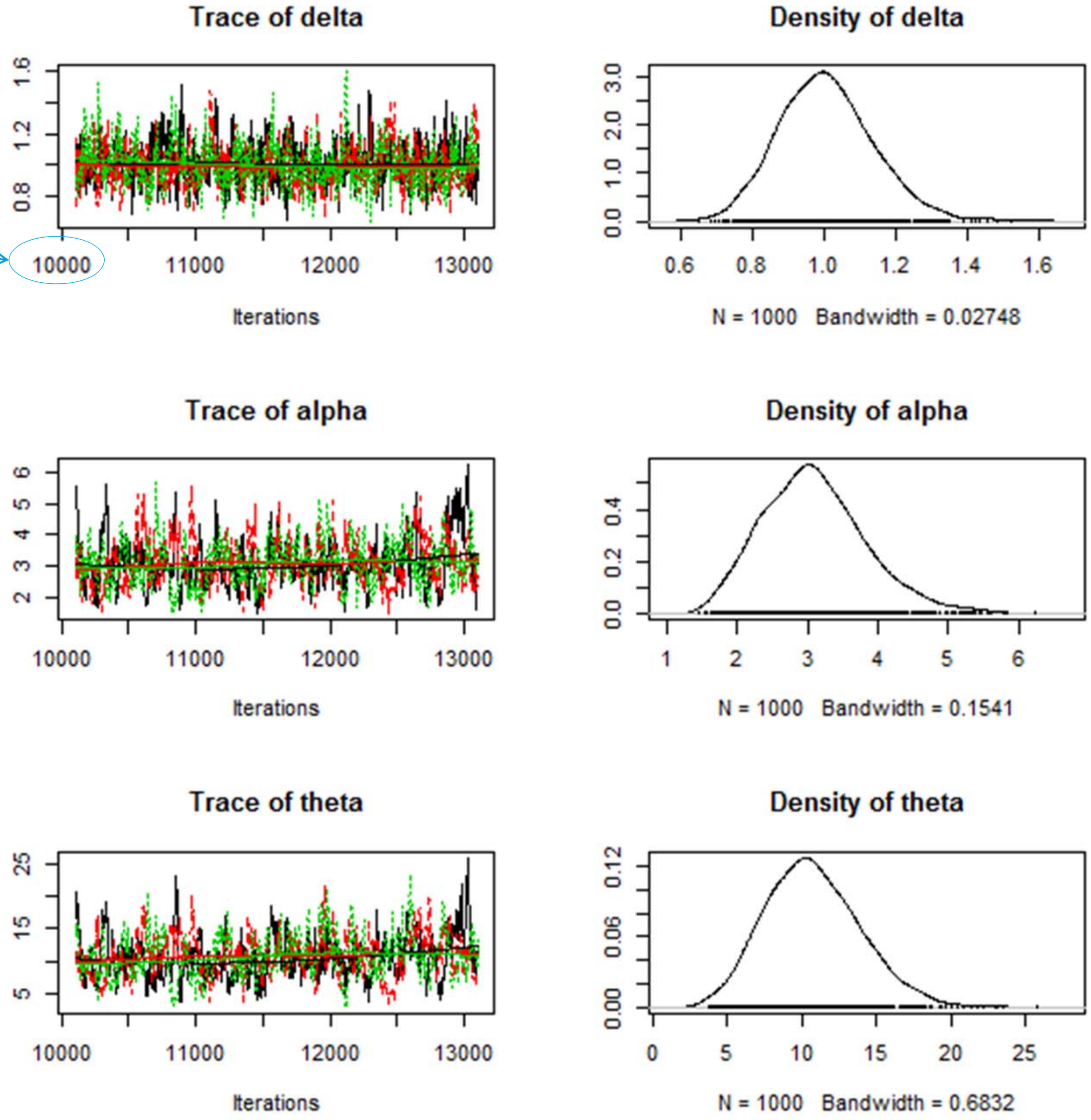
α is all-important… corresponds to dispersion in the underlying loss-generating processes.

Results

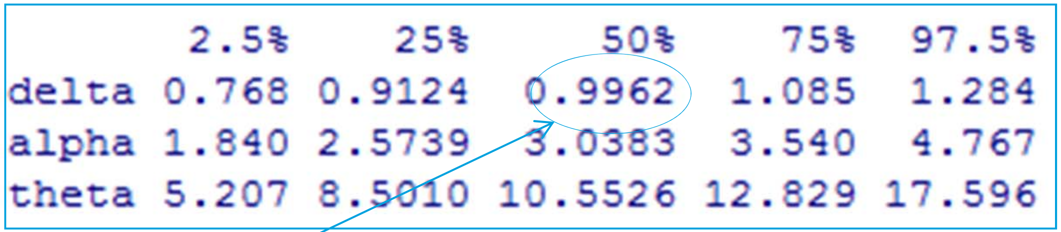


```
         2.5%    25%     50%    75%  97.5%
delta 0.768 0.9124  0.9962  1.085  1.284
alpha 1.840 2.5739  3.0383  3.540  4.767
theta 5.207 8.5010 10.5526 12.829 17.596
```

- Well, this is nice:
    - The 3 different random walks settled down after 10,000 burn-in iterations

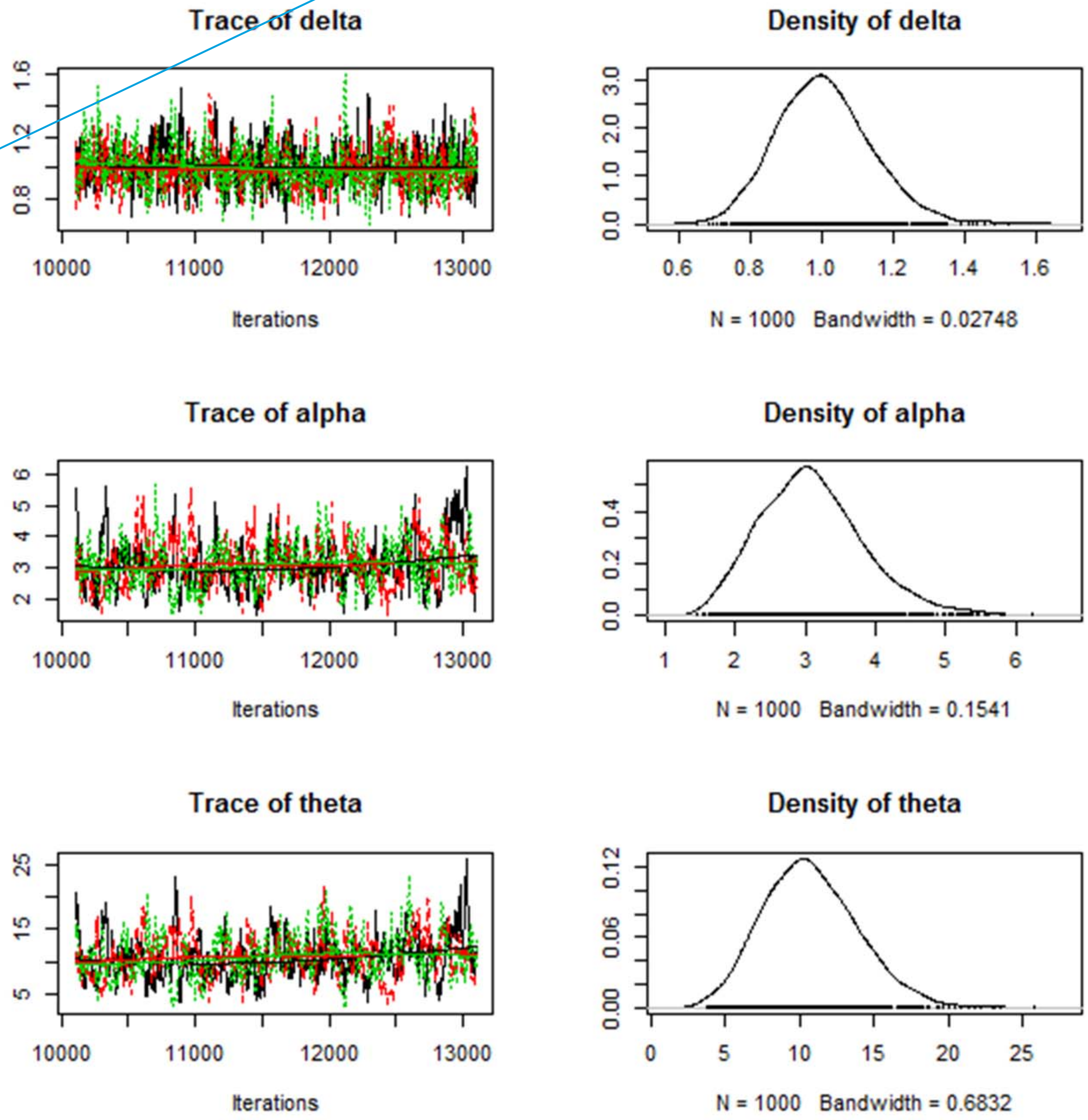


Results

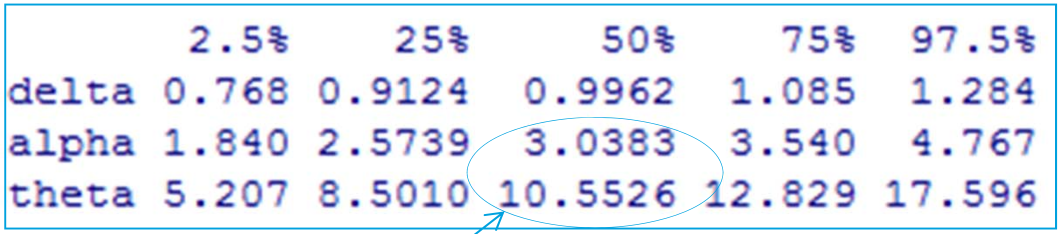


```
         2.5%    25%     50%     75%   97.5%
delta 0.768 0.9124  0.9962  1.085  1.284
alpha 1.840 2.5739  3.0383  3.540  4.767
theta 5.207 8.5010 10.5526 12.829 17.596
```

- Well, this is nice:
  - The 3 different random walks settled down after 10,000 burn-in iterations
  - Recall δ=1 implies a gamma mixtures of exponentials… which is Pareto.

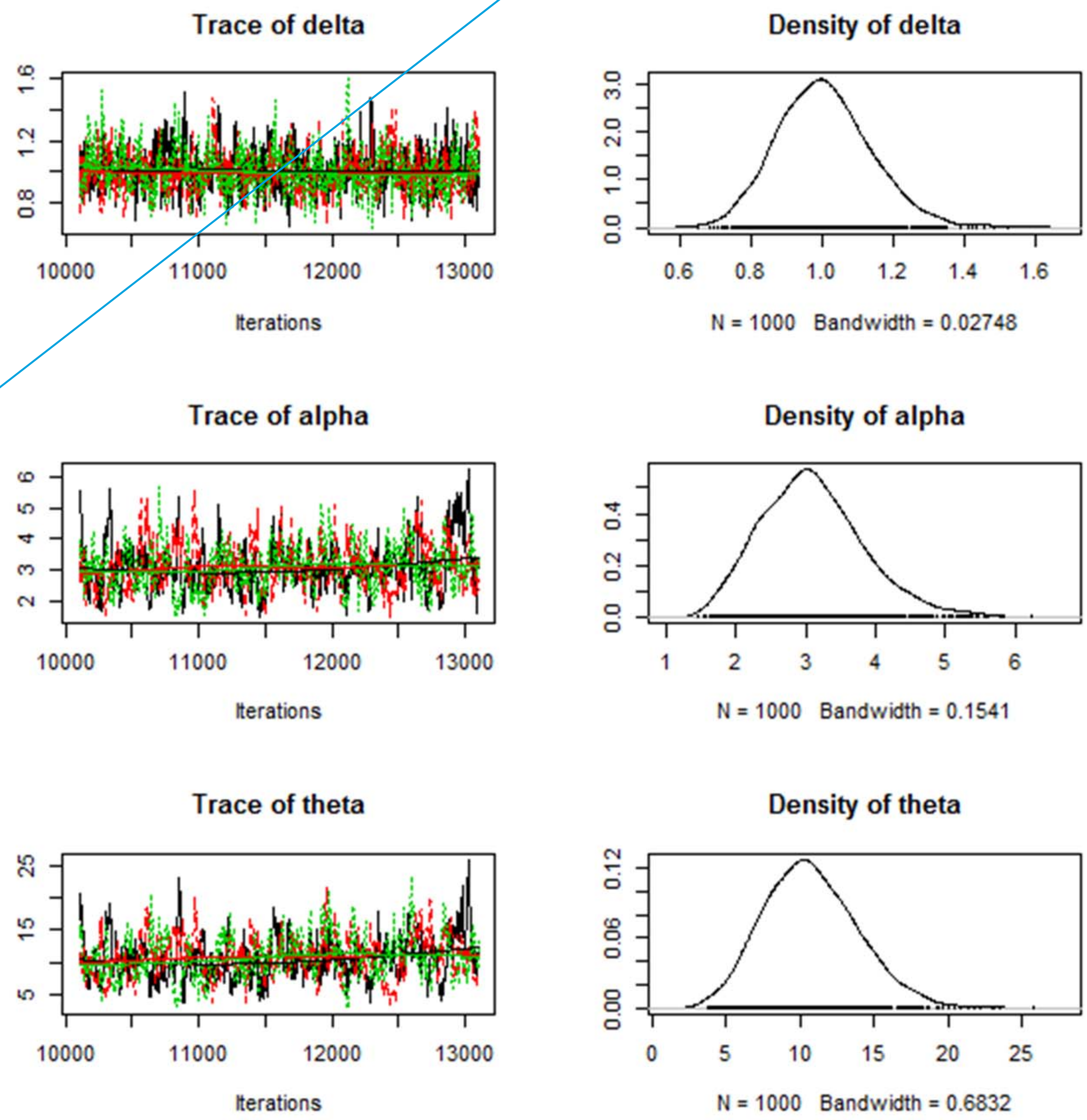


Results



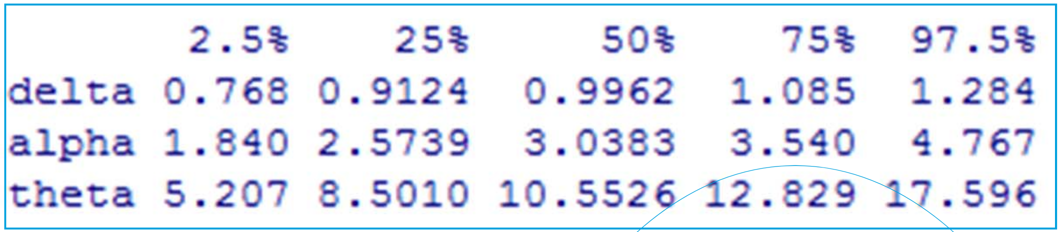
```
        2.5%    25%     50%    75%  97.5%
delta 0.768 0.9124  0.9962  1.085  1.284
alpha 1.840 2.5739  3.0383  3.540  4.767
theta 5.207 8.5010 10.5526 12.829 17.596
```

- Well, this is nice:
  - The 3 different random walks settled down after 10,000 burn-in iterations
  - Recall δ=1 implies a gamma mixtures of exponentials… which is Pareto.
  - The mean and variance c a Pareto (3,10) are 5 and 33.3 respectively… close to the data’s sample averages.

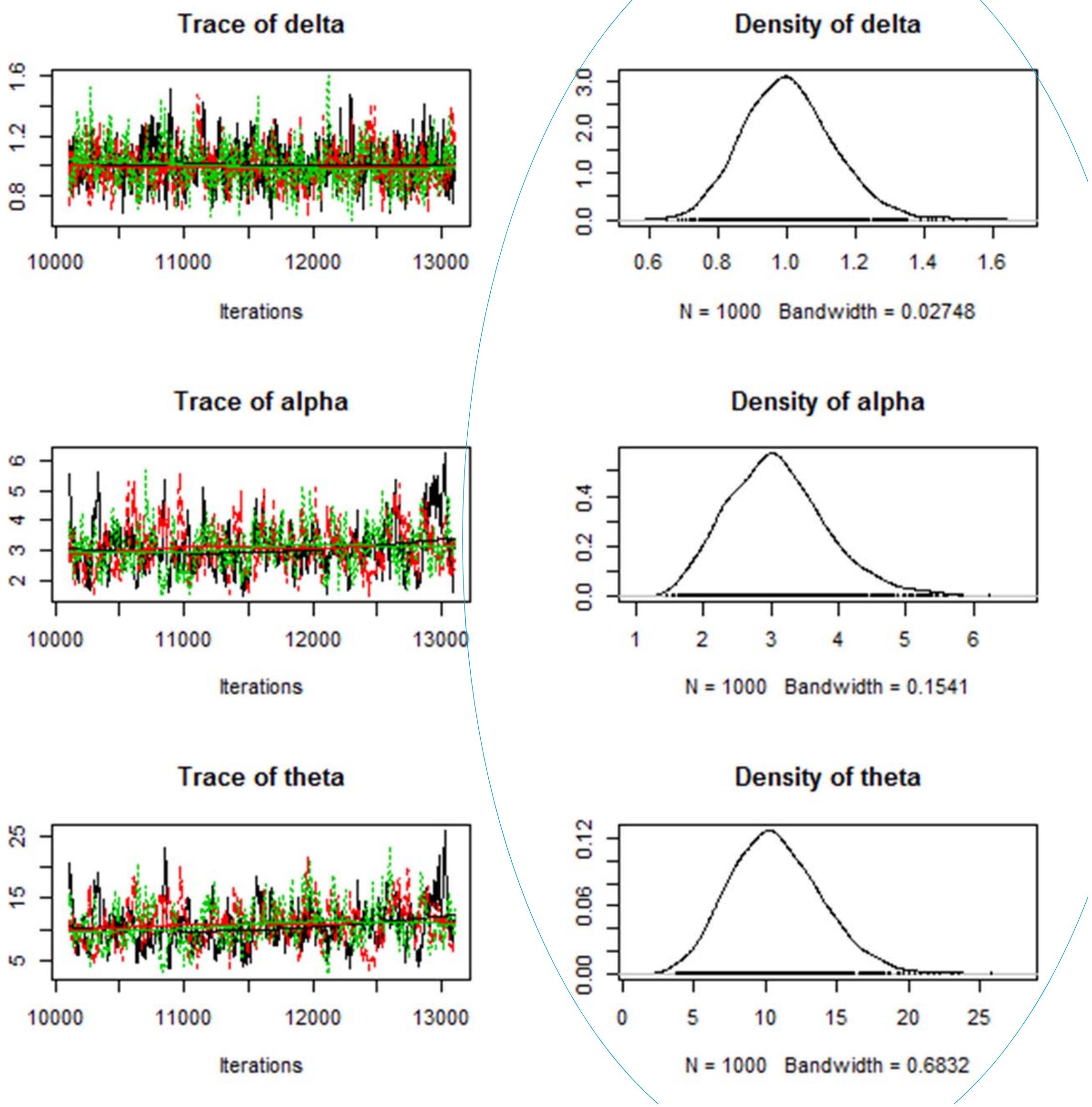


Results

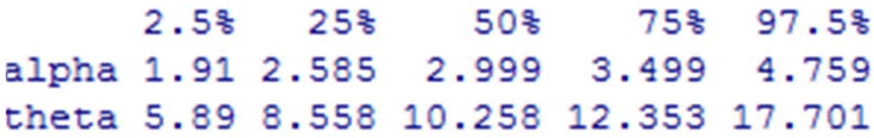
- Well, this is nice:
  - The 3 different random walks settled down after 10,000 burn-in iterations
  - Recall δ=1 implies a gamma mixtures of exponentials… which is Pareto.
  - The mean and variance c a Pareto (3,10) are 5 and 33.3 respectively… close to the data’s sample averages.
  - And we get a 3D posterior distribution… reflecting our uncertainty.



```
          2.5%    25%     50%     75%   97.5%
delta 0.768 0.9124  0.9962  1.085  1.284
alpha 1.840 2.5739  3.0383  3.540  4.767
theta 5.207 8.5010 10.5526 12.829 17.596
```



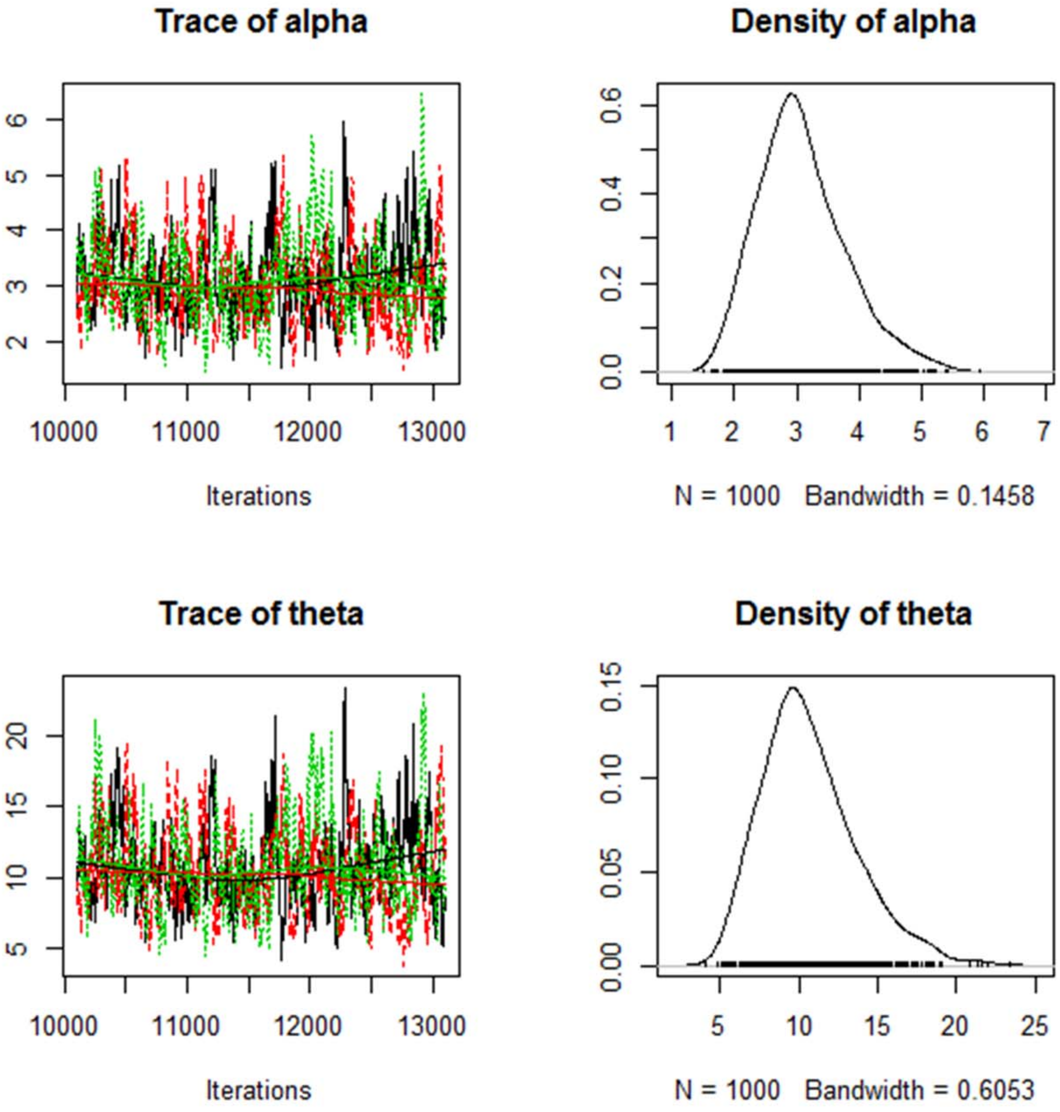
Now Simplify

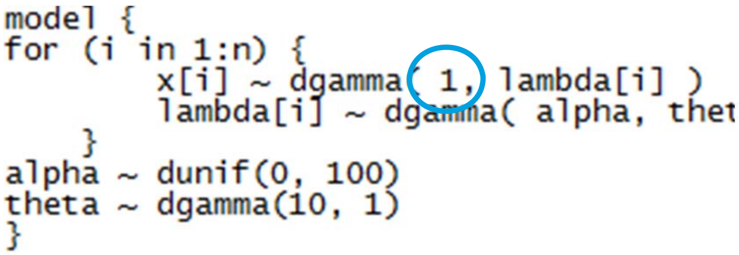


```
        2.5%   25%    50%    75%  97.5%
alpha 1.91 2.585  2.999  3.499  4.759
theta 5.89 8.558 10.258 12.353 17.701
```

- Let’s just assume that the data is Pareto (δ=1).
    - Purely for illustration
    - May be unjustified

- Rerunning the model yields broadly consistent results.



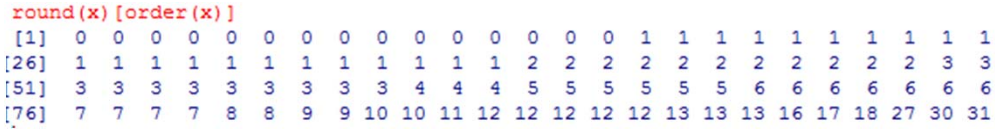


```
model {
for (i in 1:n) {
        x[i] ~ dgamma( 1, lambda[i] )
        lambda[i] ~ dgamma( alpha, thet
    }
alpha ~ dunif(0, 100)
theta ~ dgamma(10, 1)
}
```

Posterior Distribution VaR_{99} Estimates

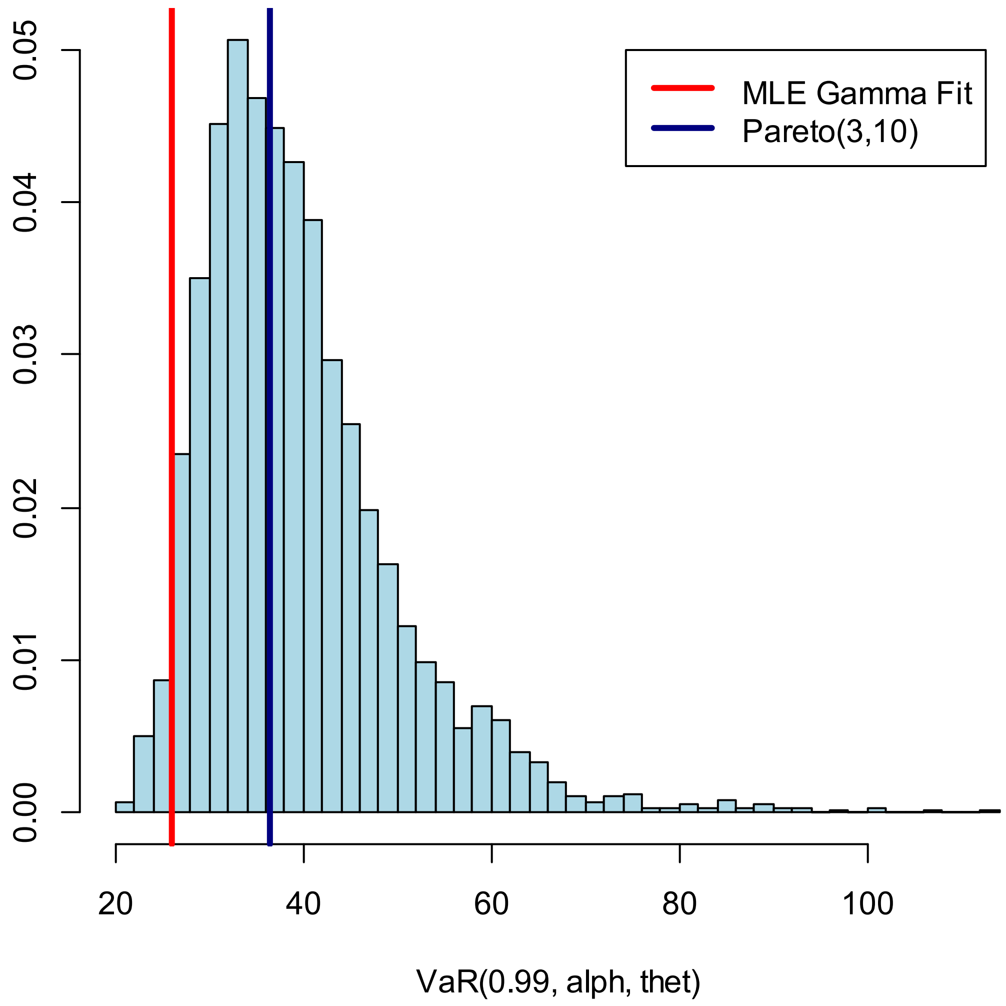
- If we had settled for our initial Gamma MLE fit, our estimate would have likely been way too low.

- Just reporting the VaR for a Pareto(3,10) fit doesn't tell the whole story either.
  - Parameter uncertainty results in widely divergent VaR estimates.
  - In real life, the next step would be to specify more informative priors…



```
round(x)[order(x)]
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  1  1
[26]  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  3  3
[51]  3  3  3  3  3  3  3  3  3  4  4  4  5  5  5  5  5  5  6  6  6  6  6  6  6
[76]  7  7  7  7  8  8  9  9 10 10 11 12 12 12 12 12 13 13 13 16 17 18 27 30 31
```

Estimated Bayesian Posterior Distribution of 99% VaR

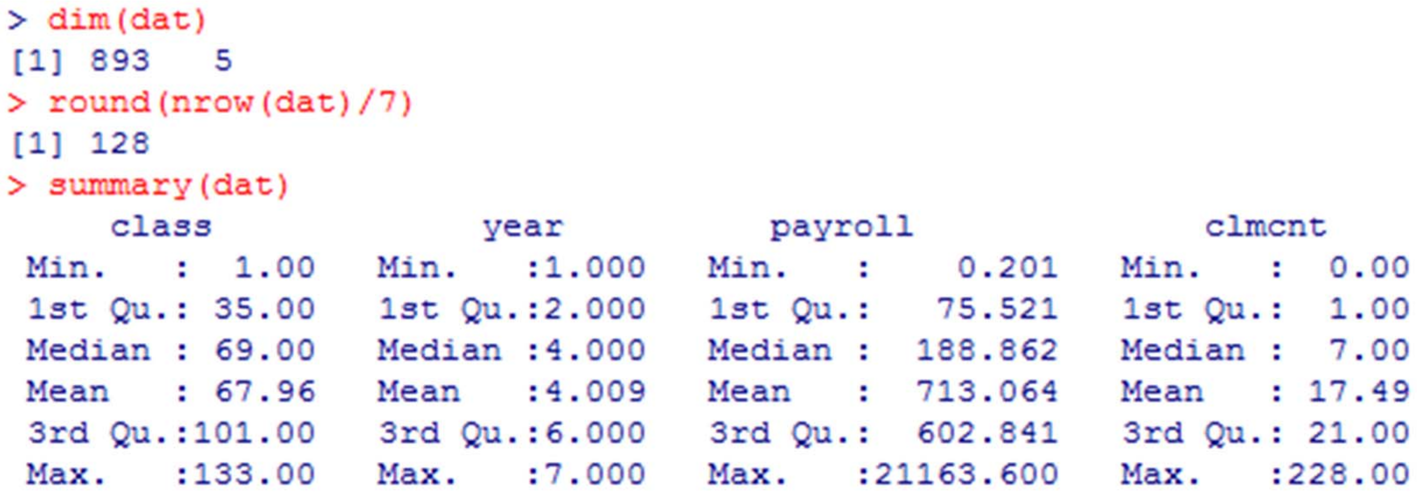


Case Study #2

Workers Comp Claim Frequency

Data and Problem

- We have 7 years of Workers Comp data
  - For each of 7 years we are given payroll and claim count by class.
  - Let’s build a Bayesian hierarchical Poisson GLM model on years 1-6 and compare the result with the actual claim counts from year 7.
  - Data is from Start Klugman 1992 book on Bayesian Statistics for actuarial science.



```
> dim(dat)
[1] 893   5
> round(nrow(dat)/7)
[1] 128
> summary(dat)
     class             year          payroll              clmcnt     
 Min.   :  1.00   Min.   :1.000   Min.   :    0.201   Min.   :  0.00  
 1st Qu.: 35.00   1st Qu.:2.000   1st Qu.:   75.521   1st Qu.:  1.00  
 Median : 69.00   Median :4.000   Median :  188.862   Median :  7.00  
 Mean   : 67.96   Mean   :4.009   Mean   :  713.064   Mean   : 17.49  
 3rd Qu.:101.00   3rd Qu.:6.000   3rd Qu.:  602.841   3rd Qu.: 21.00  
 Max.   :133.00   Max.   :7.000   Max.   :21163.600   Max.   :228.00  
```

Exploratory Data Analysis

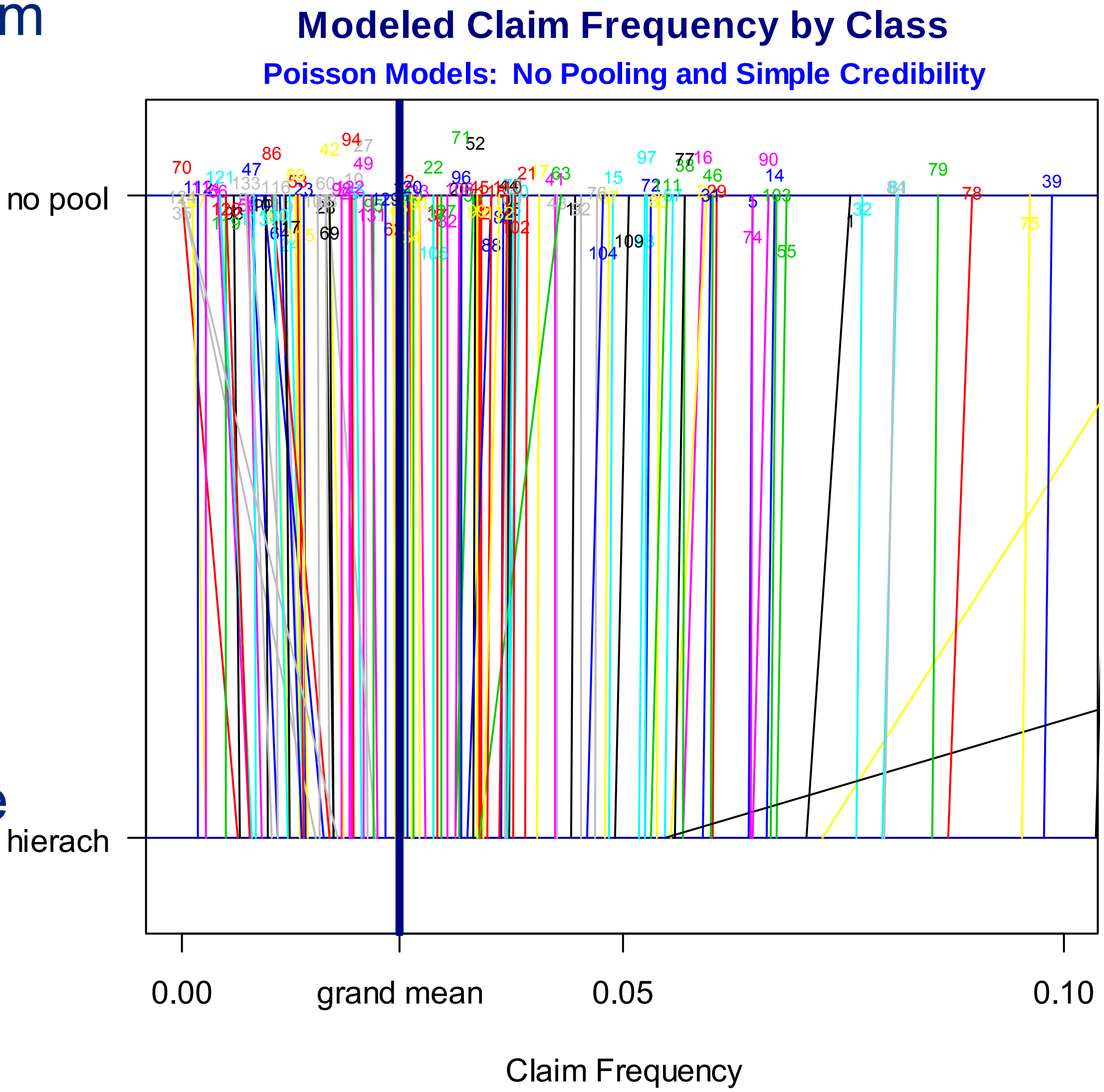
- The endgame is to build a Bayesian hierarchical GLM model.

- But in the spirit of data exploration, it makes sense to built empirical Bayes models first.
  - This is essentially a Bühlmann-Straub type credibility model.
  - This will help us get a feel for how much "shrinkage" (credibility-weighting) is called for.
  - Compare credibility weighted result with simply calculating empirical 6-year claim frequency by class.

$$\boxed{\begin{gathered} clmcnt_i \sim Poi\left(payroll_i \lambda_{j[i]}\right) \\ \lambda_j \sim N\left(\mu_\lambda, \sigma_\lambda^2\right) \end{gathered}}$$

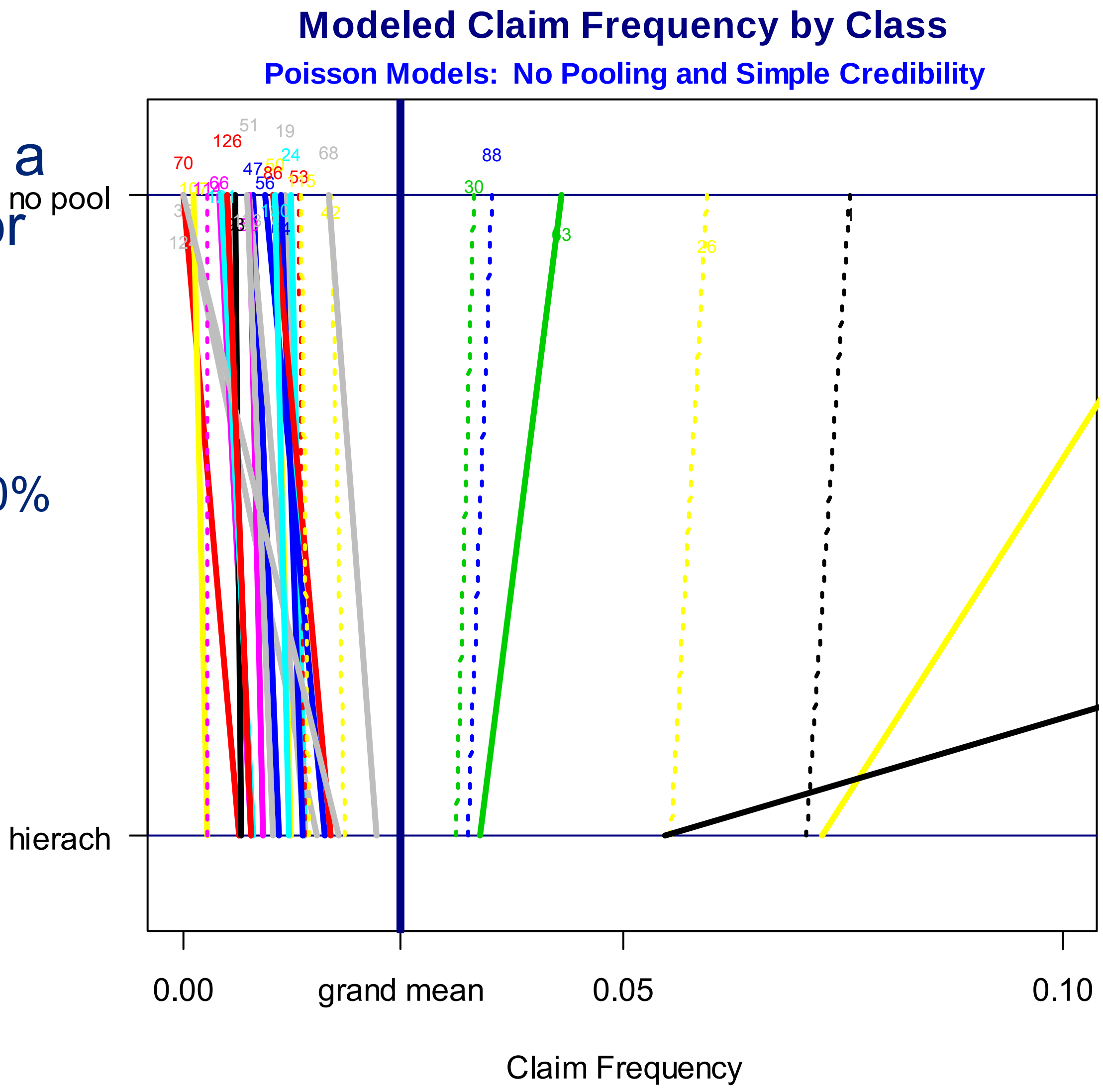
Shrinkage Effect of Hierarchical Model

- Top row: estimated claim frequencies from un-pooled model.
    - Separately calculate #claims/payroll by class
- Bottom row: estimated claim frequencies from Poisson hierarchical (credibility) model.
- Credibility estimates are “shrunk” towards the grand mean.



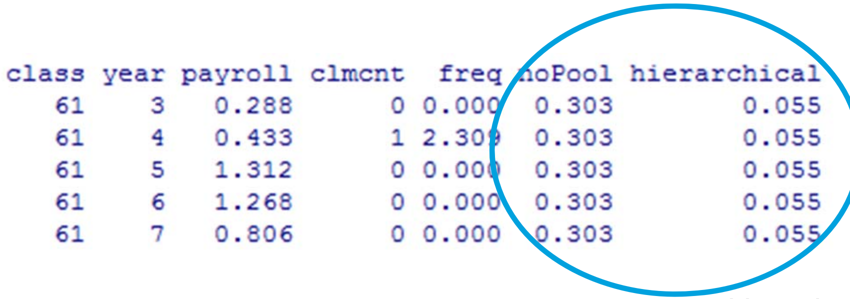
Shrinkage Effect of Hierarchical Model

- Let's plot the claim frequencies only for classes that experience a shrinkage effect is 5% or greater.
  - Dotted line: shrinkage between 5=10%.
  - Solid line: shrinkage > 10%

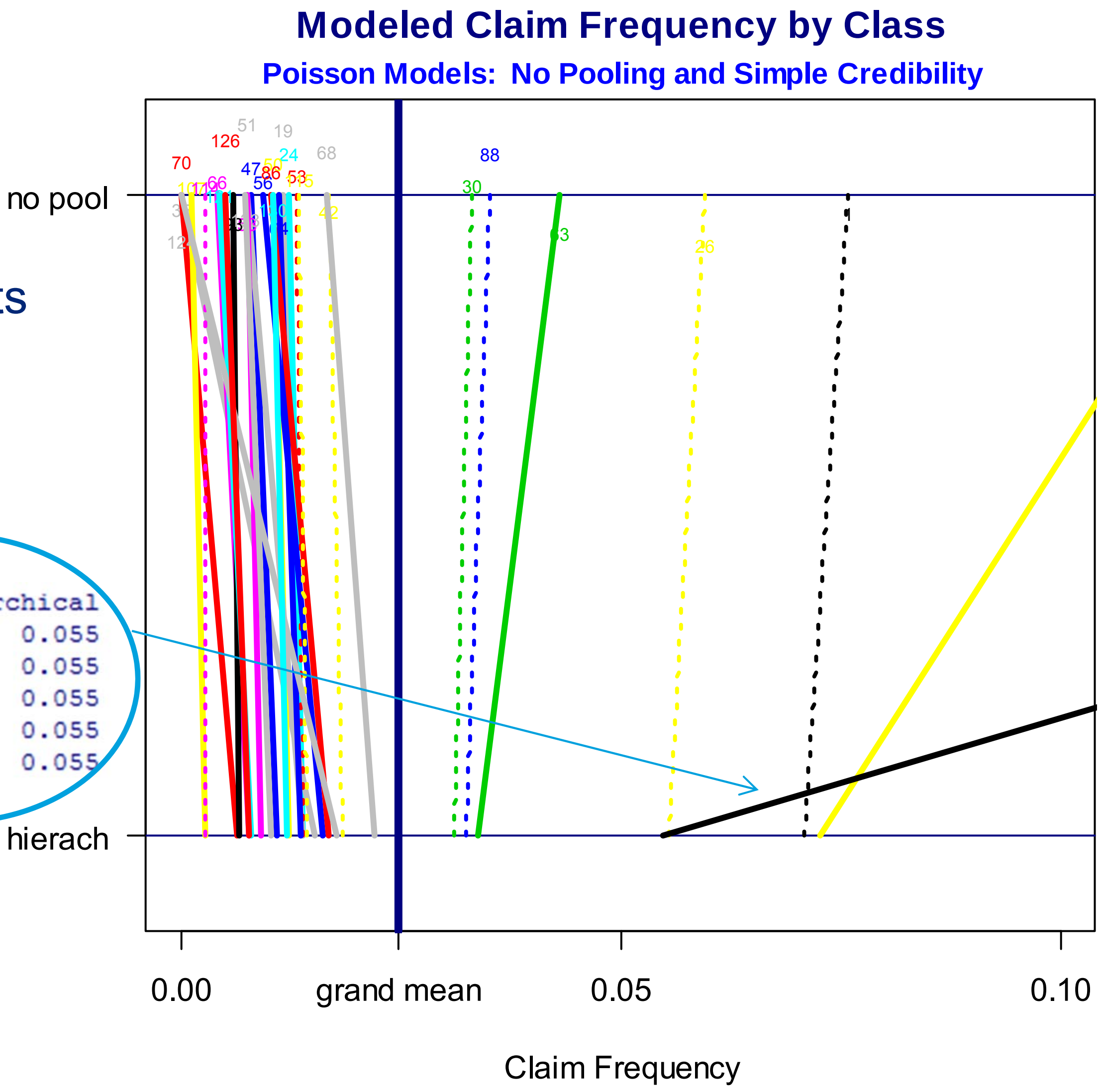


Shrinkage Effect of Hierarchical Model

- The most extreme shrinkage occurs for class 61.
  - Only 1 claim in years 3-6.
  - But very low payroll results in a large pre-shrunk estimated frequency.

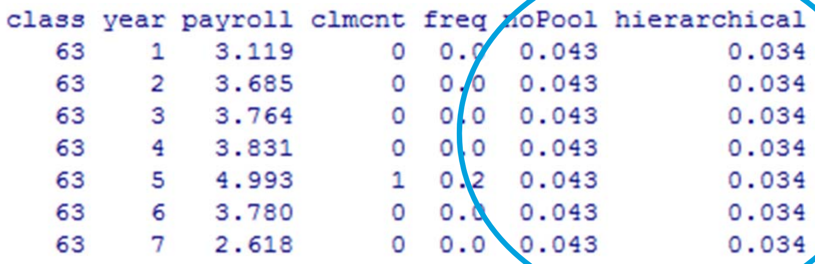


```
class year payroll clmcnt  freq noPool hierarchical
   61    3   0.288      0 0.000  0.303        0.055
   61    4   0.433      1 2.309  0.303        0.055
   61    5   1.312      0 0.000  0.303        0.055
   61    6   1.268      0 0.000  0.303        0.055
   61    7   0.806      0 0.000  0.303        0.055
```

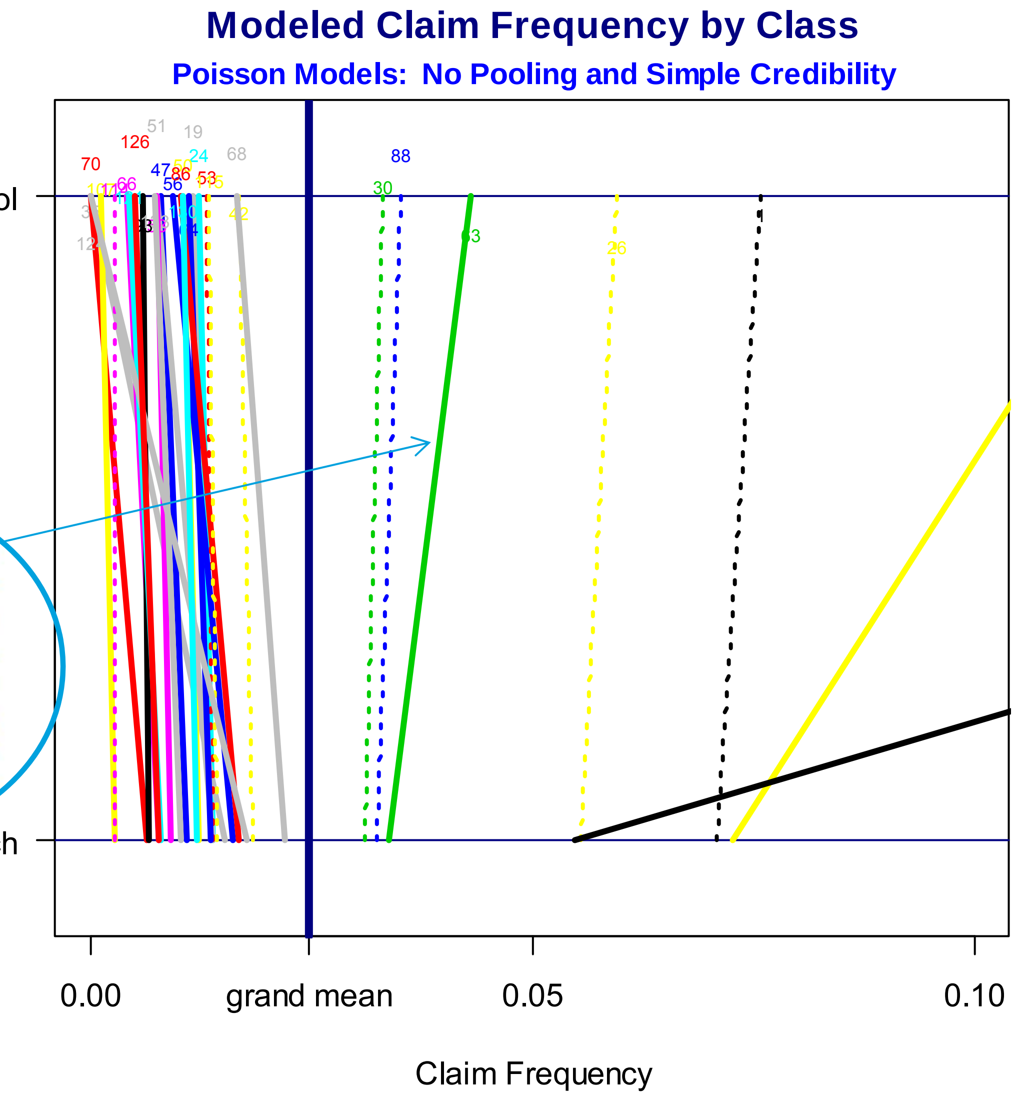


Shrinkage Effect of Hierarchical Model

- Shrinkage also occurs for class 63.
  - More payroll than class 61 but similar logic.

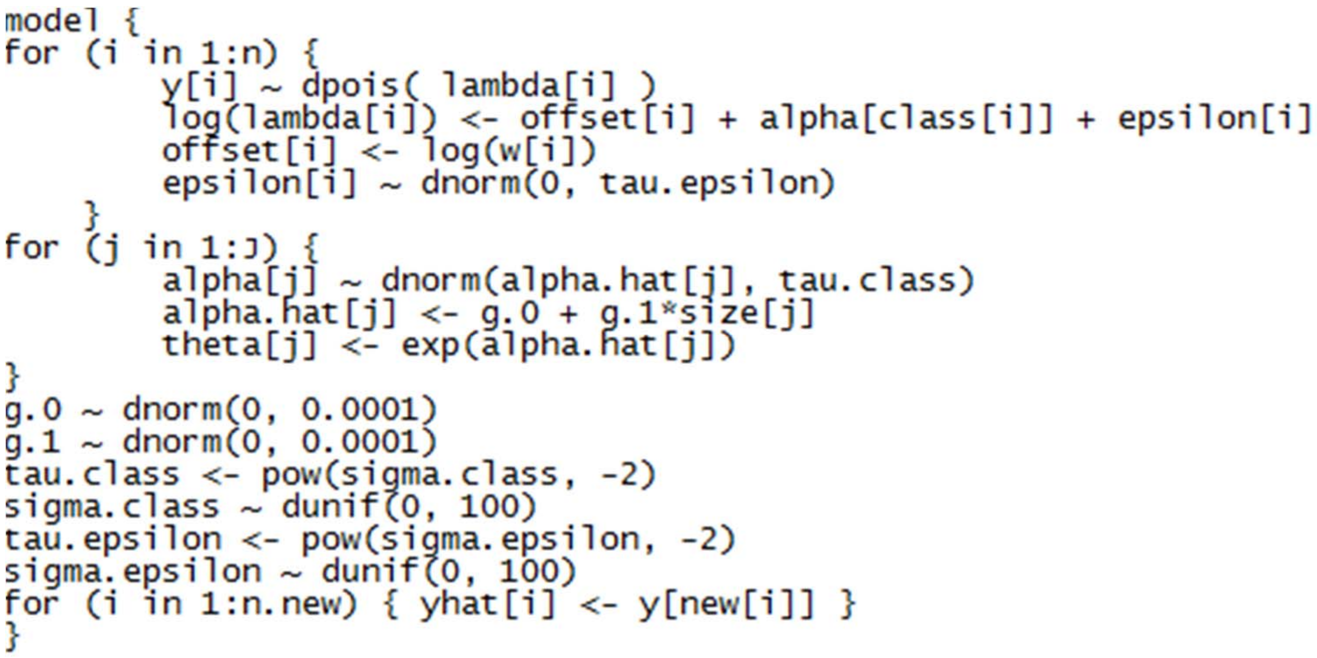


```
class year payroll clmcnt freq noPool hierarchical
   63    1   3.119      0  0.0  0.043        0.034
   63    2   3.685      0  0.0  0.043        0.034
   63    3   3.764      0  0.0  0.043        0.034
   63    4   3.831      0  0.0  0.043        0.034
   63    5   4.993      1  0.2  0.043        0.034
   63    6   3.780      0  0.0  0.043        0.034
   63    7   2.618      0  0.0  0.043        0.034
```



Now Specify a Fully Bayesian Model

- Here we specify a fully Bayesian model.
  - Still Poisson regression with an offset (*y*[*i*] is claim count)
  - Throw in a class-level covariate (relative “size” of the class).
  - Replace year-7 actual values with missing values so that we <u>model</u> the year-7 results and can compare actual with posterior credible interval.
  - Very flexible framework… could add in time trend as next step.



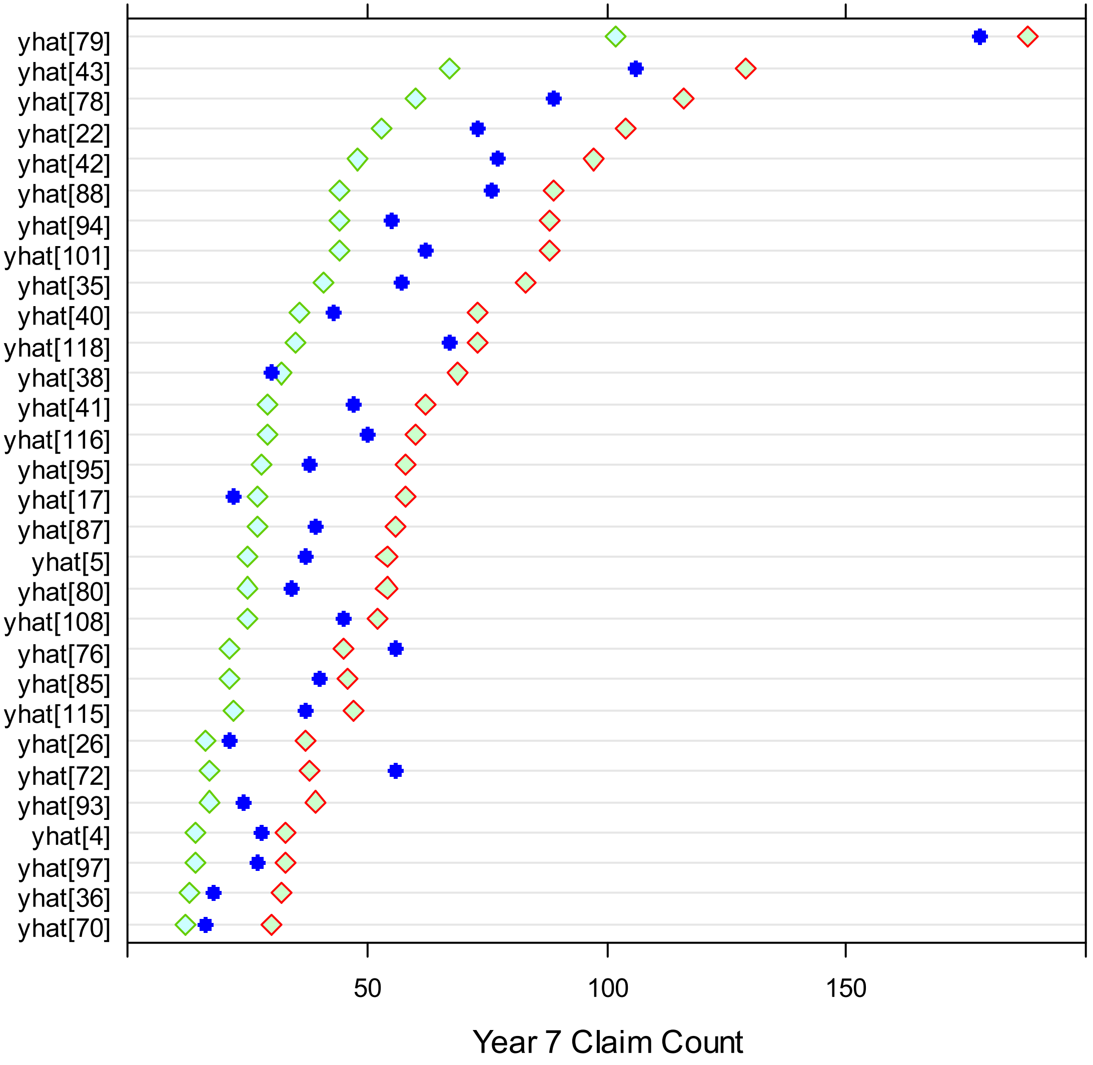
```
model {
for (i in 1:n) {
        y[i] ~ dpois( lambda[i] )
        log(lambda[i]) <- offset[i] + alpha[class[i]] + epsilon[i]
        offset[i] <- log(w[i])
        epsilon[i] ~ dnorm(0, tau.epsilon)
    }
for (j in 1:J) {
        alpha[j] ~ dnorm(alpha.hat[j], tau.class)
        alpha.hat[j] <- g.0 + g.1*size[j]
        theta[j] <- exp(alpha.hat[j])
}
g.0 ~ dnorm(0, 0.0001)
g.1 ~ dnorm(0, 0.0001)
tau.class <- pow(sigma.class, -2)
sigma.class ~ dunif(0, 100)
tau.epsilon <- pow(sigma.epsilon, -2)
sigma.epsilon ~ dunif(0, 100)
for (i in 1:n.new) { yhat[i] <- y[new[i]] }
}
```

A Credible Result

- Let’s rank the top 30 WC classes by the median of the posterior predictive density of year-7 claim count.

- **87%** of the top 30 classes have actual year-7 claim count that falls within the 90% posterior credible interval.

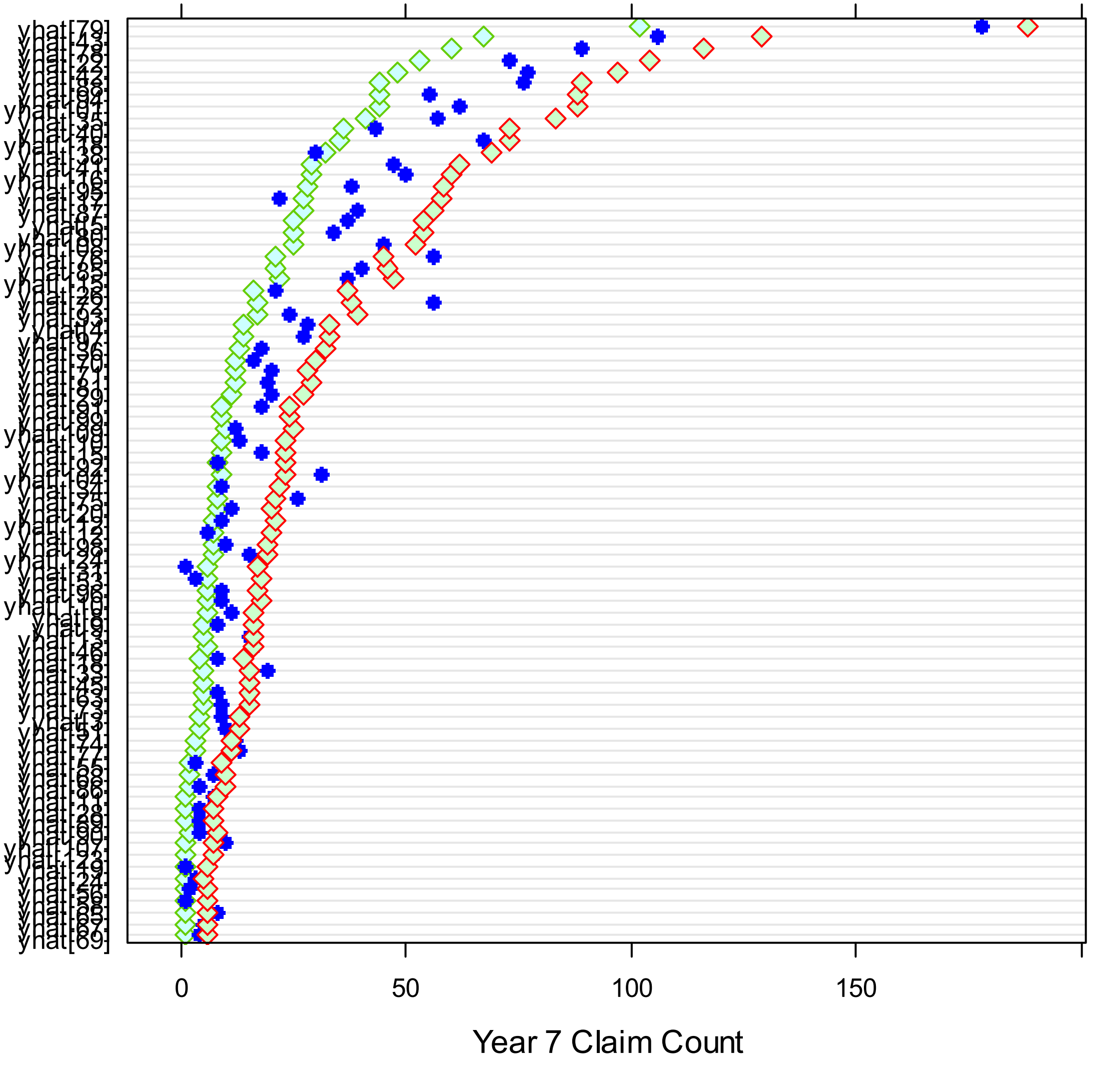
Top 30 WC Classes Ranked by Median Predicted Claim Count



A Credible Result

- If we increase this to the top-80, the corresponding number drops to **74%**.

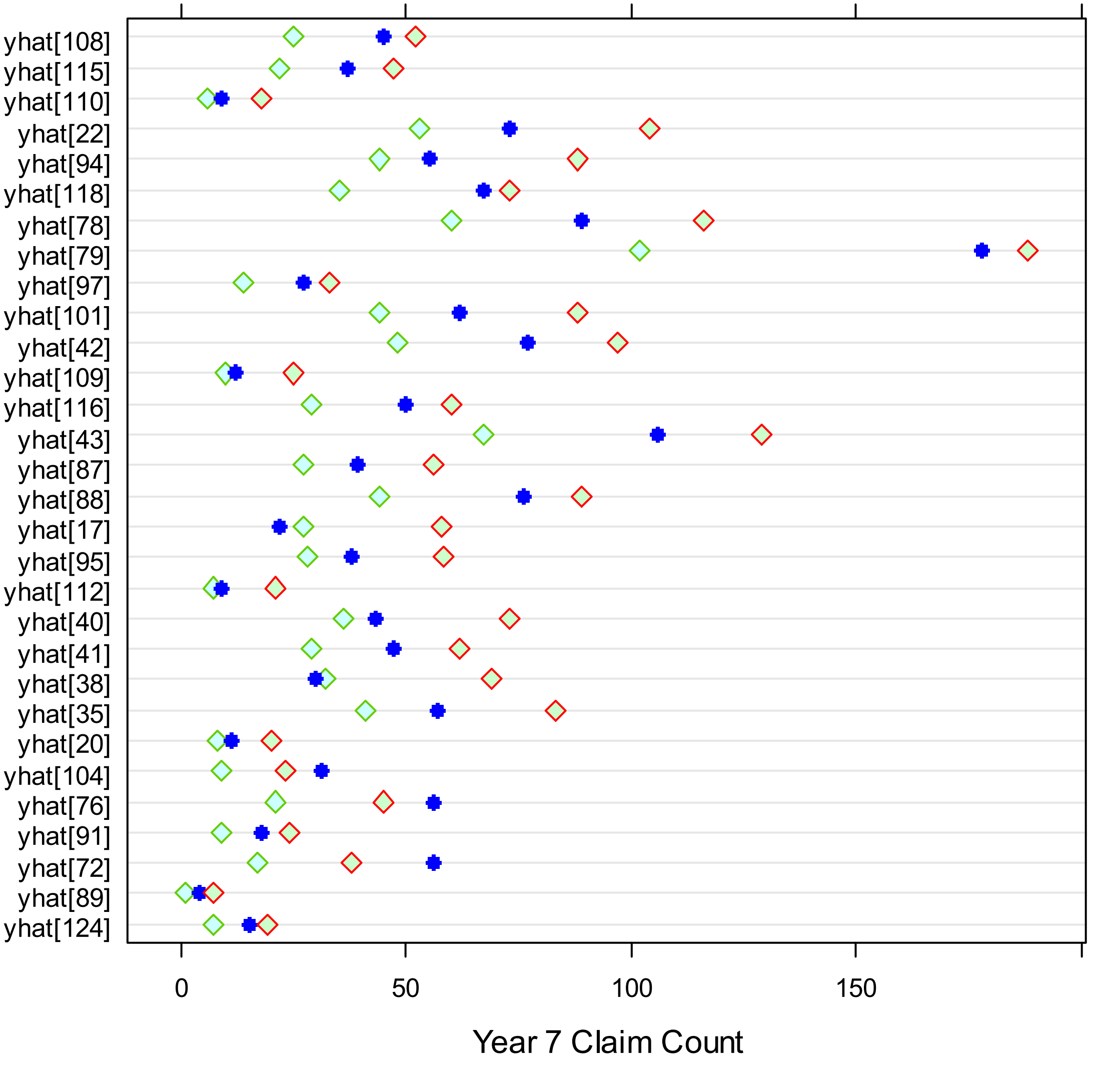
Top 80 WC Classes Ranked by Median Predicted Claim Count



A Credible Result

- Now we look at the top-30, ranked in descending order by payroll.
- **83%** of the top 30 classes have actual year-7 claim count that falls within the 90% posterior credible interval.

Top 30 WC Classes Ranked by Payroll


yhat[108]
yhat[115]
yhat[110]
yhat[22]
yhat[94]
yhat[118]
yhat[78]
yhat[79]
yhat[97]
yhat[101]
yhat[42]
yhat[109]
yhat[116]
yhat[43]
yhat[87]
yhat[88]
yhat[17]
yhat[95]
yhat[112]
yhat[40]
yhat[41]
yhat[38]
yhat[35]
yhat[20]
yhat[104]
yhat[76]
yhat[91]
yhat[72]
yhat[89]
yhat[124]
Year 7 Claim Count