Week5-Lab V

Aamol Gote

10/3/2019

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

names(nc)

## [1] "fage" "mage" "mature" "weeks"   
## [5] "premie" "visits" "marital" "gained"   
## [9] "weight" "lowbirthweight" "gender" "habit"   
## [13] "whitemom"

dim(nc)

## [1] 1000 13

**Excercise 1**

What are the cases in this data set? How many cases are there in our sample?

Dataset contains cases for the birth records of the children born in state of North Carolina. There are 1000 sample cases. Dataset contains 13 variables.

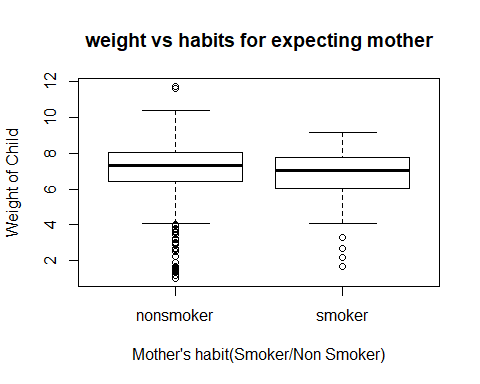
summary(nc)

## fage mage mature weeks   
## Min. :14.00 Min. :13 mature mom :133 Min. :20.00   
## 1st Qu.:25.00 1st Qu.:22 younger mom:867 1st Qu.:37.00   
## Median :30.00 Median :27 Median :39.00   
## Mean :30.26 Mean :27 Mean :38.33   
## 3rd Qu.:35.00 3rd Qu.:32 3rd Qu.:40.00   
## Max. :55.00 Max. :50 Max. :45.00   
## NA's :171 NA's :2   
## premie visits marital gained   
## full term:846 Min. : 0.0 married :386 Min. : 0.00   
## premie :152 1st Qu.:10.0 not married:613 1st Qu.:20.00   
## NA's : 2 Median :12.0 NA's : 1 Median :30.00   
## Mean :12.1 Mean :30.33   
## 3rd Qu.:15.0 3rd Qu.:38.00   
## Max. :30.0 Max. :85.00   
## NA's :9 NA's :27   
## weight lowbirthweight gender habit   
## Min. : 1.000 low :111 female:503 nonsmoker:873   
## 1st Qu.: 6.380 not low:889 male :497 smoker :126   
## Median : 7.310 NA's : 1   
## Mean : 7.101   
## 3rd Qu.: 8.060   
## Max. :11.750   
##   
## whitemom   
## not white:284   
## white :714   
## NA's : 2   
##   
##   
##   
##

**Excercise 2**

Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables?

boxplot(nc$weight~nc$habit, main="weight vs habits for expecting mother", xlab = "Mother's habit(Smoker/Non Smoker)", ylab = "Weight of Child")



Based on the side by side box plot, it is clearly evident that

1. Median weight of the child of the smoking mother is less than that of non smoking mother.
2. Non smokers data has lot of outliars compared that to smoking mother’s. It has more outliars across both whiskers.
3. There is high degree of variance in the non smoking category than that of smoking category.

by(nc$weight, nc$habit, mean)

## nc$habit: nonsmoker  
## [1] 7.144273  
## --------------------------------------------------------   
## nc$habit: smoker  
## [1] 6.82873

**Excercise 3**

Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same by command above but replacing mean with length.

by(nc$weight, nc$habit, length)

## nc$habit: nonsmoker  
## [1] 873  
## --------------------------------------------------------   
## nc$habit: smoker  
## [1] 126

Conditions for inferences: Independence and Sample Size

Sampling is random.

Sample size is 1000, so it can be safely assumed that it is less than 10% of the overall population.

**Excercise 4**

Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

Null Hypothesis : There is **no difference** in the average weights of babies born to smoking and non-smoking mothers

Alternative Hypothesis : There is **difference** in the average weights of babies born to smoking and non-smoking mothers

======================================================================

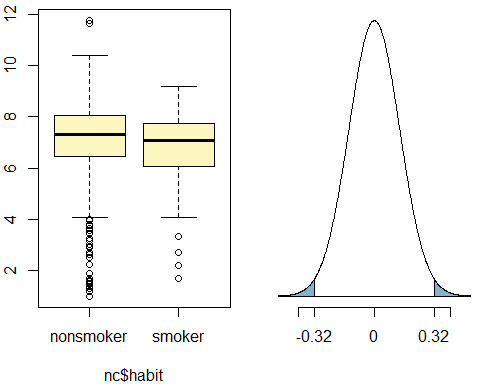
weight => Response variables

habit => explantory variable

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

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187  
## n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862

## Observed difference between means (nonsmoker-smoker) = 0.3155  
##   
## H0: mu\_nonsmoker - mu\_smoker = 0   
## HA: mu\_nonsmoker - mu\_smoker != 0   
## Standard error = 0.134   
## Test statistic: Z = 2.359   
## p-value = 0.0184

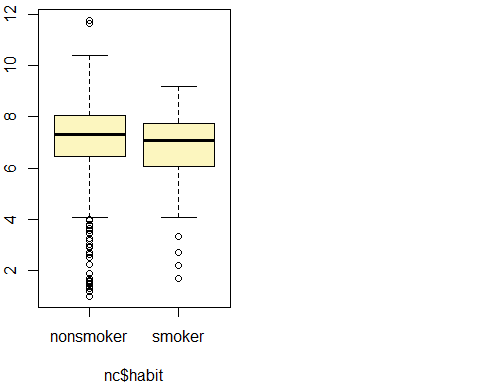


**Excercise 5**

Change the type argument to “ci” to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

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

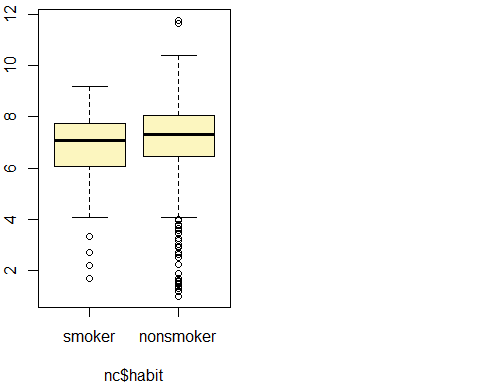
## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187  
## n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862



## Observed difference between means (nonsmoker-smoker) = 0.3155  
##   
## Standard error = 0.1338   
## 95 % Confidence interval = ( 0.0534 , 0.5777 )

inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,   
 alternative = "twosided", method = "theoretical",   
 order = c("smoker","nonsmoker"))

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862  
## n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187



## Observed difference between means (smoker-nonsmoker) = -0.3155  
##   
## Standard error = 0.1338   
## 95 % Confidence interval = ( -0.5777 , -0.0534 )

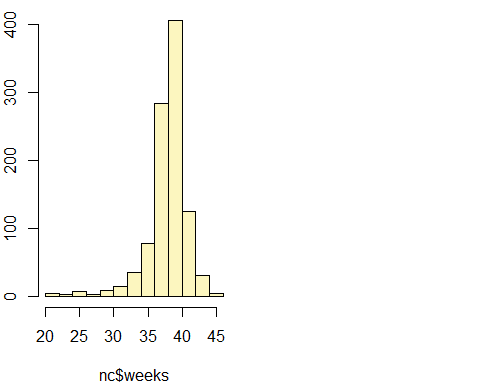
Confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers is between -0.5777 and -0.0534

**On your own**

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")

## Single mean   
## Summary statistics:



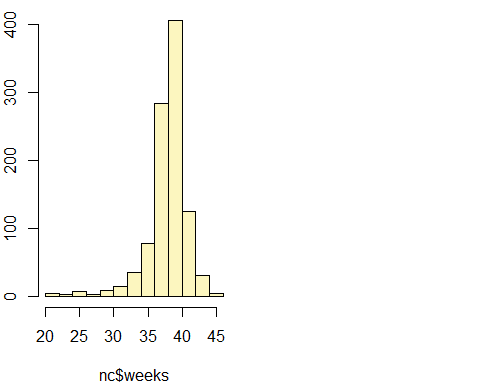
## mean = 38.3347 ; sd = 2.9316 ; n = 998   
## Standard error = 0.0928   
## 95 % Confidence interval = ( 38.1528 , 38.5165 )

So we are 95% confidence that average length of pregnancies in weeks falls between 38.1528 and 38.5165

1. 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 )

So we are 90% confidence that average length of pregnancies in weeks falls between 38.182 and 38.4873

1. Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

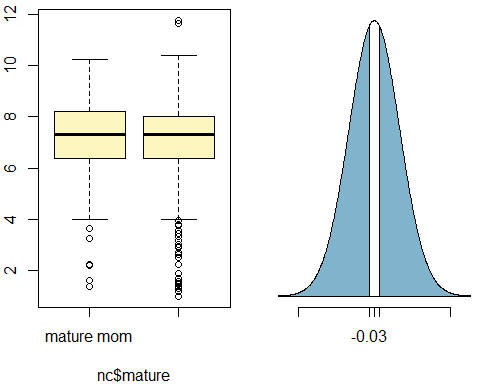
Ho: There **is no difference** in the average weight gained by young mothers and mature mothers

Ho: There **is difference** in the average weight gained by young mothers and mature mothers

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

## 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



P-value (0.8526) is higher than significance level of 0.05 so **fail to reject NULL hypothesis**. So there is not much evidence to prove that average weight gained by younger mothers is different than the average weight gained by mature mothers.

1. Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

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

Max age of younger mothers is: 34

Min age of mature mothers is: 35

So mother with age > 34 which is 35 are considered to be mature and age <=34 are considered young mothers. So the age cutoff for younger and mature mothers is **35**

1. 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.

Numerical variable : Length of pregnancy in weeks Categorical Variable: habit (Smoking/Non Smoking)

Question: Is there a difference in length of preganacy in week for smoking and non smoking mothers?

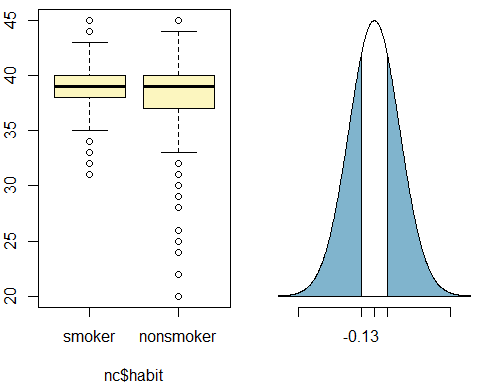
Ho: There **is no difference** in average length of preganacy in weeks for smoking and non smoking mothers

Ha: There **is difference** in average length of preganacy in weeks for smoking and non smoking mothers

inference(y = nc$weeks, x = nc$habit, est = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical",order = c("smoker","nonsmoker"))

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_smoker = 126, mean\_smoker = 38.4444, sd\_smoker = 2.4676  
## n\_nonsmoker = 872, mean\_nonsmoker = 38.3188, sd\_nonsmoker = 2.9936

## Observed difference between means (smoker-nonsmoker) = 0.1256  
##   
## H0: mu\_smoker - mu\_nonsmoker = 0   
## HA: mu\_smoker - mu\_nonsmoker != 0   
## Standard error = 0.242   
## Test statistic: Z = 0.519   
## p-value = 0.6038



In this case p-value(0.6038) is greater than 0.05 (significance level), so we fail to reject Null hypothesis. So there is not sufficient evidence to provie that there is a difference in length of preganacy in week for smoking and non smoking mothers.

We affirm the null hypothesis.