# Psych 100A Spring 2019: Week 7 Slides

Amanda Montoya

May 14, 2019

```r
NLSdata <- read.csv("http://bit.ly/NLSdata", header = TRUE)
```

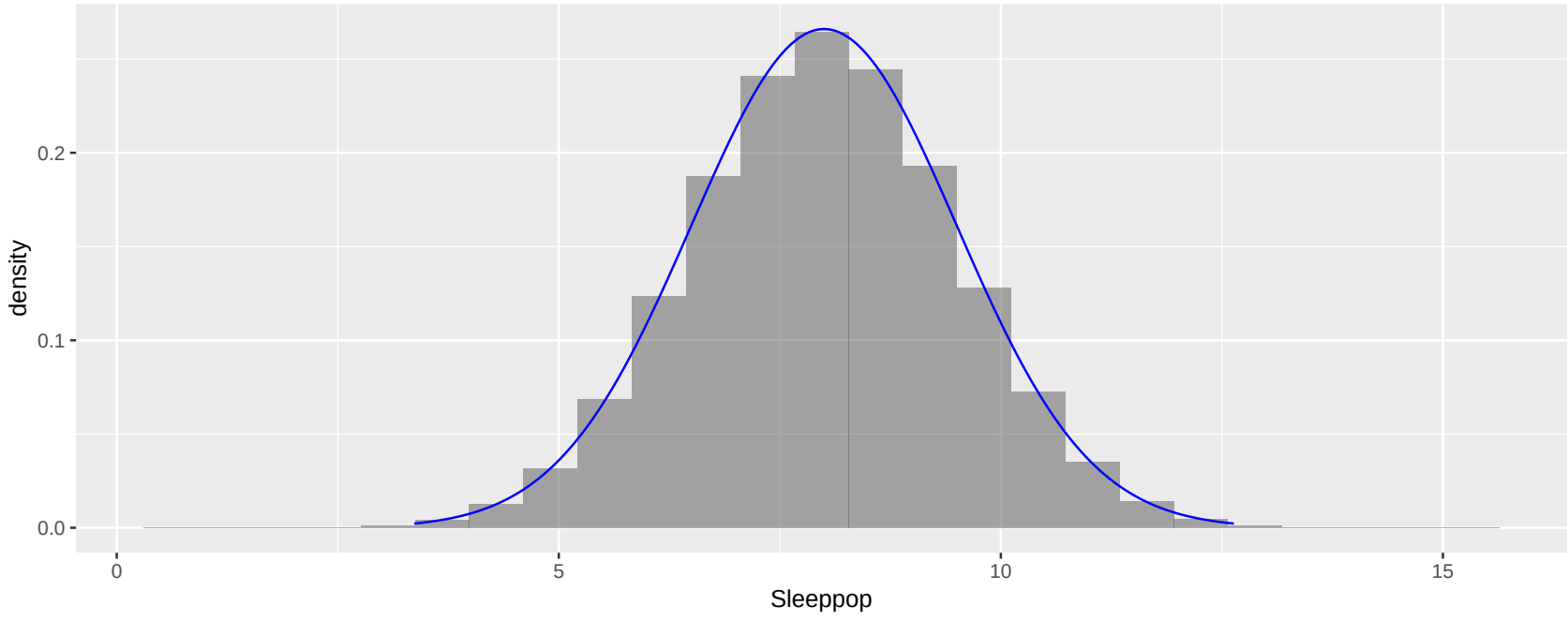
# Learning Outcomes Today

- Define a sampling distribution
- Describe what a sampling distribution is useful for
- Explain how to simulate a sampling distribution

# Hours of Sleep: The Data Generating Process

Imagine we live in a (wonderful) world, where on average people sleep 8 hours ($\mu = 8$), and the population standard deviation of hours of sleep is 1.5 ($\sigma = 1.5$). Let's also assume for now that hours of sleep is normally distributed.

```
Sleeppop <- rnorm(1000000, mean = 8, sd = 1.5)
gf_dhistogram(~Sleeppop) %>%
gf_dist( "norm" , color = "blue" , params =list(8,1.5))
```



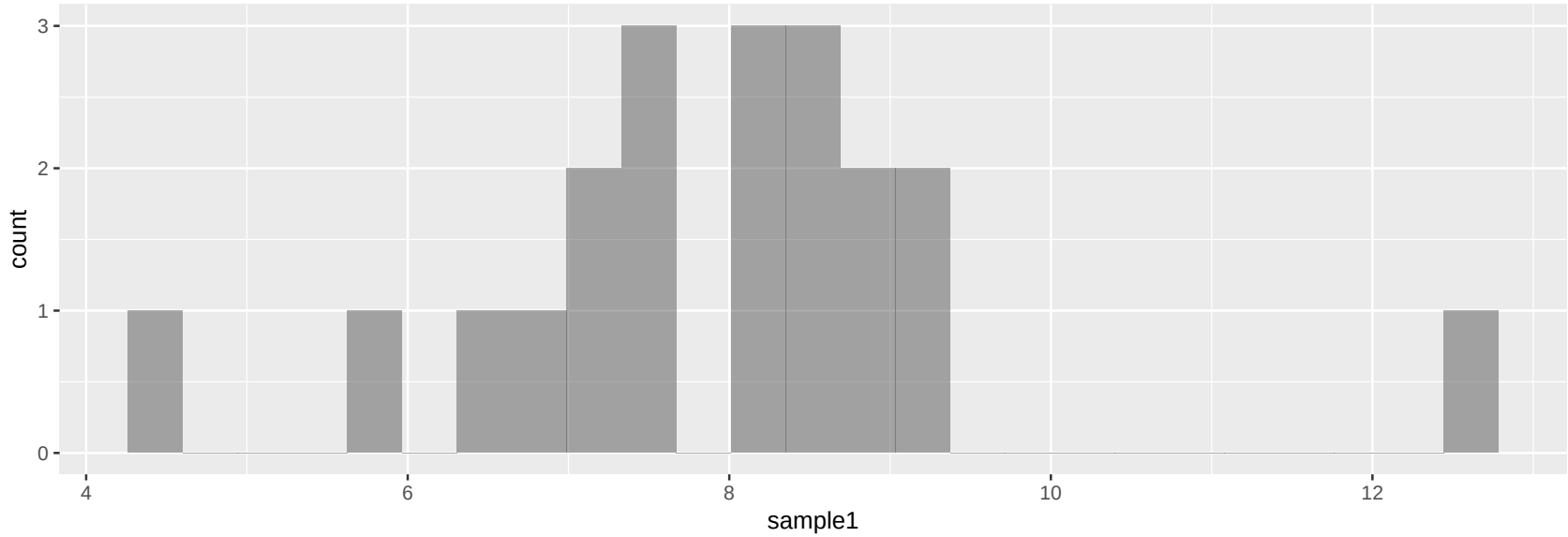
# Sampling from the Population

If I took a random sample of 20 people from the population of individuals, and asked how many hours they sleep and took the mean, we would expect that the mean would be close to 8, but perhaps not exactly 8.

```
sample1 <- sample(Sleeppop, 20)
mean(sample1)
```

```
## [1] 7.972252
```

```
gf_histogram(~sample1)
```



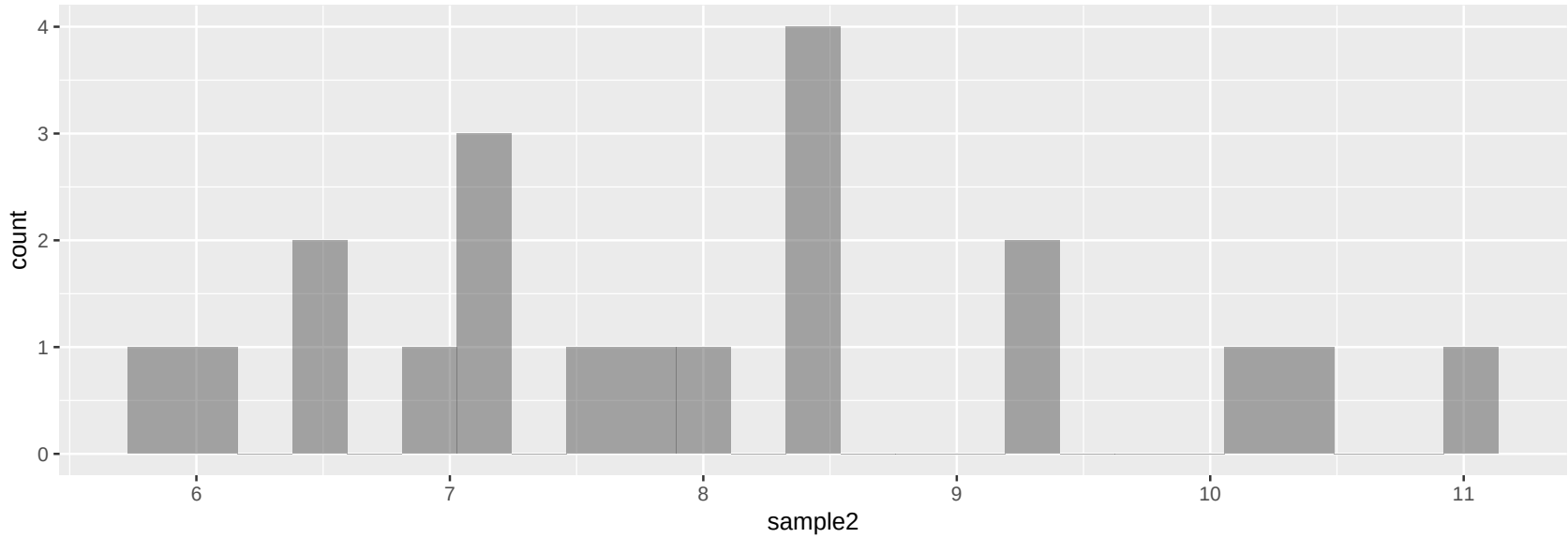
# Taking another sample

Similarly, if I took another random sample of 20 people, the mean would be close to 8, not exactly 8, and not exactly the same as from the first sample.

```
sample2 <- sample(Sleeppop, 20)
mean(sample2)
```

```
## [1] 8.031556
```

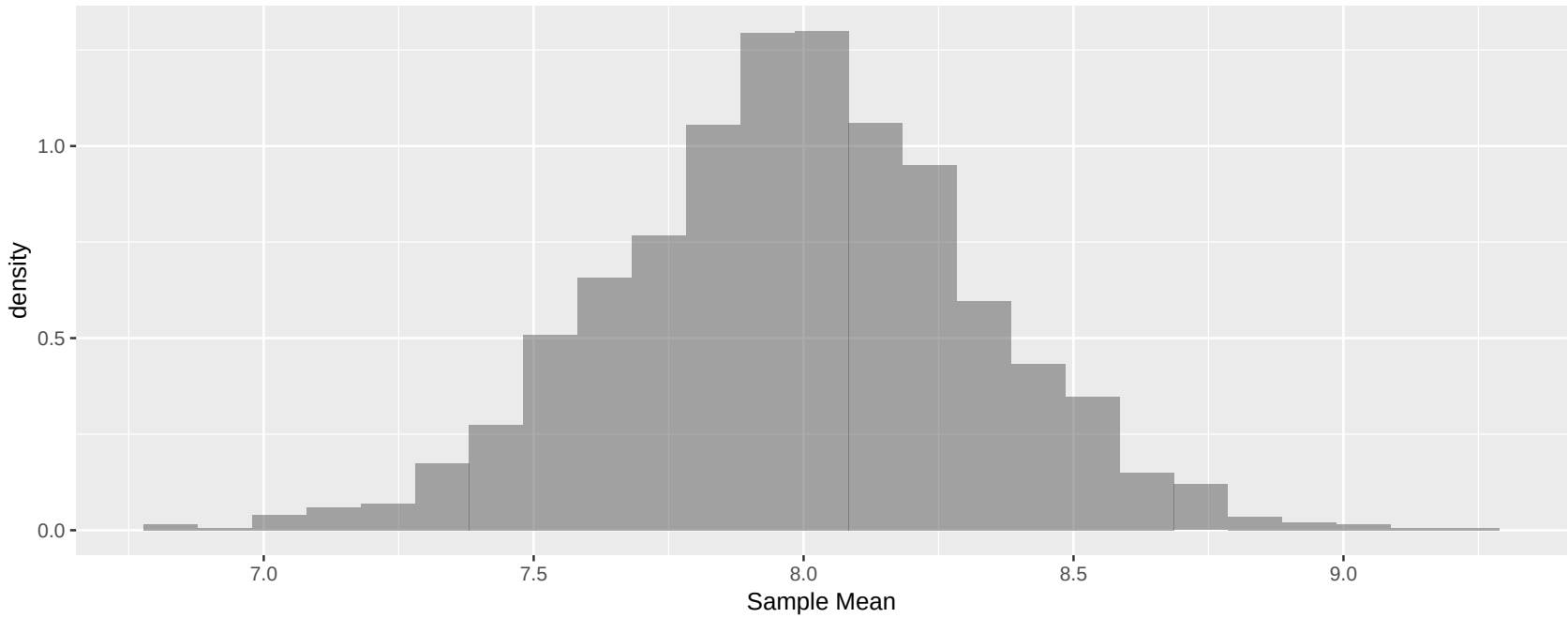
```
gf_histogram(~sample2)
```



# A Distribution of Estimates

I can think about taking many many samples from the population, to consider the distribution of means that are possible when sampling from a population with mean 8 and standard deviation 1.5.

```
SDoM <- do(2000)*mean(sample(Sleeppop, 20))
gf_dhistogram(~mean, data = SDoM, xlab = "Sample Mean")
```

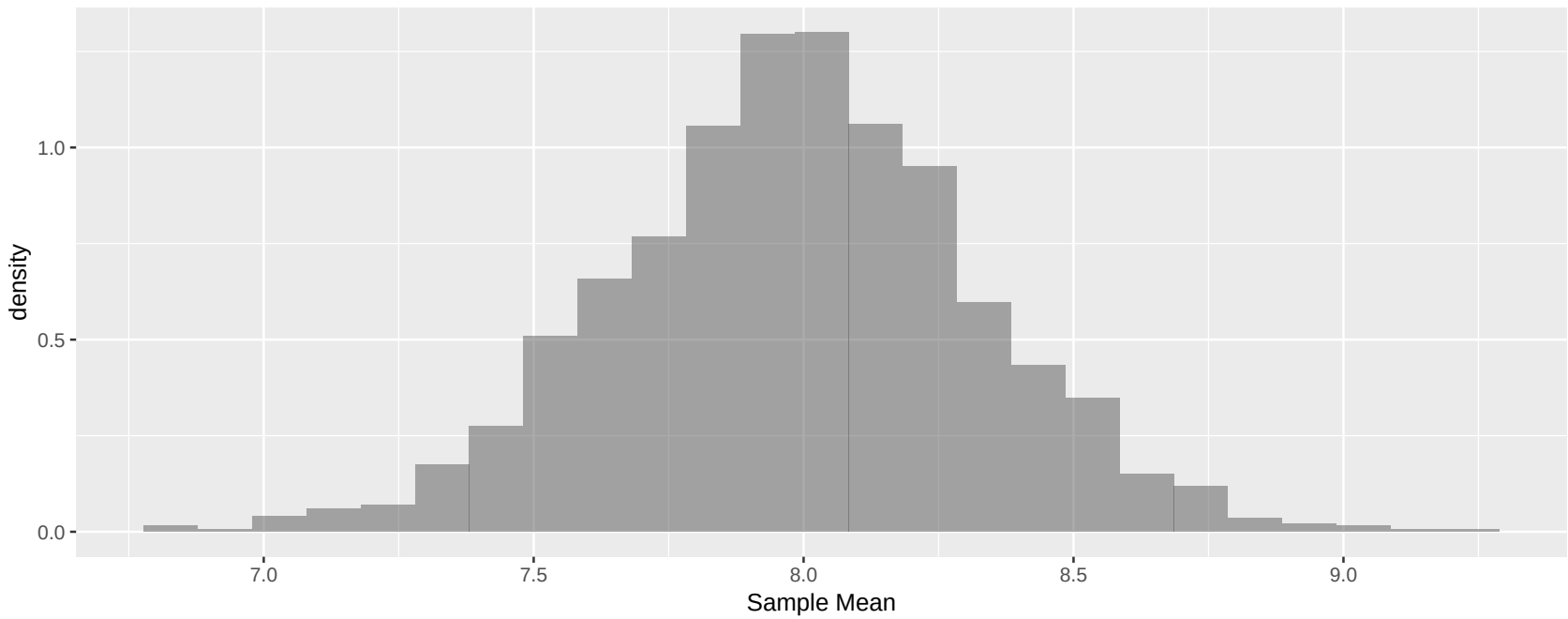


# What do we learn from the distribution?

Does it seem likely or unlikely to get a sample with a mean greater than 9?

Does it seem likely or unlikely to get a sample with a mean between 7 and 8?

```
gf_dhistogram(~mean, data = SDoM, xlab = "Sample Mean")
```



# Using a sampling distribution

Each sample provides us with our best guess about the population, but that guess can vary from sample to sample

**Sampling variability** is the idea that randomly sampling results in different estimates from each sample

**Sampling distributions** are our way of visualizing and quantifying that sampling variability

When we collect a sample, we can only see one observation from the sampling distribution. However we must always remember that this was **one of many possible estimates**.

# Sampling distributions and Z-scores

The purpose of a Z-score was to tell us how **unusual** an observation is. We compare the observation to the sample mean and scale by the standard deviation.

The sampling distribution is similarly used to tell us how unusual **a sample mean** is. We use the mean and standard deviation from the **sampling distribution** to tell us about our observed mean (from our sample).

The information we need for a Z-score is **mean and standard deviation**, let's see how we can get that information for the sampling distribution.

# To create a sampling distribution we need a DGP

Most of the time we don't know what the DGP is, so we'll need to imagine a DGP, and think about the implications for our sample.

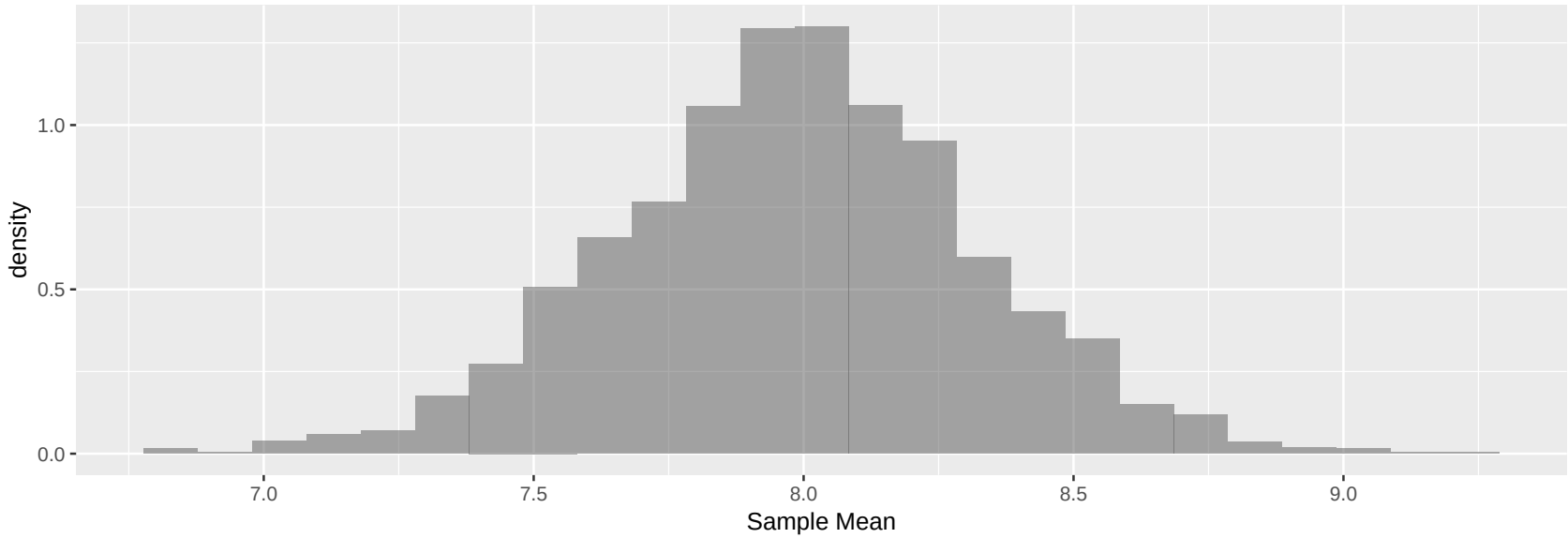
Remember at the beginning of class I said "Imagine we live in a (wonderful) world, where on average people sleep 8 hours ($\mu = 8$), and the population standard deviation of hours of sleep is 1.5 ($\sigma = 1.5$)." This is imagining a DGP.

I generated samples from the DGP, to tell me about the sampling distribution of the mean **if that DGP is true**. If that DGP is not true, the sampling distribution of the mean will be different.

# Other possible DGPs

We considered a world where $\mu = 8$, $\sigma = 1.5$. The sampling distribution for a sample of 20 looked like this:

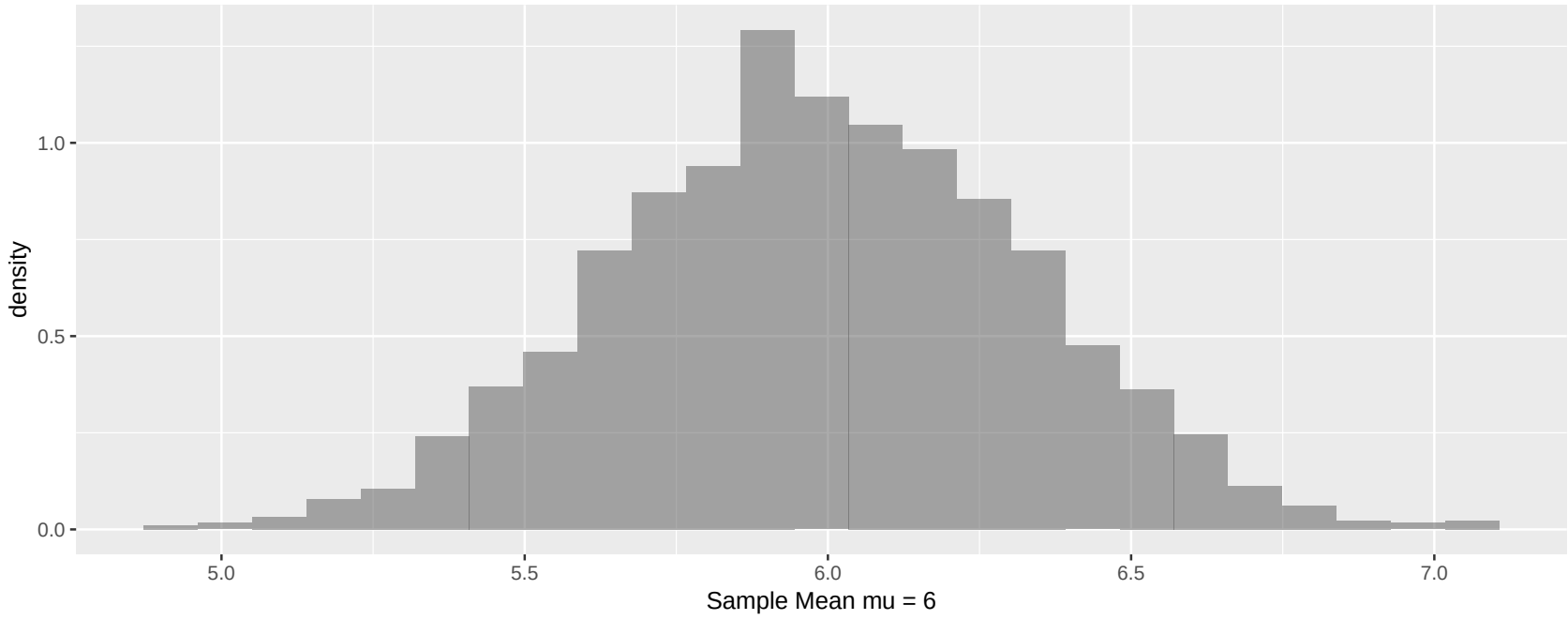
```
gf_dhistogram(~mean, data = SDoM, xlab = "Sample Mean")
```




What if we consider a different world: $\mu = 6$, $\sigma = 1.5$. How will the sampling distribution be different?

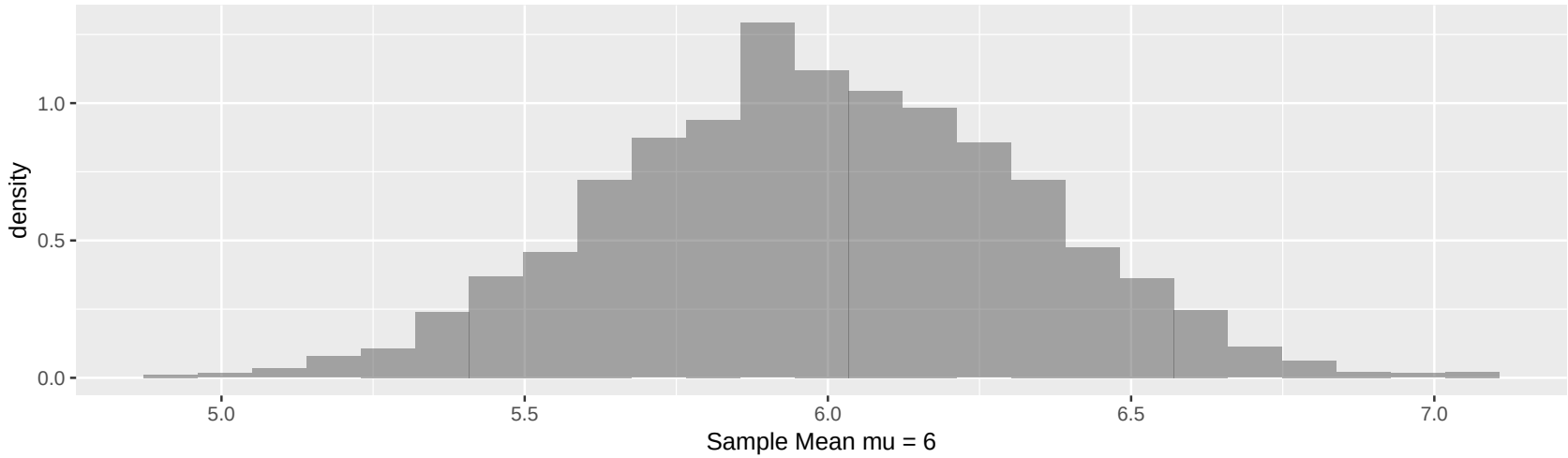
# Other possible DGPs

```
Sleeppop2 <- rnorm(1000000, 6, 1.5)
SDoM2 <- do(2000)*mean(sample(Sleeppop2, 20))
gf_dhistogram(~mean, data = SDoM2, xlab = "Sample Mean mu = 6")
```

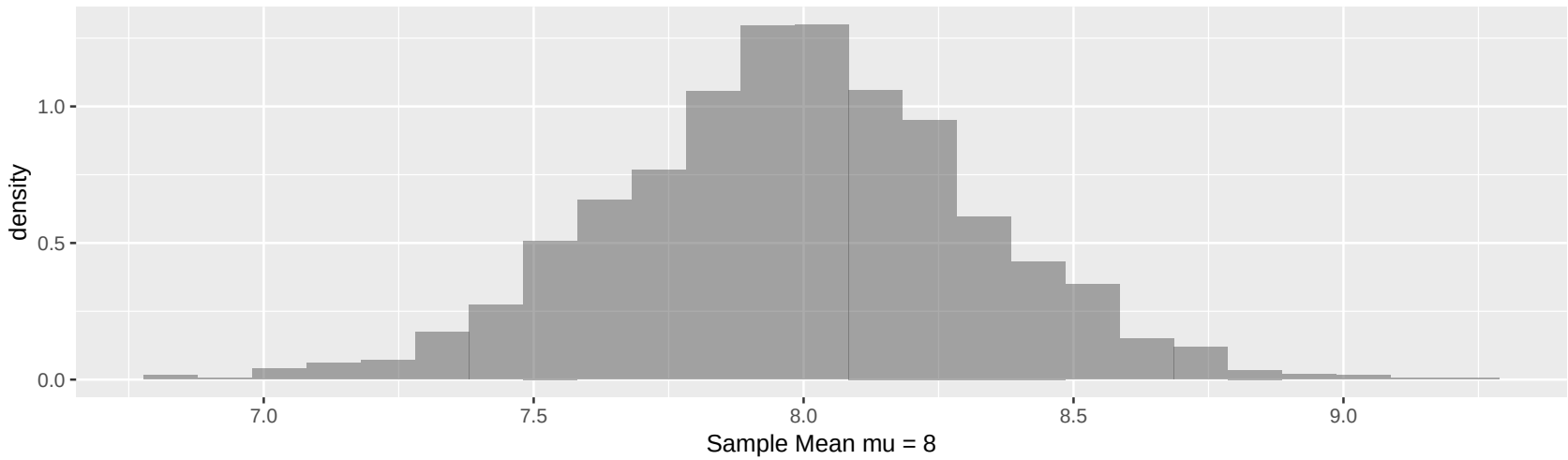


# Comparing the two sampling distributions

```
gf_dhistogram(~mean, data = SDoM2, xlab = "Sample Mean mu = 6")
```




```
gf_dhistogram(~mean, data = SDoM, xlab = "Sample Mean mu = 8")
```



# Other possible DGPs

Notice that the mean of the sampling distribution is pretty much equal to the mean of the population distribution.

```
mean(SDoM$mean)
```

```
## [1] 7.984728
```

```
mean(SDoM2$mean)
```

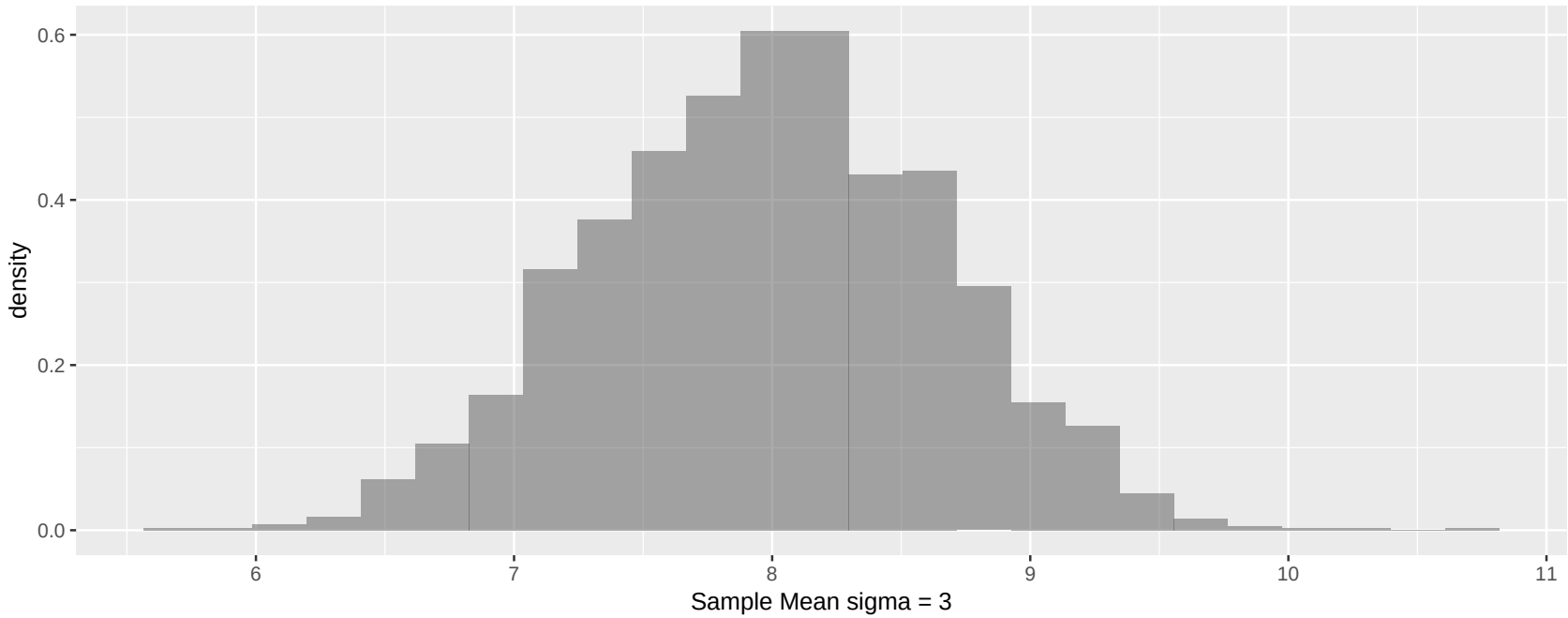
```
## [1] 5.987993
```

What if we consider a different world (with greater variance): $\mu = 8$, $\sigma = 3$. How will the sampling distribution be different?

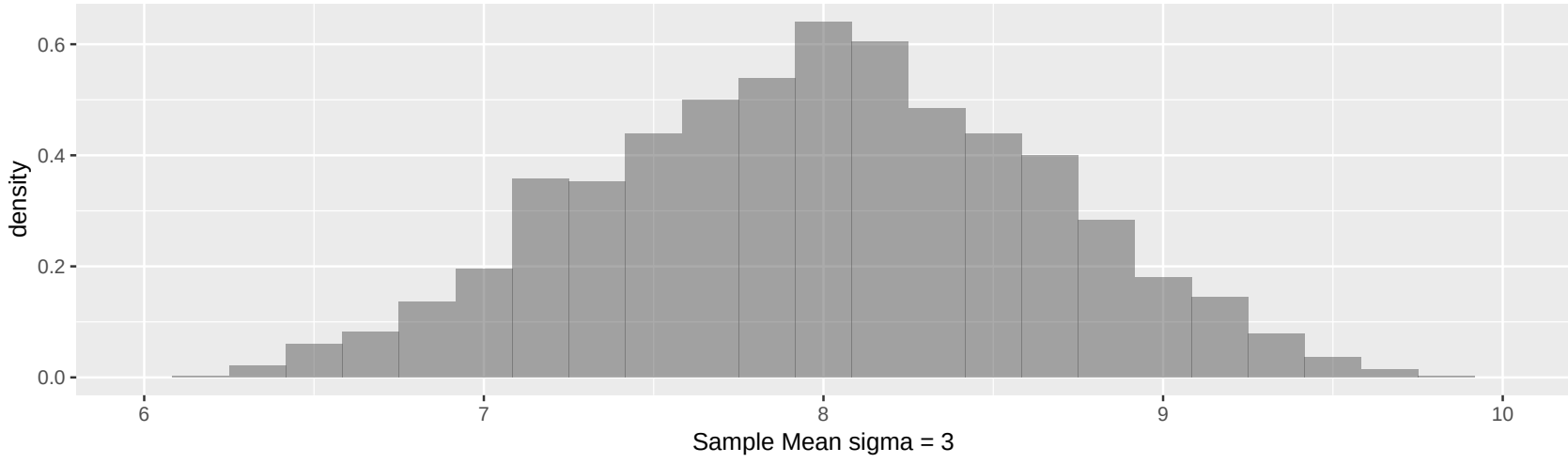
# Other possible DGPs

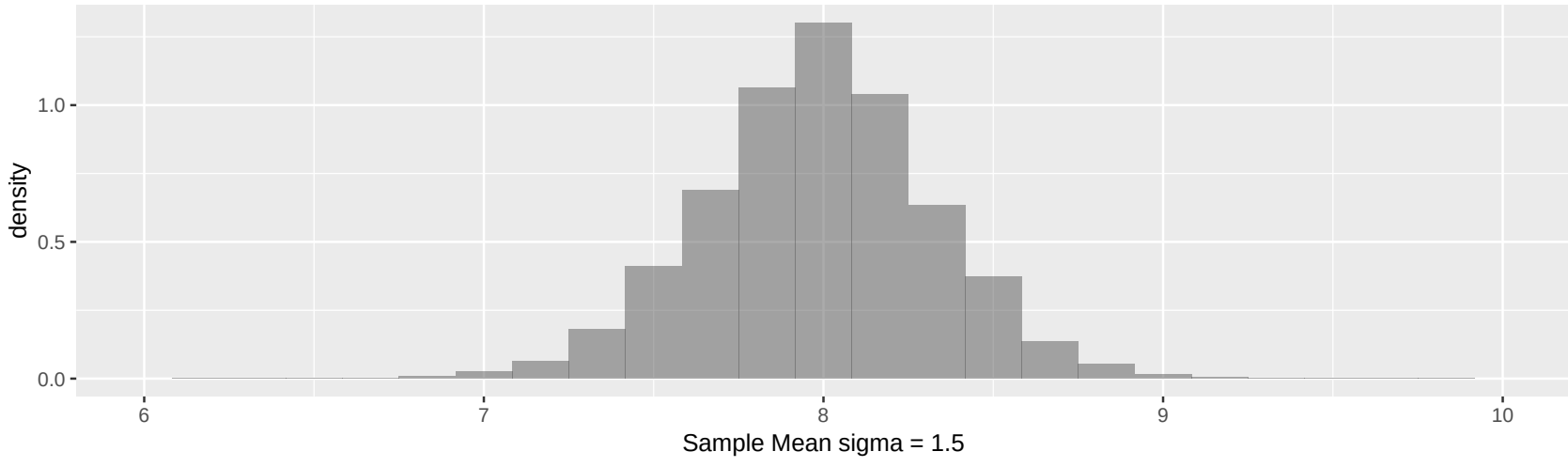
What if we consider a different world: $\mu = 8$, $\sigma = 3$. How will the sampling distribution be different? How will it be the same?

```
Sleeppop3 <- rnorm(1000000, 8, 3)
SDoM3 <- do(2000)*mean(sample(Sleeppop3, 20))
gf_dhistogram(~mean, data = SDoM3, xlab = "Sample Mean sigma = 3")
```



# Comparing sampling distributions



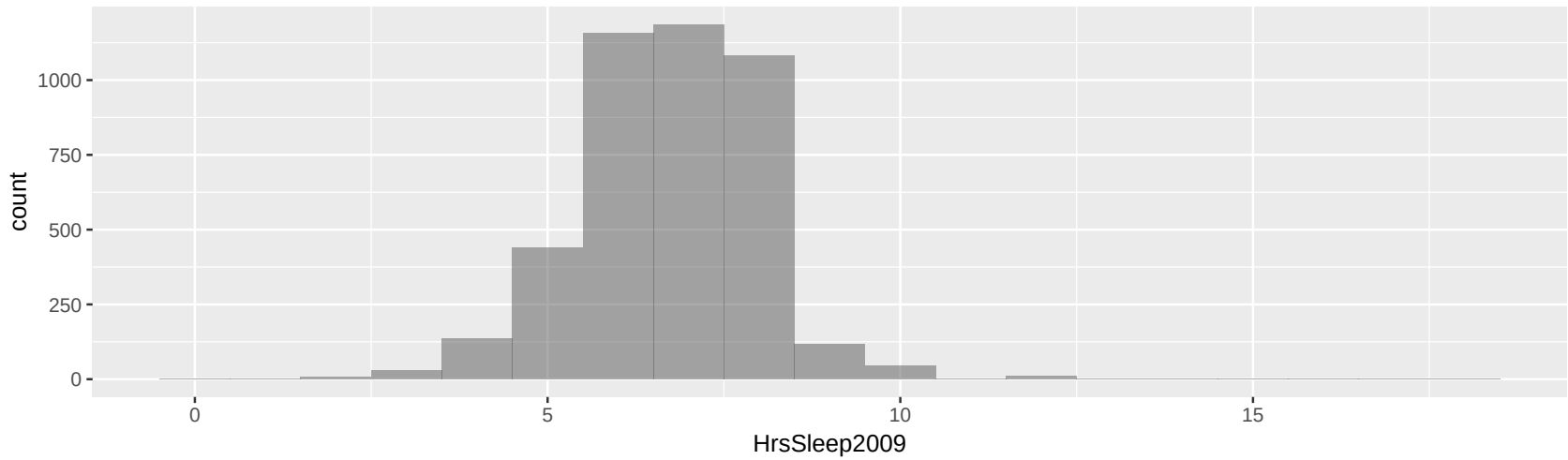
# Thinking backwards

What if I have a random sample of 4224 individuals and I wanted to know if it seems likely this sample came from a population where the average hours of sleep was 8 hours?

```
mean(NLSdata$HrsSleep2009)
```

```
## [1] 6.741951
```

```
gf_histogram(~HrsSleep2009, data = NLSdata, binwidth = 1)
```




We can use what we know about samples that come from a population DGP with mean 8 to evaluate whether this sample seems particularly unusual. Before we looked at samples with N = 20. Let's look again with N = 4224.

# Thinking backwards

Let's assume for a moment that the population standard deviation is pretty close to the estimated standard deviation.

```
sd(NLSdata$HrsSleep2009)
```

```
## [1] 1.305077
```

Let's say the population standard deviation is 1.5. Based on what we know about samples from a distribution with a mean of 8 and standard deviation of 1.5, does this seem like a likely sample from this distribution?

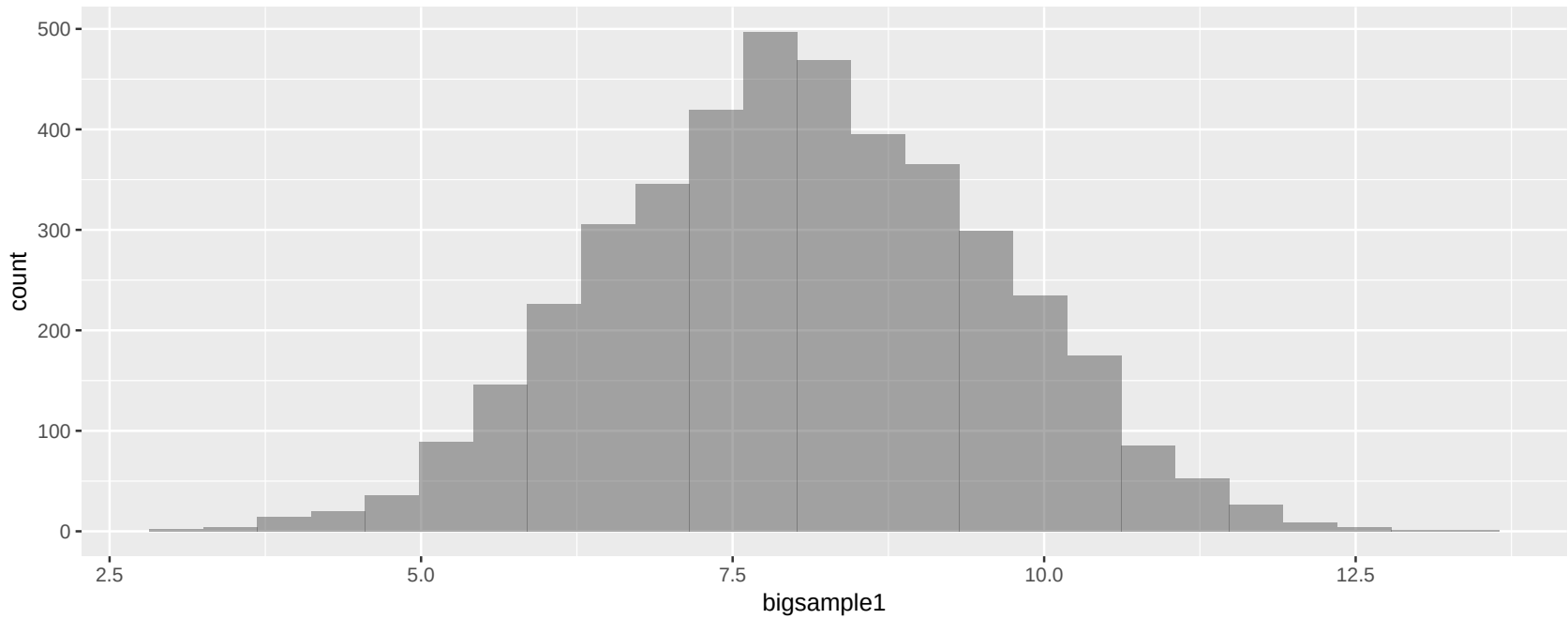
# Generating samples

If I took a random sample of 4224 people from the population of individuals, and asked how many hours they sleep and took the mean, we would expect that the mean would be close to 8, but perhaps not exactly 8.

```
bigsample1 <- sample(Sleeppop, 4224)
mean(bigsample1)
```

```
## [1] 8.049069
```

```
gf_histogram(~bigsample1)
```



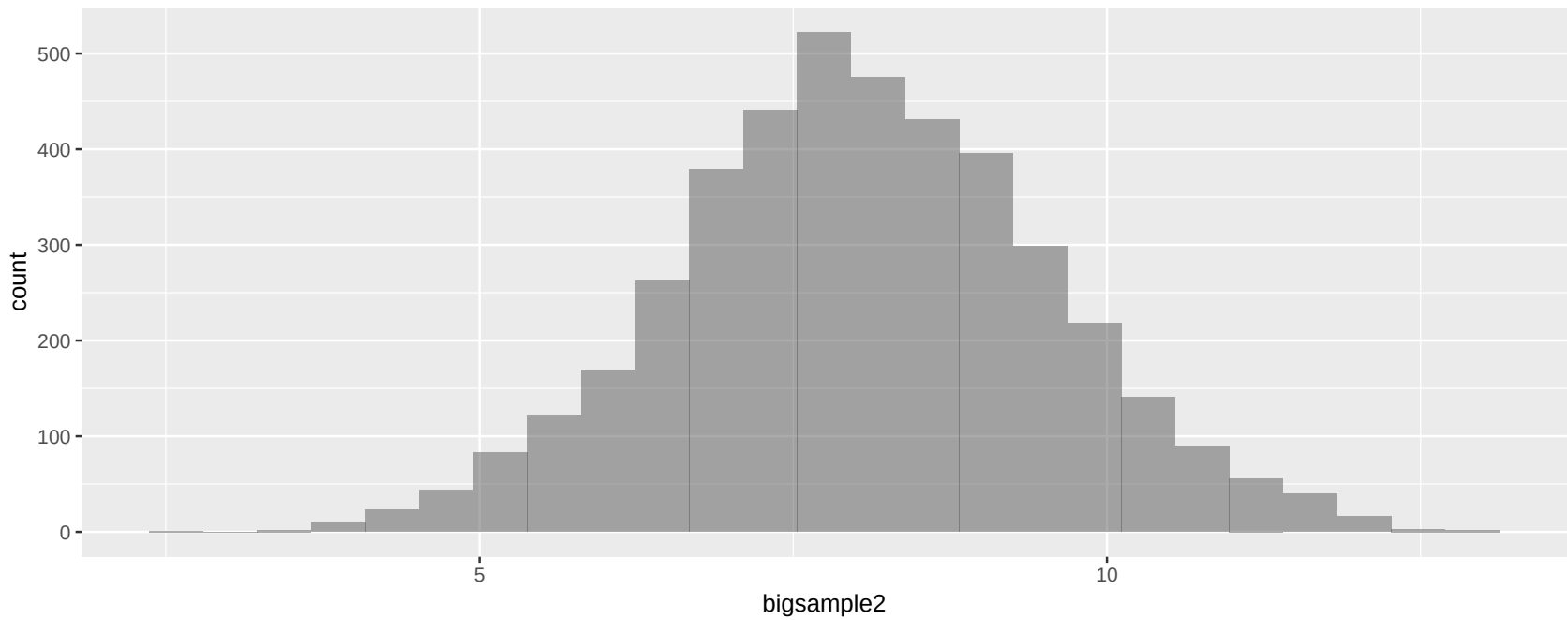
# Taking another sample

Similarly, if I took another random sample of 4224 people, the mean would be close to 8, not exactly 8, and not exactly the same as from the first sample.

```
bigsample2 <- sample(Sleeppop, 4224)
mean(bigsample2)
```

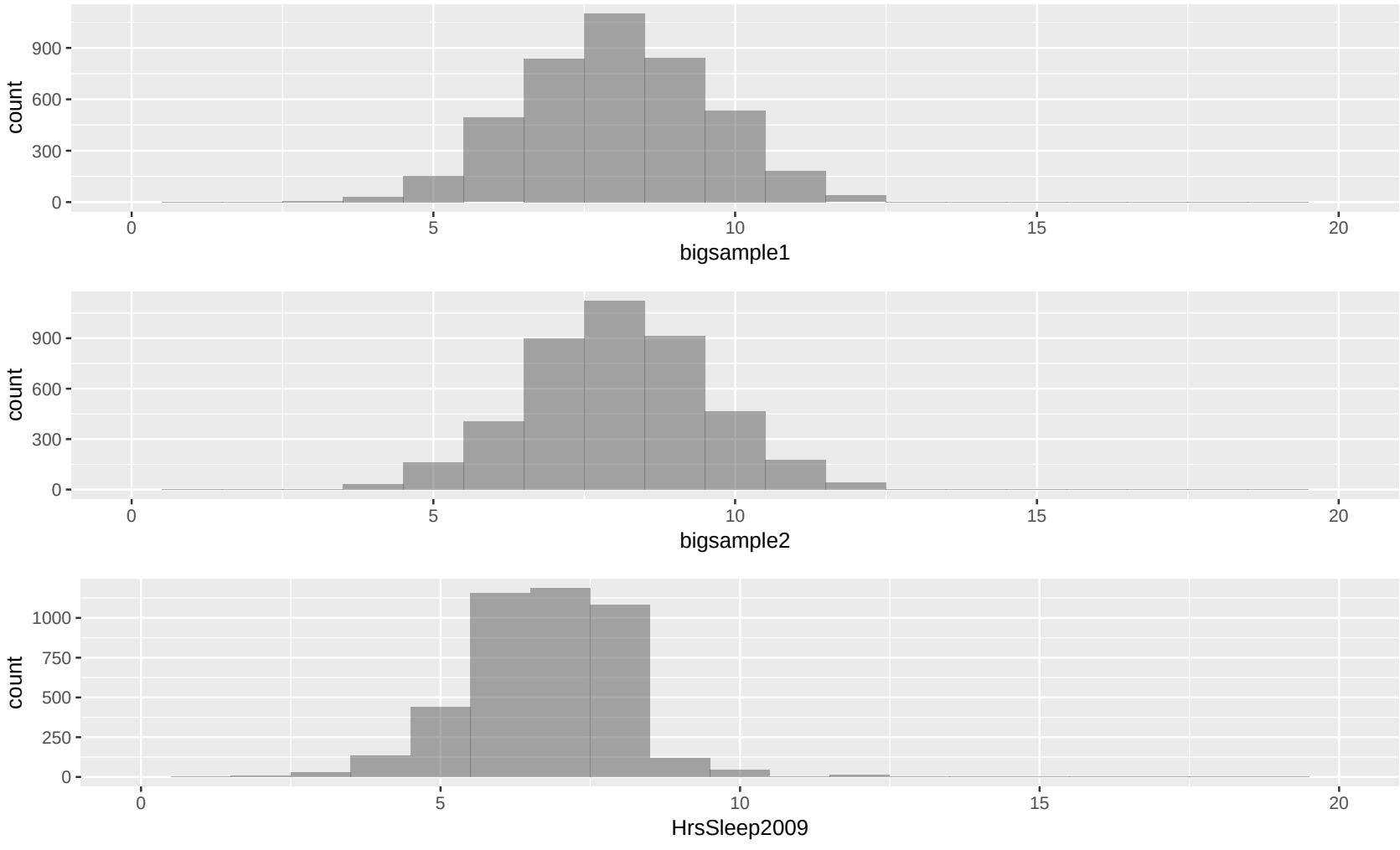
```
## [1] 8.049194
```

```
gf_histogram(~bigsample2)
```



# Comparing our sample to the generated samples

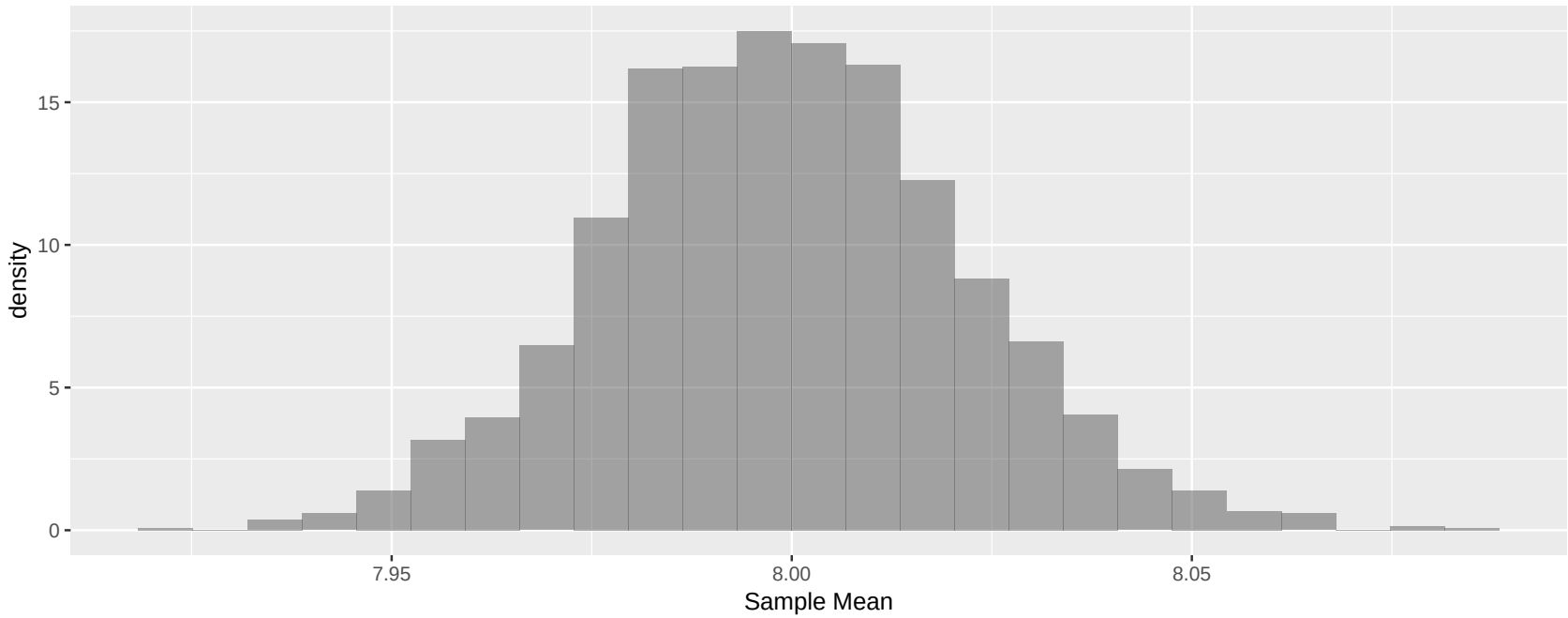
Does it look like our sample comes from a distribution with the same mean as the other distributions? How about the same standard deviation? How about the same size of distribution? Discuss with your neighbors.



# A Distribution of Estimates

We can take many many samples from the population, to consider the distribution of means that are possible when sampling from a population with mean 8 and standard deviation 1.5.

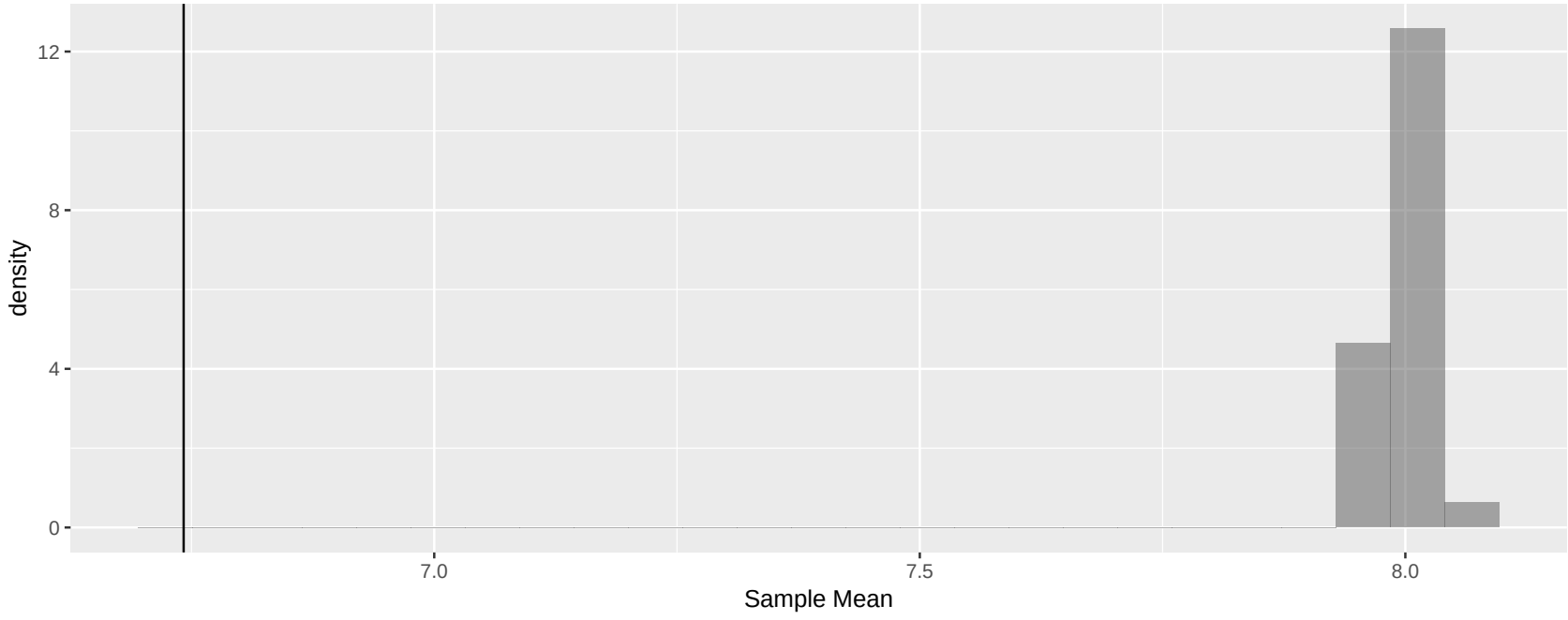
```
bigSDoM <- do(2000)*mean(sample(Sleeppop, 4224))
gf_dhistogram(~mean, data = bigSDoM, xlab = "Sample Mean")
```



# Comparing observed sample to sampling distribution

Where does the observed sample fall on this distribution?

```
gf_dhistogram(~mean, data = bigSDoM, xlab = "Sample Mean")%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```



# Quantifying likelihood

What proportion of observed sample means from the DGP ($\mu = 8$, $sd = 1.5$) are less than our observed mean from the NLSdata (6.7419508)?

```
tally(~(mean<mean(NLSdata$HrsSleep2009)), data = bigSDoM)
```

```
## (mean < mean(NLSdata$HrsSleep2009))
##  TRUE FALSE
##     0  2000
```

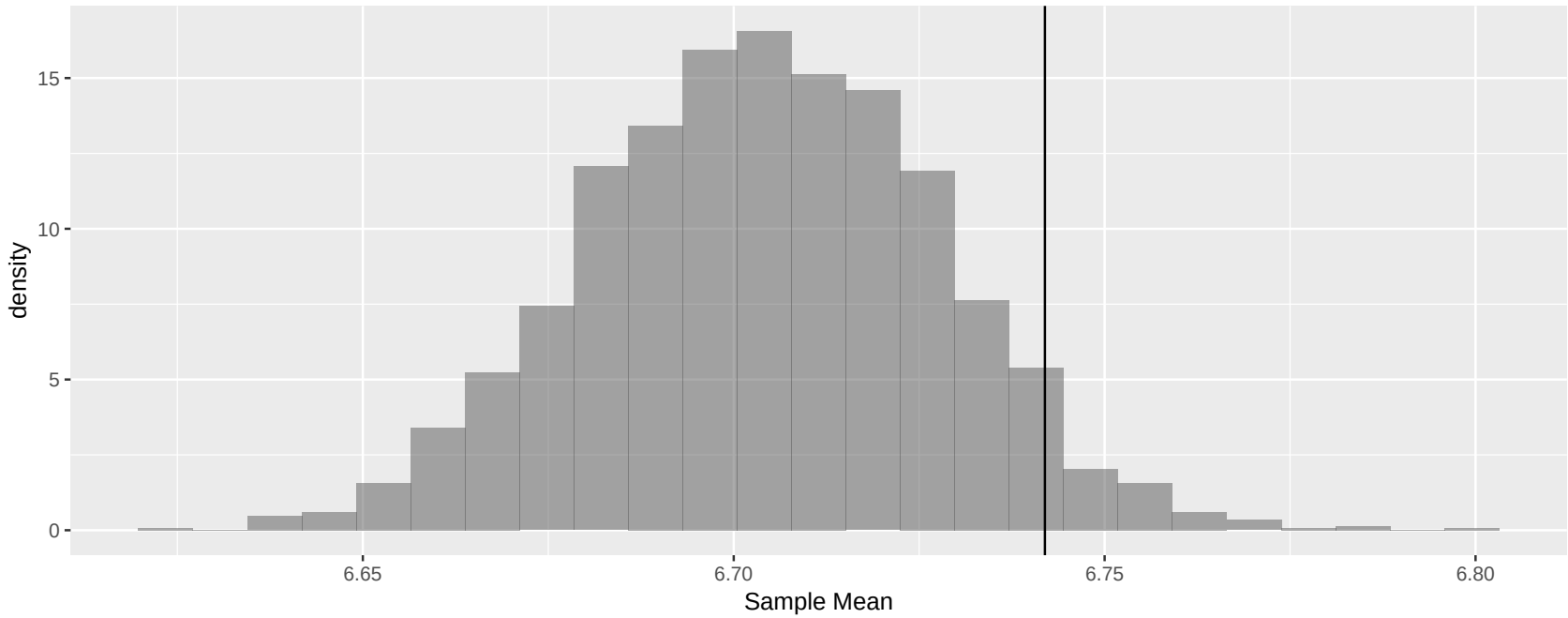
Not a single simulated sample mean was as low or lower than our observed mean!

Our observed mean seems pretty unlikely under this specific DGP

# Imagine a different DGP

Let's consider a DGP with a smaller mean (e.g., $\mu = 6.7$)

```
Sleeppop67 <- rnorm(1000000, mean = 6.7, sd = 1.5)
SDoM67 <- do(2000)*mean(sample(Sleeppop67, 4224))
gf_dhistogram(~mean, data = SDoM67, xlab = "Sample Mean")%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```



# Quantifying Likelihood

```
tally(~(mean<mean(NLSdata$HrsSleep2009)), data = SDoM67)
```

```
## (mean < mean(NLSdata$HrsSleep2009))
##  TRUE FALSE
##  1897   103
```

```
tally(~(mean<mean(NLSdata$HrsSleep2009)), data = SDoM67,
      format = "proportion")
```

```
## (mean < mean(NLSdata$HrsSleep2009))
##   TRUE  FALSE
## 0.9485 0.0515
```

Even for a DGP mean which is **pretty close** to our sample mean, getting a sample mean that is 6.74 or bigger is pretty unusual. Why do you think that is?

# The role of sample size in sampling distributions

The spread of the sampling distribution is influenced by the size of the sample.

In fact, we can directly calculate the standard deviation of the sampling distribution based on the standard deviation of the population distribution $\sigma$ **and the sample size** ($N$).

We call the standard deviation of the sampling distribution **standard error**(of the mean) and use the notation $\sigma_{\bar{Y}}$.

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$$

# Breaking down this equation

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$$

If $\sigma$ increases, and $N$ stays the same, $\sigma_{\bar{Y}}$ will _____.

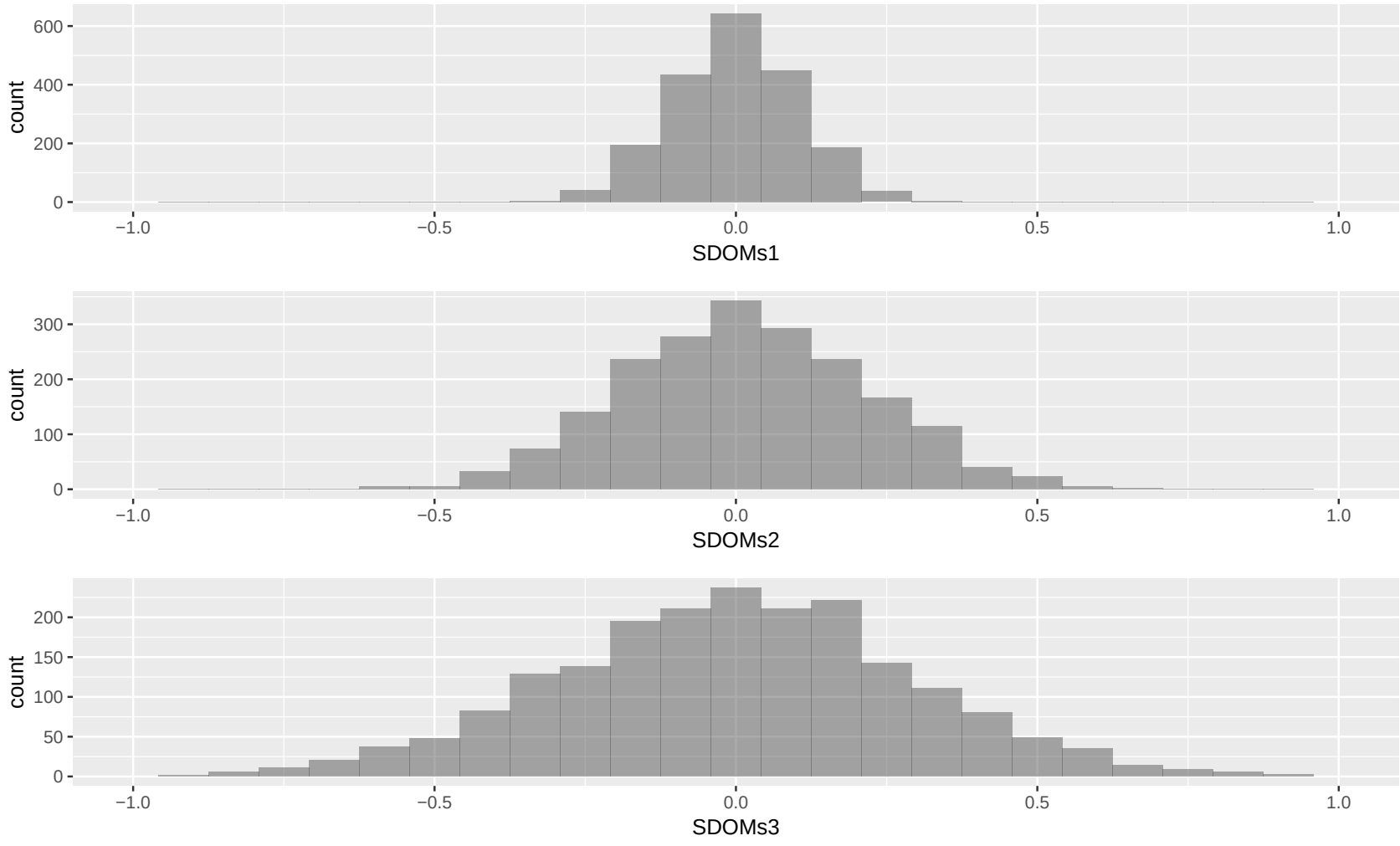
This means, if I'm comparing the sampling distribution of the mean from two variables and the same sample size, the one with the smaller population standard deviation will have a _____ distribution.

# Visualizing the change

Let's look at sampling distributions from 3 different population standard deviations, and the same sample size (N = 100)

```
SDOMs1 <- do(2000)*mean(rnorm(100, mean = 0, sd = 1))
SDOMs2 <- do(2000)*mean(rnorm(100, mean = 0, sd = 2))
SDOMs3 <- do(2000)*mean(rnorm(100, mean = 0, sd = 3))
```

# Visualizing the change



# Comparing the Simulation to Equation

```r
#sigma = 1, N = 100
sd(SDOMs1$mean)
```

```
## [1] 0.103365
```

```r
1/sqrt(100)
```

```
## [1] 0.1
```

```r
#sigma = 2, N = 100
sd(SDOMs2$mean)
```

```
## [1] 0.2010556
```

```r
2/sqrt(100)
```

```
## [1] 0.2
```

```r
#sigma = 3, N = 100
sd(SDOMs3$mean)
```

```
## [1] 0.2947739
```

```r
3/sqrt(100)
```

# Breaking down this equation

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$$

If $\sigma$ stays the same, and $N$ increases, $\sigma_{\bar{Y}}$ will _____.

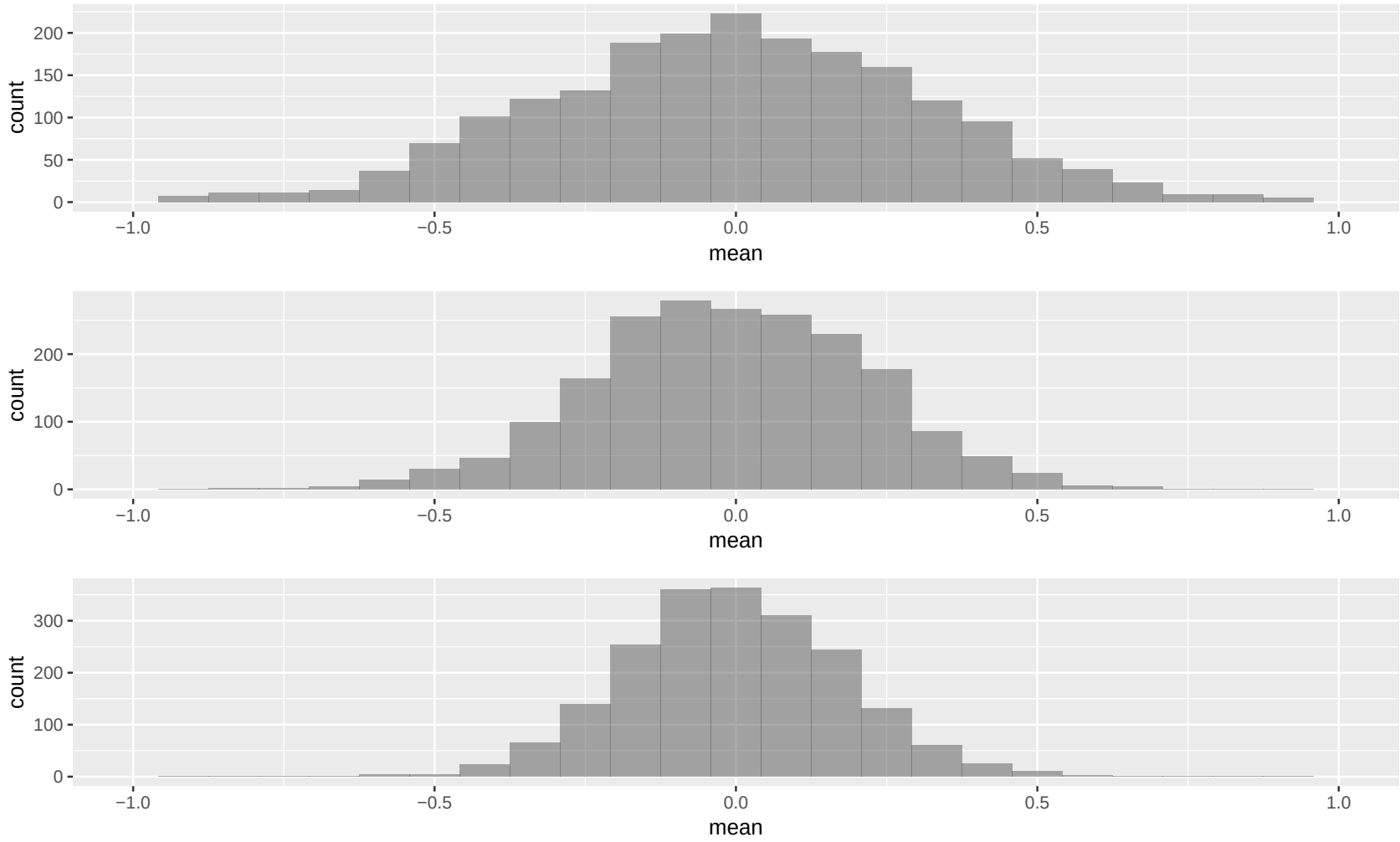
This means, if I'm comparing the sampling distribution of the means from two studies of the same variable (same $\sigma$), the one with the smaller sample will have a _____ distribution.

# Visualizing the change

Let's look at sampling distributions from 3 different population standard deviations, and the same sample size (N = 100)

```
SDOMN10 <- do(2000)*mean(rnorm(10, mean = 0, sd = 1))
SDOMN20 <- do(2000)*mean(rnorm(20, mean = 0, sd = 1))
SDOMN30 <- do(2000)*mean(rnorm(30, mean = 0, sd = 1))
```

# Visualizing the change



# Comparing the Simulation to Equation

```r
#sigma = 1, N = 10
sd(SDOMN10$mean)
```

```
## [1] 0.3166566
```

```r
1/sqrt(10)
```

```
## [1] 0.3162278
```

```r
#sigma = 1, N = 20
sd(SDOMN20$mean)
```

```
## [1] 0.2257526
```

```r
1/sqrt(20)
```

```
## [1] 0.2236068
```

```r
#sigma = 1, N = 30
sd(SDOMN30$mean)
```

```
## [1] 0.1772286
```

```r
1/sqrt(30)
```

# Wrapping Up

- Sampling distributions are a representation of what a statistic (so far we've looked at the mean) might look like when many random samples of the same size are drawn from the same data generating process.
  - The mean of a sampling distribution of a mean $\mu_{\bar{Y}}$ is equal to the population mean $\mu$
  - The standard deviation of a sampling distribution of a mean $\sigma_{\bar{Y}}$ is equal to the popululation standard deviation scaled by the square root of the sample size $\sigma/\sqrt{N}$
- Sampling distributions are used to help us assess the likelihood of an observed statistic under different data generating processes. Given a DGP we can evaluate if this sample is usual or unusual (likely or unlikely).
- We can use simulation, `do()` to create sampling distributions of any statistic we might want to learn about!

# Thursday 5/16

```
NLSdata <- read.csv("http://bit.ly/NLSdata", header = TRUE)
```

# Learning Outcomes

- Build a mathematical understanding of likely and unlikely outcomes
- Define a confidence interval
- Compute a confidence interval for a mean using simulation
- Describe two interpretations of a confidence interval

# What is likely? What is unlikely?

We as people have an intuition about what we think is likely and what is unlikely to happen. That perception may vary across people. It also varies across situations.

At what point would you say something is likely?


Figure 1:

# Statistically likely

Much of statistics relies on "If. . . " thinking, and categorizing outcomes as "unlikely" under certain hypotheses.

Statisticians need to agree on an approximate range of unlikely that they are willing to accept, and this should not differ much across situations for the purpose of science (there is still some variability).

In general, psychology (and other behavioral science fields) generally works on a norm of unlikely being about 5%.

Medical research/Neuroscience is held to a higher standard (with lives and money at stake) 1% or .1%

Exploratory or new research lines have a lower standard 10%

## Splitting Up 5%

Typically, in psychology we're curious about unlikely outcomes in both directions (too high or too low), so we split the 5% on either side of the distribution.

Any observations in the outer-most 5% of the distribution are considered unlikely.

Just because an observation is classified as unlikely doesn't mean it's **impossible** just **improbable**.

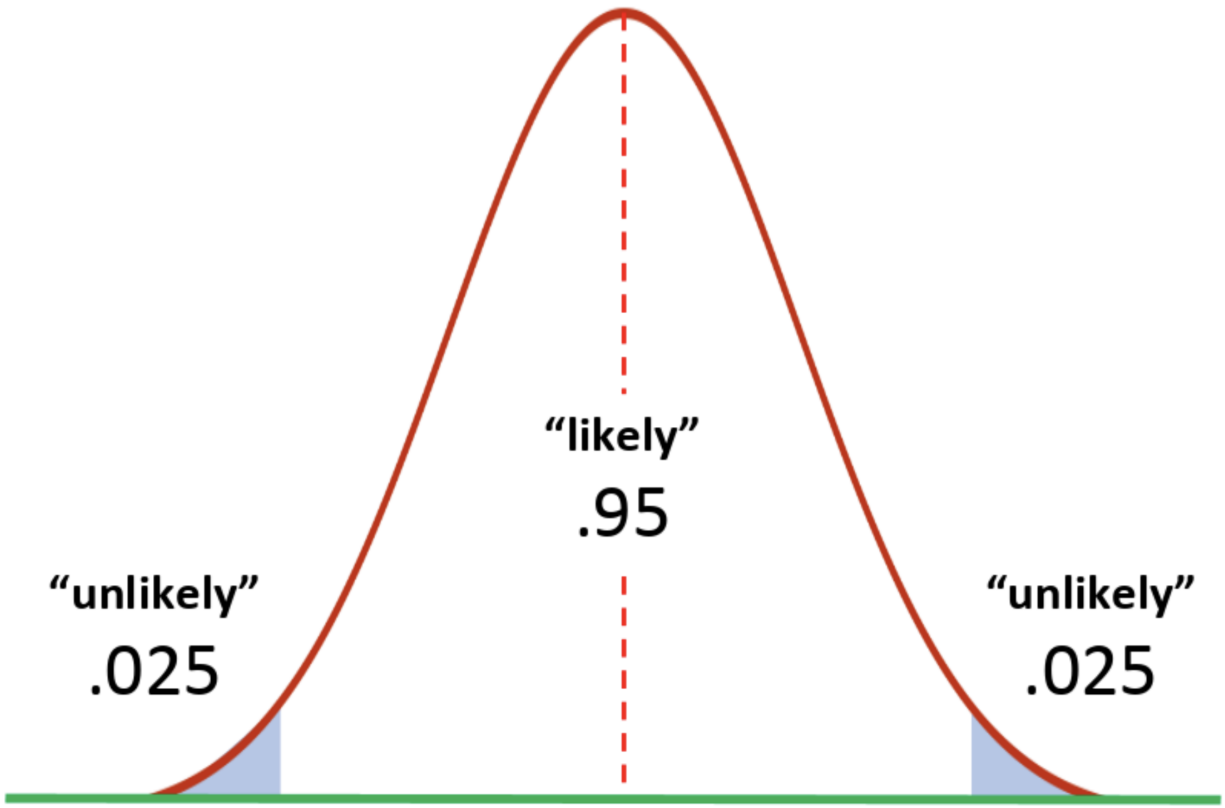


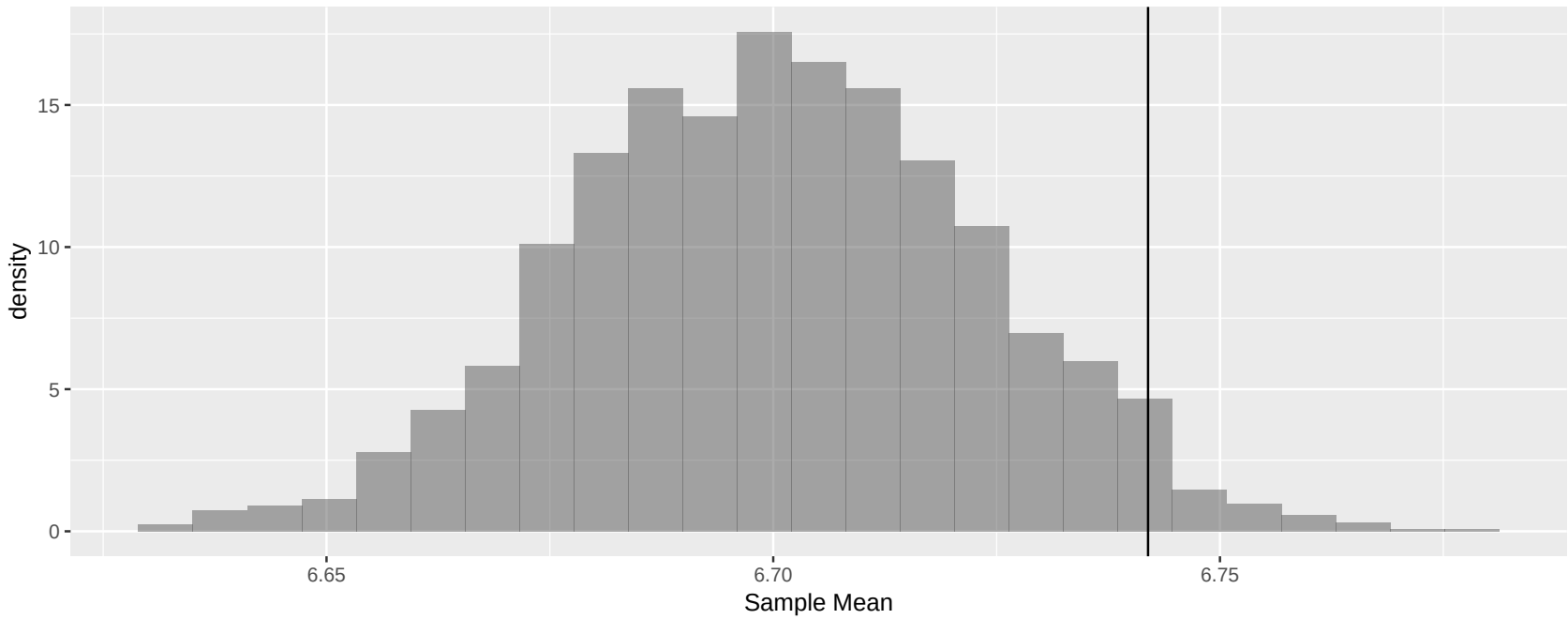

Figure 2:

# Example of likelihood

Last time we discussed the likelihood of observing our NLS sample mean for HrsSleep2009 (6.74) **IF** the population mean is 6.7 and the population standard deviation is 1.5.

Is getting a sample mean of 6.74 or larger a likely or unlikely event?

```
set.seed(12345)
Sleeppop67 <- rnorm(1000000, mean = 6.7, sd = 1.5)
SDoM67 <- do(2000)*mean(sample(Sleeppop67, 4224))
gf_dhistogram(~mean, data = SDoM67, xlab = "Sample Mean")%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```



# Quantifying Likelihood

Is getting a sample mean of 6.74 or larger a likely or unlikely event?

We can use the tally function to calculate the proportion of simulated means greater than the observed mean.

Using our standard of 5% we would say this is an unlikely event. (Using a standard of 2.5% we would say it's likely)

```
tally(~(mean<mean(NLSdata$HrsSleep2009)), data = SDoM67,
      format = "proportion")
```

```
## (mean < mean(NLSdata$HrsSleep2009))
##   TRUE  FALSE
## 0.9645 0.0355
```

# Which events are likely (or not unlikely)?

We can use our standard of 5% (2.5% on each side of the distribution) to come up with a set of sample means which seem likely under a given DGP.

First we sort the means. Then we look at the bottom 2.5% and the top 2.5% of our simulated means and the top 2.5%.

```
SDoM67 <- arrange(SDoM67, desc(mean))
cbind(head(SDoM67), tail(SDoM67))
```

```
##       mean     mean
## 1 6.775199 6.637430
## 2 6.774652 6.637258
## 3 6.764896 6.637043
## 4 6.763948 6.633733
## 5 6.763598 6.632002
## 6 6.762983 6.629039
```

# Which events are likely (or not unlikely)?

Next we look at the bottom 2.5% and the top 2.5% of our simulated means.

I have 2000 means, so the top 2.5% is 2000*0.025 = 50 and the bottom 2.5% is 2000*.975=1950

```
SDoM67$mean[50]
```

```
## [1] 6.743872
```

```
SDoM67$mean[1950]
```

```
## [1] 6.655068
```

Based on these results what do we know?

# Probability

Probability is defined as a long term frequency. After many many trials, what proportion of outcomes would be a certain way.

If I flip a coin many many times the proportion which end up heads will be 0.5.

If I draw a card from a deck of 52 cards repeatedly, the proportion which end up spades will be 0.25.

In order for there to be a probability there must be a random process which generates the outcome.

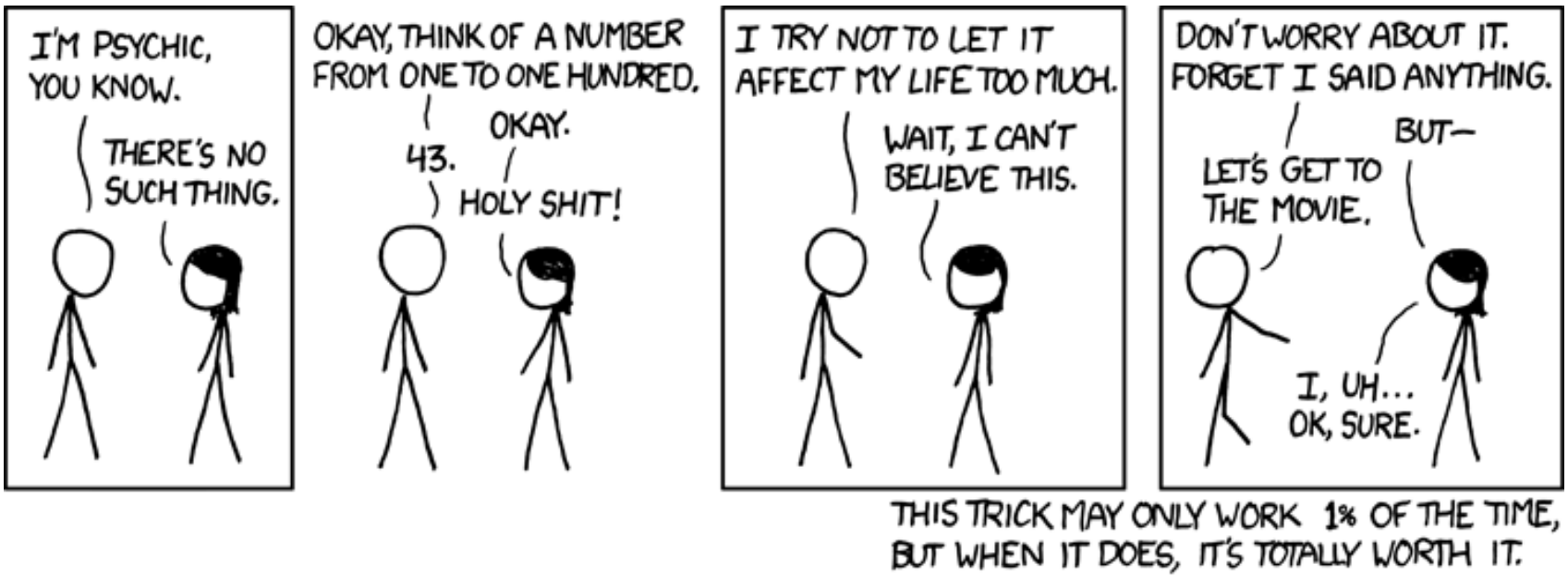



Figure 3:

# Making statements about population means

Usually we **want** to think about the population/DGP instead of possible sample means.

We already have an observed mean, but we want to think about the possible population means which could have generated our data.

I want to say "There is a 95% probability that the population mean is between 6.55068 and 6.743872"

But! The population mean is not a probabilistic value! It is a fixed but unknown thing!

# The Fixed Unknown

What is the probability that a flip of a coin results in a head?

Once I've flipped a coin, but you have not seen the outcome, what is the probability that it is a head?

If I randomly choose someone from the class, what is the probability that their birthday is in May?

Once I choose the person, what is the probability that their birthday is in May?

# The Fixed Unknown

After the random process (or if there is no random process) has occured, there is no probability.

The coin is either heads or tails, we just don't know what it is.

An individual's birthday either is or is not in May, we just don't know.

The population mean either is or is not between 6.55068 and 6.743872, we just don't know.

Population parameters are a fixed but unknown thing, so we can't make probability statements about them.

# Making statements about our data

We can make probability statements about our data, based on "if" statements about the means.

We can say "If the population mean is this. . . . the probability of getting my result or something like it is 95%"

We can also define the range of population means for which our data is at least 95% likely, this is a confidence interval.

# Confidence Intervals

The purpose of confidence intervals are to identify a range of **population means** from which are data are considered likely.

Note here that the *likelihood* is attached to the data, not the population means!

```
mean(NLSdata$HrsSleep2009)
```

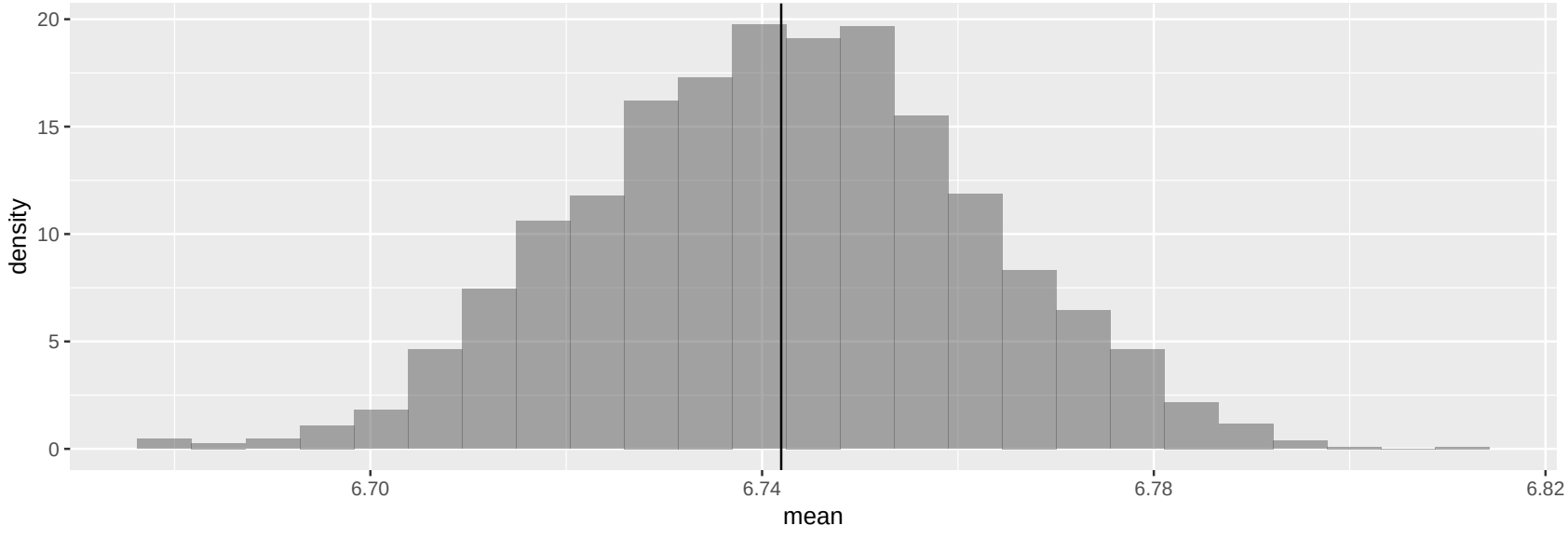
What population means do you think could have generated this sample?

One guess is the mean we observed, if the population mean is 6.7419508 then getting samples like our sample would be very probable.

# Confidence Intervals

If we consider a population with a mean equal to the observed mean, then our data (and other datasets like our data) seem very likely.

```
#make a population with the observed mean and sd from the data
Sleeppop4 <- rnorm(1000000, mean = mean(NLSdata$HrsSleep2009), sd = sd(
#create a set of samples from that population to consider sampling vari
SDoM4 <- do(2000)*mean(sample(Sleeppop4, 4224))
#Visualize observed sample vs. sampling distribution
gf_dhistogram(~mean, data = SDoM4)%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```



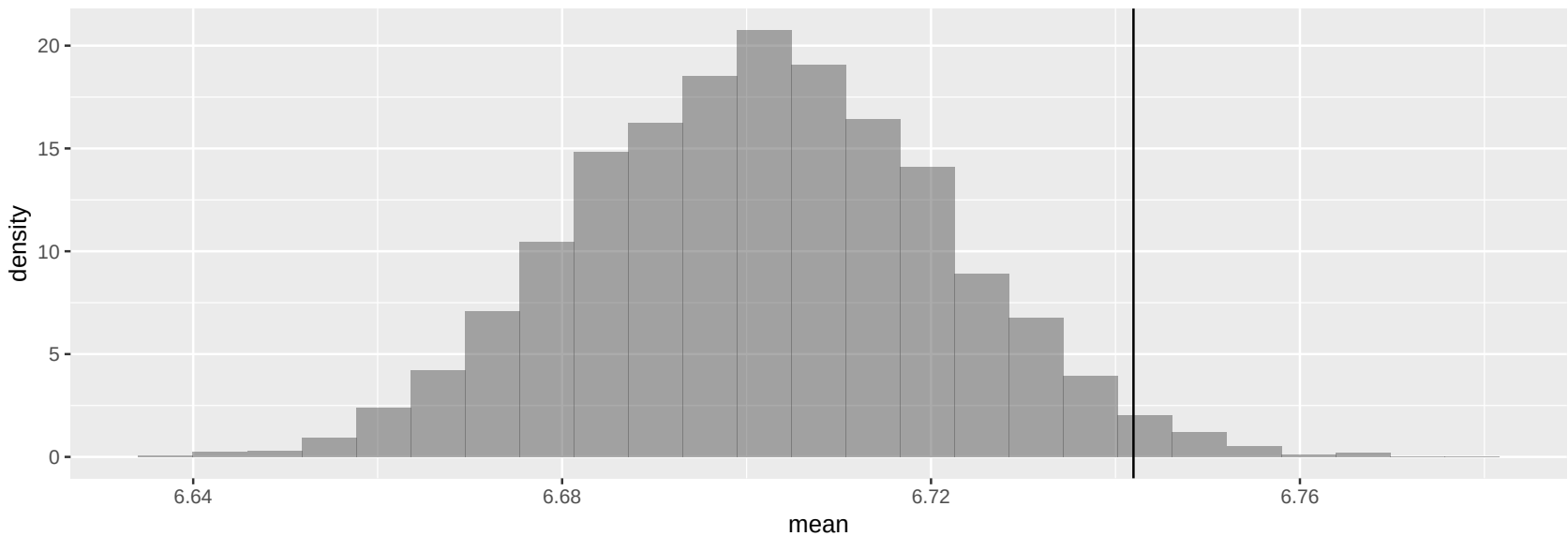
# Defining the edges

Let's consider some perhaps less perfect options

```
Sleeppop5 <- rnorm(1000000, mean = mean(NLSdata$HrsSleep2009)-0.0393,
                   sd = sd(NLSdata$HrsSleep2009))
SDoM5 <- do(4000)*mean(sample(Sleeppop5, 4224))
tally(~(mean > mean(NLSdata$HrsSleep2009)), data = SDoM5, format = "pro
```

```
## (mean > mean(NLSdata$HrsSleep2009))
##   TRUE  FALSE
## 0.0195 0.9805
```

```
gf_dhistogram(~mean, data = SDoM5)%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```

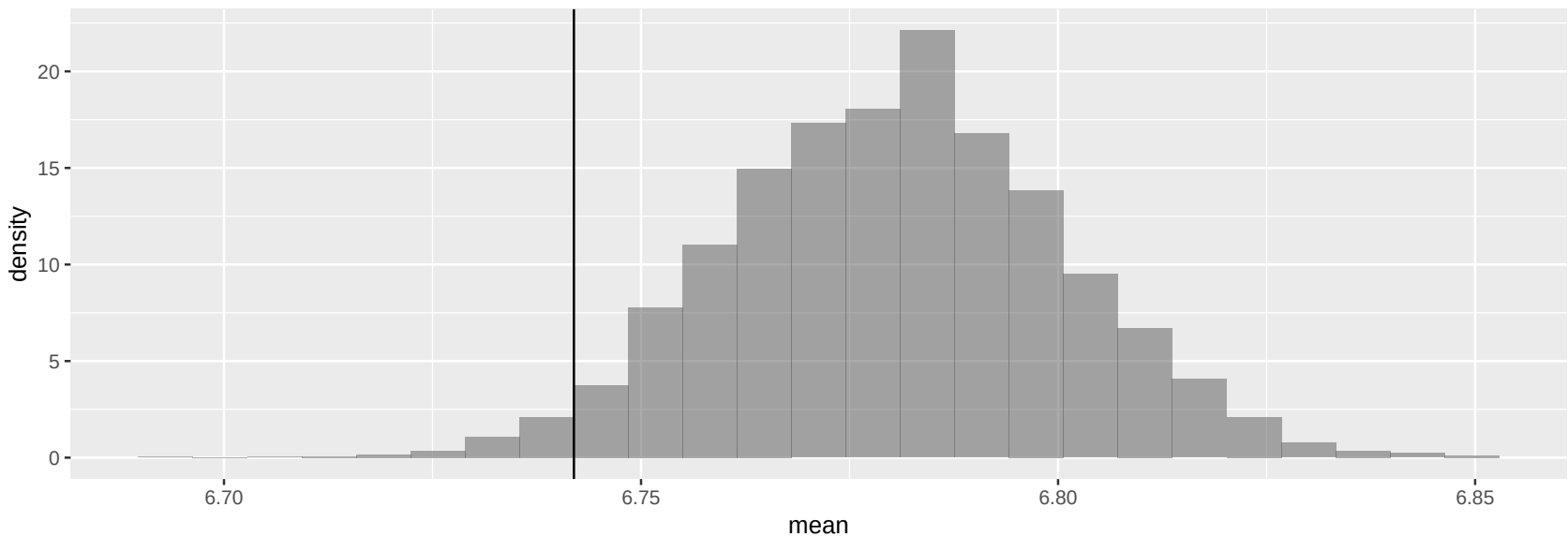


# Defining the edges

```
Sleeppop6 <- rnorm(1000000, mean = mean(NLSdata$HrsSleep2009)+0.0393,
                   sd = sd(NLSdata$HrsSleep2009))
SDoM6 <- do(4000)*mean(sample(Sleeppop6, 4224))
tally(~(mean < mean(NLSdata$HrsSleep2009)), data = SDoM6, format = "pro
```

```
## (mean < mean(NLSdata$HrsSleep2009))
##  TRUE FALSE
## 0.025 0.975
```

```
gf_dhistogram(~mean, data = SDoM6)%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```



# Likely or unlikely outcomes

Use the below code to try a population mean outside of the range 6.7026508 - 6.7812508. Is the observed sample mean a likely or unlikely outcome for this population mean? Now try a population mean inside the range, is the observed sample mean likely or unlikely?

Write an open response about what you found.

```
#make a population with the high mean and sd from the data
popmean <-
Sleeppop7 <- rnorm(1000000, mean = popmean,
                   sd = sd(NLSdata$HrsSleep2009))
#create a set of samples to consider sampling variability
SDoM7 <- do(4000)*mean(sample(Sleeppop7, 4224))
#tally sample means greater than or less than observed mean
tally(~(mean > mean(NLSdata$HrsSleep2009)), data = SDoM7,
      format = "proportion")
tally(~(mean < mean(NLSdata$HrsSleep2009)), data = SDoM7,
      format = "proportion")
#Visualize observed sample vs. sampling distribution
gf_dhistogram(~mean, data = SDoM7)%>%
  gf_vline(xintercept = mean(NLSdata$HrsSleep2009))
```

# Confidence Intervals

The goal of a confidence interval is to define an **upper and lower limit** of the population mean, for which the observed data seems likely (or not too unlikely).

Populations means between 6.7026508 and `mean(NLSdata$HrsSleep2009)+0.0393` seem like the edges of population means for which are data is likely, this means that our data is unlikely for means outside of this range.

Notice that 8 is not included in this range. If we live in a world where the average hours of sleep is 8, the data we observed (or others like it) are unlikely.

# Confidence intervals as a process

A second way to think about a confidence interval is to think of it like a process.

Probability can come into play when there is a random process (and before that process is realized).

Random sampling is the process that creates a specific dataset.

If we calculate a 95% confidence interval for each of many random samples, 95% of the time it will contain the population mean.

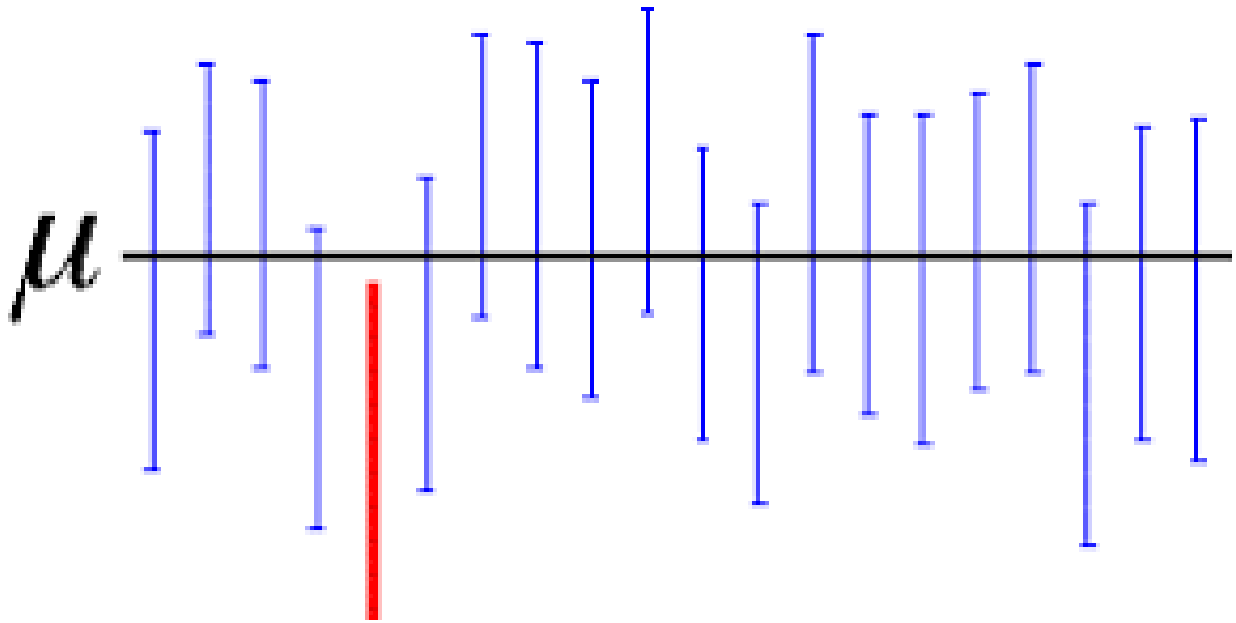



Figure 5:

## The language "confidence"

We can't say that there is a 95% "probability" that the population mean is contained in a confidence interval, after the sample is taken, because now it's a fixed unknown.

The random part of a confidence interval is not the mean, but the interval itself.

Statisticians say they are 95% "confident" that confidence interval contains the population mean. Our confidence is in the interval (the random part).

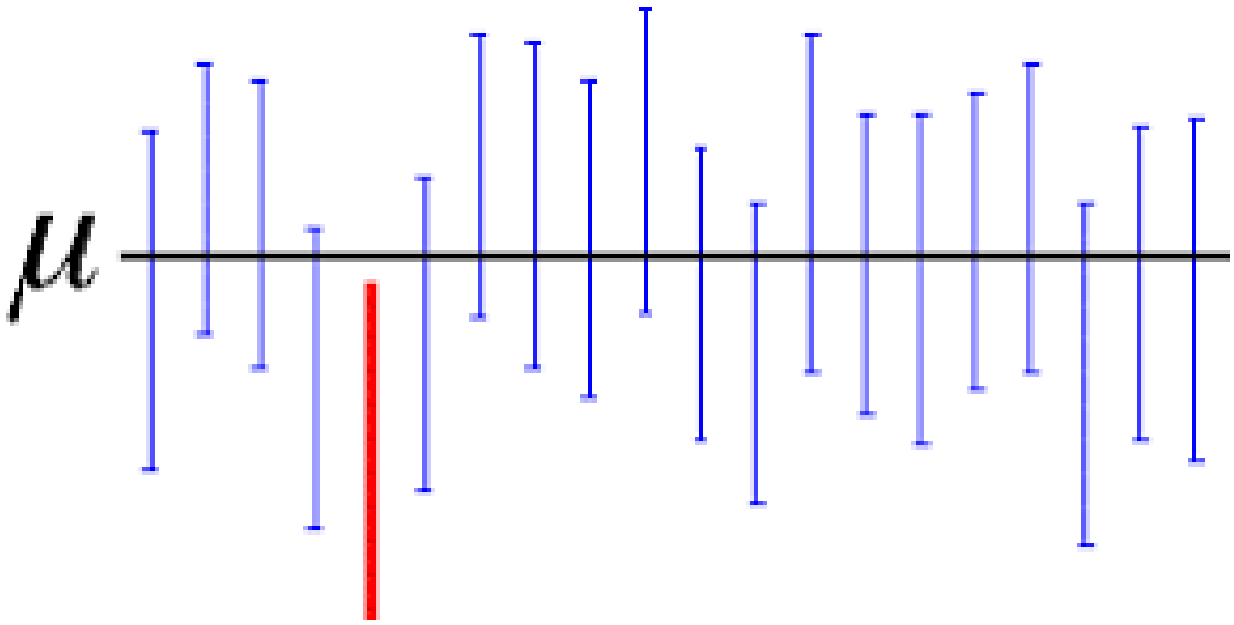



Figure 6:

# Wrapping Up

- Probability/likelihood, requires a random process. Once that process is realized (or if there is no process) there is no probability.
- Sampling distributions are used to think about a range of likely and unlikely outcomes of a process (DGP)
- Confidence intervals can tell us the range of population of population means for which are data seem "likely"
- We can think of confidence intervals as a *process* which captures to population mean 95% of the time