Every Day Bayes:
A brief tutorial on the Bayesian way

Mark J. Meyer

September 28, 2018

GEORGETOWN UNIVERSITY

Georgetown College

Department of Mathematics and Statistics

Introduction to Bayesian Thought

Let's begin with a population we would like to study:

Suppose we want to study blood pressure in this group

What do we do?

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

- How do we decide what summaries are important?

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

- How do we decide what summaries are important?
- How do we go about estimating those summaries?

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

- How do we decide what summaries are important?
- How do we go about estimating those summaries?
- How do we draw inference on our findings or make conclusions using statistical evidence?

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

- How do we decide what summaries are important?
- How do we go about estimating those summaries?
- How do we draw inference on our findings or make conclusions using statistical evidence?

Introduction to Bayesian Thought

We collect a sample (ideally random) and measure the blood pressure of the study participants

Next we take that data and begin to summarize it

This should lead us to a series of questions:

- How do we decide what summaries are important?
- How do we go about estimating those summaries?
- How do we draw inference on our findings or make conclusions using statistical evidence?

Question: What parameter should we focus on to summarize our data?

Introduction to Bayesian Thought

Introduction to Bayesian Thought

(I'm hoping you said average or mean)

Introduction to Bayesian Thought

(I'm hoping you said average or mean)

How do we estimate the parameter, call it θ, that represents the average or mean?

Introduction to Bayesian Thought

(I'm hoping you said average or mean)

How do we estimate the parameter, call it θ, that represents the average or mean?

(You probably said "the sample mean")

Introduction to Bayesian Thought

(I'm hoping you said average or mean)

How do we estimate the parameter, call it θ, that represents the average or mean?

(You probably said "the sample mean")

How would you conduct inference on θ?

Introduction to Bayesian Thought

(I'm hoping you said average or mean)

How do we estimate the parameter, call it θ, that represents the average or mean?

(You probably said "the sample mean")

How would you conduct inference on θ?

(You probably said "confidence interval, p-value, z-test")

Introduction to Bayesian Thought

Now think about *why* you said the sample mean, confidence interval, p-value, and z-test

Introduction to Bayesian Thought

Now think about *why* you said the sample mean, confidence interval, p-value, and z-test

This choice is ultimately a philosophical question but one we don't normally think of as philosophical

Introduction to Bayesian Thought

Now think about *why* you said the sample mean, confidence interval, p-value, and z-test

This choice is ultimately a philosophical question but one we don't normally think of as philosophical

In most statistics classes, we're taught one philosophy of statistical estimation and inference referred to as Frequentisit

Introduction to Bayesian Thought

Now think about *why* you said the sample mean, confidence interval, p-value, and z-test

This choice is ultimately a philosophical question but one we don't normally think of as philosophical

In most statistics classes, we're taught one philosophy of statistical estimation and inference referred to as Frequentisit

Today, I will attempt to introduce one alternative

Introduction to Bayesian Thought

SECTIONS HOME SEARCH

The New York Times

SCIENCE

The Odds, Continually Updated

By F. D. FLAM SEPT. 29, 2014

Introduction to Bayesian Thought

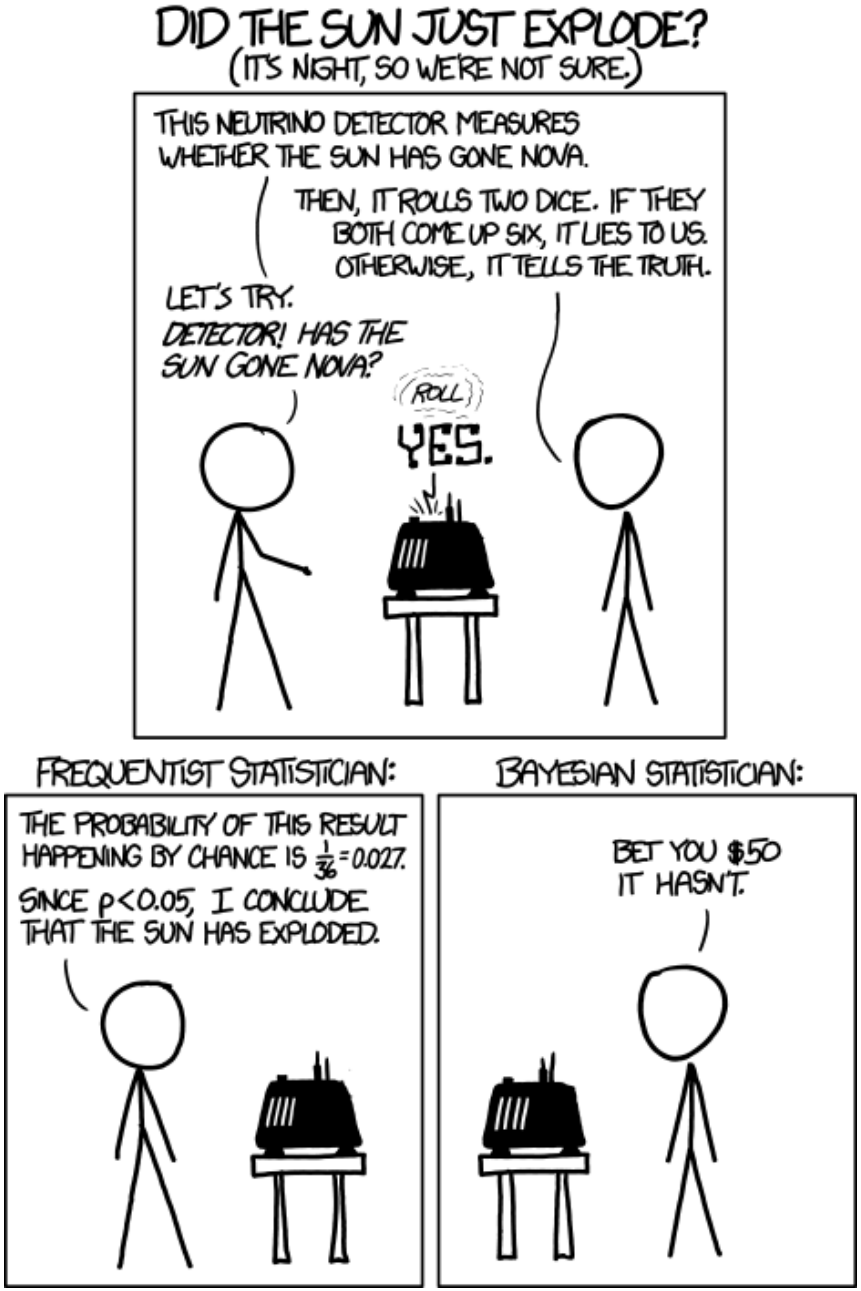
"...Bayesian statistics are rippling through everything from physics to cancer research, ecology to psychology."

Flam's definition of Bayesian Statistics

"A set of mathematical rules for using new data to continuously update beliefs or existing knowledge"

This definition misses the bigger picture...

Introduction to Bayesian Thought



Introduction to Bayesian Thought

But seriously...

In Bayesian statistics we begin with the premise that while the true value of θ is some unknown constant, we can treat θ as if it were a random variable that is dependent on the data and therefore describe its behavior mathematically

Once we have this mathematical form or distribution, we can generate any quantity of interest and conduct inference accordingly

The primary question is how do we determine the distribution?

Introduction to Bayesian Thought

Some definitions and notation:

Likelihood or $\mathcal{L}$

A mathematical formula describing the data, X, that is dependent upon the parameter of interest, θ. The nature of X should determine the likelihood but we do have some choice.

If the data is the number of successes out of a total number of independent trials, $\mathcal{L}$ should be Binomial

If the data is symmetric and unimodal, then it may be reasonable to assume the likelihood for X is normal but other distributions could suffice like the t

Introduction to Bayesian Thought

Prior or π

A mathematical formula describing our belief or knowledge about the parameter θ. We select this too but there is a lot of flexibility in what to choose.

We have the most control over the selection of the prior

There are priors that provide mathematical convenience but they may not always be best

The prior can include information about θ from past studies or experience but we want to balance just how much information we include so to not inadvertently guide the analysis in one direction

A Note on the Prior

There are several different kinds of prior but the ones you may hear most commonly talked about are non-informative versus informative priors

These names are self explanatory:

- Informative priors give information about θ regarding its value
- Non-informative priors do not

We will encounter both types today

Introduction to Bayesian Thought

The Basic Bayesian Model

Combining the Likelihood and the Prior, we get the Posterior:

$$\text{Posterior} = C \times \text{Likelihood} \times \text{Prior}$$

where C is some constant. Thus the Posterior is the product of the Likelihood and the Prior.

A more mathematical definition is

$$p(\theta|X) \propto \mathcal{L}(X|\theta)\pi(\theta)$$

where "$\propto$" means "proportional to" and "$|$" means "given"

Introduction to Bayesian Thought

The Posterior is the distribution of the parameter of interest given the data—the mathematical formula that describes the behavior of θ that is also dependent upon the data, X

The goal of a Bayesian analysis is to first estimate the Posterior and then generate random samples from it using software

From these samples, we can calculate the quantities we desire to describe the parameter

Introduction to Bayesian Thought

A Quick Historical Aside

Bayesian Statistics is named for Thomas Bayes from whom we get Bayes' Rule (sometimes erroneously called Bayes' *Theorem*)

Bayes never published his eponymous rule, that was done by Richard Price who was asked to edit Bayes' unpublished works after his death in 1761

Pierre-Simon Laplace popularized the rule and laid the groundwork for what is now Bayesian Statistics—although it took nearly 200 years to catch on

Introduction to Bayesian Thought

This is partly because many Bayesian models require a fair amount of computing power, relatively speaking, that simply didn't exist until 30 years ago

While some Bayesian models result in what we call a "recognizable" Posterior, most won't

Of the examples we consider today, only one will be "recognizable"

Road Map

1. One Sample Models
2. Two Sample Models
3. Regression

We will save computer illustrations of each of these for the workshop portion of the tutorial

One Sample Models

Example I.1

Consider a study of 37 bone marrow transplant patients. One outcome of interest is whether or not the subject developed acute graft versus host disease (GVHD) where the donated marrow views the recipient's body as foreign and begins to attack it.

Of the 37 study subjects, 17 developed GVHD by the end of the study period

Using this data, we wish to estimate the proportion of bone marrow recipients who will develop GVHD

One Sample Models

Continuing Example I.1, we must now select the likelihood and the prior

Selecting the Likelihood

We have 37 total binary (presumably independent) observations and the sum of those observations is 17 this suggests a *Binomial* likelihood

Selecting the Prior

We have no prior knowledge so we select a prior that is non-informative, in this case the prior equals 1

One Sample Models

Mathematical Form of Posterior

Combining the likelihood and the prior, we have

$$p(\theta|X, N) \propto \theta^X (1-\theta)^{N-X} \text{ or } \theta^{17}(1-\theta)^{20}$$

While it may not be obvious, this mathematical form is in fact recognizable as something called a *Beta* distribution meaning it is easy to use `R` to draw samples from it

Now that we have the form, we can easily generate a posterior sample

One Sample Models

In R, I generate 20,000 samples from the posterior for the proportion of bone marrow transplant patients who will develop GVHD

The median posterior estimate is 0.461 with a 95% *Credible* Interval of $(0.307, 0.619)$

Based on our sample, it appears 46.1% of bone marrow transplant patients are likely to develop GVHD

One Sample Models

Note that in Bayesian Statistics, we don't use *confidence* intervals but instead use *credible* intervals

Interpretation of the Credible Interval

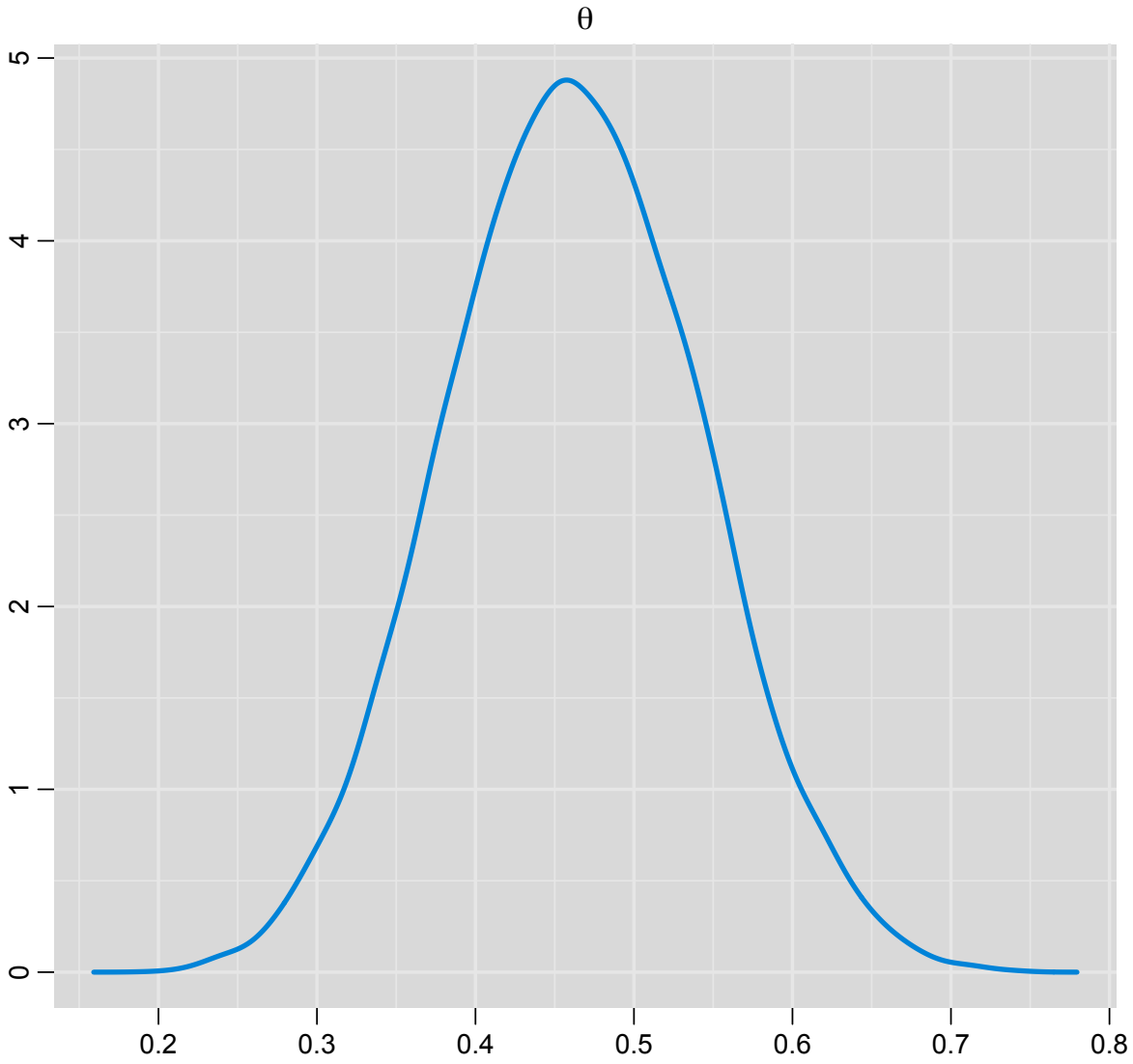
95% of the posterior estimates of θ are between 0.307 and 0.619

In context, the posterior estimate for the proportion of bone marrow transplant patients who will develop GVHD falls between 0.307 and 0.619 on 95% of samples

We can also say the posterior probability that θ is between 0.307 and 0.619 is 0.95

One Sample Models

We can plot the posterior distribution of θ



One Sample Models

Take Home Message from Example I.1

- Used non-informative prior
  - Best to do in absence of prior knowledge
- Estimate of θ determined from posterior samples
  - Used median, could also mean
- To measure uncertainty, used the middle 95% of distribution
  - Could also calculate standard deviation of posterior samples

One Sample Models

Take Home Message from Example I.1

- Used non-informative prior
  - Best to do in absence of prior knowledge
- Estimate of θ determined from posterior samples
  - Used median, could also mean
- To measure uncertainty, used the middle 95% of distribution
  - Could also calculate standard deviation of posterior samples

Question: How do we incorporate prior information?

One Sample Models

The short answer:

One Sample Models

The short answer:
Use a different prior that allows for it

One Sample Models

The short answer:
Use a different prior that allows for it

The long answer:

One Sample Models

The short answer:
Use a different prior that allows for it

The long answer:
We instead place what is called a *Beta* prior based on the information at hand, either from a previous study or somehow based on "expert" knowledge

One Sample Models

The short answer:
Use a different prior that allows for it

The long answer:
We instead place what is called a *Beta* prior based on the information at hand, either from a previous study or somehow based on "expert" knowledge

The *Beta* prior is based off of the *Beta* distribution which we use for its mathematical convenience

One Sample Models

The short answer:
Use a different prior that allows for it

The long answer:
We instead place what is called a *Beta* prior based on the information at hand, either from a previous study or somehow based on "expert" knowledge

The *Beta* prior is based off of the *Beta* distribution which we use for its mathematical convenience

Other priors are possible, but the posteriors become more difficult to handle

One Sample Models

Suppose we conduct a second study of GVHD in bone marrow transplant patients this time recruiting 200 subjects

Of those 200, we find that 83 develop GVHD by the end of the study period

To analyze this data, we could use the Posterior from the first study as our prior

One Sample Models

Mathematical forms for new analysis:

Informative Prior

$$\pi(\theta) \propto \theta^{17}(1-\theta)^{20}$$

Likelihood

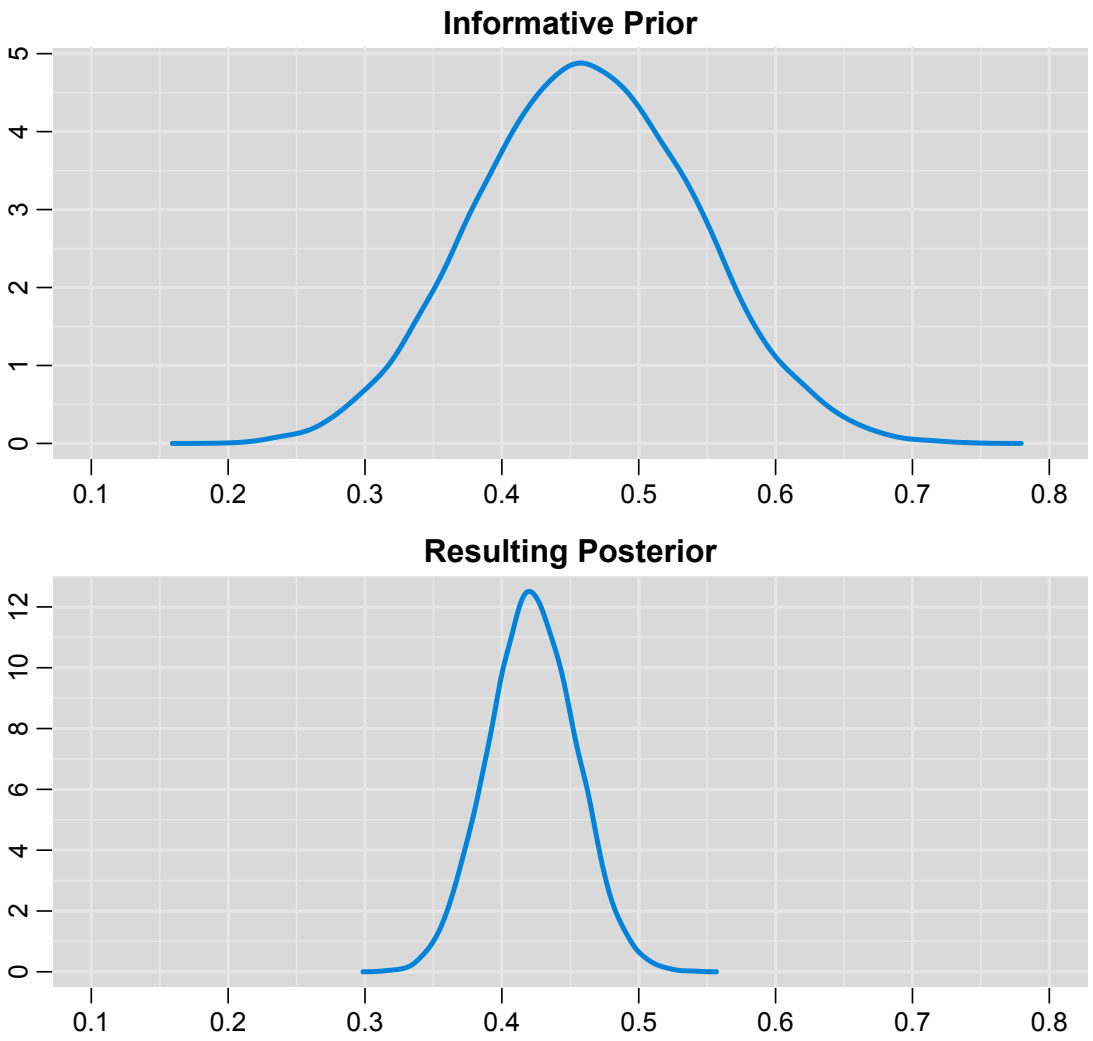
$$\mathcal{L}(X|\theta, N) \propto \theta^{83}(1-\theta)^{117}$$

Posterior

$$\begin{aligned} P(\theta|X, N) &\propto \mathcal{L}(X|\theta, N) \times \pi(\theta) \\ &= \theta^{83+17}(1-\theta)^{117+20} \end{aligned}$$

One Sample Models

The proportion of bone marrow transplant that will develop GVHD is estimated to be 42.2% (36.1%, 48.5%)



One Sample Models

We now consider the Bayesian equivalent of the paired-t test

These models get slightly more complicated mathematically, but the general intuition remains the same: select a likelihood and prior, multiply the two and draw samples from the posterior

Example I.2

Consider a classic study of sleeping agents from 1905. Ten subjects were each given two soporific drugs, on different nights, and asked to record their change in hours of sleep. Each student serves as its own control and the difference in the change in hours of sleep is the data of interest (first drug − second drug).

One Sample Models

Differences tend to be normally distributed, so we assume the likelihood is normal with some unknown mean μ and unknown variance σ^2—turning a two sample case into a one sample

Using a non-informative prior for both μ and σ^2 jointly, the posterior distribution is complex but can be sampled easily using an iterative procedure known as Gibbs Sampling

In a Gibbs Sampler, the code will alternate between draws from the posterior for μ and the posterior for σ until we obtain the desired number of posterior samples

One Sample Models

Sketch of Sampler

At each step of the code, we first randomly take a new μ from a *Normal* distribution with mean equal to $\bar{X}$ and variance equal to the previous step's variance estimate divided by N

Then we randomly take a new σ^2 from an *Inverse Gamma* distribution where the parameters are based on the data as well as the previous step's estimate of the mean

One Sample Models

Posterior Summaries from Sleep Data

The mean difference in the change in the number of hours of sleep is -1.575 $(-2.458, -0.695)$

We can also summarize the variance: the estimated variance is 1.642 (0.717, 5.10)

The paired setting has a natural null hypothesis to test, $\mathrm{H}_0 : \delta = 0$ where δ is the unknown mean difference

One Sample Models

Posterior Summaries from Sleep Data

The mean difference in the change in the number of hours of sleep is -1.575 $(-2.458, -0.695)$

We can also summarize the variance: the estimated variance is 1.642 (0.717, 5.10)

The paired setting has a natural null hypothesis to test, $\mathrm{H}_0 : \delta = 0$ where δ is the unknown mean difference

<u>Question</u>: How do we test this hypothesis in the Bayesian setting?

One Sample Models

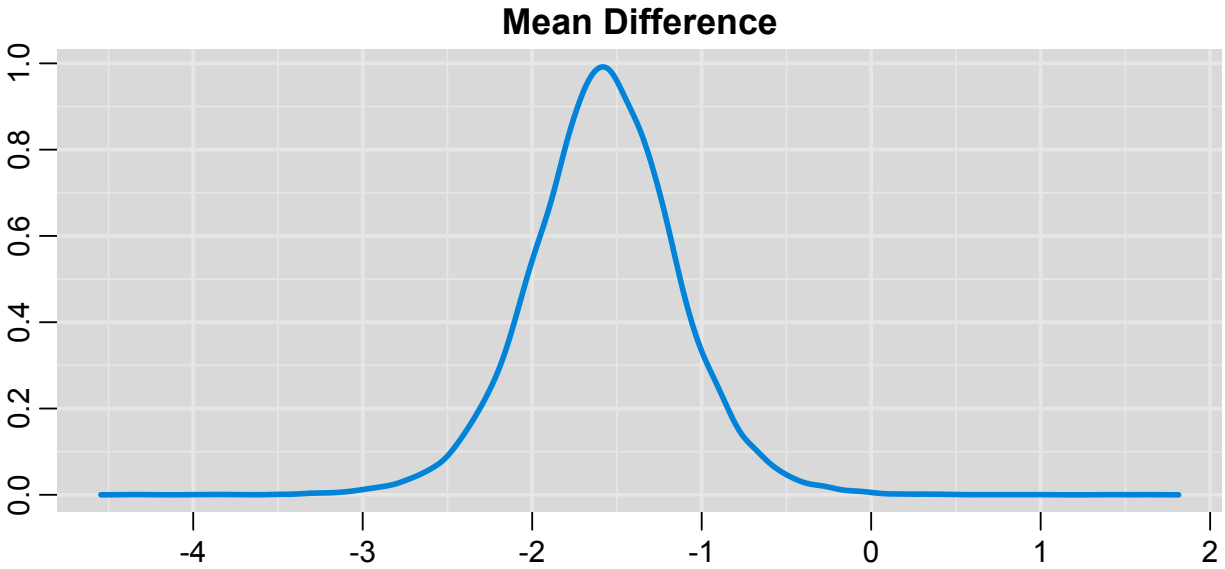
There are several ways to conduct hypothesis testing, but the easiest is to use the Credible Interval to test the null (similar to how one might use a confidence interval)

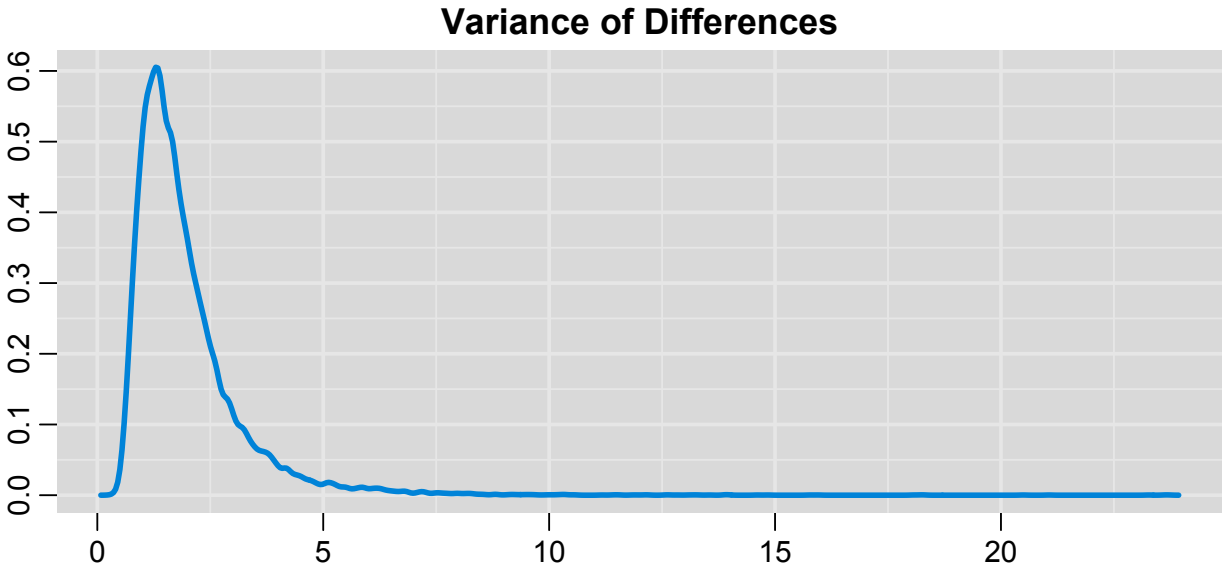
The 95% Credible Interval is $(-2.458, -0.695)$ which we see excludes zero, therefore we may consider the result to be statistically significant at the $\alpha = 0.05$ level

We could also find the posterior probability that the mean difference is greater than 0, i.e. find the proportion of posterior samples that are greater 0 which is 0.0012 or 0.12%

One Sample Models

Posterior distributions of parameters:





Two Sample Models

Let's now look at the Bayesian equivalent of the two sample t test

Example II.1

Suppose we want to look at the cold tolerance of a specific species of grass, *Echinochloa crusgalli*, that was subjected to two different treatments. CO2 uptake was measured in both treatment groups with one group of plants being chilled overnight before the measurements were taken.

Our goal is to compare mean CO2 uptake between plants that were chilled overnight and plants that were not

Two Sample Models

There are several ways to proceed:

1. We could fit two separate normal models to each group and compare the difference in the posterior means—corresponds to two sample t with *unequal* variance
2. Noting the relationship between the two-sample t and regression, we could fit a Bayesian regression model with treatment as a predictor—corresponds to two sample t with *equal* variance

We will use the former to illustrate a point about Bayesian statistics, specifically that we can use posterior samples to calculate almost any statistic of interest

Two Sample Models

The Bayesian model then proceeds as it did for the paired example in I.2 but in replicate: one model for each group

So we assume mean CO2 uptake in the non-chilled group is normal with mean μ_N and variance σ_N^2

Similarly, we assume mean CO2 uptake in the chilled group is normal mean μ_C and variance σ_C^2

Two Sample Models

Separate Summaries

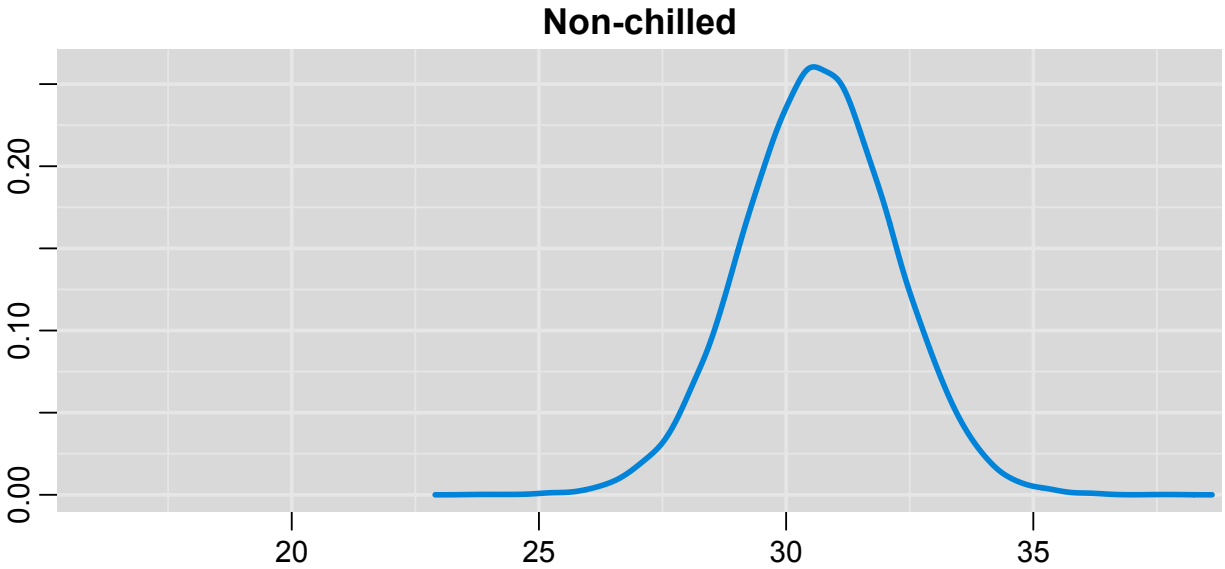
We first summarize the findings separately:

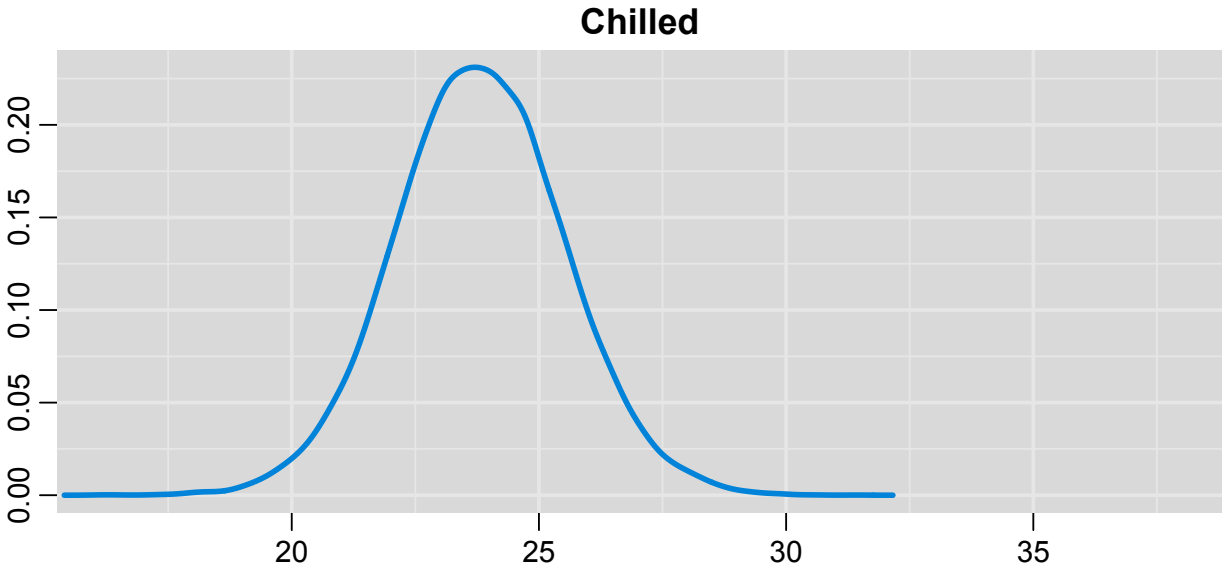
- The estimated mean CO2 uptake in non-chilled plants is 30.638 $\mathrm{umol/m^2}$ sec (27.578, 33.644)
- In the chilled plants, mean CO2 uptake is 23.781 $\mathrm{umol/m^2}$ sec (20.450, 27.171)

The minimal overlap in the credible intervals suggests a possible difference

Two Sample Models

Graphically, we examine a plot of both estimates





Two Sample Models

Summary of Difference

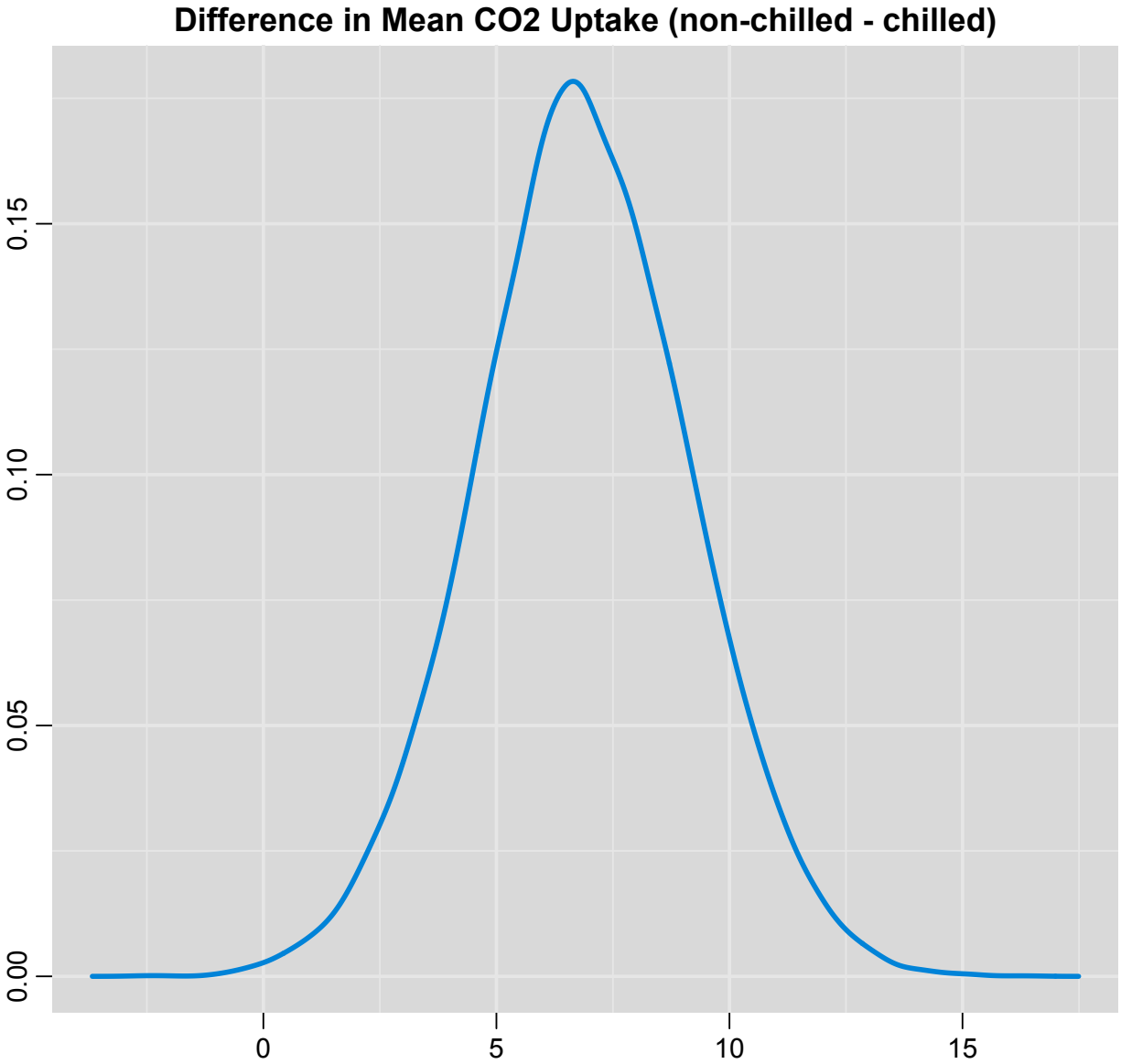
The difference in mean CO2 uptake comparing non-chilled to chilled plants is 6.815 (2.262, 11.378) where the difference is taken as non-chilled − chilled

Under the null, $H_0 : \mu_N - \mu_C = 0$, since 0 is excluded from the interval, we have evidence to suggest a difference in mean CO2 uptake with non-chilled plants having a larger uptake

The posterior probability that $\mu_N - \mu_C > 0$ is 0.9984 or 99.84%

Two Sample Models

Graphically, we the posterior distribution of the difference is



Two Sample Models

We now consider the Bayesian analog to the two proportion test

Example II.2

In a study of stroke victims, suppose we wish to compare the proportion of deaths by end of follow-up between those subjects who had prior history of hypertension and those who did not. Patients were collected over the course of three years and followed for up to five years from enrollment.

We follow the same strategy as in the two sample t model where we fit separate models to each proportion and then take the difference of the posterior samples

Two Sample Models

The 2×2 table summarizing the data is

	Dead		
Hyp.	Yes	No	Total
Yes	225	186	411
No	253	157	410

For those with a history of hypertension, $X_1 = 225$ and $N_1 = 411$ and for those without, $X_2 = 253$ and $N_2 = 410$

From before, using non-informative priors, we know the posteriors have the forms $p(\theta_1|X_1, N_1) = \theta_1^{225}(1-\theta_1)^{411-225}$ and $p(\theta_2|X_2, N_2) = \theta_2^{253}(1-\theta_2)^{410-253}$ where θ_1 and θ_2 is the probability of dying for each group

Two Sample Models

Separate Summaries

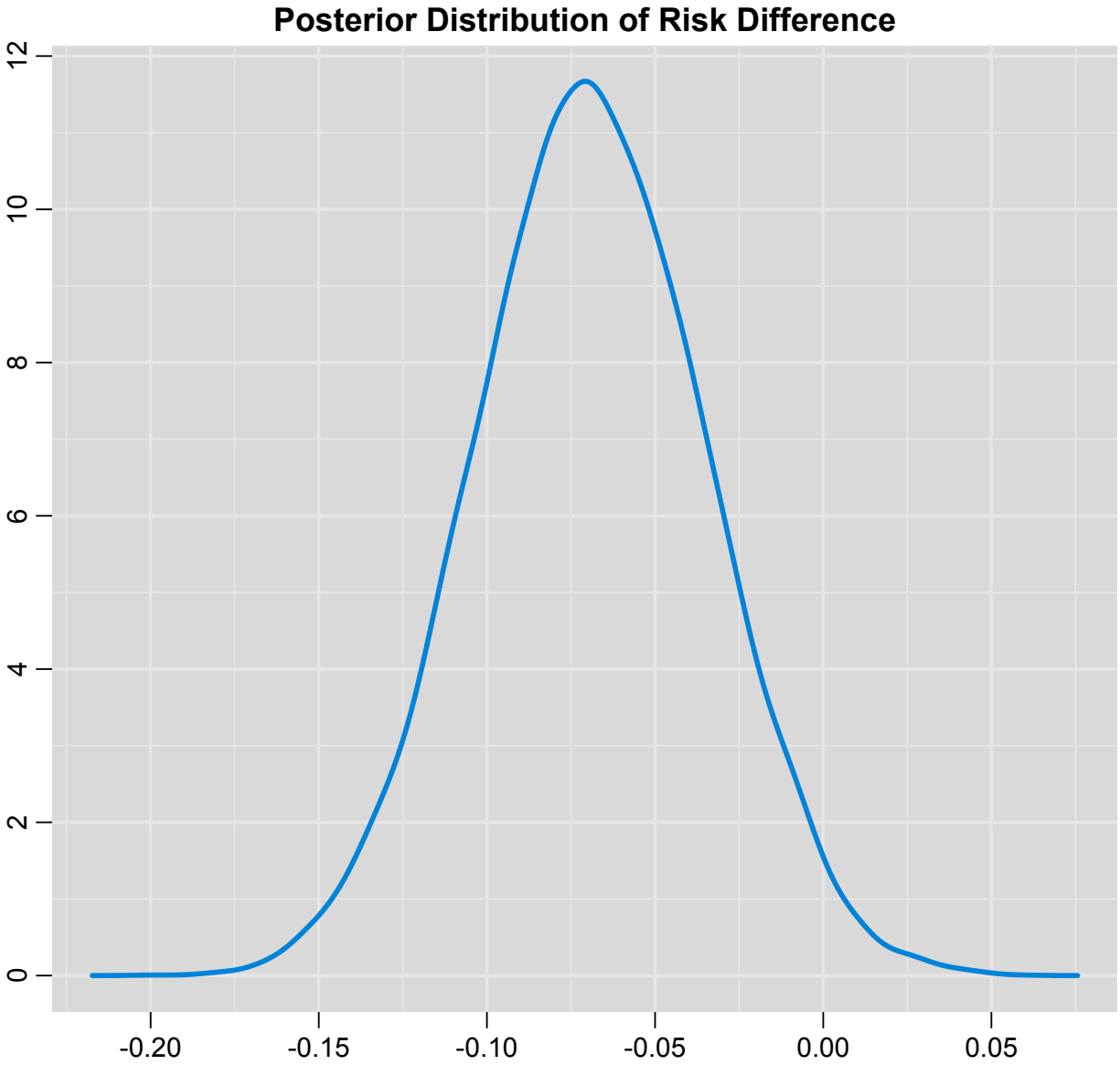
The estimated probability of dying for stroke patients with a prior history of hypertension is 0.547 (0.499, 0.595) while the probability of dying for stroke patients without a prior history of hypertension is 0.617 (0.570, 0.663).

Difference Summary

The difference in probability of death, or risk difference, is -0.069 $(-0.136, -0.003)$. Given a null of no difference, there may be some evidence to suggest those without a prior history of hypertension are at higher risk of dying following a stroke (posterior probability of 0.020).

Two Sample Models

Graphically, the posterior distribution of the difference is



Regression

We now examine a Bayesian regression model

Example III.1

Consider the classic data from Longley on seven macroeconomic indicators in the US observed yearly from 1947 to 1962. We're interested in using some of those indicators to predict the number of people who were employed in each year.

The model we will build is

$$\text{Employed}_i = \beta_0 + \beta_1 \text{GNP}_i + \beta_2 \text{Armed.forces}_i + \beta_3 \text{Population}_i + \epsilon_i$$

Regression

Likelihood

We make the typical regression assumption that the errors, ϵ_i, are normally distributed with mean zero and variance σ^2 thus the outcome, Employed_i, is normal with mean $\beta_0 + \beta_1 \text{GNP}_i + \beta_2 \text{Armed.forces}_i + \beta_3 \text{Population}_i$ and variance σ^2

Prior

The standard regression prior is to assume the joint prior of the coefficients and the variance is proportional to σ^{-2}, which is non-informative

Once again, we need a Gibbs Sampler to estimate posterior samples for the β's and σ^2

Regression

Note on the Longley Data

This dataset is a common example of data with large amounts of multicollinearity, i.e. correlated predictors

In the Frequentist setting, one could use Ridge Regression to account for multicollinearity

In the Bayesian setting, Ridge Regression corresponds to placing a normal prior on the parameters collinear covariates, so β_k is normal, mean zero with variance σ^2_β

Regression

Below is a table of the posterior estimates

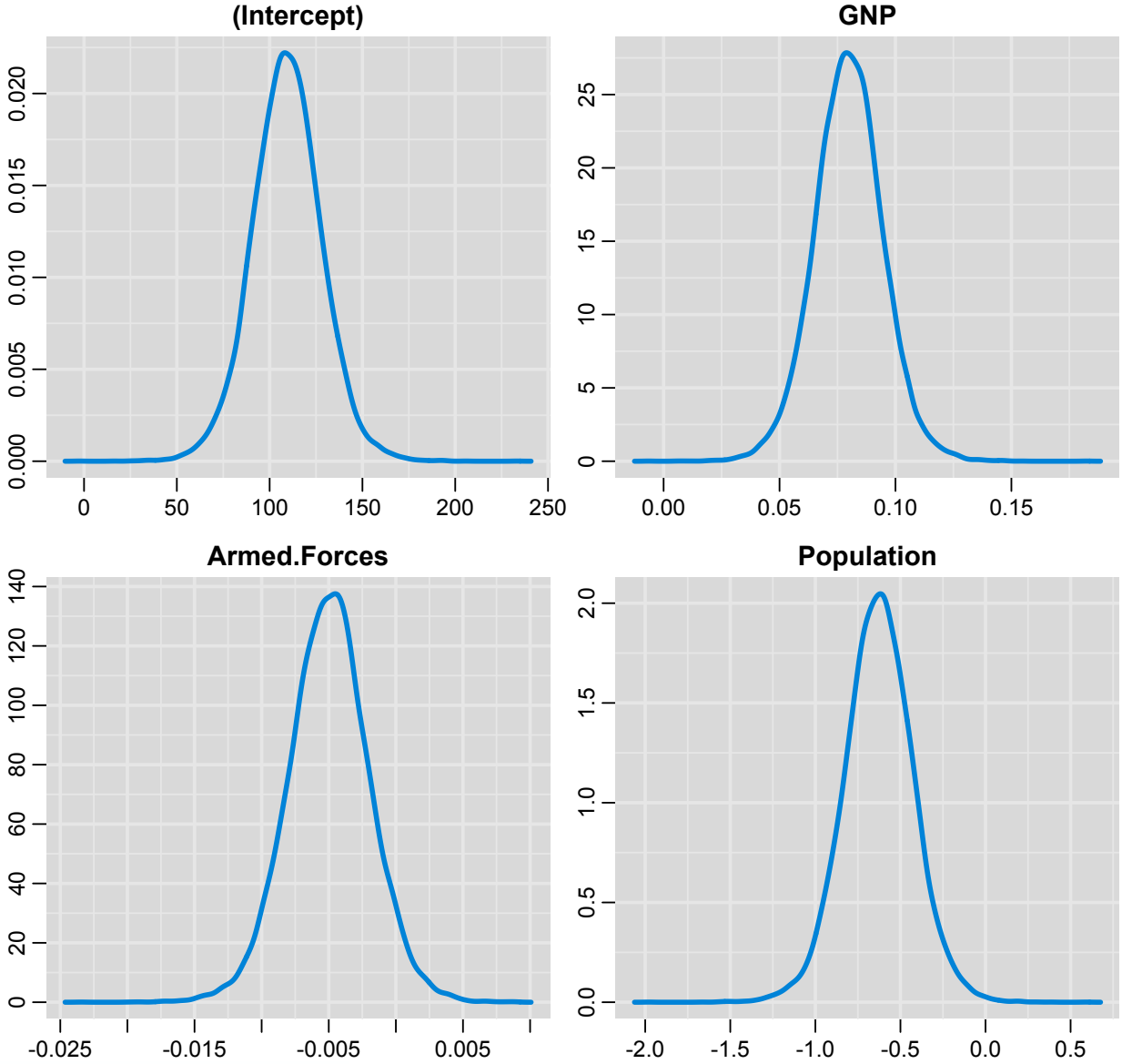
Covariate	Parameter	Estimate	2.5%	97.5%
(Intercept)	β_0	109.277	72.280	146.682
GNP	β_1	0.080	0.050	0.110
Armed.Forces	β_2	-0.005	-0.011	0.001
Population	β_3	-0.627	-1.032	-0.224
	σ^2	0.270	0.131	0.693

Given the scale, we would likely prefer to interpret these for changes in each covariate that are larger than 1 (say 10 or 100 even)

To test $H_0 : \beta_k = 0$, use the credible intervals

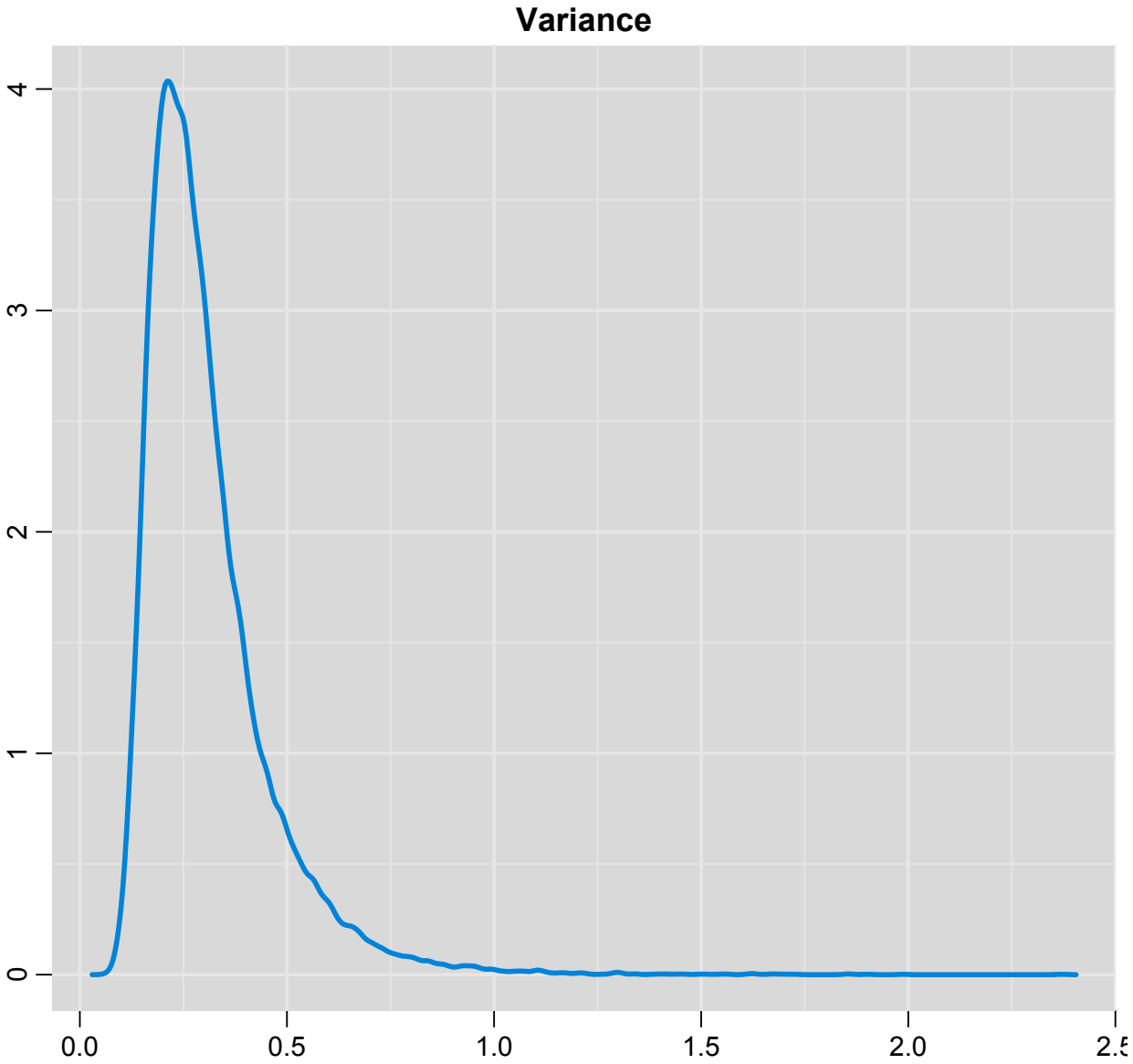
Regression

Graphically, the posterior distributions of the coefficients are



Regression

Graphically, the posterior distribution of the variance is



Discussion

We've covered a fair amount of ground, but there's a lot more:

- Informative priors for the normal likelihood models (hint: it's normal)
- Bayes Factors and additional Bayesian Inference techniques
- Other regression types, logistic or poisson or mixed effects
- Bayesian models without such "easy" posteriors
- Sampling techniques such as Metropolis-Hastings, Slice Sampling, Importance Sampling, Rejection Sampling, Adaptive Rejection Sampling, Hamiltonian Monte Carlo, or Variational Bayesian Inference

But hopefully this is a good start!

Discussion

Discussion

Many of these techniques produce similar results to their Frequentist analogs

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

- Bayesian models can give us greater flexibility

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

- Bayesian models can give us greater flexibility
- We don't have to perform analyses in a vacuum

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

- Bayesian models can give us greater flexibility
- We don't have to perform analyses in a vacuum
- Some models are theoretically simpler in the Bayesian setting

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

- Bayesian models can give us greater flexibility
- We don't have to perform analyses in a vacuum
- Some models are theoretically simpler in the Bayesian setting
- Don't have to rely on the Central Limit Theorem

Discussion

Many of these techniques produce similar results to their Frequentist analogs

So why choose Bayes?

- Bayesian models can give us greater flexibility
- We don't have to perform analyses in a vacuum
- Some models are theoretically simpler in the Bayesian setting
- Don't have to rely on the Central Limit Theorem

But ultimately, it is a philosophical choice

Discussion

Tutorial Time

Tutorial Time

Wait, there's one more thing...

Discussion

Discussion

What you've all been waiting for

Discussion

What you've all been waiting for

The question of the hour

Discussion

What you've all been waiting for

The question of the hour

How do you pronounce Bayesian?

Discussion

What you've all been waiting for

The question of the hour

How do you pronounce Bayesian?

This matter is still an open question

Tutorial Time

Tutorial code is available at `https://github.com/markjmeyer/`