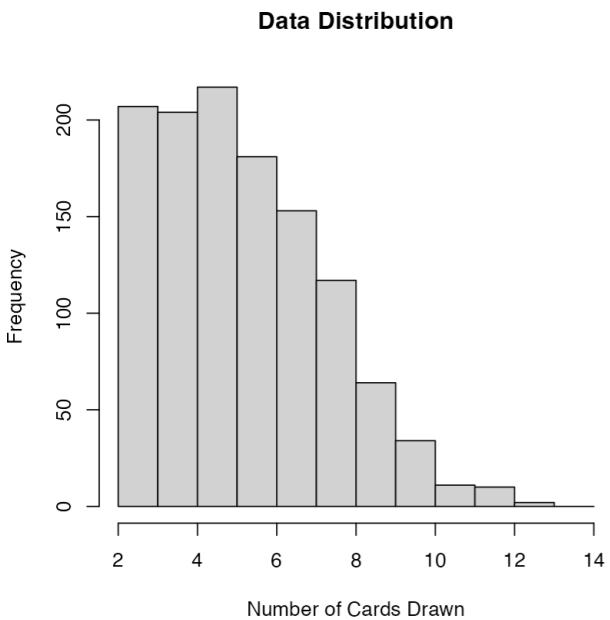
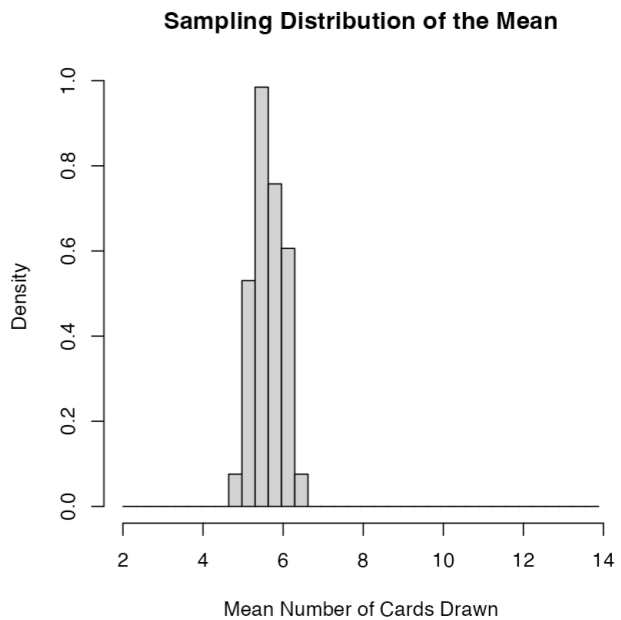
IMS 12 (ish): Sampling Distributions and Bootstrapping

Last Time

Previously, we simulated how many cards we have to draw in a well shuffled deck of cards until we saw two cards that had the same suit. If we pooled the results from all three sections of 220. We would have something like the following:





Last Time

- Describe the two distributions: center, shape/skewness, spread.
- Suppose there where 60 groups in total across the three sections of STAT 220. Use the web-widget online and observe what happens when you change the "*Number of simulations per mean estimated*". Try multiple values: very small, very very large, etc. Which distribution changes? How?

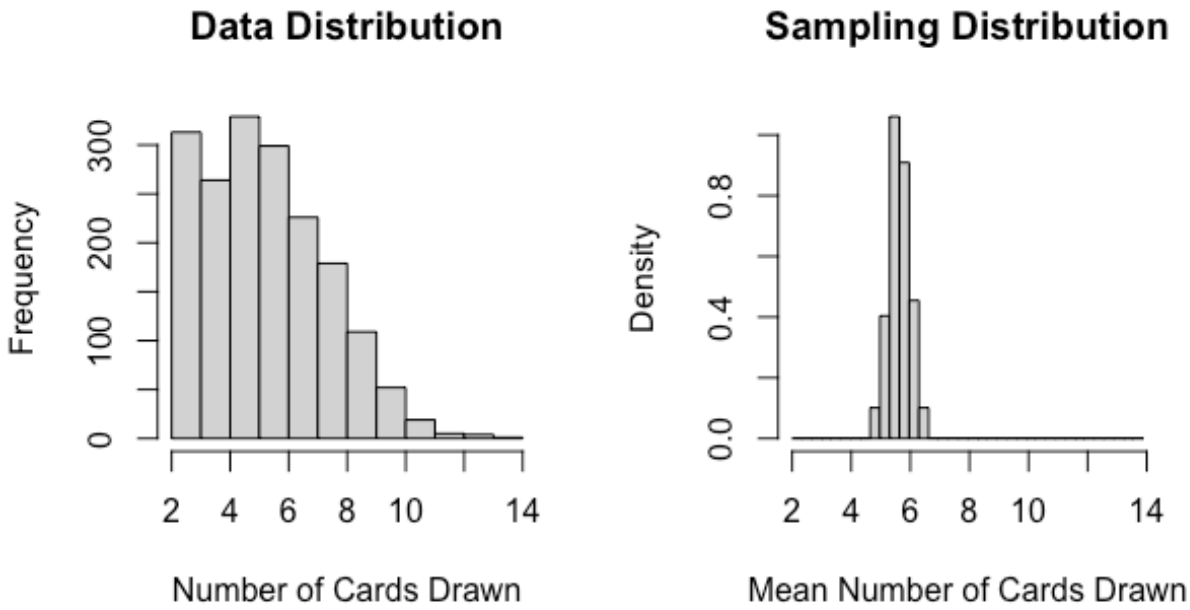
Recall: Statistics Estimate Parameters

Definition

A **statistic** is a numerical summary of sample data. A **parameter** is a numerical summary of a population.

- A *statistic*, is an estimate for a *parameter.*
- Common statistics are:
  - sample proportion ($\hat{p}$)
  - sample mean ($\overline{x}$ or $\hat{\mu}$).
- These statistics estimate the
  - population proportion (p)
  - population mean (μ).

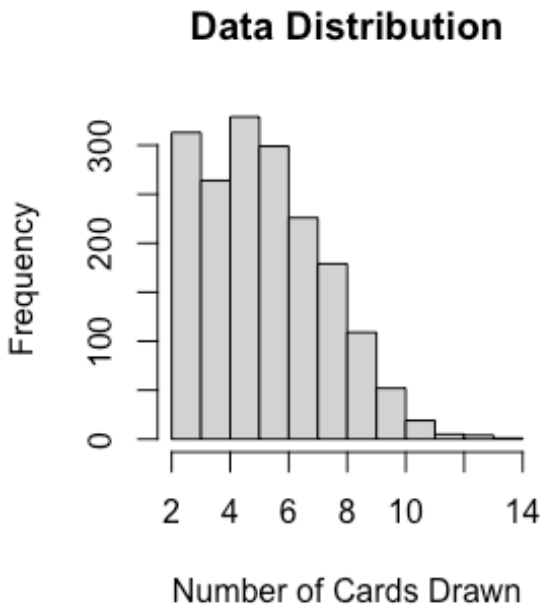
Sampling Distributions

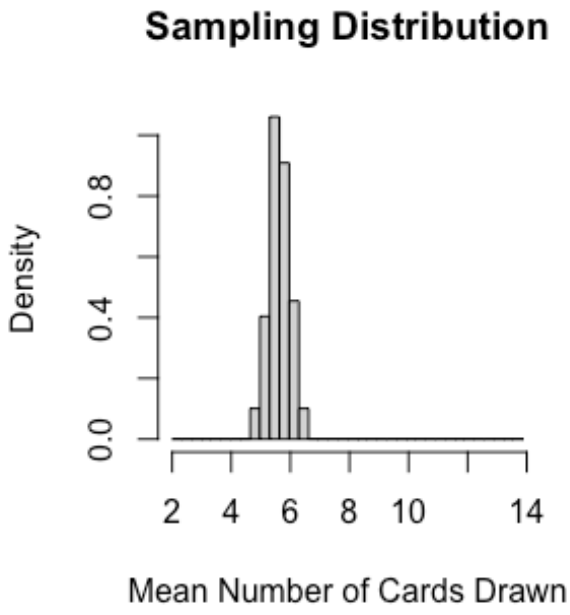


Definition

The **sampling distribution** of a statistic describes the range of possible values that can be observed for a statistic when taking a random sample of size n from the population, and their associated probabilities.

Sampling Distributions





- The distribution of individual observations is called the *data distribution*.
- A *sampling distribution* is the distribution of a statistic.

Sampling Distributions

Sampling distributions help us answer the questions:

- How much might a *statistic* vary from sample to sample?
- How would we describe the shape, center, and variability of the possible values for our *statistic*?
- What is the effect of the sample size n on the shape of the *sampling distribution*.

Sampling Distributions

- The distribution of means we simulated helped us understand the concept of sampling distributions.
- Sampling distributions for means often do NOT behave like the distributions like the data distribution.
- In fact, sampling distributions for means have a particular behavior that *always* happens.

FTS Revisited

We recall the Fundamental Theorem of Statistics (also called the Central Limit Theorem), a bit more formally this time.

Fundamental Theorem of Statistics (for proportions)

Suppose we have an independent random sample taken from a given population. The estimated proportion of a particular outcome will converge to the true proportion as the sample size (n) increases

$$\hat{p} \to p.$$

Furthermore, the sampling distribution of $\hat{p}$ convervges to a **normal distribution** as n increases.

FTS Revisited

We recall the Fundamental Theorem of Statistics (also called the Central Limit Theorem), a bit more formally this time.

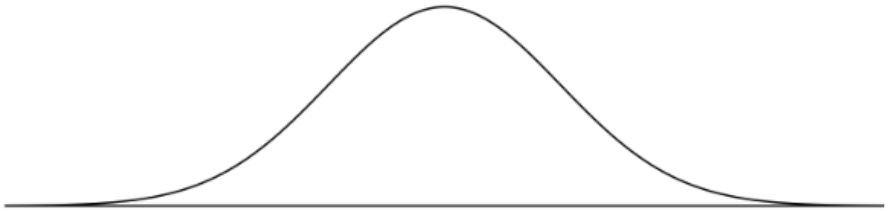
Fundamental Theorem of Statistics (for means)

Suppose we have an independent random sample taken from a given population. The estimated mean will converge to the true mean as the sample size (n) increases

$$\overline{x} \to \mu.$$

Furthermore, the sampling distribution of $\overline{x}$ convervges to a **normal distribution** as n increases.

Normal Distribution



- The **normal distribution** is a symmetric, unimodal, bell-shaped continuous probability distribution.
- The normal distribution is the most famous (and perhaps the most important) distribution in statistics because of its relationship with the FTS.
- Note, $\overline{x}$ and $\hat{p}$ will have a sampling distribution that is normally distributed centered around the **parameter** (true population mean/proportion)
- Note, not all statistics are normally distributed.

Sampling Distributions Continued

- We cannot always generate a sampling distribution.
- For example, we may want the sampling distribution for:
  - **Political polls**. How many people voted for a specific candidate?
  - **Taxi Cab Fares**. How much would I spend on taxi cabs in NYC?
  - **Biological data**. What is the growth rate for a plant with a new fertilizer?
  - **Health data**. What percent of patients respond well to new treatment?
- It is not feasible to conduct multiple samples, multiple statistics, and then look at the sampling distribution to understand the behavior of your statistic.

Instead we can consider bootstrapping.

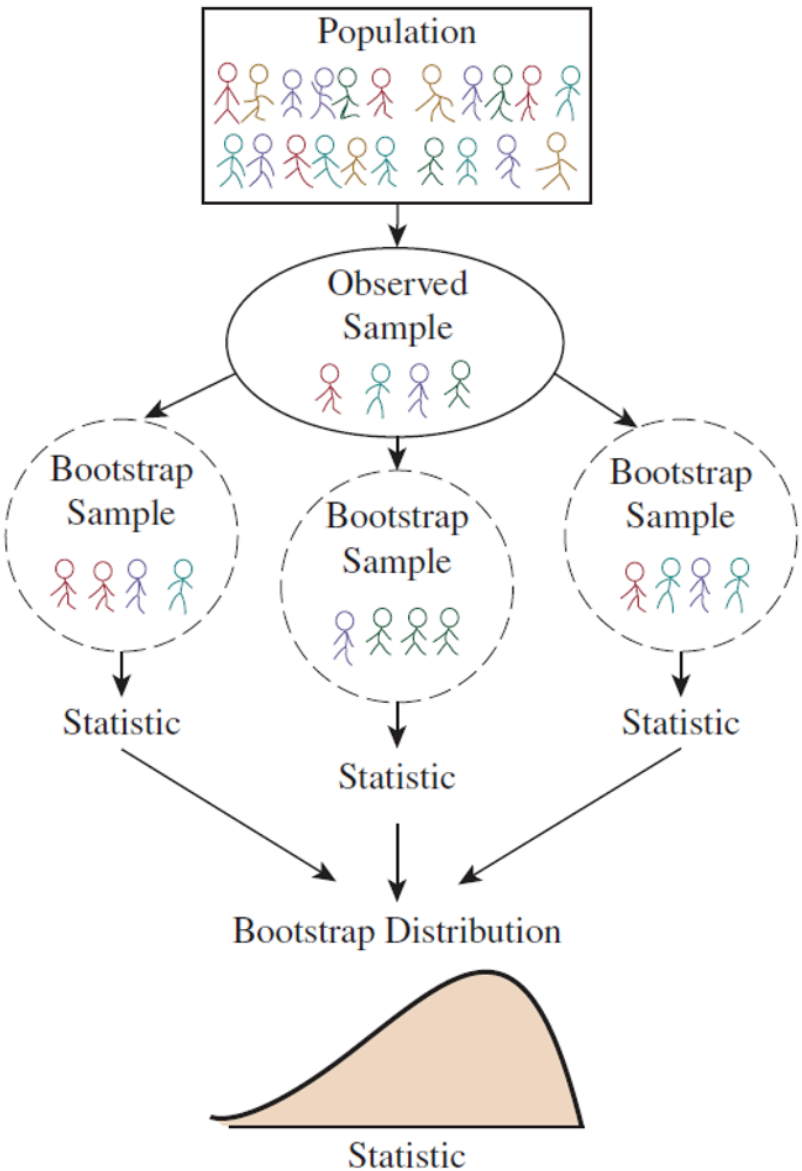
Bootstrapping

In bootstrapping we simulate drawing a random sample from the *observed data*. That is, we resample by repeatedly taking samlpes from the original sample.

Definition

A **bootstrap sample** is a sample drawn *with replacement* from the original sample, and of the same sample size as the original sample.

Bootstrap



We can 'pull ourselves up by our bootstraps' to attack the problem by using computer simulations.

Bootstrap Resampling

To use the bootstrap method with your original data set with n observations:

1) You resample, with replacement, n observations from the data distribution.
2) For the new bootstrap sample of size n, construct the point estimate of the parameter of interest.
3) Repeat the process a very large number of times, B (e.g., selecting $B = 10{,}000$ separate samples of size n and calculating the 10,000 corresponding parameter estimates).

Bootstrap Distribution

For each bootstrap sample, we can compute a statistic of interest, such as a mean.

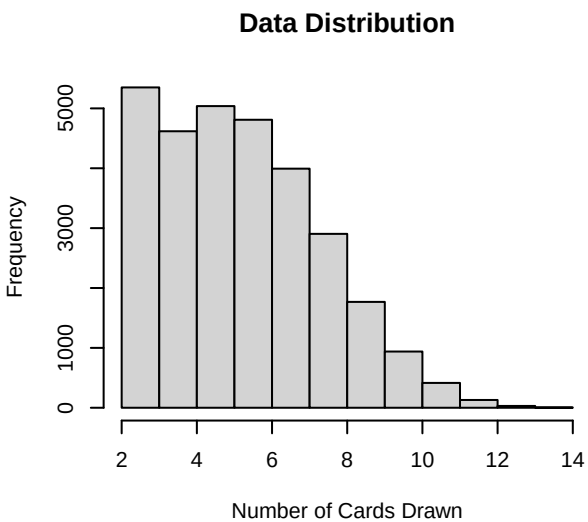
We compute the sample mean from thousands of bootstrap samples. The distribution of all these means, called the **bootstrap distribution**, will help us estimate the *sampling distribution* of the sample mean without having to take samples over and over again.

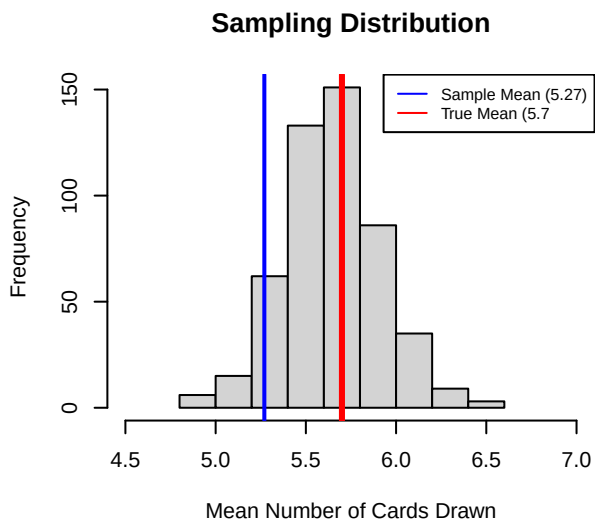
Features of the Bootstrap Distribution

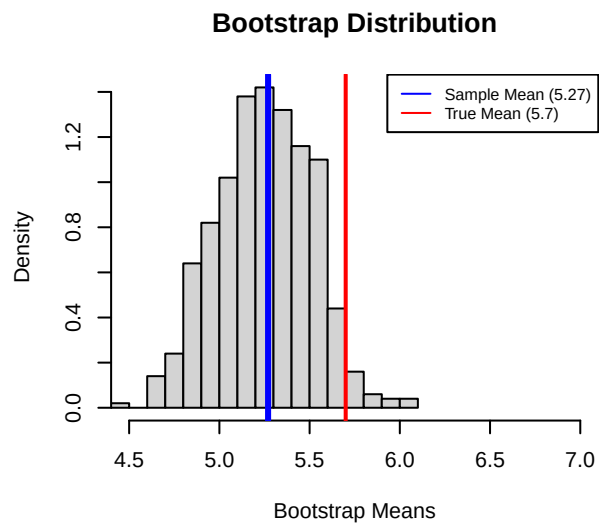
- **Center**: the observed sample statistic
  - This differs from the sampling distribution, which is centered around the true population parameter.
- **Spread**: even though the means of the bootstrap distribution and the sampling distribution are not the same, their spreads are.
  - The bootstrapped statistic vary about original sample statistic in the same way that the original sample proportions vary about true parameter.

Features of the Bootstrap Distribution

Suppose when we sampled cards last lecture one of our groups had a mean 5.27, but the *true* mean was 5.7 cards.







Limitations of the Bootstrap

- It is essential that the original sample is a random sample from the population, or at least representative of it.
- Larger samples (big n) are typically better for bootstraps.
- When the bootstrap distribution consists of only a few values and is highly discrete, it is of limited use and should not be used.
- We typically need a large number of bootstrap resamples ($B \geq 5000$) to obtain a bootstrap distribution that reasonably approximates the key features of the sampling distribution

Why are these things important?