Course Review

Clark

12/9/2019

Six Step Investigative Process

State the Research Question

Design a study and collect data

Explore the data

Draw inferences beyond the data

Formulate Conclusions

Look back and ahead

First thing we do is define our variables. Is our response variable categorical or continuous?

The second thing to do is determine if we have one group, two groups, or a regression? In other words, what is our explanatory variable.

The difference situations we could be in, with appropriate parameters, are:

The next step is to explore the data and calculate our statistic. Depending on our situation, the different statistics we could calculate are:

The next thing we want to do is to draw inferences beyond the data. This involves testing a hypothesis. Our hypothesis we might be testing are:

Next we want to calculate a P-value to determine the probability we would have observed something as extreme (or more) from statistic assuming H_0 is true. We can either simulate from H_0 and compare our statistic to the values obtained, or we could use a theory based method. The validity conditions depending on our situation are:

Next we want to formulate conclusionsHere we report a P-value as well as potentially a Confidence Interval. Recall that a confidence interval is the range of plausable values. We often say a 95% CI gives us 95% confidence that our parameter is contained within that interval.

We also want to determine causality. The best picture for this is Lesson 13 of the course guide.

Finally we want to think about how we can improve/extend the study.

Some example problems

A Pew Research survey of nationally representative 242 cell phone useres, ages 16 to 17 years, found that 126 had talked on the phone while driving.

Use these data to estimate with 95% confidence, the proportion of all 16 to 17 year old cell phone useres who talk on the phone while driving

Do these data provide evidence that a majority of 16 to 17 year old cell phone useres talk on the phone while driving?

If we sampled more users would our 95% CI increase or decrease in width?

Are our validity conditions met to use a theory based test here?

On average Americans have visited 16 states in the United States. In a survey of 50 students, I found the average number of states the students had visited to be 9.48 and the standard deviation to be 7.13

What are the observational units for this study?

Is the variable of interest categorical or quantitative?

State the null and alternative hypothesis in symbols to test whether the average number of states all students visited is different then 16.

What are our validity conditions to use a theory based method?

Find a 95% CI using the 2SD method. From the CI can we determine if our H_a is true or not?

From a 2013 Gallup poll that asked randomly selected US adults whether they wanted to stay at their current body weight or change, of 562 men surveyed, 242 wanted to stay at their current weight, whereas of the 477 women surveyed, 172 wanted to stay at their current weight.

Define the parameters of interest of this study. Assign appropriate symbols to the parameters

State the appropriate null and alternative hypothesis if we are interested in determining if there's a difference between mena nd women

Is it valid to use a theory-based approach? Why or why not?

What R command would you use in order to find the P-value?

On the blackboard site there's a file called 'candy.txt'. Use this to answer the following questions. Researchers were interested in whether Payday or Mounds candy bars weighed differently, so they randomly selected 20 of each variety and weighed them.

What are the explanatory variables and what are the response variables for this study?

Write down the null and alternative hypothesis (in symbols) to determine if the candy bars weighed different?

Are your validity conditions met to use a theory based test?

Use a theory based test to find a p-value and answer the researchers question

A team of reserachers investigated whether hours slept or gender impacted a students GPA

What are the observational/experimental units?

The data are located at the blackboard site in a file called `sleep.csv`.

```
sleep.dat<-read.csv("sleep.csv")
sleep.lm<-lm(GPA~Sex+Sleep_hours,data=sleep.dat)
summary(sleep.lm)
```

```
##
## Call:
## lm(formula = GPA ~ Sex + Sleep_hours, data = sleep.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90894 -0.20926  0.01496  0.25868  0.65702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.85832    0.21028  13.593  < 2e-16 ***
## SexMale     -0.10422    0.05921  -1.760  0.08066 .
## Sleep_hours  0.08413    0.02975   2.827  0.00541 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3374 on 134 degrees of freedom
## Multiple R-squared:  0.07487,    Adjusted R-squared:  0.06106
## F-statistic: 5.422 on 2 and 134 DF,  p-value: 0.005441
```

Write out the fitted least squares regression model.

What does it mean that `Sleep_hours` has a P value of 0.00541?

What is the difference in Males and Females GPA for two students who sleep the same amount?

What does it mean that `Sleep_hours` has a coefficient of 0.08413?

Next the model

```
sleep2.lm<-lm(GPA~Sex+Sleep_hours+Sex:Sleep_hours,data=sleep.dat)
summary(sleep2.lm)
```

```
##
## Call:
## lm(formula = GPA ~ Sex + Sleep_hours + Sex:Sleep_hours, data = sleep.dat)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90829 -0.20944  0.01525  0.25833  0.65701
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.8559851  0.2615221  10.921   <2e-16 ***
## SexMale             -0.0976344  0.4401701  -0.222   0.8248
## Sleep_hours          0.0844606  0.0372069   2.270   0.0248 *
## SexMale:Sleep_hours -0.0009414  0.0623836  -0.015   0.9880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3387 on 133 degrees of freedom
## Multiple R-squared:  0.07487,    Adjusted R-squared:  0.054
## F-statistic: 3.588 on 3 and 133 DF,  p-value: 0.01554
```

Is fit.

What would it mean if there was an interaction effect between Sex and Sleep_hours?

Do the data suggest that there is an interaction effect?

Does this model do a good job explaning the data? Why or why not?