Sampling Distributions and Central Limit Theorem

Ryan Miller

Introduction

- Video #1
  - Sampling Distributions
- Video #2
  - Central Limit Theorem
- Video #3
  - Interval Estimation

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*
- The act of data collection is a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*
- The act of data collection is a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group
- Further, *any descriptive summary* of our sample data (ie: means, proportions, correlations, etc.) is the *observed outcome* of a *random variable*

The Sample Average as a Random Variable

- Consider a sample of n cases from a population, the sample average is calculated:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

- In addition to usefulness in describing the center of a quantitative variable's distribution, lots of statistical theory has been developed for understanding variability in sample averages

Proportions are Averages

- Now, consider a *binary categorical* variable
  - Because binary variables involve only two categories, we can map their outcomes to the numeric values of 0 and 1
  - For example, consider a coin flip, we could map the outcome "Heads" to "1" and the outcome "Tails" to "0"

Proportions are Averages

- Now, consider a *binary categorical* variable
  - Because binary variables involve only two categories, we can map their outcomes to the numeric values of 0 and 1
  - For example, consider a coin flip, we could map the outcome "Heads" to "1" and the outcome "Tails" to "0"
- By coding the outcomes using 1s and 0s, we can see that the *sample proportion* is also an average:

$$\hat{p} = \frac{1+0+1+1+0+...+1}{n}$$

- If we mapped "Heads" to a value of "1", $\hat{p}$ would refer to the proportion of heads in our sample

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates
- Randomly sampling n adults represents a *random process*
  - The proportion of college graduates in a sample, $\hat{p}$, is a *random variable*

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates
- Randomly sampling n adults represents a *random process*
  - The proportion of college graduates in a sample, $\hat{p}$, is a *random variable*
- Let's explore some different outcomes of this random variable for two different sampling protocols: random samples of size $n = 10$, and random samples of size $n = 100$

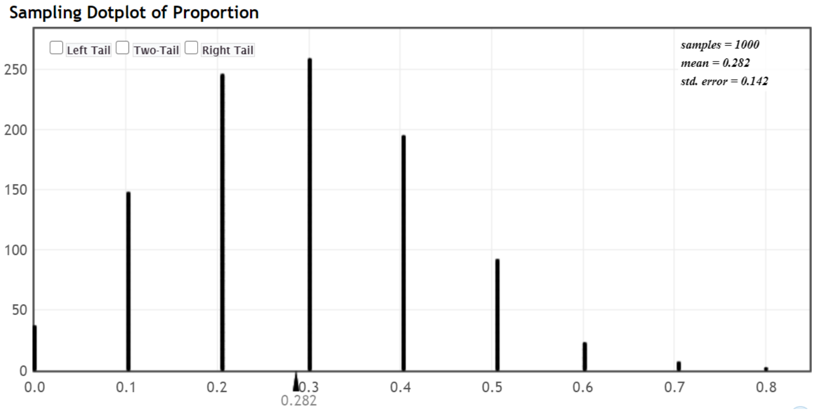
Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that could occur
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$

Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that could occur
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$
- Rather than trying to perform probability calculations, we'll instead look at repeatedly drawing different random samples (of size $n = 10$) to judge the likelihood of each of these outcomes

Random Samples of size $n = 10$



- Each dot represents the proportion of college graduates *in a different random sample* of size $n = 10$

Random Samples of size $n = 10$

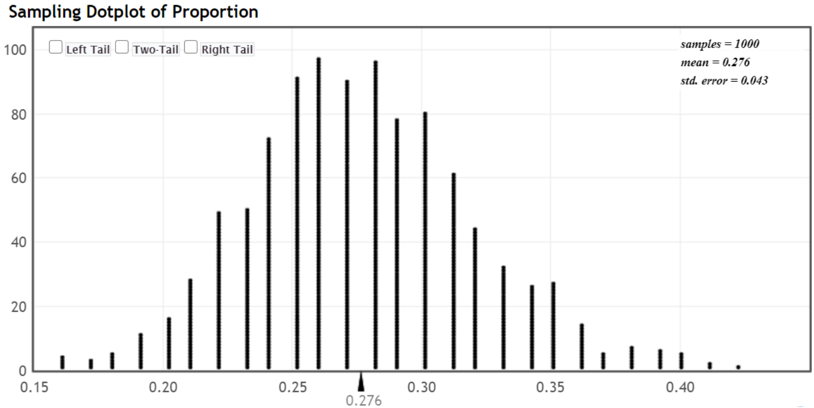
- Due to the relatively small number of discrete outcomes, it's reasonable to use a table to convey a probability model for the sample proportion:

Sample Proportion ($n = 10$)	Probability
$0/10$	$40/1000 = 0.04$
$1/10$	$150/1000 = 0.15$
$2/10$	$250/1000 = 0.25$
$3/10$	$270/1000 = 0.27$
$4/10$	$190/1000 = 0.19$
...	...
$10/10$	$0/1000 = 0$

Random Samples of size $n = 100$

- For a random sample of $n = 100$, there are now 101 discrete outcomes that could be observed for the sample proportion $\{0/100, 1/100, 2/100, \ldots, 100/100\}$
  - At this point, it's impractical to write-out probabilities using a table, instead it makes more sense to treat the sample proportion as a *continuous random variable*

Random Samples of size $n = 100$



- Notice this distribution is roughly *bell-shaped*, it's *centered* at the population proportion (approximately), and has a spread described by the *standard error*

A Normal Model?

- You might be thinking that we can apply a Normal model here, but getting the proper Normal distribution requires us to get the center and spread correct
  - StatKey will report these values (which it finds via simulation), but we'll discuss their origin at length in the next couple of videos

Sampling Distributions

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*)
  - Sampling variability is quantified by the **standard error**, which describes the expected average distance of a sample estimate from its expected value

Sampling Distributions

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*)
  - Sampling variability is quantified by the **standard error**, which describes the expected average distance of a sample estimate from its expected value
- For example, different random samples of size $n = 100$ (of US adults) yield sample proportions (of college graduates) that are *on average* 0.043 off from their expected value of 0.275
  - Different random samples of size $n = 10$ yield sample proportions that are on average 0.142 off from their expected value of 0.275

Sampling Distributions

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*)
  - Sampling variability is quantified by the **standard error**, which describes the expected average distance of a sample estimate from its expected value
- For example, different random samples of size $n = 100$ (of US adults) yield sample proportions (of college graduates) that are *on average* 0.043 off from their expected value of 0.275
  - Different random samples of size $n = 10$ yield sample proportions that are on average 0.142 off from their expected value of 0.275
  - This should make sense, larger samples contain more information about the population and therefore provide estimates that are more reliable (ie: tend to have less sampling error)

Closing Remarks

- So far, we've seen that descriptive statistics calculated using sample data can be viewed as a realized outcome of a *random variable*
- To understand the role of random chance in observed sample data, we can explore the sampling distributions of these random variables
  - If we can come up with an accurate probability model for the sampling distribution, we can use it to evaluate random chance as a viable explanation for trends that occur in our sample data

John Kerrich

- John Kerrich, a South African mathematician, was visiting Copenhagen in 1940
- When Germany invaded Denmark he was sent to an internment camp, where he spend the next five years
- To pass time, Kerrich conducted experiments exploring sampling and probability theory
  - One of these experiments involved flipping a coin 10,000 times

Kerrich's Experiment and Probability

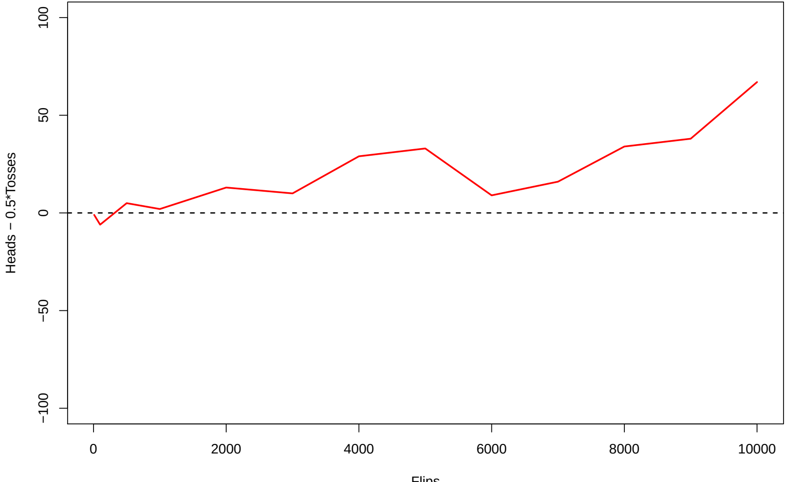
- We know that a fair coin shows "Heads" with a probability of 50%
- So, in a random sample of n coin flips, we'd expect roughly even numbers of "Heads" and "Tails"
  - We'll explore the results of Kerrich's experiment to see why the *sample average* is so special

Kerrich's Results

Number of Tosses (n)	Number of Heads	Heads - 0.5*Tosses
10	4	-1
100	44	-6
500	255	5
1,000	502	2
2,000	1,013	13
3,000	1,510	10
4,000	2,029	29
5,000	2,533	33
6,000	3,009	9
7,000	3,516	16
8,000	4,034	34
9,000	4,538	38
10,000	5,067	67

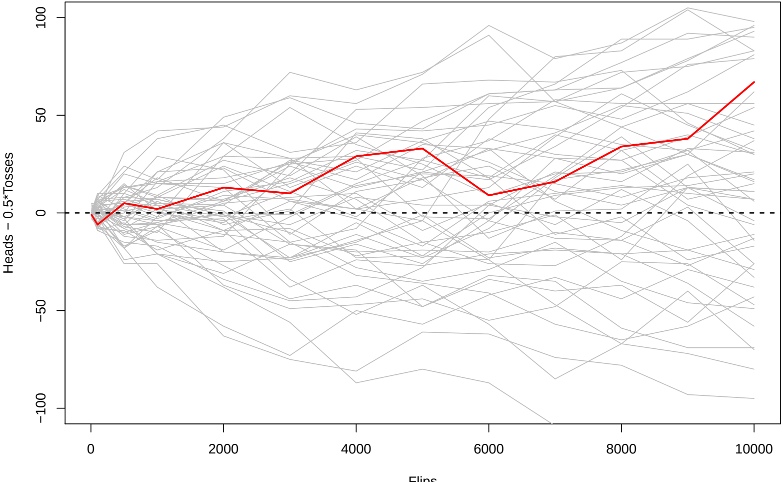
Kerrich's Results

It seems like the number of heads and tails are actually getting further apart... could this be a fluke?



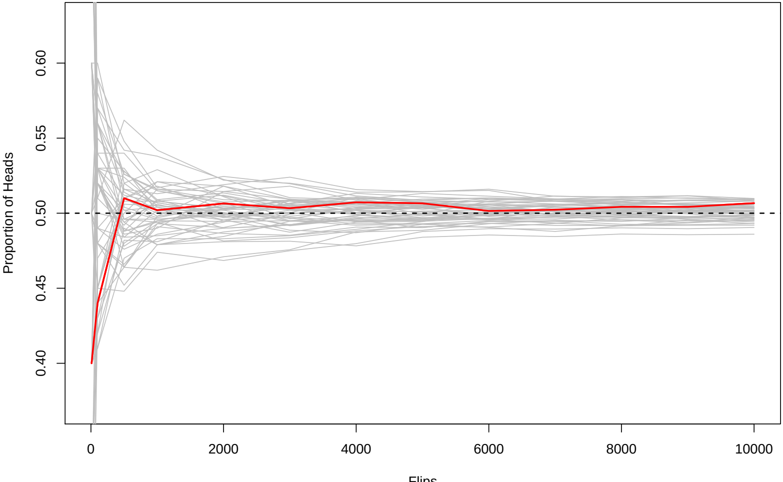
Kerrich's Experiment (repeated 50 times)

No, the phenomenon occurs systematically when repeating Kerrich's experiment



Kerrich's Experiment (sample proportions)

The *sample proportion* of heads behaves exactly as we'd expect, but why?



Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \to N(0,1)$$

Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \to N(0, 1)$$

- Often it is more useful to think of CLT in the following way (which abuses notation):

$$\bar{X} \sim N\left(E(X), \frac{SD(X)}{\sqrt{n}}\right)$$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$
- Thus, the *sampling distribution* of sample proportions is:

$$\hat{p} \sim N(p, \sqrt{p(1 - p)/n})$$

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)
- But by focusing on the *sample average* this isn't an issue, as CLT provides us the distribution of sample averages
  - That is, we are able to use CLT to understand the *sampling variability* of our study, despite only getting to see a single sample!

Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...

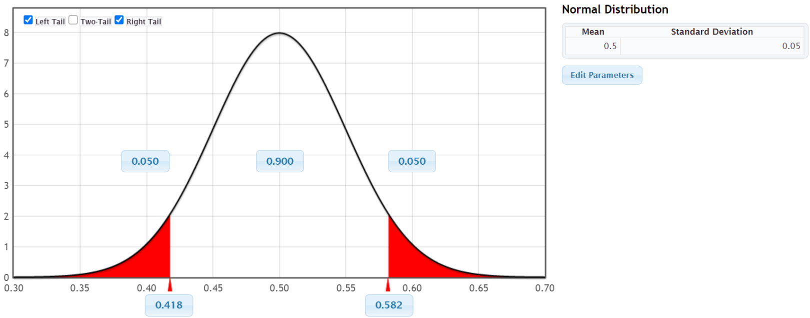
Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...
- Each coin flip is a random variable an expected value of 0.5, so Central Limit Theorem tells us that proportion of heads in random samples of $n = 100$ coin flips follows a Normal distribution:

$$\hat{p} \sim N(0.5, \sqrt{0.5(1-0.5)/100})$$

- To understand the sampling variability of $n = 100$ coin flips, we might look at the *interval* that defines what we'd expect to see 90% of the time

Example



- We'd expect 90% of different random samples to result in sample proportions between 0.418 and 0.582

Assumptions

Using the Central Limit theorem to determine the distribution of sample averages is only appropriate when the following conditions are met:

1) *Independence* - the cases in the sample (ie: the individual contributions to the sample average) are not related to each other
2) *Large population* - less that 10% of the population is being sampled (otherwise removing the already sampled individuals has too much of an impact on the probability of selection)
3) *Large sample* - $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$

Most of the time, it's only the third condition that is problematic

Closing Remarks

- Central Limit theorem provides a theoretical basis for focusing on *sample averages* when attempting to characterize a population
  - Put differently, CLT allows us to understand the *sampling variability* of the sample average without needing to actually take multiple different samples!

Sampling and Statistical Inference

- A *fundamental goal* of statisticians is to use information from a sample to make *reliable* statements about a population
  - This idea is called **statistical inference**

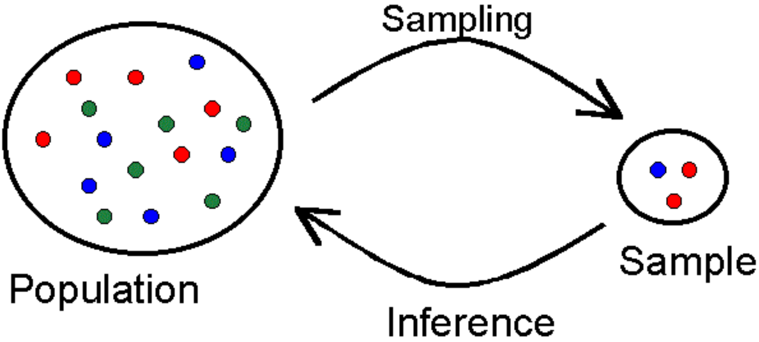


Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

Statistical Inference - Notation

Statisticians use different notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample). For a few common measures, this notation is summarized below:

	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

For example, μ is the mean of the target population, while $\bar{x}$ is the mean of the cases that ended up in the sample

Point Estimation

- If a sampling protocol is *unbiased*, the sample average is a sensible estimate of the population mean
  - This is called a **point estimate**, referring to the fact that it is a single value

Point Estimation

- If a sampling protocol is *unbiased*, the sample average is a sensible estimate of the population mean
  - This is called a **point estimate**, referring to the fact that it is a single value
- From our study of *sampling distributions*, we know that the existence of sampling variability means a point estimate is almost certainly wrong (at least to some degree)
  - This suggests that we can more appropriately describe what we think is true of the population by reporting an **interval estimate** that accounts for *sampling variability*

Point vs. Interval Estimation

To summarize:

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

Point vs. Interval Estimation

To summarize:

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

An analogy:

> *Using only a point estimate is like fishing in a murky lake with a spear. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.*

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

- We'd like the *margin of error* to be constructed in way that carries a *quantifiable* claim of precision
  - ie: 80% of the time an interval with this type of margin of error will contain the population characteristic
  - Without an accompanying claim regarding precision, reporting a margin of error is not particularly useful

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"

Statistical Inference

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
  - False - point estimates have variability

Statistical Inference

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
  - False - point estimates have variability
- True or false? "We observed $\hat{p} = 0.14$, it's probably true 14% of all babies are born prematurely"

Statistical Inference

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
  - False - point estimates have variability
- True or false? "We observed $\hat{p} = 0.14$, it's probably true 14% of all babies are born prematurely"
  - False - sampling distributions show the point estimate is likely off by some degree

Statistical Inference

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
  - False - point estimates have variability
- True or false? "We observed $\hat{p} = 0.14$, it's probably true 14% of all babies are born prematurely"
  - False - sampling distributions show the point estimate is likely off by some degree
- True or false? "Although we don't know p, if we attach a large margin error to our point estimate,the interval estimate $14\% \pm 10\% = (4\%, 24\%)$ probably contains p"

Statistical Inference

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
  - False - point estimates have variability
- True or false? "We observed $\hat{p} = 0.14$, it's probably true 14% of all babies are born prematurely"
  - False - sampling distributions show the point estimate is likely off by some degree
- True or false? "Although we don't know p, if we attach a large margin error to our point estimate,the interval estimate $14\% \pm 10\% = (4\%, 24\%)$ probably contains p"
  - False - we don't know how reliable this margin of error is, perhaps an MOE of 10% is not wide enough

Conclusion

- This presentation introduces the idea of interval estimation
  - The key concept is that point estimates are almost always off, but by attaching a margin of error we can more reliably describe the population of interest
- In class this week, we'll further explore this concept and learn how to use sampling distributions to come up with interval estimates that have *meaningful margins of error*