Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Hand in HW1.
2. 1.96 SE and theory based confidence intervals and used cars example.
3. t versus normal and assumptions.
4. Factors affecting the width of a CI.
5. Bradley effect.
6. Statistical and practical significance, and longevity example.
7. Observational studies: association, confounding, and nightlights example.

Finish reading chapter 4.
http://www.stat.ucla.edu/~frederic/13/F18 .
HW2 is due Thu Oct25 and is problems 2.3.15, 3.3.18, and 4.1.23.
2.3.15 starts "Consider a manufacturing process that is producing hypodermic needles that will be used for blood donations."
3.3.18 starts "Reconsider the investigation of the manufacturing process that is producing hypodermic needles. Using the data from the most recent sample of needles, a 90% confidence interval for the average diameter of needles is...."
4.1.23 starts "In November 2010, an article titled 'Frequency of Cold Dramatically Cut with Regular Exercise' appeared in *Medical News Today*."

1.96 SE method for 95% CIs for a proportion

- The 1.96 SE method only gives us a 95% confidence interval
- The method is valid provided n is large, which in the case of 0-1 data means there are at least 10 successes and 10 failures in your sample.

- The value 1.96 comes from using the normal distribution to approximate our simulated null distribution.
- This gives us a formula for confidence intervals.

$$\hat{p} \pm \textit{multiplier} \times \sqrt{\hat{p}(1-\hat{p})/n}.$$

For a 95% CI, the book suggests a multiplier of 2.
Actually it should be 1.96, not 2.

qnorm(.975) = 1.96.

qnorm(.995) = 2.58.

- Let’s check out this example using the theory-based method.
- Remember 69% of 1034 respondents were not affected.

$\widehat{p} \pm multiplier \times \sqrt{\hat{p}(1-\hat{p})/n}$

$= 69\% \pm 1.96 \times \sqrt{.69(1-.69)/1034}$

$= 69\% \pm 2.82\%$.

With 2 instead of 1.96 it would be $69\% \pm 2.88\%$.

1.96 SE and formulas for confidence intervals for the mean of a quantitative variable.

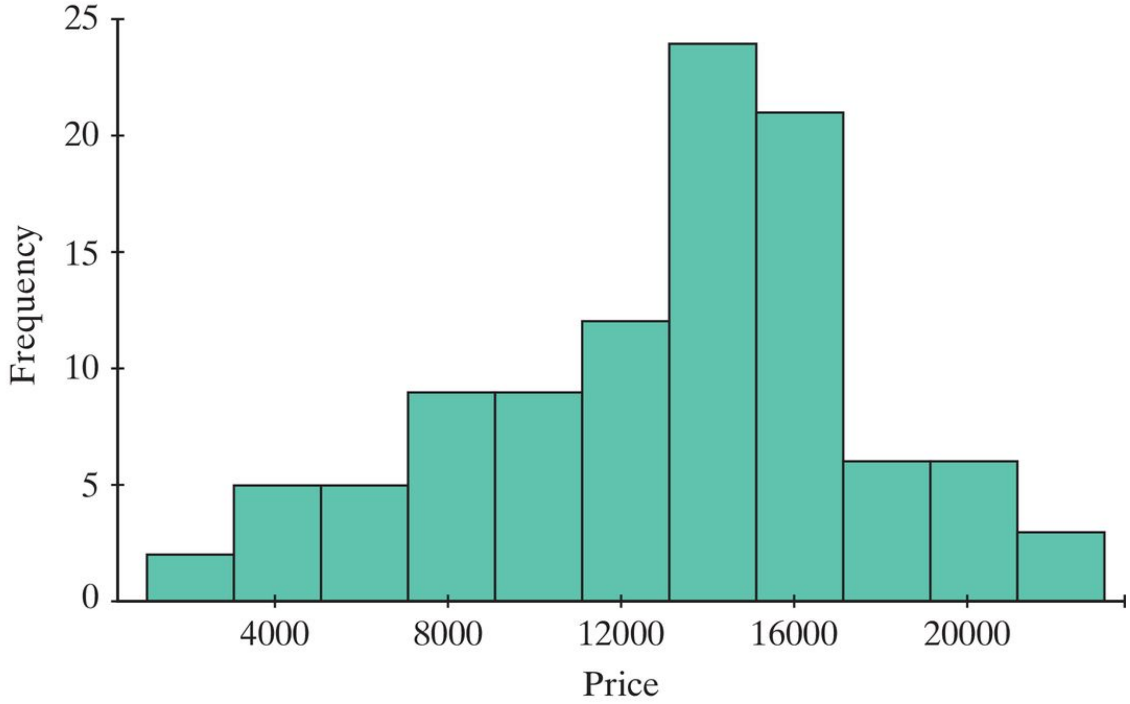
Section 3.3

Used Cars

Example 3.3

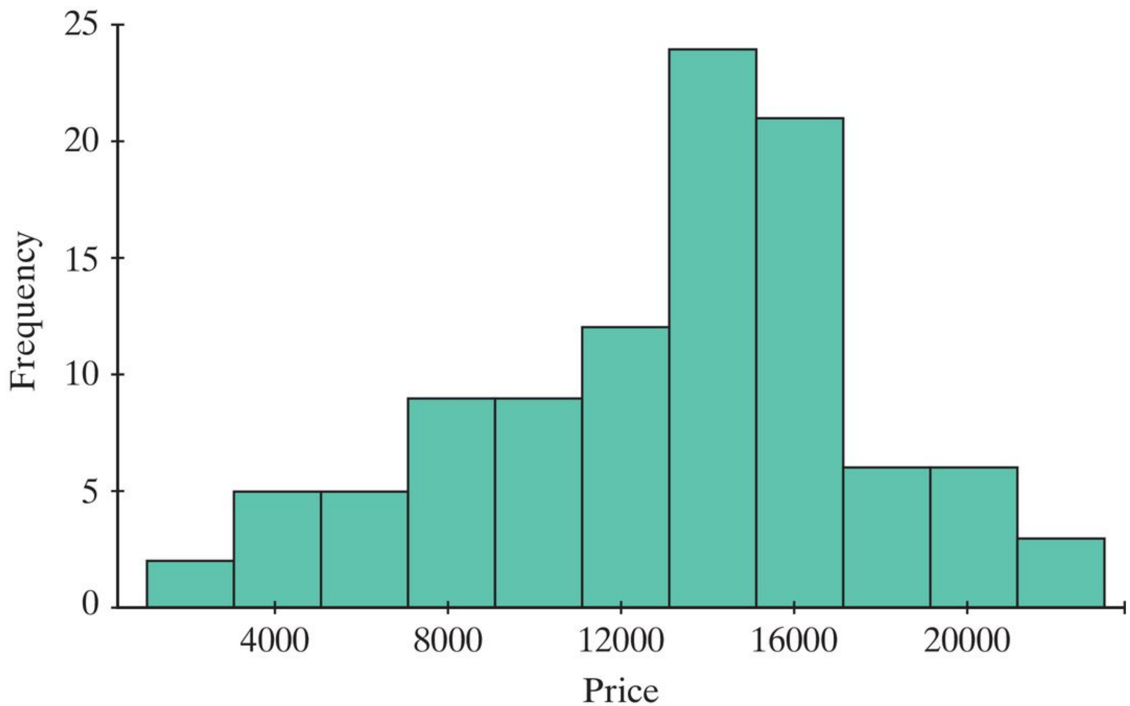
Used Cars

The following histogram displays data for the selling price of 102 Honda Civics that were listed for sale on the Internet in July 2006.



Used Cars

- The average of this sample is $\bar{x}$ = \$13,292 with a standard deviation of s = \$4,535.
- What can we say about μ, the average price of all used Honda Civics?



Used Cars

- While we should be cautious about our sample being representative of the population, let's treat it as such.
- μ might not equal $13,292 (the sample mean), but it should be close.
- To determine how close, we can construct a confidence interval.

Confidence Intervals

- Remember the basic form of a confidence interval is:

statistic ± multiplier × SE

The book sometimes uses the term SD of statistic instead of Standard Error (SE).

- In our case, the statistic is $\bar{x}$ and for a 95% confidence interval, when n is large, the multiplier is 1.96 so we write our 95% confidence interval as:

$$\bar{x} \pm 1.96(\text{SE})$$

Confidence Intervals

- It is important to note that the SE, i.e. the SD of $\bar{x}$, and the SD of our sample (*s* = $4,535) are not the same.
- There is more variability in the data, the car-to-car variability, than in sample means.
- The SE for a sample mean is $s/\sqrt{n}$. This means we can write a 1.96 SE confidence interval for the sample mean as:

$$\bar{x} \pm 1.96 \times \frac{s}{\sqrt{n}}$$

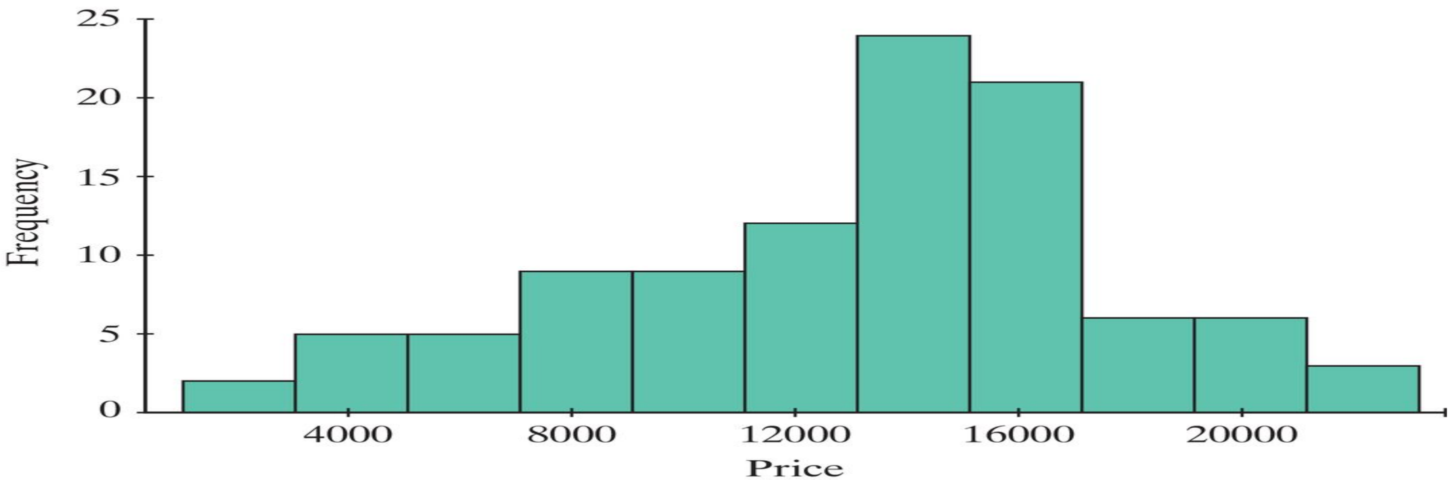
- This method will be valid when n is large. the null distribution is normally distributed.

Summary Statistics

- What your book calls a theory-based confidence interval for the mean of a quantitative variable is quite similar except it uses a multiplier that is based on a *t*-distribution and is dependent on the sample size and confidence level.
- For a theory-based confidence interval for a population mean using the t distribution (called a one-sample t-interval) to be valid, the observations should be approximately independent, and **the population should be normal**. Check the sample distribution for skew and asymmetry.

Confidence Intervals

- We find our 95% CI for the mean price of all used Honda Civics is from $12,411.90 to $14,172.10.
- Notice that this is a much narrower range than the prices of all used Civics.
- For a 99% confidence interval, it would be wider. The multiplier would be 2.58 instead of 1.96.



t versus normal and assumptions.

Why do we sometimes use the t distribution and sometimes the normal distribution in testing and confidence intervals?

The central limit theorem (de Moivre 1733) states that, for any iid random variables $X_1, \ldots, X_n$ with mean μ and SD σ,

$(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal, as $n \to \infty$.

iid means independent and identically distributed, like draws from the same large population.

standard means mean 0 and SD 1.

t versus normal and assumptions.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal, as n -> ∞.

If Z is std. normal, then $P(|Z| < 1.96) = 95\%$.

So, if n is large, then

$P(|(\bar{x} - \mu) \div (\sigma/\sqrt{n})| < 1.96) \sim 95\%$.

Mult. by $(\sigma/\sqrt{n})$ and get

$P(|\bar{x} - \mu| < 1.96\ \sigma/\sqrt{n}) \sim 95\%$.

$P(\mu - \bar{x}$ is in the range $0 \pm 1.96\ \sigma/\sqrt{n}) \sim 95\%$.

$P(\mu$ is in the range $\bar{x} \pm 1.96\ \sigma/\sqrt{n}) \sim 95\%$.

This all assumes n is large. What if n is small?

t versus normal and assumptions.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ -> standard normal, as $n \to \infty$.

What about if n is small?

A property of the normal distribution is that the sum of independent normals is also normal, and from this it follows that if $X_1, \ldots, X_n$ are iid and normal, then $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ is standard normal.

So again P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) = 95%.

This assumes you know σ. What if σ is unknown?

t versus normal and assumptions.

Suppose $X_1, ..., X_n$ are iid with mean μ and SD σ.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ ~ std. normal, as $n \to \infty$.

If $X_1, ..., X_n$ are normal, then $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ is std. normal.

σ is the SD of the population from which $X_1, ..., X_n$ are drawn. s is the SD of the sample, $X_1, ..., X_n$.

Gosset/student (1908) showed that replacing σ with s,

if $X_1, ..., X_n$ are normal, then $(\bar{x} - \mu) \div (s/\sqrt{n})$ is t distributed.

So in this situation we need the multiplier from the t distribution.

t versus normal and assumptions.

To sum up,

if the observations are iid and n is large, then

P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) ~ 95%.

If the observations are iid and normal, then

P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) ~ 95%.

If the obs. are iid and normal, then

P(μ is in the range $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$) ~ 95%.

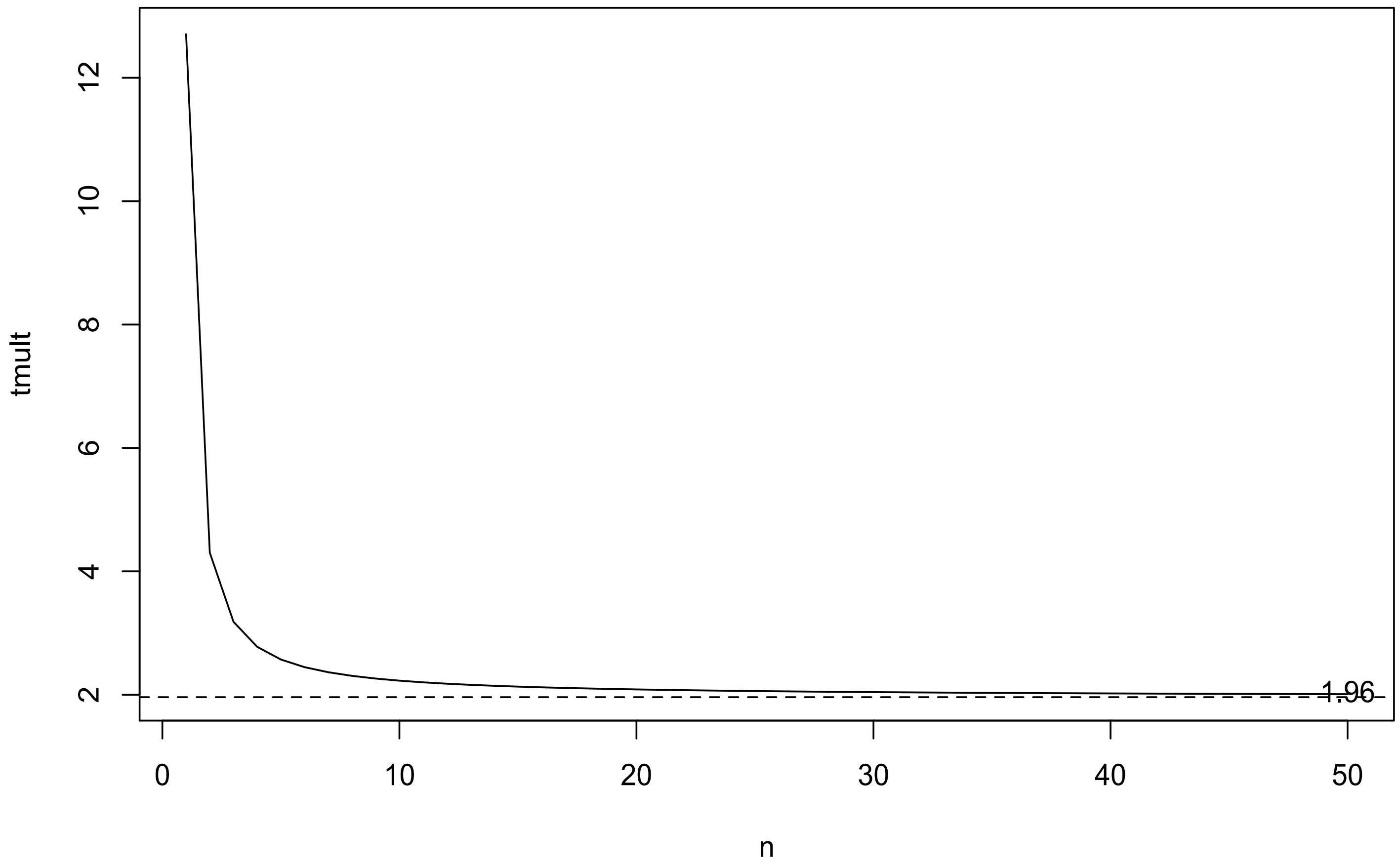
where t_{mult} is the multiplier from the t distribution.

This multiplier depends on n. [qt(.975,df=n-1)]

If n is large (like 30 or more) then t_{mult} is close to 1.96 anyway.

If obs. are iid, n is not large, and the obs. are not normal? Then you need to do simulations to get a 95% CI.

t versus normal and assumptions.



Factors that Affect the Width of a Confidence Interval

Section 3.4

Factors Affecting Confidence Interval Widths

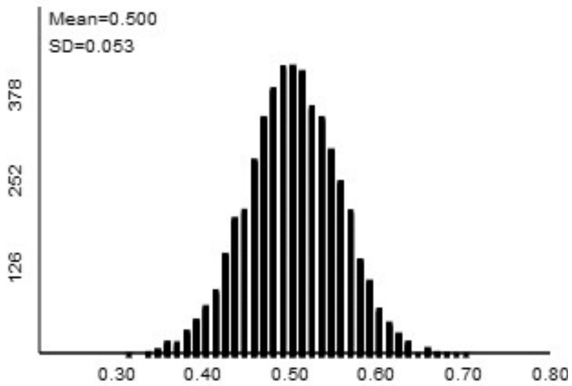
- **Level of confidence** (e.g., 90% vs. 95%)
  - As we increase the confidence level, we increase the width of the interval.
- **Sample size**
  - As sample size increases, variability decreases and hence the standard error will be smaller. This will result in a narrower interval.
- **Sample standard deviation**
  - A larger standard deviation, s, will yield a wider interval.
  - For sample proportions, wider intervals when $\hat{p}$ is closer to 0.5. s = √ [$\hat{p}$ (1-$\hat{p}$)].

Level of Confidence

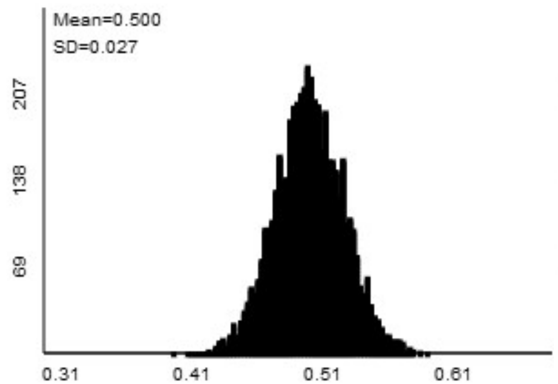
- If we have a wider interval, we should be more confident that we have captured the population proportion or population mean.
- We could see this with repeated tests of significance.
  - A higher confidence level corresponds to a lower significance level, and one must go farther to the left and farther to the right in our tables to get our confidence interval.

Sample Size

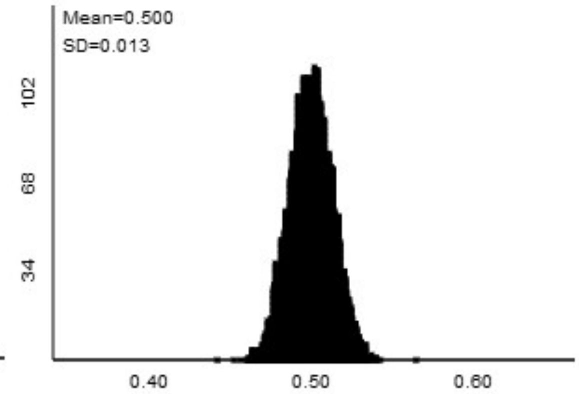
- We know as sample size increases, the variability (and thus standard error) in our null distribution decreases



n = 90 (SD = 0.054)



n = 361 (SD = 0.026)



n = 1444 (SD = 0.013)

Sample size	90	361	1444
SE	0.053	0.027	0.013
Margin of error	2 x SD = 0.106	2 × SD = 0.054	2 × SD = 0.026
Confidence interval	(0.091, 0.303)	(0.143, 0.251)	(0.171, 0.223)

Sample Size

- (With everything else staying the same) increasing the sample size will make a confidence interval narrower.

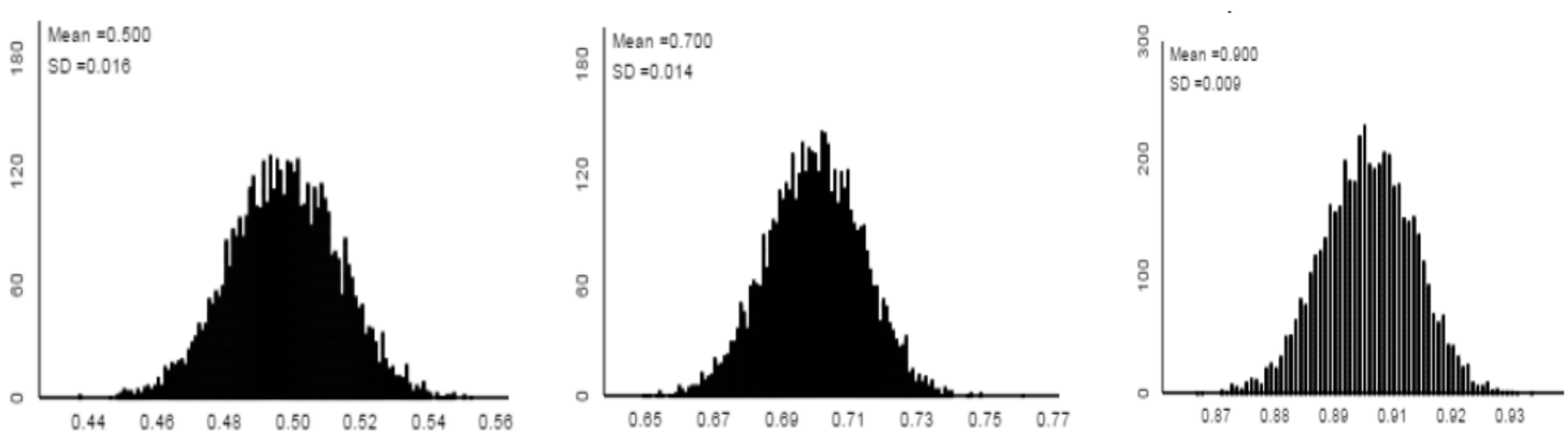
Notice:

- The observed sample proportion is the midpoint. (that won't change)
- Margin of error is a multiple of the standard deviation so as the standard deviation decreases, so will the margin of error.

Value of $\hat{p}$

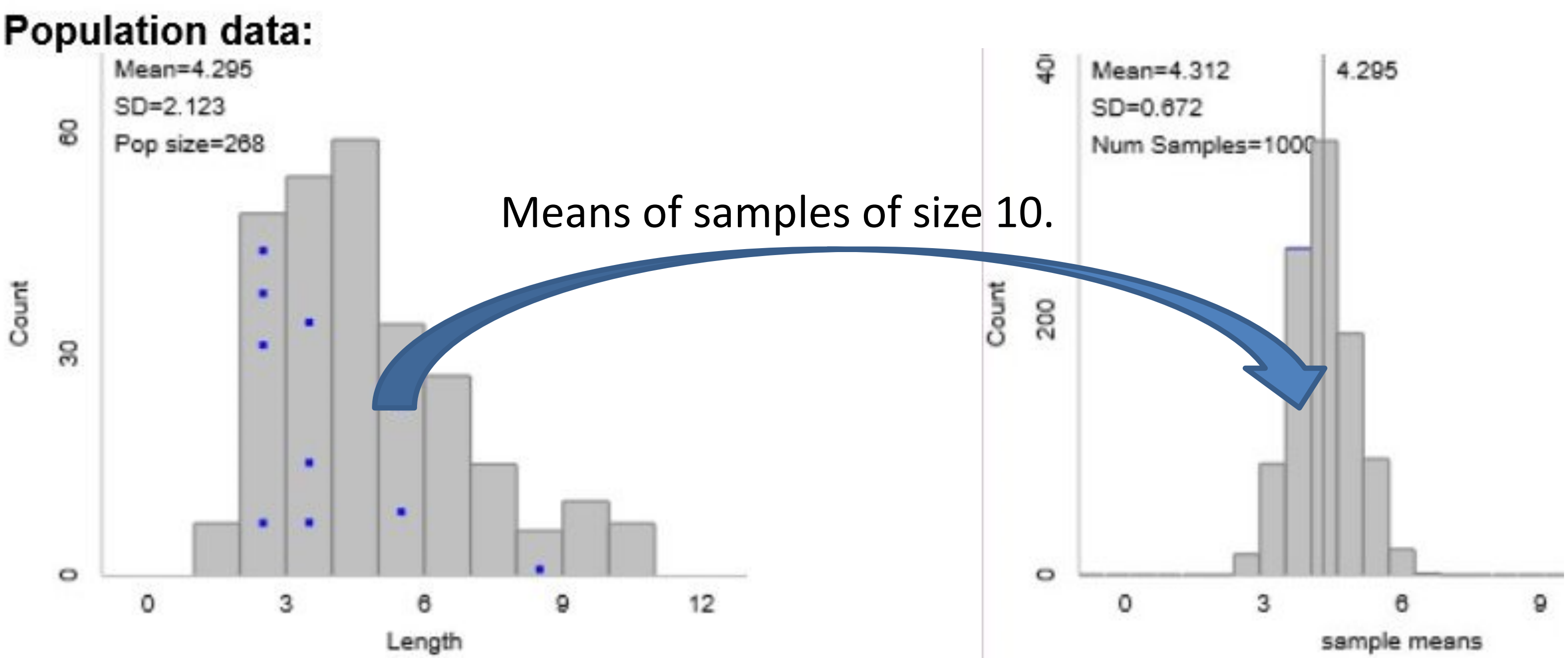
(or the value used for π under the null)

- As the value that is used under the null gets farther away from 0.5, the standard error decreases.
- When this standard error is used in the 1.96 SE method, the interval gets gradually narrower.



Standard Deviation

- Suppose we are taking repeated samples of a population.
- How do we estimate what the standard deviation of the sample mean will be? This is the SE, $s/\sqrt{n}$.



Standard Deviation

- The SE, or SD of the null distribution, is approximated by $s/\sqrt{n}$.
- Remember that $1.96(s/\sqrt{n})$ is approximately the margin of error for a 95% confidence interval for the mean, so as the standard deviation of the data (s) increases so does the width of the confidence interval.
- Intuitively this should make sense, as more variability in the data should be reflected by a wider confidence interval.

Formulas for Theory-Based Confidence Intervals

$$\hat{p} \pm multiplier \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \bar{x} \pm multiplier \times \frac{s}{\sqrt{n}}$$

- The width of the confidence interval increases as level of confidence increases (multiplier)
- The width of the confidence interval decreases as the sample size increases
- The value $\hat{p}$ also has a more subtle effect. The farther it is from 0.5 the smaller the width.
- The width of the confidence interval increases as the sample standard deviation increases.

What does 95% confidence mean?

- If we repeatedly sampled from a population and constructed 95% confidence intervals, 95% of our intervals will contain the population parameter.
- Notice the interval is the random event here.

What does 95% confidence mean?

- Suppose a 95% confidence interval for a mean is 2.5 to 4.3. We would say we are 95% confident that the population mean is between 2.5 and 4.3.
  - Does that mean that 95% of the data fall between 2.5 and 4.3?
    - No
  - Does that mean that in repeated sampling, 95% of the sample means will fall between 2.5 and 4.3?
    - No
  - Does that mean that there is a 95% chance the population mean is between 2.5 and 4.3?
    - Not quite but close.

What does 95% confidence mean?

- What does it mean when we say we are 95% confident that the population mean is between 2.5 and 4.3?
  - It means that if we repeated this process (taking random samples of the same size from the same population and computing 95% confidence intervals for the population mean) repeatedly, 95% of the confidence intervals we find would contain the population mean.
  - P(confidence interval contains μ) = 95%.

Cuations when conducting inference, and the controversial "Bradley Effect"

Example 3.5A

The "Bradley Effect"

- Tom Bradley, long-time mayor of Los Angeles, ran as the Democratic Party's candidate for Governor of California in 1982.
  - Political polls of likely voters showed Bradley with a significant lead in the days before the election
  - Exit polls favored Bradley significantly
  - Many media outlets projected Bradley as the winner
- Bradley narrowly lost the overall race

The “Bradley Effect”

- After the election, research suggested a smaller percentage of white voters had voted for Bradley than polls predicted
- A very large proportion of undecided voters voted for Deukmejian.

The “Bradley Effect”

- What are explanations for this discrepancy?
  - Likely voters answered the questions with a “social desirability bias”
  - They answered polling questions the way they thought the interviewer wanted them to.
- Discrepancies in polling and elections has since been called the “Bradley effect.”
- It has been cited in numerous races and has included gender and other stances on political issues.

Clinton vs. Obama

- In the 2008 New Hampshire democratic primary
  - Obama received 36.45% of the primary votes
  - Clinton received 39.09%.
- This result shocked many since Obama seemed to hold a lead over Clinton.
- USA Today/Gallup poll days before the primary, $n = 778$.
  - 41% of likely voters said they would vote for Obama
  - 28% of likely voters said they would vote for Clinton
- How unlikely are the Clinton and Obama poll numbers given that 39.09% and 36.45% of actual primary voters voted for Clinton and Obama?

Clinton vs. Obama

- We're assuming that the 778 people in the survey are a good representation of those who will vote.
  - The 778 people aren't a simple random sample.
  - Need to have a list of all voters in the election, and randomly choose some.
- Pollsters used random digit dialing and asked if respondents planned to vote in the Democratic primary.
  - 9% (a total of 778) agreed to participate.
  - 319 said that they planned to vote for Obama and 218 for Clinton.

Clinton vs. Obama

Suppose we make the following assumptions:

1. Random digit dialing is a reasonable way to get a sample of likely voters.
2. The 9% who participated are like the 91% who didn't.
3. Voters who said they planned to vote actually voted in the primary.
4. Answers to who they say they will vote for match who they actually vote for.

Then we expect the sample proportion to agree with the final vote proportion.

Clinton vs. Obama

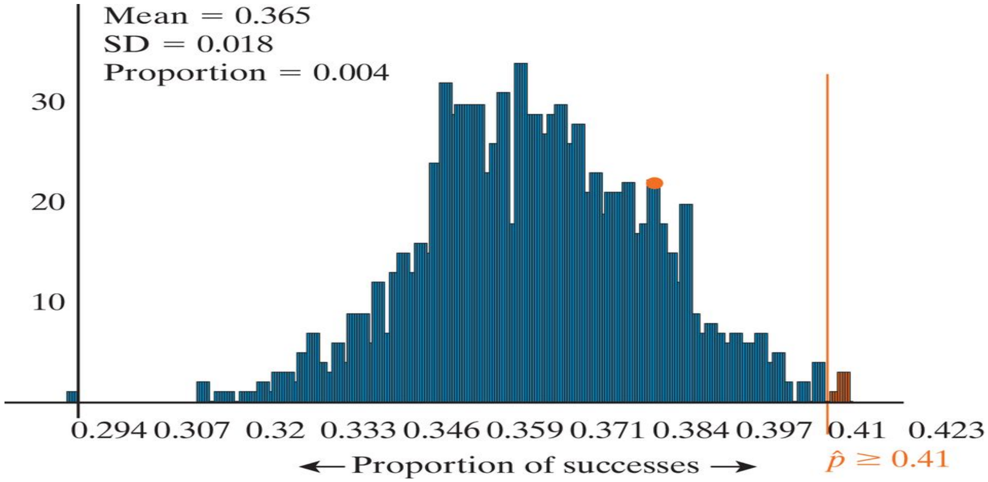
- One question is whether the proportion of likely voters who say they will vote for Obama is the same as the proportion of likely voters who actually vote for Obama (observed on primary day to be 0.3645).
- What would the Bradley Effect do in this case?
  - The proportion who say they will vote for Obama would be larger than 0.3645.

Clinton vs. Obama

- State the Null and Alternative hypotheses
  - Null: The proportion of likely voters who would claim to vote for Obama is 0.3645.
  - Alternative: The proportion of likely voters who would claim to vote for Obama is higher than 0.3645.

Clinton vs. Obama

- Simulation of 778 individuals randomly chosen from a population where 36.45% vote for Obama
- The chance of getting a sample proportion of 0.41 successes or higher is very small. 0.004.



Clinton vs. Obama

- Convincing evidence that the discrepancy between what people said and how they voted is not explained by random chance alone.
- At least one of the 4 model assumptions is not true.

Clinton vs. Obama

1. **Random digit dialing is a reasonable way to get a sample of likely voters**
   - Roughly equivalent to a SRS of New Hampshire residents who have a landline or cell phone
   - Slight over-representation of people with more than one phone

Clinton vs. Obama

2. **The 9% of individuals reached by phone who agree to participate are like the 91% who didn't**
   - 91% includes people who didn't answer their phone and who didn't participate
   - Assumes that respondents are like non-respondents.
   - The ***response rate*** was very low, but typical for phone polls
   - No guarantee that the 9% are representative.

Clinton vs. Obama

3. **Voters who said they plan to vote in the Democratic primary will vote in the primary**
   – There is no guarantee.
4. **Respondent answers to who they say they will vote for matches who they actually vote for.**
   There is no guarantee.

Clinton vs. Obama

Because of the wide disparity between polls and the primary, an independent investigation was done with the following conclusions:

1. People changed their opinion at the last minute
2. People in favor of Clinton were more likely not to respond
3. The Bradley Effect
4. Clinton was listed before Obama on every ballot

These are examples of **nonrandom errors.**

Statistical and Practical significance.

- *Statistically significant* means that the results are unlikely to happen by chance alone.
- *Practically important* means that the difference is large enough to matter in the real world.

Cautions

- Practical importance is context dependent and somewhat subjective.
- Well designed studies try to equate statistical significance with practical importance, but not always.
- Look at the sample size.
  - If n is very large, even small effect sizes will yield significant results.
  - If n is very small, don’t expect significant results. (A lot of missed opportunities---type II errors.)

Longevity example.

According to data from the WHO (2014) and World Cancer Report (2014), the average number of cigarettes smoked per adult per day in the U.S. is 2.967, and in Latvia it is 2.853.

The sample sizes are huge, so even this little difference is stat. sig. (In the U.S., the National Health Interview Survey has $n > 87000$).

If you do not like cigarette smoke around you, should you move to Latvia?

The difference is statistically significant, but not practically significant for most purposes.

Causation.

Chapter 4

- Previously research questions focused on **one** proportion
  - What proportion of the time did Marine choose the right bag?
- We will now start to focus on research questions comparing **two** groups.
  - Are smokers more likely than nonsmokers to have lung cancer?
  - Are children who used night lights as infants more likely to need glasses than those who didn't use night lights?

- Typically we observe two groups and we also have two variables (like smoking and lung cancer).
- So with these comparisons, we will:
  - determine when there is an association between our two variables.
  - discuss when we can conclude the outcome of one variable causes an outcome of the other.

Observational studies and confounding.

Types of Variables

- When two variables are involved in a study, they are often classified as explanatory and response
- **Explanatory variable** (Independent, Predictor)
  - The variable we think may be causing or explaining or used to predict a change in the response variable. (Many times, this is the variable the researchers are manipulating.)
- **Response variable** (Dependent)
  - The variable we think may be being impacted or changed by the explanatory variable.

Roles of Variables

- Choose the explanatory and response variable:
  - Smoking and lung cancer
  - Heart disease and diet
  - Hair color and eye color

- Sometimes there is a clear distinction between explanatory and response variables and sometimes there isn't.

Observational Studies

- The norovirus study is an example of an **observational study**.
- In observational studies, researchers *observe* and measure the explanatory variable but do not set its value for each subject.
- Examples:
  - A significantly higher proportion of individuals with lung cancer smoked compared to same-age individuals who don't have lung cancer
  - College students who spend more time on Facebook tend to have lower GPAs

Observational Studies

Do these studies prove that smoking *causes* lung cancer or Facebook *causes* lower GPAs?

- Many people who see these types of studies think so…
- It depends on the study design

Night Lights and Nearsightedness

Example 4.1

Nightlights and Near-Sightedness

- Near-sightedness often develops in childhood
- Recent studies looked to see if there is an association between near-sightedness and night light use with infants
- Researchers interviewed parents of 479 children who were outpatients in a pediatric ophthalmology clinic
- Asked whether the child slept with the room light on, with a night light on, or in darkness before age 2
- Children were also separated into two groups: near-sighted or not near-sighted based on the child's recent eye examination

Night-lights and near-sightedness

	Darkness	Night Light	Room Light	Total
Near-sighted	18	78	41	137
Not near-sighted	154	154	34	342
Total	172	232	75	479

The largest group of near-sighted kids slept in rooms with night lights. It might be better to look at the data in terms of proportions.

Conditional proportions

$18/172 \approx 0.105$ $78/232 \approx 0.336$ $41/75 \approx 0.547$

Night lights and near-sightedness

	Darkness	Night Light	Room Light	Total
Near-sighted	**10.5%** 18/172	**33.6%** 78/232	**54.7%** 41/75	137
Not near-sighted	154	154	34	342
Total	172	232	75	479

- Notice that as the light level increases, the percentage of near-sighted children also increases.
- We say there is an **association** between near-sightedness and night lights.
- Two variables are **associated** if the values of one variable provide information (help you predict) the values of the other variable.

Night lights and near-sightedness

- While there is an association between the lighting condition nearsightedness, can we claim that night lights and room lights *caused* the increase in near-sightedness?
- Might there be other reasons for this association?

Night lights and near-sightedness

- Could parents' eyesight be another explanation?
  - Maybe parents with poor eyesight tend to use more light to make it easier to navigate the room at night and parents with poor eyesight also tend to have children with poor eyesight.
  - Now we have a third variable of *parents' eyesight*
  - *Parents' eyesight* is considered a ***confounding variable.***
  - Other possible confounders? Wealth? Books? Computers?

Confounding Variables

- A **confounding variable** is associated with both the explanatory variable and the response variable.
- We say it is confounding because its effects on the response cannot be separated from those of the explanatory variable.
- Because of this, we can’t draw cause and effect conclusions when confounding variables are present.

Confounding Variables

- Since confounding variables can be present in observational studies, we can’t conclude causation from these kinds of studies.
- This doesn’t mean the explanatory variable isn’t influencing the response variable. **Association may not imply causation, but can be a pretty big hint.**