

# A Discussion of the Bayesian Approach

Reference: Chapter 10 of *Theoretical Statistics*,  
Cox and Hinkley, 1974

and Sujit Ghosh's lecture notes

David Madigan

# Statistics

The subject of statistics concerns itself with using data to make inferences and predictions about the world

Researchers assembled the vast bulk of the statistical knowledge base prior to the availability of significant computing

Lots of assumptions and brilliant mathematics took the place of computing and led to useful and widely-used tools

Serious limits on the applicability of many of these methods: small data sets, unrealistically simple models,

Produce hard-to-interpret outputs like p-values and confidence intervals

# Bayesian Statistics

The Bayesian approach has deep historical roots but required the algorithmic developments of the late 1980's before it was of any use

The old sterile Bayesian-Frequentist debates are a thing of the past

Most data analysts take a pragmatic point of view and use whatever is most useful

Think about this...

	<i>Hospital</i>											
	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>J</b>	<b>K</b>	<b>L</b>
<b>No. of ops. <math>n</math></b>	47	148	119	810	211	196	148	215	207	97	256	360
<b>No. of deaths <math>r</math></b>	0	18	8	46	8	13	9	31	14	8	29	24

Denote  $\theta$  the probability that the next operation in hospital A results in a death

Use the data to estimate (i.e., guess the value of)  $\theta$

# Introduction

Classical approach treats  $\theta$  as fixed and draws on a repeated sampling principle

Bayesian approach regards  $\theta$  as the realized value of a random variable  $\Theta$ , with density  $f_{\Theta}(\theta)$  (“the prior”)

This makes life easier because it is clear that if we observe data  $X=x$ , then we need to compute the conditional density of  $\Theta$  given  $X=x$  (“the posterior”)

The Bayesian critique focuses on the “legitimacy and desirability” of introducing the rv  $\Theta$  and of specifying its prior distribution

# Bayesian Estimation

e.g. beta-binomial model:

$$\begin{aligned} p(\theta | D) &\propto p(D | \theta) p(\theta) \\ &= \theta^r (1 - \theta)^{n-r} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &= \theta^{r+\alpha-1} (1 - \theta)^{n-r+\beta-1} \end{aligned}$$

Predictive distribution:

$$\begin{aligned} p(x(n+1) | D) &= \int p(x(n+1), \theta | D) d\theta \\ &= \int p(x(n+1) | \theta) p(\theta | D) d\theta \end{aligned}$$

# Interpretations of Prior Distributions

1. As frequency distributions
2. As normative and objective representations of what is rational to believe about a parameter, sometimes in a state of ignorance
3. As a subjective measure of what a particular individual, “you,” actually believes

# Prior Frequency Distributions

- Sometimes the parameter value may be generated by a stable physical mechanism that may be known, or inferred from previous data
- e.g. a parameter that is a measure of a properties of a batch of material in an industrial inspection problem. Data on previous batches allow the estimation of a prior distribution
- Has a physical interpretation in terms of frequencies

# Normative/Objective Interpretation

- Central problem: specifying a prior distribution for a parameter about which nothing is known
- If  $\theta$  can only have a finite set of values, it seems natural to assume all values equally likely *a priori*
- This can have odd consequences. For example specifying a uniform prior on regression models:

$[], [1], [2], [3], [4], [12], [13], [14], [23], [24], [34], [123], [124], [134], [234], [1234]$

assigns prior probability 6/16 to 3-variable models and prior probability only 4/16 to 2-variable models

# Continuous Parameters

- Invariance arguments. e.g. for a normal mean  $\mu$ , argue that all intervals  $(a, a+b)$  should have the same prior probability for any given  $b$  and all  $a$ . This leads a uniform prior on the entire real line (“improper prior”)
- For a scale parameter,  $\sigma$ , may say all  $(a, ka)$  have the same prior probability, leading to a prior proportional to  $1/\sigma$ , again improper

# Continuous Parameters

- Natural to use a uniform prior (at least if the parameter space is of finite extent)
- However, if  $\Theta$  is uniform, an arbitrary non-linear function,  $g(\Theta)$ , is not
- Example:  $p(\theta)=1, \theta>0$ . Re-parametrize as:

$$\gamma = g(\theta) = \log(\theta)$$

then:  $p'(\gamma) = |J|p(e^\gamma)$  where  $J = d\theta/d\gamma$   
 $\theta = g^{-1}(\gamma) = e^\gamma \Rightarrow J = e^\gamma$

so that:  $p'(\gamma) = e^\gamma$

- “ignorance about  $\theta$ ” does not imply “ignorance about  $\gamma$ .”  
The notion of prior “ignorance” may be untenable?

# The Jeffreys Prior

(single parameter)

- Jeffreys prior is given by:  $p(\theta) \propto [I(\theta)]^{1/2}$

where

$$I(\theta) = -E_{X|\theta} \left[ \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right]$$

is the expected Fisher Information

- This is invariant to transformation in the sense that all parametrizations lead to the same prior
- Can also argue that it is uniform for a parametrization where the likelihood is completely determined except for its location (see Box and Tiao, 1973, Section 1.3)

# Jeffreys for Binomial

$$\log p(x|\theta) \propto x \log(\theta) + (n - x) \log(1 - \theta)$$

so that

$$\frac{\partial}{\partial \theta} \log p(x|\theta) = \frac{x}{\theta} - \frac{n - x}{1 - \theta}$$

and

$$\frac{\partial^2}{\partial \theta^2} \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2}.$$

$$I(\theta)^{1/2} \propto \theta^{-1/2} (1 - \theta)^{-1/2}$$

which is a beta density with parameters  $1/2$  and  $1/2$

# Other Jeffreys Priors

$$\text{Poisson}(\lambda): \pi(\lambda) \propto \lambda^{-1/2}$$

$$\text{Normal}(\mu): \pi(\mu) = 1, \mu \in \mathfrak{R}$$

$$\text{Normal}(\sigma): \pi(\sigma) = 1/\sigma, \sigma > 0$$

# Improper Priors $\Rightarrow$ Trouble (sometimes)

- Suppose  $Y_1, \dots, Y_n$  are independently normally distributed with constant variance  $\sigma^2$  and with:

$$E(Y_j) = \gamma + \beta \rho^{x_0 + ja}, j = 1, \dots, n.$$

- Suppose it is known that  $\rho$  is in  $[0,1]$ ,  $\rho$  is uniform on  $[0,1]$ , and  $\gamma$ ,  $\beta$ , and  $\sigma$  have improper priors
- Then for any observations  $\mathbf{y}$ , the marginal posterior density of  $\rho$  is proportional to

$$(\rho^{x_0} (1 - \rho^a) h(\rho, \mathbf{y}))^{-1}$$

where  $h$  is bounded and has no zeroes in  $[0,1]$ . This posterior is an improper distribution on  $[0,1]$ !

# Improper prior usually $\Rightarrow$ proper posterior

Suppose given  $\theta$ ,  $x_1, \dots, x_n$  are i.i.d.  $N(\theta, 1)$ . Here,  $\Theta = \{\theta : -\infty < \theta < \infty\}$ .

Suppose  $\pi(\theta) \propto 1$ . Then the posterior distribution of  $\theta$  is given by

$$\begin{aligned} p(\theta | x) &\propto p(x | \theta) \pi(\theta) \\ &\propto \exp \left\{ -\frac{1}{2} \sum (x_i - \theta)^2 \right\} \times 1 \\ &\propto \exp \left\{ -\frac{n}{2} [\theta - \bar{x}]^2 \right\} \Rightarrow \theta | x \sim N(\bar{x}, \frac{1}{n}) \end{aligned}$$

# Another Example

Given  $\theta$ , suppose  $x_1, \dots, x_n$  are i.i.d.  $\text{Poisson}(\theta)$ .

Suppose  $\pi(\theta) \propto \theta^{-\frac{1}{2}}$ . Here  $\Theta = \{\theta : 0 < \theta < \infty\}$ , and so

$$\int_0^{\infty} \pi(\theta) d\theta = \int_0^{\infty} \theta^{-\frac{1}{2}} d\theta = \infty.$$

Thus  $\pi(\theta)$  is improper for  $\theta$ .

$$\begin{aligned} p(\theta | x) &\propto \left( \theta^{\sum x_i} e^{-n\theta} \right) \left( \theta^{-\frac{1}{2}} \right) \\ &= \theta^{\sum x_i - \frac{1}{2}} e^{-n\theta} \\ &= \theta^{\sum x_i + \frac{1}{2} - 1} e^{-n\theta} \end{aligned}$$

Thus  $\theta | x \sim \text{gamma} \left( \frac{1}{2} + \sum x_i, n \right)$ .

# Subjective Degrees of Belief

- Probability represents a subjective degree of belief held by a particular person at a particular time
- Various techniques for eliciting subjective priors. For example, Good's device of imaginary results.
- e.g. binomial experiment. beta prior with  $a=b$ .  
“Imagine” the experiment yields 1 tail and  $n-1$  heads. How large should  $n$  be in order that we would just give odds of 2 to 1 in favor of a head occurring next? (eg  $n=4$  implies  $a=b=1$ )

$$\frac{n-1+a}{n+2a} = \frac{2}{3}$$

# Problems with Subjectivity

- What if the prior and the likelihood disagree substantially?
- The subjective prior cannot be “wrong” but may be based on a misconception
- The model may be substantially wrong
- Often use hierarchical models in practice:

$$p(\beta_i | \tau_i) = N(\mathbf{0}, \tau_i)$$

$$p(\tau_i) \propto 1/\tau_i \qquad p(\tau_i | \gamma) = \frac{\gamma}{2} \exp\left(-\frac{\gamma}{2}\tau_i\right)$$

# General Comments

- Determination of subjective priors is difficult
- Difficult to assess the usefulness of a subjective posterior
- Don't be misled by the term “subjective”; all data analyses involve appreciable personal elements

# EVVE

$$E(u) = E_v(E_{u|v}(u|v))$$

$$E(u) = \int \int up(u, v)dudv = \int \int up(u|v)dup(v)dv = \int E(u|v)p(v)dv$$

$$\text{var}(u) = E(\text{var}(u|v)) + \text{var}(E(u|v))$$

$$\begin{aligned} & \text{since } E[\text{var}(u|v)] + \text{var}[E(u|v)] \\ &= E[E(u^2|v) - (E(u|v))^2] + E[(E(u|v))^2] - (E[E(u|v)])^2 \\ &= E(u^2) - E[(E(u|v))^2] + E[(E(u|v))^2] - (E(u))^2 \\ &= E(u^2) - (E(u))^2 = \text{var}(u) \end{aligned}$$

# Bayesian Compromise between Data and Prior

$$\text{var}(\theta) = E(\text{var}(\theta|y)) + \text{var}(E(\theta|y))$$

- Posterior variance is on average smaller than the prior variance
- Reduction is the variance of posterior means over the distribution of possible data

# Posterior Summaries

- Mean, median, mode, etc.
- Central 95% interval versus highest posterior density region (normal mixture example...)

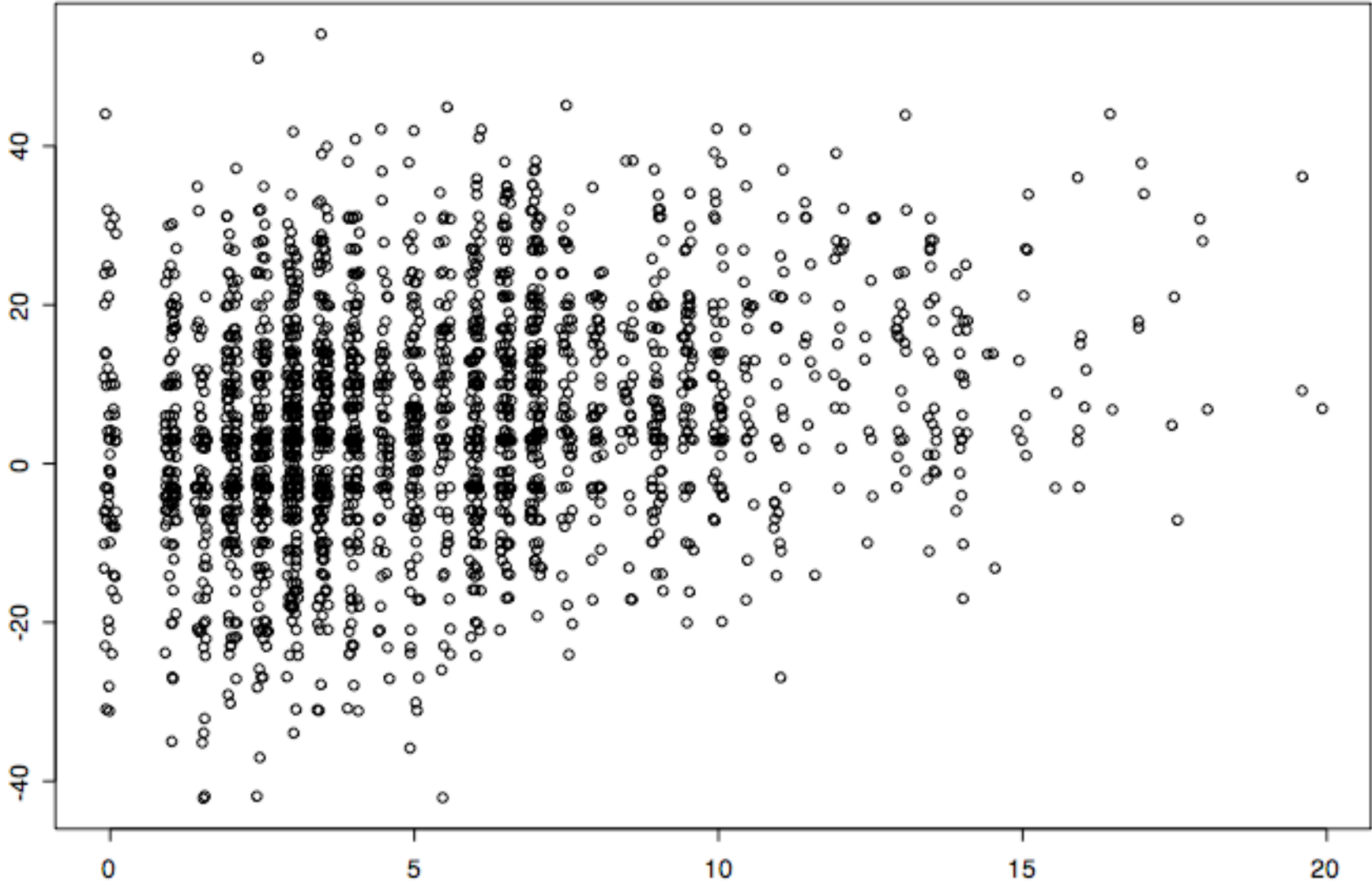
# Conjugate priors

<u>Family</u>	<u>Conjugate Prior</u>
Binomial( $N, \theta$ )	$\theta \sim \text{beta}(\alpha, \lambda)$
Poisson( $\theta$ )	$\theta \sim \text{gamma}(\delta_0, \gamma_0)$
$N(\mu, \sigma^2)$ , $\sigma^2$ known	$\mu \sim N(\mu_0, \sigma_0^2)$
$N(\mu, \sigma^2)$ , $\mu$ known, $\tau = 1/\sigma^2$	$\tau \sim \text{gamma}(\delta_0, \gamma_0)$
$\text{gamma}(\alpha, \lambda)$ , $\alpha$ known	$\lambda \sim \text{gamma}(\delta_0, \gamma_0)$
Beta( $\alpha, \lambda$ ), $\lambda$ known	$\alpha \sim \text{gamma}(\delta_0, \gamma_0)$

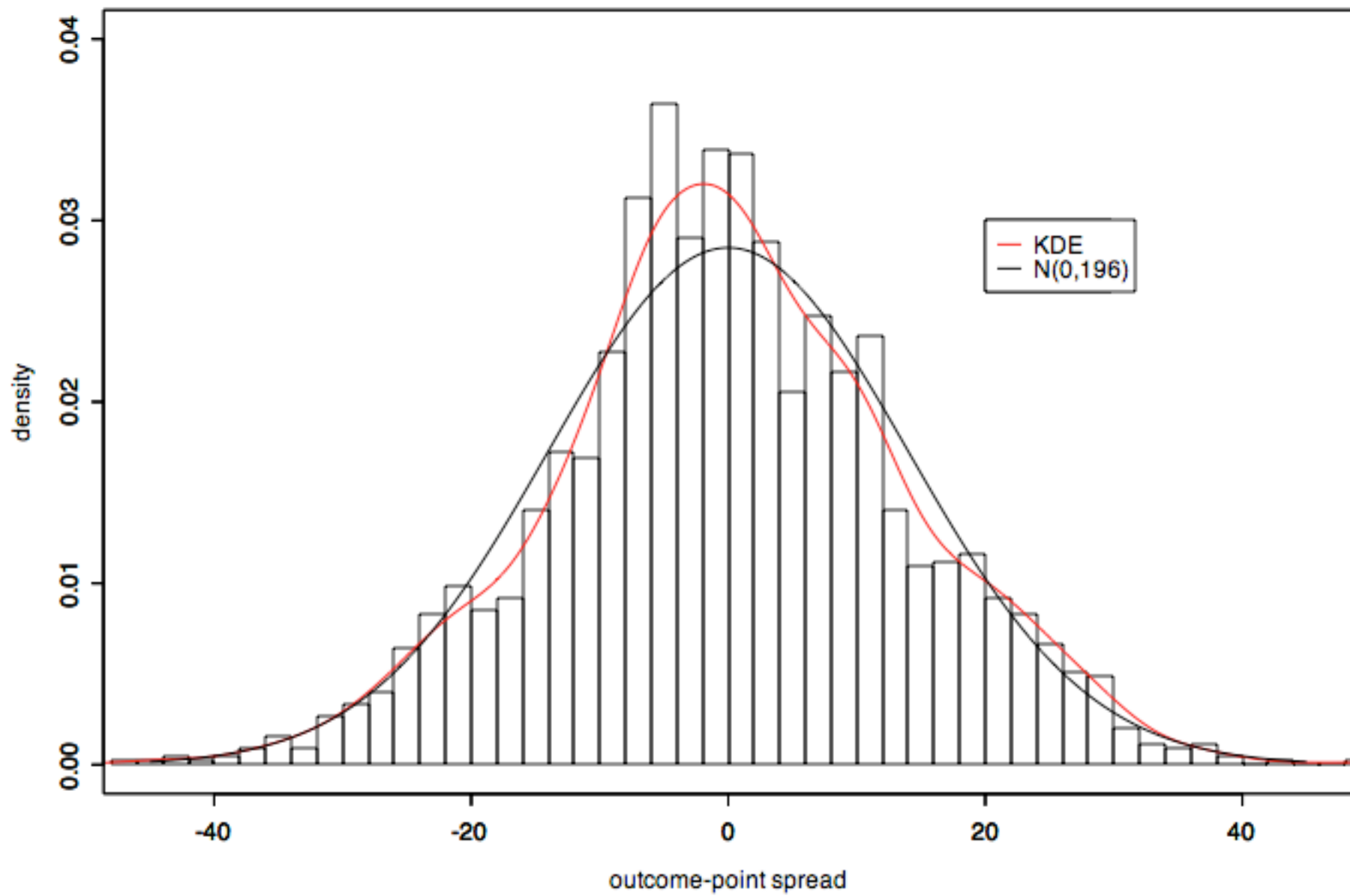
# Example: Football Scores

- “point spread”
- Team A might be favored to beat Team B by 3.5 points
- “The prior probability that A wins by 4 points or more is 50%”
- Treat point spreads as given – in fact there should be an uncertainty measure associated with the point spread

favorite - underdog + runif(favorite, -0.2, 0.2)



spread + runif(spread, -0.1, 0.1)



# Example: Football Scores

- outcome-spread seems roughly normal, e.g.,  
 $N(0,14^2)$
- $\Pr(\text{favorite wins} \mid \text{spread} = 3.5)$   
 $= \Pr(\text{outcome-spread} > -3.5)$   
 $= 1 - \Phi(-3.5/14) = 0.598$
- $\Pr(\text{favorite wins} \mid \text{spread} = 9.0) = 0.74$

# Example: Football Scores, cont

- Model:  $(X=)$ outcome-spread  $\sim N(0, \sigma^2)$
- Prior for  $\sigma^2$  ?
- The inverse-gamma is conjugate...

For  $N(y, \theta, \sigma^2)$  with  $\theta$  known and  $\sigma^2$  unknown, the likelihood for a vector  $y$  of  $n$  iid observations is:

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2\sigma^2} v\right) \end{aligned}$$

where  $v$  is:

$$v = \frac{1}{n} \sum_{i=1}^n (y_i - \theta)^2.$$

The corresponding conjugate prior density is the inverse-gamma:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

with hyperparameters  $\alpha$  and  $\beta$ .

A more convenient parameterization is the “scaled inverse- $\chi^2$  distribution” denoted:  
 $\sigma^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)$ :

$$p(\sigma^2) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \sigma_0^\nu (\sigma^2)^{-(\nu/2+1)} e^{-\nu\sigma_0^2/(2\sigma^2)}.$$

This is the same as the inverse gamma with  $\alpha = \frac{\nu}{2}$  and  $\beta = \frac{\nu}{2}\sigma_0^2$ .

The posterior density for  $\sigma^2$  is now:

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{\nu}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-((n+\nu_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + n\nu)\right). \end{aligned}$$

Thus:

$$\sigma^2|y \sim \text{Inv} - \chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right).$$

Thus the prior can be thought of as  $\nu_0$  observations with average squared deviation  $\sigma_0^2$ .

## Example: Football Scores, cont

- $n = 2240$  and  $\nu = 187.3$

Prior

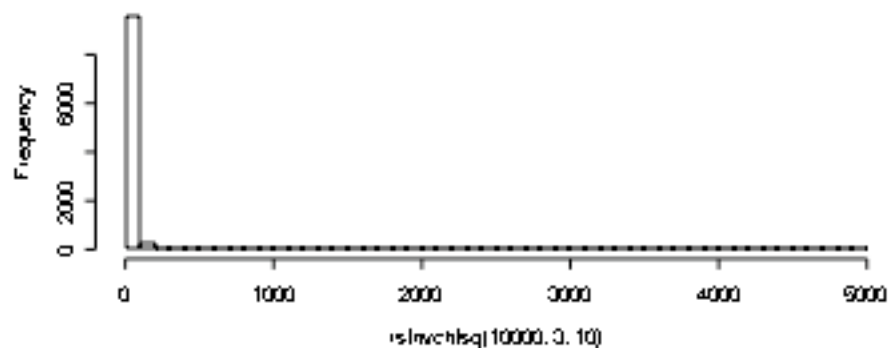
Posterior

$$\text{Inv-}\chi^2(3,10) \Rightarrow \text{Inv-}\chi^2(2243,187.1)$$

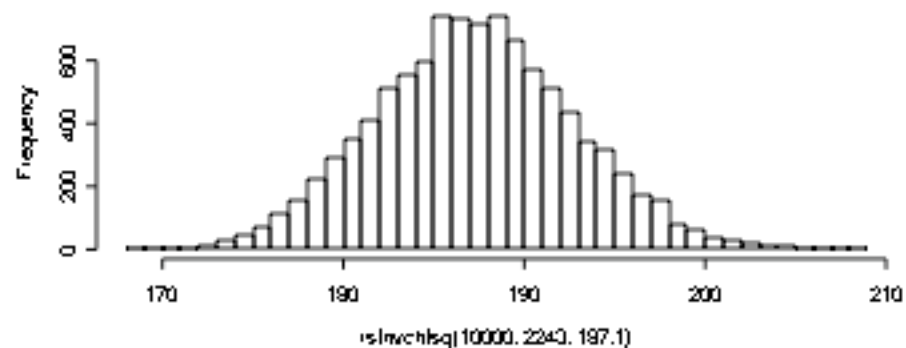
$$\text{Inv-}\chi^2(1,50) \Rightarrow \text{Inv-}\chi^2(2241,187.2)$$

$$\text{Inv-}\chi^2(100,180) \Rightarrow \text{Inv-}\chi^2(2340,187.0)$$

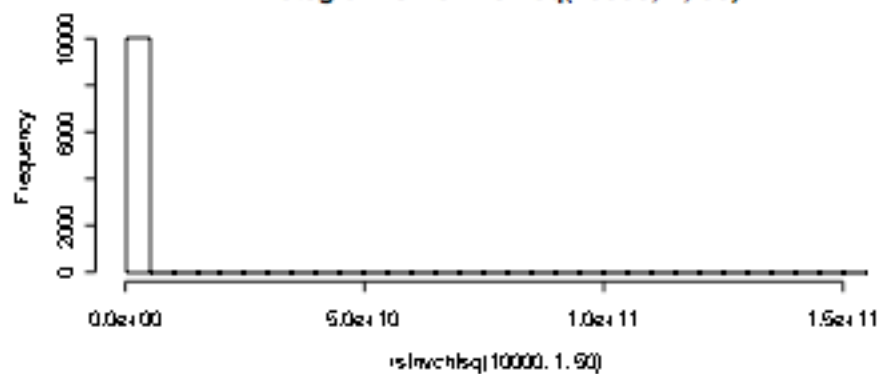
Histogram of  $\text{rsinvchisq}(10000, 3, 10)$



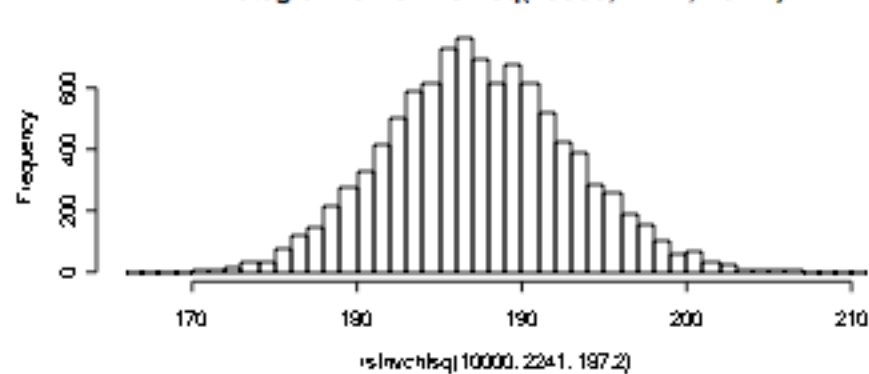
Histogram of  $\text{rsinvchisq}(10000, 2243, 187.1)$



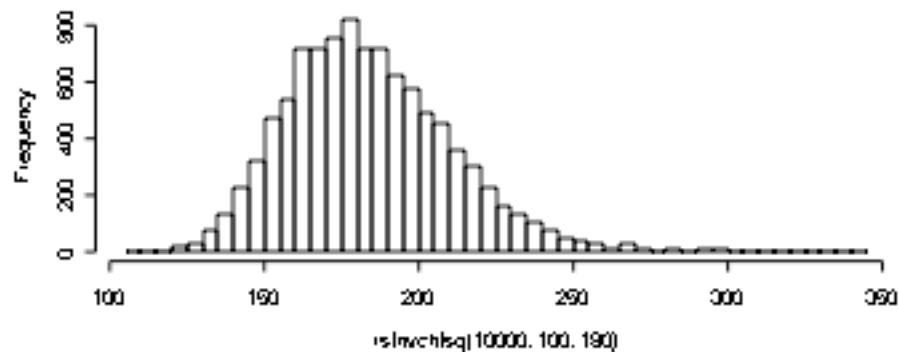
Histogram of  $\text{rsinvchisq}(10000, 1, 50)$



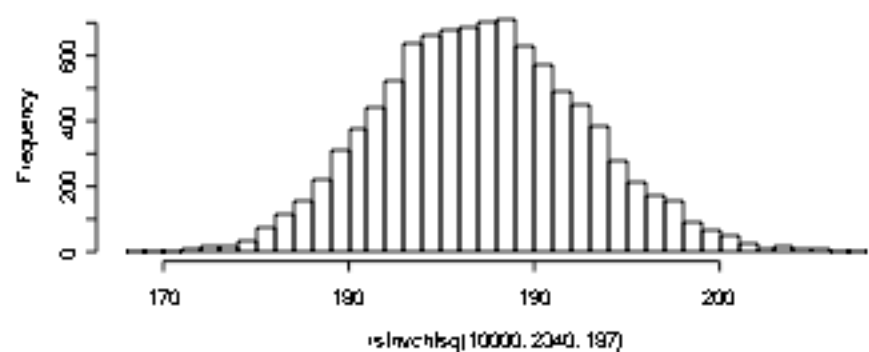
Histogram of  $\text{rsinvchisq}(10000, 2241, 187.2)$



Histogram of  $\text{rsinvchisq}(10000, 100, 180)$



Histogram of  $\text{rsinvchisq}(10000, 2340, 187)$

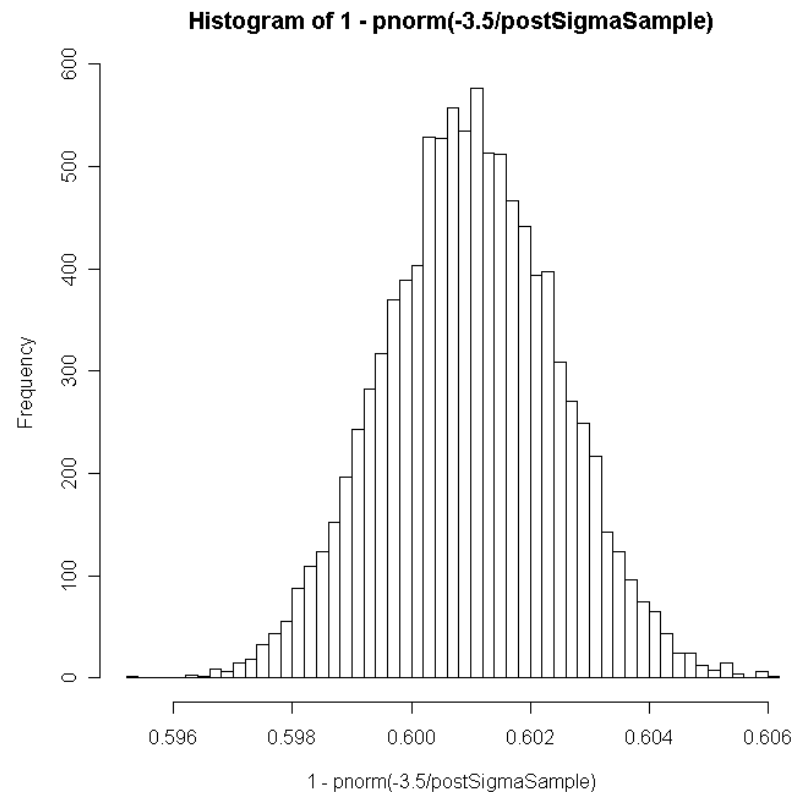


# Example: Football Scores

- $\Pr(\text{favorite wins} \mid \text{spread} = 3.5)$   
 $= \Pr(\text{outcome} - \text{spread} > -3.5)$   
 $= 1 - \Phi(-3.5/\sigma) = 0.598$

- Simulate from posterior:

```
postSigmaSample <-  
  sqrt(rsinvchisq(10000,2340,187.0))  
hist(1-pnorm(-  
  3.5/postSigmaSample),nclass=50)
```



## Example: Football Scores, cont

- $n = 10$  and  $\nu = 187.3$

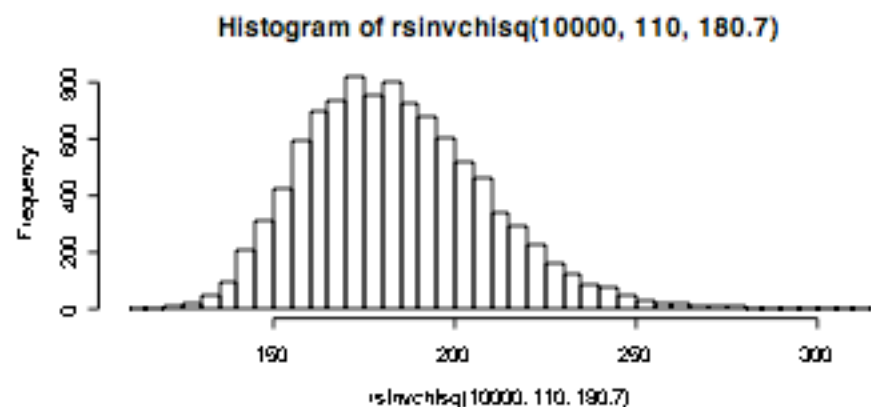
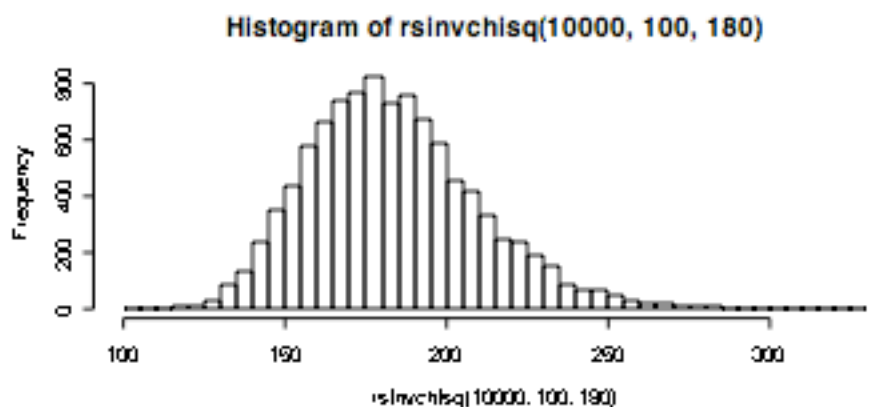
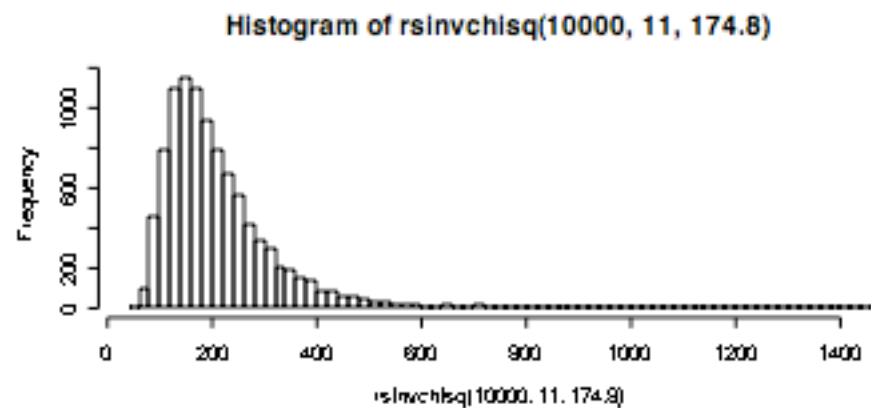
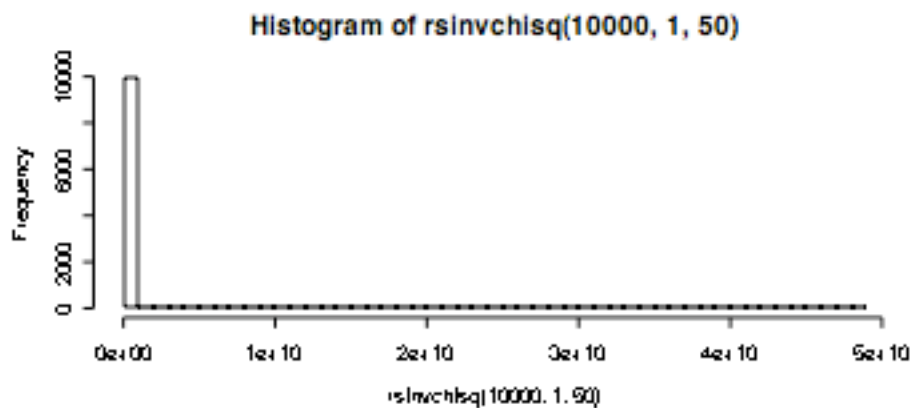
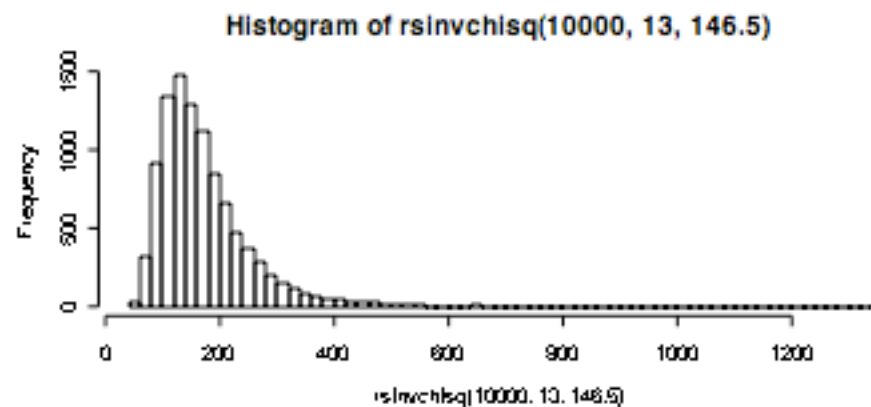
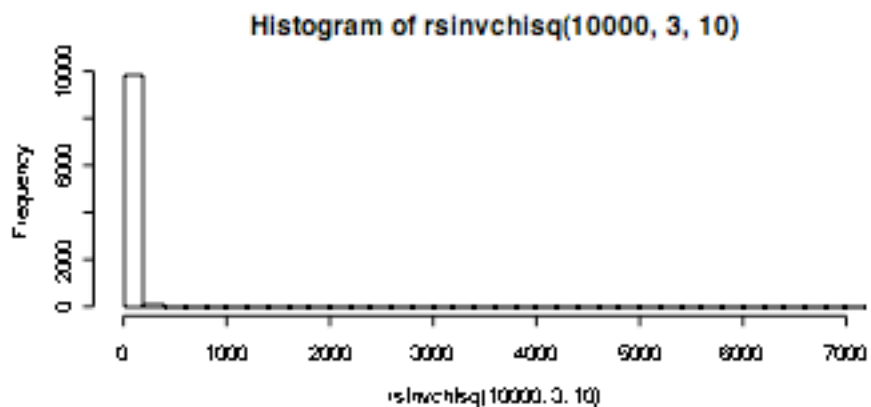
Prior

Posterior

$$\text{Inv-}\chi^2(3,10) \Rightarrow \text{Inv-}\chi^2(13,146.4)$$

$$\text{Inv-}\chi^2(1,50) \Rightarrow \text{Inv-}\chi^2(11,174.8)$$

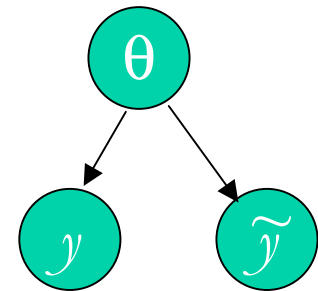
$$\text{Inv-}\chi^2(100,180) \Rightarrow \text{Inv-}\chi^2(110,180.7)$$



# Prediction

“Posterior Predictive Density” of a future observation

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta$$



binomial example,  $n=20$ ,  $x=12$ ,  $a=1$ ,  $b=1$

$$p(\tilde{y} = 1|\mathbf{y}) = \int \theta \frac{\Gamma(22)}{\Gamma(13)\Gamma(9)} \theta^{12}(1 - \theta)^8 d\theta = E[\theta|\mathbf{y}] = \frac{13}{22}$$

# Prediction for Univariate Normal

Consider a single observation  $y$  from  $N(\theta, \sigma^2)$ ,  $\sigma^2$  known. Let:

$$[\theta] \sim N(\mu_0, \tau_0^2)$$

Then:

$$[\theta|y] \sim N(\mu_1, \tau_1^2)$$

$$\mu_1 = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{1}{\sigma^2} y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right) \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right) d\theta \end{aligned}$$

# Prediction for Univariate Normal

- Posterior Predictive Distribution is Normal

$$E(\tilde{y}|y) = E_{\theta}(E_{\tilde{y}|\theta}(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_1$$

$$\begin{aligned}\text{var}(\tilde{y}|y) &= E(\text{var}(\tilde{y}|\theta, y)|y) + \text{var}(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + \text{var}(\theta|y) \\ &= \sigma^2 + \tau_1^2\end{aligned}$$

# Prediction for a Poisson

Suppose  $x_1, \dots, x_n$  are i.i.d.  $\text{Poisson}(\theta)$ . Suppose  $\theta \sim \text{gamma}(\alpha, \lambda)$ . Find  $p(z | x)$ .

$$p(\theta | x) \propto \theta^{\sum x_i} \exp\{-n\theta\} \theta^{\alpha-1} \exp\{-\lambda\theta\}$$

Thus  $\theta | x \sim \text{gamma}(\sum x_i + \alpha, \lambda + n)$

$$p(z | x)$$

$$= \int_0^{\infty} \frac{\theta^z \exp\{-\theta\}}{z!} \frac{(\lambda + n)^{\Sigma x_i + \alpha}}{\Gamma(\Sigma x_i + \alpha)} \theta^{\Sigma x_i + \alpha - 1} \exp\{-\theta(\lambda + n)\} d\theta$$

$$= \frac{(\lambda + n)^{\Sigma x_i + \alpha}}{z! \Gamma(\Sigma x_i + \alpha)} \int_0^{\infty} \theta^{z + \Sigma x_i + \alpha - 1} \exp\{-\theta(\lambda + n + 1)\} d\theta$$

$$= \frac{(\lambda + n)^{\Sigma x_i + \alpha}}{z! \Gamma(\Sigma x_i + \alpha)} \frac{\Gamma(z + \Sigma x_i + \alpha)}{(\lambda + n + 1)^{z + \Sigma x_i + \alpha}}$$

$$= \binom{z + \Sigma x_i + \alpha - 1}{z} \left( \frac{\lambda + n}{\lambda + n + 1} \right)^{\Sigma x_i + \alpha} \left( \frac{1}{\lambda + n + 1} \right)^z, \quad z = 0, 1, 2,$$

Thus

$$z|x \sim \text{neg-binomial} \left( \Sigma x_i + \alpha, \frac{1}{\lambda + n + 1} \right).$$