Recap

Announcements

Lab 9 due now (extra credit)

Projects are due this Friday 5pm on Sakai:
- Write up (5 page, double spaced, including figures)
- Code (as a .R file)
- Data (as a .csv file, or if you have recreated your data in R this can be included in the code)

Midterm 2 next Tuesday (Nov 8):
- Focuses on chapters 4 - 6
- Thursday’s class will be mostly review as well
- Optional review session: Sat, Nov 5, 1pm - 2:30pm, place TBA

HW 7 due next Tuesday along with next Monday’s lab

Office hours:
- Today after class
- Thursday 2:30 - 4:30pm
- Friday 9:00 - 11:00am
- Monday 12:00 - 2:30pm and 4pm - 5pm

Review question

You may skip Section 6.2.4 in your textbook, it’s about pooling the standard deviation when testing for a difference between the means. We would only do this if we know the two samples come from distributions with equal standard deviations. This is so difficult to test for that we barely ever use it.

Given below are some sample statistics on maximum cranial breadth of 30 randomly sampled male Egyptian skulls from 4000 BCE and 150 CE. Which of the following is the correct calculation of a 90% confidence interval for the difference between the two means?

<table>
<thead>
<tr>
<th></th>
<th>4000 BCE</th>
<th>150 CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>131 mm</td>
<td>136 mm</td>
</tr>
<tr>
<td>SD</td>
<td>5 mm</td>
<td>5.35 mm</td>
</tr>
</tbody>
</table>

(a) \((131 - 136) \pm 1.65 \times \sqrt{\frac{5^2}{30} + \frac{5.35^2}{30}}\)

(b) \((131 - 136) \pm 1.65 \times \sqrt{\frac{5^2}{30} + \frac{5.35^2}{30}}\)

(c) \((131 - 136) \pm 1.70 \times \sqrt{\frac{5^2}{30} + \frac{5.35^2}{30}}\)

(d) \((131 - 136) \pm 1.70 \times \sqrt{\frac{5^2}{30} + \frac{5.35^2}{30}}\)

Thomson, Randall-MacIver. The ancient races of the Thebaid, 1905.
A variety of studies suggest that 10% of the world population is left-handed.

From Wikipedia.

**Clicker question**

Are you right-handed, left-handed, or ambidextrous?

(a) right  
(b) left  
(c) ambidextrous

**Hypotheses**

Assuming that this class is a representative sample of Duke students, which of the following are the correct set of hypotheses for testing if the proportion of Duke students who are left-handed is different than the proportion of left-handed people in the world.

(a) \(H_0 : p = 0.10\) 
\(H_A : p < 0.10\)

(b) \(H_0 : p = 0.10\) 
\(H_A : p \neq 0.10\)

(c) \(H_0 : p_{Duke} = p_{world}\) 
\(H_A : p_{Duke} \neq p_{world}\)

(d) \(H_0 : \hat{p}_{Duke} = \hat{p}_{world}\) 
\(H_A : \hat{p}_{Duke} \neq \hat{p}_{world}\)

**Assumptions and conditions**

- **Independence:** We are assuming the sample is representative of the Duke population, and \(n < 10\%\) of all Duke students.
- **Normality:** The number of expected successes is smaller than 10.

So what do we do?

Since the sample size isn’t large enough to use CLT based methods, we use a simulation method instead.

**Set up**

1. Generate a random sample of size \(n\) with probability of success \(p = 0.10\) using simulation.
2. Calculate the proportion of successes, \(\hat{p}_{sim}\) and record this value.
3. Repeat steps 1 and 2 many times and build a sampling distribution of \(\hat{p}_{sim}\).
4. Calculate the p-value as the proportion of simulations in which \(\hat{p}_{sim}\) is beyond \(\hat{p}_{observed}\) in the direction of the alternative hypothesis.
5. If the p-value is small, reject the null hypothesis.
There are many ways of generating a random sample of size \( n \) with probability of success \( p = 0.10 \).

- Roll a 10 sided die: if it lands on 0 record success, if it lands on any other number (1-9) record failure. Do this \( n \) times.
- In a bag put 10 chips, 1 red 9 blue. Draw a chip, if it’s red record success, if it’s blue record failure. Do this \( n \) times. Sample with replacement.
- Use computational resources to generate a random number between 0 and 9. If the number is 0 record success, if 1-9 record failure. Do this \( n \) times.

Clicker question

Is the last digit of your phone number 0?

(a) Yes
(b) No

\( \hat{p}_{\text{sim},1} = \)

Clicker question

Is the last digit of your Duke unique ID 0?

(a) Yes
(b) No

\( \hat{p}_{\text{sim},2} = \)

Clicker question

Is the last digit of your SSN 0?

(a) Yes
(b) No

\( \hat{p}_{\text{sim},3} = \)

Clicker question

Is the last digit of your Duke Net ID 0?

(a) Yes
(b) No

\( \hat{p}_{\text{sim},4} = \)

Since this process is very slow, we can do a few more samples using R to generate random numbers.

```r
phatSim <- c( , , , , rep(NA, 96))
n <-
for(i in 5:100){
  # generate n uniformly distributed random numbers between 0 and 9
  samp <- round(runif(n, min = 0, max = 9))
  # find the proportion of 0s in the sample
  pr <- sum(samp == 0) / n
  # record phatSim
  phatSim[i] <- pr
}
```
Calculating the p-value & conclusion

Duke from the Northeast

At the beginning of the semester you filled out a survey where you were asked about your home region. 48 students in this class responded to this question. The table below summarizes these responses.

<table>
<thead>
<tr>
<th>International</th>
<th>Midwest</th>
<th>Northeast</th>
<th>Southeast</th>
<th>Southwest</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>4</td>
<td>17</td>
<td>9</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

We would like to test if these data provide convincing evidence that more than 20% of students at Duke are from the Northeast.

- What is the parameter of interest?
- What is the point estimate?

Hypothesis test

\[ H_0 : p = 0.20 \]
\[ H_A : p > 0.20 \]

Assuming that this is a random sample and since 48 < 10% of all Duke students, whether or not one student in the sample is from the Northeast is independent of another. Sample size is small (fewer than 10 expected successes) so we should use a simulation method.

\[ 48 \times 0.20 = 9.6 \quad \text{and} \quad 48 \times 0.80 = 38.4 \]
Clicker question

Which of the following is correct?

\( n = 48, \text{ successes} = 17, \hat{p} = 0.354, p - \text{value} = 0.0089, \alpha = 0.025 \)

(a) If in fact 20% of Duke students are from the Northeast, the probability of getting a random sample of 48 Duke students where 17 or more are from the Northeast is 0.0089.

(b) We are 95% confident that 20% of Duke students are from the Northeast.

(c) The data do not provide convincing evidence to suggest that more than 20% of Duke students are from the Northeast.

(d) This hypothesis test is not valid since there are fewer than 10 expected successes.

For constructing a confidence interval for a proportion we **bootstrap** instead of **randomize**.

- This means we sample with replacement from the existing sample and create a bootstrapping distribution of the \( \hat{p}_{\text{sim}} \).
- Then we find the boundaries of the middle XX% of the distribution to construct a XX% confidence interval.

Bootstrapping

```r
boot1 <- sample(homeRegion, size = 48, replace = T)
success1 <- sum(boot1 == "Northeast") # 14
phatSim1 <- success1 / 48 # 0.291

boot2 <- sample(homeRegion, size = 48, replace = T)
success2 <- sum(boot2 == "Northeast") # 16
phatSim2 <- success2 / 48 # 0.333
```

```
simCI(homeRegion, cat = "Northeast", conf.level = 0.95)
# Simulated 95 % interval for the proportion = ( 0.2292 , 0.5 )
```

Conclusions

Clicker question

Which of the following is false?

(a) The hypothesis test and the confidence interval agree with each other.

(b) We are 95% confident that the proportion of Duke students from the Northeast is between 22.92% and 50%.

(c) If we were to simulate another bootstrap confidence interval, i.e. run the function `simCI` again, we would get the same exact values for the boundaries.

(d) A 99% confidence interval would be wider.
In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.

The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.

It was randomly determined which supervisors got “male” applications and which got “female” applications.

Of the 48 files reviewed, 35 were promoted.

The study is testing whether females are unfairly discriminated against.

The table below shows the gender distribution of the promoted files.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Promoted</th>
<th>Not Promoted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>21</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
</tr>
</tbody>
</table>

Parameter and point estimate

Do these data provide convincing evidence that females are unfairly discriminated against?

- **Parameter of interest**: Difference between the proportions of all males and females who are equally qualified who get promoted.
  
  \[ p_m - p_f \]

- **Point estimate**: Difference between the proportions of equally qualified males and females in the sample who get promoted.
  
  \[ \hat{p}_m - \hat{p}_f \]

Hypotheses

If the study is testing whether females are unfairly discriminated against, what are the appropriate hypotheses for this study?

- \( H_0 : p_m - p_f = 0 \)  
  \( H_A : p_m - p_f \neq 0 \)
- \( H_0 : \hat{p}_m - \hat{p}_f = 0 \)  
  \( H_A : \hat{p}_m - \hat{p}_f > 0 \)
- \( H_0 : p_m - p_f = 0 \)  
  \( H_A : \hat{p}_m - \hat{p}_f < 0 \)
- \( H_0 : \hat{p}_m - \hat{p}_f = 0 \)  
  \( H_A : \hat{p}_m - \hat{p}_f \neq 0 \)
**Assumptions and conditions**

- **Independence**: Since it was randomly determined which supervisors got “male” applications and which got “female” applications, we can assume that the decisions are independent.
- **Normality**: There are only 3 male files that did not get promoted.

So what do we do?

Since the sample size isn’t large enough to use CLT based methods, we use a simulation method instead.

**Simulation results - shuffling**

**Clicker question**

What is the difference you calculated?

(a) More than 0.3  
(b) Between 0.1 and 0.2  
(c) Between -0.1 and 0.1  
(d) Between -0.1 and -0.2  
(e) Less than -0.3

**Simulation results - R**

```r
# data
gender <- as.factor(c(rep("male", 24), rep("female",24)))
promoted <- as.factor(c(rep("promoted", "notPromoted"), c(21, 3)), rep(c("promoted", "notPromoted"), c(14,16))))

# randomization
reallocate(promoted,gender, order = c("male","female"), outcome = "promoted", alternative = "greater")
# Observed difference in sample proportions of "promoted", male-female: 0.292  
# p-value for a one sided test, greater: 0.0232
```

**Randomization Distribution**

- Frequency distribution of simulated differences in sample proportions, male-female.
Small sample inference for difference between two proportions

Bootstrap confidence interval

```r
resample(promoted, gender, order = c("male","female"), outcome = "promoted")
# Observed difference in sample proportions of "promoted",
# male-female: -0.292
# SE = 0.14
# 95 % Bootstrap interval = ( -0.583 , -0.042 )
```

How do we interpret this interval?