Moodle Questions Chapter 9

Homer White

Thursday, June 19, 2014

# Making a Confidence Interval: Spider Speeds

Remember the data frame SpiderSpeed from package abd?

require(abd)  
data(SpiderSpeed)  
View(SpiderSpeed)  
help(SpiderSpeed)

Recall that in this study, the researchers took 32 *Tidarren* spiders and recorded how fast each one could run, in centimeters per second. Then they amputated a [pedipalp](http://simple.wikipedia.org/wiki/Pedipalp) from each one of the spiders, and again recorded how fast each spider could run. This was a *repeated measures* study!

The researchers wanted to know if the spiders could run faster, on average, without the hindrance of a pedipalp.

Let's make a 95% confidence interval for mu\_d, the mean difference in running speed (after amputation minus before amputation).

Which of the following R functions will deliver the confidence interval you want?

Option A:

ttestGC(~ speed.after - speed.before, data=SpiderSpeed)

Feedback: Good job! Notice that you want a 95% confidence interval, you don't have to say so. R will give you a 95% confidence interval by default.

Option B:

ttestGC(~ speed.before - speed.after, data=SpiderSpeed)

Feedback: Close, but no cigar. In the set-up mu\_d refers to speed after minus speed before, whereas this code works with speed before minus speed after. All the numbers will be the "negative" of what you were expecting.

Option C:

ttestGC(~ speed.after, data=SpiderSpeed)

Feedback: This gives a confidence interval for the mean speed of all spiders AFTER amputation. It doesn't give an interval for the mean of differences mu\_d.

Option D:

ttestGC(speed.after ~ speed.before, data=SpiderSpeed)

Feedback: Running this code results in R throwing an error at you!

# Using an Interval to Answer a Research Question: Spider Speeds

Again we will work with the data frame SpiderSpeed from package abd:

require(abd)  
data(SpiderSpeed)  
View(SpiderSpeed)  
help(SpiderSpeed)

Recall that in this study, the researchers took 32 *Tidarren* spiders and recorded how fast each one could run, in centimeters per second. Then they amputated a [pedipalp](http://simple.wikipedia.org/wiki/Pedipalp) from each one of the spiders, and again recorded how fast each spider could run. This was a *repeated measures* study!

The researchers wanted to know if the spiders could run faster, on average, without the hindrance of a pedipalp.

In the output below, a 95%-confidence interval is given for mu\_d.

*Which of the following options is the best way to use the confidence interval to answer the Research Question?*

##   
##   
## Inferential Procedures for the Difference of Means mu-d:  
## speed.before minus speed.after   
##   
##   
## Descriptive Results:  
##   
## Difference mean.difference sd.difference n  
## speed.after - speed.before 1.186 1.074 16  
##   
##   
## Inferential Results:  
##   
## Estimate of mu-d: 1.186   
## SE(d.bar): 0.2684   
##   
## 95% Confidence Interval for mu-d:  
##   
## lower.bound upper.bound   
## 0.613449 1.757801

Option A:

If amputation made no difference in running speed, on average, then the mean of differences would be 0. However, our 95%-confidence interval for mu\_d contains only *postive* numbers (from .61 to 1.76 cm/sec). Hence the we are pretty sure -- based on the data -- that the mean of differences is a positive number, meaning that on average *Tidarren* spiders run faster when they are missing a pedipalp.

Option B:

The interval tells us that mu\_d is positive.

Feedback: This is vague, leaving out crucial steps in the argument. It also gives the misleading impression that we are ABSOLUTELY certain that mu\_d is positive, whereas in fact there is always a possibility (however slight) that you can get a really bizarre, unlikely sample that leads you to the wrong conclusion.

Option C:

The interval tells us that about 95% of all*Tidarren* spiders will run somewhere between .61 and 1.76 cm/sec faster after amputation, so on mean difference for the whole species must be positive, too.

Feedback: This conclusion is based on a mis-interpretation of the confidence interval. Confidence intervals don't try to capture a certain percentage of the population values; instead they try to capture just one number: the population parameter for which they were made.

# Napkins I: The Code

In the tigerstats package we find the data frame napkins:

data(napkins)  
View(napkins)  
help(napkins)

this was an observational study performed at Georgetown College. A small team of student researchers visited the Cafe several times during the lunch hour and recorded the sex of each student in the study as well as how many napkins each student used during the meal.

The Research Question was:

*Who uses more napkins, on average, during lunch: a GC male or a GC female?*

The researchers decided to study their question by making a confidence interval. First they defined some parameters:

Let

mu1 = the mean number of napkins per student, for all male students at Georgetown.

mu2 = the mean number of napkins per student, for all female students at Georgetown.

The researchers would like to make a 95%-confidence interval for mu1 - mu2.

Help them out: which of the following bits of R-code will give the researchers the results they need?

ttestGC(~napkins,data=napkins)

Feedback: This produces a 95% confidence interval for mu, the mean number of napkins used by all GC students (females and males together). It does not help you see which sex uses more napkins.

ttestGC(sex~napkins,data=napkins)

Feedback: If you run this code, R will throw an error at you: the variables in the formula are the wrong way around.

ttestGC(napkins~sex,conf.level=0.95)

Feedback: If you run this code, R will throw an error at you: you need to specify that the variables **napkins** and **sex** come from the napkins data frame, otherwise R won't be able to find them.

ttestGC(napkins~sex,data=napkins)

Feedback: Close, but not quite. The way we defined the parameters, the GC males are the first population, and we need to let R know this. Otherwise R sees "female" before "male" in alphabetical order, and thinks of the females in the study as the first sample.

ttestGC(napkins~sex,data=napkins,  
 first="male")

Feedback: Good job. Notice that for a 95%-confidence interval you don't need to specify the confidence level: R's default confidence level is 95%.

# Napkins II: Safety

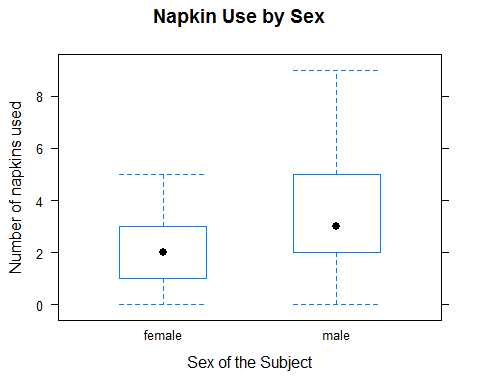
We continue with the napkins study.

The researchers ran code to make a 95%-confidence interval, getting the following output:

##   
##   
## Inferential Procedures for the Difference of Two Means mu1-mu2:  
## (Welch's Approximation Used for Degrees of Freedom)  
## napkins grouped by sex   
##   
##   
## Descriptive Results:  
##   
## group mean sd n  
## male 3.628 2.381 43  
## female 2.279 1.453 43  
##   
##   
## Inferential Results:  
##   
## Estimate of mu1-mu2: 1.349   
## SE(x1.bar - x2.bar): 0.4253   
##   
## 95% Confidence Interval for mu1-mu2:  
##   
## lower.bound upper.bound   
## 0.500453 2.197222

They also decided to look at graphs of the samples (this is always a good idea):

bwplot(napkins~sex,data=napkins,  
 main="Napkin Use by Sex",  
 xlab="Sex of the Subject",  
 ylab="Number of napkins used")



Run the boxplot code yourself and study the graphs. Regarding the question of whether or not we can trust the results of confidence interval procedure, which of the following analyses is best?

Option A:

Ideally, both populations would be exactly normally distributed, but this never happens in practice. The distribution of napkins for the males is a bit right-skewed, indicating that the same may be true for the male population, but the skewness is not extreme, and the sample size is 43, which is large enough to keep the results reliable even in the presence of skewness. The females also had a large enough sample size, and even less skewness. If the two samples are reasonably like simple random samples from their respective populations, then we can trust the confidence interval!

Option B:

For the confidence intervals to work exactly as advertised, both populations would have be exactly normally distributed. The distribution of napkins for the males is a bit right-skewed, indicating that the same may be true for the male population. We cannot rely on the confidence interval.

Option C:

Both samples sizes are bigger than 30, so we can trust the confidence interval.

Feedback: Don't treat 30 as a sacred cut-off. Even at large sample sizes, extreme skewness can be a problem. Also, this response ignores the really important issue: are the two samples like random samples from their respective populations?

# Napkins III: Interval and Interpretation

We continue with the napkins study.

The researchers have made the confidence interval:

##   
##   
## Inferential Procedures for the Difference of Two Means mu1-mu2:  
## (Welch's Approximation Used for Degrees of Freedom)  
## napkins grouped by sex   
##   
##   
## Descriptive Results:  
##   
## group mean sd n  
## male 3.628 2.381 43  
## female 2.279 1.453 43  
##   
##   
## Inferential Results:  
##   
## Estimate of mu1-mu2: 1.349   
## SE(x1.bar - x2.bar): 0.4253   
##   
## 95% Confidence Interval for mu1-mu2:  
##   
## lower.bound upper.bound   
## 0.500453 2.197222

Find the interval in the output above and say which of the following statements provide the best interpretations of it:

Option A:

We are 95% confident that the mean number of napkins used at lunch by all GC males exceeds the mean number used by all GC females, by an amount that is somewhere between 0.5 and 2.2 napkins.

Option B:

There is about a 95% chance that the mean number of napkins used at lunch by all GC males exceeds the mean number used by all GC females, by an amount that is somewhere between 0.5 and 2.2 napkins.

Feedback: This is tempting, but a bit wrong. mu1 - mu2 is a definite number that is either in the interval or not -- no "chances" about it.

Option C:

If you look at all pairs of (one GC male and one GC female), you would find that in about 95% of those pairs the male uses between 0.5 and 2.2 more (i.e. 1 or 2 more) napkins than the female does.

Feedback: The confidence interval does not try to capture individual values: it only tries to capture mu1 - mu2.

# Napkins IV: Drawing a Conclusion

Let's finish up the napkins example. Again, here is the output:

##   
##   
## Inferential Procedures for the Difference of Two Means mu1-mu2:  
## (Welch's Approximation Used for Degrees of Freedom)  
## napkins grouped by sex   
##   
##   
## Descriptive Results:  
##   
## group mean sd n  
## male 3.628 2.381 43  
## female 2.279 1.453 43  
##   
##   
## Inferential Results:  
##   
## Estimate of mu1-mu2: 1.349   
## SE(x1.bar - x2.bar): 0.4253   
##   
## 95% Confidence Interval for mu1-mu2:  
##   
## lower.bound upper.bound   
## 0.500453 2.197222

Recall that the Research Question was:

*Who uses more napkins at lunch, on average: a GC male or a GC female?*

Which of the following is the best way to use the confidence interval to answer the Research Question?

Option A:

If the two sexes used the same number of napkins on average, then the difference in means mu1 - mu2 would be 0. All of the values in the 95% confidence interval lie above 0, so we are quite confident, based on the data, that mu1 is bigger than mu2, which means that on average GC guys use more napkins than GC gals do.

Option B:

Guys use more napkins than gals do, because the interval lies above 0.

Feedback: Help the reader out a little bit: tell him/her what the value 0 means, in the context of our Research Question, otherwise he/she won't be able to follow your reasoning.

Option C:

We are 95% confident that the mean number of napkins used at lunch by all GC males exceeds the mean number used by all GC females, by an amount that is somewhere between 0.5 and 2.2 napkins.

Feedback: This is a good interpretation of the confidence interval, but it doesn't address the Research Question directly.

# ACT Scores

In the tigerstats package we find the data frame hair\_and\_act:

data(hair\_and\_act)  
View(hair\_and\_act)  
help(hair\_and\_act)

This was an observational study performed at Georgetown College. A pair of student researchers attempted to make a random selection of their fellow students. One of variables they recorded was **act**, the ACt score of each student.

Assume that it is known that the mean ACT score of all student at the University of Kentucky is 24.

We are interested in the Research Question:

*Is the mean ACT score of all GC students the same as that of all UK students, do the two schools differ in their mean ACT scores?*

To investigate this Research Question, we define:

mu = the mean ACT score of all GC students.

We decide to make a 95%-confidence interval for mu, as follows:

ttestGC(~act,data=hair\_and\_act)

##   
##   
## Inferential Procedures for One Mean mu:  
##   
##   
## Descriptive Results:  
##   
## variable mean sd n  
## act 23.83 3.806 100  
##   
##   
## Inferential Results:  
##   
## Estimate of mu: 23.83   
## SE(x.bar): 0.3806   
##   
## 95% Confidence Interval for mu:  
##   
## lower.bound upper.bound   
## 23.074799 24.585201

Based on the confidence interval, what is the best conclusion we can make concerning our Research Queston?

Option A:

The mean ACT for all UK students is 24, and our interval contains this value. Hence we cannot conclude that GC students differ from UK students, as far as mean ACT score is concerned.

Option B:

We are 95% confident that the mean ACT score for all GC student is somewhere between 23 and 24.6.

Feedback: This is a good interpretation of the interval itself, but it does not address the Research Question.

Option C:

UK's mean is 24, and this is close to the upper bound of our interval. It looks like UK's mean is higher than GC's mean ACT score.

Feedback: Granted, UK's mean is not in the very middle of the interval, but it is inside the interval at least. Isn't the interval supposed to provide a range of reasonable values, based on the data, for what the mean GC ACT score could be? So doesn't 24 have to be counted as a "reasonable" value?

# Summary Data I

We would like to estimate mu, the mean length of all Great White sharks, so we sample 16 Great Whites at random and find that the sample mean length is 15 feet. The standard deviation of the sample lengths is 3 feet. Compute a 90% (\*\*not a 95%) confidence interval for mu, using this summary data.

What is the upper bound for the interval? Write your answer below, correct to at least two decimal places.

ttestGC(mean=15,sd=3,n=16, conf.level=.90)

Answer is 16.314788

# Summary Data II

Thinking back on the Great White study from the previous question, what would you say is the most nagging problem in using the confidence interval we obtained?

Option A:

The sample size of 16 is fairly small. In order to be sure that the interval is reliable, we should make a graph of the sample and check it for skewness and outliers, but we can't do this when we have only summary data.

Feedback: Yes, since we know we took a random sample of sharks, the remaining "safety" issue is whether or not the population is "too far" from being normally distributed. To check this you look at a graph of your sample, and we cannot do this!

Option B:

Since the sample size is small, the confidence interval will be rather wide. We won't have a very precise idea of what mu could actually be.

Feedback: Yes, the interval will be wide, but that doesn't by itself mean that we can't be 90%-confident that mu lies within it.

Option C:

Males and females probably differ in length, but with a small sample it's possible that by chance we got mostly males, or mostly females. This would result in poor estimates of mu.

Feedback: This is a possibility, but confidence intervals do take sampling variability into account. At small sample sizes they will be wider, partly in order to account for the increased likelihood that the sample won't be a good cross-section of the population.

# GC Voting I

It is known that 60% of all Ohio State students plan to vote for the Democratic candidate in the next presidential election. In a survey of 200 randomly-selected Georgetown College students, it is found that 101 of them plan to vote for the Democratic candidate.

The Research Question is:

*Is the level of support for the Democratic candidate lower at GC than it is at OSU?*

To investigate this question, let's define:

p = the proportion of ALL GC students who plan to vote for the Democratic candidate.

We'll use our summary data to make a confidence interval for p.

Which of the following bits of R-code would produce an acceptable interval? There are two correct answers.)

binomtestGC(x=101,n=200)

Feedback: Yes, this would be the most commonly-used option, since it produces "exact" confidence intervals.

proptestGC(x=101,n=200)

Feedback: This option is also acceptable. Although it relies on the normal approximation, the approximation is quite good, as there are 101 success and 99 failures in the sample (so both successes and failures are at least 10).

ttestGC(mean=101,sd=0.60,n=200)

Feedback: This doesn't make sense. When you are studying a categorical variable, ttestGC() should not be used: it is for numerical variables.

binomtestGC(x=200,n=101)

Feedback: R will throw an error at you: you switched around the number of successes and the sample size.

# GC Voting II

The output for the confidence interval turns out to be:

##   
## Exact Binomial Procedures for a Single Proportion p:  
## Results based on Summary Data  
##   
##   
## Descriptive Results: 101 successes in 200 trials  
##   
## Inferential Results:  
##   
## Estimate of p: 0.505   
## SE(p.hat): 0.0354   
##   
## 95% Confidence Interval for p:  
##   
## lower.bound upper.bound   
## 0.433587 0.576263

Recall that we have defined p to be the proportion of all GC students who plan to vote Democratic.

Which of the following statements is correct? (More than one could be correct.)

Option A:

We estimate that p is around 50.5%, but we could easily be off by 3.54% or so.

Feedback: That's a good way to use p-hat, the sample proportion, together with the SE for p-hat. Note that we converted form proportions to percents: that's fine.

Option B:

We estimate that p is around 50.5%, and our estimate cannot be in error by more than 3.54%.

Feedback: No, in principal we could be off by much, much more. We aren't likely to be more than a couple of standard errors off, though.

Option C:

A rough 68% confidence interval for p would be 46.96% to 54.04%.

Feedback: Yes, that's p-hat plus/minus one SE for p-hat.

# GC Voting III

Recall that it was known that 60% of all OSU students would vote Democratic, and that the Research Question was:

*Is the level of support for the Democratic candidate lower at GC than it is at OSU?*

Which of the following is the best way to answer the Research Question?

Option A:

All of the values in the 95% confidence interval are less than 0.60, so we are confident that the proportion of all students who would vote Democratic is lower at GC than it is at OSU.

Option B:

We are 95% confident that somewhere between 43.57% and 57.43% of all GC students vote Democratic.

Feedback That's a good interpretation of the confidence interval, but it does not address the Research Question.

Option C:

A GC student is 43.57% to 57.4% less likely to vote Democratic than an OSU student is.

Feedback: Remember that the confidence interval is trying to capture p, not to compare chances for Gc students with chances for OSU students.

Option D:

There is a 95% chance that Gc students vote Democratic at a lower rate than OSU students do.

Feedback: The confidence interval doesn't give "chances" that p is less than 0.60.

# Home Runs I

Babe Ruth and Ralph Kiner are two Hall of Fame baseball legends who both hit home runs frequently. The Babe, who played in the Majors from 1914 to 1935, smacked 714 homers in a total of 8399 at-bats, for a home-run rate of 714/8399 = 0.07089, or about 8.5%. Ralph Kiner, who was active from 1946 to 1955, hit 369 homers in 52095 at-bats, for a home-run rate of 7.09%.

Our Research Question is:

*Did the Babe actually have a higher probability of hitting a home run when he stepped up to the plate, or could the difference in performance between the two be explained in terms of chance variation?*

After all, chance is at play when a hitter goes to bat: whether or not one hits a home run is due to all sorts of factors -- wind direction and speed, whether or not the pitcher throws a poorly-chosen pitch, etc.

In order to investigate this question, let's define:

p1 = probability that the Babe hits a home run when he goes to bat

p2 = probability that Ralph hits a home run when he steps up to the plate

We'll make a 95% confidence interval for p1 - p2.

What's the right R-code?

proptestGC(x=c(714,369),n=c(8399,5205))

proptestGC(x=c(369,714),n=c(5205,8399))

Feedback: Close, but not quite. As we defined the parameters, this will give you a confidence interval for p2 - p1, not for p1 - p2.

proptestGC(x=c(714,369),n=c(5205,8399))

Feedback: You've got the Babe's HR count mixed up with Ralph's at-bat count.

binomtestGC(x=c(714,369),n=c(8399,5205))

Feedback: binomtestGC() is only available for studying one proportion at a time.

# Home Runs II

Here are the results in the study involving the Babe and Ralph:

##   
##   
## Inferential Procedures for the Difference of Two Proportions p1-p2:  
## Results taken from summary data.  
##   
##   
## Descriptive Results:  
##   
## successes n estimated.prop  
## Group 1 714 8399 0.08501  
## Group 2 369 5205 0.07089  
##   
##   
## Inferential Results:  
##   
## Estimate of p1-p2: 0.01412   
## SE(p1.hat - p2.hat): 0.004681   
##   
## 95% Confidence Interval for p1-p2:  
##   
## lower.bound upper.bound   
## 0.004941 0.023292

Select the best conclusion, one that uses the confidence interval to answer the original Research Question.

Option A:

If Ralph and the Babe both had the same chance of hitting a home run, then p1 - p2 would be 0. But the confidence interval contains only values that are above 0, so we are confident that Babe's chance of a home run is higher than Ralph's.

Option B:

We are 95% confident that the difference in their HR probabilities (Babe minus Ralph) is somewhere between 0.5% and 2.3%.

Feedback: That's a great interpretation of the confidence interval, but it does not directly address the Research Question.

Option C:

The Babe's chance of hitting a home run is about 1.4% higher than Ralph's chance of hitting a home run.

Feedback: It's true that p1-hat - p2-hat is about 0.014, or 1.4%, but that's just the difference in sample proportions. You need to use the confidence interval if you want to address the question about the difference in underlying HR probabilities.

Option D:

We estimate that the Babe's chance of a HR exceeds Ralph's chance by about 1.4%, give or take 0.47% or so.

Feedback: That's a wonderful use of standard error to talk about how far off your estimate of p1 -p2 might be, but you need to say something that resolves the Research Question directly.

# 95% Confidence

Two hundred statisticians are interested in estimating q, the proportion of all U.S. adults who plan to vote Democratic in the next presidential election. Each one, independently of the others, take a simple random sample of size 2500 from a population, and on the basis of that sample computes a 95%-confidence interval for q using binomtestGC().

About how many of these statisticians would you expect to get an interval that contains q? Give or take about how many?

Option A:

About 190 of them, five or take 3.08 or so.

Feedback: Right. Each interval will either contain the mean or it won't. Each statistician has a 95% chance of getting an interval that contains q, independently of the other statisticians. Hence the number X of statisticians who get an interval that contains q is a binomial random variable, with n = 200 trials and chance of success p = 0.95. The expected value of X is n*p = 200*0.95 = 190, and the SD of X is sqrt(200*p*(1-p)) = 3.08.

Option B:

All of them, no give or take about it. Each one took a simple random sample, so you can trust the intervals they obtained.

Feedback: Even if your statistical method is a good one, couldn't you still get a really unlucky sample that leads you to a wrong conclusion? Review where "95%-confidence" comes from.

Option C:

About half of them, give or take 10 or so.

Feedback: Say that you are going to make a 95%-confidence interval. Before you take the sample, what's the chance that the interval you make will contain the parameter you are trying to estimate?

Option D:

In order to answer this question, we would have to know the value of q.

Feedback: You have enough information. Try again.

# UC-Davis Seating I: the Code

Recall the ucdavis1 data frame:

data(ucdavis1)  
View(ucdavis1)  
help(ucdavis1)

We may consider this data to come from a simple random sample of all UC-Davis students. The variable **Seat** records where a student prefers to sit in a classroom: front, back, or middle.

We would like to know if more than half of all UC-Davis students prefer to sit in the middle.

For descriptive purposes, here is a two way table:

ucdSeat <- xtabs(~Seat,data=ucdavis1)  
ucdSeat

## Seat  
## Back Front Middle   
## 37 41 93

And here is a table of percentages:

rowPerc(ucdSeat)

## Back Front Middle Total  
## 21.64 23.98 54.39 100.00

More than half of the sample prefers the middle, but can we conclude that more than half of the *population* prefers the middle?

Which of the following bits of R-code will produce a 95%-confidence interval for the proportion p of all UC-Davis students who prefer the middle? There are two correct responses.

binomtestGC(~Seat,data=ucdavis1,success="Middle")

Feedback: Yes, this is probably the easiest approach.

binomtestGC(~Seat,data=ucdavis1,success="middle")

Feedback: Be very careful about spelling, and remember that R is case-sensitive!

binomtestGC(~Seat,data=ucdavis1)

Feedback: If you try this code, R will say that no successes were found. You need to specify what counts as a success!

binomtestGC(x=93,n=171)

Feedback: Right. You could use the results of the two-way table to feed summary data to R. You have to do some math though. The sample size is n = 37+41+93 = 171.

binomtestGC(x=93,n=78)

Feedback: You don't have the right sample size for n.

# UC-Davis Seating II: Answering the Question

Here is the output for a confidence interval in the UC-Davis seating study:

##   
## Exact Binomial Procedures for a Single Proportion p:  
## Variable under study is Seat   
##   
##   
## Descriptive Results: 93 successes in 171 trials  
##   
## Inferential Results:  
##   
## Estimate of p: 0.5439   
## SE(p.hat): 0.0381   
##   
## 95% Confidence Interval for p:  
##   
## lower.bound upper.bound   
## 0.466083 0.620090

What's the best way to use the interval to answer the original research question?

Option A:

If more than half of all UC-Davis students prefer the middle, then the proportion p that prefer the front would be bigger than 0.50. However, the confidence interval begins at 0.466, so it contains plenty of values that are LESS than 0.50. Hence we cannot conclude with confidence that p is greater than 0.5.

Option B:

If more than half of all UC-Davis students prefer the middle, then the proportion p that prefer the front would be bigger than 0.50. The confidence interval ends at 0.62, so it contains plenty of values that are bigger than 0.50. Hence we can conclude with confidence that p is bigger than 0.5.

Feedback: Yes but the interval also contains values that are less than 0.5. Since values in the interval are considered "reasonable", based on the data, we cannot rule out the notion that p is actually less than 0.5.

Option C:

We are 95% confident that the proportion of all UC-Davis student who prefer the middle is somewhere between 46.6% and 62%.

Feedback: That's a very nice interpretation of the confidence interval, but it does not grapple with the question of whether we can conclude that the proportion is bigger than 0.5.

Option D:

It is not appropriate to study our research question with this type of confidence interval.

Feedback: Sure it is.

# First-Sight Love I

Recall the m111survey data frame:

data(m111survey)  
View(m111survey)  
help(m111survey)

It contains information on the sex of each subject in the survey (**sex**, with values "female" and "male") and whether or not the subject believes in love at first sight (**love\_first**, with values "yes" and "no").

Here is a two-way table of the responses:

sexLove <- xtabs(~sex+love\_first,data=m111survey)  
sexLove

## love\_first  
## sex no yes  
## female 22 18  
## male 23 8

Here are the row percentages:

rowPerc(sexLove)

## no yes Total  
## female 55.00 45.00 100  
## male 74.19 25.81 100

That looks like an impressive percentage difference (45% yes for the gals, vs. only 25.81% for the guys). Does this difference mean that GC gals are more likely than GC guys to believe in love at first sight, or could the difference be due just to chance variation in the sampling process?

To answer this question, we define:

p1 = the proportion of all GC females who believe in love at first sight

p2 = the proportion of all GC males who believe in love at first sight

figure out the code to make a 90% confidence interval for p1 - p2. (**Note**: it's 90%, not the usual 95%.) Report the lower bound of the confidence interval below, making sure your answer is correct to at least two decimal places.

0.009039

# First-Sight Love II

Assuming the students in the m111survey data frame are like a random sample from the population of all GC students, what do you think might be the most important concerns about the reliability of a confidence interval for p1 - p2 obtained using proptestGC()? Choose the most reasonable answer below.

Option A:

In order for the normal approximation to yield reliable confidence intervals, you need lots of successes and failures in each sample. Ten or more of each is generally considered "enough." Since there were only eight males who believed in love at first sight, the confidence interval made by proptestGC() for p1 - p2 should be used with some caution.

Feedback: This makes the most sense. After all, only one of the four counts in the two-way table was below 10, and 8 is pretty close to 10.

Option B:

In order for the normal approximation to yield reliable confidence intervals, you need lots of successes and failures in each sample. Ten or more of each is generally considered "enough." Since there were only eight males who believed in love at first sight, the confidence interval made by proptestGC() for p1 - p2 should be ignored altogether.

Feedback: That's too strict. After all, 8 is pretty close to 10.

Option C:

You can't use proptestGC() to study a question about the difference of two proportions.

Feedback: Sure you can!

Option D:

If you are concerned about proptestGC(), why not use binomtestGC() instead? It delivers "exact" confidence intervals!

Feedback: But only for one proportion, not for the difference of two proportions.