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

1. Comparing 2 props. with theory-based testing, smoking and gender example.
2. Causation and prediction.
3. Review of relationship between CIs and tests.
4. Review list.
5. Example problems.

Read ch6. The midterm is on chapters 1-6, but only on material we have discussed in class, including items 1 and 2 above.
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. There is also no office hour Tue Nov7.

1. 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

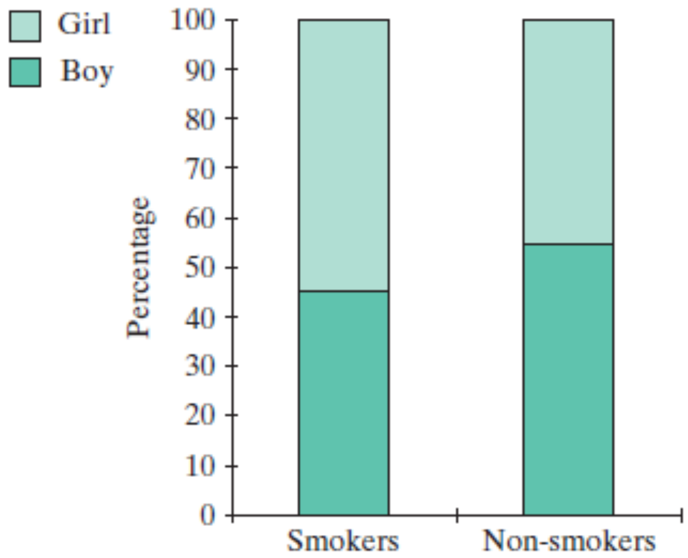
Example 5.3

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 significant.



	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.
- $\pi_{smoking} - \pi_{nonsmoking} = 0$

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

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

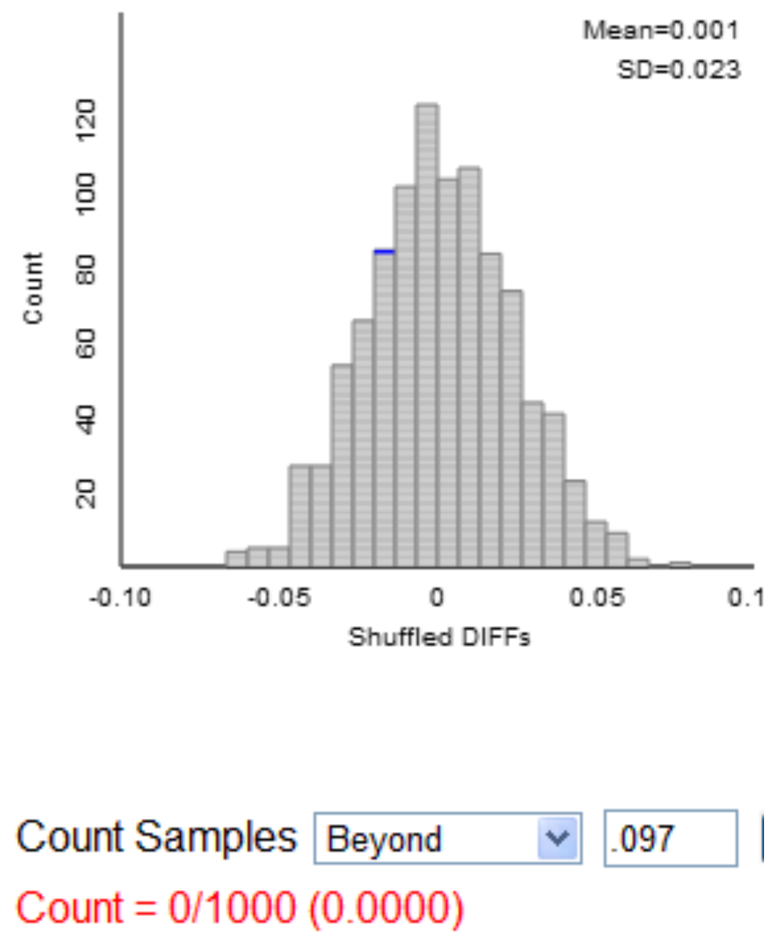
2. Simulate:

- 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

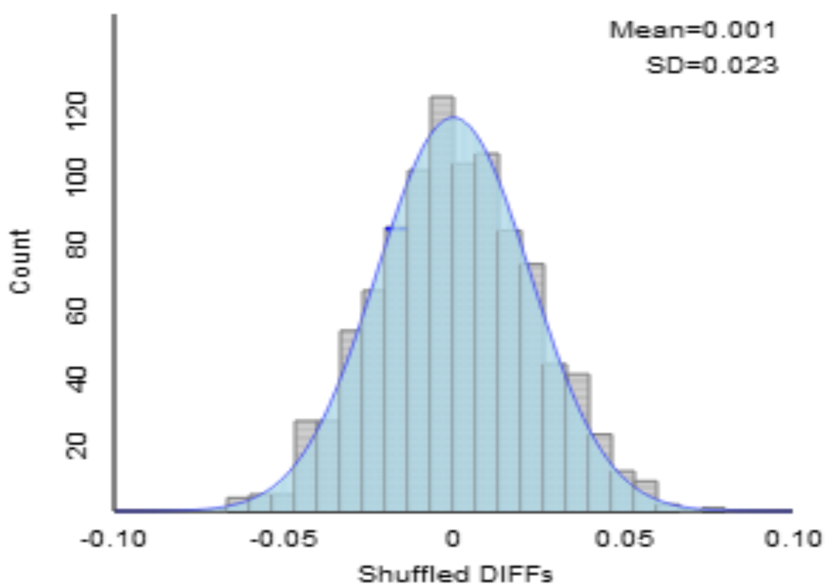
3. Strength of evidence:

- 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%.
-

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 very 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.

2. Causation and prediction.

Note that for prediction, you sometimes do not care about confounding factors.

- Forecasting wildfire activity using temperature. Warmer weather may directly cause wildfires via increased ease of ignition, or due to confounding with people chooseing to go camping in warmer weather. It does not really matter for the purpose of merely *predicting* how many wildfires will occur in the coming month.
- The same goes for predicting lifespan, or liver disease rates, etc., using smoking as a predictor variable.

3. Review of CIs and tests.

Suppose we are comparing death rates in a treatment group and a control group. We observe a difference of 10.2%, do a 2-sided test, and find a p-value of 8%.

Does this mean the 95%-CI for the difference in death rates between the two groups would contain zero?

Review of CIs and tests.

Suppose we are comparing death rates in a treatment group and a control group. We observe a difference of 10.2%, do a 2-sided test, and find a p-value of 8%.

Does this mean the 95%-CI for the difference in death rates between the two groups would contain zero?

yes. We cannot rule out 0 as a plausible null value for the difference between the two groups.

CIs and tests.

Suppose we are comparing blood pressures in a treatment group and a control group. We observe a difference of 10.2 mm, do a 2-sided test, and find a p-value of 3%.

Would the 95%-CI for the difference in blood pressures between the two groups contain zero?

CIs and tests.

Suppose we are comparing blood pressures in a treatment group and a control group. We observe a difference of 10.2 mm, do a 2-sided test, and find a p-value of 3%.

Would the 95%-CI for the difference in blood pressures between the two groups contain zero or not?

No. It would not contain zero.

For what confidence level would the CI just barely contain 0?

97%.

CIs and tests.

The p-value is 3%. A 97%-CI would just contain zero.

```
          Ho
     (    95%       )  (     95%-CI        )
---------------------------------------------------------------------
           |                      |
           0                   10.2mm


               Ho
(              97%                 )
               (                   97%-CI                     )
---------------------------------------------------------------------
               |                  |
               0                10.2mm
```

4. Review list.

1. Meaning of SD.
2. Parameters and statistics.
3. Z statistic for proportions.
4. Simulation and meaning of pvalues.
5. SE for proportions.
6. What influences p-values.
7. CLT and validity conditions for tests.
8. 1-sided and 2-sided tests.
9. Reject the null vs. accept the alternative.
10. Sampling and bias.
11. Significance level.
12. Type I, type II errors, and power.
13. CIs for a proportion.
14. CIs for a mean.
15. Margin of error.
16. Practical significance.
17. Confounding.
18. Observational studies and experiments.
19. Random sampling and random assignment.
20. Two proportion CIs and testing.
21. Placebo effect, adherer bias, and nonresponse bias.
22. Prediction and causation.
23. Assumptions for Z and t tests and CIs.

5. Example problems.

Some good hw problems from the book are 1.2.18, 1.2.19, 1.2.20, 1.3.17, 1.5.18, 2.1.38, 2.2.6, 2.2.24, 2.3.3, 2.3.25, 3.2.11, 3.2.12, 3.3.8, 3.3.19, 3.3.22, 3.5.23, 4.1.14, 4.1.18, 5.2.2, 5.2.10, 5.2.24, 5.3.11, 5.3.21, 5.3.24, 6.2.23, 6.3.1, 6.3.12, 6.3.22, 6.3.23, 7.2.20, 7.2.24, 7.3.7, 7.3.24.

5. Example problems.

NCIS was a top-rated tv show in 2014. It is currently 5th in 2017.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

5. Example problems.

NCIS was a top-rated tv show in 2014.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

No. Age is a confounding factor. The median age of a viewer is 61 years old.

5. Example problems.

What is a plausible explanation for why more 2008 New Hampshire Democratic primary voters voted for Hillary Clinton compared to what the surveys predicted?

a. The surveys sampled fewer Obama voters by chance.

b. Obama voters did not respond to the surveys in as high percentages as Clinton voters did.

c. A substantial number of Clinton voters claimed in the surveys they supported Obama.

d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.

e. None of the above.

5. Example problems.

What is a plausible explanation for why more 2008 New Hampshire Democratic primary voters voted for Hillary Clinton compared to what the surveys predicted?

a. The surveys sampled fewer Obama voters by chance.

b. Obama voters did not respond to the surveys in as high percentages as Clinton voters did.

c. A substantial number of Clinton voters claimed in the surveys they supported Obama.

d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.

e. None of the above.

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

$2.0 \pm 1.96 \sqrt{1.5^2/100 + 2.2^2/80} = 2.0 \pm 0.564.$

5. Example problems.

Suppose you sample 100 UCLA students and 80 2^{nd} graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2^{nd} grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in 2^{nd} grade statistically significant?

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in 2nd grade statistically significant?

Yes. The 95%-CI does not come close to containing 0.

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

5. Example problems.

Suppose you sample 100 UCLA students and 80 2^{nd} graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2^{nd} grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

Observational study.

5. Example problems.

Suppose you sample 100 UCLA students and 80 2^{nd} graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2^{nd} grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

No. Age is a confounding factor. Young kids eat more candy.

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter   random variable   t-test

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by?

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by?

$SE = \sqrt{1.5^2/100 + 2.2^2/80} = .288$ mmol/L

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

5. Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

1.5 mmoL/L.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

1. Which of the following is true?

a. A **parameter** of interest is the percentage in the sample who have diabetes, and it is used to estimate the percentage of Type A blooded Americans who have diabetes, which is a **random variable.**

b. A **parameter** of interest is the percentage of Type A blooded Americans who have diabetes, and the **random variable** used to estimate it is the sample size of 144.

c. A **parameter** of interest is the central limit theorem, and the **random variable** used to estimate it is the conditional proportion of the t-test confidence interval.

d. A **parameter** of interest is the percentage of Type A blooded Americans who have diabetes, and the **random variable** used to estimate it is the percentage in the sample who have diabetes.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

1. Which of the following is true?

a. A **parameter** of interest is the percentage in the sample who have diabetes, and it is used to estimate the percentage of Type A blooded Americans who have diabetes, which is a **random variable.**

b. A **parameter** of interest is the percentage of Type A blooded Americans who have diabetes, and the **random variable** used to estimate it is the sample size of 144.

c. A **parameter** of interest is the central limit theorem, and the **random variable** used to estimate it is the conditional proportion of the t-test confidence interval.

d. A **parameter** of interest is the percentage of Type A blooded Americans who have diabetes, and the **random variable** used to estimate it is the percentage in the sample who have diabetes.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

2. Based only on the sample of 144 Type A blooded Americans, find a 95% CI for the percentage of Type A blooded Americans with diabetes.

a. 9.03% +/- 4.7%.

b. 9.03% +/- 5.8%.

c. 9.03% +/- 6.1%.

d. 9.03% +/- 7.8%.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

2. Based only on the sample of 144 Type A blooded Americans, find a 95% CI for the percentage of Type A blooded Americans with diabetes.

a. 9.03% +/- 4.7%.

b. 9.03% +/- 5.8%.

c. 9.03% +/- 6.1%.

d. 9.03% +/- 7.8%.

$1.96 \sqrt{(.093 \times .907 / 144)} = 0.04744.$

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

3. In order to test whether the diabetes rate depends on blood type, what would the **null hypothesis** be?

a. The percentage of Americans with Type A blood who have diabetes is **not** 9.3%.

b. The percentage of Americans with Type A blood is 9.3%.

c. 9.3% of Americans with Type A blood have diabetes.

d. 9.3% of Americans with diabetes have Type A blood.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

3. In order to test whether the diabetes rate depends on blood type, what would the **null hypothesis** be?

a. The percentage of Americans with Type A blood who have diabetes is **not** 9.3%.

b. The percentage of Americans with Type A blood is 9.3%.

c. 9.3% of Americans with Type A blood have diabetes.

d. 9.3% of Americans with diabetes have Type A blood.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

4. Under this null hypothesis, what is the size of the **standardized** Z statistic, summarizing the difference between the sample percentage with diabetes and the population percentage with diabetes?

a. 0.0321.

b. 0.0505.

c. 0.112.

d. 0.676.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

4. Under this null hypothesis, what is the size of the **standardized** Z statistic, summarizing the difference between the sample percentage with diabetes and the population percentage with diabetes?

a. 0.0321.

b. 0.0505.

c. 0.112.

d. 0.676.

Z = (13/144 – 9.3%) ÷ √(.093 x .907/144) = -.112476.

So the size of Z is 0.112476.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

5. Which of the following can be concluded?

a. The difference between the observed diabetes percentage in the sample of Type A blooded Americans and the diabetes percentage among all Americans is not statistically significant.

b. The diabetes percentage for the population of Type A blooded Americans is 9.30%.

c. The diabetes percentage for the population of Type A blooded Americans is 9.03%.

d. Diabetes rates appear to vary with blood type and the difference is statistically significant.

Consider the following for the next 5 problems. About 9.30% of Americans have diabetes, according to the CDC. A scientist wants to see if the frequency of diabetes among Americans with Type A blood is different from the overall percentage of 9.30%. She takes a simple random sample of 144 patients with Type A blood, and finds that 13 of them have diabetes.

5. Which of the following can be concluded?

a. The difference between the observed diabetes percentage in the sample of Type A blooded Americans and the diabetes percentage among all Americans is not statistically significant.

b. The diabetes percentage for the population of Type A blooded Americans is 9.30%.

c. The diabetes percentage for the population of Type A blooded Americans is 9.03%.

d. Diabetes rates appear to vary with blood type and the difference is statistically significant.

In the Physician's Health Study I on aspirin's effect on reducing the risk of heart attacks, which of the following was not a reason for randomly assigning people to treatment or control in this experiment?

a. To ensure that the treatment and control groups are similar with respect to known potential confounders such as diet and exercise.

b. To ensure that the treatment and control groups are similar with respect to unknown confounding factors.

c. To ensure that the sample is more representative of the overall population.

Suppose a random sample of 100 college students is asked if they regularly eat breakfast. A 95% confidence interval for the proportion of all students that regularly eat breakfast is found to be 0.70 to 0.90. If a 99% confidence interval was calculated instead, how would it differ from the 95% confidence interval?

a. The 99% confidence interval would have the same width as the 95% confidence interval.

b. The 99% confidence interval would be wider.

c. The 99% confidence interval would be narrower.

d. More information is needed in order to determine if the 99% CI would be wider or narrower than the 95% CI.

Which of the following does not influence the p-value when testing the difference between two means?

a. The sample size.

b. The effect size.

c. The standard deviations of the two groups.

d. The choice of which group is considered the treatment and which is the control.

Suppose you take a simple random sample of 15 residents of a certain city and find their mean income is $42,000 and the SD is $10,000. You want to find a 95% CI for the mean income of residents in this city, and are considering using $42,000 +/- 1.96 x $10,000/√15. Which of the following is true?

a. You should use this formula because of the central limit theorem, since the sample size is large.

b. You should use this formula because incomes are known to be normally distributed.

c. You should not use this formula because the study is an observational study rather than an experiment.

d. You should not use this formula because the sample size is small, the population SD is unknown and incomes are probably not normally distributed.