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

1. 1.96SE and formula-based CIs for a proportion, ACA example.
2. Formulas for CIs for a quantitative variable and used car example.
3. When to use which multiplier.
4. Statistical and practical significance, and longevity example.
5. Causation, observational studies, and confounding. Smoking and facebook examples.

Read chapter 4.

Midterm is Tue May5 in class.

HW2, due Fri, May8, 1159pm. 2.3.15, 3.3.18, and 4.1.23.

The course website is http://www.stat.ucla.edu/~frederic/13/S26 .

1. 1.96SE and formula-based Confidence Intervals for a Single Proportion and ACA example.

Section 3.2

Introduction

- Previously we found confidence intervals by doing repeated tests of significance (changing the value in the null hypothesis) to find a range of values that were plausible for the population parameter.
- This is a very tedious way to construct a confidence interval.
- We will now look at two others way to construct confidence intervals [1.96SE and Theory-Based].

The Affordable Care Act

Example 3.2

The Affordable Care Act

- A November 2013 Gallup poll based on a random sample of 1,034 adults asked whether the Affordable Care Act had affected the respondents or their family.
- 69% of the sample responded that the act had no effect. (This number went down to 59% in May 2014 and 54% in Oct 2014.)
- What can we say about the proportion of all adult Americans that would say the act had no effect?

The Affordable Care Act

- We could construct a confidence interval just like we did last time. We get (0.661, 0.717).
- We are 95% confident that the proportion of all adult Americans that felt unaffected by the ACA is between 0.661 and 0.717.

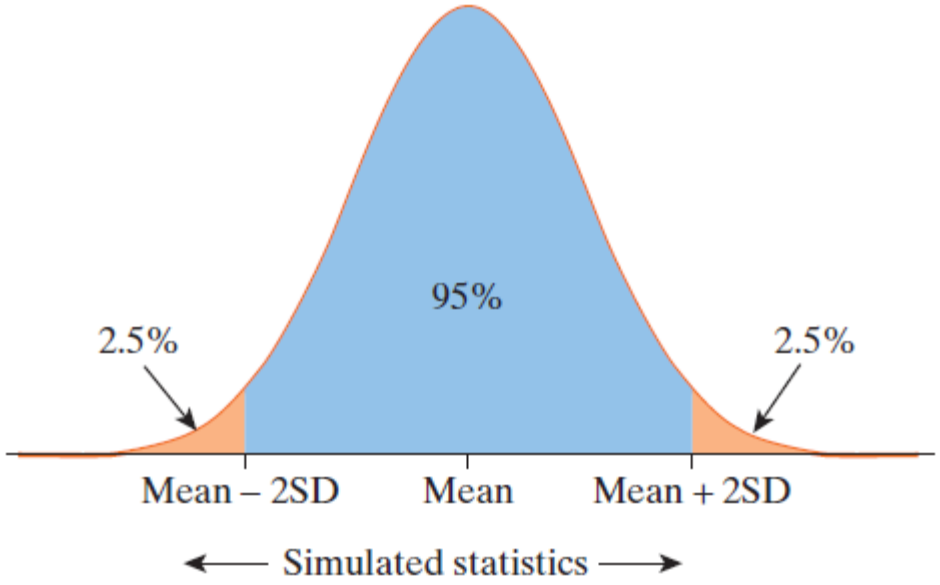
Probability under null	0.659	0.660	0.661	.....	0.717	0.718	0.719
Two-sided p-value	0.0388	0.0453	0.0514	.....	0.0517	0.0458	0.0365
Plausible value (0.05)?	No	No	Yes	.....	Yes	No	No

Short cut?

- The method we used last time to find our interval of plausible values for the parameter is tedious and time consuming.
- Might there be a short cut?
- Our sample proportion should be the middle of our confidence interval.
- We just need a way to find out how wide it should be.

1.96SE method

- When a statistic is normally distributed, about 95% of the values fall within 1.96 standard errors of its mean with the other 5% outside this region



1.96SE method

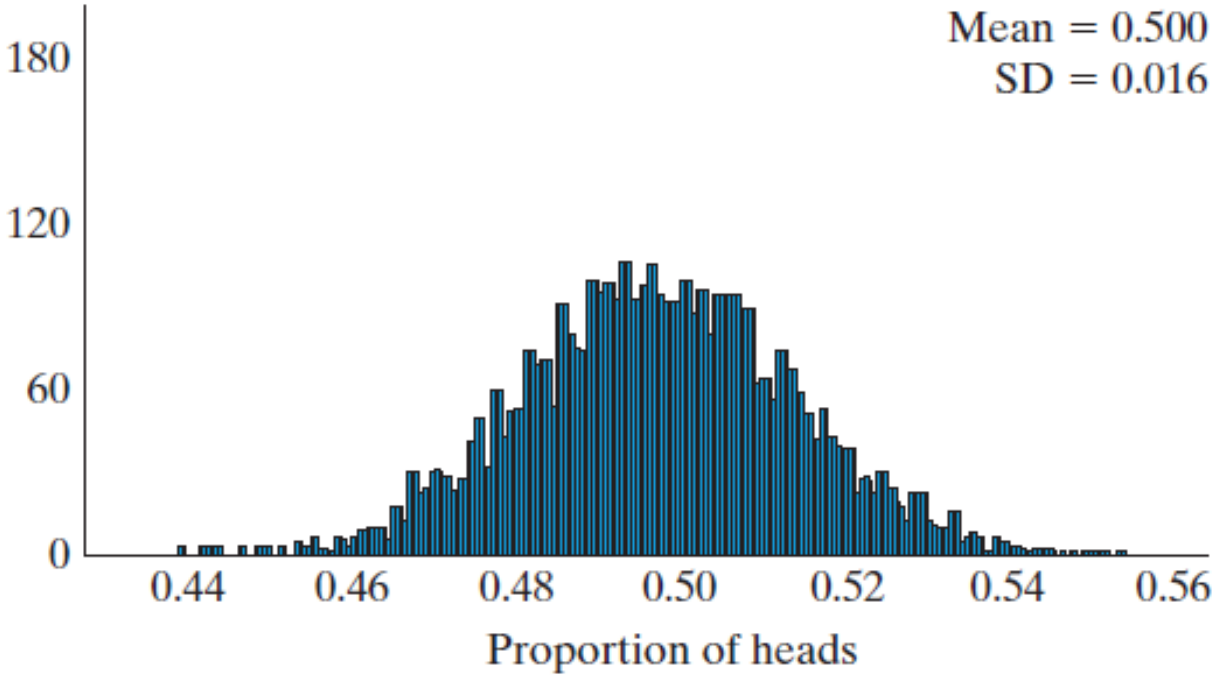
- So we could say that a parameter value is plausible if it is within 1.96 standard errors from our best estimate of the parameter, which is our observed sample statistic.
- This gives us the simple formula for a 95% confidence interval of

$$\hat{p} \pm 1.96SE$$

Note that your book calls this the 2SD method but it really should be called the 1.96SE method.

Where do we get the SE?

- One way is via simulation.
- When the null hypothesis is $\pi = 0.5$, the SE = 0.016.



1.96SE method

- Using the 1.96SE method on our ACA data we get a 95% confidence interval

$$0.69 \pm 1.96(0.016)$$
$$0.69 \pm 0.031$$

- The $\pm$ part, like the 0.031 in the above, is called the **margin of error.**
- The interval can also be written as we did before using just the endpoints; (0.659, 0.721)
- This is approximately what we got using simulations, with our range of plausible values method. We had (0.661, 0.717).

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 (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.

- On the previous slides, 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 $\hat{p}$ in its place:

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

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 $\hat{p}$, use

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

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

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

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

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

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

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 formula based Confidence Intervals for a mean of a quantitative variable and used car example.

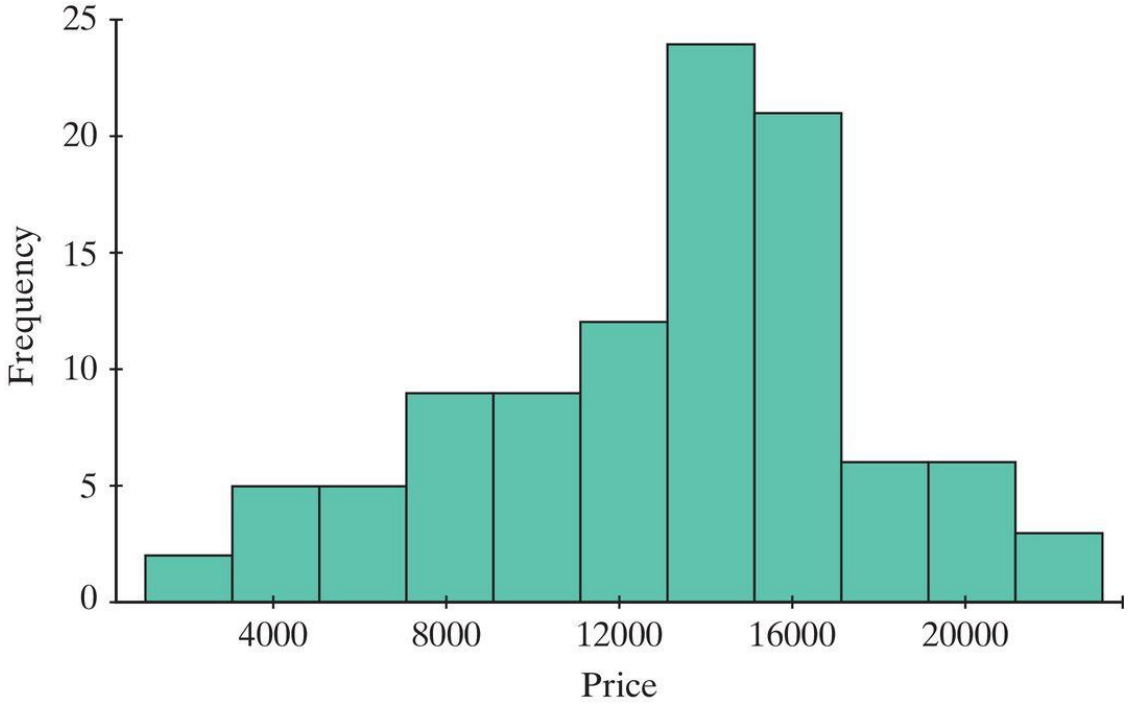
Section 3.3

Used Cars

Example 3.3

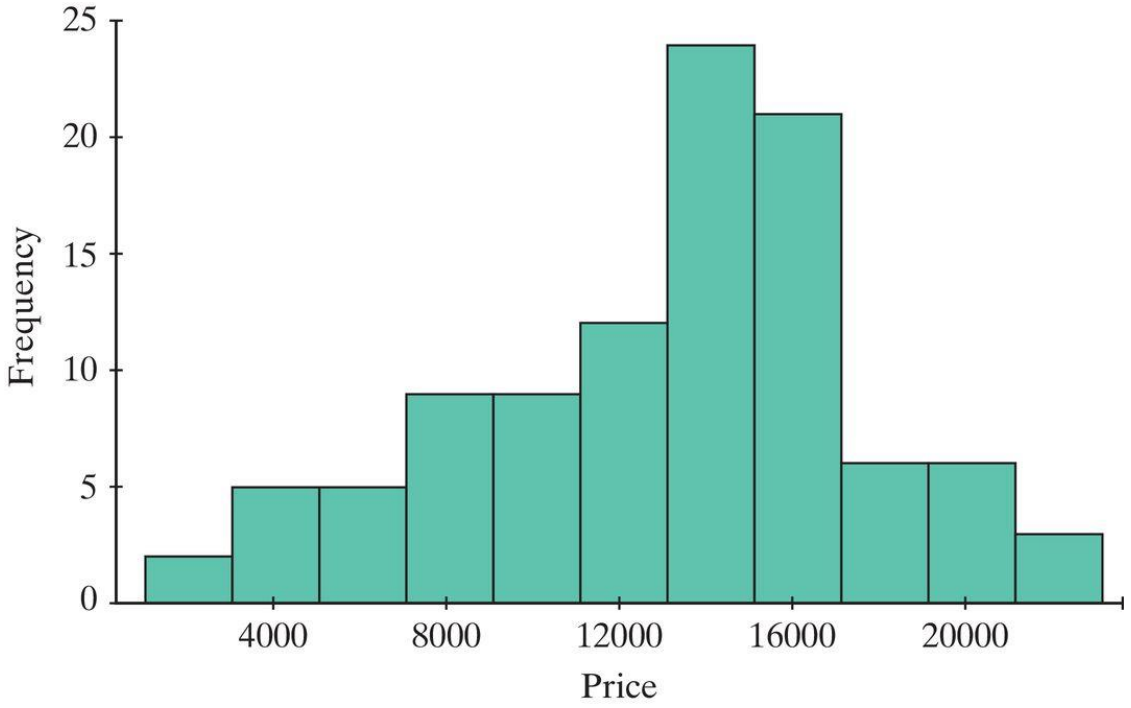
Used Cars

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:

$$\text{statistic} \pm \text{multiplier} \times \text{SE}$$

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 $\bar{x}$, 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

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

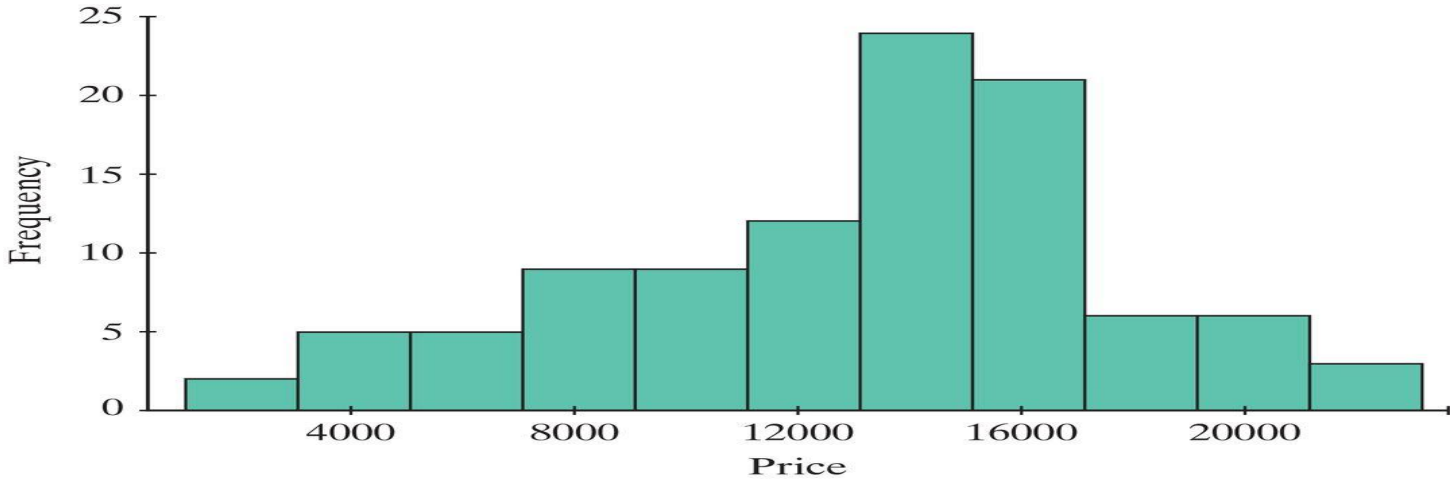
This 1.96 multiplier may be valid when n is large.

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 σ/√n) ~ 95%.

If the observations are iid and normal, and σ is known, then

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

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

This multiplier depends on n.

For quantitative symmetric data, n ≥ 30 is large.

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

4. 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.

5. 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?