Lin-Lab5

Bin Lin

2016-10-29

```
library(IS606)
```

```
##
## Welcome to CUNY IS606 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='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

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

```
library(ggplot2)
library(inference)
```

```
## Loading required package: sandwich
```

```
#startLab('Lab5')
#setwd('C:/Users/blin261/Documents/Lab5')
#install.packages("inference")
load("nc.RData")
```

Exercise 1: What are the cases in this data set? How many cases are there in our sample?

The birth information of babies and their parents recorded in North Carolina. Total 1000 cases are in the sample.

```
head(nc)
```

```
##   fage mage      mature weeks    premie visits marital gained weight
## 1   NA   13 younger mom    39 full term     10 married     38   7.63
## 2   NA   14 younger mom    42 full term     15 married     20   7.88
## 3   19   15 younger mom    37 full term     11 married     38   6.63
## 4   21   15 younger mom    41 full term      6 married     34   8.00
## 5   NA   15 younger mom    39 full term      9 married     27   6.38
## 6   NA   15 younger mom    38 full term     19 married     22   5.38
##   lowbirthweight gender     habit  whitemom
## 1        not low   male nonsmoker not white
## 2        not low   male nonsmoker not white
## 3        not low female nonsmoker     white
## 4        not low   male nonsmoker     white
## 5        not low female nonsmoker not white
## 6            low   male nonsmoker not white
```

```
str(nc)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ fage          : int  NA NA 19 21 NA NA 18 17 NA 20 ...
##  $ mage          : int  13 14 15 15 15 15 15 15 16 16 ...
##  $ mature        : Factor w/ 2 levels "mature mom","younger mom": 2 2 2 2 2 2 2 2 2 2 ...
##  $ weeks         : int  39 42 37 41 39 38 37 35 38 37 ...
##  $ premie        : Factor w/ 2 levels "full term","premie": 1 1 1 1 1 1 1 2 1 1 ...
##  $ visits        : int  10 15 11 6 9 19 12 5 9 13 ...
##  $ marital       : Factor w/ 2 levels "married","not married": 1 1 1 1 1 1 1 1 1 1 ...
##  $ gained        : int  38 20 38 34 27 22 76 15 NA 52 ...
##  $ weight        : num  7.63 7.88 6.63 8 6.38 5.38 8.44 4.69 8.81 6.94 ...
##  $ lowbirthweight: Factor w/ 2 levels "low","not low": 2 2 2 2 2 1 2 1 2 2 ...
##  $ gender        : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ habit         : Factor w/ 2 levels "nonsmoker","smoker": 1 1 1 1 1 1 1 1 1 1 ...
##  $ whitemom      : Factor w/ 2 levels "not white","white": 1 1 2 2 1 1 1 1 2 2 ...
```

```
summary(nc)
```

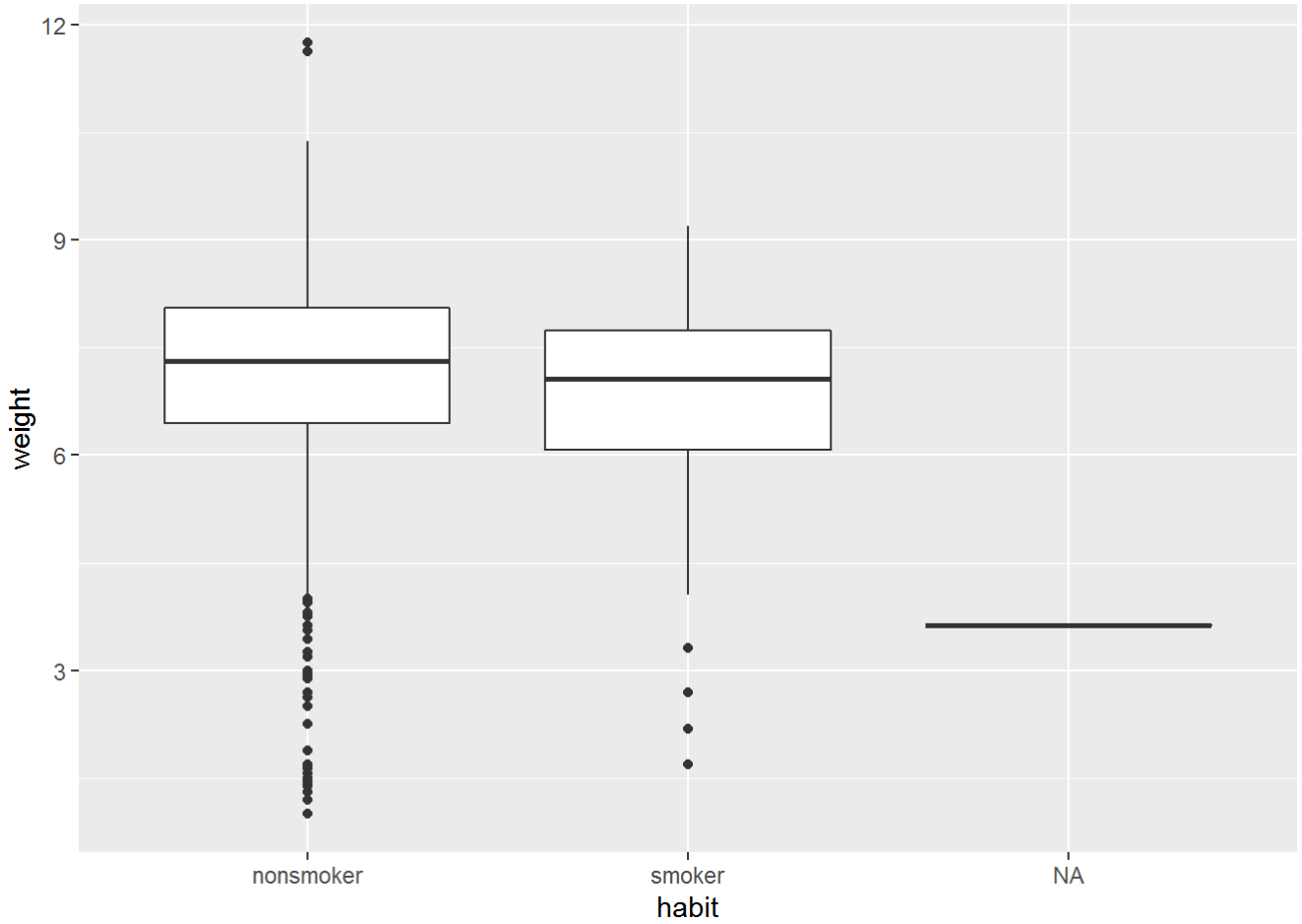
```
##       fage            mage            mature        weeks      
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00  
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00  
##  Median :30.00   Median :27                     Median :39.00  
##  Mean   :30.26   Mean   :27                     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  
##                 
##                 
##                 
##
```

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

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

Exercise 2: Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables?

```
ggplot(data = nc, aes(x = habit, y = weight)) + geom_boxplot()
```



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.

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

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

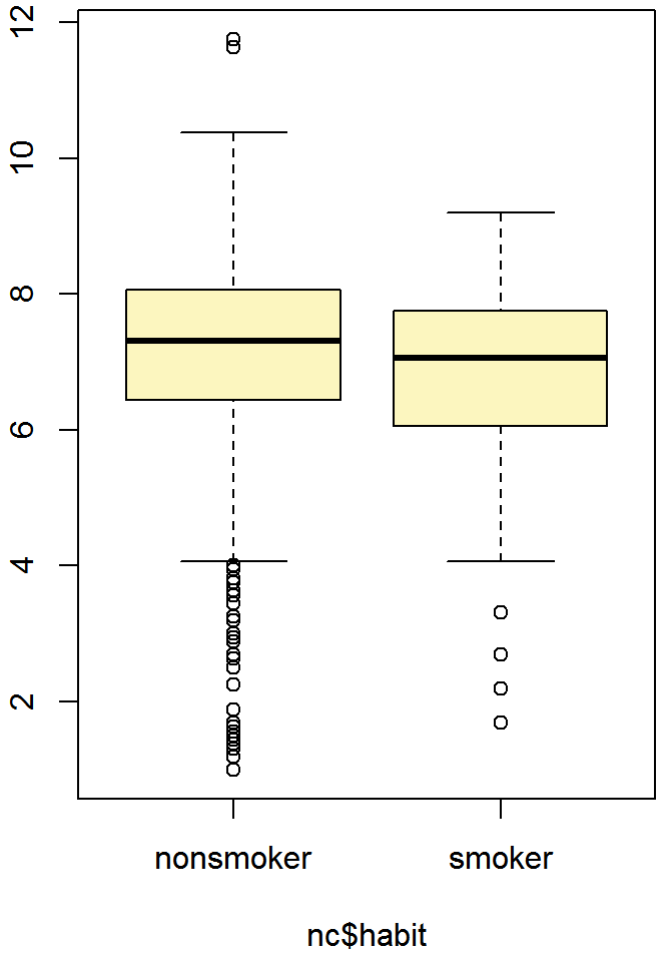
Exercise 4: Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

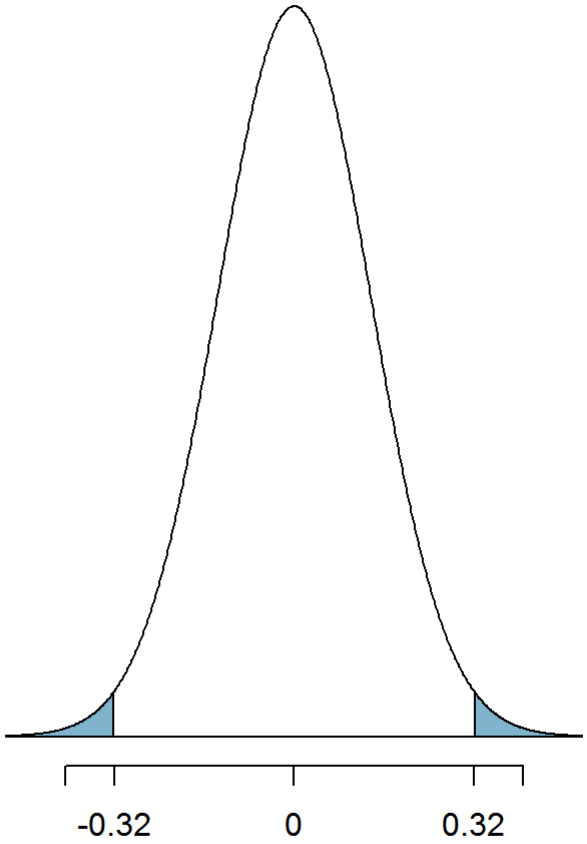
H0: Baby's Weight (smoking mother) = Baby's Weight (non-smoking mother) HA: Baby's Weight (smoking mother) != Baby's Weight (non-smoking mother)

```
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
```



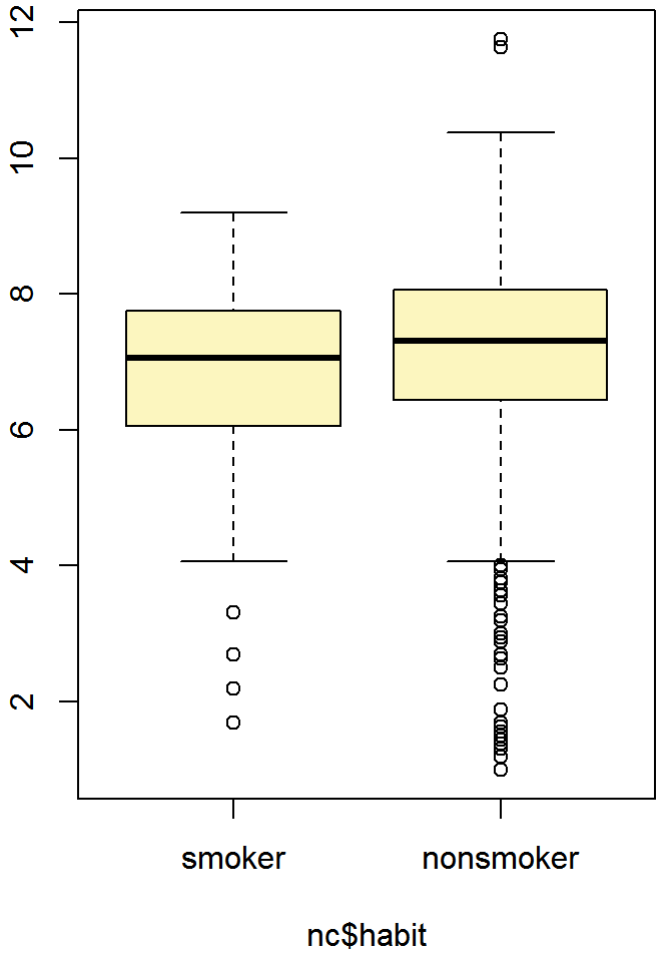


Exercise 5: Change the type argument to “ci” to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

95 % Confidence interval = (-0.5777 , -0.0534)

```
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
```



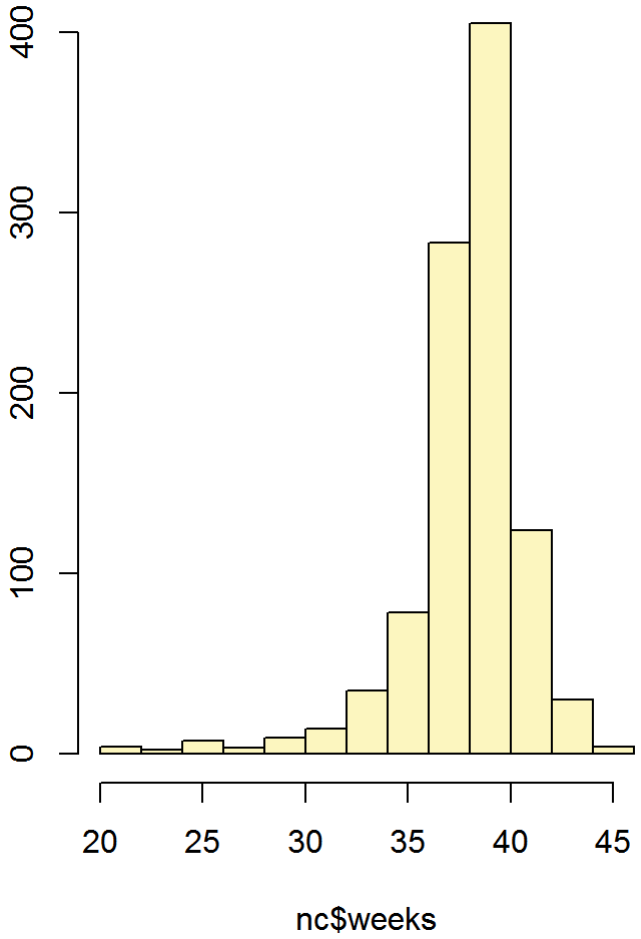

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

On your own 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.

95 % Confidence interval = (38.1528 , 38.5165), which means there is 95% chance that this interval is going to catch the true population mean.

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

```
## Single mean
## Summary statistics:
```



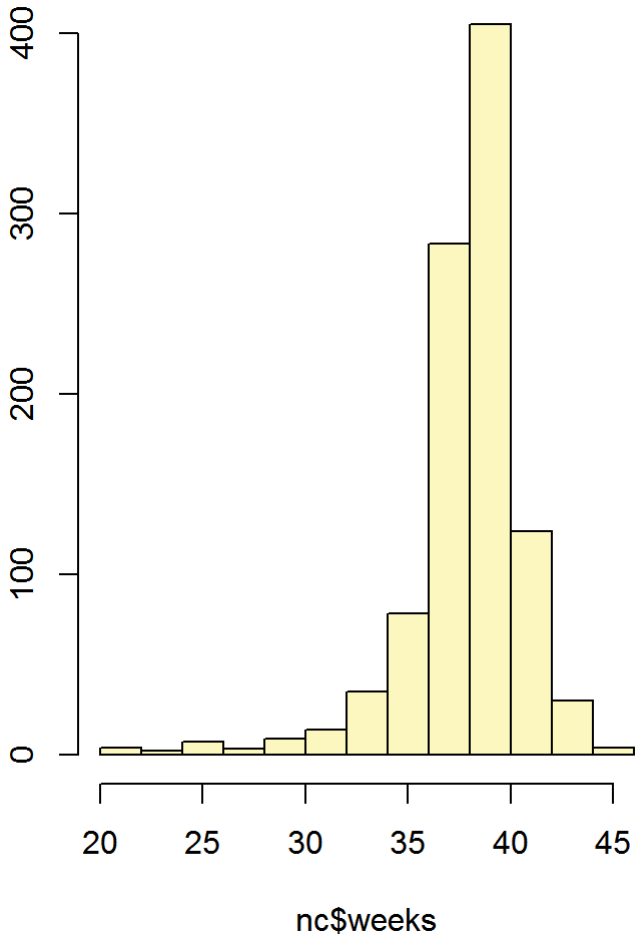
```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

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.

90 % Confidence interval = (38.182 , 38.4873)

```
inference(y = nc$weeks, est = "mean", type = "ci", conflevel = 0.9, null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## 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.

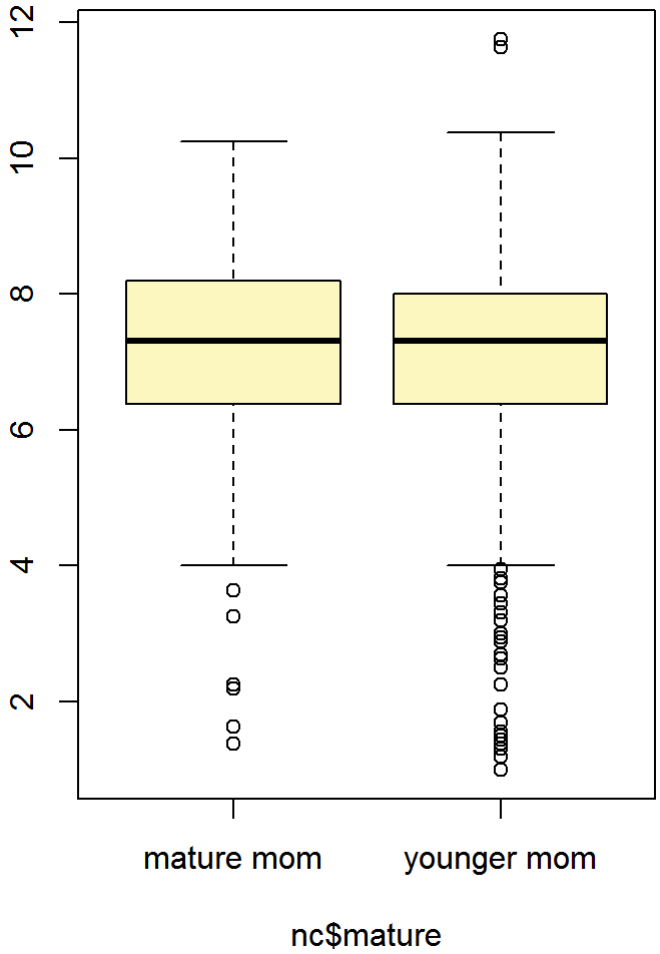
H0: mu_mature mom - mu_younger mom = 0 HA: mu_mature mom - mu_younger mom != 0 Standard error = 0.152 Test statistic: Z = 0.186 p-value = 0.8526

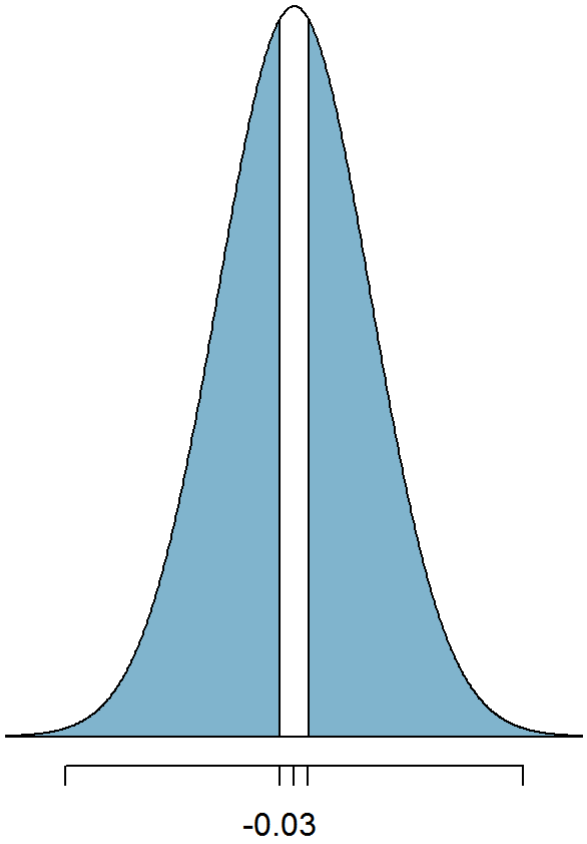
Since the p-value is greater than 0.05, we fail to reject the null hypothesis, which means the average weight gained by younger mothers could be same as the average weight gained by mature mothers.

```
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
```





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.

I think the cut off for younger mothers is 34(inclusive) and the cut off for the mature mothers is 35(inclusive). The method I use is just to determine the mother age above which is considered to be mature and under which is considered young.

max(nc$mage[$nc$mature == “younger mom”]) min(nc$mage[$nc$mature == “mature mom”])

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.

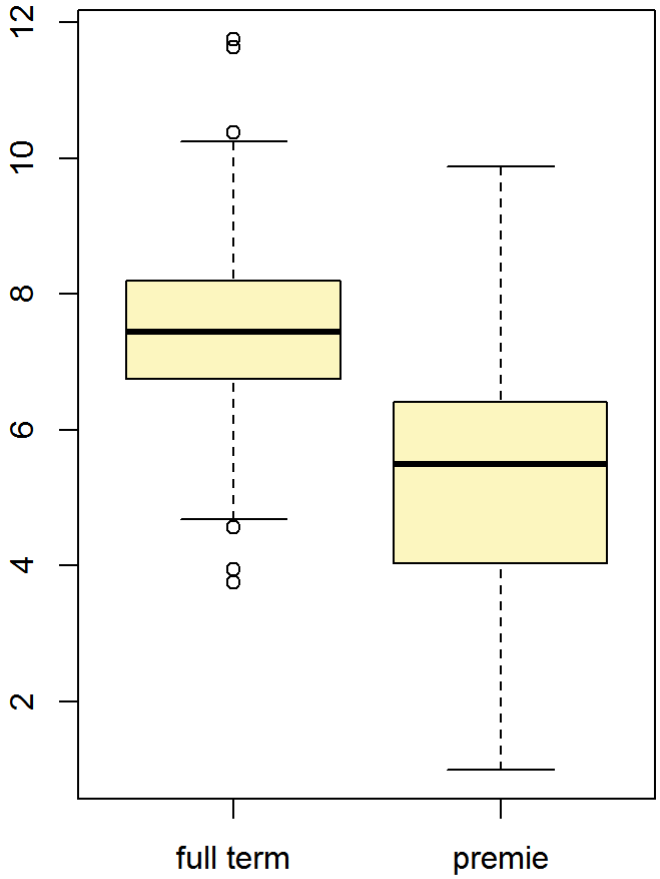
I will pick premie and weight as my variables for the hypothetical testing.

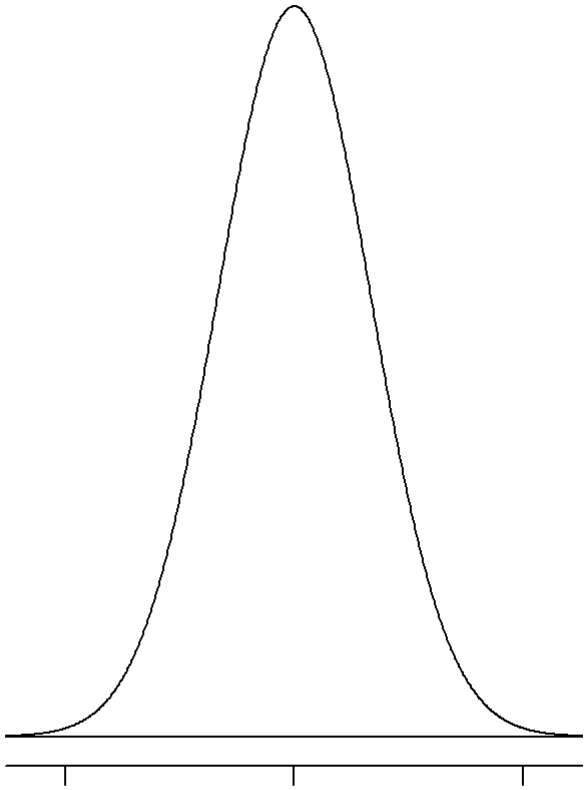
H0: Premature babies and full term babies have same average birth weight HA: Premature babies and full term babies have different average birth weight.

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_full term = 846, mean_full term = 7.4594, sd_full term = 1.075
## n_premie = 152, mean_premie = 5.1284, sd_premie = 1.9696
```

```
## Observed difference between means (full term-premie) = 2.331
##
## H0: mu_full term - mu_premie = 0
## HA: mu_full term - mu_premie != 0
## Standard error = 0.164
## Test statistic: Z =  14.216
## p-value =  0
```





H0: mu_full term - mu_premie = 0 HA: mu_full term - mu_premie != 0 Standard error = 0.164 Test statistic: Z = 14.216 p-value = 0 Because p-value is very closed to 0, which means the difference of average birth weight between ppremature babies and full term babies are very statistical significant at any alpha levels. So that the hypothesis test reject the null hypothesis, therefore the alternative hypothesis is true.