Lab5

Priya Shaji

3/15/2019

## Inference for numerical data

### Lab

#### North Carolina births

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

#### Exploratory analysis

Load the nc data set into our workspace.

1. Download the file and Load the data

library(DATA606)

## Loading required package: shiny

## Loading required package: openintro

## Please visit openintro.org for free statistics materials

##   
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':  
##   
## cars, trees

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: bitops

## Loading required package: maps

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':  
##   
## diamonds

## Loading required package: markdown

##   
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics   
## This package is designed to support this course. The text book used   
## is OpenIntro Statistics, 3rd Edition. You can read this by typing   
## vignette('os3') or visit www.OpenIntro.org.   
##   
## The getLabs() function will return a list of the labs available.   
##   
## The demo(package='DATA606') will list the demos that are available.

##   
## Attaching package: 'DATA606'

## The following object is masked from 'package:utils':  
##   
## demo

download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")  
load("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/Labs/Lab5/nc.RData")

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

variable description

1. fage father’s age in years.
2. mage mother’s age in years.
3. mature maturity status of mother.
4. weeks length of pregnancy in weeks.
5. premie whether the birth was classified as premature (premie) or full-term.
6. visits number of hospital visits during pregnancy.
7. marital whether mother is married or not married at birth.
8. gained weight gained by mother during pregnancy in pounds.
9. weight weight of the baby at birth in pounds.
10. lowbirthweight whether baby was classified as low birthweight (low) or not (not low).
11. gender gender of the baby, female or male.
12. habit status of the mother as a nonsmoker or a smoker.
13. whitemom whether mom is white or not white.

As a first step in the analysis, we should consider summaries of the data. This can be done using the summary command:

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   
##   
##   
##   
##

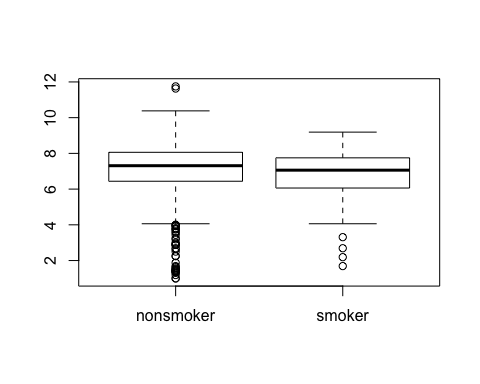
As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren’t sure or want to take a closer look at the data, make a graph.

Consider the possible relationship between a mother’s smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

Answer) fage father’s age in years. : Numerical mage mother’s age in years. : Numerical mature maturity status of mother. : Categorical weeks length of pregnancy in weeks. : Numerical premie whether the birth was classified as premature (premie) or full-term. : Categorical visits number of hospital visits during pregnancy. :Numerical marital whether mother is married or not married at birth. :Categorical gained weight gained by mother during pregnancy in pounds. :Numerical weight weight of the baby at birth in pounds. : Numerical lowbirthweight whether baby was classified as low birthweight (low) or not (not low). : Categorical gender gender of the baby, female or male. : Categorical habit status of the mother as a nonsmoker or a smoker.: Categorical whitemom whether mom is white or not white. : Categorical

Plotting a graph between a mother’s smoking habit and the weight of her baby :

plot(nc$habit,nc$weight)



non-smoker moms has few outliers for babies weighing 11 pounds and more which tells us their is less or weak association between a non smoking mother and high weight of her baby whihc is a very less likely event. According to the graph we can see, there is a strong association between mom’s who are non-smoker and the weight of their babies who weigh between approx 0.5 to 4 pounds.

Research question:

Is the baby of a mother who smokes has lower chances of being born healthy? Does the baby of a motther who smokes has higher chances of risk of diseases and incresed risk of disability?

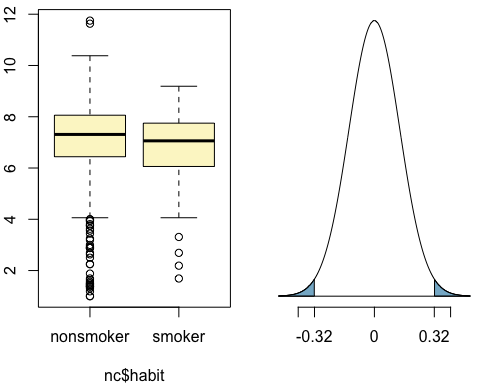
#### Inference

Next, we introduce a new function, inference, that we will use for conducting hypothesis tests and constructing confidence intervals.

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



Let’s pause for a moment to go through the arguments of this custom function. The first argument is y, which is the response variable that we are interested in: nc$weight. The second argument is the explanatory variable, x, which is the variable that splits the data into two groups, smokers and non-smokers: nc$habit. The third argument, est, is the parameter we’re interested in: "mean" (other options are "median", or "proportion".) Next we decide on the type of inference we want: a hypothesis test ("ht") or a confidence interval ("ci"). When performing a hypothesis test, we also need to supply the null value, which in this case is 0, since the null hypothesis sets the two population means equal to each other. The alternative hypothesis can be "less", "greater", or "twosided". Lastly, the method of inference can be "theoretical" or "simulation" based.

### Exercises

#### Exercise 1

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

Answer 1)

nrow(nc)

## [1] 1000

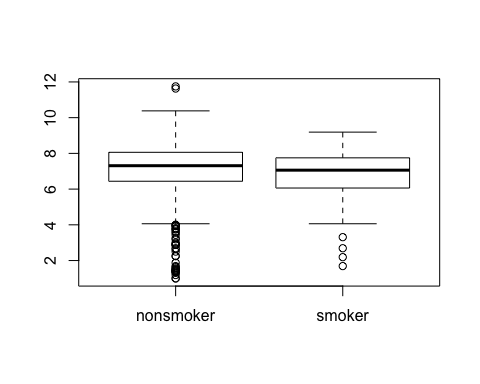
Generally, In statiscs, cases are called as observations of the dataset. Therefore, in nc dataset there are 1000 rows in a dataset

#### Exercise 2

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

Answer 2)

plot(nc$habit,nc$weight)



non-smoker moms has few outliers for babies weighing 11 pounds and more which tells us their is less or weak association between a non smoking mother and high weight of her baby whihc is a very less likely event. According to the graph we can see, there is a strong association between mom’s who are non-smoker and the weight of their babies who weigh between approx 0.5 to 4 pounds.

The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following function to split the weight variable into the habit groups, then take the mean of each using the mean function.

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

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

#### Exercise 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.

Answer 3)

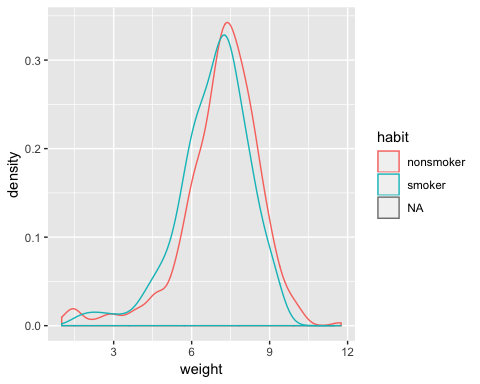
by(nc$weight, nc$habit, length)

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

Let’s check if the distribution is normal

g <- ggplot(nc, aes(x = weight, colour = habit)) + geom\_density()  
g

## Warning: Groups with fewer than two data points have been dropped.



From the above we can note the sample size is sufficiently large (>30), and their distributions approach normal, although left-skewed. We can say that the events are independent.

#### Exercise 4

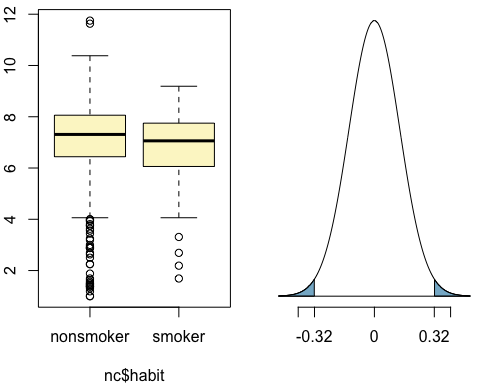
Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

Answer 4)

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



#### Exercise 5

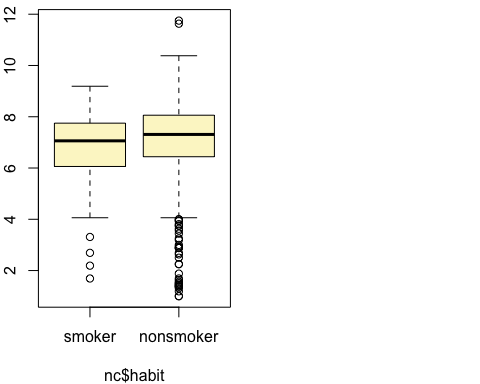
Change the type argument to “ci” to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

Answer 5)

By default the function reports an interval for (μnonsmoker−μsmoker) . We can easily change this order by using the order argument:

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 )

### On Your Own

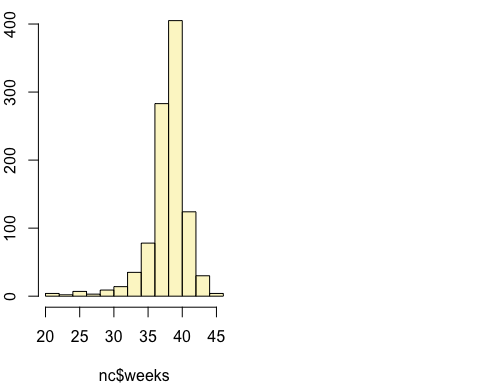
#### Question 1

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.

Answer 1)

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

## Single mean   
## Summary statistics:



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

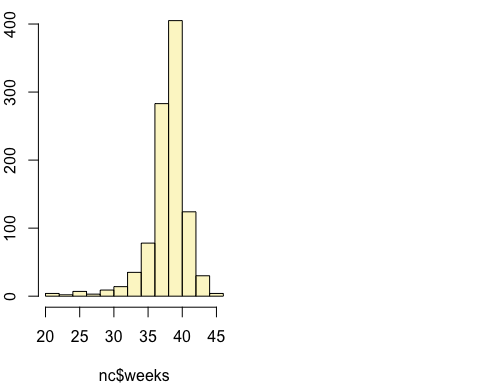
#### Question 2

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.

Answer 2)

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

#### Question 3

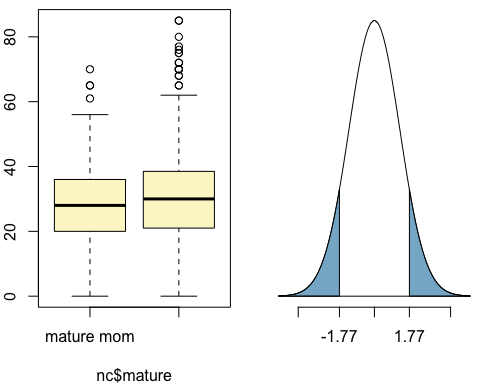
Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

Answer 3)

inference(y = nc$gained, x = nc$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 = 129, mean\_mature mom = 28.7907, sd\_mature mom = 13.4824  
## n\_younger mom = 844, mean\_younger mom = 30.5604, sd\_younger mom = 14.3469

## Observed difference between means (mature mom-younger mom) = -1.7697  
##   
## H0: mu\_mature mom - mu\_younger mom = 0   
## HA: mu\_mature mom - mu\_younger mom != 0   
## Standard error = 1.286   
## Test statistic: Z = -1.376   
## p-value = 0.1686



#### Question 4

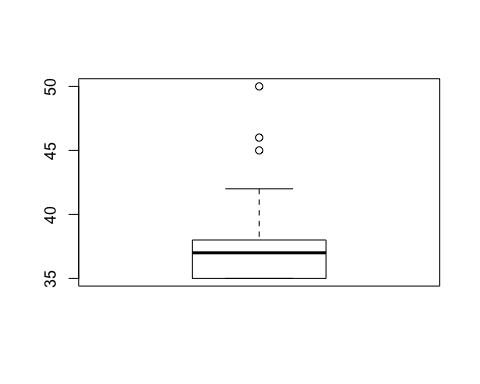
Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

Answer 4)

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

boxplot(nc$mage[nc$mature == "mature mom"])



I have used the method by() and boxplot() method to get the age cutoff for younger and mature mothers. nc$mature: mature mom , therefore for mature mom’s the minimum age cutoff is 35 and for younger mom the minimum age cutoff is 13.

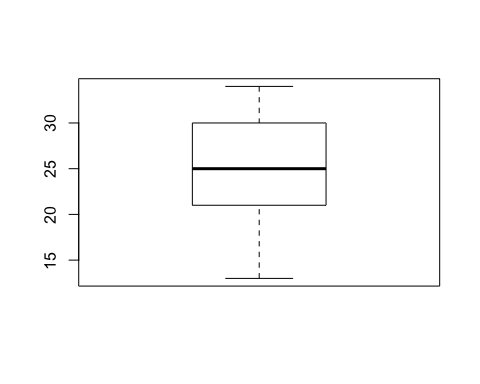
using the by() method:

compare the summary of the distributions using the following function to split the mage variable into the maturity groups, then take the summarize each using the summary function.

using the boxplot() method:

by plotting the boxplot for the mature mom, we get the minimum age as 35

boxplot(nc$mage[nc$mature == "younger mom"])



by plotting the boxplot for the younger mom, we get the minimum age as 13

#### Question 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.

Answer 5)

numerical variable:

weeks | length of pregnancy in weeks.

categorical variable:

habit | status of the mother as a nonsmoker or a smoker.

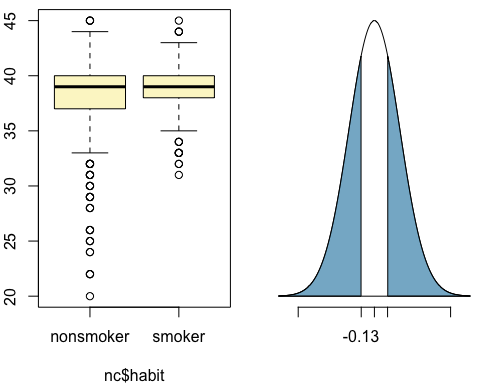
Research Question:

Conduct a hypothesis test evaluating how length of pregnency in weeks depend on the status of the mother as a smoker or nonsmoker.

inference(y = nc$weeks, x = nc$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 = 872, mean\_nonsmoker = 38.3188, sd\_nonsmoker = 2.9936  
## n\_smoker = 126, mean\_smoker = 38.4444, sd\_smoker = 2.4676

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



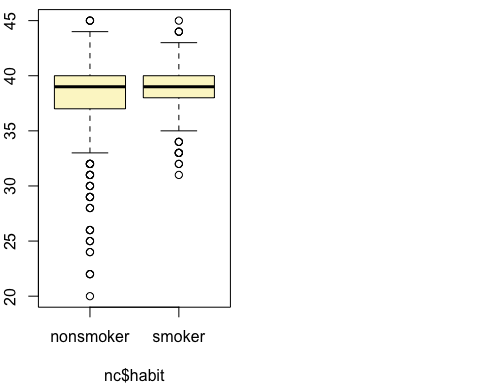
H0: µ{mu\_nonsmoker} - µ{mu\_smoker} = 0, There is no difference in the mean of the length of pregnancy in weeks between a smoker and nonsmoker mothers

HA: µ{mu\_nonsmoker} - µ{mu\_smoker} != 0, There is a difference in the mean of the length of pregnancy in weeks between a smoker and nonsmoker mothers

Calculate a 95% confidence interval how length of pregnency in weeks depend on the status of the mother as a smoker or nonsmoker.

inference(y = nc$weeks, x = nc$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 = 872, mean\_nonsmoker = 38.3188, sd\_nonsmoker = 2.9936  
## n\_smoker = 126, mean\_smoker = 38.4444, sd\_smoker = 2.4676



## Observed difference between means (nonsmoker-smoker) = -0.1256  
##   
## Standard error = 0.2421   
## 95 % Confidence interval = ( -0.6001 , 0.3488 )

Conclusion

Given a p-value of 0.6038 , we affirm the null hypothesis and fail to reject the null hypothesis determine that there is no significant difference between mean of length of pregnancy in weeks between a smoker and nonsmoker mothers.