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

0. SEs for percentages when testing and for CIs.
1. More about SEs and confidence intervals.
2. Clinton versus Obama and the Bradley effect.
3. Stat sig. versus practical significance.
4. Observational studies and confounding.

Finish reading chapter 4.
http://www.stat.ucla.edu/~frederic/13/F18 .
The midterm is Tue Nov6. There is no lecture Thu Nov1!

SE for percentages when testing and for CIs.

Typically there is some population whose mean μ we want to estimate. The sd of the population is σ. We take a sample of size n, and use the sample mean as our estimate. The sample sd is s.
The SE for our sample mean is given by $\sigma / \sqrt{n}$, but typically we do not know σ, so our estimate of the SE is $s / \sqrt{n}$. We use this formula $s / \sqrt{n}$ when computing confidence intervals for μ.

When we are testing the null hypothesis that μ = some value, then we want to *assume* μ is equal to this value. We still typically do not know σ, so again we would use $s / \sqrt{n}$ as our estimate of the SE.

<u>0-1 data</u>
In the special case where all the values are 0s and 1s, the population mean is typically called π, but really $\pi = \mu$ so it's the same thing.
The sd of a bunch of 0s and 1s is simply $\sqrt{(pq)}$, where p is the proportion of 1s and q is the proportion of 0s. So for 0-1 data, the formulas above still apply, but $\sigma = \sqrt{[\pi (1-\pi)]}$, and $s = \sqrt{[\hat{p} (1-\hat{p})]}$. Since we typically do not know π, we do not know σ, so our estimate of the SE is $s / \sqrt{n} = \sqrt{[\hat{p}(1-\hat{p}) / n]}$. This is the formula we would use in confidence intervals for π.

However, when we are testing the null hypothesis that π = some value, then in this situation, assuming π = the value, we know $\sigma = \sqrt{[\pi (1-\pi)]}$, so in testing we would use $\sqrt{[\pi (1-\pi) / n]}$ as our SE.

Clinton vs. Obama, continued.

- 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

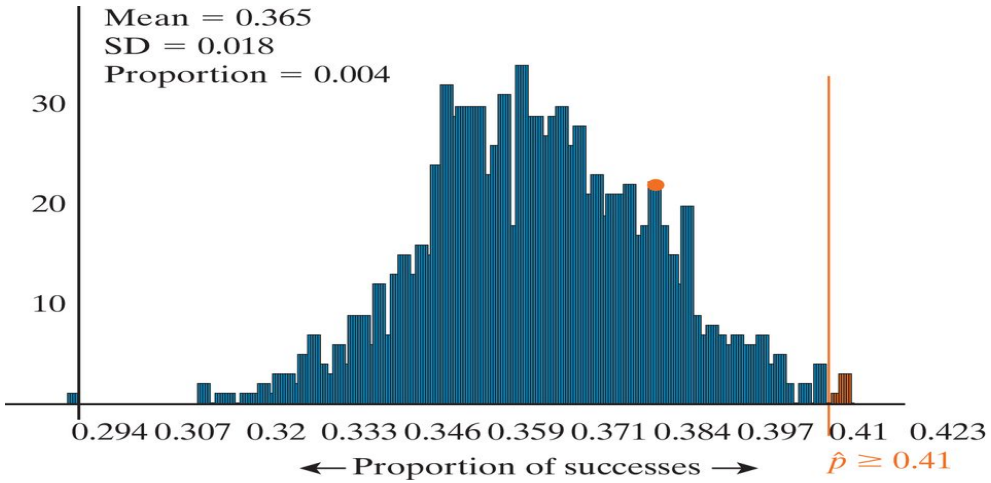
- One question is whether the proportion of likely voters who say they will vote for Obama is the same as the proportion of likely voters who actually vote for Obama (observed on primary day to be 0.3645).
- What would the Bradley Effect do in this case?
  - The proportion who say they will vote for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

Polls in 2016.

a. Why were they so far off?

www.realclearpolitics.com/epolls/2016/president/wi/wisconsin_trump_vs_cli

Polling Data

Poll	Date	Sample	MoE	Clinton (D)	Trump (R)	Johnson (L)	Stein (G)	Spread
Final Results	--	--	--	46.9	47.9	3.6	1.1	Trump +1.0
RCP Average	10/26 - 11/2	--	--	46.8	40.3	5.8	2.0	Clinton +6.5
Remington Research (R)*	11/1 - 11/2	2720 LV	1.9	49	41	3	--	Clinton +8
Loras	10/31 - 11/1	500 LV	4.4	44	38	7	2	Clinton +6
Remington Research (R)*	10/30 - 10/30	1172 LV	2.9	46	42	4	--	Clinton +4
Marquette	10/26 - 10/31	1225 LV	3.5	46	40	4	3	Clinton +6
Emerson	10/26 - 10/27	400 LV	4.9	48	42	9	1	Clinton +6
Remington Research (R)*	10/20 - 10/22	1795 LV	2.3	46	41	5	--	Clinton +5
Monmouth	10/15 - 10/18	403 LV	4.9	47	40	6	1	Clinton +7
WPR/St. Norbert	10/13 - 10/16	644 LV	3.8	47	39	1	3	Clinton +8
Marquette	10/6 - 10/9	878 LV	3.9	44	37	9	3	Clinton +7
CBS News/YouGov	10/5 - 10/7	993 LV	4.3	43	39	4	1	Clinton +4
Loras	10/4 - 10/5	500 LV	4.4	43	35	8	2	Clinton +8
Gravis	10/4 - 10/4	1102 RV	3.0	48	40	4	1	Clinton +8
Emerson	9/19 - 9/20	700 LV	3.6	45	38	11	2	Clinton +7
Marquette	9/15 - 9/18	677 LV	4.8	41	38	11	2	Clinton +3
Monmouth	8/27 - 8/30	404 LV	4.9	43	38	7	3	Clinton +5
Marquette	8/25 - 8/28	650 LV	5.0	41	38	10	4	Clinton +3
Marquette	8/4 - 8/7	683 LV	5.0	47	34	9	3	Clinton +13
Marquette	7/7 - 7/10	665 LV	4.5	43	37	8	2	Clinton +6
CBS News/YouGov*	6/21 - 6/24	993 LV	4.3	41	36	3	2	Clinton +5

2. Polls.

In total this makes 17,104 likely voters in those Wisconsin polls put together.

They averaged **40.3%** for Trump, and Clinton 46.8%. The difference is 6.5%.

Combined, the margin of error for a 95% confidence interval around Trump's percentage would be 0.735%.

The standard error is 0.375% on the estimate of Trump's percentage of 40.3%,

and he got **47.9%**. So they were off by 7.6% which is more than 20

standard errors. The probability is 1 in 10^90 that the polls would be off by that much or more just by chance, if the answers to the polls were just a random sample of how people were actually going to vote.

Technically, there are undecided voters in the polls also. Just taking the difference in percentages between Trump and Hillary Clinton rather than the percentage for Trump into account, the results were off by about 10 SEs, not 20, and this makes the probability of something this extreme or more extreme still astronomical, about 1 in 10^{23}.

The chance of a monkey randomly typing 15 letters completely at random and happening to choose "hillary r clinton" in order, would be 1 in 6 x 10^{22}.

What do we conclude?

Polls.

What do we conclude?

Either

* lots of people changed their minds,

* the polls were biased,

* the official results were incorrect,

or

* the polls weren't independent of each other.

Those are really the only tenable explanations.

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 ≈ 0.105 78/232 ≈ 0.336 41/75 ≈ 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.**