hw6

jason grahn

11/13/2018

download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")  
load("nc.RData")

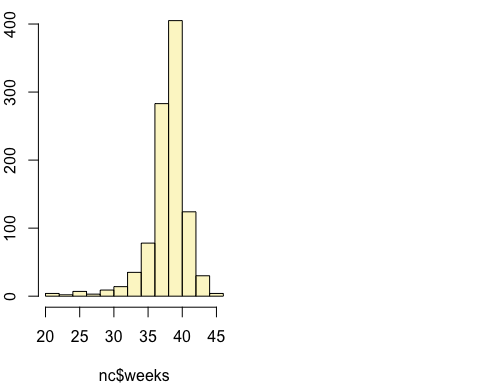
## Inference for numerical data

### 1.

Calculate a 95% confidence interval for the average length of pregnancies (weeks) and interpret it in context. Note that since you’re doing inference on a single population parameter, there is no explanatory variable, so you can omit the x variable from the function.

inference(y = nc$weeks,   
 est = "mean",   
 type = "ci",   
 method = "theoretical",  
 conflevel = 0.95)

## Single mean   
## Summary statistics:



## mean = 38.3347 ; sd = 2.9316 ; n = 998   
## Standard error = 0.0928   
## 95 % Confidence interval = ( 38.1528 , 38.5165 )

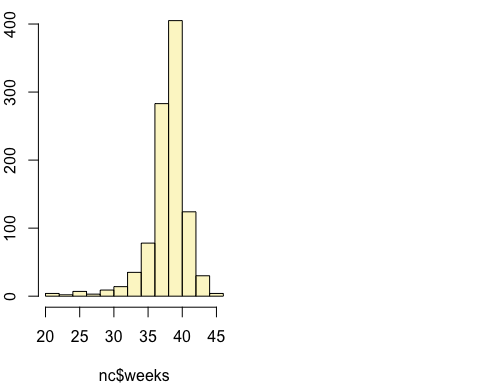
Assuming normality and independence, we are 95% confident that the true mean of pregnancy length is between 38.3347 and 38.5165 weeks.

### 2.

Calculate a new confidence interval for the same parameter at the 90% confidence level. You can change the confidence level by adding a new argument to the function: conflevel = 0.90.

inference(y = nc$weeks,   
 est = "mean",   
 type = "ci",   
 method = "theoretical",  
 conflevel = 0.90)

## Single mean   
## Summary statistics:



## mean = 38.3347 ; sd = 2.9316 ; n = 998   
## Standard error = 0.0928   
## 90 % Confidence interval = ( 38.182 , 38.4873 )

### 3.

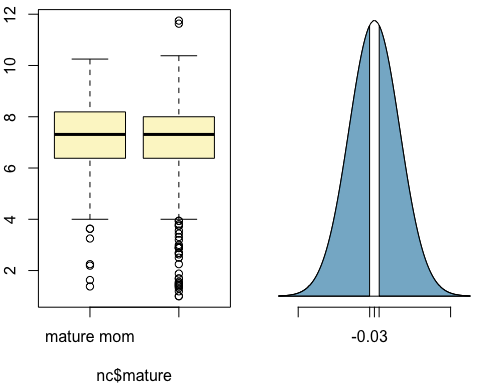
Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers. Make sure you include the hypotheses set up and final interpretation.

\* H0 = The average weight gained by younger and mature mothers is not different.   
\* HA = The average weight gained by younger and mature mothers is different.

inference(y = nc$weight,   
 x = nc$mature,   
 type="ht",   
 est="mean",   
 null=0,   
 method="theoretical",   
 alternative="twosided")

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_mature mom = 133, mean\_mature mom = 7.1256, sd\_mature mom = 1.6591  
## n\_younger mom = 867, mean\_younger mom = 7.0972, sd\_younger mom = 1.4855

## Observed difference between means (mature mom-younger mom) = 0.0283  
##   
## H0: mu\_mature mom - mu\_younger mom = 0   
## HA: mu\_mature mom - mu\_younger mom != 0   
## Standard error = 0.152   
## Test statistic: Z = 0.186   
## p-value = 0.8526



With a p-value of 0.8526, we cannot reject the null hypothesis and cannot conclude that the mean difference in weights is statistically significant.

### 4.

Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

We subset the mother’s age (mage) variable by the maturity (mature) variable then provides the numerical sumamry for the subset. The max of younger should line up with the min by mature.

by(nc$mage,   
 nc$mature,  
 summary)

## nc$mature: mature mom  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 35.00 35.00 37.00 37.18 38.00 50.00   
## --------------------------------------------------------   
## nc$mature: younger mom  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.00 21.00 25.00 25.44 30.00 34.00

Sure enough, we see the max of younger is 34 where the min age of mature is 35. This seems to be a valid cutoff age.

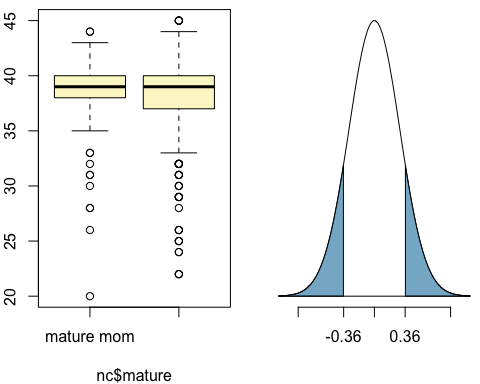
### 5.

Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the inference function, report the statistical results, and also provide an explanation in plain language.

#Question: Is there a difference between the length of pregnancy for a mature or younger mother?   
#H0 = There average length of pregnancy between younger and mature mothers is not different.   
#HA = There average length of pregnancy between younger and mature mothers is different.  
  
inference(nc$weeks,   
 nc$mature,   
 type="ht",   
 est="mean",   
 null=0,   
 method="theoretical",   
 alternative="twosided")

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_mature mom = 132, mean\_mature mom = 38.0227, sd\_mature mom = 3.2184  
## n\_younger mom = 866, mean\_younger mom = 38.3822, sd\_younger mom = 2.8844

## Observed difference between means (mature mom-younger mom) = -0.3595  
##   
## H0: mu\_mature mom - mu\_younger mom = 0   
## HA: mu\_mature mom - mu\_younger mom != 0   
## Standard error = 0.297   
## Test statistic: Z = -1.211   
## p-value = 0.2258



With a p-value of 0.2258, we cannot reject the null hypothesis and cannot conclude that the mean difference in weights is statistically significant.

## Inference for categorical data

download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")  
load("atheism.RData")

### 6.

Answer the following two questions using the inference function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

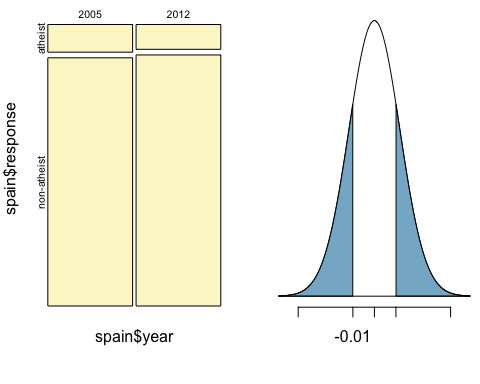
#### a.

Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012? Hint: Create a new data set for respondents from Spain. Then use their responses as the first input on the inference, and use year as the grouping variable.

#H0 = There is no evidence that the atheism index in Spain has seen a change from 2005 to 2012; that is p(2005) is equal to p(2012).  
#HA = There is evidence that the atheism index in Spain has seen a change from 2005 to 2012; that is p(2005) is not equal to p(2012).  
spain <- atheism %>%   
 filter(nationality == "Spain",  
 year == 2005 | 2012) %>%   
 mutate(year = as.factor(year))  
  
inference(y = spain$response,   
 x = spain$year,   
 est = "proportion",  
 type = "ht",   
 null = 0,   
 alternative = "twosided",   
 method = "theoretical",   
 success = "atheist")

## Response variable: categorical, Explanatory variable: categorical  
## Two categorical variables  
## Difference between two proportions -- success: atheist  
## Summary statistics:  
## x  
## y 2005 2012 Sum  
## atheist 115 103 218  
## non-atheist 1031 1042 2073  
## Sum 1146 1145 2291

## Observed difference between proportions (2005-2012) = 0.0104  
##   
## H0: p\_2005 - p\_2012 = 0   
## HA: p\_2005 - p\_2012 != 0   
## Pooled proportion = 0.0952   
## Check conditions:  
## 2005 : number of expected successes = 109 ; number of expected failures = 1037   
## 2012 : number of expected successes = 109 ; number of expected failures = 1036   
## Standard error = 0.012   
## Test statistic: Z = 0.848   
## p-value = 0.3966



With a p-value of 0.3966, we fail to reject the null hypothesis. We conclude that the evidence supports there is no change in Spain’s atheism index between 2005 and 2012.

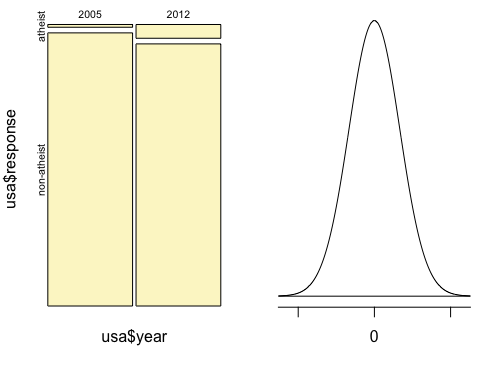
#### b.

Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

#H0 = There is no evidence that the atheism index in USA has seen a change from 2005 to 2012; that is p(2005) is equal to p(2012).  
#HA = There is evidence that the atheism index in USA has seen a change from 2005 to 2012; that is p(2005) is not equal to p(2012).  
usa <- atheism %>%   
 filter(nationality == "United States",  
 year == 2005 | 2012) %>%   
 mutate(year = as.factor(year))  
  
inference(y = usa$response,   
 x = usa$year,   
 est = "proportion",  
 type = "ht",   
 null = 0,   
 alternative = "twosided",   
 method = "theoretical",   
 success = "atheist")

## Response variable: categorical, Explanatory variable: categorical  
## Two categorical variables  
## Difference between two proportions -- success: atheist  
## Summary statistics:  
## x  
## y 2005 2012 Sum  
## atheist 10 50 60  
## non-atheist 992 952 1944  
## Sum 1002 1002 2004

## Observed difference between proportions (2005-2012) = -0.0399  
##   
## H0: p\_2005 - p\_2012 = 0   
## HA: p\_2005 - p\_2012 != 0   
## Pooled proportion = 0.0299   
## Check conditions:  
## 2005 : number of expected successes = 30 ; number of expected failures = 972   
## 2012 : number of expected successes = 30 ; number of expected failures = 972   
## Standard error = 0.008   
## Test statistic: Z = -5.243   
## p-value = 0



With a p-value of 0, we reject the null hypothesis and conclude there is a change in the United States atheism index between the years of 2005 and 2012.

### 7.

If in fact there has been no change in the atheism index in the countries listed in Table 4, in how many of those countries would you expect to detect a change (at a significance level of 0.05) simply by chance?

#countries on the table in the pdf = 39, this would be our (n)  
#significance level of 0.05 denotes percentage of successes, binomially & randomly; this would be our (p)  
  
#n \* p  
#so effectively, all we need to do is multiply 39 by 0.05   
+ 39 \* 0.05

## [1] 1.95

Since .95 countries can’t be a change, we round up to 2 countries expected to have a change in atheism index by chance.

### 8.

Suppose you’re hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. According to the guidelines, the estimate must have a margin of error no greater than *1%* with *95%* confidence. You have no idea what to expect for pp. How many people would you have to sample to ensure that you are within the guidelines? Hint: Refer to your plot of the relationship between pp and margin of error. Do not use the data set to answer this question.

The two variables we don’t know are p and n. When we don’t have an estimate for p we utilize 0.5 in order to find the largest margin of error; this provides us with a “worst case scenario” for these kinds of estimates.

The formula to use is ME = z \* SE;   
and SE = sqrt(p\*(1-p)/n)

so

ME = z \* sqrt(p\*(1-p)/n)

Because we are limited to a margin of error of no greater than 1% (0.01) for the estimate, and a 95% confidence level (a 0.05 confidence interval), we use

0.01 = 1.96 \* sqrt(0.5\*(1-0.5)/n)

Therefore,

n = 0.5 \* (1 - 0.5) \* 1.96 ^ 2 / 0.01 ^ 2

#filling in variables for formula   
p <-0.5   
z <- qnorm(1 - 0.05/2)  
ME <- 0.01   
n <- p \* (1 - p) \* z ^ 2 / ME ^ 2  
  
n

## [1] 9603.647

Because we can’t have only a portion of a human, we round up from 9603.6470517 to 9604 people that would have to be sampled in order to ensure we are within the guidelines.