Population, Sample, and Sampling Distributions

Ryan Miller

Three Distributions Activity

We've previously discussed samples vs. populations (the Candy Activity). Today we will learn about the age of our class (in days):

- The entire class represents the **population**
- A **parameter**, or characteristic of the population that we want to determine, could be the population mean, denoted by μ

- Suppose we randomly sample $n = 10$ students
- We can estimate μ using a **statistic** calculated from the sample, $\bar{x}$, but how accurate is this estimate?

Three Distributions Activity

Thinking about the different sources of variability in this example, there are 3 different distributions at play:

1. The distribution of the cases in our population
2. The distribution of the cases in our sample
3. The distribution of $\bar{x}$

Question: Why doesn't μ, the population mean, have a distribution?

Three Distributions Activity

- I'd like each of you to calculate your age (in days) using the following link: https://www.calculator.net/age-calculator.html
- Once you have it, write your age below any under open ID number on the board
- We'll use these data to determine μ and σ
- We'll also construct a *dot plot* of our *population's age distribution*

Three Distributions Activity

Now the most interesting part, we need random samples of $n = 10$ students

1. Draw six different random samples of $n = 10$ using: https://www.random.org/sequences/
2. Construct a dot plot for each sample
3. Calculate $\bar{x}$ and s for each sample
4. Construct a dot plot of your 6 different sample means (this is the sampling distribution!)
5. Because it's hard to get a clear picture of the sampling distribution with just 6 samples, write your sample means on the board and add them to the class dot plot

Three Distributions Activity - Questions

1. Usually we only see one sample, what is the estimate that is most likely to occur for a single sample?

- The mean of the sampling distribution is the most likely estimate

2. How does the most likely estimate compare with the true population parameter?

- Provided there is no sampling bias, the mean of the sampling distribution is *unbiased* for the population parameter it is estimating

3. How reliable do you think your best estimate is? Is there a way that you could quantify its variability?

- The reliability depends on the variability (standard error) of the sampling distribution, if we knew the standard error we'd have a sense of the estimate's variability

The Role of Sample Size

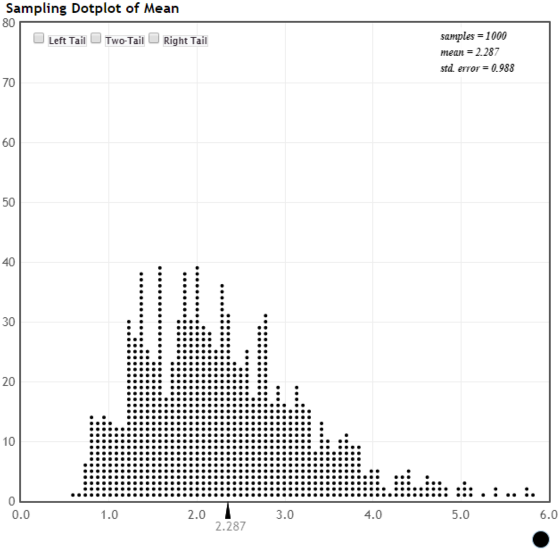
The sampling distribution depends upon the parameters of the population distribution, as well as the sample

- Let's 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

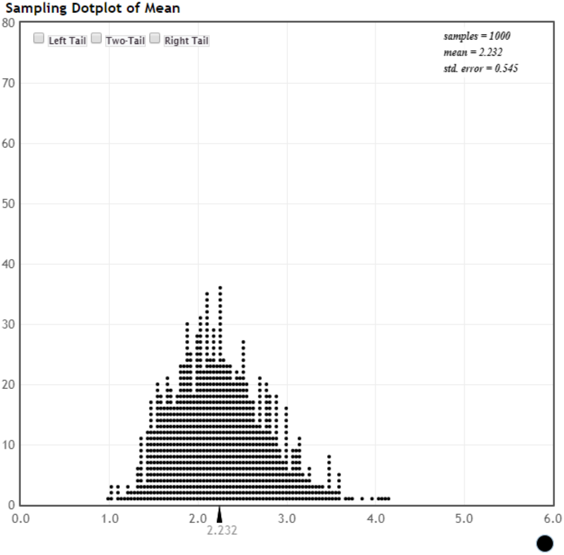
The Role of Sample Size

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



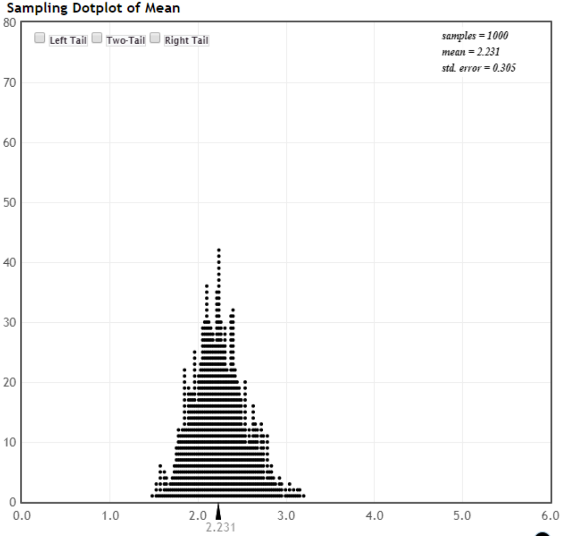
The Role of Sample Size

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



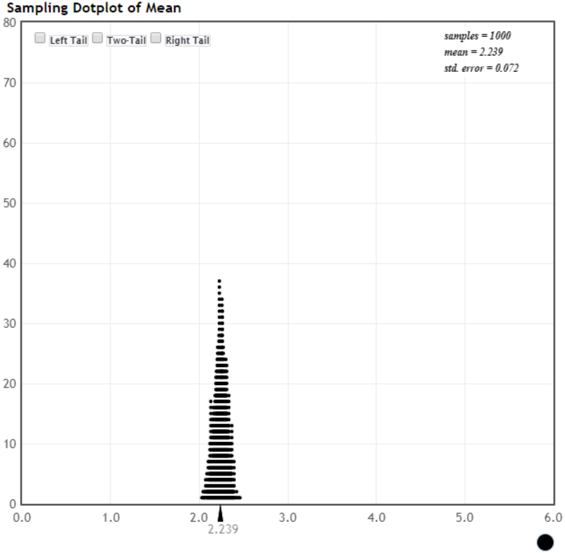
The Role of Sample Size

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



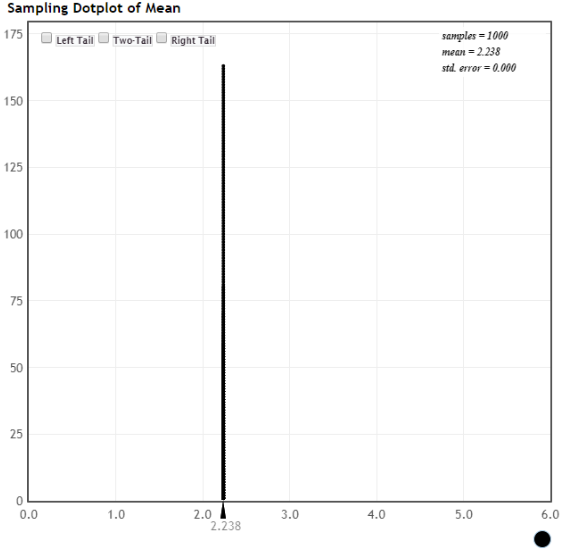
The Role of Sample Size

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



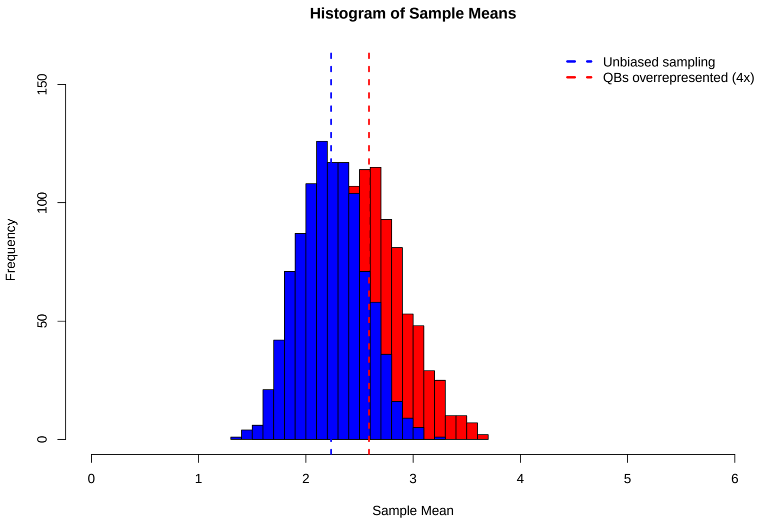
The Role of Sample Size

- 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)

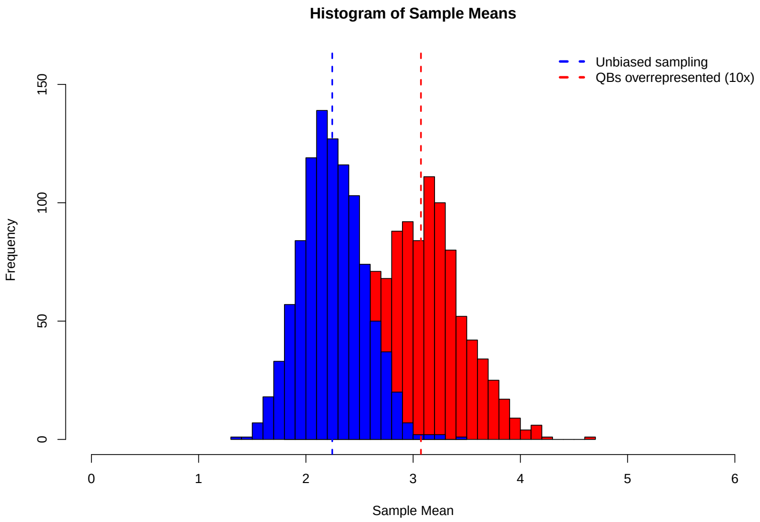
Sampling Bias

- 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