Sampling Distributions

Ryan Miller

Statistical Inference

A major goal of statistics is *inference*, or using a sample to learn about a population. Today we will walk through the train-of-thought of how statisticians developed formal approaches to inference.

- In today's activity, the population will be the end of semester grades (percentages) of my previous Sta-209 students
- I won't give you the population, but I'll let you take as many *random samples* of size $n = 10$ as you want
- The goal will be to find a logically sound method for describing an aspect of the population in the more realistic setting where we only have *one random sample*

Distributions

- If we want to learn about a population, the most informative thing we could possibly ask for is the full population distribution
    - ie: a histogram or dotplot of *all* the end of semester grades
- Instead, for various reasons, we typically focus on a single number that *summarizes* an aspect of the population distribution that we're most interested in
- Thinking about our population of interest, what summary measures might we want to know?

Estimation

One goal of statistical inference is **estimation**:

- Suppose we're interested in the mean of the population (μ)
- How might we estimate μ from a random sample?
- How certain are we that this estimate will be close to μ?

It is logical to estimate μ using $\bar{x}$; but with only a single sample, the accuracy of our estimate is a bit of a mystery. However, if we repeatedly draw random samples, we can study how an estimate will behave!

Sampling Distribution Activity - Directions

For this activity we will use a statistics program known as `R`, this is likely the only time we'll use `R` in this class, but it is software of choice for many statisticians.

1. Open `RStudio` and type: "source("https://remiller1450.github.io/s209s19/funs.R")"
2. Enter "sample_grades()" to generate a random sample of 10 student's end of semester grades
3. Find the mean of your random sample and record it on the dotplot on the board
4. Repeat steps 1 and 2 until you've recorded the means of six different random samples on the board

This dotplot represents the distribution of different sample means that we could pontentially see when taking a random sample of size $n = 10$ from this population. With your group, discuss why it might be valuable to study this distribution (think about *inference* and *estimation*).

Sampling Distribution Activity - Some Questions

1. Based off the **sampling distribution** (the dotplot on the board), what do you think μ is?
2. Had you only collected a single random sample of size 10, what value do you think is most likely to be that sample's mean?
3. How much variability is there in the different sample means that we could possibly see?
4. Could we use this to provide an **interval estimate** of μ?

Sampling Distribution Activity - Answers

1. Assuming the samples are *representative*, μ is the center of the sampling distribution! This is because the sample statistic $\bar{x}$ is **unbiased**
2. We are most likely to see a sample mean at the very center of sampling distribution, so μ is the *most likely* mean of any particular sample
3. We can assess the variability of the possible sample means that we could see by looking at the standard deviation of the sampling distribution, this is called the **standard error** (SE) of the sample mean.
4. We could provide estimates of μ that look like $\bar{x} \pm b * SE$. The 68-95-99 rule can help us choose b (at least for sampling distributions with a certain shape)

The Role of Sample Size

The sampling distribution depends upon both:

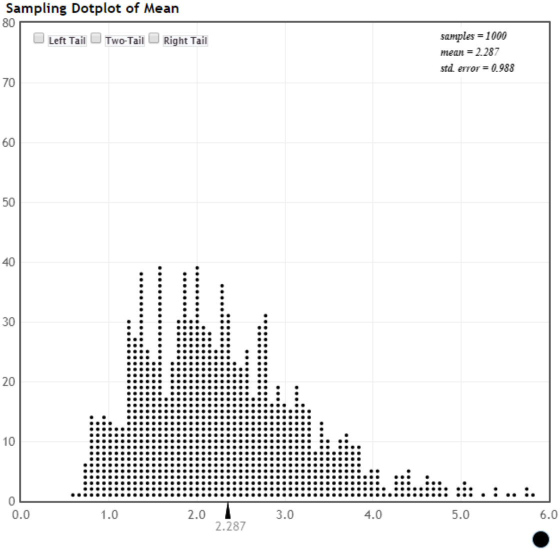
1. The parameters of the population distribution
2. The size of the sample, and how it was collected

- We'll investigate the role of sample size using *StatKey*, a free online companion to the Lock5 textbook: StatKey Link
- We'll look at the "NFL Contracts" dataset that comes pre-loaded in StatKey

StatKey allows us to quickly generate many random samples from a dataset

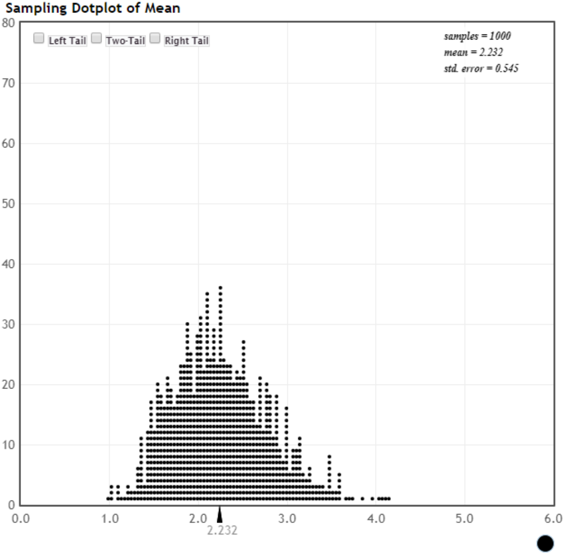
The Role of Sample Size

Sampling distribution of $\bar{x}$ for 1000 samples of size $n = 10$



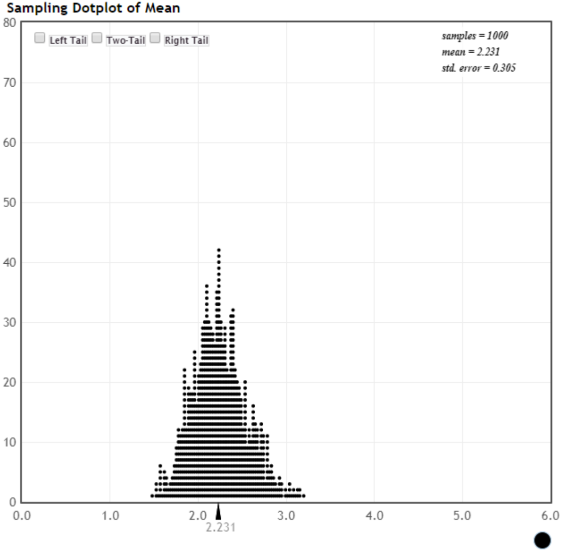
The Role of Sample Size

Sampling distribution of $\bar{x}$ for 1000 samples of size $n = 30$



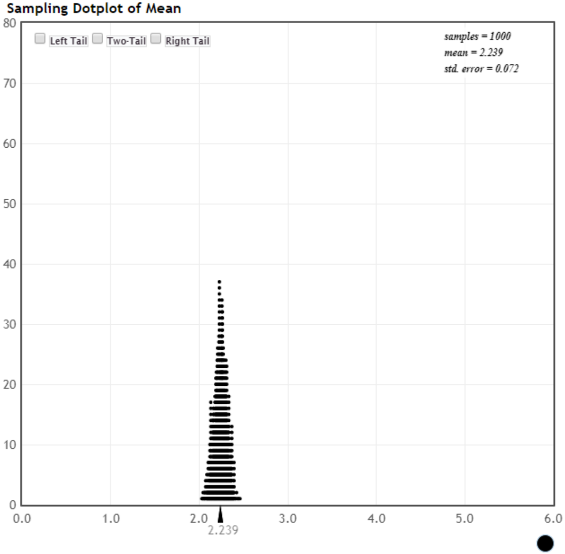
The Role of Sample Size

Sampling distribution of $\bar{x}$ for 1000 samples of size $n = 100$



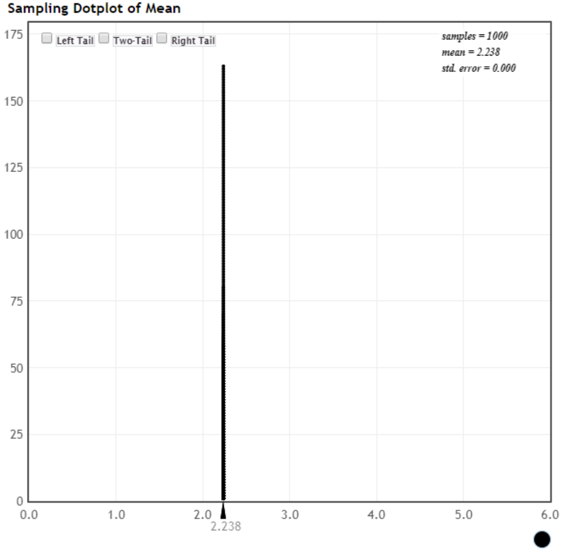
The Role of Sample Size

Sampling distribution of $\bar{x}$ for 1000samples of size $n = 1000$



The Role of Sample Size

Sampling distribution of $\bar{x}$ when the entire population is sampled



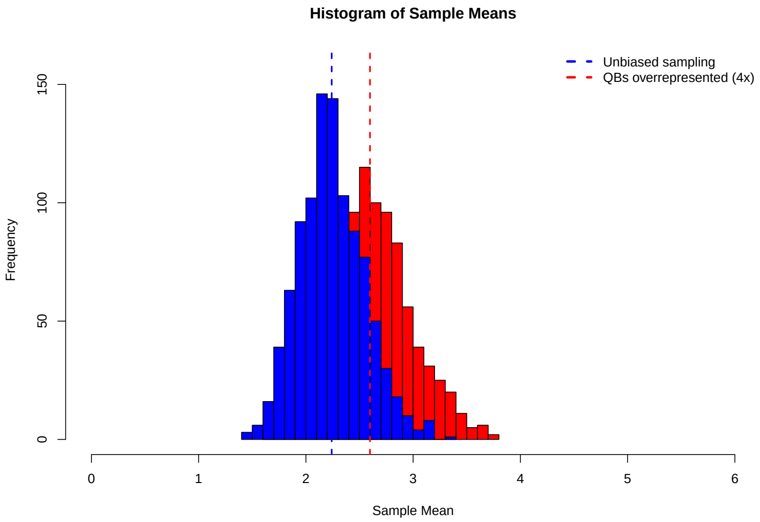
The Role of Sample Size

- As the size of our sample increases, the **standard error**, denoted SE, of our sample statistic decreases
- Standard error is the standard deviation of a sample statistic (ie: it describes variability in the sampling distribution)

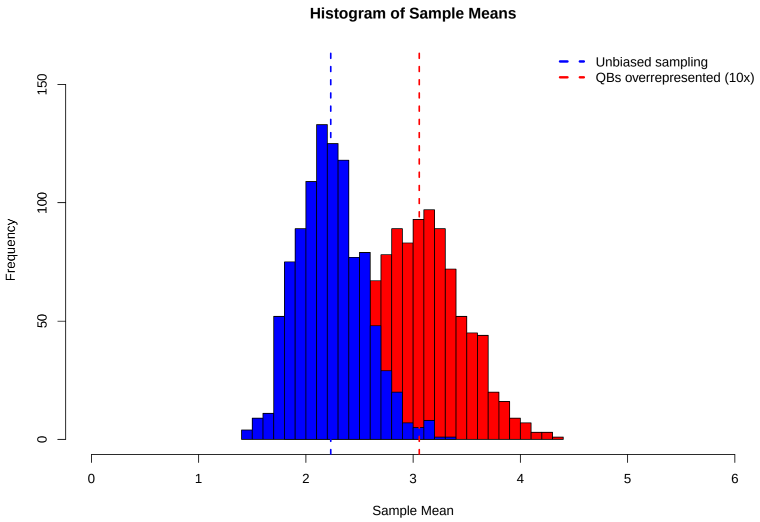
Sampling Bias

- Quarterbacks represent 4.3% of NFL players but often to receive a disproportionate amount of attention and also tend to be paid higher salaries than other positions
- Suppose we sample in a way that makes QBs four times more likely to be sampled than other positions, how might this influence our samples?
- What if QBs were ten times more likely to be sampled?

Sampling Bias



Sampling Bias



Conclusion

Right now you should...

1. Understand the relationships between the **population distribution**, the **sample distribution**, and the **sampling distribution**
2. Be comfortable with the terminology of **parameters** and **statistics**
3. Understand, when we only have one sample, the sample statistic is our best guess at the population parameter
4. Understand the impact of bias and sample size (variability) on the sampling distribution

These notes cover Section 3.1 of the textbook, I encourage you to read through the section and its examples