Sampling Distributions

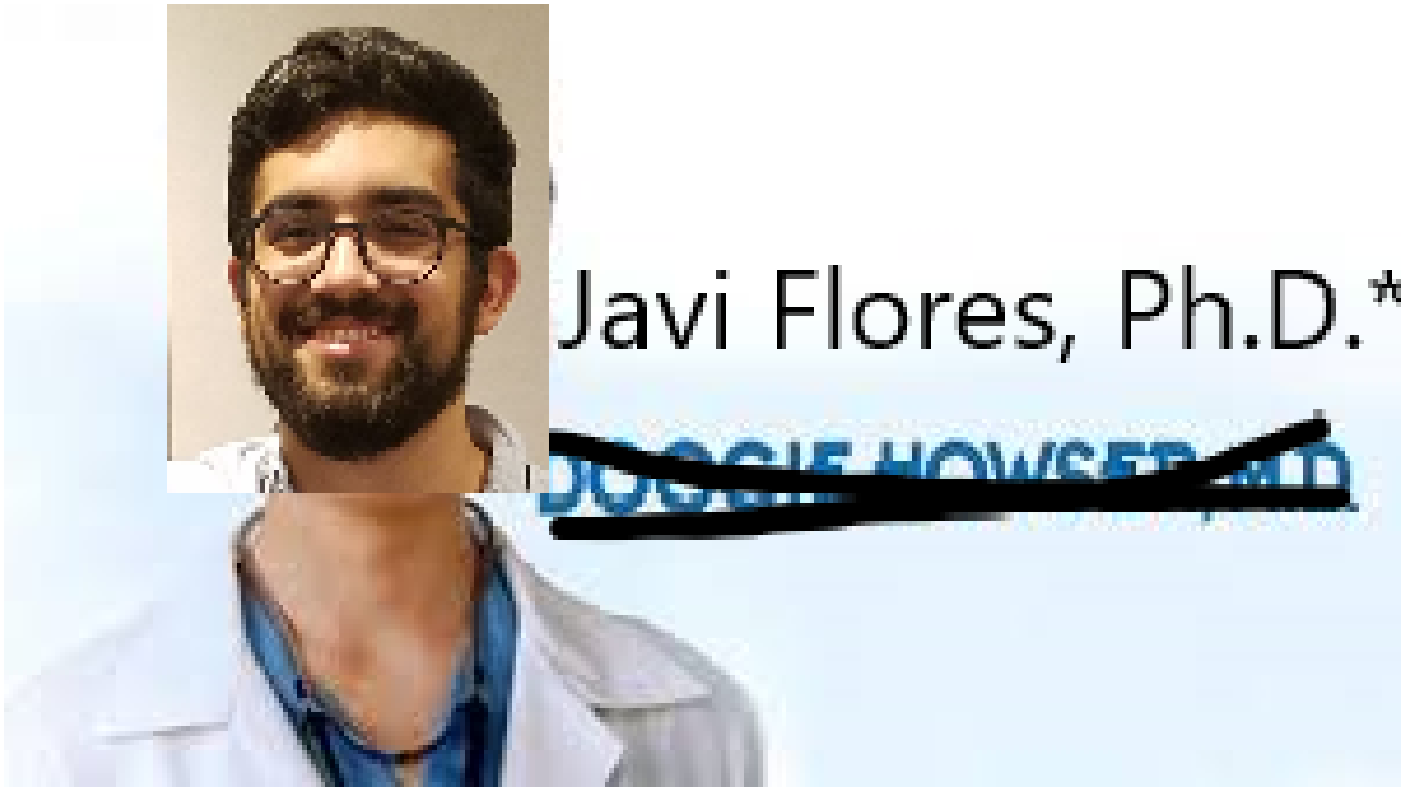
Javier E. Flores

February 13, 2019

Doogie Howser, M.D.

How old is this guy anyway?

After serving as your professor for the past three weeks or so, have any of you ever wondered how old I was?



* I have not yet received my PhD.

How old is this guy anyway?

Well whether or not you have, today you'll find out!

And as an added bonus, we're going to learn about sampling distributions along the way (exciting)!

Activity

It wouldn't be fun (and I wouldn't be able to teach you anything) if I just told you how old I was.

Instead, I want you each to write on the board how old you think I am (in months). Write your guess under any open ID number.

We'll then use these data to determine the best guess of my age.

Parameters and Statistics

For the purposes of this activity, this entire class represents the **population**.

We call any characteristic of a population a **parameter**. A common parameter of interest is the population mean, which is denoted by μ.

In this activity, μ represents the mean guess of my age.

If there were students absent to class today, we would no longer have access to the entire population. As a result, we wouldn't know the true value of μ.

However, we would still have a sample of students from the entire class. We can then calculate the *sample* mean $\bar{x}$, which is a **statistic**, to *estimate* the true population parameter μ.

How accurate is the estimate?

When we talk about the accuracy of a statistic, we are talking about how close the statistic is to the true population parameter.

In this case, we are talking about how close the average guess in our sample is to the average guess of the entire class.

In thinking about the accuracy of our statistic, it is important to determine its distribution. The distribution of our statistic, otherwise known as the **sampling distribution**, is dependent on:

1) The distribution of cases in our population
2) The distribution of cases in our sample

Question: How do we determine the sampling distribution of a statistic?

Activity (continued)

On the board are all of your guesses of my age. Using these data, which are from the entire population, we can determine μ and the population standard deviation, σ.

We'll also use these data to construct a *dot plot* of the population distribution of guesses.

In order to determine the sampling distribution of $\bar{x}$, I want each group to:

1) Draw six different random samples of size 10 using this web app (https://www.random.org/sequences/).
2) Construct a dot plot for each sample and calculate $\bar{x}$ and the sample standard deviation s. Write your sample means on the board and add them to the population dot plot.
3) Construct a dot plot of your six different sample means (this is the sampling distribution of $\bar{x}$).

Activity (continued)

Q: In reality, we usually have access to only a single sample. What is the estimate that is most likely to occur for a single sample?

A: The mean of the sampling distribution.

Q: How does the most likely estimate compare with the true population parameter?

A: Without sampling bias, the mean of the sampling distribution is *unbiased* for the true population parameter.

Q: How reliable do you think your best estimate is? Is there a way you could quantify its variability?

A: The variability of the sampling distribution, otherwise known as the **standard error**, is used to determine how reliable our estimate is.

Standard Error and Sample Size

The standard error and mean of our sampling distribution are both dependent on the population parameters, but they are also affected by sample size.

To investigate the role that sample size plays in our sampling distribution we'll use *StatKey*, a free online companion to the course textbook.

The pre-loaded dataset, "Percent with Internet Access (Countries)", will be used for our investigation.

Standard Error and Sample Size Activity

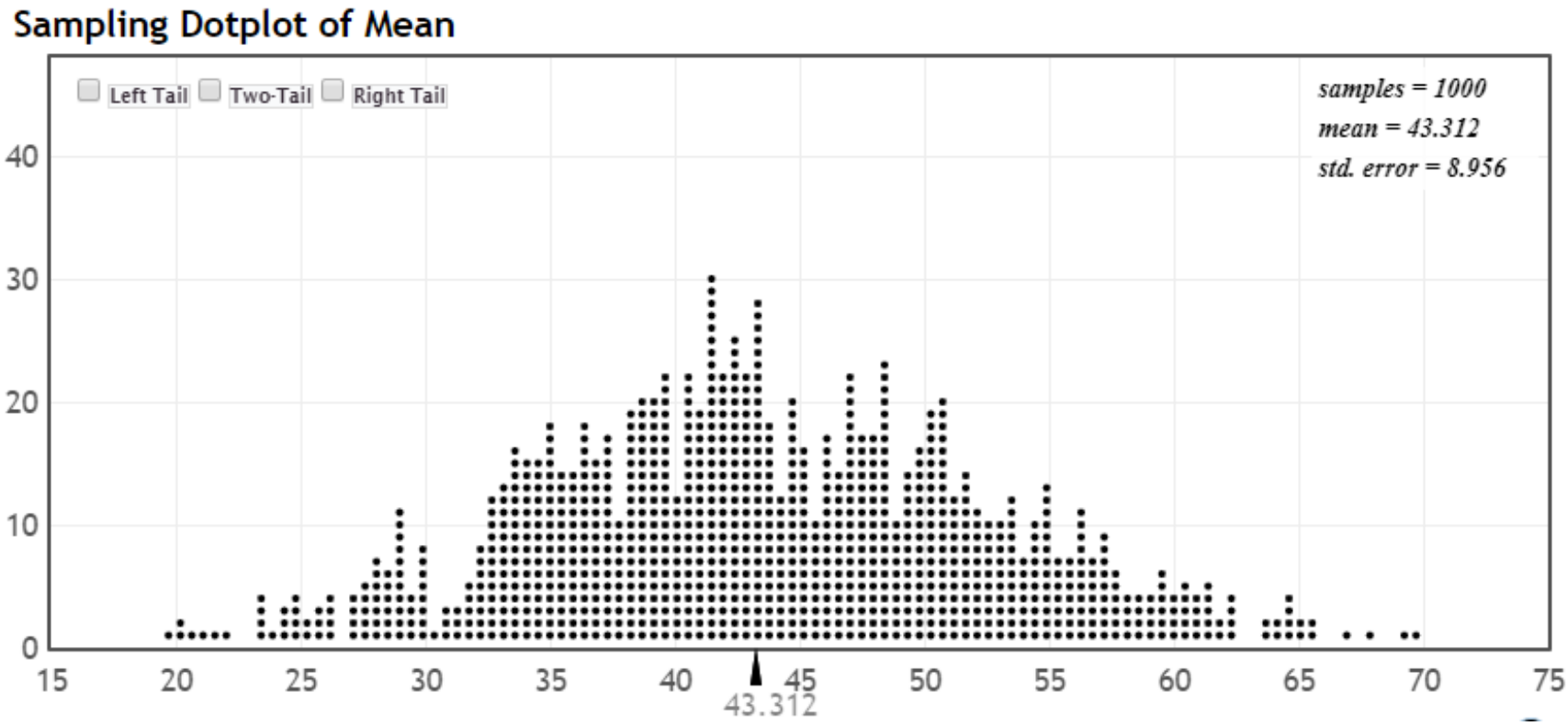
With your groups, generate the following using *StatKey*:

1) A dotplot of the sampling distribution generated by 1000 samples of size 10.
2) A dotplot of the sampling distribution generated by 1000 samples of size 30.
3) A dotplot of the sampling distribution generated by 1000 samples of size 150.
4) A dotplot of the sampling distribution generated by 1000 samples of size 203.

Discuss your findings among your groups.

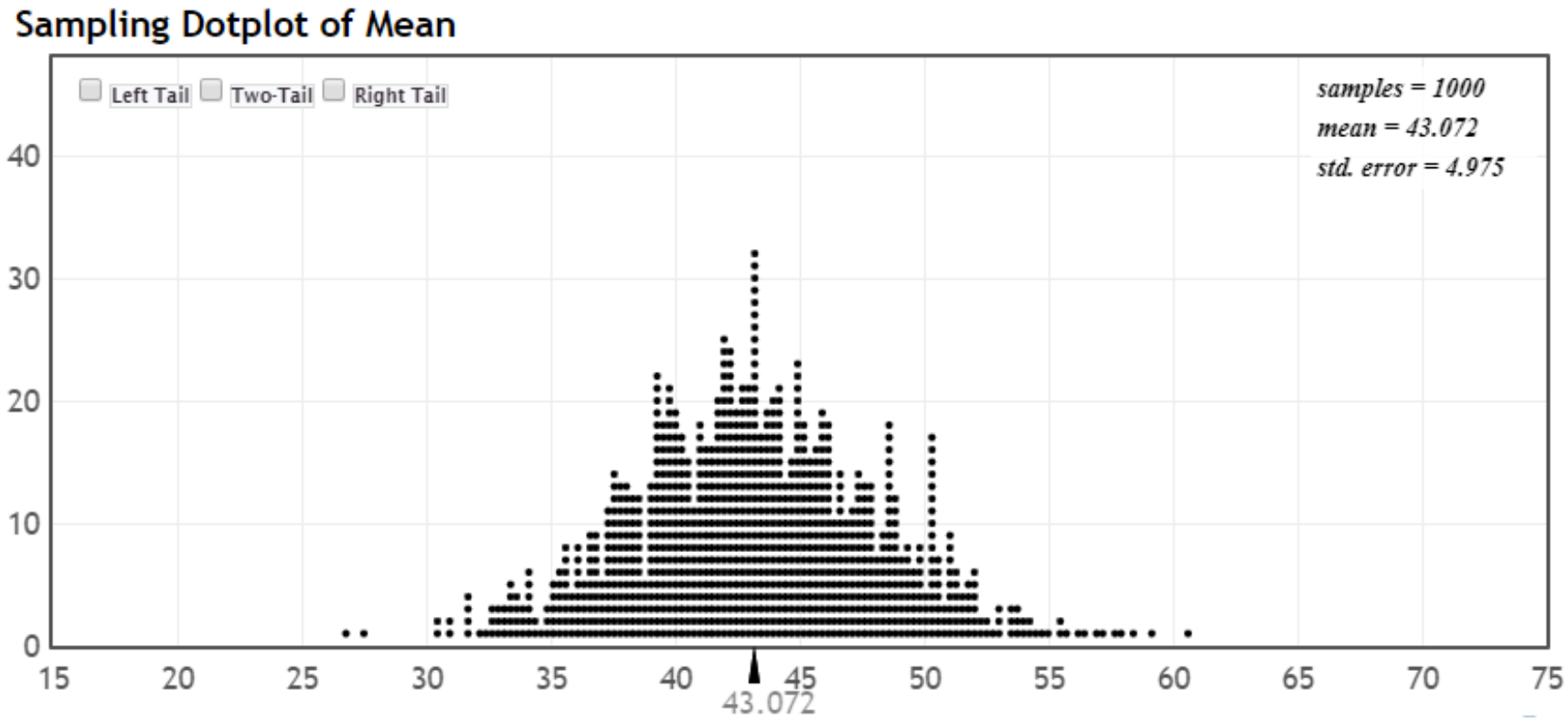
Standard Error and Sample Size

Shown below is the sampling distribution of the mean generated by 1000 samples of size 10.



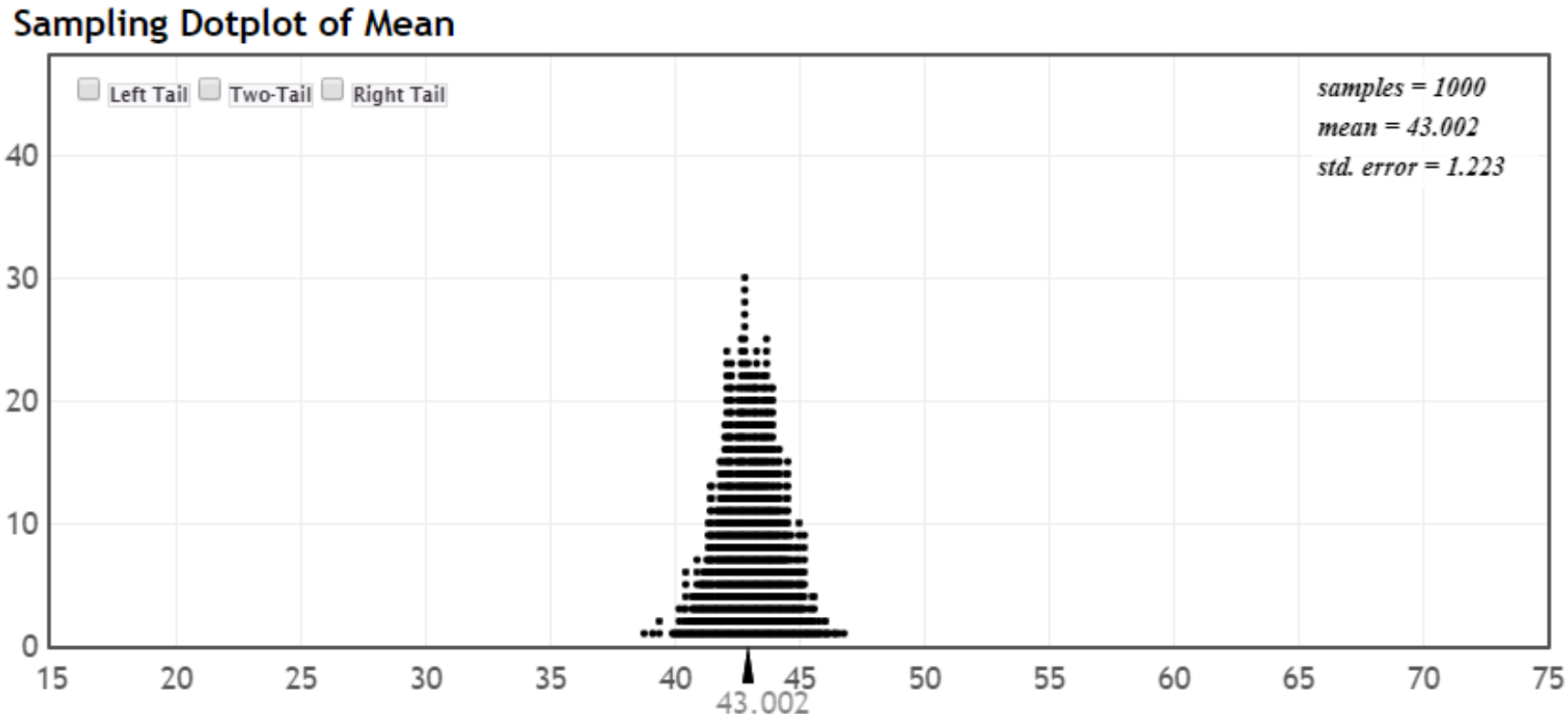
Standard Error and Sample Size

After tripling the sample size of each of our 1000 samples, we obtain the following:



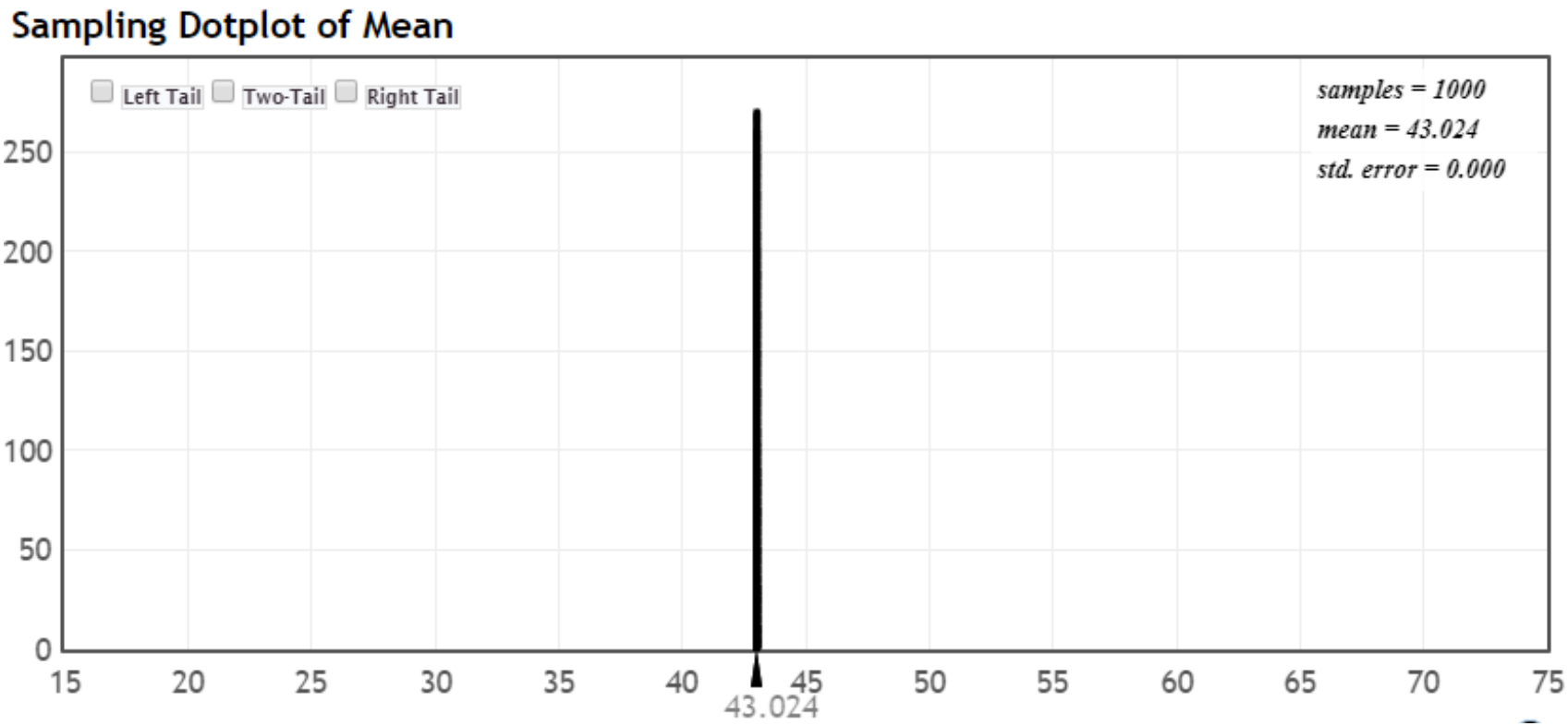
Standard Error and Sample Size

After increasing the sample size of each of our 1000 samples to 150, we obtain the following:



Standard Error and Sample Size

Finally, if we use the entire population's data:



Standard Error and Sample Size

As the size of our sample increases, the standard error of our sample statistic decreases.

The standard error may be interpreted as the standard deviation of a sample statistic.

In the past, we've discussed the role sample size plays in reducing variability. As this exercise has demonstrated, the property still holds for the sampling distribution standard error.

Related to discussions of variability were discussions of bias.

Question: Does increasing sample size affect bias?

Sampling Bias

Earlier I mentioned that the mean of our sampling distribution is *unbiased* for the population mean in the absence of sampling bias.

This then implies that if we did have sampling bias, the mean of our sampling distribution would be biased in some way.

The direction and magnitude of this bias is completely dependent on the way and extent to which sampling bias was introduced.

Consider the following example...

Sampling Bias

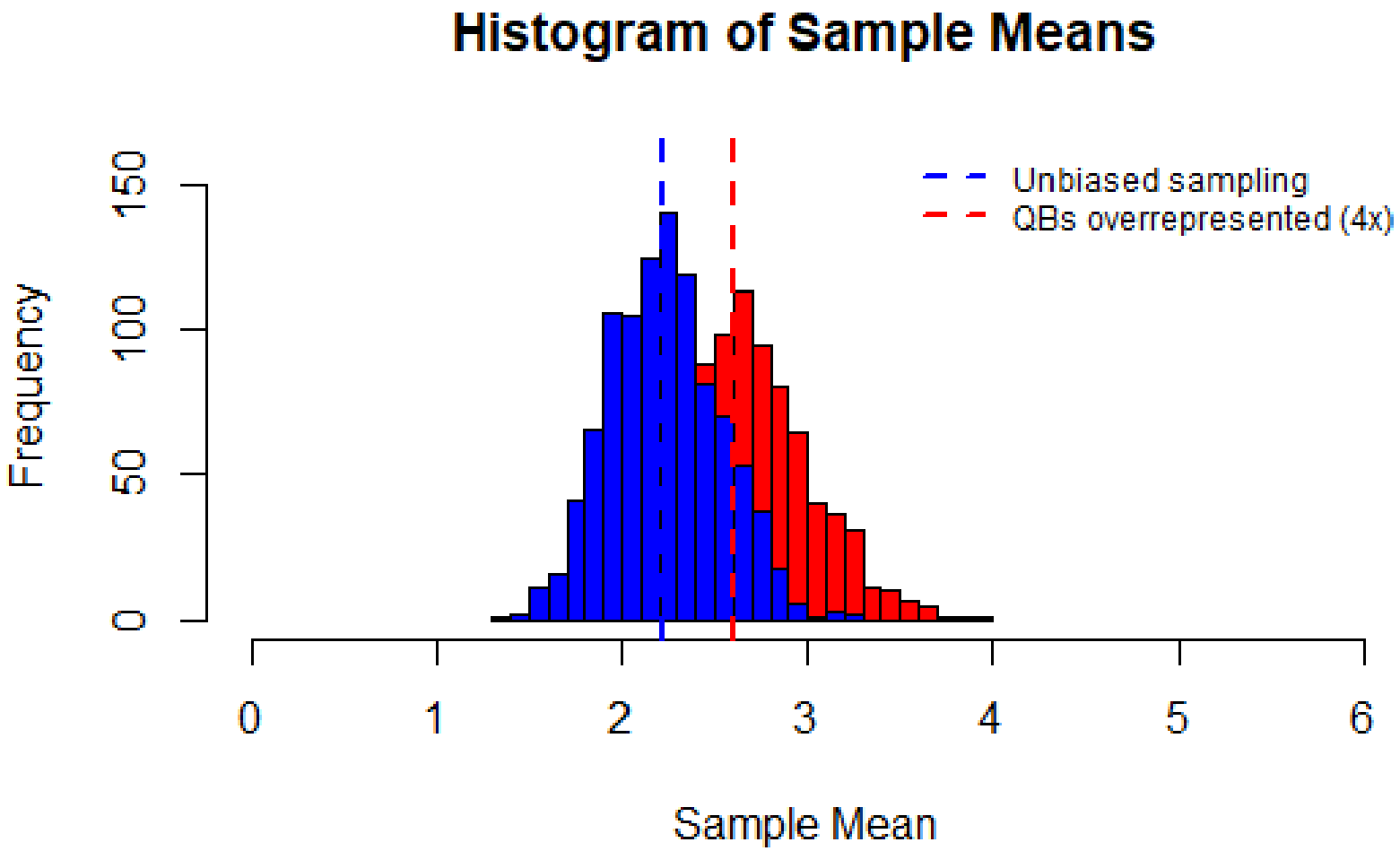
Suppose we were interested in determining the average salary of a football player in the NFL.

Suppose also that we hadn't yet learned about proper sampling methods and "accidentally" sampled in such a way that quarterbacks were four times as likely to be sampled than other positions.

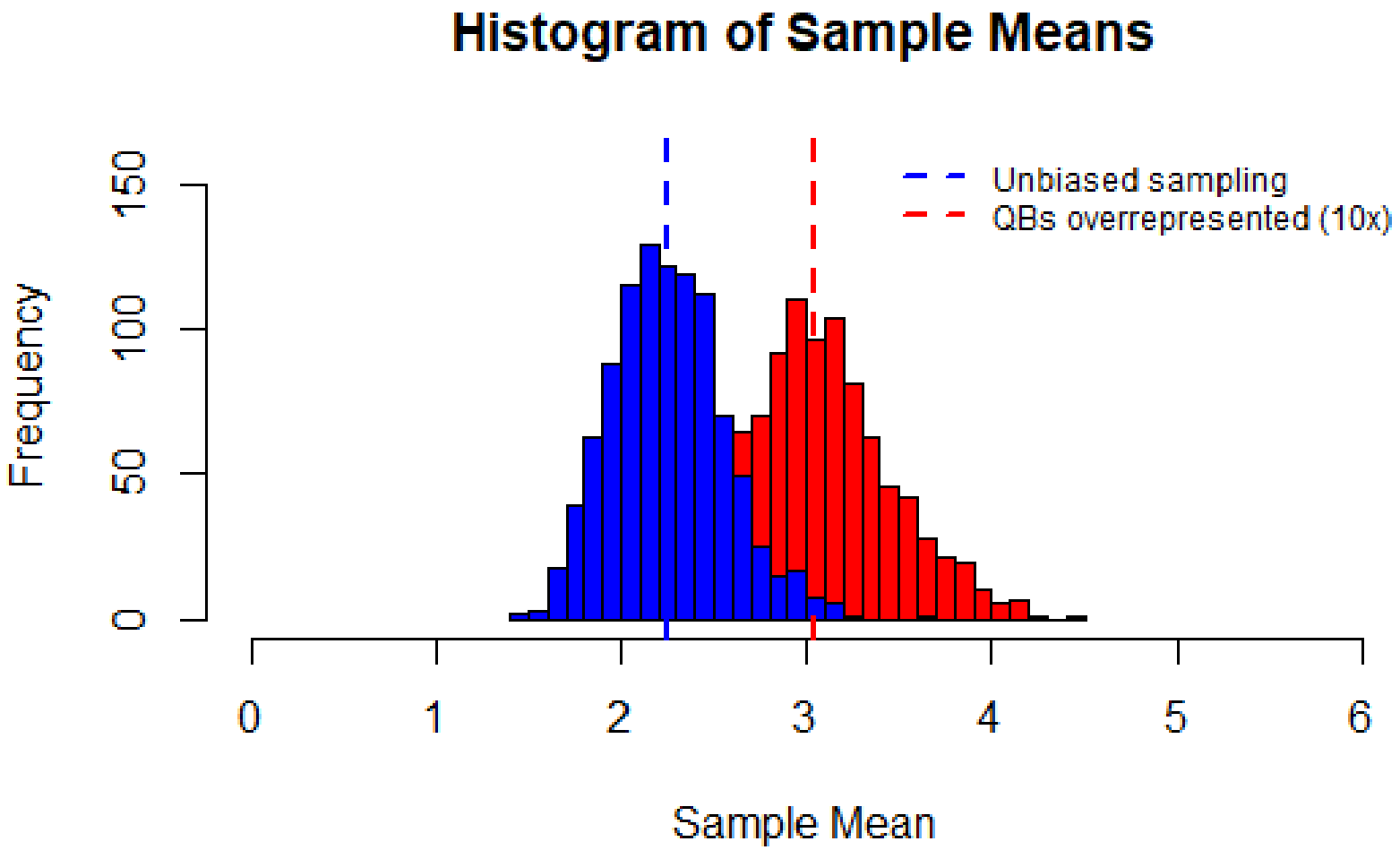
How would this affect the mean of our sampling distribution relative to the true population mean?

What if quarterbacks were ten times more likely to be sampled?

Sampling Bias



Sampling Bias



Wrap-Up

Right now, you should...

- Understand the difference between a parameter and a statistic.
- Be able to explain what a sampling distribution is and describe how it relates to the population distribution and sample distribution.
- Know the effect sample size has on the bias and variability of the sampling distribution.

These notes cover section 3.1 of the textbook. Please read through the section and its examples along with any links provided in this lecture.