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

1. Formula for 1.96SE and theory-based CIs for a proportion.
2. 1.96SE and theory-based CIs for a mean, used car example.
3. When to use which multiplier.
4. Bradley effect.
5. Statistical and practical significance, and longevity example.
6. Causation, observational studies, and confounding. Smoking and facebook examples.

http://www.stat.ucla.edu/~frederic/13/W24 .
HW2 due Mon, Feb12, 1159pm. 2.3.15, 3.3.18, and 4.1.23.
Read chapter 4.
Midterm is Mon Feb26 in class.

1. CIs for a proportion. Formula, or Theory-Based Method

- The 1.96SE method is for a 95% confidence interval.
- If we want a different level of confidence, we can use the range of plausible values (hard) or theory-based methods (easy).
- The theory-based method is valid for CIs for a proportion, provided it's a Simple Random Sample (SRS) and there are at least 10 successes and 10 failures in your sample.

FORMULA FOR CIs FOR A PROPORTION.

- Last time, we relied on simulations to tell us that the SE was 0.016. But we don't need this. In general for testing a proportion, under the null hypothesis, SE = $\sqrt{\pi(1-\pi)/n}$.
- For confidence intervals, we do not assume the null hypothesis, and since π is unknown, use $\widehat{p}$ in its place:

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

For a 95% CI, the book suggests a multiplier of 2. Actually people use 1.96, not 2. This comes from a property of the normal distribution.

qnorm(.975) = 1.96.

qnorm(.995) = 2.58, the multiplier for a 99% CI.

- Going back to the ACA example, recall

69% of 1034 respondents were not affected.

With no default value of π, to get a 95% CI for $\widehat{p}$, use

$\widehat{p}$ ± *multiplier* × $\sqrt{\hat{p}(1-\hat{p})/n}$

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

= 69% ± 2.82%.

With 2 instead of 1.96 it would be 69% ± 2.88%.

This is the formula we actually use for CIs for a proportion.

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

To review, the book first explains how to get a CI by repeated testing, then using the "2 SE" method where the SE is found via simulation, then gives you this formula. But the formula is actually the correct answer. The others are approximations and require simulation.

2. 1.96SE and Theory-Based Confidence Intervals for a Single Mean and used car example.

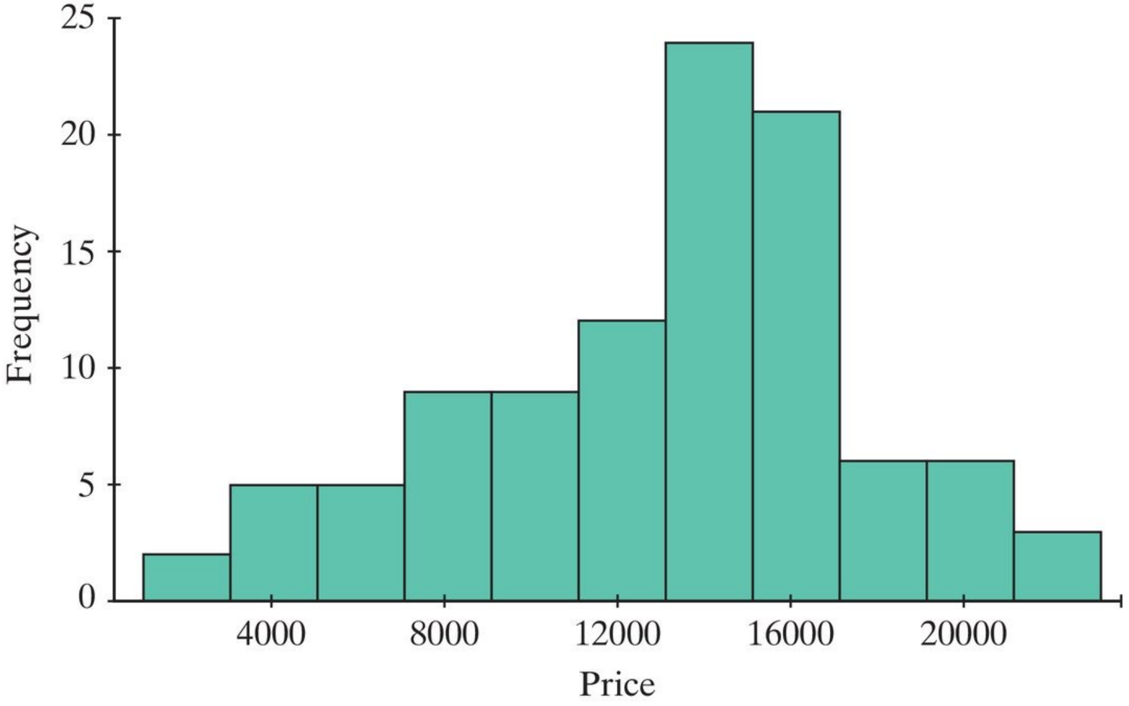
Section 3.3

Used Cars

Example 3.3

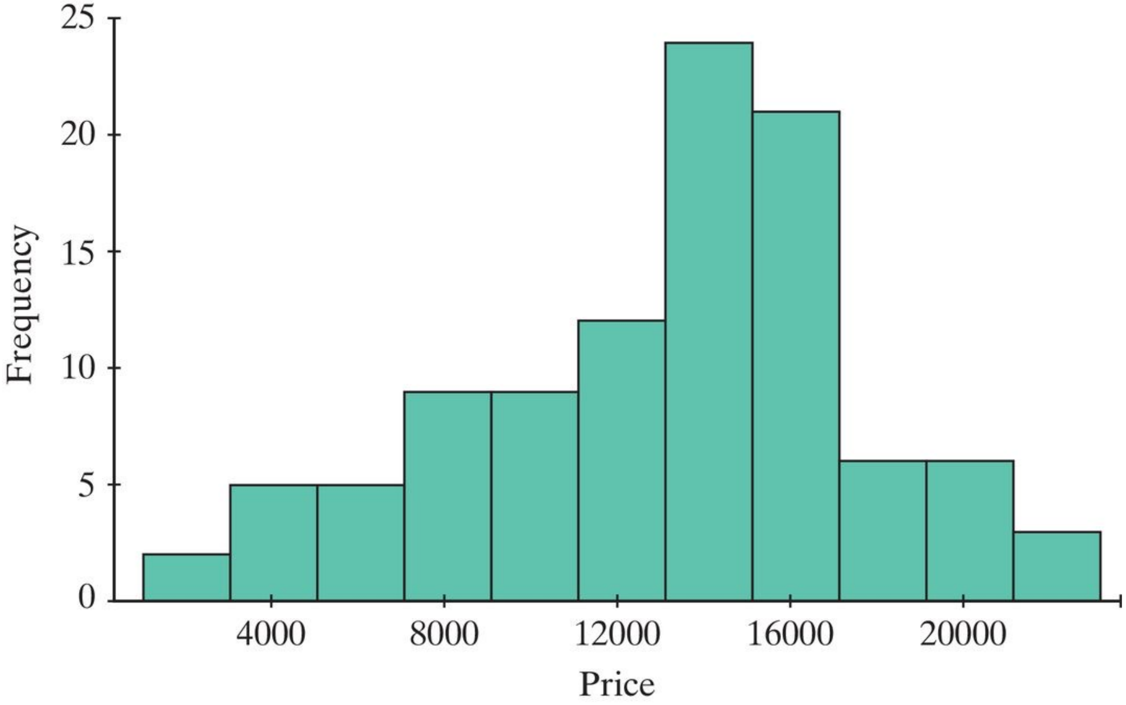
Used Cars

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

  SE is called by the book "SD of statistic".

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

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

Confidence Intervals

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

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

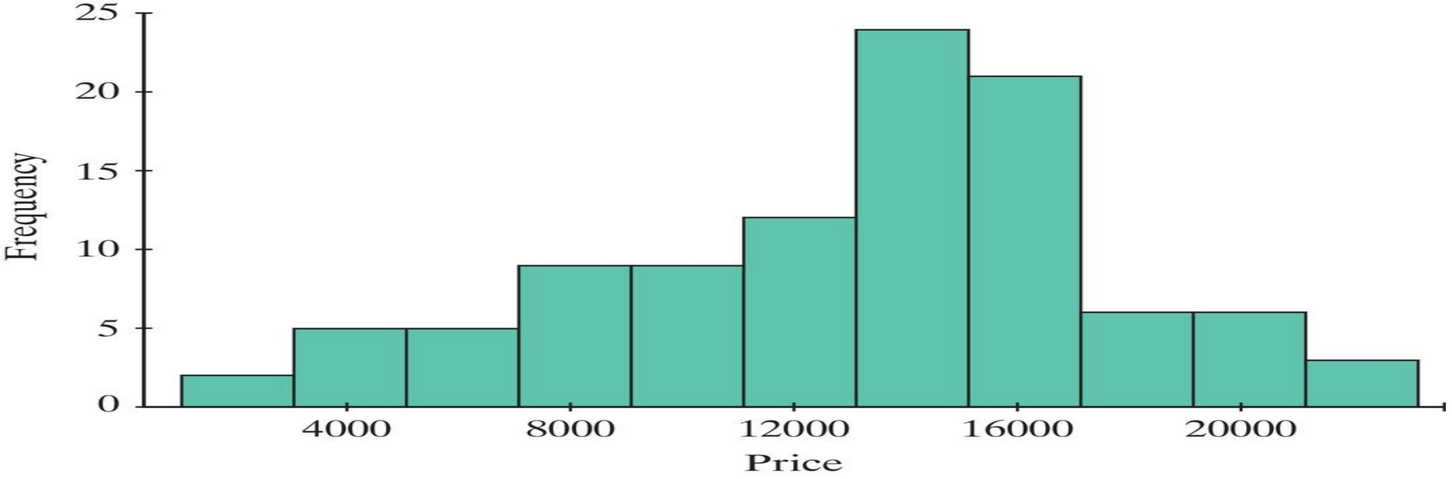
This 1.96 multiplier may be valid when n is large.

Summary Statistics

- When n is small and the population is approximately normal, we will use a multiplier that is based on a *t*-distribution, instead of 1.96. The t multiplier is dependent on the sample size and confidence level.
- For a theory-based confidence interval for a population mean (called a one-sample t-interval) to be valid, the observations should be approximately iid (independent and identically distributed), and either the population should be normal or n should be large. 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 $12,401.20 to $14,182.80.
- 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.



3. For CIs, when to use 1.96 from the normal, & when to use a multiplier based on the t distribution.

iid = independent and identically distributed.

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, and σ is known, then

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

If the obs. are iid and normal and σ is unknown, 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.

For quantitative symmetric data, book says n ≥ 20 is large.

For proportions, need ≥ 10 of each type, in your sample.

4. Cautions 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.
- 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 roughly to agree with the final vote proportion.

Clinton vs. Obama

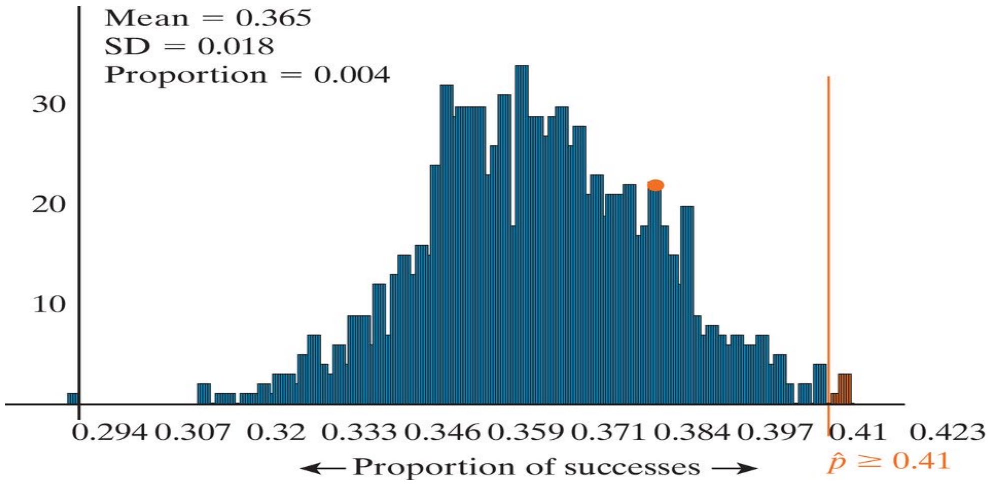
- One question is whether the proportion of likely voters who say they will vote for Obama is the same as the proportion of likely voters who actually vote for Obama (observed on primary day to be 0.3645).
- What would the Bradley Effect do in this case?
  - The proportion who say they will vote for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

5. 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 very large, expect significant results.
  - If 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.

6. Causation, observational studies, and confounding. Smoking and facebook examples.

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 a change in 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. (Often 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.
  - The one we are interested in predicting.

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

- 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.

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