Module 4: Introduction to the Normal Gamma Model

Rebecca C. Steorts and Lei Qian

Agenda

- Continue with more conjugate distributions
- Define the precision
- Introduce the Normal-Gamma model
- Consider our first hiearchical model with more than two levels
- Deriving the posterior for the Normal likelihood and Normal-Gamma prior
- Application to IQ scores (spurters versus controls)

NormalGamma distribution

Let

$X_1, \ldots, X_n \mid \mu, \lambda \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$ and assume **both**

- the mean μ and
- the precision $\lambda = 1/\sigma^2$ are **unknown**.

The NormalGamma(m, c, a, b) distribution, with $m \in \mathbb{R}$ and $c, a, b > 0$, is a joint distribution on (μ, λ) obtained by letting

$$\begin{aligned} \boldsymbol{\mu}|\lambda &\sim \mathcal{N}(m, (c\lambda)^{-1}) \\ \boldsymbol{\lambda} &\sim \text{Gamma}(a, b) \end{aligned}$$

In other words, the joint p.d.f. is

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu \mid m, (c\lambda)^{-1}) \, \text{Gamma}(\lambda \mid a, b)$$

which we will denote by NormalGamma$(\mu, \lambda \mid m, c, a, b)$.

NormalGamma distribution (continued)

It turns out that this provides a **conjugate prior** for (μ, λ).

One can show the posterior is

$$\boldsymbol{\mu}, \boldsymbol{\lambda} | x_{1:n} \sim \text{NormalGamma}(M, C, A, B) \tag{1}$$

i.e., $p(\mu, \lambda | x_{1:n}) = \text{NormalGamma}(\mu, \lambda \mid M, C, A, B)$, where

$$\begin{aligned} M &= \frac{cm + \sum_{i=1}^{n} x_i}{c + n} \\ C &= c + n \\ A &= a + n/2 \\ B &= b + \tfrac{1}{2}(cm^2 - CM^2 + \textstyle\sum_{i=1}^{n} x_i^2). \end{aligned}$$

NormalGamma distribution (continued)

For interpretation, B can also be written (by rearranging terms) as

$$B = b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2}\frac{cn}{c+n}(\bar{x} - m)^2. \tag{2}$$

In this module, we will derive the posterior derivation of this hierarchical model, which has three levels (or layers) to it.[1]

[1] The data is considered the first level, the prior on λ is considered the second level, and the prior on μ is considered the third level of the hierarchical model.

NormalGamma distribution (continued)

- M: Posterior mean for μ. It is a weighted average (convex combination) of the prior mean and the sample mean:

$$M = \frac{c}{c+n} m + \frac{n}{c+n} \bar{x}.$$

- C: "Sample size" for estimating μ. (The standard deviation of $\mu|\lambda$ is $\lambda^{-1/2}/\sqrt{C}$.)
- A: Shape for posterior on λ. Grows linearly with sample size.
- B: Rate (1/scale) for posterior on λ. Equation 2 decomposes B into the prior variation, observed variation (sample variance), and variation between the prior mean and sample mean:

$$B = (\text{prior variation}) + \tfrac{1}{2} n (\text{observed variation}) \quad (3)$$
$$+ \tfrac{1}{2} \tfrac{cn}{c+n} (\text{variation bw means}). \quad (4)$$

Posterior derivation

First, consider the NormalGamma density. Dropping constants of proportionality, multiplying out $(\mu - m)^2 = \mu^2 - 2\mu m + m^2$, and collecting terms, we have

$$\begin{aligned}
&\text{NormalGamma}(\mu, \lambda \mid m, c, a, b) && (5)\\
&= \mathcal{N}(\mu \mid m, (c\lambda)^{-1}) \, \text{Gamma}(\lambda \mid a, b) \\
&= \sqrt{\frac{c\lambda}{2\pi}} \exp\Big(-\tfrac{1}{2} c\lambda(\mu - m)^2 \Big) \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda) \\
&\underset{\mu,\lambda}{\propto} \lambda^{a-1/2} \exp\Big(-\tfrac{1}{2} \lambda (c\mu^2 - 2cm\mu + cm^2 + 2b) \Big). && (6)
\end{aligned}$$

Posterior derivation (continued)

Similarly, for any x,

$$\begin{aligned}\mathcal{N}(x \mid \mu, \lambda^{-1}) &= \sqrt{\frac{\lambda}{2\pi}} \exp\Big(-\tfrac{1}{2}\lambda(x-\mu)^2 \Big) \\ &\underset{\mu,\lambda}{\propto} \lambda^{1/2} \exp\Big(-\tfrac{1}{2}\lambda(\mu^2 - 2x\mu + x^2) \Big). \qquad (7)\end{aligned}$$

Posterior derivation (continued)

$$
\begin{aligned}
& p(\mu, \lambda | x_{1:n}) \\
& \underset{\mu,\lambda}{\propto} p(\mu, \lambda) p(x_{1:n} | \mu, \lambda) \\
& \underset{\mu,\lambda}{\propto} \lambda^{a-1/2} \exp\Big(-\tfrac{1}{2}\lambda(c\mu^2 - 2cm\mu + cm^2 + 2b)\Big) \\
& \qquad \times \lambda^{n/2} \exp\Big(-\tfrac{1}{2}\lambda(n\mu^2 - 2(\textstyle\sum x_i)\mu + \sum x_i^2)\Big) \\
& = \lambda^{a+n/2-1/2} \exp\Big(-\tfrac{1}{2}\lambda((c+n)\mu^2 - 2(cm + \textstyle\sum x_i)\mu + cm^2 + 2b + \sum x_i^2)\Big) \\
& \overset{(a)}{=} \lambda^{A-1/2} \exp\Big(-\tfrac{1}{2}\lambda\big(C\mu^2 - 2CM\mu + CM^2 + 2B\big)\Big) \\
& \overset{(b)}{\propto} \text{NormalGamma}(\mu, \lambda \mid M, C, A, B)
\end{aligned}
$$

where step (b) is by Equation 6, and step (a) holds if $A = a + n/2$, $C = c + n$, $CM = (cm + \sum x_i)$, and

$$CM^2 + 2B = cm^2 + 2b + \sum x_i^2.$$

Posterior derivation (continued)

We have two equations and two unknowns (B and M):

$$CM = (cm + \sum x_i)$$

$$CM^2 + 2B = cm^2 + 2b + \sum x_i^2.$$

Solving for M and B, (verify this on your own), we can find that

$$M = (cm + \sum x_i)/(c + n)$$

and

$$B = b + \tfrac{1}{2}(cm^2 - CM^2 + \sum x_i^2).$$

Alternative derivation: complete the square all way. This is in the reading notes if you'd like to see a different derivation.

Do a teacher's expectations influence student achievement?

Do a teacher's expectations influence student achievement? In a famous study, Rosenthal and Jacobson (1968) performed an experiment in a California elementary school to try to answer this question. At the beginning of the year, all students were given an IQ test. For each class, the researchers randomly selected around 20% of the students, and told the teacher that these students were "spurters'' that could be expected to perform particularly well that year. (This was not based on the test—the spurters were randomly chosen.) At the end of the year, all students were given another IQ test. The change in IQ score for the first-grade students was:[2]

This applied example corresponds to lab 4 and homework 5.

[2] The original data is not available. This data is from the `ex1321` dataset of the `R` package `Sleuth3`, which was constructed to match the summary statistics and conclusions of the original study.

Do a teacher's expectations influence student achievement?

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
```

Spurters/Control Data

```
#spurters
x <- c(18, 40, 15, 17, 20, 44, 38)
#control
y <- c(-4, 0, -19, 24, 19, 10, 5, 10,
       29, 13, -9, -8, 20, -1, 12, 21,
       -7, 14, 13, 20, 11, 16, 15, 27,
       23, 36, -33, 34, 13, 11, -19, 21,
       6, 25, 30,22, -28, 15, 26, -1, -2,
       43, 23, 22, 25, 16, 10, 29)
iqData <- data.frame(Treatment =
       c(rep("Spurters", length(x)),
       rep("Controls", length(y))),
       Gain = c(x, y))
```

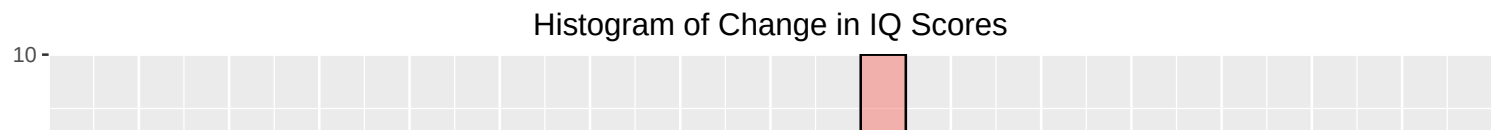
An initial exploratory analysis

Plot the **number of students** versus the **change in IQ score** for the two groups. How strongly does this data support the hypothesis that the teachers' expectations caused the spurters to perform better than their classmates?

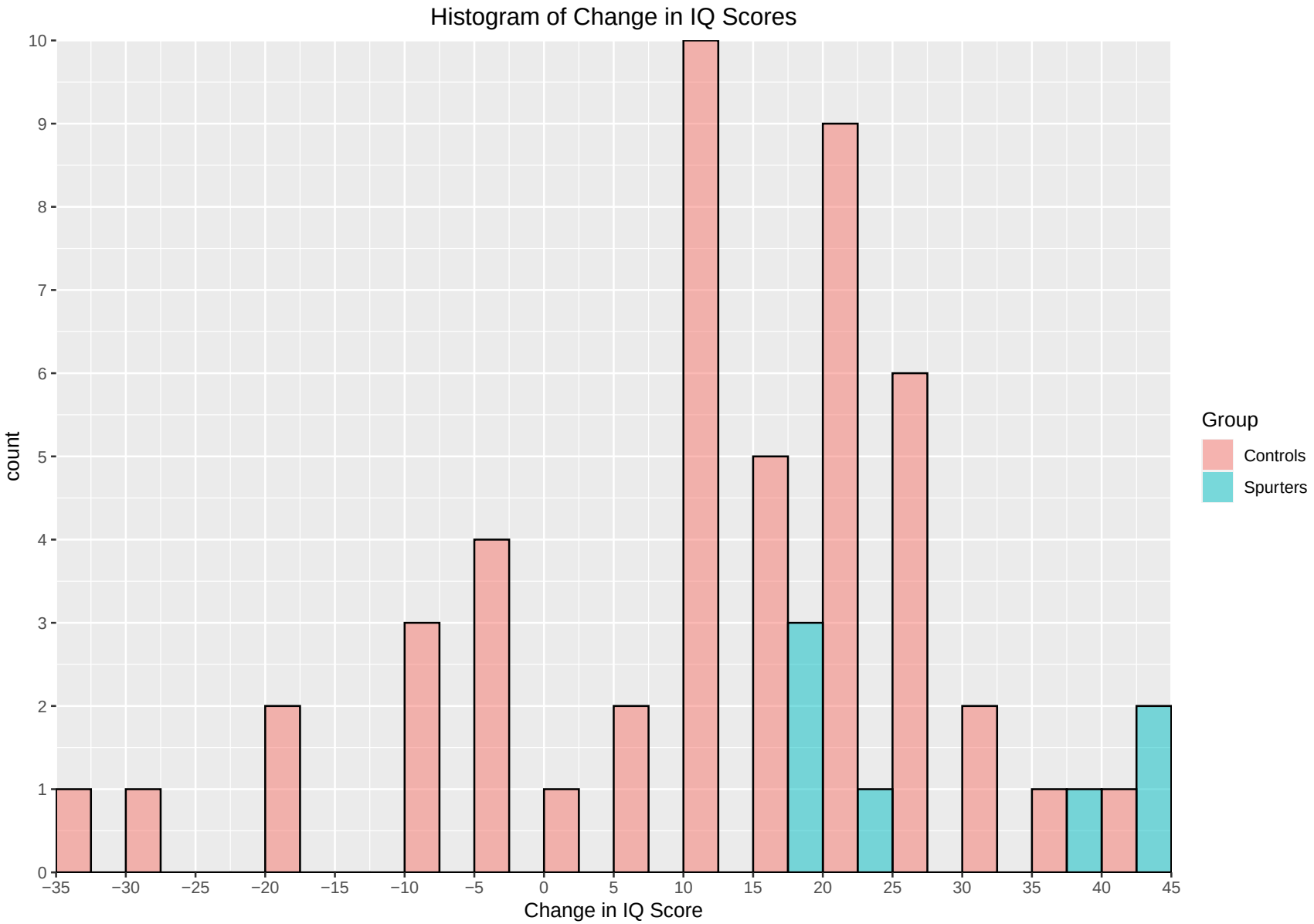
Histogram of Change in IQ Scores

```
xLimits = seq(min(iqData$Gain) - (min(iqData$Gain) %% 5),
              max(iqData$Gain) + (max(iqData$Gain) %% 5),
              by = 5)

ggplot(data = iqData, aes(x = Gain,
                          fill = Treatment,
                          colour = I("black"))) +
  geom_histogram(position = "dodge", alpha = 0.5,
                 breaks = xLimits, closed = "left")+
  scale_x_continuous(breaks = xLimits,
                     expand = c(0,0))+
  scale_y_continuous(expand = c(0,0),
                     breaks = seq(0, 10, by = 1))+
  ggtitle("Histogram of Change in IQ Scores") +
  labs(x = "Change in IQ Score", fill = "Group") +
  theme(plot.title = element_text(hjust = 0.5))
```



Histogram of Change in IQ Scores



IQ Tests and Modeling

IQ tests are purposefully calibrated to make the scores normally distributed, so it makes sense to use a normal model here:

$$\text{spurters: } X_1, \ldots, X_{n_S} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_S, \lambda_S^{-1})$$
$$\text{controls: } Y_1, \ldots, Y_{n_C} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_C, \lambda_C^{-1}).$$

- We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$?
- We don't know the standard deviations $\sigma_S = \lambda_S^{-1/2}$ and $\sigma_C = \lambda_C^{-1/2}$, and the sample seems too small to estimate them very well.

IQ Tests and Modeling

On the other hand, it is easy using a Bayesian approach: we just need to compute the posterior probability that $\mu_S > \mu_C$:

$$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid x_{1:n_S}, y_{1:n_C}).$$

Let's use independent NormalGamma priors:

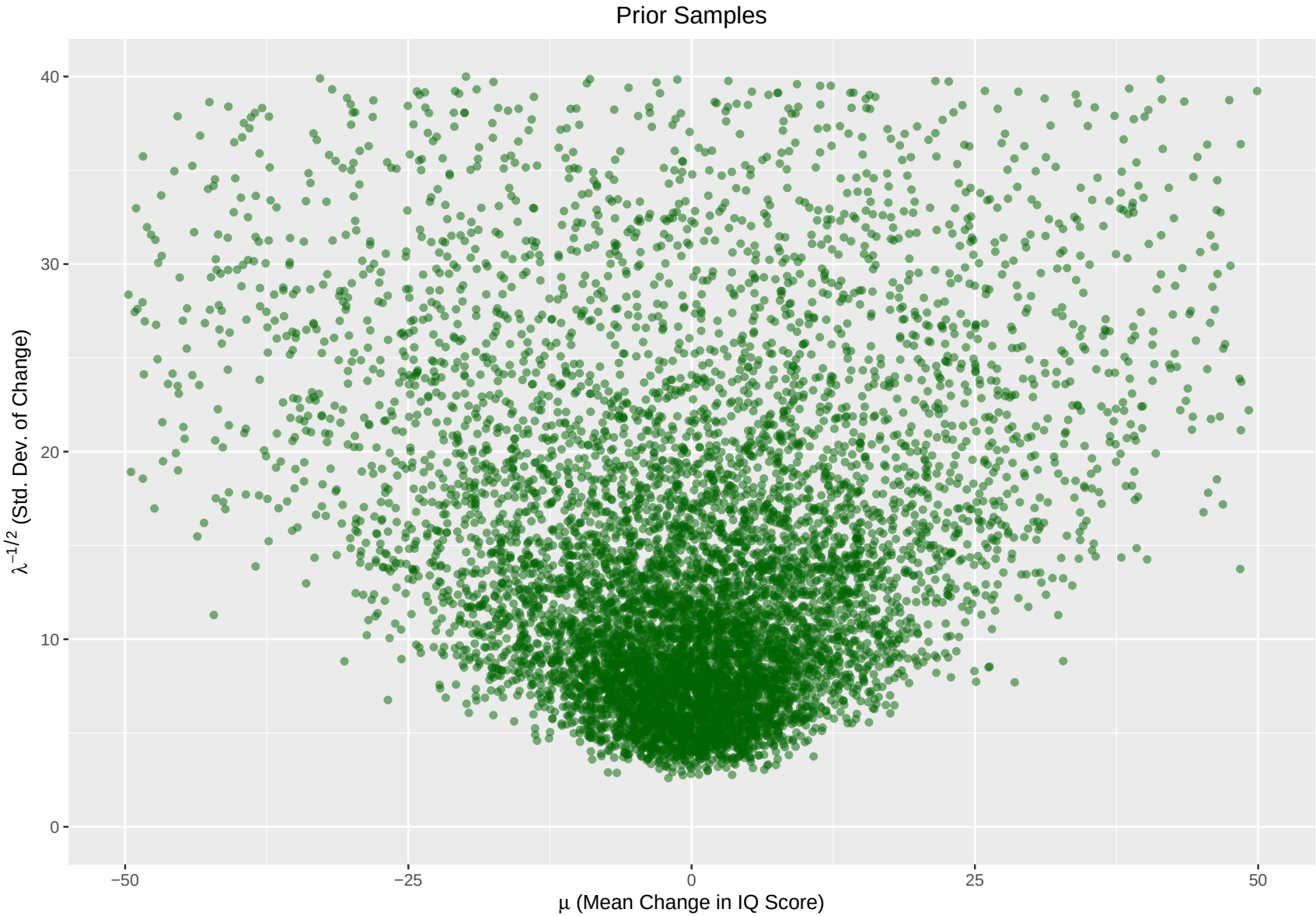
$$\text{spurters: } (\boldsymbol{\mu}_S, \boldsymbol{\lambda}_S) \sim \text{NormalGamma}(m, c, a, b)$$
$$\text{controls: } (\boldsymbol{\mu}_C, \boldsymbol{\lambda}_C) \sim \text{NormalGamma}(m, c, a, b)$$

Hyperparameter settings

- $m = 0$ (Don't know whether students will improve or not, on average.)
- $c = 1$ (Unsure about how big the mean change will be—prior certainty in our choice of m assessed to be equivalent to one datapoint.)
- $a = 1/2$ (Unsure about how big the standard deviation of the changes will be.)
- $b = 10^2 a$ (Standard deviation of the changes expected to be around $10 = \sqrt{b/a} = \mathbb{E}(\lambda)^{-1/2}$.)

Prior Samples

```
## Warning: Removed 2065 rows containing missing values (ge
```



Original question

"What is the posterior probability that $\mu_S > \mu_C$?"

- The easiest way to do this is to take a bunch of samples from each of the posteriors, and see what fraction of times we have $\mu_S > \mu_C$.
- This is an example of a Monte Carlo approximation (much more to come on this in the future).
- To do this, we draw $N = 10^6$ samples from each posterior:

$$(\mu_S^{(1)}, \lambda_S^{(1)}), \ldots, (\mu_S^{(N)}, \lambda_S^{(N)}) \sim \text{NormalGamma}(A_1, B_1, C_1, D_1)$$
$$(\mu_C^{(1)}, \lambda_C^{(1)}), \ldots, (\mu_C^{(N)}, \lambda_C^{(N)}) \sim \text{NormalGamma}(A_2, B_2, C_2, D_2)$$

Original question (continued)

What are the updated posterior values?

and obtain the approximation

$$\mathbb{P}(\mu_S > \mu_C \mid x_{1:n_S}, y_{1:n_C}) \approx \frac{1}{N} \sum_{i=1}^{N} I(\mu_S^{(i)} > \mu_C^{(i)}) = ??.$$

You will explore this more in lab 4 and homework 5.

Takeaways

- We introduced the Normal-Gamma distribution.
- We defined the precision (inverse of the variance).
- We defined a hiearchical model with three levels.
- We derived the posterior for the Normal likelihood and the Normal-Gamma prior. What did this involve?
- We utilized our modeling techniques this week on an application to IQ scores. How realistic was this model in practice?
- If you made changes to the model, what changes would you make and what complications might you run into?

Detailed Takeaways

- In order to understand this module, you must be able to derive the normal-normal model.
- You must have a solid foundation regarding the normal distribution and its properties.
- You must know the relationship between the precision and the variance.
- You should be able to derive the normal-normal-gamma model.
- You should conceptually understand why this model is important and when it's used for a case study.
- You should understand conceptually what each of the four parameters of the normal-gamma distribution represent.
- The case study for this problem (on IQ scores) is a good review problem regarding when you should use the normal-normal-gamma model over the normal-normal model. This is something that you should be able to conceptually walk through for this case study or a similar case study.

Module 4 Notes

Module 4 Class Notes can be found here:

https://github.com/resteorts/modern-bayes/tree/master/lectures ModernBayes20/lecture-4/04-class-notes