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

1. t versus normal and assumptions.
2. Confounding and lefties example.
3. Comparing two proportions using numerical and visual summaries,
   good or bad year example.
4. Comparing 2 proportions with CIs + testing using simulation, dolphin example.
5. Comparing 2 props. with theory-based testing, smoking and gender example.

Read ch5.
http://www.stat.ucla.edu/~frederic/13/F17 .
Bring a PENCIL and CALCULATOR and any books or notes you want to the exams.
The midterm is Tue Nov 7.
There is no lecture Thu Nov 2 because of faculty retreat and also no lecture Thu Nov 9.
The first exam will be on everything through today. We will do review next class.

1. 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 states that, for any iid random variables $X_1, ..., X_n$ with mean μ and SD σ, $(\bar{x} - \mu) \div (\sigma/\sqrt{n}) \to$ 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: $(\bar{x}$ - μ) ÷ (σ/√n) -> standard normal.

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

So, if n is large, then

P(|$(\bar{x}$ - μ) ÷ (σ/√n)| < 1.96) ~ 95%.

Mult. by (σ/√n) and get

P(|$\bar{x}$ - μ| < 1.96 σ/√n) ~ 95%.

P(μ $-$ $\bar{x}$ is in the range 0 +/- 1.96 σ/√n) ~ 95%.

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

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

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

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

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

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

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

So 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 and σ is unknown, then

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

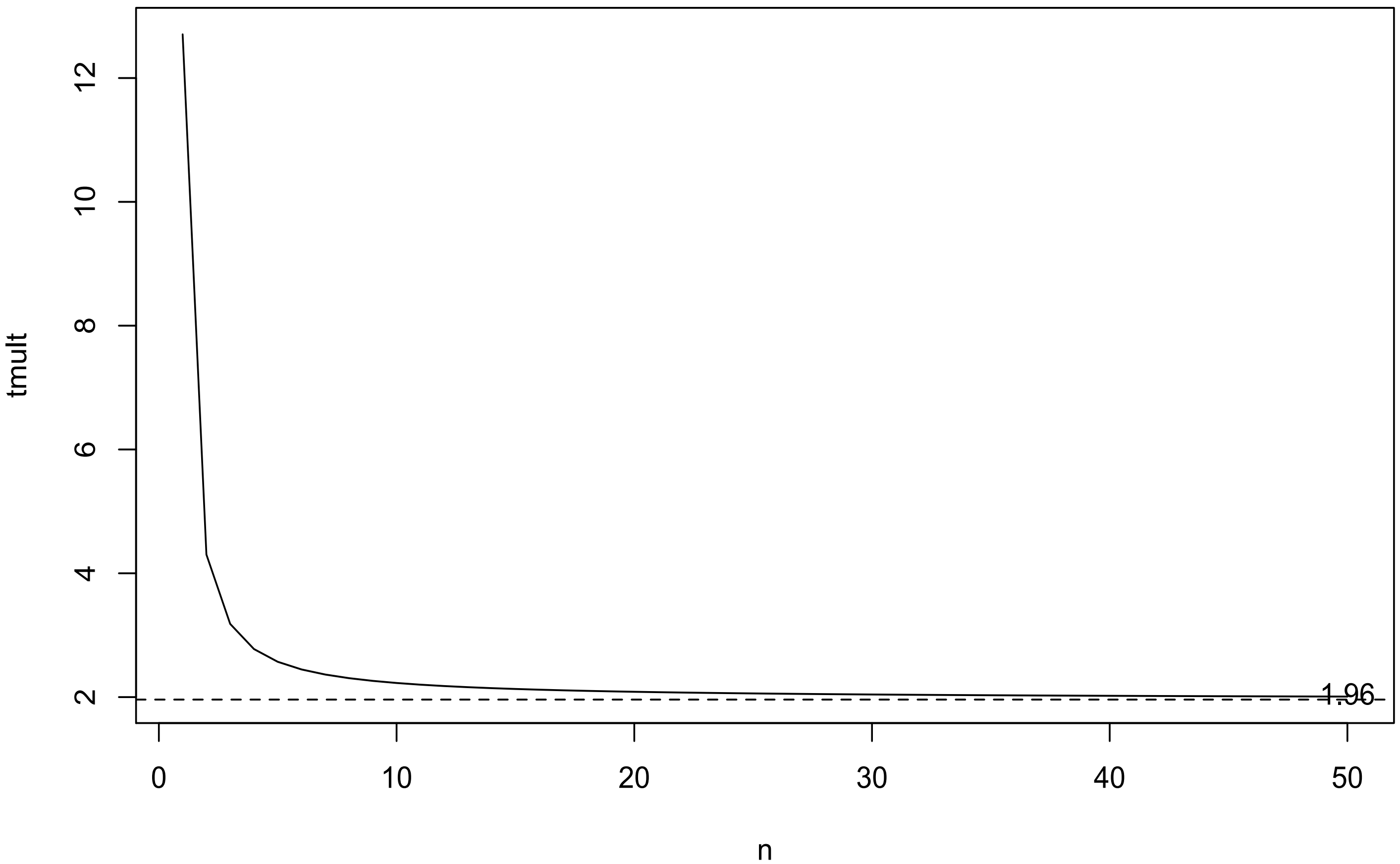
where t_{mult} is the multiplier from the t distribution.

This multiplier depends on n.

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.

t versus normal and assumptions.

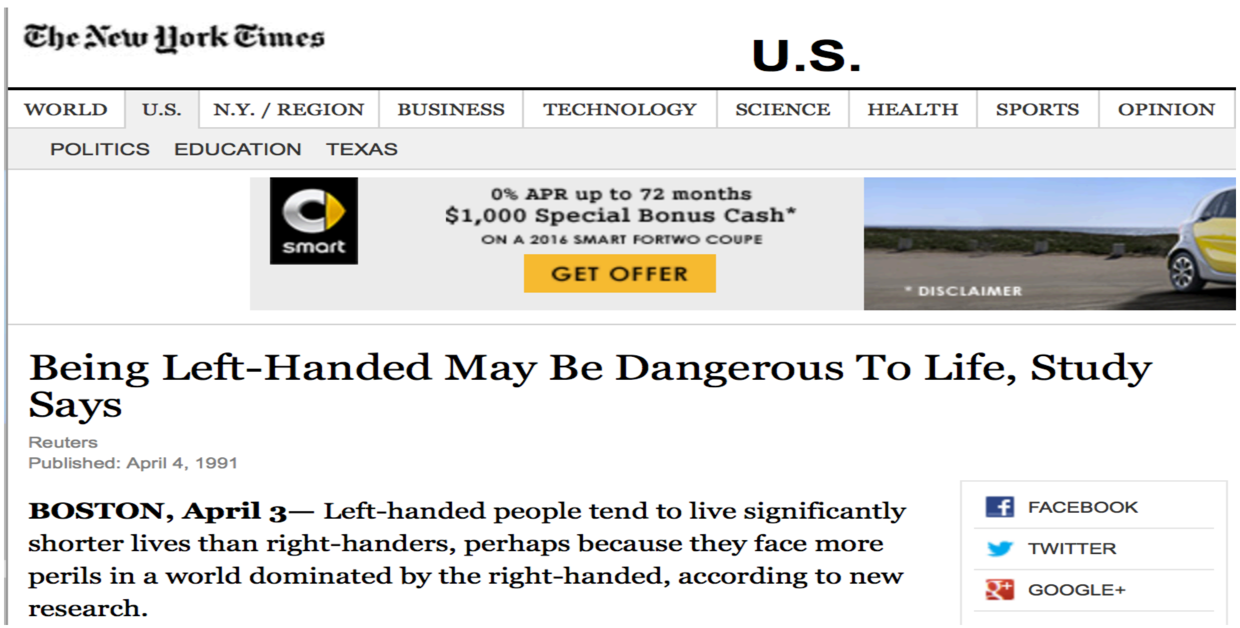


2. More about confounding factors and lefties example.

- A confounding factor must be plausibly linked to both the explanatory and response variables. So for instance saying "perhaps a higher proportion of the smokers are men" would not be a very convincing confounding factor, unless you have some reason to think gender is strongly linked to liver cancer.
- Another example: left-handedness and age at death. Psychologists Diane Halpern and Stanley Coren looked at 1,000 death records of those who died in Southern California in the late 1980s and early 1990s and contacted relatives to see if the deceased were righthanded or lefthanded. They found that the average ages at death of the lefthanded was 66, and for the righthanded it was 75. Their results were published in prestigious scientific journals, Nature and the New England Journal of Medicine.

More about confounding factors.

All sorts of causal conclusions were made about how this shows that the stress of being lefthanded in our righthanded world leads to premature death.



The New York Times

U.S.

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICS EDUCATION TEXAS

Being Left-Handed May Be Dangerous To Life, Study Says

Reuters
Published: April 4, 1991

BOSTON, April 3— Left-handed people tend to live significantly shorter lives than right-handers, perhaps because they face more perils in a world dominated by the right-handed, according to new research.

FACEBOOK
TWITTER
GOOGLE+

More about confounding factors.

- Is this an observational study or an experiment?

More about confounding factors.

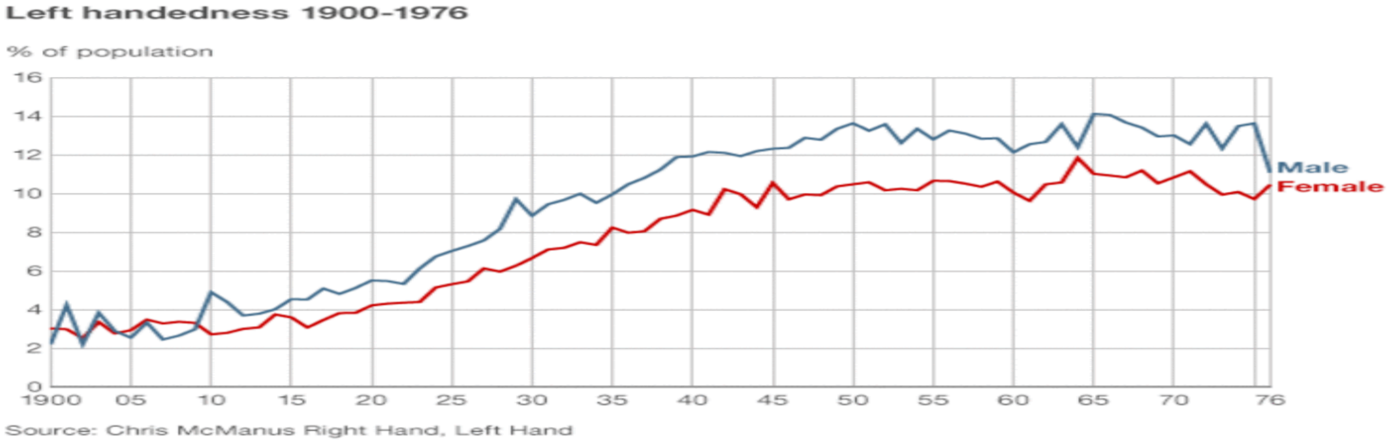
- Is this an observational study or an experiment?

It is an observational study.

- Are there plausible confounding factors you can think of?

More about confounding factors.

- A confounding factor is the age of the two populations in general. Lefties in the 1980s were on average younger than righties. Many old lefties were converted to righties at infancy, in the early 20th century, but this practice has subsided. Thus in the 1980s and 1990s, there were relatively few old lefties but many young lefties in the overall population. This alone explains the discrepancy.



Unit 2: Comparing Two Groups

- In Unit 1, we learned the basic process of statistical inference using tests and confidence intervals. We did all this by focusing on a single proportion.
- In Unit 2, we will take these ideas and extend them to comparing two groups. We will compare two proportions, two independent means, and paired data.

Chapter 5: Comparing Two Proportions

5.1: Comparing Two Groups: Categorical Response
5.2: Comparing Two Proportions: Simulation-Based Approach
5.3: Comparing Two Proportions: Theory-Based Approach

3. Comparing two proportions using numerical and visual summaries, and the good or bad year example.

Section 5.1

Example 5.1: Positive and Negative Perceptions

- Consider these two questions:
  - Are you having a good year?
  - Are you having a bad year?

- Do people answer each question in such a way that would indicate the same answer? (e.g. Yes for the first one and No for the second.)

Positive and Negative Perceptions

- Researchers questioned 30 students (randomly giving them one of the two questions).
- They then recorded if a positive or negative response was given.
- They wanted to see if the wording of the question influenced the answers.

Positive and negative perceptions

- Observational units
  - The 30 students
- Variables
  - Question wording (good year or bad year)
  - Perception of their year (positive or negative)
- Which is the explanatory variable and which is the response variable?
- Is this an observational study or experiment?

Raw Data in a Spreadsheet

Individual	Type of Question	Response
1	Good Year	Positive
2	Good Year	Negative
3	Bad Year	Positive
4	Good Year	Positive
5	Good Year	Negative
6	Bad Year	Positive
7	Good Year	Positive
8	Good Year	Positive
9	Good Year	Positive
10	Bad Year	Negative
11	Good Year	Negative
12	Bad Year	Negative
13	Good Year	Positive
14	Bad Year	Negative
15	Good Year	Positive

Individual	Type of Question	Response
16	Good Year	Positive
17	Bad Year	Positive
18	Good Year	Positive
19	Good Year	Positive
20	Good Year	Positive
21	Bad Year	Negative
22	Good Year	Positive
23	Bad Year	Negative
24	Good Year	Positive
25	Bad Year	Negative
26	Good Year	Positive
27	Bad Year	Negative
28	Good Year	Positive
29	Bad Year	Positive
30	Bad Year	Negative

Two-Way Tables

- A **two-way table** organizes data
  - Summarizes *two* categorical variables
  - Also called contingency table
- Are students more likely to give a positive response if they were given the good year question?

	Good Year	Bad Year	Total
Positive response	15	4	19
Negative response	3	8	11
Total	18	12	30

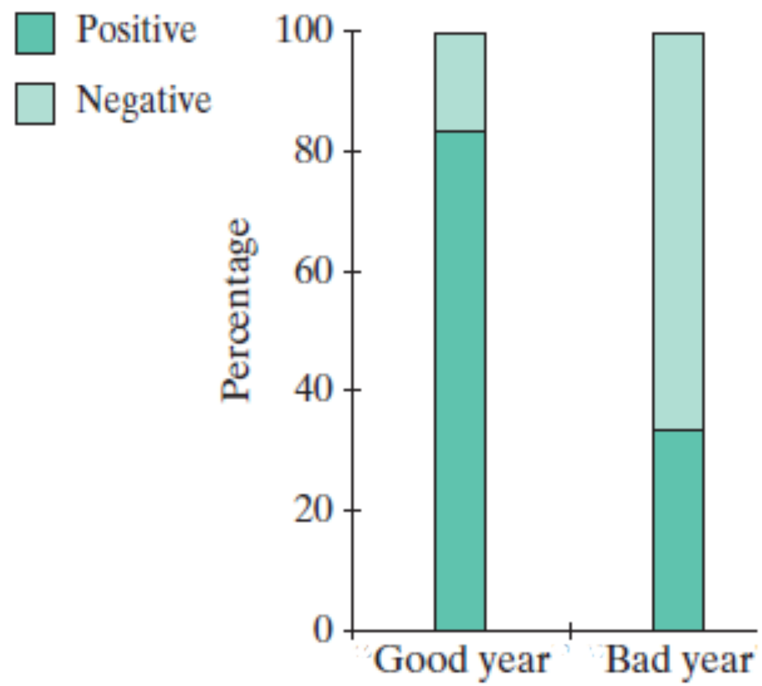
Two-Way Tables

- Conditional proportions will help us better determine if there is an association between the question asked and the type of response.
- We can see that the subjects with the positive question were ***more likely*** to respond positively.

	Good Year	Bad Year	Total
Positive response	15/18 ≈ 0.83	4/12 ≈ 0.33	19
Negative response	3	8	11
Total	18	12	30

Segmented Bar Graphs

- We can also use segmented bar graphs to see this association between the "good year" question and a positive response.



Statistic

- The statistic we will mainly use to summarize this table is the difference in proportions of positive responses is 0.83 – 0.33 = 0.50.

	Good Year	Bad Year	Total
Positive response	15 (83%)	4 (33%)	19
Negative response	3	8	11
Total	18	12	30

Another Statistic

- Another statistic that is often used, called **relative risk**, is the ratio of the proportions: 0.83/ 0.33 = 2.5.
- We can say that those who were given the good year question were 2.5 times as likely to give a positive response.

	Good Year	Bad Year	Total
Positive response	15 (83%)	4 (33%)	19
Negative response	3	8	11
Total	18	12	30

4. Comparing 2 proportions with CIs and testing using simulation, dolphin example.

Section 5.2

Swimming with Dolphins

Example 5.2

Swimming with Dolphins

Is swimming with dolphins therapeutic for patients suffering from clinical depression?

- Researchers Antonioli and Reveley (2005), in British Medical Journal, recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression
- Discontinued antidepressants and psychotherapy 4 weeks prior to and throughout the experiment
- 30 subjects went to an island near Honduras where they were randomly assigned to two treatment groups

Swimming with Dolphins

- Both groups engaged in one hour of swimming and snorkeling each day
- One group swam in the presence of dolphins and the other group did not
- Participants in both groups had identical conditions except for the dolphins
- After two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study
- The response variable is whether or not the subject achieved substantial reduction in depression

Swimming with Dolphins

Null hypothesis: Dolphins do not help.

- Swimming with dolphins is not associated with substantial improvement in depression

Alternative hypothesis: Dolphins help.

- Swimming with dolphins **increases** the probability of substantial improvement in depression symptoms

Swimming with Dolphins

- The parameter is the (long-run) difference between the probability of improving when receiving dolphin therapy and the prob. of improving with the control ($\pi_{\text{dolphins}} - \pi_{\text{control}}$)
- So we can write our hypotheses as:

H_0: $\pi_{\text{dolphins}} - \pi_{\text{control}} = 0$.

H_a: $\pi_{\text{dolphins}} - \pi_{\text{control}} > 0$.

or

H_0: $\pi_{\text{dolphins}} = \pi_{\text{control}}$

H_a: $\pi_{\text{dolphins}} > \pi_{\text{control}}$

(Note: we are not saying our parameters equal any certain number.)

Swimming with Dolphins

Results:

	Dolphin group	Control group	Total
Improved	10 (66.7%)	3 (20%)	13
Did Not Improve	5	12	17
Total	15	15	30

The difference in proportions of improvers is:

$\hat{p}_d - \hat{p}_c$ **= 0.667 – 0.20 = 0.467.**

Swimming with Dolphins

- There are two possible explanations for an observed difference of 0.467.
  - A tendency to be more likely to improve with dolphins (alternative hypothesis)
  - The 13 subjects were going to show improvement with or without dolphins and random chance assigned more improvers to the dolphins (null hypothesis)

Swimming with Dolphins

- If the null hypothesis is true (no association between dolphin therapy and improvement) we would have 13 improvers and 17 non-improvers regardless of the group to which they were assigned.
- Hence the assignment doesn’t matter and we can just randomly assign the subjects’ results to the two groups to see what would happen under a true null hypothesis.

Swimming with Dolphins

- We can simulate this with cards
  - 13 cards to represent the improvers
  - 17 cards represent the non-improvers
- Shuffle the cards
  - put 15 in one pile (dolphin therapy)
  - put 15 in another (control group)

Swimming with Dolphins

- Compute the proportion of improvers in the Dolphin Therapy group
- Compute the proportion of improvers in the Control group
- The difference in these two proportions is what could just as well have happened under the assumption there is no association between swimming with dolphins and substantial improvement in depression.

Dolphin Therapy

Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver

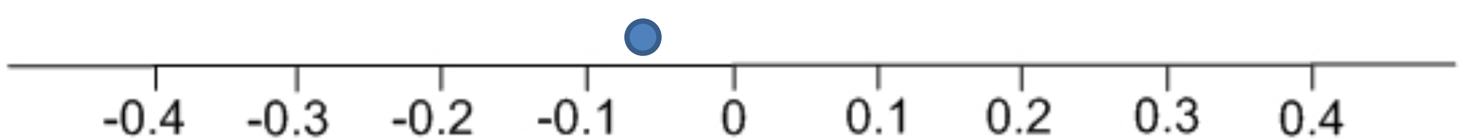
40.0% Improvers

Control

Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

46.7% Improvers

0.400 – 0.467 = -0.067



Difference in Simulated Proportions

Dolphin Therapy

Non-improver	Non-improver	Non-improver
Non-improver	Improver	Improver
Improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

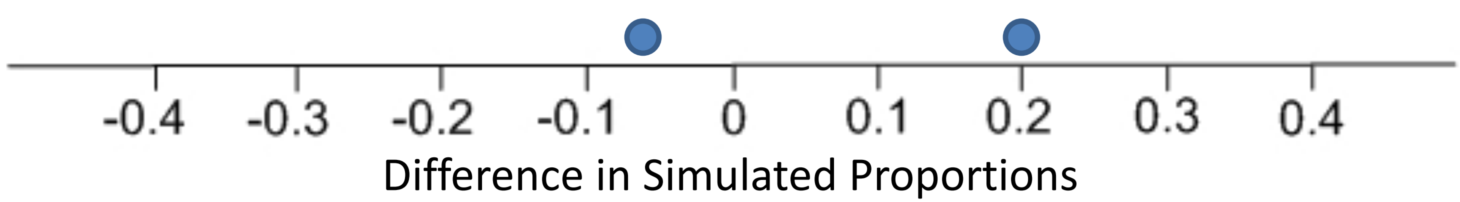
53.3% / 40.0% Improvers

Control

Non-improver	Improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Improver	Non-improver	Improver
Improver	Improver	Non-improver

33.3% / 46.7% Improvers

0.533 – 0.333 = 0.200



Dolphin Therapy

Non-improver	Non-improver	Non-improver
Non-improver	Improver	Improver
Improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

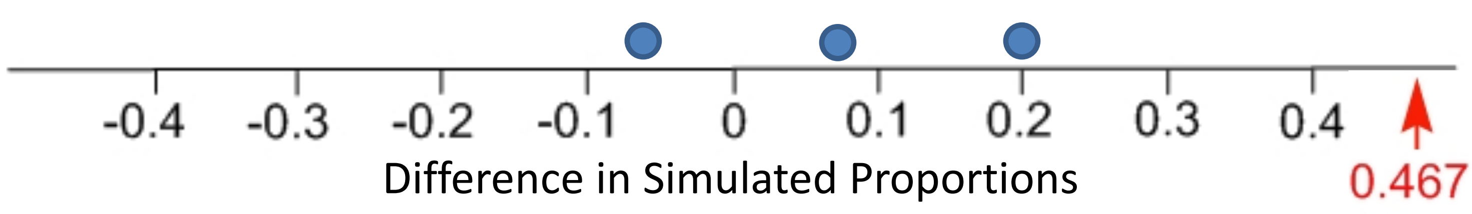
46.7% Improvers

Control

Non-improver	Improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Improver	Non-improver	Improver
Improver	Improver	Non-improver

40.0% Improvers

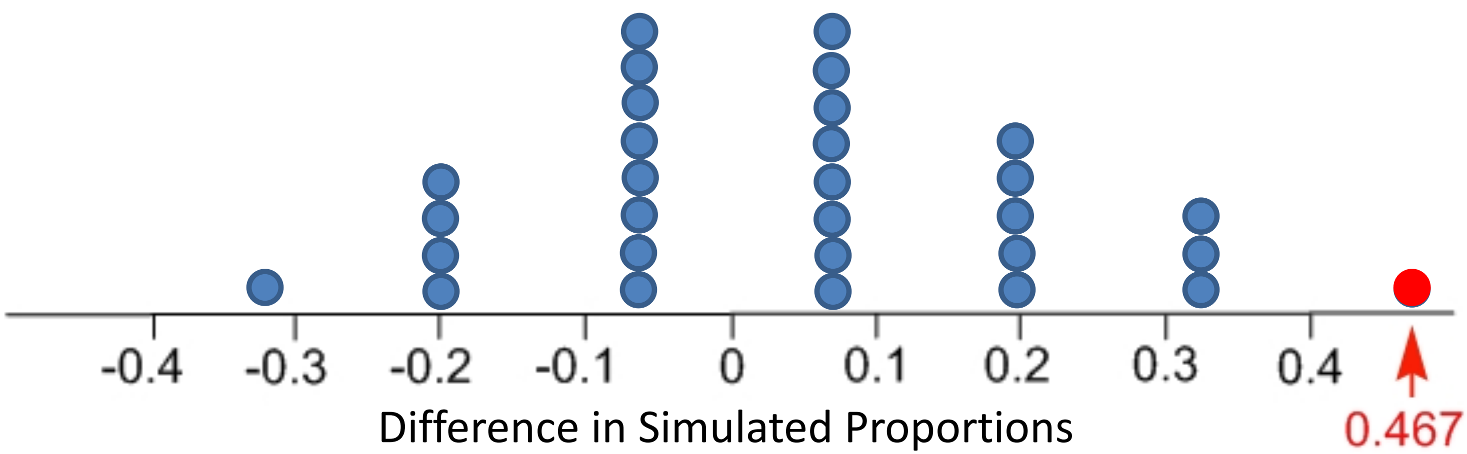
0.467 – 0.400 = 0.067



More Simulations

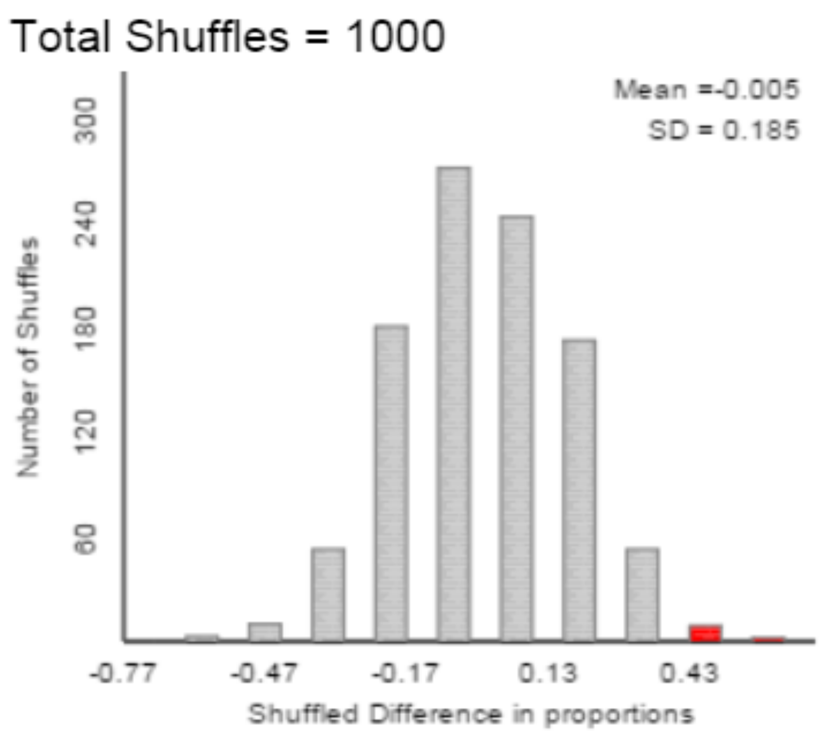
0.067 -0.067 0.067 0.200 0.067
-0.333 -0.200 0.333 0.333
-0.067 0.333 0.067 -0.200 -0.200
0.067 -0.067 0.467
0.200 0.067 0.067
-0.067 0.200
0.200 0.067 -0.067 0.067
-0.200

Only one simulated statistics out of 30 was as large or larger than our observed difference in proportions of 0.467, hence our p-value for this null distribution is 1/30 ≈ 0.03.



Swimming with Dolphins

- We did 1000 repetitions to develop a null distribution.



Swimming with Dolphins

- 13 out of 1000 results had a difference of 0.467 or higher (p-value = 0.013).
- 0.467 is $\frac{0.467-0}{0.185} \approx 2.52$ SD above zero.
- Using either the p-value or standardized statistic, we have strong evidence against the null and can conclude that the improvement due to swimming with dolphins was statistically significant.

Swimming with Dolphins

- A 95% confidence interval for the difference in the probability using the standard deviation from the null distribution is 0.467 ± 2(0.185) = 0.467 ± 0.370 or (0.097 to 0.837)
- We are 95% confident that when allowed to swim with dolphins, the probability of improving is between 0.097 and 0.837 higher than when no dolphins are present.
- How does this interval back up our conclusion from the test of significance?
- See if 0 is in the interval.

Swimming with Dolphins

- Can we say that the presence of dolphins *caused* this improvement?
  - Since this was a randomized experiment, and assuming everything was identical between the groups, we have strong evidence that dolphins were the cause
- Can we generalize to a larger population?
  - Maybe mild to moderately depressed 18-65 year old patients willing to volunteer for this study
  - We have no evidence that random selection was used to find the 30 subjects. "Outpatients, recruited through announcements on the internet, radio, newspapers, and hospitals."

5. Comparing Two Proportions: Theory-Based Approach, and smoking and gender example.

Section 5.3

Introduction

- Just as with a single proportion, we can often predict results of a simulation using a theory-based approach.
- The theory-based approach also gives a simpler way to generate confidence intervals.
- The main new mathematical fact to use is the SE for the difference between two proportions is

$\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}\right)}$ for testing

or $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$ for CIs.

A clarification on the new formula

- The margin of error for the difference in proportions is $Multiplier \times \text{SE}$, where $\text{SE} = \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$.

  In testing, the null hypothesis is no difference between the two groups, so we use the SE

$$\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}\right)}$$

  where $\hat{p}$ is the proportion in both groups combined. But in CIs, we use the formula $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$ because we are not assuming $\hat{p}_1 = \hat{p}_2$ with CIs.

Parents’ Smoking Status and their Babies’ Gender

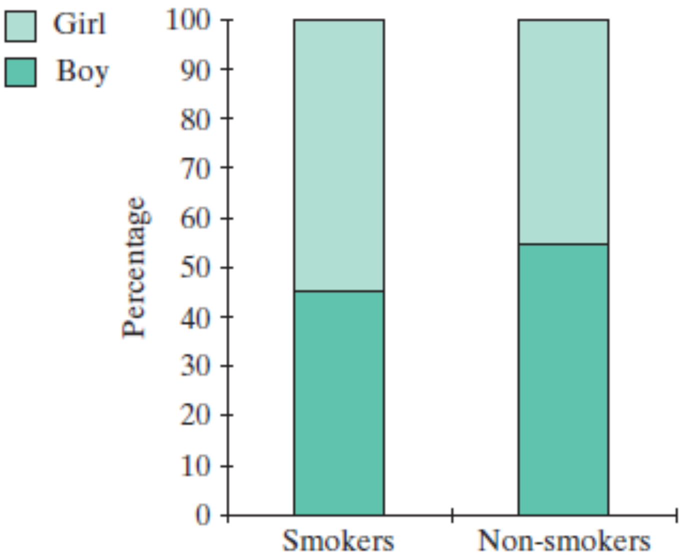
Example 5.3

Smoking and Gender

- How does parents’ behavior affect the gender of their children?
- Fukuda et al. (2002) found the following in Japan.
  - Out of 565 births where both parents smoked more than a pack a day, 255 were boys. This is 45.1% boys.
  - Out of 3602 births where both parents did not smoke, 1975 were boys. This 54.8% boys.
  - In total, out of 4170 births, 2230 were boys, which is 53.5%.
- Other studies have shown a reduced male to female birth ratio where high concentrations of other environmental chemicals are present (e.g. industrial pollution, pesticides)

Smoking and Gender

- A segmented bar graph and 2-way table
- Let's compare the proportions to see if the difference is statistically significantly.



	Both Smoked	Neither Smoked
Boy	255 (45.1%)	1,975 (54.8%)
Girl	310	1,627
Total	565	3,602

Smoking and Gender

Null Hypothesis:

- There **is no association** between smoking status of parents and sex of child.
- The probability of having a boy **is the same** for parents who smoke and don’t smoke.
- $\pi_{\text{smoking}} - \pi_{\text{nonsmoking}} = 0$

Smoking and Gender

Alternative Hypothesis:

- There **is an association** between smoking status of parents and sex of child.
- The probability of having a boy **is not the same** for parents who smoke and don’t smoke
- $\pi_{\text{smoking}} - \pi_{\text{nonsmoking}} \neq 0$

Smoking and Gender

- What are the observational units in the study?
- What are the variables in this study?
- Which variable should be considered the explanatory variable and which the response variable?
- What is the parameter of interest?
- Can you draw cause-and-effect conclusions for this study?

Smoking and Gender

Using the 3S Strategy to asses the strength

1. Statistic:

- The proportion of boys born to nonsmokers minus the proportion of boys born to smokers is 0.548 – 0.451 = 0.097.

Smoking and Gender

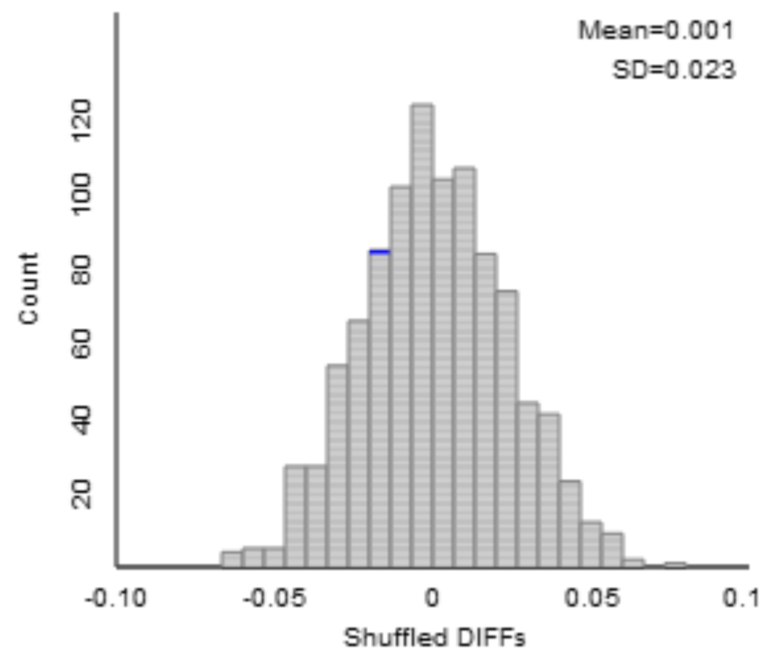
2. Simulate:

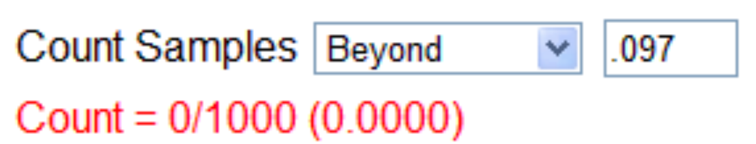
- Use the **Two Proportions** applet to simulate
- Many repetitions of shuffling the 2230 boys and 1937 girls to the 565 smoking and 3602 nonsmoking parents
- Calculate the difference in proportions of boys between the groups for each repetition.
- Shuffling simulates the null hypothesis of no association

Smoking and Gender

3. Strength of evidence:

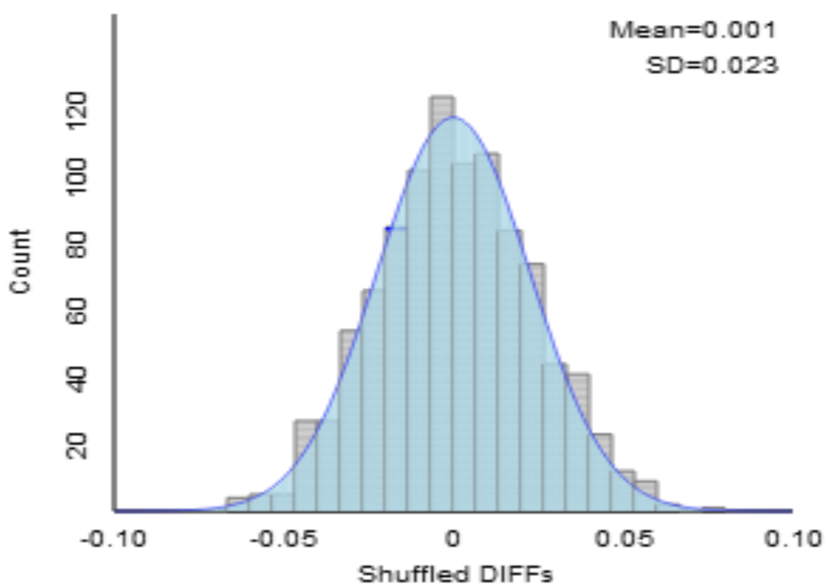
- Nothing as extreme as our observed statistic (≥ 0.097 or ≤ -0.097) occurred in 5000 repetitions,
- How many SDs is 0.097 above the mean?
  $Z = 0.097/0.023 = 4.22$
  using simulations. What about using the theory-based approach?





Smoking and Gender

- Notice the null distribution is centered at zero and is bell-shaped.
- This can be approximated by the normal distribution.



Formulas

- The theory-based approach yields z = 4.30.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Here $z = \frac{.548 - .451}{\sqrt{.535\,(1-.535)\left(\frac{1}{3602}+\frac{1}{565}\right)}}$ = 4.30.
- p-value is 2*(1-pnorm(4.30)) = 0.00171%.
-

Smoking and Gender

- Fukuda et al. (2002) found the following in Japan.
  - Out of 3602 births where both parents did not smoke, 1975 were boys. This 54.8% boys.
  - Out of 565 births where both parents smoked more than a pack a day, 255 were boys. This is 45.1% boys.
  - In total, out of 4170 births, 2230 were boys, which is 53.5% boys.

Formulas

- How do we find the margin of error for the difference in proportions?

$$Multiplier \times \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$$

- Here n is large and obs. are iid, so we use the multipliers of

  1.645 for 90% confidence

  1.96 for 95% confidence

  2.576 for 99% confidence

- We can write the confidence interval in the form:
  – statistic ± margin of error.

Smoking and Gender

- Our statistic is the observed sample difference in proportions, 0.097.
- Plugging in $1.96 \times \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} = 0.044$, we get 0.097 ± 0.044 as our 95% CI.
- We could also write this interval as (0.053, 0.141).
- We are 95% confident that the probability of a boy baby where neither family smokes minus the probability of a boy baby where both parents smoke is between 0.053 and 0.141.