Applied Bayesian Methods

Michael Escobar

University of Toronto

m.escobar@utoronto.ca

http://individual.utoronto.ca/Escobar/

Course Philosophy

Applied Bayesian methods is a rather large area of knowledge. Here are some presentation principles that I am going to follow:

- **This course will not involve complex math.** I recognize that most people here are probably comfortable if I did do a lot of heavy math. However, please note a) that this presents a barrier for the use of these methods to others (who probably do maybe 99% of all the real data analysis in the world), b)that this will provide a teaching resource for others who need to transfer/translate this information to research colleagues or to students, and c)Brad Efron's story of why he created the bootstrap.

Course Philosophy

(continued)

- **I am not going to be arguing why Bayesian methods are better or worse than Frequentist methods.** I assume that people are here because they would like to learn more about Bayesian methods. However, in different places I will point out the differences between the two approaches.

Course Philosophy

(continued)

- **It is not my purpose to provide a large catalog of examples showing many different models.** I will concentrate on explaining how to use the new Bayesian methods. There exist several places where one can find many different examples. I hope that after this course, you will be well on your way to do these examples once you understand the basic mechanisms.

Outline of the Course

Part 1: Basic Bayesian inference. Quantifying uncertainty as probability. Some simple Bayesian models.

Part 2: The basics of Markov chain Monte Carlo (MCMC) methods. Why does it work? What can go wrong? Simple examples with the basic normal models.

Part 3: Using MCMC method on normal models. Introduction to a Bayesian software package. Model diagnostics and non-normal models.

Part 4: Other topics...

Part 1

A Basic Probability Model

1. Quantifying uncertainty with probability
2. Review of the rules of probability
3. Bayes theorem and how to update beliefs

B Binomial models.

C Normal models.

D Modelling prior belief

Measuring Uncertainty

- In this course, statistics is the quantification of uncertainty.
- There are some things which we are more uncertain about then others.
- Also, as we learn from observations/experiments, our uncertainty changes. (It usually gets smaller.)

Measuring Uncertainty)

(continue)

- What is the probability of drawing a ball of a particular color from an Urn?

Measuring Uncertainty)

(continue)

- Consider the height of the next person to walk through a door. We might have a lot of uncertainty about this.
  - However, we could make a guess as to the "typical" height. Also, we could make a statement about the range of heights which seem most likely.
  - The belief may change with more information.

Measuring Uncertainty)

(continue)

- The first task is to describe the uncertainty in terms of a mathematical expression.
- Once the notion of uncertainty is expressed as a mathematical expression, then mathematical tools can be used to update one's uncertainty when new data is presented.
- Then, probability theory is used to mathematically capture the notion of uncertainty, and Bayes theorem is used to model the change in uncertainty to account for new information.

To illustrate this, a simple example is presented. This example will contain the basic elements which are in a typical Bayesian problem.

Example 1: Which Box type was Picked?

Consider a large container which contains several boxes which from the outside all look alike.

Part 1: There are 20 boxes in total. In 10 of the boxes, there are 1 red ball and 5 blue balls in each box. Call theses type I boxes. In the other 10 boxes there are 3 red balls and 3 blue balls in each box. Call these type III boxes. Pick one of the boxes out of the container.

Part 2: Pick a ball out of the box chosen above. It is noted that the ball is red and the ball is put back into the box.

Part 3: Which type of box do you think you picked, type I or III?

Comments on Example 1

1. Please note that this “basic” model is more complicated than the usual basic model in introductory probability.

2. There are 2 parts to this model:

   Part i: This is the *prior* information about the chance of getting a box of either type I or III.

   Part ii : This gives the information as to how *likely* it is to see a red ball from a box of either type I or III.

3. So, there is prior opinion, then an observation is made and the opinion is updated in light of this new information.

So, now, we will transfer this information into mathematics.

Probability: Converting sets to numbers

Mathematically, a probability function is a function which maps a set of events to a number. It follows these three axioms:

1. The probability of the set of all events is one. That is, $P(S) = 1$.

2. The probability of any set of events is greater than or equal to zero. That is, $\forall A \subseteq S,\ P(A) \geq 0$.

3. The probability of the union of disjoint sets is equal to the sum of the probability of the individual sets. That is, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$ where $\forall i \neq j, A_i \cap A_j = \phi$.

Assigning Probability Values

- Please note that the axioms of probability do not state what values to assign. In fact, there are many ways to assign probability values so that the axioms of probability are satisfied.
- In Example 1, there are three “sub-experiments”. In each of these sub-experiments, probabilities can be assigned based on *physical symmetries*. That is, for each of these devices, there exist physical symmetries, so one could equally assign probability values to each of the different possible outcomes.

Probabilities by Physical Symmetries

One can assign probabilities using physical symmetries. The belief is that the different events are all physically symmetric, so each event would have equal probability.

Red ball from box type I: $P(\text{Red}) = 1/6$.

Red ball from box type III: $P(\text{Red}) = 3/6 = 1/2$.

Picking box type out of container:

$P(\text{Box type I}) = 10/20 = P(\text{Box type III})$.

Conditional Probability: Definition

- In order to link the probabilities together and to relate them, we need conditional probability.
- Conditional probability is the probability of a set of events under the knowledge that another event has occurred.
- So, we can define the conditional probability of A given the knowledge that an event in the set B has occurred by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note that the function $P(\cdot|B)$ defines a probability function

Conditional Probability: Bayes Theorem

Given the axioms of probability and the definition of conditional probability, we can relate the conditional probability of A given B to the probability of B given A. It is related by the following formula which is called Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

Also note, that the purpose of the denominator in the right hand side of the above equation is to insure that the total probability $P(S|B)$ equals one. With this in mind, it is sometimes written:

$$P(A|B) \propto P(B|A)P(A)$$

Reverse Probabilities

- The question in this example is which box type was drawn given information about the second stage.
- Before knowing anything about the second stage, the probability that the box of type III is $\frac{1}{2}$. This is what the value of $P(\text{III})$ equals.
- Now we want to know about the conditional probability of $\{\text{III}\}$ given information from the second stage. That is, we want to know what $P(\text{III}|\text{Red})$ is if the second stage was a success and we want to know what $P(\text{III}|\neg\text{Red})$ is otherwise.

Reverse Probabilities

Therefore, by Bayes theorem, we have:

$$\begin{aligned} P(\mathrm{III}|\mathrm{Red}) &= \frac{P(\mathrm{Red}|\mathrm{III})P(\mathrm{III})}{P(\mathrm{Red}|\mathrm{III})P(\mathrm{III}) + P(\mathrm{Red}|\neg\mathrm{III})P(\neg\mathrm{III})} \\ &= \frac{\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}\times\frac{1}{2}+\frac{1}{6}\times\frac{1}{2}} = \frac{3}{4} \end{aligned}$$

$$\begin{aligned} P(\neg\mathrm{III}|\mathrm{Red}) &= \frac{P(\mathrm{Red}|\neg\mathrm{III})P(\neg\mathrm{III})}{P(\mathrm{Red}|\mathrm{III})P(\mathrm{III}) + P(\mathrm{Red}|\neg\mathrm{III})P(\neg\mathrm{III})} \\ &= \frac{\frac{1}{6}\times\frac{1}{2}}{\frac{1}{2}\times\frac{1}{2}+\frac{1}{6}\times\frac{1}{2}} = \frac{1}{4} \end{aligned}$$

Reverse Probabilities

(continued)

- So, when there is a red ball drawn at the second stage, then the probability that a box of type III was drawn goes from 50% to 75%.
- Using a similar calculation, when their is a blue ball drawn at the second stage, then the probability that box type III was draw goes from 50% to 37.5%.

A: Basic Probability Model - Summary

- At stage one, something happens which you don’t know, but you have some opinion on.
- At stage two, something happens which gives clues as to what happens at the first stage.
- Mathematical probability can be used to formalize these beliefs.
- Also, one can use the rules of probability to update ones belief for stage one given information at stage two.

B: The Binomial Model

Outline:

- The likelihood function: the binomial distribution.
- Convenient prior distribution and the resulting posterior distributions.
- Inference statements.
- Simple calculations.
- Other types of priors.

The Binomial Model: Introduction

- The simplest observation is to observe whether an event occurs or not. For example did the patient die or not. Is there an adverse effect to a drug? Does a subject have a particular disease or not.
- With these models, we are interested in understanding the frequency that something occurs. In the previous example, we looked at the frequency of drawing a red ball.
- Like in the previous model, we will be interested in making statements about the frequency based on observing the number of times something occurred.

The Binomial Model: Math Notation

- First, let X take on the value 1 if the event occurs and take on the value 0 if the event does not occur. This is called the Bernoulli distribution.

- It has one parameter, θ. The probability that $X = 1$ is equal to θ. The probability that X equals 0 is $1 - \theta$. The mean of X is θ. The variance of X is $\theta(1 - \theta)$.

- Consider a sequence of X_i's where $i = 1, \ldots, n$ and if given knowledge of θ, then these X_i observations are independent. That is, the random variables X_i's are conditionally independent given θ.

The Binomial Model: Math Notation

- Let Y be the sum of the X_i's. The random variable Y has a binomial distribution, and it has two parameters.
- The first parameter is from the individual Bernoulli distribution and it is the parameter θ.
- The second parameter is n which is the number of individual Bernoulli's which are summed together.
- The probability that the random variable Y has the value y is:

$$\Pr(Y = y|\theta, n) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y}.,$$

- The mean of Y is $n\theta$ and the variance is $n\theta(1-\theta)$.

Bayes Theorem

Bayes Theorem for this model is:

$$
\begin{aligned}
f(\theta|Y) &= \frac{f(Y|\theta)f(\theta)}{f(Y)} \\
&= \frac{f(Y|\theta)f(\theta)}{\int f(Y|\theta)f(\theta)d\theta} \quad \text{(if } \theta \text{ is continuous)} \\
&= \frac{f(Y|\theta)f(\theta)}{\sum_{j=1}^{m} f(Y|\theta_j)f(\theta_j)}
\end{aligned}
$$

(if θ is discrete with values at $\theta_1, \theta_2, \ldots, \theta_m$)

Bayes Theorem

(continue)

- The main purpose of the denominator in the above equation is to ensure that the total probability of $f(\cdot|Y)$ sums/integrates to one.
- This denominator is sometimes called the proportionality constant.
- This is because the total probability equals one. Therefore, the above definition is sometimes written as:

$$\begin{aligned} f(\theta|Y) &= C\, f(Y|\theta)f(\theta), \text{ or} \\ f(\theta|Y) &\propto f(Y|\theta)f(\theta), \end{aligned}$$

where C is some constant and $\propto$ means “proportional to”.

Bayes Theorem

(continue)

- The posterior distribution is equal to a constant times the product of the likelihood function and the prior distribution.
- The constant term in the above expression can be a function of Y but it cannot be a function of θ.
- Therefore, when calculating the posterior distribution, it is sometimes helpful to drop all terms in the prior distribution and the likelihood function which are not function of the parameter of interest.

Example 2: Anti-Cancer Drug Efficiency

- Suppose there is a new anti-cancer drug and there is an interest in knowing if this drug is effective. The interest is to determine the probability that the drug treatment will result in an improvement in the cancer.
- Suppose that one observes 5 out of 20 patients who showed an improvement.
- The statistical question is to determine our new belief as to what the efficacy is.

Approach to Example 2

- Using Bayes theorem, we first consider our prior opinion about the value of the efficacy.
- Then, we consider the data, and we see how likely the data is for different possible values of the efficacy.
- Bayes theorem is then used to combine the prior distribution and the likelihood function to obtain the posterior distribution which represents our new belief as to what the efficacy is.
- In considering the mechanics of this calculation, it is more convenient to first look at the structure of the likelihood and then to look at the structure of the prior and posterior distribution.

Likelihood

- In the example, the likelihood function is based on $f(y = 5|n = 20, \theta)$.
- The probability function is a function of Y for fixed values of θ, but the likelihood function is a function of θ for fixed values of the data Y.
- First, let's look at the value of $f(y|n = 20, \theta)$ over the different possible values for y for a fixed set of values of θ.

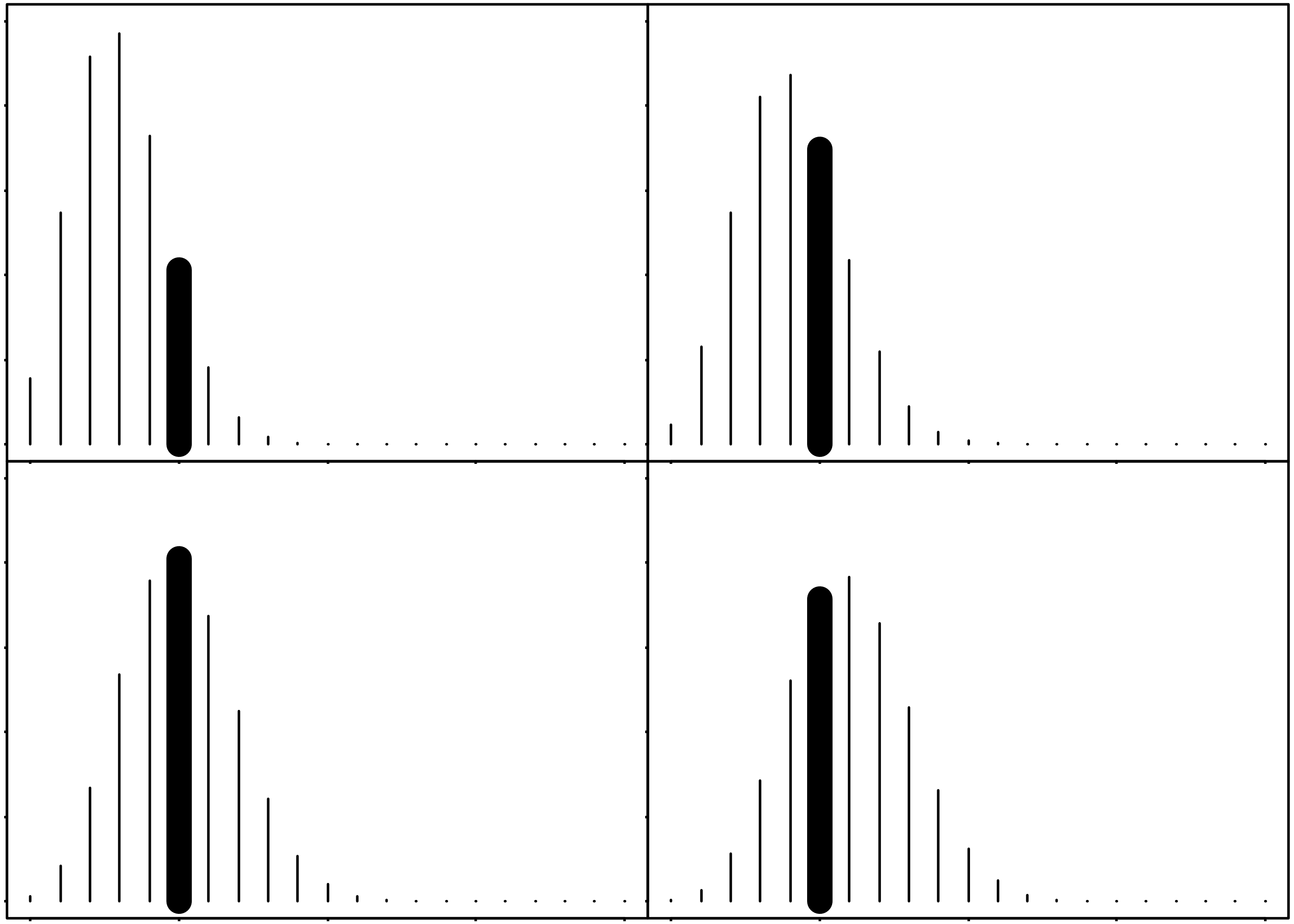


Figure 1: Probability distribution function for different values of θ. (a) $\theta = .15$, (b) $\theta = .20$, (c) $\theta = .25$, and (d) θ=.30

Likelihood

(continue)

- Figure 1 shows the probability distribution function of seeing different values of Y when θ takes on the values .15, .20, .25, and .30.
- Looking at these figures we see that seeing a data value of 5 is fairly common for these value of θ.

Now, we will look at $f(Y = 5|n = 20, \theta)$ as a function of θ where θ takes on values between 0 and 1.

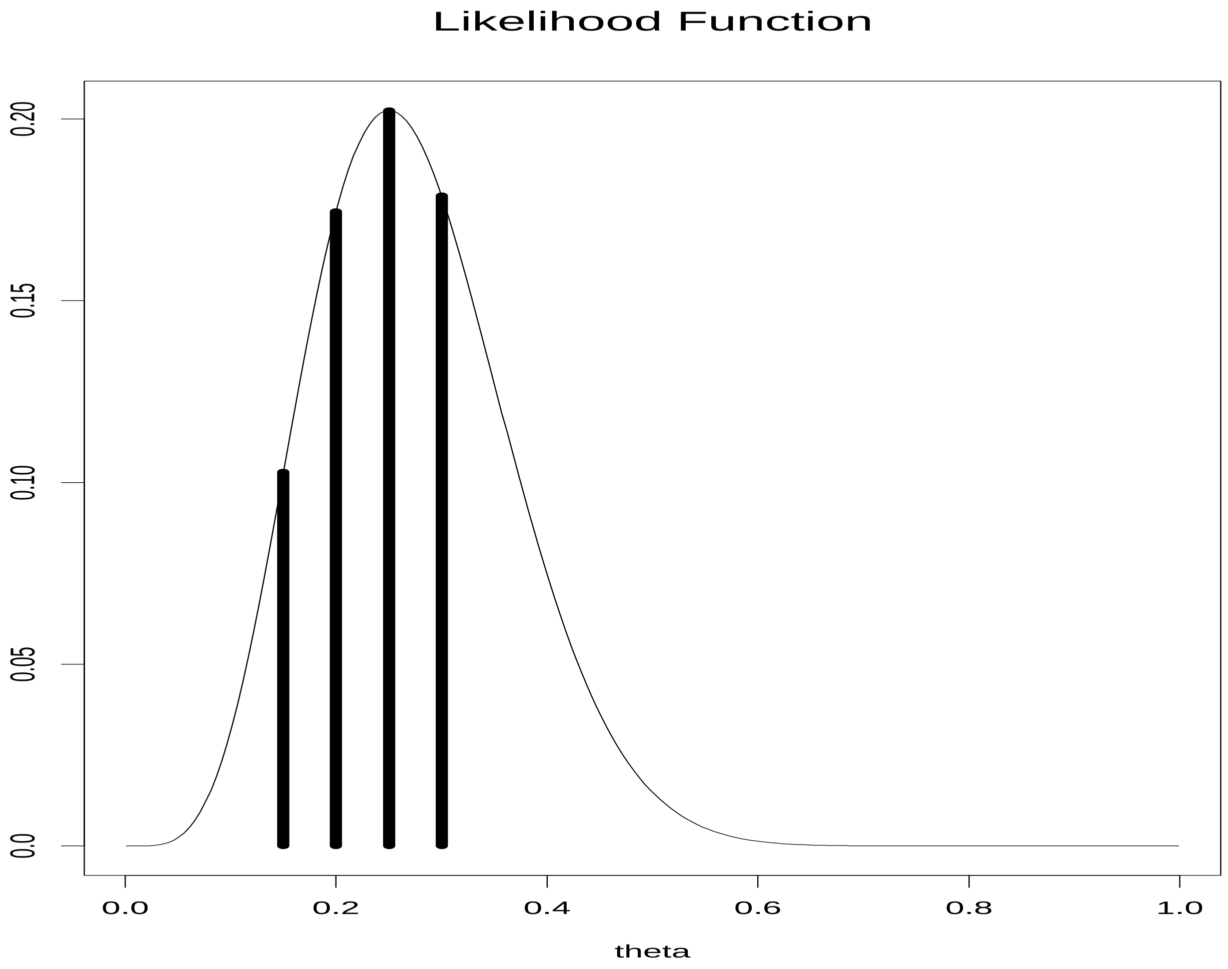


Figure 2: The likelihood function of binomial parameter θ when $Y = 5$ and $n = 20$. Solid lines are for the values of $\theta = .15, .20, .25, .30$.

Likelihood

(continue)

The probability function is a function of Y for fixed values of θ, but the likelihood function is a function of θ for fixed values of Y. Figure 2 shows the likelihood function of θ when $n = 20$ and $Y = 5$.

- From the graph, .25 is the highest point and it might be called the “most likely”. In fact, this value is the maximum likelihood estimator or the MLE.
- However, we also see that several other values are extremely likely also. Values of .20 or .30 for example also have high likelihood values.

Now we need to think about the prior so that it can be combined with the likelihood in order to calculate the posterior.

Calculating the Posterior Distribution

Here are some ways to calculate the posterior distribution:

1. Analytically: Some special priors allow for a closed form solution for the posterior distribution. This can be done with "conjugate priors". For conjugate priors, the mathematics work out nicely.

2. Estimate on a grid: One can simply multiply the prior and likelihood for each individual point on a grid of points. This will give the plot of the posterior distribution.

3. Sample from the posterior distribution: This can be done with newer algorithms such as Markov chain Monte Carlo methods.

Conjugate Priors

- Given a model with a certain likelihood function, a conjugate family of distributions for this likelihood is a family of distributions such that if the prior is member of the family then the posterior is a member also for all possible, sampled values
- In this case, the prior distribution is called the *conjugate prior*.

Conjugate Priors

- A conjugate prior is a convenient prior, since it allows one to easily calculate the posterior distribution.
- Hopefully the family is flexible enough to approximate one's belief. If this is not enough, then more flexible priors can be used by using a fine grid of discrete points or by using a more complex prior and using the methods which are discussed later on today.

Beta distribution

When the data is from a binomial distribution, the beta distribution is a convenient prior.

- If θ has a beta distribution with parameters $\alpha = a$ and $\beta = b$, the it has the following density function:

$$f(\theta|\alpha = a, \beta = b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\{0 \leq \theta \leq 1\}.$$

- The function $\Gamma(\cdot)$ is the gamma function, and it is not needed in the calculation of the posterior distribution.
- For notation, we call this distribution Beta(a,b).

Updating with the Beta Prior

- If the prior distribution is a beta distribution, then the posterior distribution is also a beta distribution.
- The posterior distribution is proportional to the prior distribution times the likelihood function.
- Therefore, we have the following:

$$\begin{aligned} f(\theta|\alpha = a, \beta = b, Y = y, n) &\propto f(Y = y|\theta, n) f(\theta|\alpha = a, \beta = b) \\ &\propto \theta^y (1-\theta)^{n-y} \quad \theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ &\propto f(\theta|\alpha = a + y, \beta = b + n - y) \end{aligned}$$

- So, if the prior is a beta distribution with parameter $\alpha = a$ and $\beta = b$ then the posterior distribution is a beta distribution with parameters $\alpha = a + y$ and $\beta = b + n - y$.

Updating with the Beta Prior

(continued)

Key idea

If we have the following:

- Data, y, is from a binomial distribution with parameters θ and n.
- The parameter θ has a prior distribution which is a Beta(a,b).
- Then the posterior distribution of θ is a Beta(a+y, b+$n-y$).

Comments about the Beta Prior

- The sample size of the experiment increases, then the values of y and $n - y$ will increase. Therefore, as the sample size increases, the influence of the data will dominate the calculation of the posterior.
- So, for large amounts of data, the parameters of the prior will have less and less influence on determining the posterior distribution.

Approximating with a Beta

- The beta distribution is the conjugate prior distribution for this family.
- Hopefully, it is a close to the the actual prior that you might have. If it is, it allows you to quickly calculate the posterior distribution.
- Since it can be expressed mathematically, it can be used to understand the model better.
- If your prior is not close to a beta distribution, then one needs more work to find the posterior distribution. Later in this course, such methods are discussed.

The Shape of the Beta

To use the beta distribution to model our prior belief, it is necessary to know how the parameters affect the shape.

- When both α and β are less than one, the density has a U-shape.
- When either α or β are not greater one then but not both, then the shape is a J-shape or a backward J-shape.
- When both α and β are greater than one, then the beta density has a unimodal shape.
- When both α and β are equal to one, then the beta density is flat and is the uniform distribution.

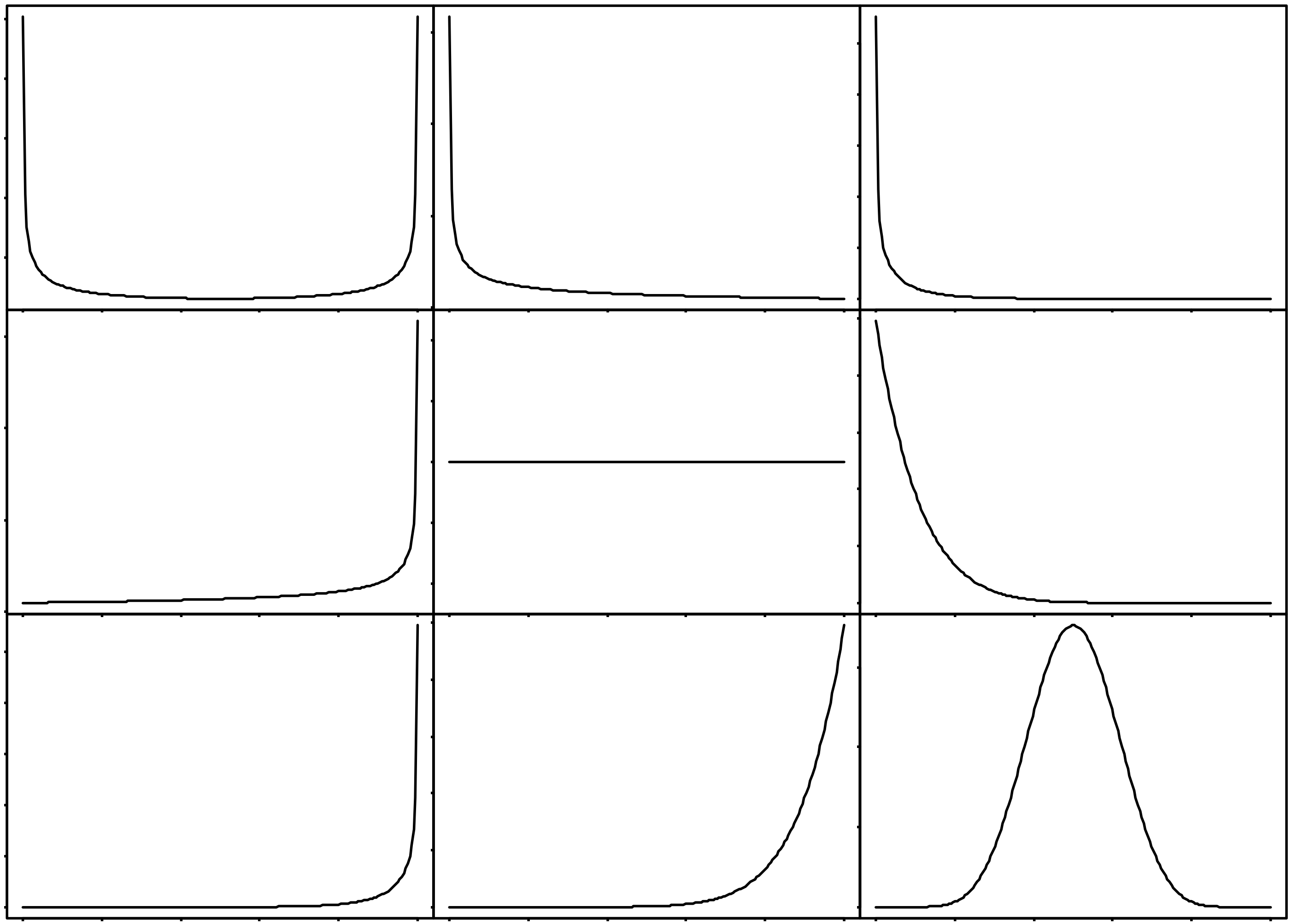


Figure 3: The beta distribution for different values of α and β. The rows (t-b) are for $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$. The columns (l-r) are for $\beta < 1$, $\beta = 1$, $\beta > 1$.

Beta Distribution

- The mean is $\alpha/(\alpha+\beta)$.
- The variance is $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.
- The precision increases as the sum $\alpha+\beta$ increases and the peak of the mode of the beta density gets sharper as the sum $\alpha+\beta$ increases.
- The statistical functions for the beta distribution (such as the "area under the curve") are contained in most statistical packages.

Flat Priors

- If there are no preferences on the value for θ, then one might consider using a "flat prior". That is, the uniform distribution between 0 and 1. {This is also a Beta(1,1).}
- In some sense, this implies that each value of θ as equally likely.
- Since the prior is flat, the posterior distribution has the same shape as the likelihood multiplied be a constant.

Flat Priors

(continued)

- Since the posterior is proportional to the product of the likelihood and the prior, then this prior will result in a posterior which has the same shape as the likelihood. That is,

$$f(\theta|Y) \propto f(Y|\theta)f(\theta) = f(Y|\theta) * C$$

- Figure 4 shows the likelihood, prior, and posterior distribution when we use a flat prior and we see 5 out of 20 events in the experiment. The likelihood function in figure 4 is the same as the likelihood in figure 2.

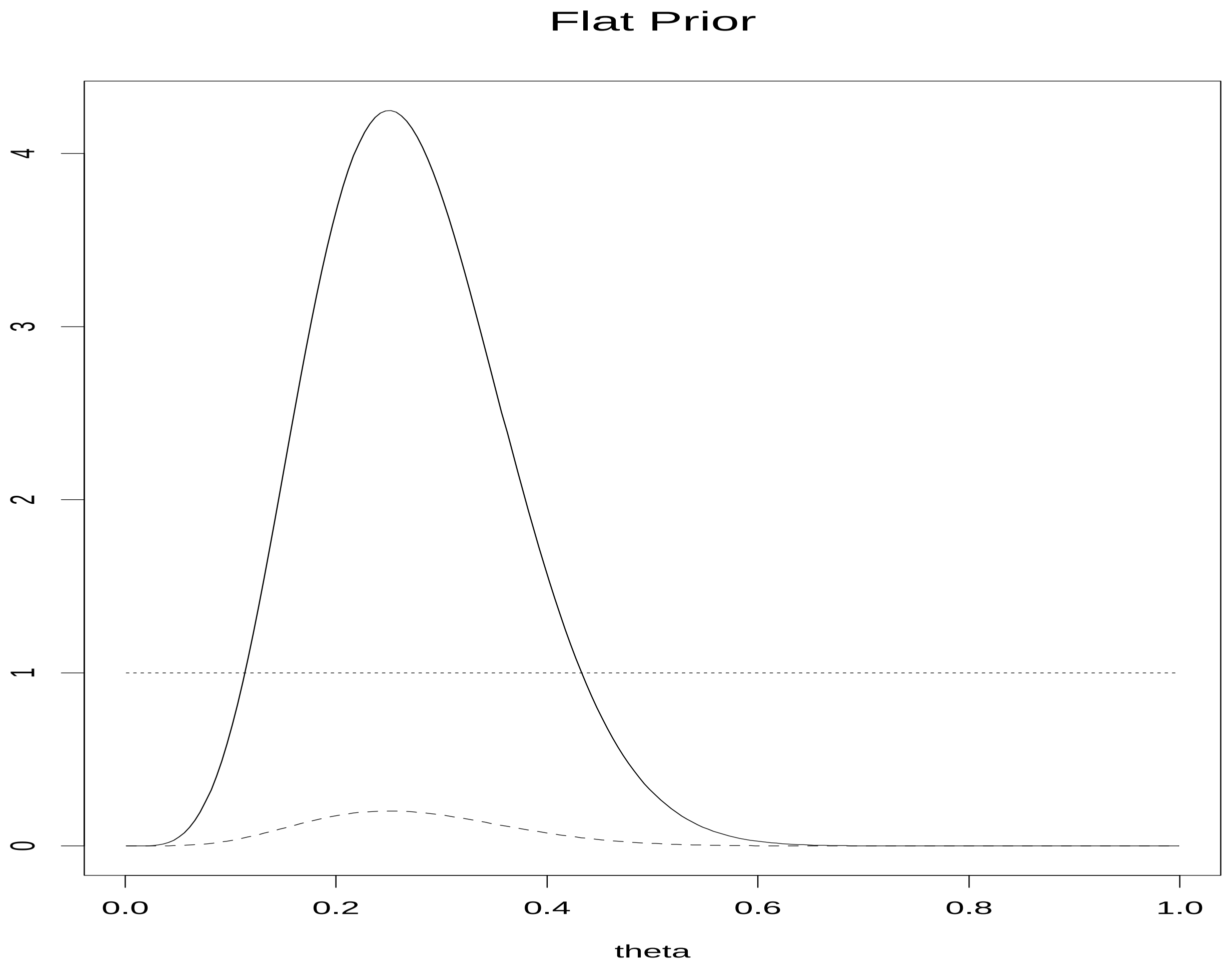


Figure 4: The prior is a Beta(1,1). Dotted line: likelihood, dashed-dotted line: prior, and solid line: posterior distribution.

Flat Priors: Analytical Calculations

- Since the uniform distribution is a Beta(1,1), then the posterior distribution is a Beta($\alpha + 1, \beta + 1$).
- So, if we observe 5 successes out of 20 attempts, then the posterior distribution is a Beta(5+1, 15+1) which is a Beta(6,16).
- Therefore, the mean of the posterior distribution is 6/22 which is .273. Also, the posterior standard deviation is .093.
- Most statistical software packages have a program which can be used to find the percentiles of a beta distribution and using such a program we find that the 95% credible region is the interval $(.1128, .4716)$.

Flat Priors: Analytical Calculations

(continued)

- So, if we had a flat prior for the value θ which is the efficacy of the anti-cancer drug and we observe 5 of the 20 patients improve, the posterior distribution in figure 4 would represent our belief as to the efficacy of this drug.
- Furthermore, we might wish to report that our posterior mean would be .273 with posterior standard deviation of .093 and with a 95% credible region of $(.1128, .4716)$.

Other Flat Priors: Jeffrey's Prior

The uniform prior like it is somewhat equivalent to the "physical symmetry" assumption, but...

- It might be argued that a change of .04 might be different for different parts of the range of values for θ.
- For example, a difference between .52 to .48 might not be as big as a difference between .99 and .95.
- Some might feel that the differences in the far end of the range are in some sense "bigger" than in the middle. One way to think about this is to think that the variances for the estimate of θ are smaller when θ is near the end points.

Jeffrey's Prior

(continued)

- Because of this, people often work with a "variance stabilizing transformation" of θ.
- In this model, the variance stabilizing transformation is $\arcsin(\sqrt{\theta})$. For the random variable $\arcsin(\sqrt{\theta})$, a Beta(1/2,1/2) is a flat prior.
- Using this type of argument, Sir Harold Jeffery, a famous pioneer of Bayesian methods, thought that a Beta(1/2, 1/2) would be good "reference" prior for the parameters of a binomial.
- We will call this the Jeffery's prior for this type of data. He suggested that a reference prior might be used as starting point to compare results between two people.

Jeffrey's Prior

(continued)

- Mathematically, it is a bit challenging to show that when θ follows a Beta(1/2,1/2) distribution then the distribution for the transformed random variable is flat. However, it is fairly easy to show this is approximately correct through a computer simulation.
- A large random sample from a distribution can be used to approximate the properties of the distribution.
- We use a computer to draw a large sample (say 50,000) from the distribution of interest. Then, we can transform the data and look at a histogram of the transformed data. A histogram of this data is an approximation of the density.

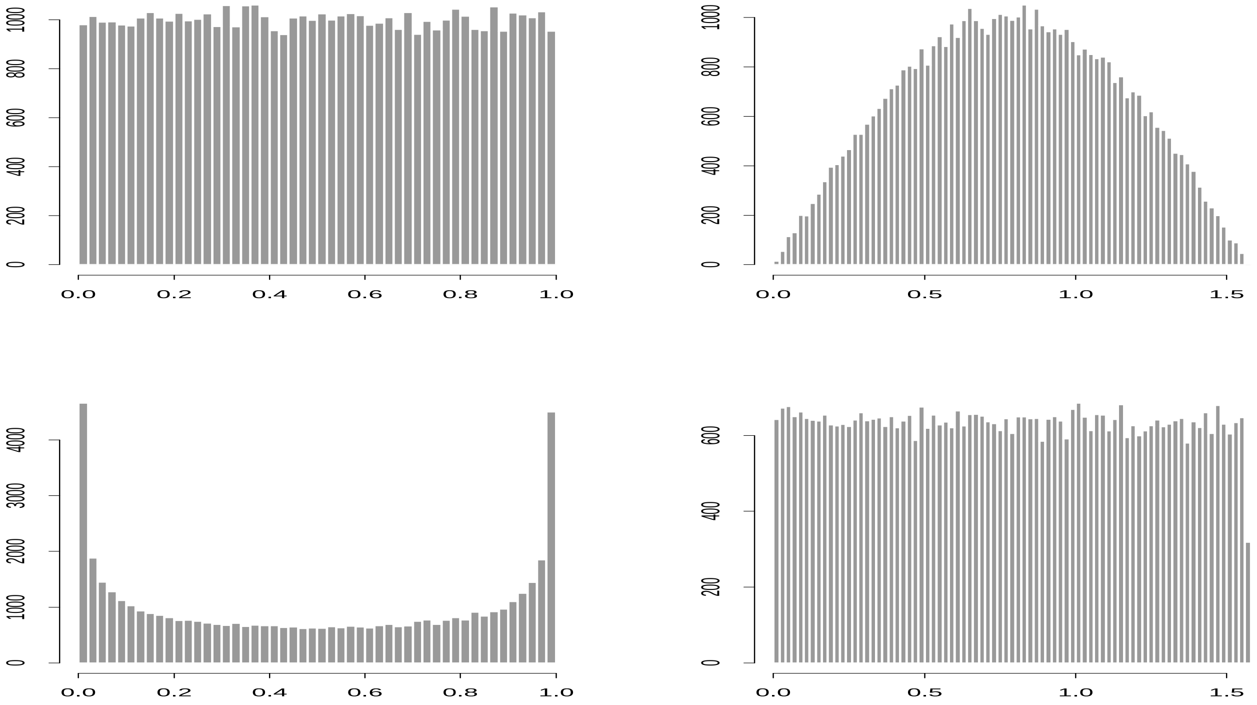


Figure 5: Each plot contains a histogram of a sample of 50,000 observations from the various distributions. The top row is from a Beta(1,1) and the bottom row is for Beta(1/2,1/2). The left column is for θ and the right column is for $\arcsin(\sqrt{\theta})$.

Jeffrey's Prior

(continued)

- Figure 5 contains the plots of θ from either a Beta(1,1) or Beta(1/2,1/2) for the values of θ transformed and not transformed.
- Note that in the transformed scale, the Beta(1/2,1/2) is flat while the Beta(1,1) down weights the extreme values for θ. Please note that in general a prior which is "flat" for some random variable will not be flat for a transformation of that random variable.
- Also, note that the sample size of the prior, the sum of $(\alpha + \beta)$, is 1 here while the flat prior had a sum of 2. So, in some sense, the prior has less weight than the flat prior.

“Very” Uninformative Priors???

- When we observe y successes out of n improvement in a binomial experiment, then the posterior distribution is Beta(a+y, b+n-y).
- So the α parameter for the posterior is “a” plus the number of successes and the β parameter is “b” plus the number of failures.
- So, the α parameter is summing up successes and the β parameter is summing up failures.
- It is as if the prior is contributing “a” successes and “b” failures and in total it is adding a prior data set of size (a+b).

"Very" Uninformative Priors???

(continued)

- The uniform distribution corresponds to a Beta(1,1) prior. So, it is like having a prior sample size of 2.
- The Jeffrey's prior is a Beta(1/2,1/2). So it is like having a prior sample size of 1.
- Thinking along these lines, it has been suggested that maybe a Beta(0,0) would be a good prior. It would have a prior sample size of 0. However, a Beta(0,0) is not a proper distribution. The "area under the curve" is infinity. This can cause problems when doing calculations.

Making Bayesian Inferences

The posterior distribution contains the results of the analysis. It represents the current belief in the parameter of interest. There is a question as to how to communicate this result.

- Present the entire posterior distribution. It could be plotted or one could give the parameters of the beta (if beta prior was used).
- Give summary statistics of the posterior: for example the mean, standard deviation, etc. As a point estimate, one could give posterior mean or median (or other location parameter).

Making Bayesian Inferences

(continued)

- For a range of values one could give a credible region (intervals). This is an interval which contains a certain amount of posterior probability. Giving a 95% probability region is common. (Note that this is different than a confidence interval.) Two common types:
  1. The smallest interval which contains the given amount of probability. (Highest posterior density).
  2. Equal tail area. This is often easy to calculate.

Results for Example 2

- In example 2, there were 5 successes seen out of 20 tries.
- Here, we look at 3 different priors: the uniform prior (Beta(1,1)), the Jeffrey's prior (Beta(.5,.5)), and the improper prior (Beta(0,0)).
- Since these are conjugate priors, then these 3 priors lead to the following 3 posterior distributions: Beta(6,16), Beta(5.5, 15.5), and Beta(5,15).
- The next slide contains a plot of these three posteriors, and the slide after that contains some of the summary values for these posteriors.

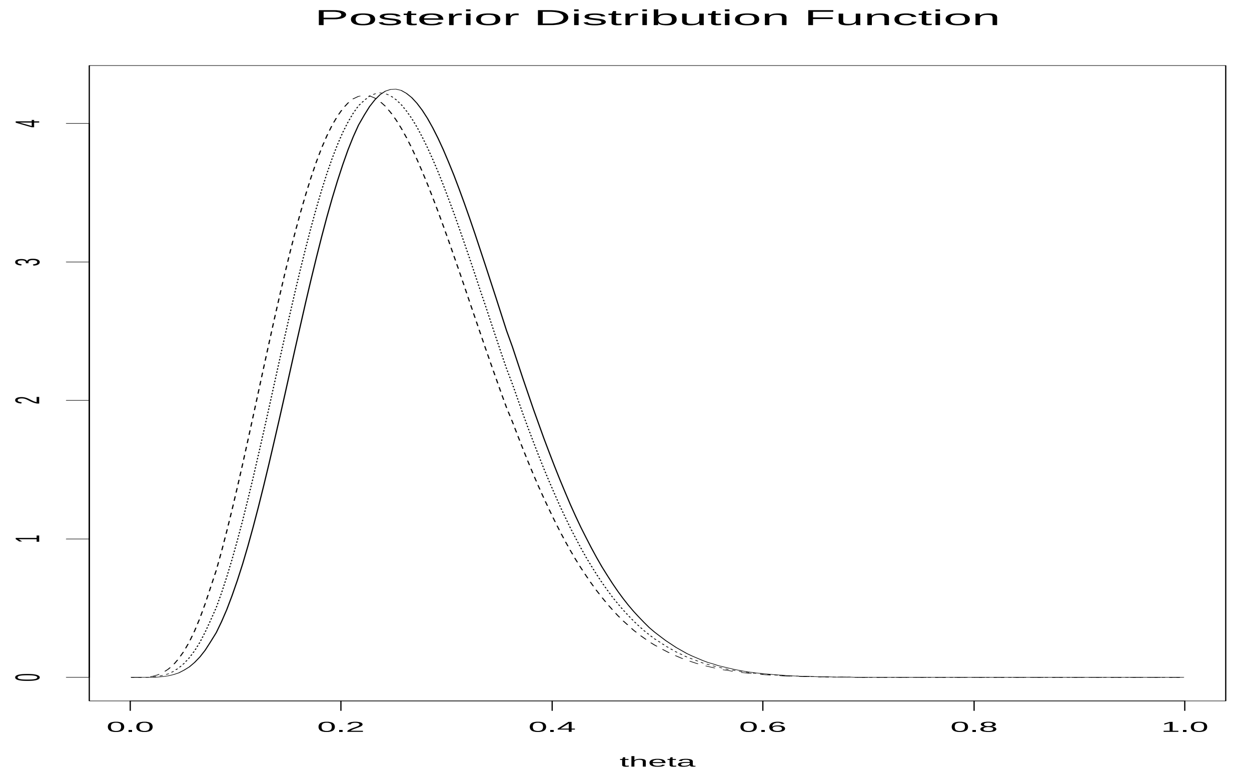


Figure 6: The posterior distributions for the priors: Beta(1,1)- solid line; Beta(.5,.5)- dotted line; Beta(0,0)- dashed line.

Results for Example 2

(continued)

Data: 5 successes out of 20

	Uniform	Jeffrey's	Improper
Prior:	Beta(1,1)	Beta(.5,.5)	Beta(0,0)
Posterior	Beta(6,16)	Beta(5.5, 15.5)	Beta(5, 15)
Mean:	.273	.262	.250
95% Credible Region:	(.113, .472)	(.102, .464)	(.091, .456)

Results for Example 2

(continued)

- Please note that the results from the three different priors are very similar.
- Also, note that with these “uninformative” priors, the likelihood function dominates the posterior. It can be shown that as the sample size gets larger, then the resulting posteriors from these three priors will get very close together.
- A situation when the data/likelihood does not dominate the prior is when there are no successes. This situation is examined next.

Example 3: the zero cases example

- Suppose in a certain standard radiology procedure, it is noted that patients have a serious reaction about 15 times out of 10,000 patients.
- Now suppose that there is a new procedure which is being used and in 167 patients so far, there have been no serious reactions.
- What is believe about the probability of a serious reaction?

note: original example is from Hanley and Lippman-Hand (1983) *Journal of the American Medical Association*, 249, 1743-1745. The Bayesian analysis that is being followed is from Winkler, Smith, and Fryback, (2002), *American Statistician*, 56, 1-4.

Zero Cases example

- Frequentist methods have some difficulty with this situation also. The usual maximum likelihood estimate is the value 0. Also, the confidence intervals are a little more difficult to come by.
- One trick is to put a 1/2 in each group (the number of successes and the number of failures). Please note the similarity between this and the use of the Jeffrey's prior.
- Depending on the purpose of the statistical calculation, an informative prior can be useful here.

An Informative Prior

- If a Beta distribution is used, then there are two parameters which need to be specified.
- For example, might assume that the mean of the new procedure is about the same as the old. Therefore $\alpha/(\alpha + \beta)$ equals .0015.
- Another assumption which might be reasonable would be that there is about a 95% probability that reaction probability is less than 5 times the reaction probability of the old procedure. So, we want a the 95%-tile of the beta to be equal to .0075.

An Informative Prior

(continued)

- So, we want a beta distribution with a mean, $\alpha/(\alpha+\beta)$, to be .0015, and we want the 95%-tile to be .0075.
- Many computer packages, including Excel, SAS, and Splus contain the function to find the value of percentiles of Beta distributions. Using this function one can find the correct beta distribution.
- It can be shown that a Beta(.042, 27.96) has the stated mean and 95%-tile.
- Since there were no reactions in 167 observations, this results in a posterior which is a Beta(.042, 194.96).
- This gives a posterior mean of .00022. Also, a 95% credible region would be (0, .0011).

An Informative Prior

- There might be some objection to using informative priors. Some might say that this “slants” the results.
- However, if one is trying to make a decision based on the best available data, then an informative prior certainly has advantages.
- Sometimes, when the results could be controversial, maybe two priors could be used. One prior is skeptical and the other is enthusiastic. This still allows for informative priors and it shows a range of posterior beliefs which correspond to a range of prior beliefs.

B: Binomial Model - Summary

- This model is good for modelling data which is a sum of successes/failures.
- The beta distribution is conjugate prior for this model. This makes it easy to calculate the posteriors for this model.
- With the beta distribution, one can model low levels of prior information including different “non-informative” priors.
- Sometimes informative priors are very important.

C: Normal Model

Outline:

- The likelihood: the normal distribution.
- Some convenient prior distributions for this model and the associated posterior distributions.
- Some illustrative examples.

Normal models: Introduction

- This is a very common distribution to model data which is continuous and follows a “bell shape”.
- It is parametrized by two parameters: one parameter for location and the other describes the spread.
- Under very weak conditions, the sum of several similar random variables is approximately a normal distribution.
- Some examples which are often modelled by a normal distribution include height and blood pressure as well as many other measurements.

Normal Distribution: Parameters

- The shape of the distribution is a the bell shape curve.
- Describe by two parameters, a mean and a variance. Alternatively, we might describe it by a mean and a standard deviation, where the standard deviation is the square root of the variance term.
- In Bayesian statistics, sometimes it is useful to describe the normal distribution by the mean and the precision parameter, where the precision is 1/variance.

Prior Distribution

- Please note that there are two parameters which need to be considered, the mean and the precision.
- Also, note that the mean could take on values from $-\infty$ to ∞ and the precision takes on values from 0 to ∞. So, a "flat" uniform prior would be improper. (If the density function is a constant over the whole real line, then the "area under the curve" would be infinity.)
- The above fact might worry people who like a "flat earth". That is those who want very uninformative priors, but there are ways to express very little information.
- As in the case with the binomial models, it is easier to first look at the family of conjugate priors. This allows one to look at the mathematics involved. From there, we can look at ways to find priors which might suit one's beliefs.

Conjugate Prior Distributions

- The conjugate prior for the mean μ is the normal distribution.
- The conjugate prior for the precision parameter τ is the gamma distribution.

Conjugate Prior Distributions

(continued)

- The gamma distribution has two parameters, α and β.
- If τ has a gamma distribution with parameters α and β we will write, Gamma$(\tau|\alpha, \beta)$, and it has the following density:

$$f(\tau|\alpha, \beta) = C\beta^{\alpha}\tau^{\alpha-1}e^{-\beta\tau},$$

where C is the normalizing constant and $C = \Gamma(\alpha)$ where $\Gamma(\alpha)$ is the gamma function.

Conjugate Prior for the Mean

For the moment, assume that the precision, τ is known. So, the mean μ is unknown but the precision τ is known.†

- Suppose that $X_1, \ldots, X_n$ is a random sample from a normal distribution with an unknown value of the mean μ and a specific value of the precision $\tau(\tau > 0)$.
- Suppose also that the prior distribution of μ is a normal distribution with mean μ_0 and precision τ_0 such that $-\infty < \mu_0 < \infty$ and $\tau_0 > 0$.

† The derivation is available many places, I got them from DeGroot, M.H. *Optimal Statistical Decision*, (New York: McGraw-Hill, 1970), pg 167.

Conjugate Prior for the Mean

(continued)

Main Point:

- Then the posterior distribution of μ when $X_i = x_i (i = 1, \ldots, n)$ is a normal distribution with mean μ' and precision $\tau_0 + n\tau$, where

$$\mu' = \frac{\tau_0 \mu_0 + n\tau\bar{x}}{\tau_0 + n\tau}$$

Conjugate Prior for the Mean

(continued)

- The posterior precision is the sum of the precision of the prior plus the "precision" from the sample (that is, $n\tau$).
- The posterior mean μ' is a weighted average of the sample mean $\bar{x}$ and the prior mean μ_0. The weighs are proportional to the precisions from the sample and the prior.
- To represent a high uncertainty in the prior, one could specify little precision. That is, make τ_0 small. In this case, we can see that the posterior mean and posterior precision will be dominated by the sample values. (However, in his case, we are assuming that τ is known.)
- Note also, that as $n \to \infty$, the likelihood overwhelms the prior and the posterior mean converges to the sample mean.

Conjugate Prior for Variance

For the moment, assume that the mean μ is known. So, the mean μ is known but the precision τ is unknown.†

- Suppose that $X_1, \ldots, X_n$ is a random sample from a normal distribution with a specific value of the mean $\mu \quad (-\infty < m < \infty)$ and an unknown value of the precision τ.
- Suppose also that the prior distribution of τ is a gamma distribution with parameters α and β such that $\alpha > 0$ and $\beta > 0$.

† DeGroot, *op. cit.*, pg 168.

Conjugate Prior for Variance

(continued)

- The posterior distribution of τ when $X_i = x_i (i = 1, \ldots, n)$ is a gamma distribution with parameters α' and β' where:

$$\alpha' = \alpha + (n/2), \text{ and}$$

$$\beta' = \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Conjugate Prior for Variance

(continued)

- Remember, the gamma distribution is a generalized version of the chi-squared distribution.
- With that in mind, note that in the prior, 2α is like the degrees of freedom for the prior. The posterior "degrees of freedom" is equal to the sample size plus the degrees of freedom for the prior.
- The parameter β' is equal to the prior β plus half the residual sum of squares. So, the $\beta's$ plays the role of the residual sum of squares (a measure of total squared deviation).

Conjugate Prior for Variance

(continued)

- For the gamma density with parameters α' and β', the mean of this distribution is α'/β'. So, the harmonic mean of the posterior distribution of the variance is β'/α', which has the form of the "posterior residual sum of squares" divided by the "posterior degrees of freedom".
- When α and β are small or the sample size is large, the posterior is dominated by the sample statistics.

Joint Conjugate Prior

(continued)

Now, consider the case where both μ and τ are unknown.

- Suppose that $X_1, \ldots, X_n$ is a random sample from a normal distribution with an unknown value of the mean μ and an unknown value of the precision τ.
- Suppose also that the prior joint distribution of μ and τ is as follows:
  - The conditional distribution of μ when $\tau = t (t > 0)$ is a normal distribution with mean μ_0 and precision θt such that $-\infty < \mu_0 < \infty$ and $\theta > 0$, and
  - the marginal distribution of τ is a gamma distribution with parameters α and β such that $\alpha > 0$ and $\beta > 0$.

Joint Conjugate Prior

(continued)

Then the posterior joint distribution of μ and τ when $X_i = x_i (i = 1, \ldots, n)$ is as follows:

- The conditional distribution of μ when $\tau = t$ is a normal distribution with mean μ' and precision $(\theta + n)t$, where

$$\mu_0' = \frac{\theta \mu_0 + n\bar{x}}{\theta + n}.$$

- The marginal distribution of τ is a gamma distribution with parameters $\alpha + (n/2)$ and β', where

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\theta n(\bar{x} - \mu_0)^2}{2(\theta + n)}.$$

Joint Conjugate Prior

(continued)

- Also, the marginal prior distribution of μ follows a general t-distribution with 2α degrees of freedom, location parameter μ_0, and precision $\alpha\theta/\beta$.
- The posterior marginal distribution of μ is obtained by replacing μ_0, θ, α, and β by their posterior values, namely μ_0', $\theta + n$, $\alpha + (n/2)$, and β' respectively.

Joint Conjugate Prior: Comments

Where both μ and τ are unknown, please note that:

- The parameter τ is doing double duty. It is a measure of the variation between the individually observed X_i's and there mean μ. Also, it measures the uncertainty that μ is μ_0. That is, it tells how much we are allowing μ to vary from μ_0. If we had samples from more than one group where each group had a separate mean μ_i, then we would have separate information on the variation between the X_i's and their μ_i's and between the μ_i's and μ_0. However, with this data, one cannot separate out this information.
- Because of the double duty performed by τ, the parameter θ is included in the model. In effect, this parameter allows one to split up the roles that τ plays.

Joint Conjugate Prior: Comments

(continued)

- For the conditional posterior of μ given that $\tau = t$, the distribution is a normal distribution.
  - The posterior precision is the sum of the prior precision and sample "precision".
  - The mean of this distribution is again a weighted average of the prior mean and the sample mean. Note that in this ratio, there is a common value t which cancels out.

Joint Conjugate Prior: Comments

(continued)

- The marginal distribution of τ is again a gamma distribution. The new α' is the same as for the case when μ was know. The difference between the case when μ was known or unknown is the value of β'. The parameter β' now contains a term which measures the deviation between the individual mean μ and the prior mean μ_0.

Joint Conjugate Prior: Comments

(continued)

- The marginal distribution of μ has a general t-distribution. So, the following expression has a standard t-distribution on $2\alpha'$ degrees of freedom:

$$(\mu - \mu_0') \Bigg/ \sqrt{\frac{\beta'}{(\theta + n)(\alpha')}} \; .$$

  - Since β' is like an error sum of squares, then the above expression should remind one of a typical t-statistic.
  - In the above expression, it is the μ which is the random variable, not the μ_0' (which is playing the role of $\bar{x}$ here).

Example 4: Heights

Consider the height of the next person to come through a doorway. In this example, this uncertainty is modelled with a normal distribution. This example is in the supplemental slides and it contains the following steps:

- The problem is stated.
- The values for the priors are considered.
- Data is presented.
- The beliefs are updated.
- The results are discussed.

Example 4: Heights

Consider the height of the next person to come through a doorway. In this example, this uncertainty is modelled with a normal distribution. This example goes through the following steps:

- The problem is stated.
- The values for the priors are considered.
- Data is presented.
- The beliefs are updated.
- The results are discussed.

Example 4: Statement of the Problem

- Consider the height of the next person to come through a doorway. To be more precise, I am considering the the height of a student who will be walking through the doorway of a large statistics class. I am assuming that the distribution of student heights follows a normal distribution, so all I need to specify is the mean and precision parameters of this distribution. Alternatively, I could specify the mean and standard deviation of this distribution.
- Therefore, I need to put priors on my belief in the values of the mean and precision parameters. Here, I use the conjugate priors and I assume that I don't know either the mean or the precision.

Example 4: Statement of the Problem

(continued)

- So, I need to specify the values for the parameters of the conjugate prior distribution. Then, the data and the prior information is then combined to give the posterior distribution.
- Also, I do have data for 71 students, and the analysis is completed on these data and compared to the results obtained for the smaller data set.

Example 4: The Model

In symbols, we have the following model:

$$\begin{aligned} X_i|\mu,\tau &\sim N(X_i|\mu,\tau) \\ \mu|\mu_0,\tau,\theta &\sim N(\mu|\mu_0,\theta\tau) \\ \tau|\alpha,\beta &\sim \text{Gam}(\alpha,\beta) \end{aligned}$$

where $i = 1, \ldots, n$

We need to specify values for μ_0, θ, α, and β.

Example 4: Elicitation of the Priors

- Please note that these are my priors. Yours might be different.
- Here, I do an “honest” elicitation. My goal is to try and specify my beliefs without being overly conservative.
- First, I get the prior for μ, and then the prior for τ.

Example 4: Elicitation of the Priors

(continued)

For the mean:

- I guess that average height is between 5 feet (150cm) and 6 feet (180cm). Most people seem to be within that range, so the mean must be somewhere in that range.
- So, a vague guess is that it is about 5 feet 6 inches, which is 66 inches. Therefore, let $\mu_0 = 66$.

Example 4: Elicitation of the Priors

For the precision τ, one needs the values for θ, α, and β.

- For τ and θ, I first consider the distribution $(X_i|\mu)$ which has precision τ. That is, the distribution of the individual heights around the mean height (assuming that it is known).
- Consider an approximate 95% interval. That would be $\mu \pm 2\sigma$ where σ is the standard deviation.
- I think that I do not expect more than 1 out 20 people to be outside the range of 4 feet 6 inches and 6 feet 6 inches. So, I consider my 95% interval to be about 24 inches wide which corresponds to a σ of 6 inches.
- Now maybe here I should be assuming that $(X|\mu_0)$ has a standard deviation of σ, but to be conservative, let us consider the σ as the standard deviation of just $(X|\mu)$.

Example 4: Elicitation of the Priors

I still need to specify θ and the pair of parameters for the gamma distribution, α and β. Now, consider the value for θ.

- The parameter theta is the ratio of the precisions for $(\mu|\mu_0)$ to $(X_i|\mu)$. This is the same as $\mathrm{Var}(X_i|\mu)/\mathrm{Var}(\mu|\mu_0)$.
- By the above argument, I stated that I believed that $\mathrm{Var}(X_i|\mu)$ is about 6^2.
- In considering μ_0, I stated that I believed it was roughly between 5 feet and 6 feet. Let's consider that range to be a 95% range, so I believe the standard deviation is about 3 inches.
- So, the ratio $\mathrm{Var}(X_i|\mu)/\mathrm{Var}(\mu|\mu_0)$ is $6^2/3^2$ which corresponds to a θ of 4.

Example 4: Elicitation of the Priors

Now, let's look at the prior distribution for τ.

- So far, I have been thinking about σ which is $1/\sqrt{\tau}$, and I have argued that I think σ is about 6. Let's interpret this to mean that the median of σ is 6. So, this gives us one point of the gamma distribution.

Example 4: Elicitation of the Priors

(Continuing with prior for τ)

- In order to quantify the spread of this distribution, let's think about how small we might consider σ to be.
- That is, if we consider a region which contains about 95% of the heights, how small would we think that is? I think that the spread of heights must be at least 10 or 12 inches.
- So, if we consider the 95% region to be $\mu \pm 2\sigma$, then I think that σ "must be" bigger than about 2.5 inches. When I say "must be", let me interpret to mean that this is about the 2.5%-tile point. This gives me the second point for the gamma distribution.

Example 4: Elicitation of the Priors

(Continuing with prior for τ)

- Now the challenge is to find parameters α and β such that the 2.5%-tile point of σ is about 2.5 and the 50%-tile point of σ is about 6 and where $1/\sigma^2$ has a gamma distribution with parameters α and β.
- After a little bit of guess work, using $\alpha = 1$ and $\beta = 25$ seems fairly close.
- For these values, the 2.5%-tile of σ is 2.6 and the 50%-tile is 6.0.

Example 4: Distribution Function for the Priors

When looking for percentiles, moments, or other features of functions of different random variables, it is helpful to know some of the mathematical properties of the distribution of interest.

- There is a simple way to see what the features of a distribution are without resorting too much mathematics.
- One can simply generate a large number of random numbers from the distribution of interest. Then, these sampled points can be transformed anyway desired and the sample statistics will closely approximate their theoretical counterparts.
- For example first generate 10,000 t_i's from a gamma $(\alpha = 1, \beta = 25)$. Then define $s_i = 1/\sqrt(t_i)$ and find the 2.5%-tile and the 50%-tile of these s_i's.
- When I did this, I got the values 2.6 and 6.0 respectively.

Example 4: Distribution Function for the Priors

(continued)

- Knowing more mathematics means that one can take some shortcuts. For example, one can first find the 97.5%-tile and 50%-tile of the t_i's and then transform these statistics by $1/\sqrt(\cdot)$ and this would give us 2.6 and 6.0 again.
- Also, one could use the quantile function for the gamma distribution to find the value of these percentiles directly.
- However, one should know how to get things by simulation also.
- For example, using these methods, one could find out the mean and the variance of σ quite painlessly without working out the distribution of the transformed variable.

Example 4: Elicitation of the Priors

The priors are now chosen. The parameters of the prior are:

$$\begin{aligned} \mu_0 &= 66 \\ \theta &= 4 \\ \alpha &= 1 \\ \beta &= 25 \end{aligned}$$

Example 4: Data

With the prior now chosen, we can now consider the data.

- A sample of 6 students who walked through the door was collected.
- They had the following heights (in inches): 64, 73, 64, 63, 69, and 71.
- The summary statistics are: $\bar{x} = 67.333$, and $\sum(x_i - \bar{x})^2 = 89.33$.

Example 4: Posterior for Mean

So, the conditional distribution of $\mu|X, \tau, \theta, \mu_0$ has a precision of $(\theta + n)\tau$ which is 10τ and a mean of μ_0' with:

$$\begin{aligned} \mu_0' &= \frac{\theta\mu_0 + n\bar{x}}{\theta + n} \\ &= \frac{(4)(66) + (6)(67.333)}{4 + 6} \\ &= 66.80 \end{aligned}$$

Example 4: Posterior for Precision

- The marginal posterior distribution of τ is a gamma distribution with parameters α' and β'.
- Given the above data, then α' is equal to $\alpha + n/2$ which is $1 + 6/2$ or 4.
- Also, for β', there is the following:

$$\begin{aligned}
\beta' &= \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\theta n(\bar{x} - \mu_0)^2}{2((\theta + n)} \\
&= 25 + \frac{1}{2}(89.33) + \frac{(4)(6)(67.333 - 66)^2}{2(6+4)} \\
&= 25 + 44.665 + 2.132 \\
&= 71.797
\end{aligned}$$

- Therefore, it is a gamma with parameters 4 and 71.797.

Example 4: Showing the Posterior

- By generating 50,000 values of the Gam(4, 71,797), then one can see that the posterior mean of $1/\sqrt{\tau}$ is 4.69. Also, a 95%-tile credible region for $1/\sqrt{\tau}$ is $[2.87, 8.25]$.
- This is obtained from the following Splus code:

```
 tau_rgamma(10000,4,71.797)

 mean(1/sqrt(tau))
[1] 4.698416

 quantile(1/sqrt(tau),c(.025,.50,.975))
    2.5%     50.0%     97.5%
 2.87115 4.411425 8.248018
```

Example 4: Showing the Posterior

(continued)

- These samples can be used to get a plot of the density using a density estimator. The following Splus code demonstrates this.

```
plot(density(1/sqrt(xxx)),type="l",ylab=" ",
xlab="SD")
```

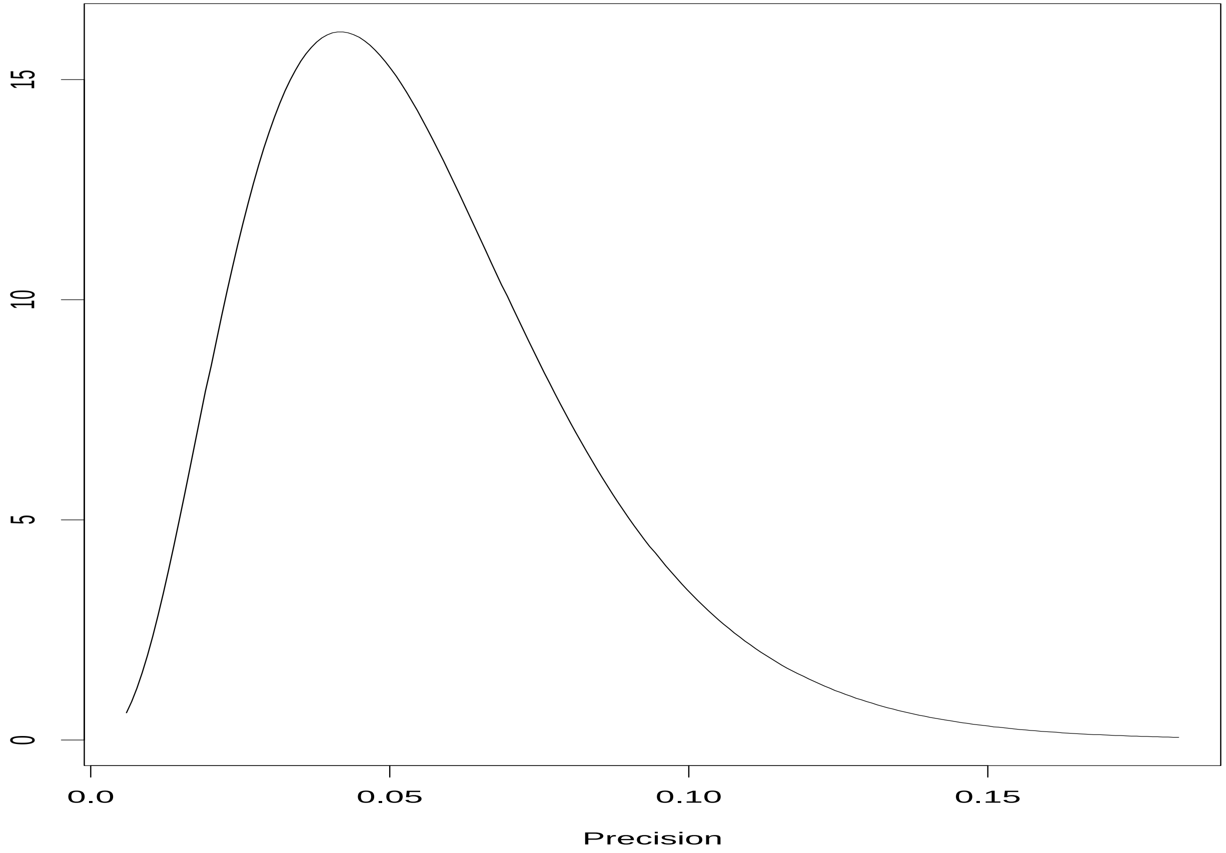


Figure 7: The Posterior density of the precision parameter τ.

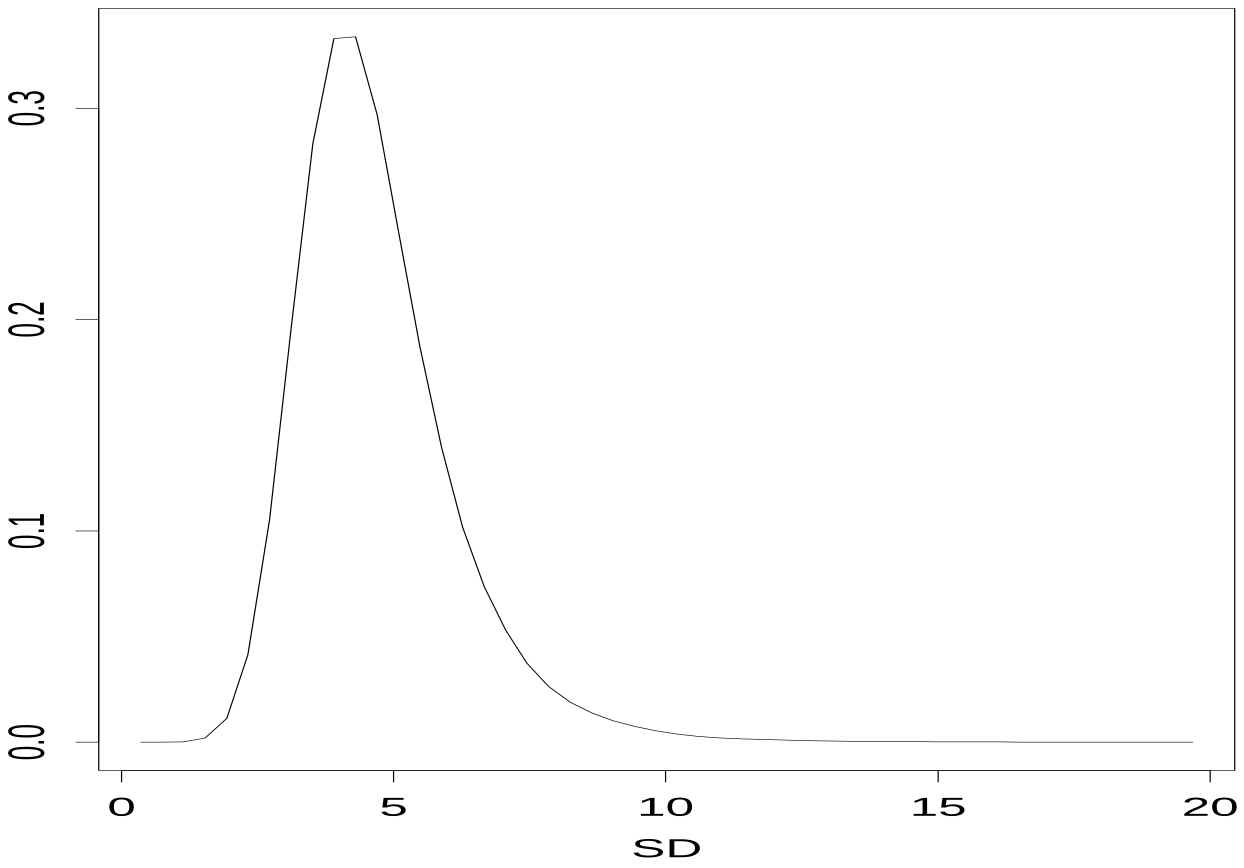


Figure 8: The Posterior density of the standard deviation which is $1/\sqrt{\tau}$.

Example 4: Showing the Posterior

(continued)

- Similarly, we know that the posterior distribution of μ given τ is normal with mean 66.80 and precision 10τ.
- So, we can generate samples of the joint distribution of (μ, τ) or we can generate samples from (μ, SD) where $\mathrm{SD} = 1/\sqrt{\tau}$.
- The following commands will generate this joint distribution:

```
tau_rgamma(10000,4,71.797)   # previously generated
mu_rnorm(10000, mean=66.80, sd=1/sqrt(10*tau))
```

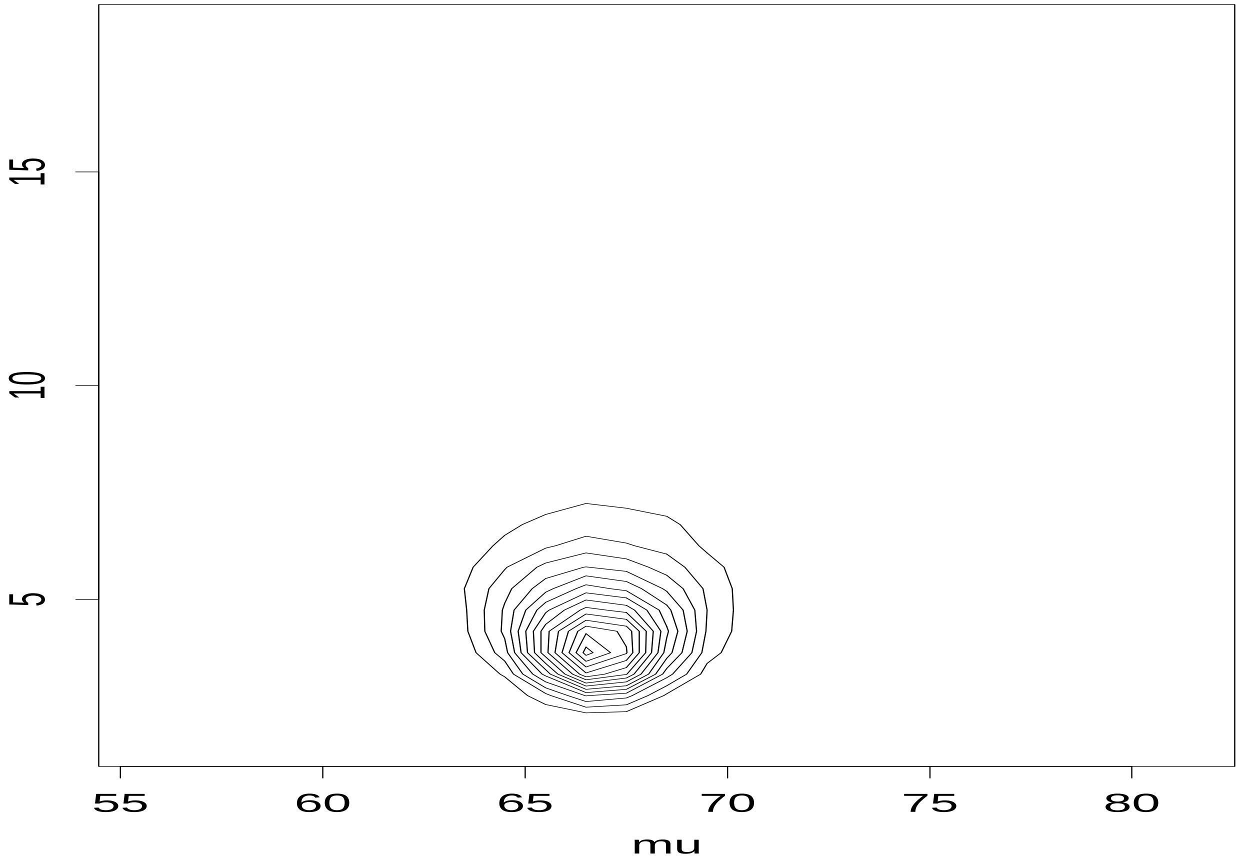


Figure 9: The joint posterior density of μ and $1/\sqrt{\tau}$. The area inside all the circles is a 95%-tile area.

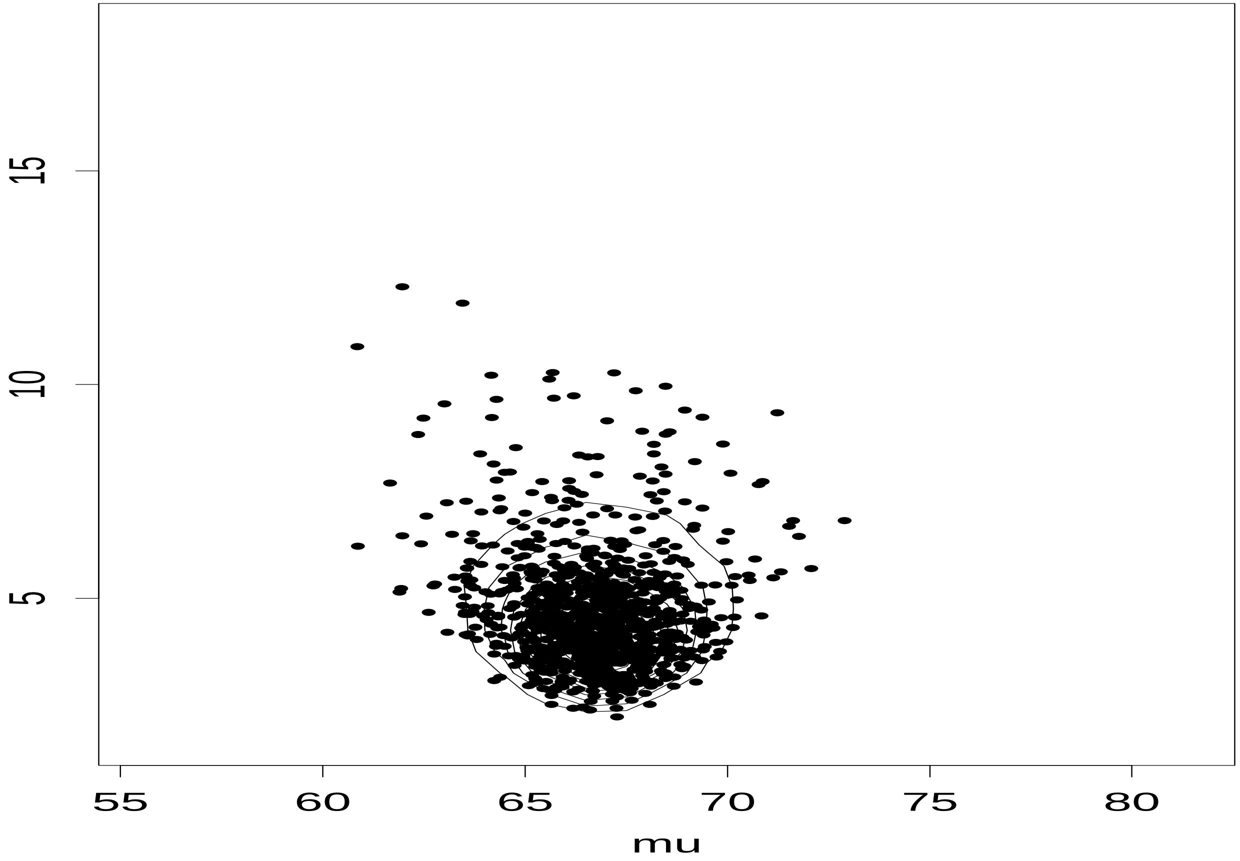


Figure 10: The joint posterior density of μ and $1/\sqrt{\tau}$, with 1,000 random points plotted.

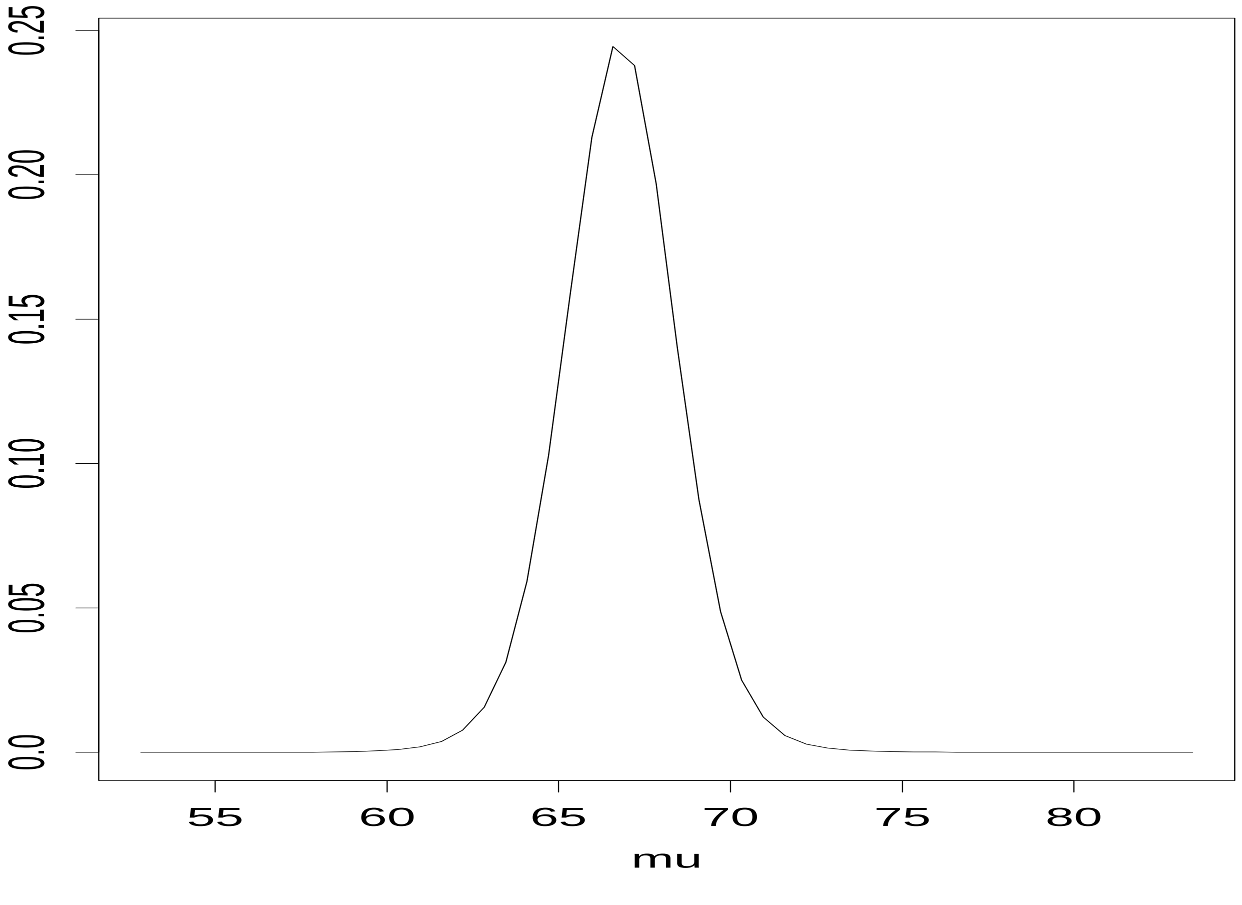


Figure 11: The marginal posterior density of μ.

Example 4: Showing the Posterior

(continued)

- Please note that although the samples of μ are drawn from a normal distribution (given the value of τ), the marginal distribution of the sampled μ's are from a t-distribution with 8 degrees of freedom.

Example 4: Showing the Posterior

(continued)

- The following Splus commands can be used to estimate the joint distribution from a sample of points from the joint distribution:

```
contour(hist2d(list(x=xmu,y=1/sqrt(10*xgam)),
   nxbin=30,nybin=30),
   xlab="mu",ylab="sd",labex=0,nlevels=19)
iind_seq(1,1000)
points(xmu[iind],1/sqrt(xgam[iind]))
```

Example 4: More Data

Now, let's look at what happens when we have more data. In this case we now have the heights of 71 students.

- The summary statistics are: $n = 71$, $\bar{x} = 67.35$, and $\sum(x_i - \bar{x})^2 = 901.33$.
- Please remember that the parameters for the prior were: $\mu_0 = 66$, $\theta = 4$, $\alpha = 1$, and $\beta = 25$.

Example 4: More Data

(continued)

Following the same methods used for the smaller data, then the conditional distribution of $\mu|X, \tau, \theta, \mu_0$ has a precision of $(\theta + n)\tau$ which is 75τ and a mean of μ_0' with:

$$\begin{aligned} \mu_0' &= \frac{\theta\mu_0 + n\bar{x}}{\theta + n} \\ &= \frac{(4)(66) + (71)(67.35)}{4 + 71} \\ &= 67.28 \end{aligned}$$

Example 4: More Data

- Again, the marginal posterior distribution of τ is a gamma distribution with parameters α' and β'.
- For α', α' is equal to $\alpha + n/2$ which is $1 + 71/2$ or 36.5.
- Also, for β', there is the following:

$$\begin{aligned}
\beta' &= \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\theta n(\bar{x} - \mu_0)^2}{2((\theta + n)} \\
&= 25 + \frac{1}{2}(901.33) + \frac{(4)(71)(67.35 - 66)^2}{2(71 + 4)} \\
&= 25 + 450.665 + 3.451 \\
&= 479.116
\end{aligned}$$

- Therefore, it is a gamma with parameters 36.5 and 479.116.

Example 4: Comparing data sets

In comparing the large data set to the small data set, please note the following two points which are discussed in more detail in the next few slides:

- As the size of the data gets bigger, then the information becomes more precise.
- As the size of the data gets bigger, then the information data overwhelms the information from the prior.

Example 4: Comparing data sets

(continued)

For larger data sets, the posterior distribution gets more concentrated. The information becomes more precise.

- The posterior distribution of μ is a normal distribution, and the precision is equal to $(n + \theta)\tau$. So, the precision becomes larger as n gets bigger.
- The posterior distribution of τ is a gamma distribution. The shape of the gamma distribution gets "tighter" as the parameter α' gets bigger. Since $\alpha' = \alpha + n/2$, then α' gets bigger as n gets bigger.

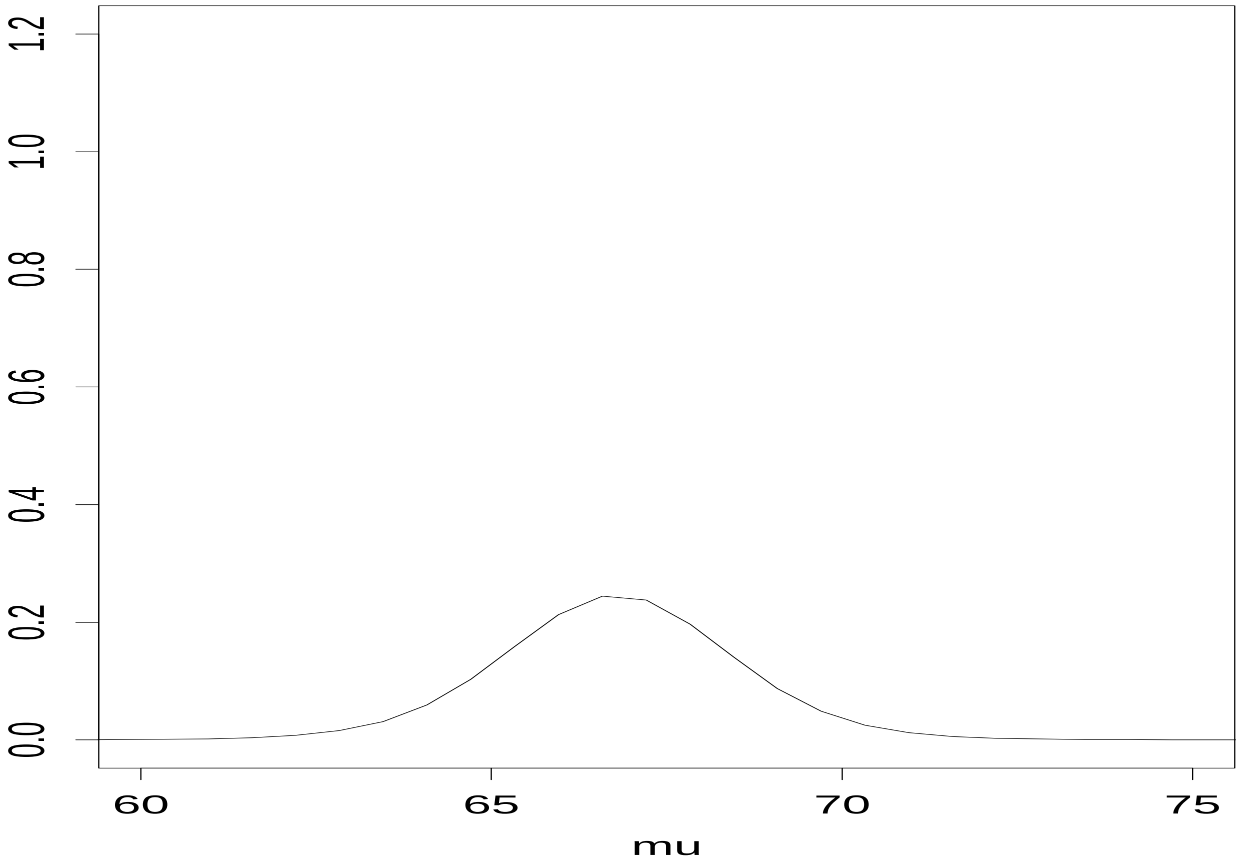


Figure 12: The marginal posterior density of μ. The solid line is for the small data set.

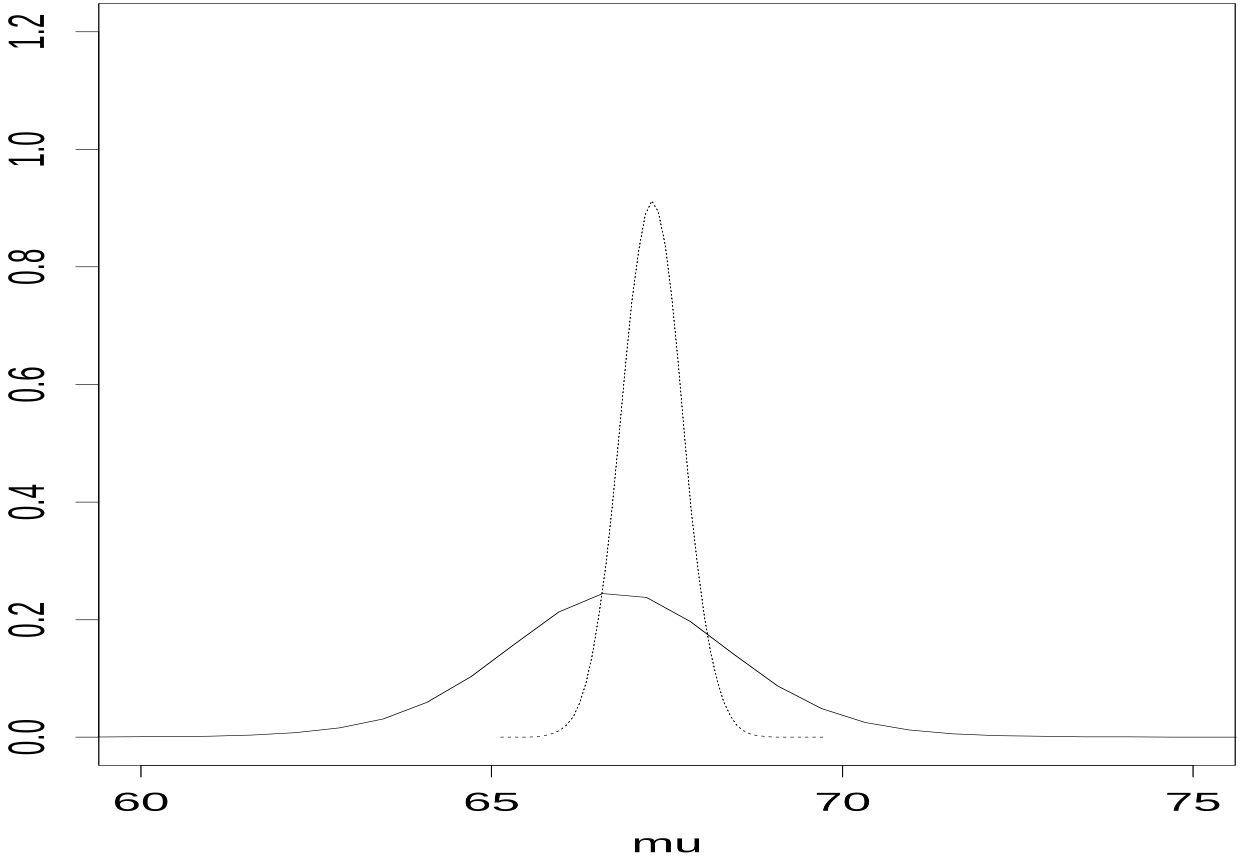


Figure 13: The marginal posterior density of μ. The solid line is for the small data set. The dashed line is for the larger data set.

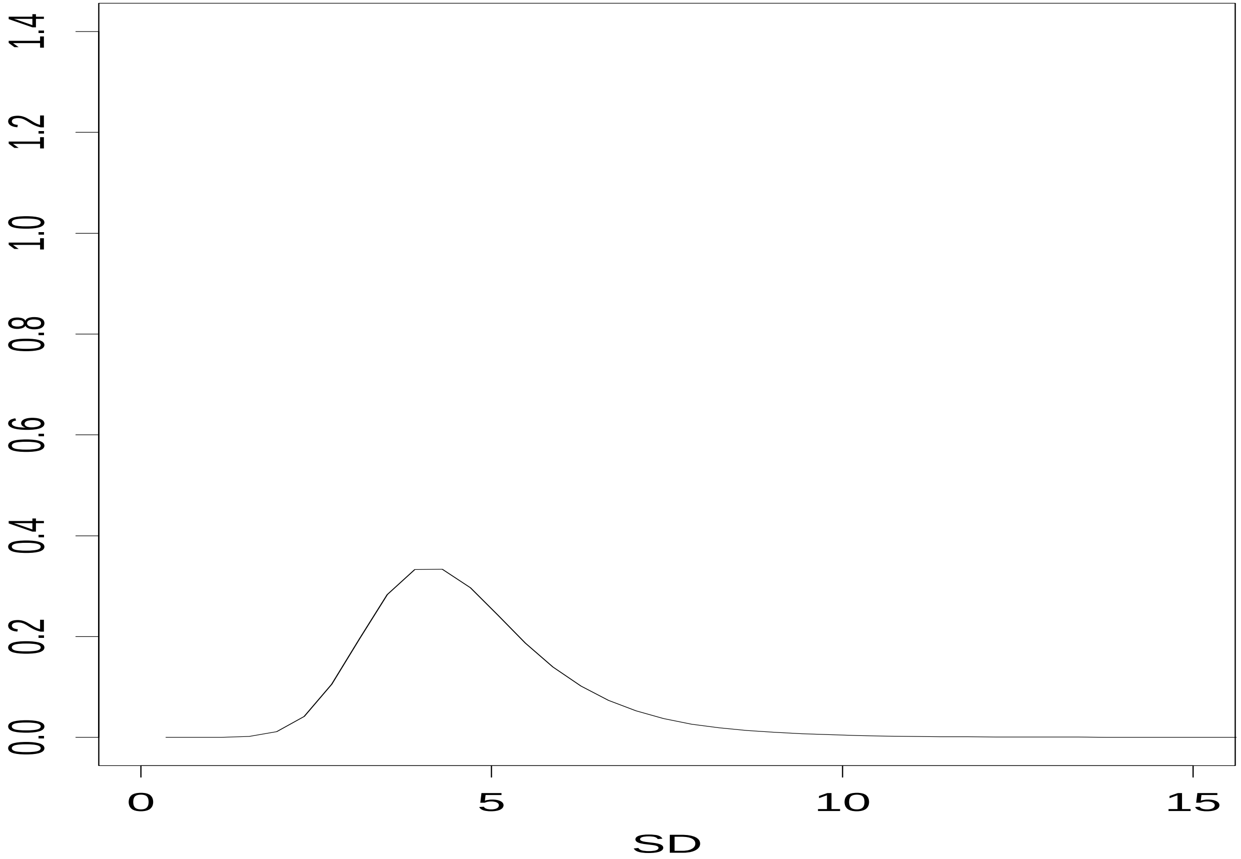


Figure 14: The marginal posterior density of $1/\sqrt{\tau}$. The solid line is for the small data set.

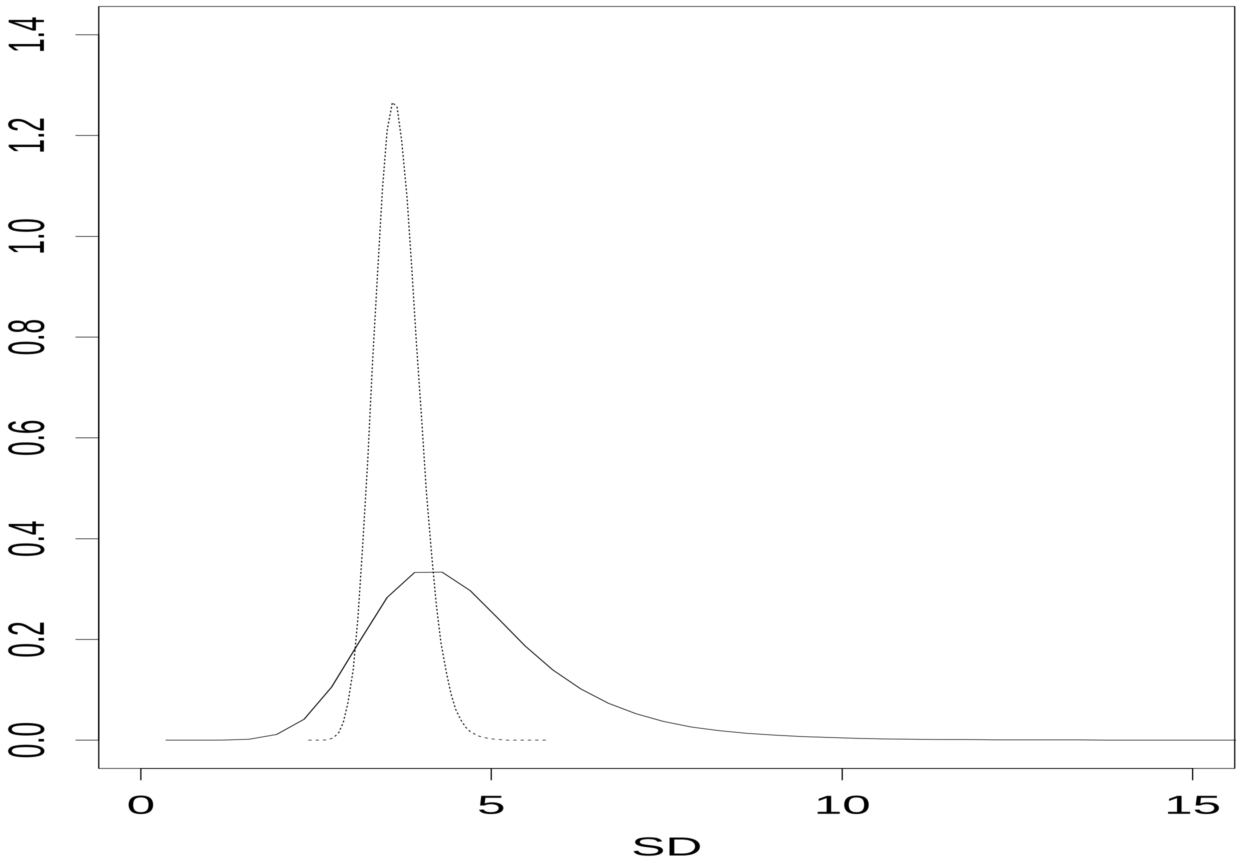


Figure 15: The marginal posterior density of $1/\sqrt{\tau}$. The solid line is for the small data set. The dashed line is for the larger data set.

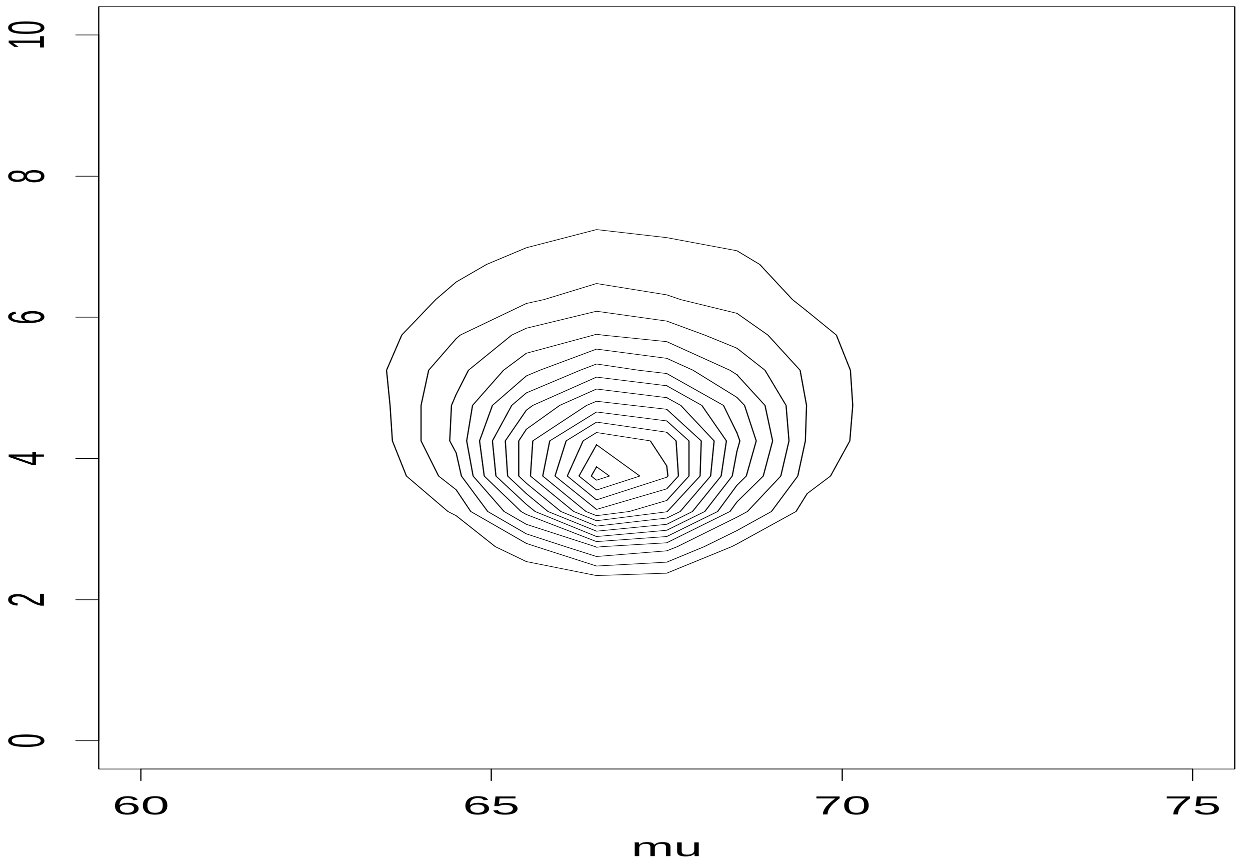


Figure 16: The joint distribution of $(\mu, 1/\sqrt{\tau})$. This is for the small data set.

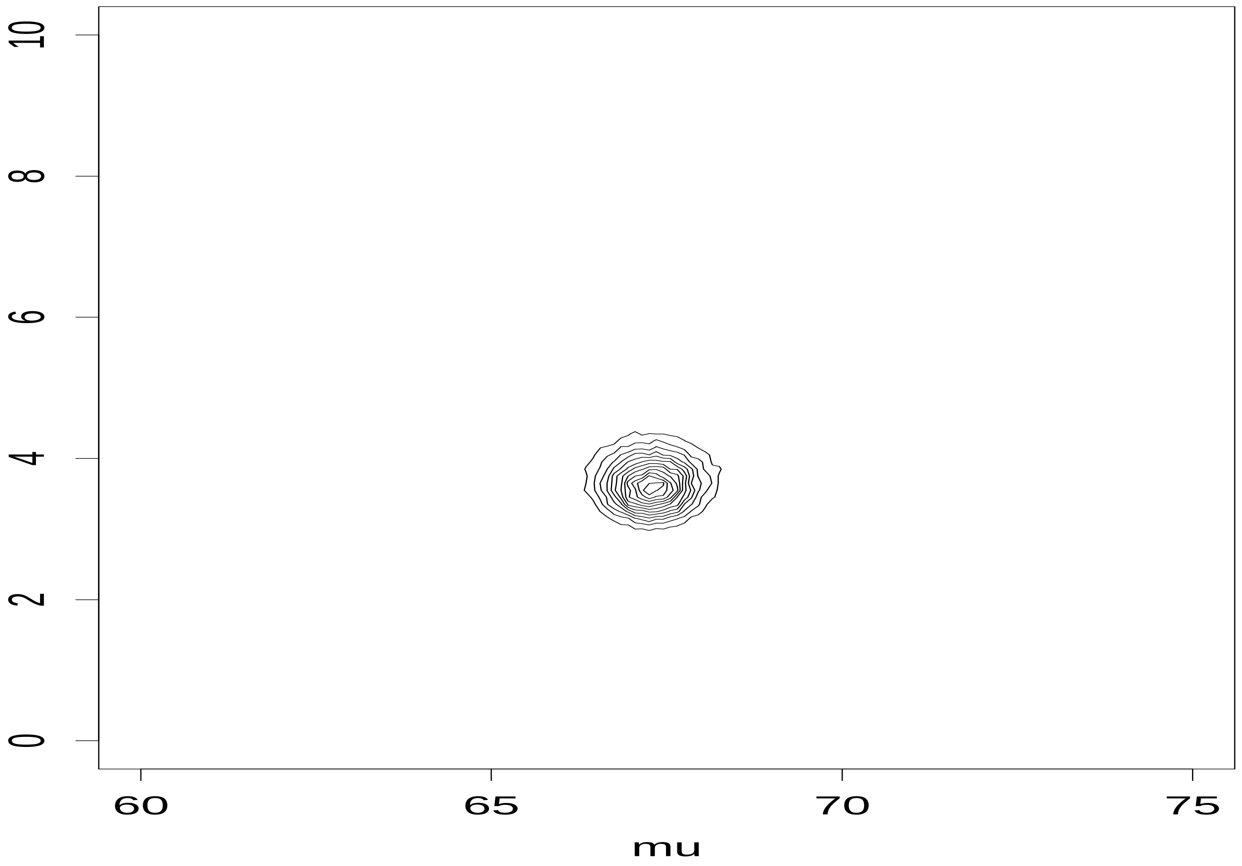


Figure 17: The joint distribution of $(\mu, 1/\sqrt{\tau})$. This is for the large data set.

Example 4: Comparing data sets

(continued)

For larger data sets, the information from the data starts to overwhelm the prior information.

- For the posterior of μ, please note that since n is much larger than θ, when calculating the parameter μ_0', the weighted average gets closer and closer to $\bar{x}$ as n gets larger.
- Also, for the posterior of τ, please note that as n gets larger, then the term $\sum_{i=1}^{n}(x_i - \bar{x})^2$ starts to dominate the parameter β'. The posterior mean of $1/\tau$ is $\beta'/(\alpha' - 1)$. This will look more and more like $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ as n gets bigger (and under certain other regularity conditions).

Low and “Non”-informative Priors

- From our understanding of the effects of the parameters of the prior and the effects of larger sample sizes, we can get an idea how to have priors with low information.
- First, a low informative prior is considered for this example. A prior is chosen so that it represents very vague information.
- From looking at the vague prior, a prior is considered which appears to have almost no information.

Low Info Prior

Now, consider a prior which has very low information.

- First consider the prior for τ. This is the inverse of the variance of x_i given μ. So, think about the possible values of the standard deviation of x_i around its mean μ.
- Since the values of the heights are reported to the nearest inch, then the standard deviation should be bigger than .25 inches.
- For the upper range for the standard deviation, we might consider some real large value. For example, the upper end of a 95%-tile range for heights is probably below 8 feet, so having 2 feet (24 inches) as the upper end of a 95%-tile value for the standard deviation is probably quite conservative.

Low Info Prior

(continue with prior for τ)

- So, from the previous slide, need the value of α and β so that the 2.5%-tile and 97.5%-tile value of $1/\sqrt{\tau}$ are 0.25 and 24.0 respectively.
- After a series of trial and error guesses, it can be shown that if τ has a gamma distribution with parameters 0.46 and 0.15, then the 2.5%-tile and 97.5%-tile of $1/\sqrt{\tau}$ are 0.25 and 23.5 respectively. So, this is close enough for this exercise.
- Please note that this is a highly skewed distribution. The expected value (mean) of $1/\sqrt{\tau}$ is 10.5 (which seems okay), but the median is 0.9. So, this might be cause for concern and the sensitivity to this prior would be checked in any final report of this analysis.

Low Info Prior

(continue)

Now consider the prior for μ.

- Being vague about the prior for μ means that one desires the precision for μ to be very small.
- Presently, the model is parametrized so that the precision of μ is $\theta\tau$. When τ has a gamma prior distribution with parameters 0.46 and 0.15, then the prior mean and prior median of τ is 3.1 and 1.3 respectively.

Low Info Prior

(continue prior for μ)

- So, to be very vague, let's consider a ridiculously large range for the possible values of μ. Say, that a 95% range would be 8 feet. This would mean that the standard deviation would be 2 feet (or 24 inches) and this corresponds to a precision of about 0.0017 (which is $1/24^2$). Since the mean and median of τ are 3.1 and 1.3, then to have $\theta\tau$ be about 0.0017, then we need θ to be about 0.002.
- So, for our vague prior, let $\theta = 0.002$.
- For μ_0, we are uncertain about this, but with such a low precision, we just need it in the general ballpark, so let $\mu_0 = 66$ like before.

Example 4: Low Info

Now, the posterior distribution is calculated with the new prior. For simplicity, only the values for the small data set with 6 subjects are calculated here. The calculations for the larger data set are similar.

- The summary statistics are: $n = 6$, $\bar{x} = 67.333$, and $\sum(x_i - \bar{x})^2 = 89.33$.
- Please remember that the parameters for the prior were: $\mu_0 = 66$, $\theta = .002$, $\alpha = 0.46$, and $\beta = 0.15$.

Example 4: Low Info

(continued)

Again, the conditional distribution of $\mu | X, \tau, \theta, \mu_0$ has a precision of $(\theta + n)\tau$ which is 6.002τ and a mean of μ_0' with:

$$\begin{aligned}
\mu_0' &= \frac{\theta\mu_0 + n\bar{x}}{\theta + n} \\
&= \frac{(.002)(66) + (6)(67.333)}{.002 + 6} \\
&= 67.3326
\end{aligned}$$

Example 4: Low Info Prior

- Again, the marginal posterior distribution of τ is a gamma distribution with parameters α' and β'.
- For α', α' is equal to $\alpha + n/2$ which is $.46 + 6/2$ or 3.46.
- Also, for β', there is the following:

$$\begin{aligned}
\beta' &= \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\theta n(\bar{x} - \mu_0)^2}{2((\theta + n)} \\
&= 0.15 + \frac{1}{2}(89.33) + \frac{(.002)(6)(67.333 - 66)^2}{2(6 + 0.002)} \\
&= 0.15 + 44.665 + 0.00178 \\
&= 44.817
\end{aligned}$$

- Therefore, it is a gamma with parameters 3.46 and 44.817.

Example 4: Low Info Prior

Please note that with this low information prior, the data dominates the parameters of the posterior distribution.

“No”- Info Prior

- By looking at the above calculations for the low information prior, we can see that we can get increasingly vague priors by letting the prior parameters α, β, and θ get smaller and smaller.
- So, in the limit, this would suggest setting α, β, and θ to zero to get a type of noninformative prior.
- One drawback of this prior is that it is not a proper probability distribution.

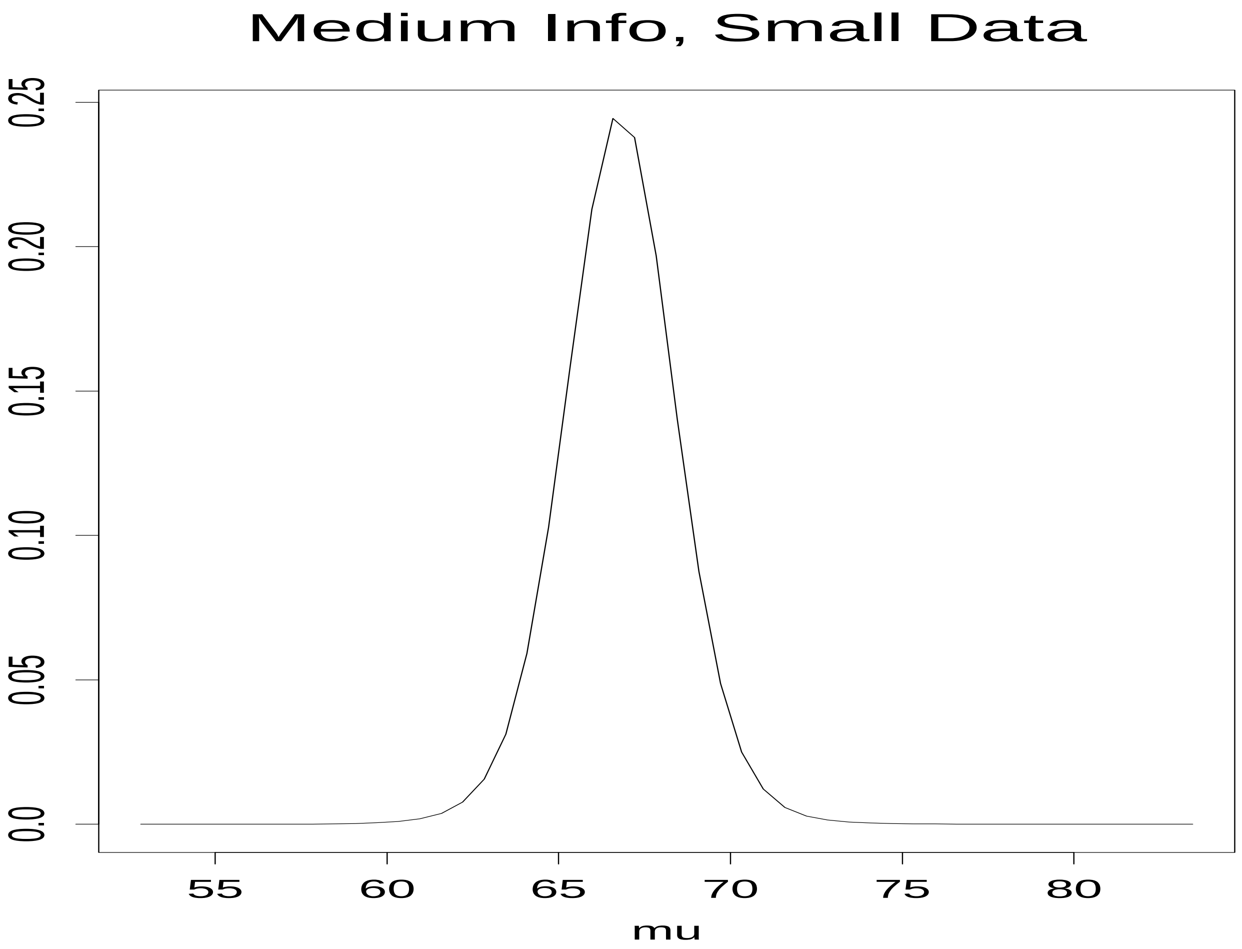
“No”- Info Prior

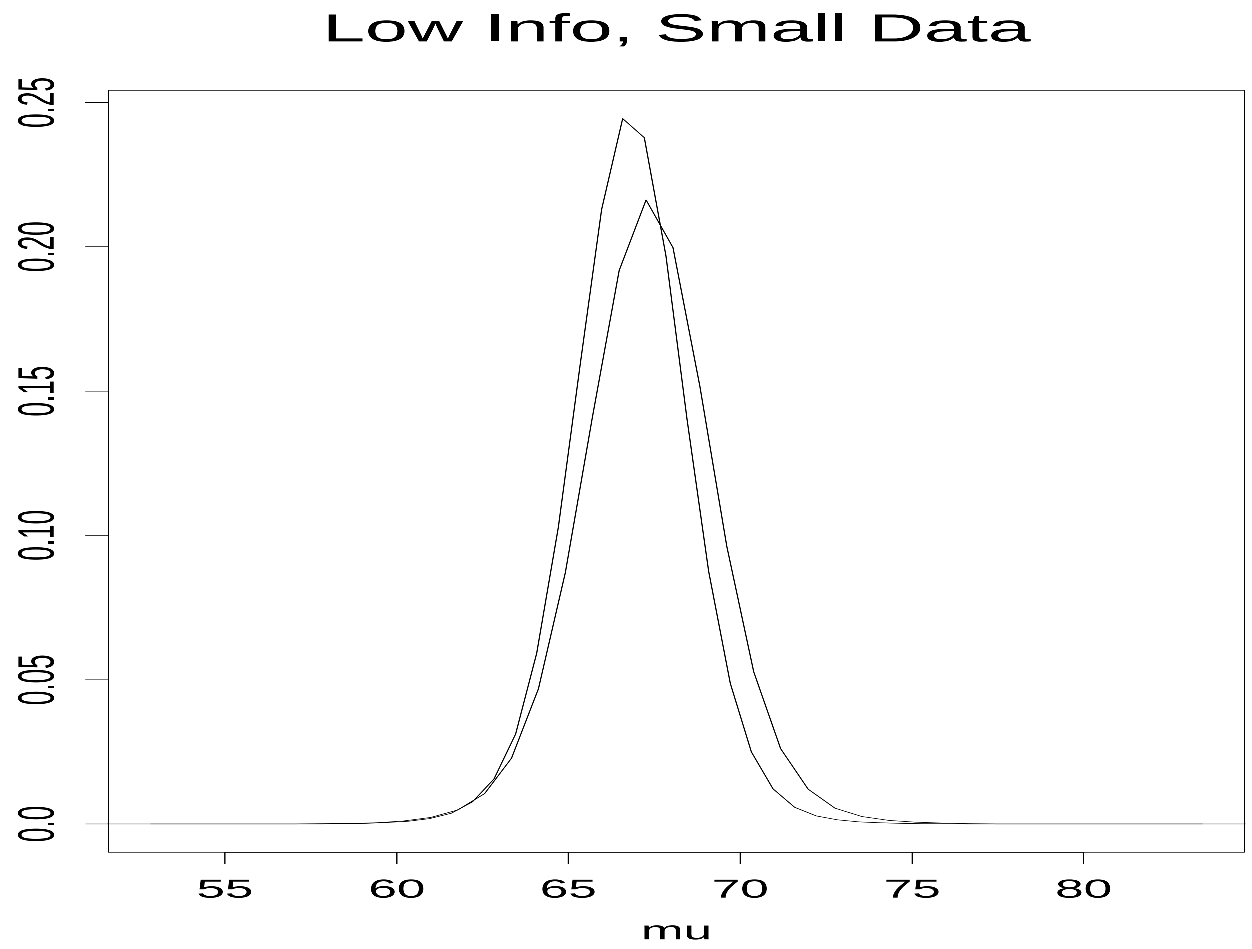
(continued)

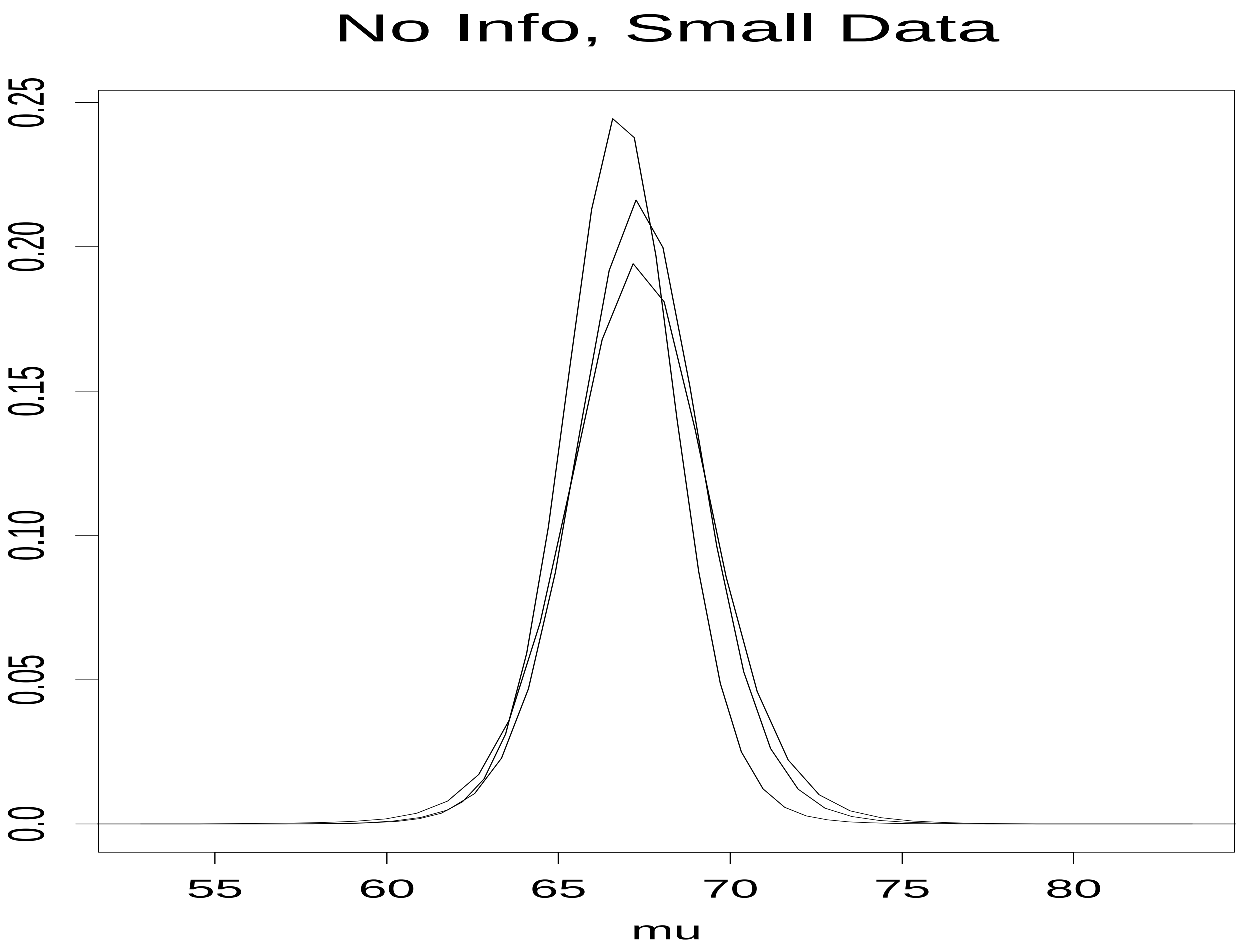
- Using this prior results in the conditional posterior distribution of μ being a normal distribution with mean $\bar{x}$ and precision $n\tau$.
- The marginal posterior distribution of τ is a gamma distribution with $\alpha' = n/2$ and $\beta' = \sum_{i=1}^{n}(x_i - \bar{x})^2/2$.
- The posterior mean of $1/\tau$ is $\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$.
- The marginal distribution of μ is a t-distribution with n degrees of freedom.

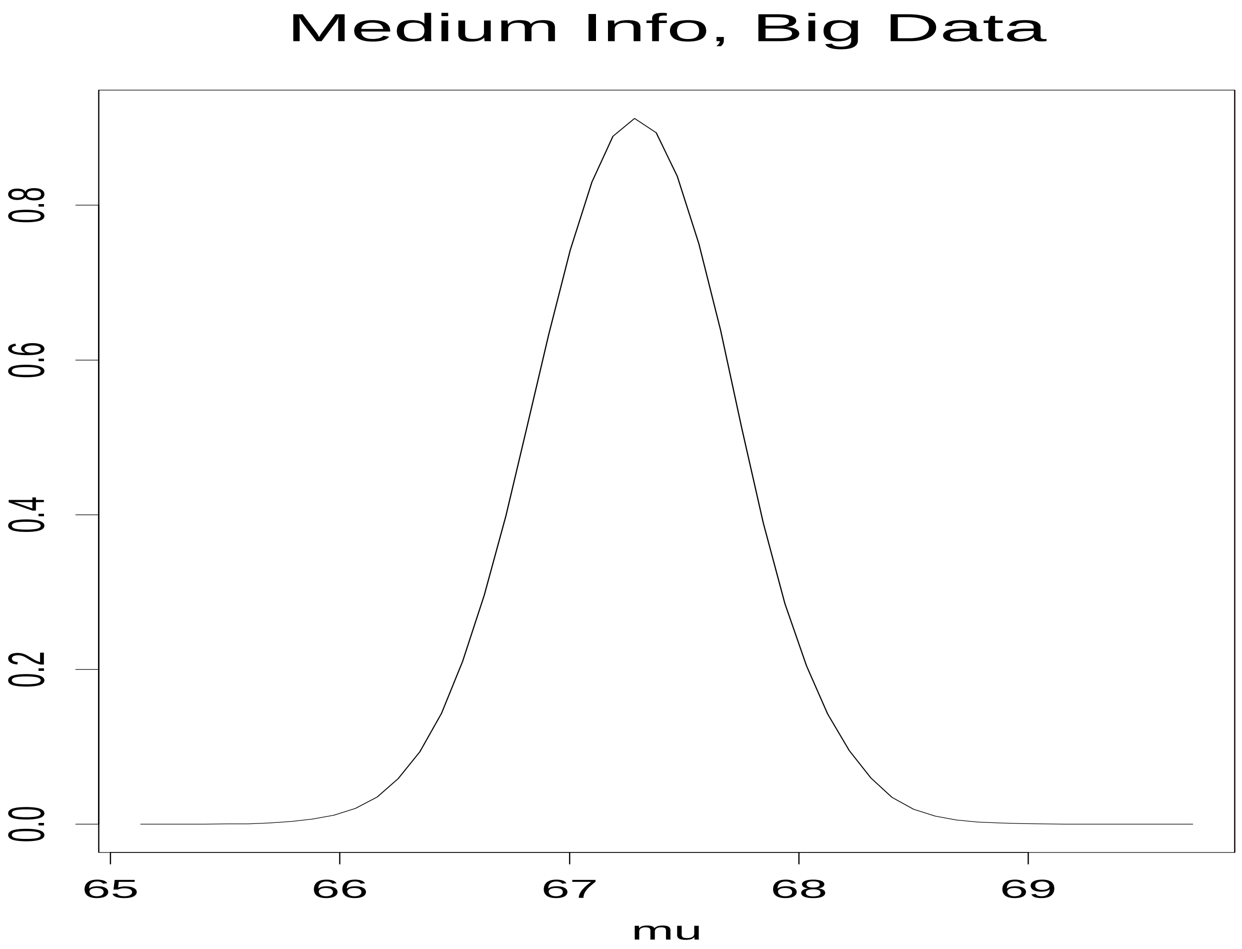
Comparing the Priors

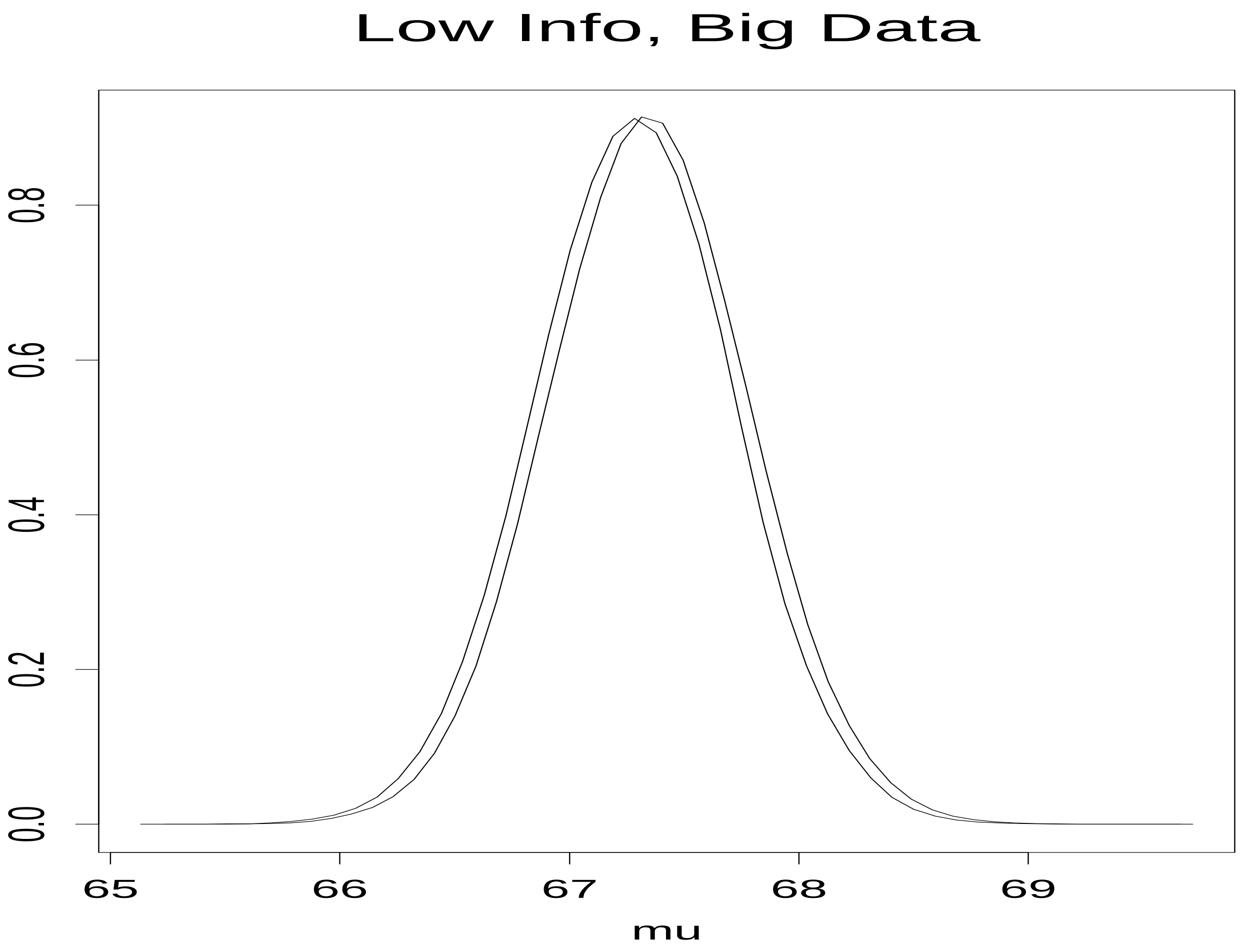
- The next series of plots contain the marginal distributions of μ and $1/\sqrt{\tau}$ and their joint distribution.
- These plots will be first for the small data of 6 subjects and then for the big data set with 71 subjects.
- For each data set, first the posterior using the first prior (the “medium info” prior) is shown, then for the prior with low prior information (the “low info” prior), and then for the improper prior where α, β, and θ all take the value zero (the “no info” prior).

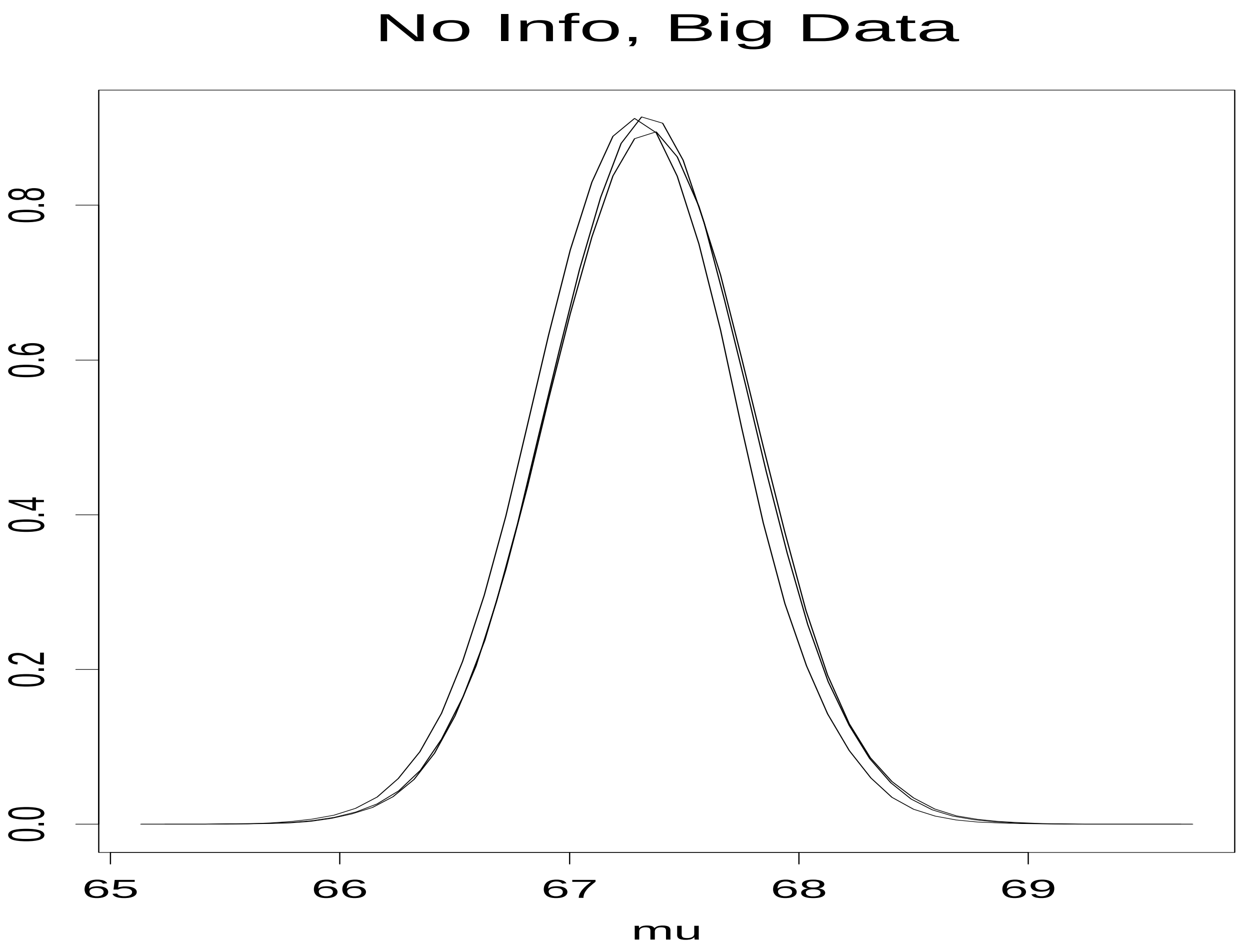

Medium Info, Small Data
0.0
0.05
0.10
0.15
0.20
0.25
55
60
65
70
75
80
mu

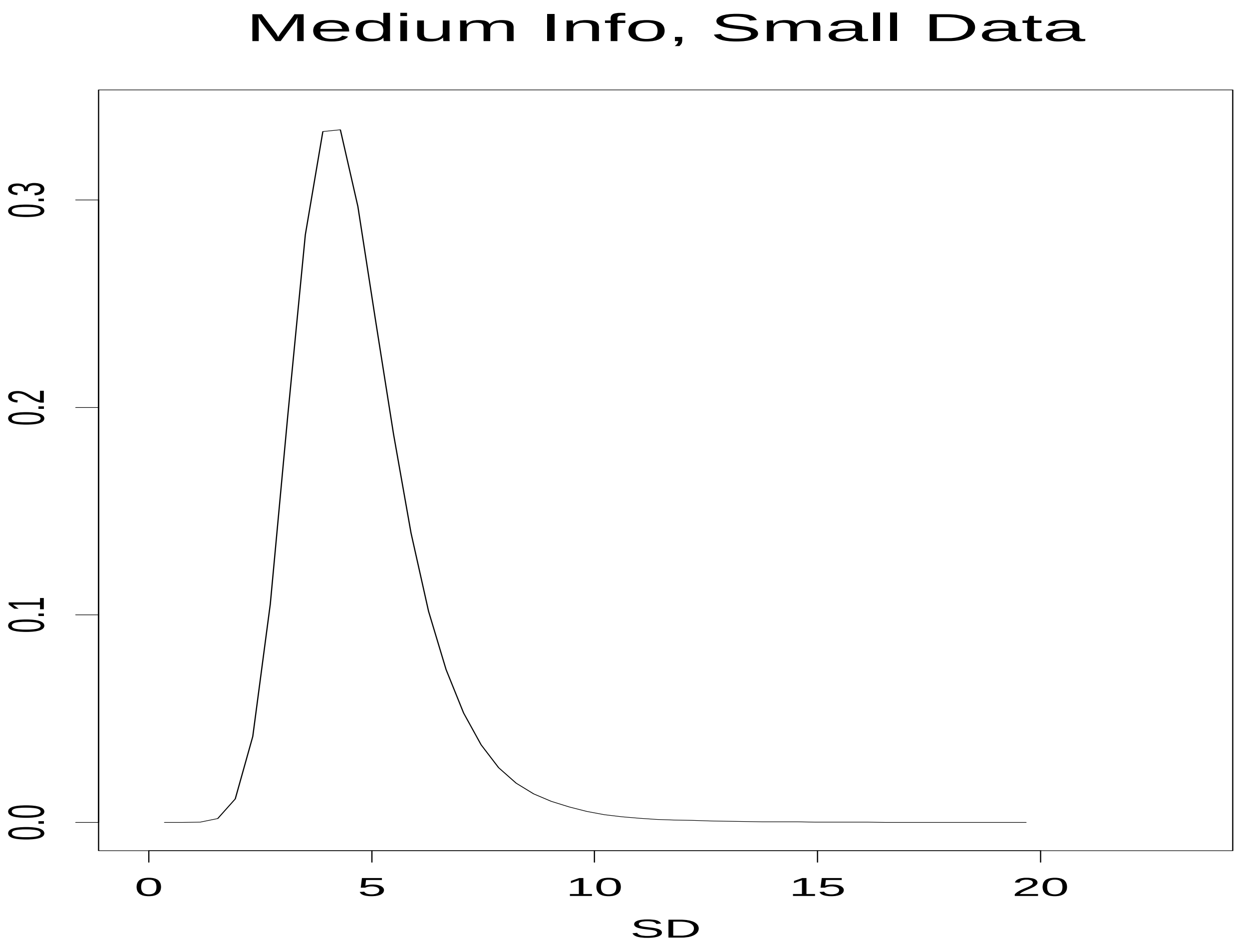

Low Info, Small Data
0.0
0.05
0.10
0.15
0.20
0.25
55
60
65
70
75
80
mu

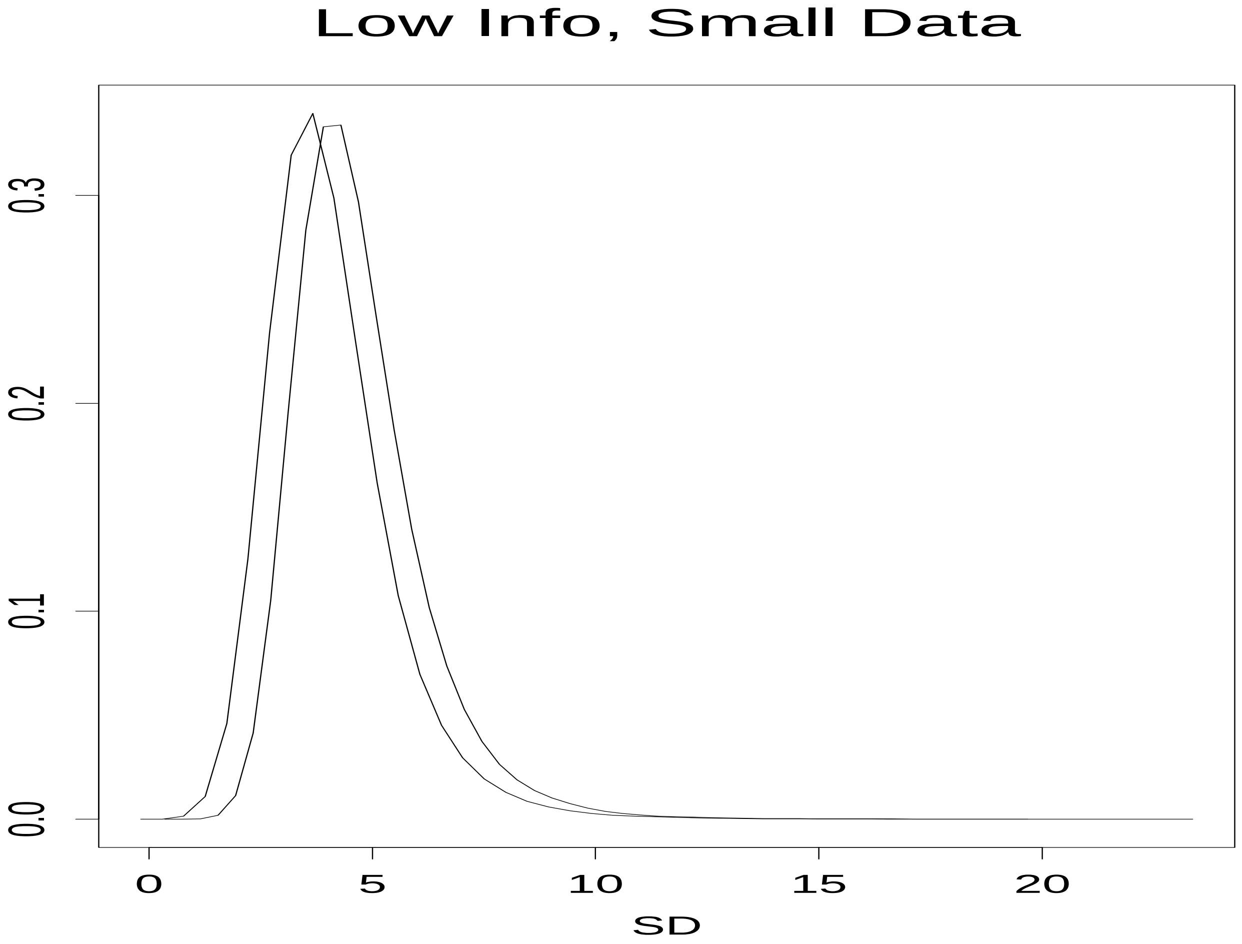

No Info, Small Data
0.0
0.05
0.10
0.15
0.20
0.25
55
60
65
70
75
80
mu

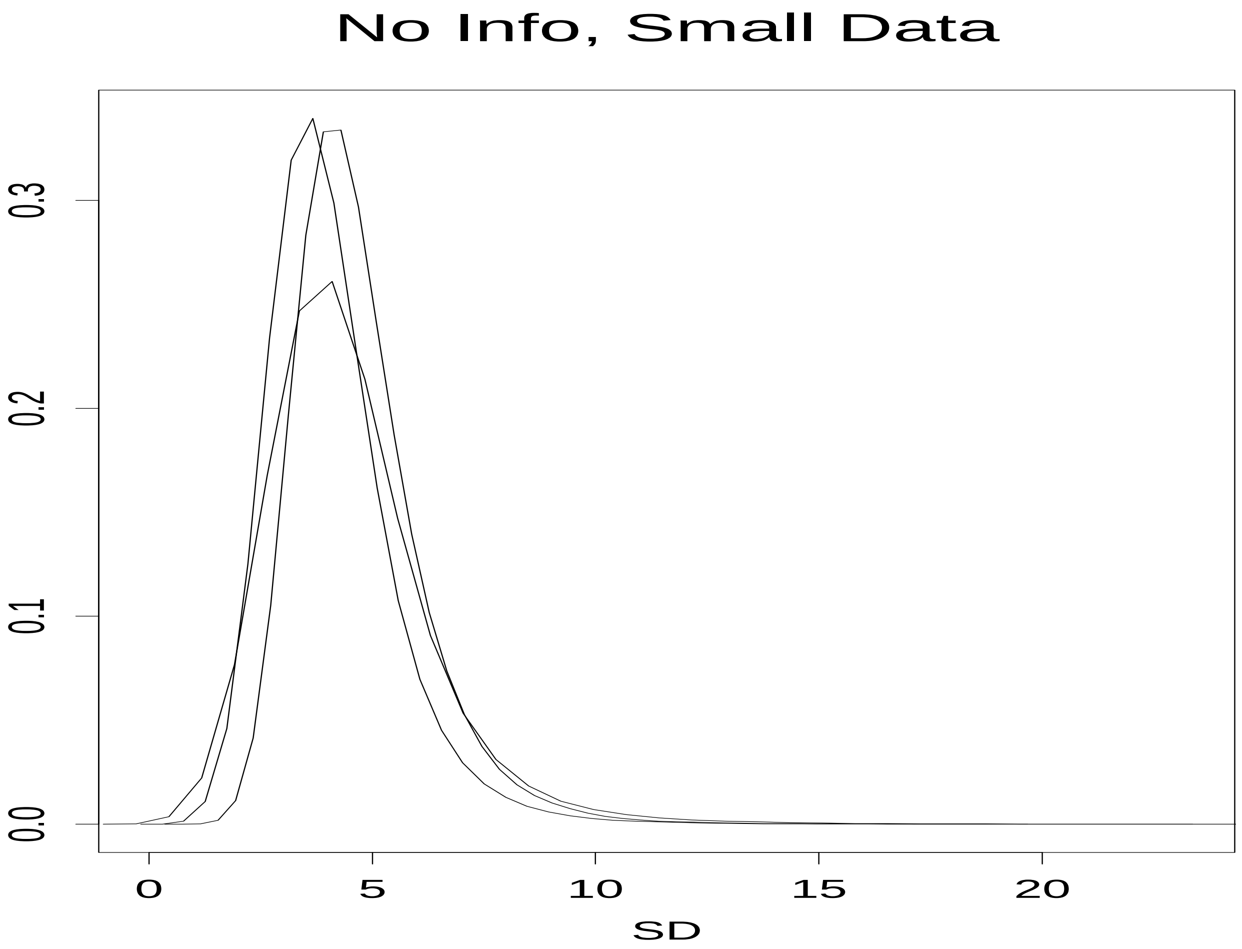

Medium Info, Big Data
0.0
0.2
0.4
0.6
0.8
65
66
67
68
69
mu

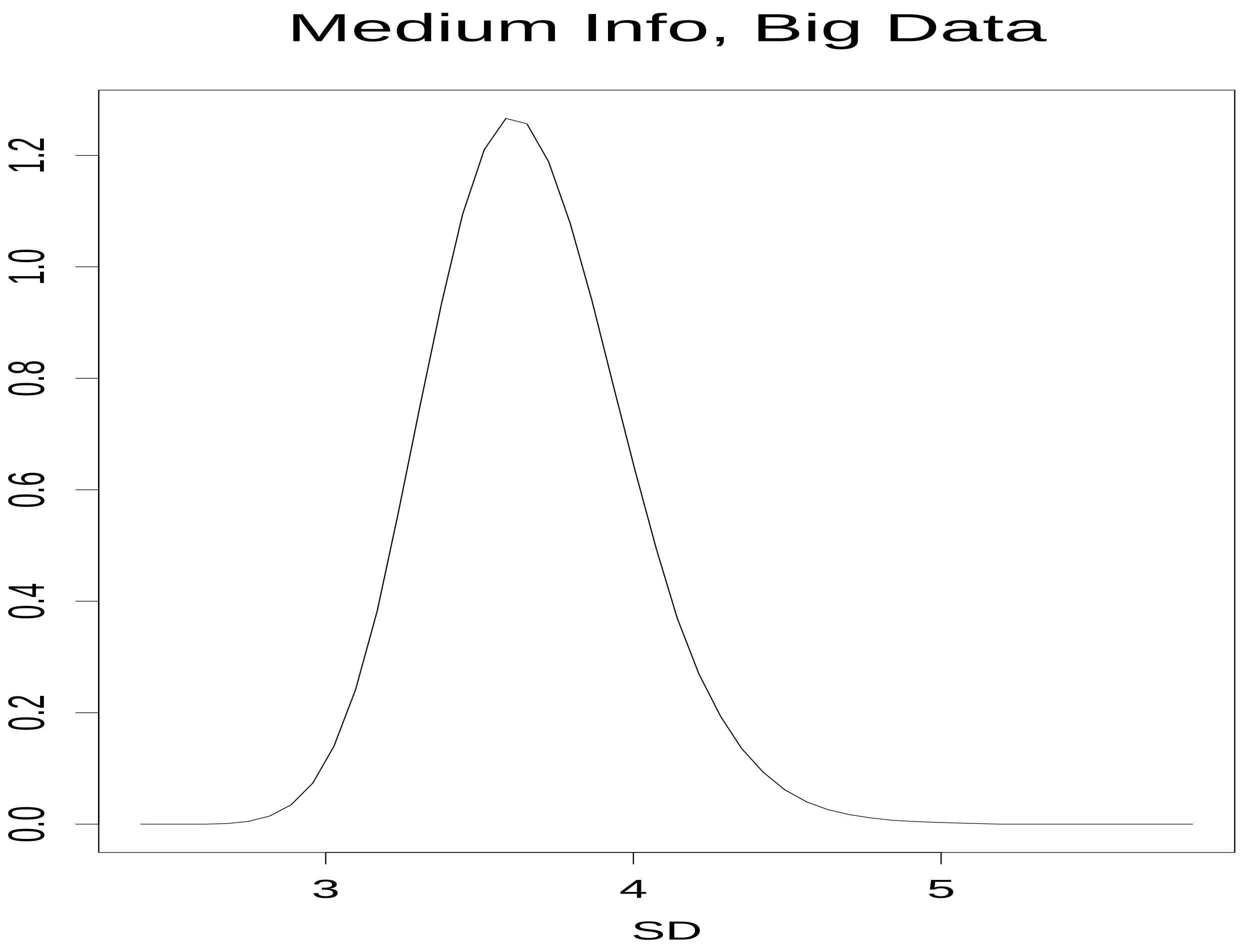

Low Info, Big Data
0.0
0.2
0.4
0.6
0.8
65
66
67
68
69
mu

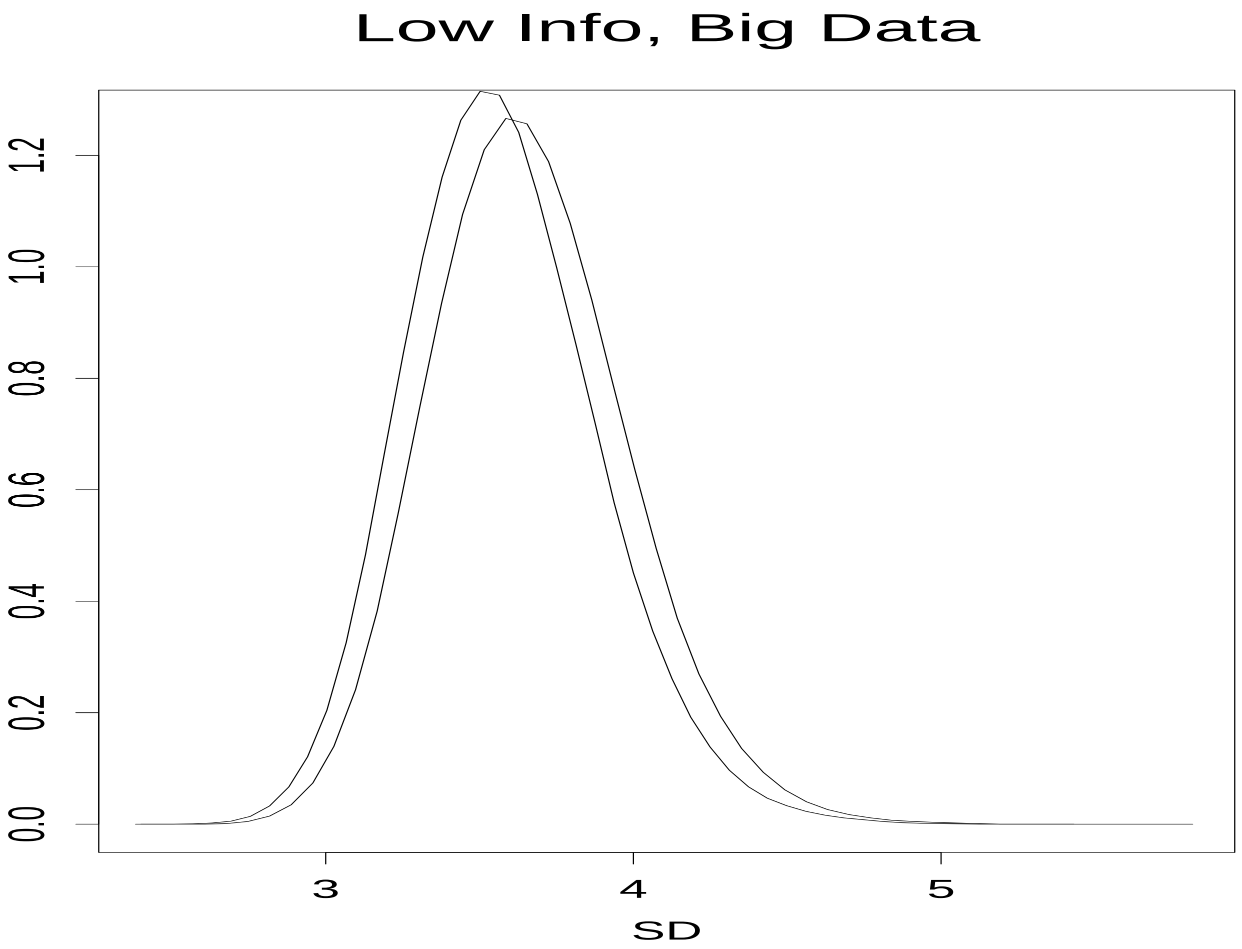

No Info, Big Data
0.0
0.2
0.4
0.6
0.8
65
66
67
68
69
mu


Medium Info, Small Data
0.0
0.1
0.2
0.3
0
5
10
15
20
SD

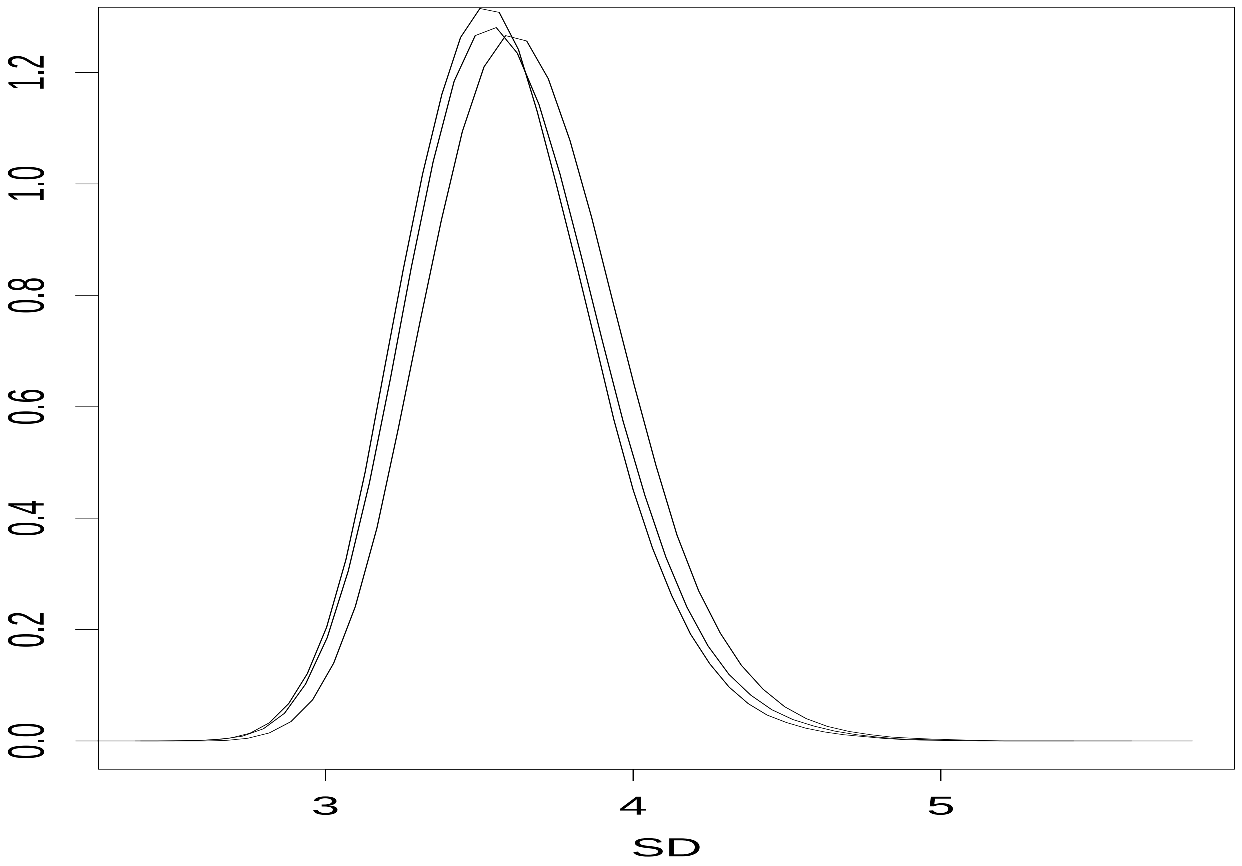

Low Info, Small Data
0.0
0.1
0.2
0.3
0
5
10
15
20
SD


No Info, Small Data
0.0
0.1
0.2
0.3
0
5
10
15
20
SD

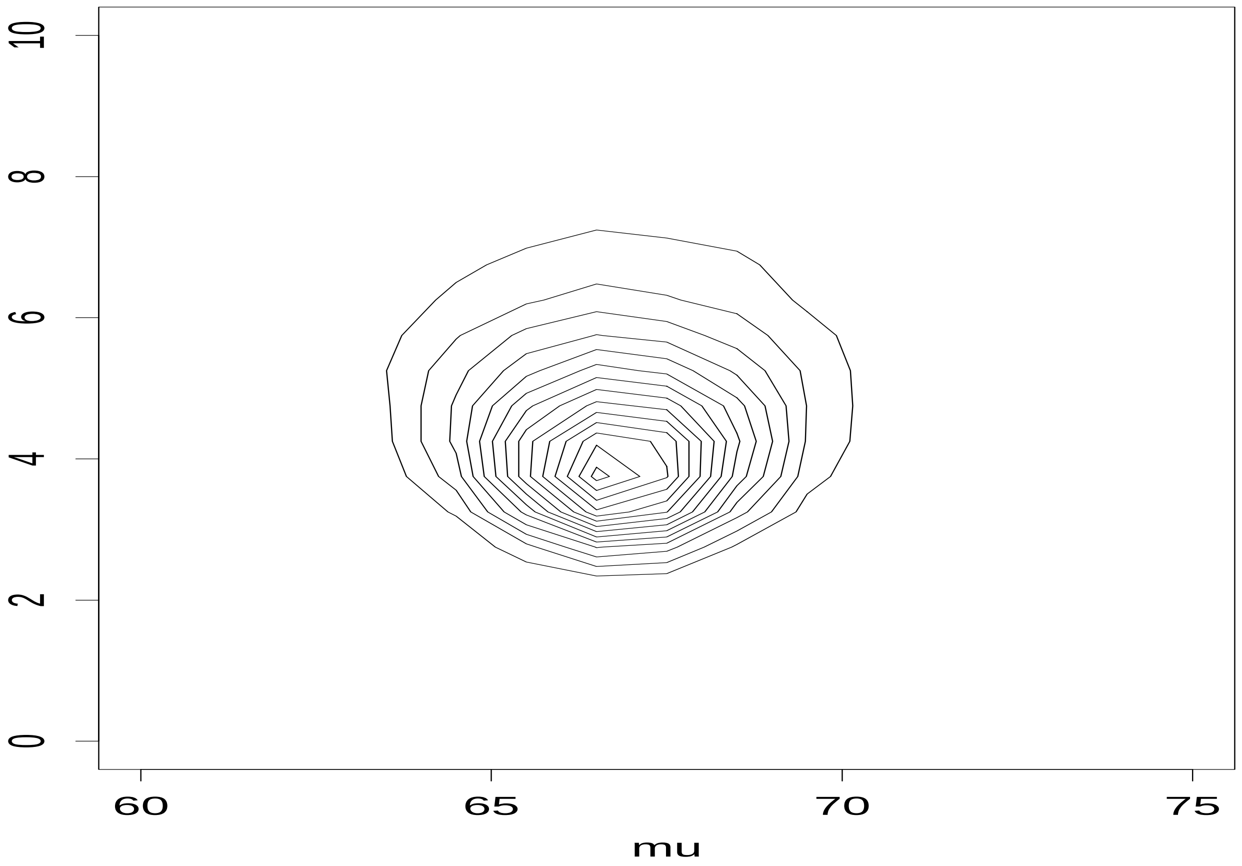

Medium Info, Big Data
0.0
0.2
0.4
0.6
0.8
1.0
1.2
3
4
5
SD


Low Info, Big Data
0.0
0.2
0.4
0.6
0.8
1.0
1.2
3
4
5
SD

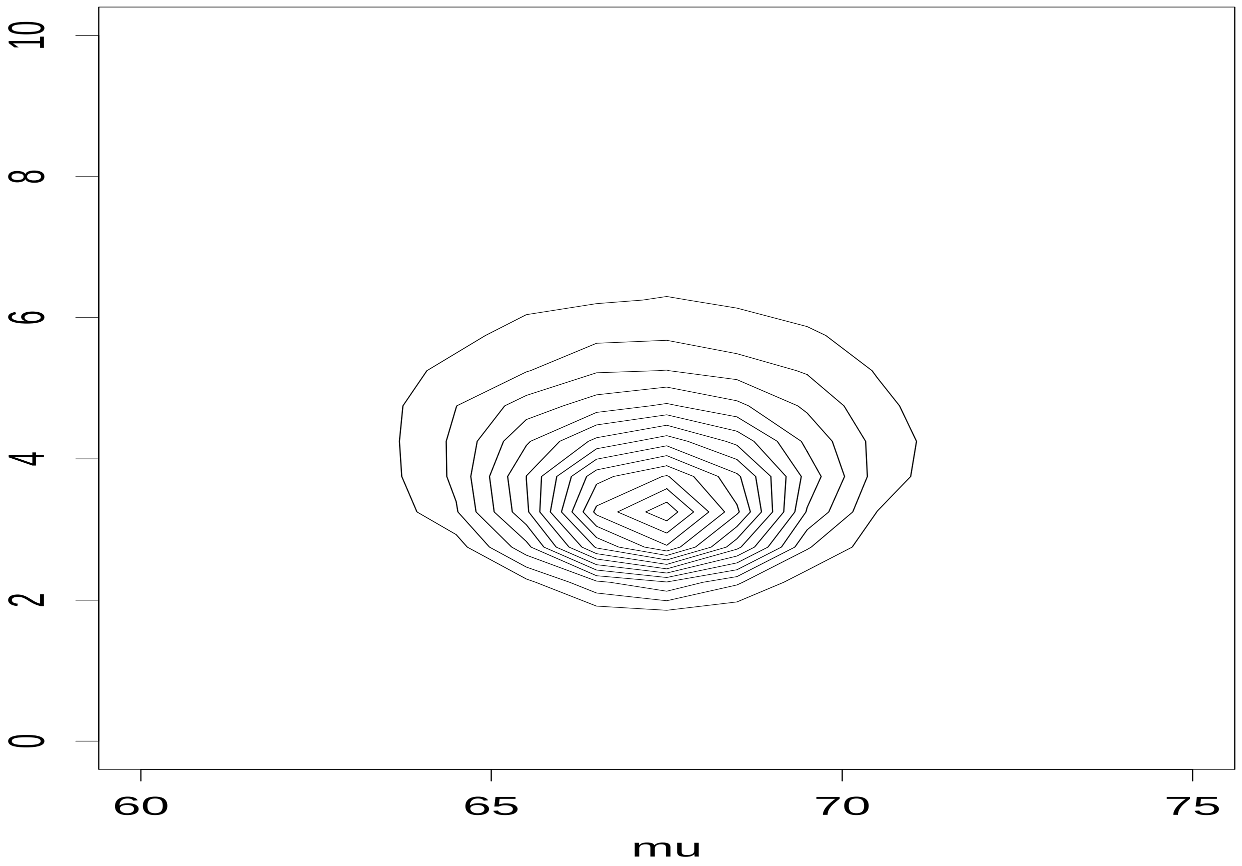
No Info, Big Data



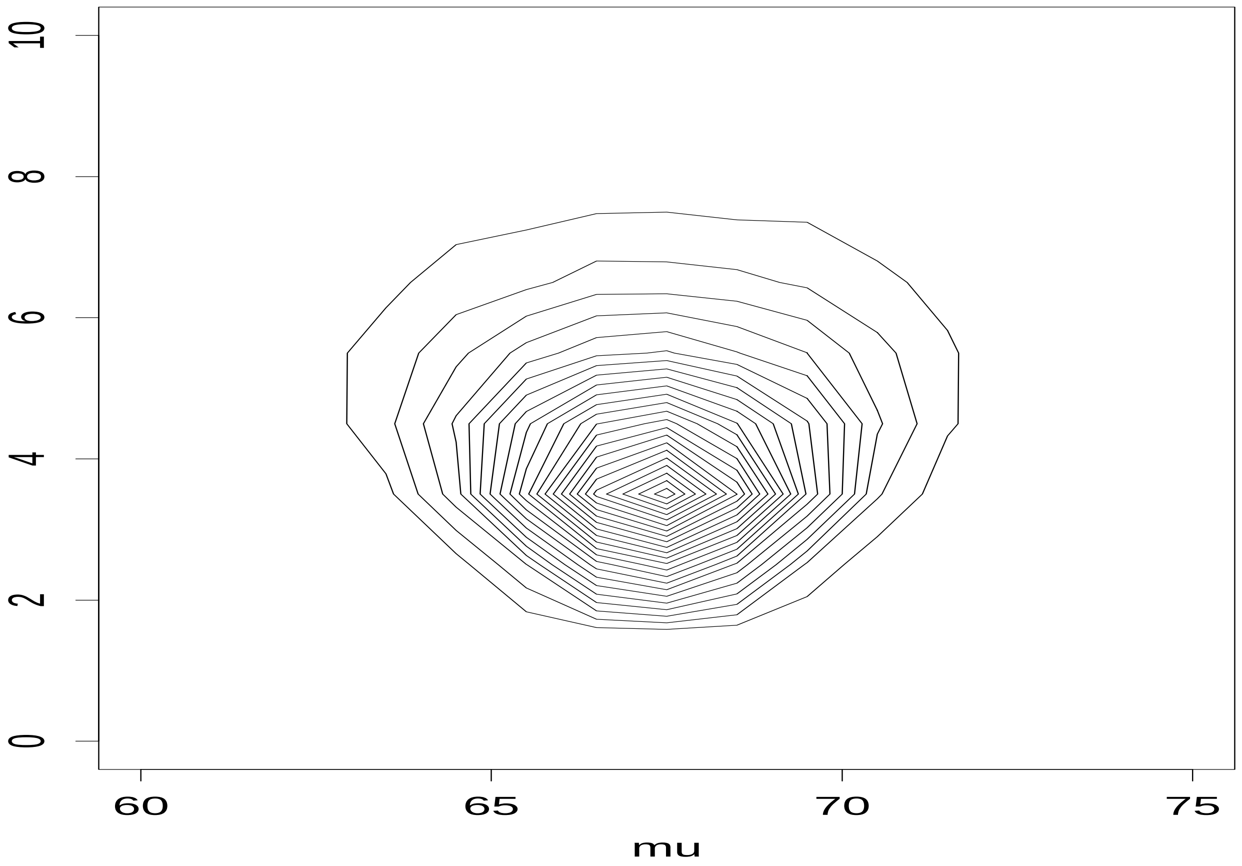
Medium Info, Small Data

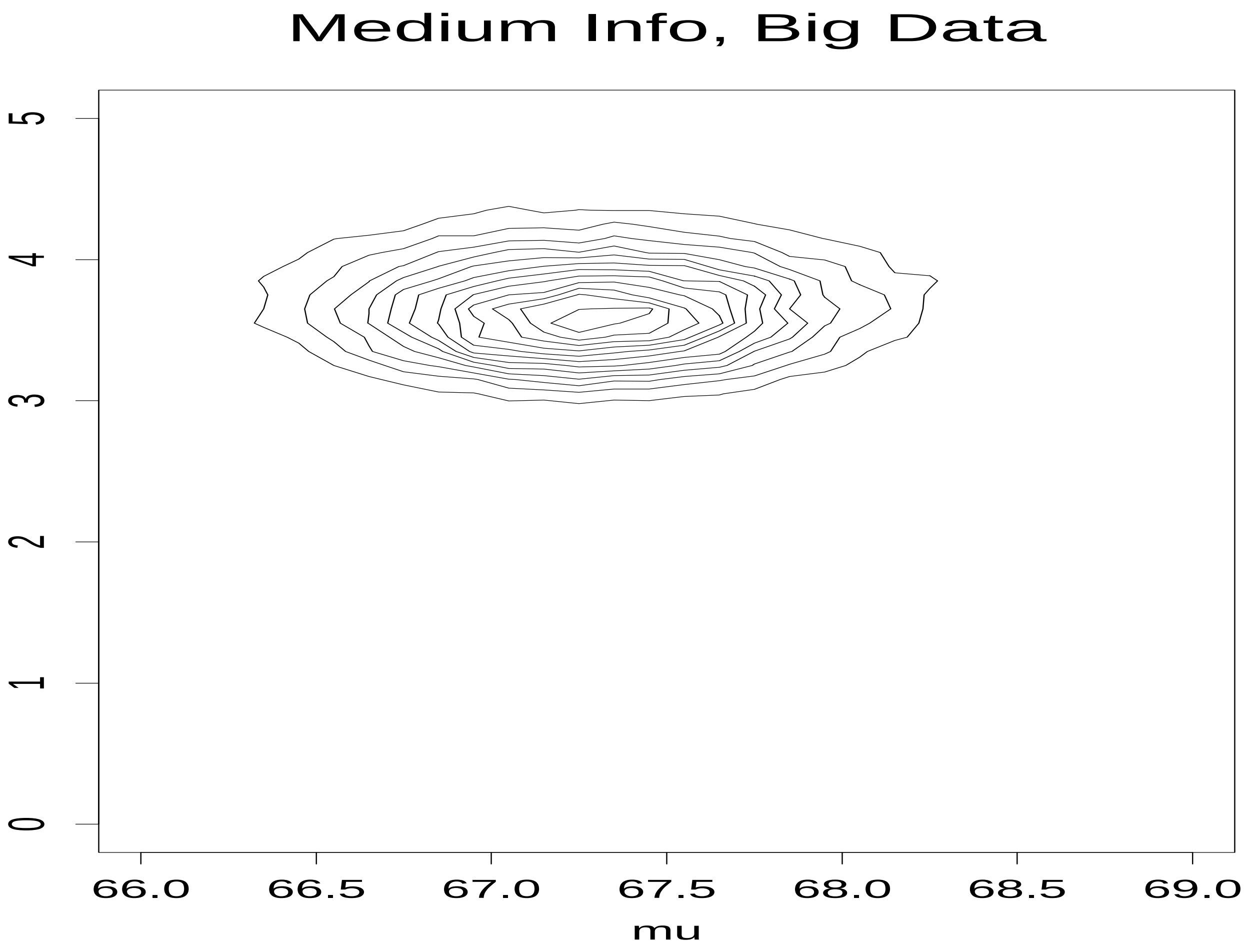


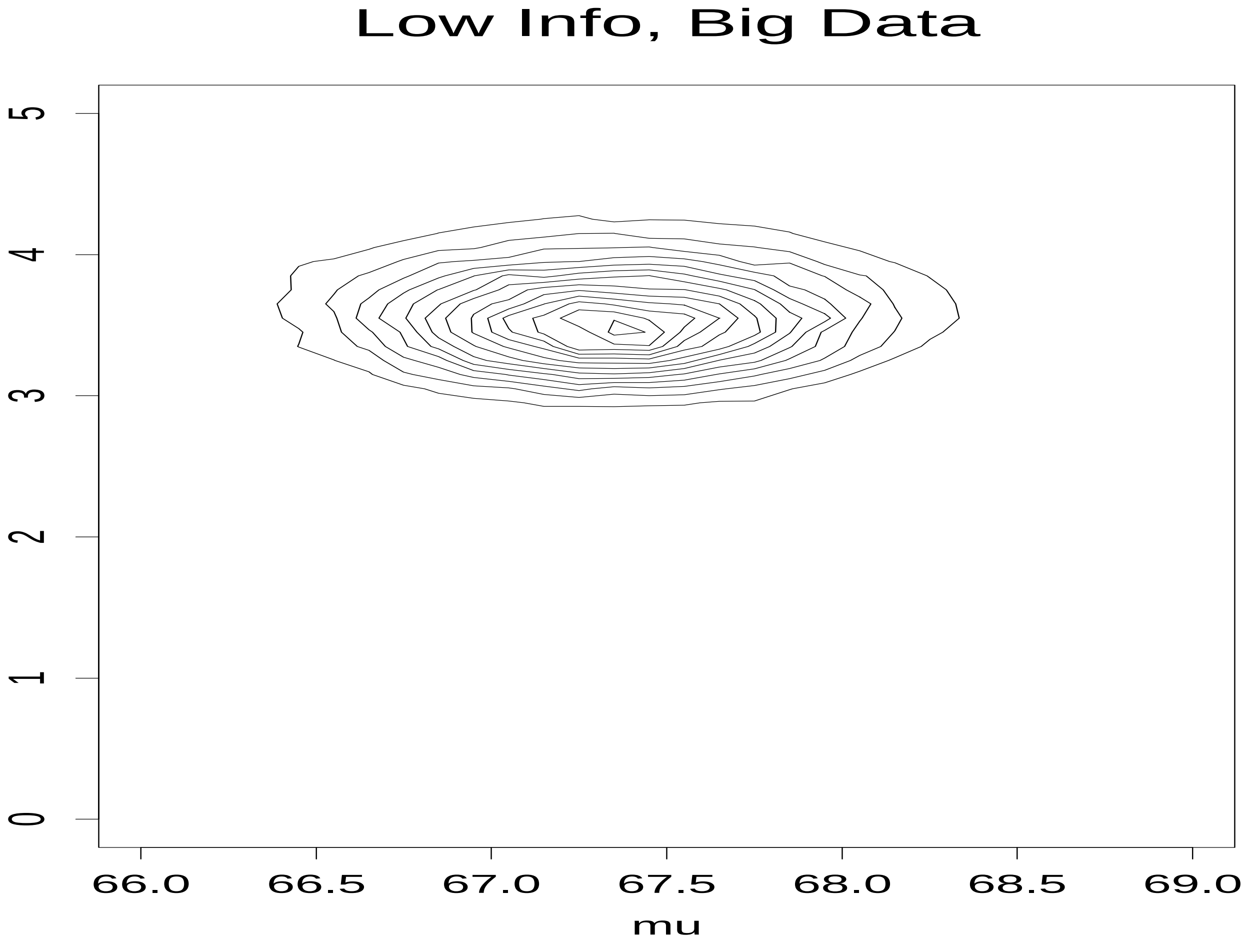
Low Info, Small Data

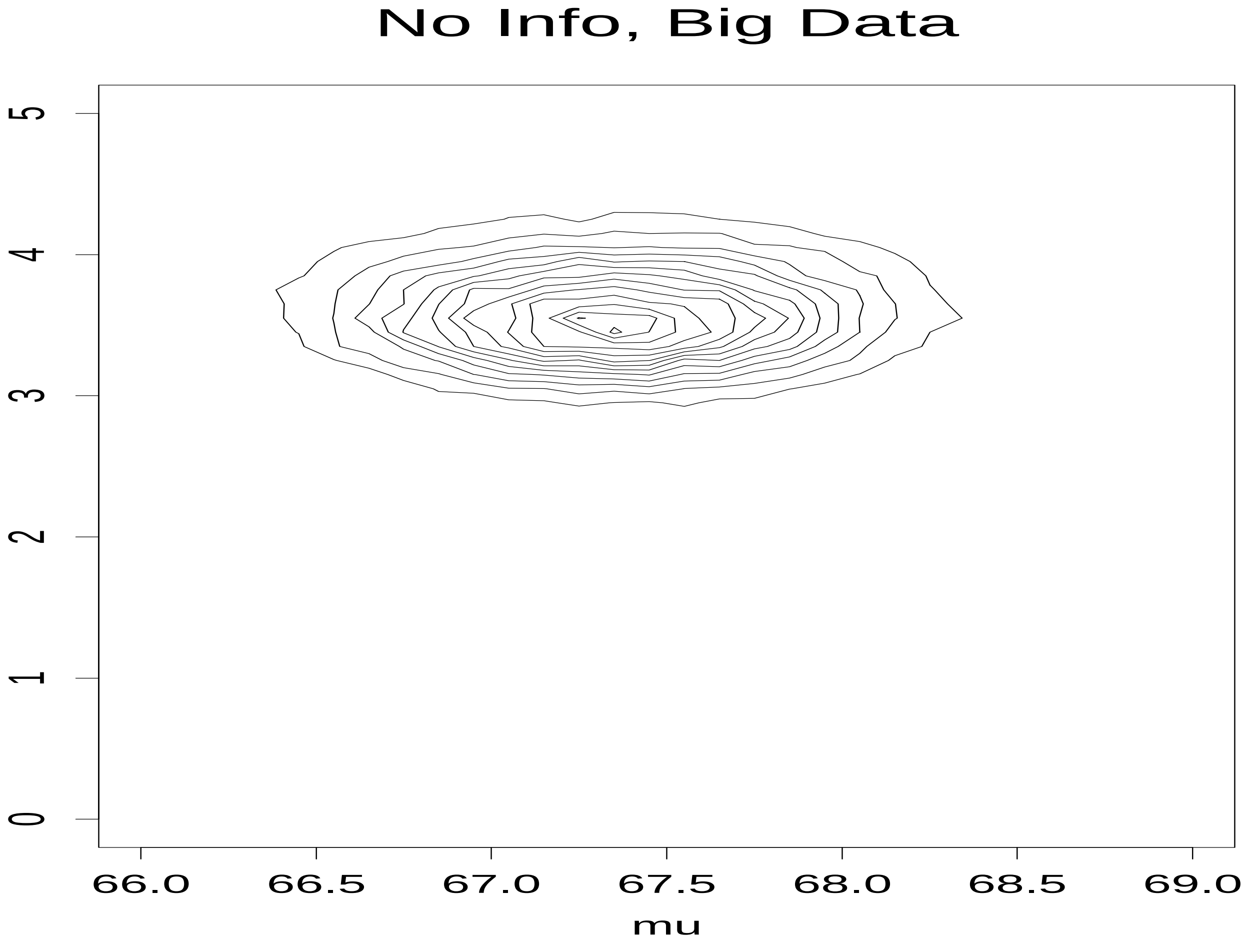


No Info, Small Data




Medium Info, Big Data
0
1
2
3
4
5
66.0
66.5
67.0
67.5
68.0
68.5
69.0
mu


Low Info, Big Data
0
1
2
3
4
5
66.0
66.5
67.0
67.5
68.0
68.5
69.0
mu


No Info, Big Data
0
1
2
3
4
5
66.0
66.5
67.0
67.5
68.0
68.5
69.0
mu

Example 4: Comments

- Calculations: The formulae for calculating the posterior distributions from the conjugate priors are demonstrated.
- Simulations: the use of computer simulation is also demonstrated. Computer simulation allows one to easily calculate certain values such as the distribution function and expected values of transformations of random variables which might have been difficult or time consuming to obtain by mathematical derivations.
- Priors: This example demonstrate how to select several different types of priors. This included the elicitation of a subjective prior, a vague prior and an improper noninformative prior.
- The effect of the sample size on the posterior distribution is also demonstrated here.

Are Subjective Priors Useful?

With the use of non-informative and vague priors, it might seem as if an informative, subjective prior is not desired. However, the next example shows some of the value of the prior.

Example 5: Blood Pressure

- In a certain population[†], it is know that the distribution of diastolic blood pressure (BP) is approximately normal with a mean of 85 and a standard deviation of 13 (so the precision is 0.006).
- Individual blood pressure varies for a person over the course of a day and over time. This variation follows a normal distribution with a standard deviation of 7 (so the precision is 0.020).

[†] The numbers from this example are from or approximated from Ingelfinger, JA, Mosteller, F, Thibodeau, LA, Ware, JH, *Biostatistics in Clinical Medicine*, (New York: MacMillan Publishing Co., Inc, 1983), pg 88 and pg 92 and references there in.

Example 5: Blood Pressure

(continue)

- Consider the case where someone is measured once and has a reading of 100 for the diastolic BP. What is the estimate of this person's "average" diastolic BP?
- One estimate of course is to estimate that the average equals the only data point that we have and estimate that the average is 100. However, using the above information, the Bayesian posterior mean would be 96.5.

Example 5: Blood Pressure

(continue)

- Let the "average" BP be the parameter μ. From the above information, the prior distribution for μ is a normal distribution with mean 85 and precision .006. Also, a single measured value, X, follows a normal distribution with mean μ and precision .020.

- So, the goal is to find the posterior mean of μ. Following the formula already given in this talk:

$$\mu' = \frac{(.006)(85) + (.020)(100)}{.006 + .020} = 96.5$$

- Also, the posterior precision is 0.026.

Example 5: Blood Pressure

(continue)

- So, with only one measurement, the Bayesian estimate is lower than the observed value. Is this a good thing?
- To see what is going on, see figure 18. This figure shows the likelihood and prior density function for our example.

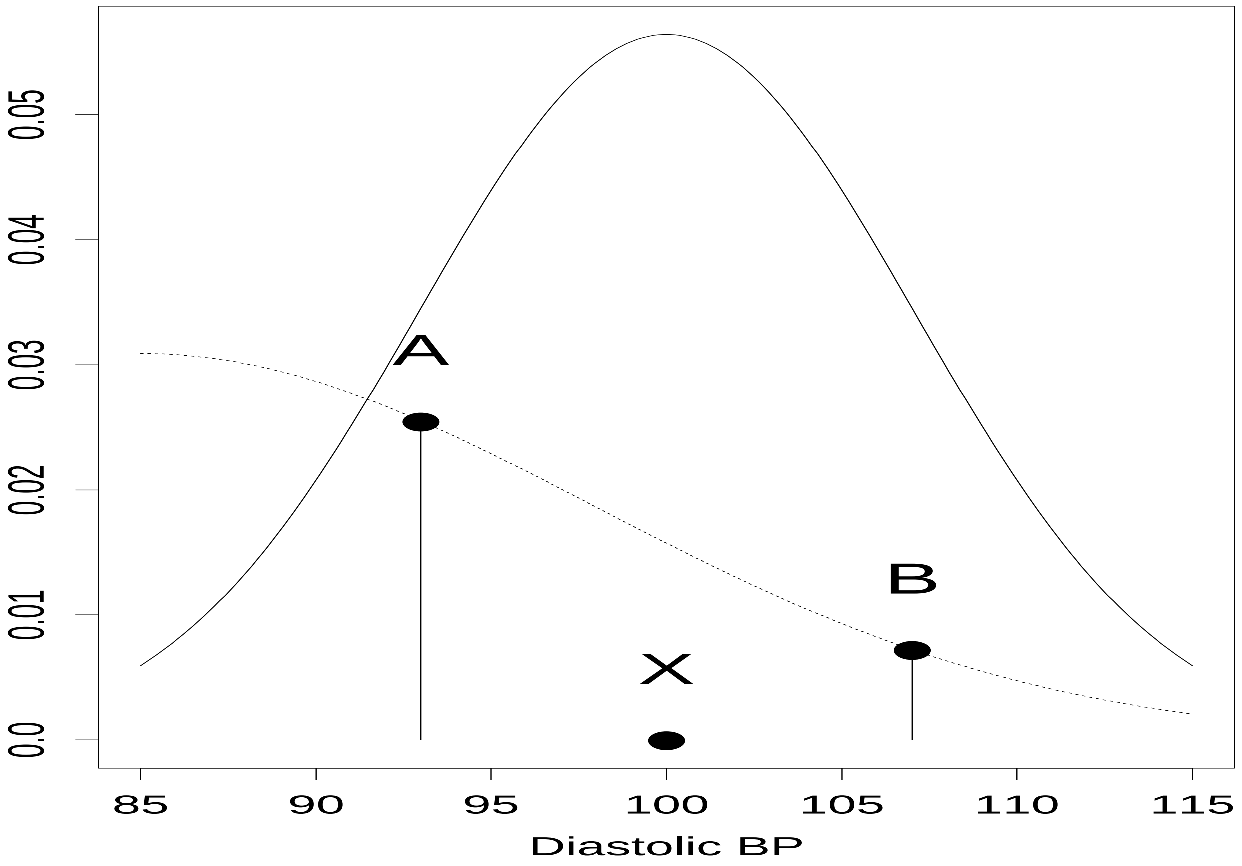


Figure 18: The likelihood and prior for diastolic BP for someone with a reading of 100. The solid line is the likelihood and the dotted line is the prior.

Example 5: Blood Pressure

(continue)

- Now, for the moment, suppose that the "true" value was either $X-\text{SD}$ or $X+SD$. According to the likelihood, either of these points are just as likely.
- Therefore, if one is only consider the information from the likelihood, then one would equally consider the value of 93 and 107 as the true value. Also, if one were to consider $X \pm 1.5\text{SD}$ one would have a similar conclusion.
- Therefore, by symmetry, one can easily see that the estimate would be 100 since all points which are equal distance from 100 has equal weight from the likelihood.

Example 5: Blood Pressure

(continue)

- Now let's consider the information from the prior. Point A in figure 18 is the value of the prior for 93 and point B is the value of the prior for 107. The point A is almost 3 times higher than the point B. So, there are 3 times more people who could have a BP of 93 than a BP of 107.
- That is, from the likelihood, there is an equal chance that it is a person with an average BP of 93 having a bad day (and having a higher BP) then it is that a person with an average BP of 107 having a good day (and seeing a lower BP than average). However, there are three times more people with an average BP of 93 than with an average of 107.

Example 5: Blood Pressure

(continue)

- Similarly, for all the pairs which are equally distant from 100, there are more people who are below 100 than above. That is, there are more people who are 0.9 SD below than 0.9 SD above or .8 SD, or .7 SD, etc.
- So, it make sense that one would give an estimated value which is lower than the observed.
- Note: this effect, that the estimated value is between the observed value and the prior mean. This effect is sometimes referred to as "regression to the mean" or a "shrinkage estimator".

C: Normal Models - Summary

- The normal distribution model is an extremely important in applied statistics, so it is important to understand how this model works.
- Methods to characterize one's belief in the values of the parameters for this model has been discussed.
- Also, the analytical formulae have been derived in order to obtained the posterior distribution of the parameters of this model when the prior distributions for these parameters belong to a certain family of distributions.
- Please note that this is a two parameter model and this adds a degree of difficulty for this model. Please also note, that applied statistician want to use much more complex versions of this model and require tools for these more complex model.

D: Modelling Prior Belief

Outline:

- Some methods for getting subjective priors.
- Some checks on modelling subjective beliefs.
- Some different types of priors.

Getting Subjective Priors

Below are several different ideas for getting subjective prior probabilities. Each of these methods is discussed in more detailed in the proceeding slides. It is a good idea to try a combination of these ideas and see if they produce approximately similar priors.[†]

- Histogram method
- Relative likelihood approach
- Match a given density function

[†] See for example J.O. Berger, *Statistical Decision theory and Bayesian Analysis, 2nd ed.*, (New York: Springer-Verlag, 1980).

Priors: Histogram Method

- Break the parameter space into intervals.
- Then assign heights to the different intervals.

Priors: Relative Likelihood Approach

- Sketch out the “likelihood” of the values of the parameter.
- You might decide to pick out the most likely and least likely points. Pick out on point of the parameter to be the reference point. Then, give the most likely point a value which represents how much more likely this is from the reference point. Add a few more points as needed.
- Then, you would sketch out the values between the points that were chosen. When connecting points, you might consider connecting the points with a straight line or maybe some other type of curve.

Priors: Relative Likelihood Approach

(continued)

- If the parameter space is unbounded, then it is necessary to have the tails fall off sufficiently fast so that it will be proper or if means and variances are desired. For example, the tails might fall away like a normal distribution or an exponential distribution.
- Also, it can be tricky having the desired ratio of mass in the central part of the prior versus out in the tails.

Priors: Match a Given Density Function

- If the prior is assumed to be a member of a parametric family of distributions, then the trick is to match properties of the distribution with our beliefs.
- For example, we might figure out what we believe is our average parameter value and match this to the mean parametric family.
- Other good choose prior so the the mean of the distribution matches our average prior belief. Also, matching percentiles like the 95-percentile is useful. It is sometimes easier to specify values like the point where you believe that the probability of being less than this point is 20%.

Priors: Match a Given Density Function

(continued)

- Moments like the variance and the skewness are sometimes more problematic because these values depend heavily on the thinness or thickness of the tails of the distribution and it is hard to see the effect of different tail statistics.
- This technique is demonstrated in Example 3 for choosing the prior on θ and Example 4 for choosing the prior on τ.

Some Checks for Subjective Priors[†]

1. Look at the "prior sample size". For example, for the beta prior this is the value $(a + b)$. Changes are not as big when this term is large. The problem with this interpretation is that the prior data/experiments might be different than the present situation (problem with historical controls or with using data from a similar product, etc). So, sometimes, one might consider "downweighting" this total. The interpretation is that we are not quite so certain that the present experiment will be similar to the prior experiment.

[†] note: the suggestions for getting subjective priors and for checking them follows from D.B. Berry, *Statistics: A Bayesian Prospective*, (Belmont, CA: Wadsworth Publishing, 1996).

Some Checks for Subjective Priors

(continued)

2. Ask the assessor to give the probability of several different sets of events. For example, ask for the probability of different tail areas. (What is the probability that of being greater than zero? What is the probability of seeing a person greater than than 6 feet? Of a person being less than 4 feet? Etc...)

3. When in doubt, be more open-minded. That is, down weight the "prior sample size". This represents a prior which is less informative.

Some Types of Priors

Note, the list below does not divide the different types of priors into non-overlapping groups of priors.

- Conjugate priors: Refers to a family of distributions such that if the prior is a member of the family, then the posterior is also a member of the family for all possible sampled values. This property for a family is specific to a the form of the likelihood. This family of distributions forms a closed system.

Some Types of Priors

(continued)

- Locally Uniform Priors: This is a prior which is roughly uniform in the area where the likelihood function has most of its weight.
  - Therefore, the posterior distribution is approximately proportional to the likelihood function.
  - An example of such a prior might be to have the prior distribution for the mean of normal distribution be a uniform with an extremely wide range. (So, maybe, have the prior distribution for the mean of the population of human heights be uniform on the interval from 0 to 12 feet.)

Some Types of Priors

(continued)

- Reference priors (eg. Jeffery's prior): A prior which is created by some underlying principle. It is meant to be used as a standard prior. Others can then compare their priors versus the reference prior and then adjust their posterior distribution accordingly.

Some Types of Priors

(continue)

- Noninformative priors: These are priors which are, in some sense, dominated by the likelihood function. Since they are dominated, they don't seem to provide much information in the creation of the posterior distribution. Note: usually noninformative priors are not really, "completely" noninformative. Sometimes they can create unintended consequences. Don't use them blindly. Also, a non-informative prior is not necessarily unique. There might be several different plausible "low-informative" priors which someone might call a noninformative prior.

Some Types of Priors

(continue)

- Improper priors: A prior distribution which is not a proper distribution. Usually this is because the area under the curve for this density is infinity. An example of this is the prior which puts equal weight on the entire real line.
- Informed priors: A prior which models an opinion.

Some Types of Priors

(continue)

- Skeptical/enthusiastic priors. This is a method allows different people to use informed priors and which allows these people to possible reach a consensus. For this method, two priors are created. One prior is made by someone who is very skeptical about some outcome and the other is fairly enthusiastic. These priors then represent the two opposite ends of the spectrum on what people tend to believe about something. If the skeptic and the enthusiast posterior distribution are fairly close, then both skeptic and enthusiast should basically agree on the results of the analyses.

Summary of Part 1

Some of the good things:

- The rules of probability can be used to quantify uncertainty/belief and the rule of probability can be used to update this belief.
- Conjugate priors can be used to make some of the calculations easier.
- Computer simulation methods can be used to calculate distributions and expected values where the mathematics might be difficult or time consuming.
- As the sample size increases, the likelihood begins to dominate the posterior.

Comments on the Part 1

Some of the problems which need to be addressed.

- The use of only conjugate priors restricts the mathematical formulation of one's belief.
- The methods presented so far are for simple models. As the models get more complex, then the computational complexity increases and getting exact analytical solutions maybe nearly impossible.
- This was basically the state of affairs of applied bayesian methods pre-1990. There were some complex mathematical approximations and numerical methods available, but even these methods would often be difficult for moderately complex models. Also, it usually required a fairly high level of statistical training to do any of these methods.