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 behind how statisticians have approached inference.

- In this activity, the population will be the end of semester grades 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
- Our short-term goal will be see what we can learn about a population by repeatedly taking random samples
  - Our long-term goal will be to apply this insight to situations involving only a single random sample

The Population Distribution

- The population distribution contains *all* of the information about the variable of interest
  - In our example, we could view the population distribution using a table or barchart of the end of semester grades
- In most situations, statisticians focus on a single statistic that summarizes the aspect of the population they are most interested in
  - In our example, the statistic of interest will be the proportion of A's

Estimation

- Suppose we're interested in the population proportion of A's, denoted p_A
- How would you estimate p_A from a single random sample?
- Is your estimate likely to be p_A *exactly*?

Estimation

- The logical estimator of p_A is the sample proportion of A's, $\hat{p}_A$
- This estimate is *unlikely* to be exactly p_A, but it should be close
  - Exactly how close is something we seek to determine
  - In today's activity, we'll take repeated random samples to gauge how close $\hat{p}_A$ tends to be to p_A

Sampling Distribution Activity - Directions

This is the only time we'll use `R` in this class, but it is the software of choice for most statisticians, and you'll use it in future stats classes (if you choose to take them).

1. Open `RStudio` and type: **source("https://remiller1450.github.io/s209f19/funs.R")**
2. Enter **sample_grades()** to generate a random sample of 10 student's end of semester grades
3. Find the proportion of A's in the sample and record it on the board
4. Repeat steps 1-3 until you've recorded results from 6 different random samples

These values represent the distribution of possible sample proportions that could occur when taking a random sample of size $n = 10$ from this population. With your group, discuss why it is important to study this distribution.

Sampling Distribution Activity - Some Questions

1. Based off the **sampling distribution** (the dotplot on the board), what do you think p_A is?
2. Had you only collected a *single* random sample of size 10, what would you expect is the *most likely* value of $\hat{p}_A$ for that sample?
3. How much variability is there across different samples?
4. Could we use this variability to come up with an **interval estimate** of p_A?

Sampling Distribution Activity - Answers

1. Assuming the samples are *representative*, p_A is the center of the sampling distribution! This is because the sample statistic $\hat{p}_A$ is **unbiased**
2. p_A is the center of the sampling distribution, so $\hat{p}_A$ is most likely to be p_A!
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) since it describes an estimate
4. We could provide estimates of p_A that look like $\hat{p}_A \pm c * SE$. The 68-95-99 rule could help us choose c (at least for sampling distributions with the right shape)

Confidence Intervals

- Intervals of the form Estimate $\pm$ MOE, where MOE is a carefully determined margin of error, are called **confidence intervals**
- We will spend the next couple of weeks studying confidence intervals
- For now, we'll investigate the influence a few different factors have an the *sampling distribution*

The Role of Sample Size

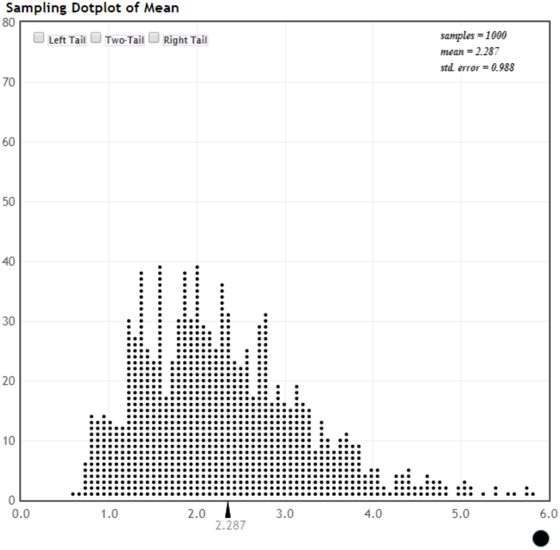
The sampling distribution depends upon:

1. The parameters of the population distribution
2. The size of the sample
3. How the sample was collected

- We'll first 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

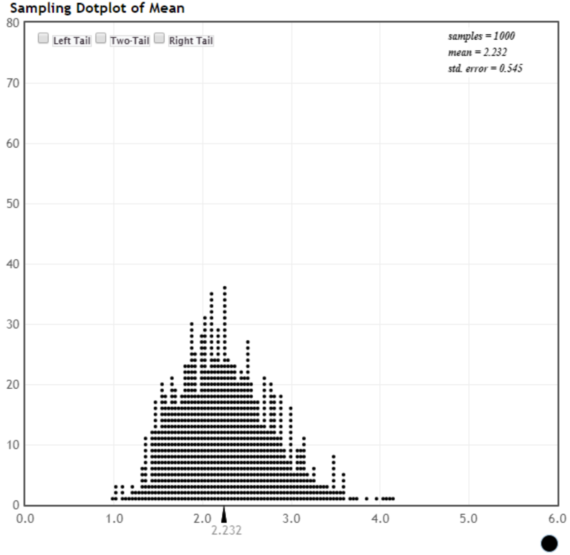
The Role of Sample Size

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



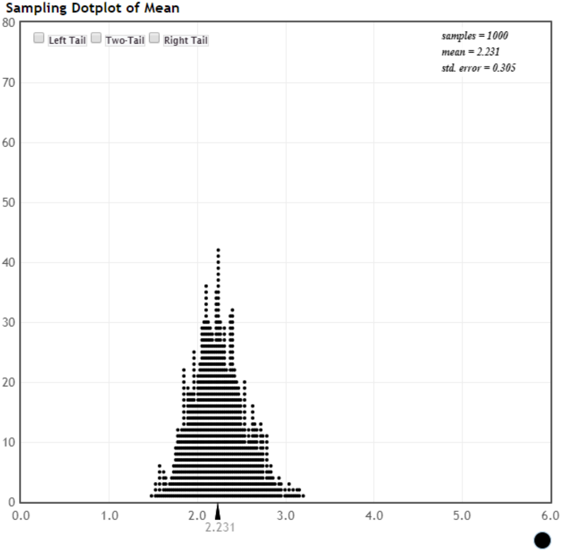
The Role of Sample Size

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



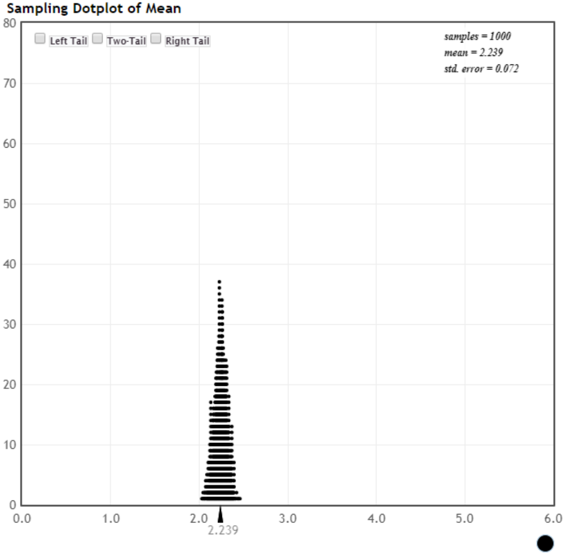
The Role of Sample Size

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



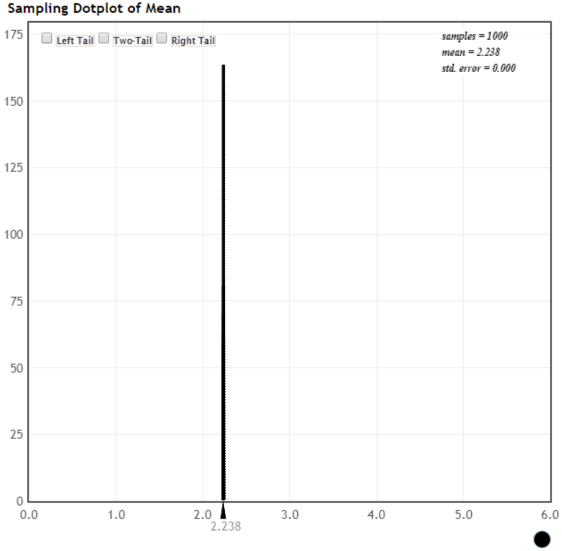
The Role of Sample Size

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



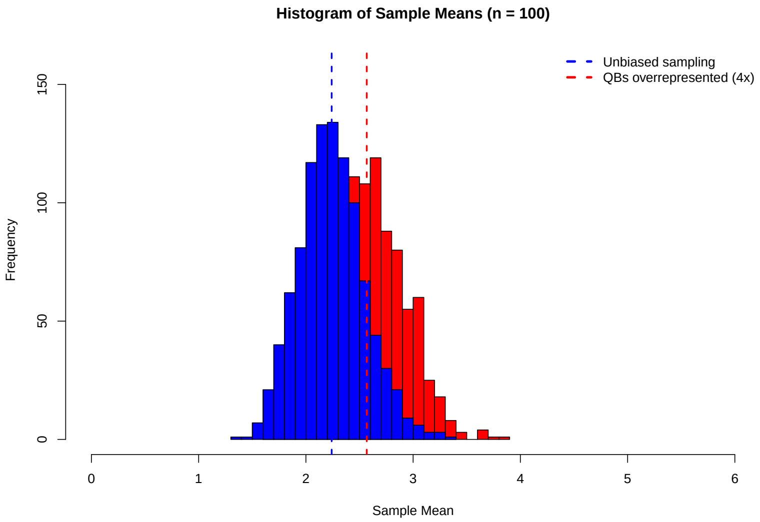
The Role of Sample Size

- 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)

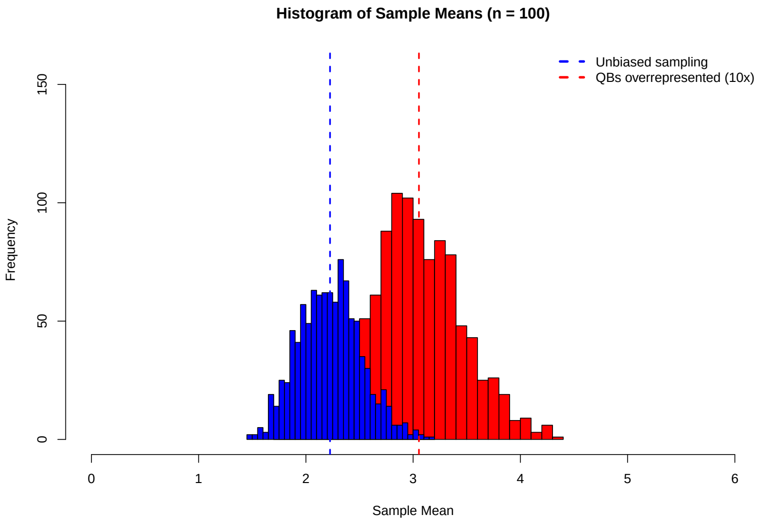
Sampling Bias

- Quarterbacks represent 4.3% of NFL players but tend to receive a disproportionate amount of media attention and are 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 the sampling distribution (for estimates, $\bar{x}$, of mean the NFL salary)?
- What if QBs were ten times more likely to be sampled?

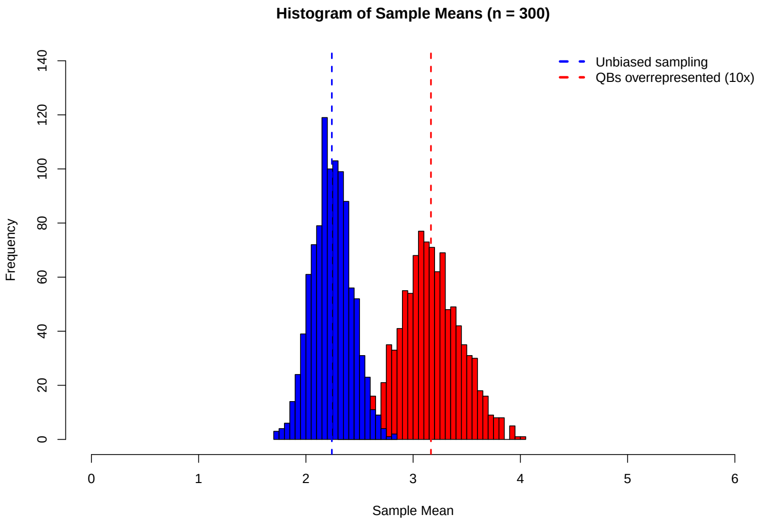
Sampling Bias



Sampling Bias



Sampling Bias



Sampling Bias

- Larger samples tend to provide better estimates if the samples are representative
  - But larger sample size cannot fix sampling bias, in fact it often exacerbates it

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

If you want more information:

- Read Ch 3.1