Inference for numerical data

Student Name : Sachid Deshmukh

Date : 11/16/2018

- GitHub Location for rmd file
- GitHub Location for pdf file
- RPubs location of published file

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.

```
load("more/nc.RData")
```

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

variable	description
`fage`	father's age in years.
`mage`	mother's age in years.
`mature`	maturity status of mother.
`weeks`	length of pregnancy in weeks.
`premie`	whether the birth was classified as premature (premie) or full-term.
`visits`	number of hospital visits during pregnancy.

variable	description
`marital`	whether mother is `married` or `not married` at birth.
`gained`	weight gained by mother during pregnancy in pounds.
`weight`	weight of the baby at birth in pounds.
`lowbirthweight`	whether baby was classified as low birthweight (`low`) or not (`not low`).
`gender`	gender of the baby, `female` or `male`.
`habit`	status of the mother as a `nonsmoker` or a `smoker`.
`whitemom`	whether mom is `white` or `not white`.

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

```
dim(nc)
```

```
## [1] 1000   13
```

Answer : Cases are birth records in the state of North Carolina in year 2004. There are total 1000 cases in our sample

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  
```

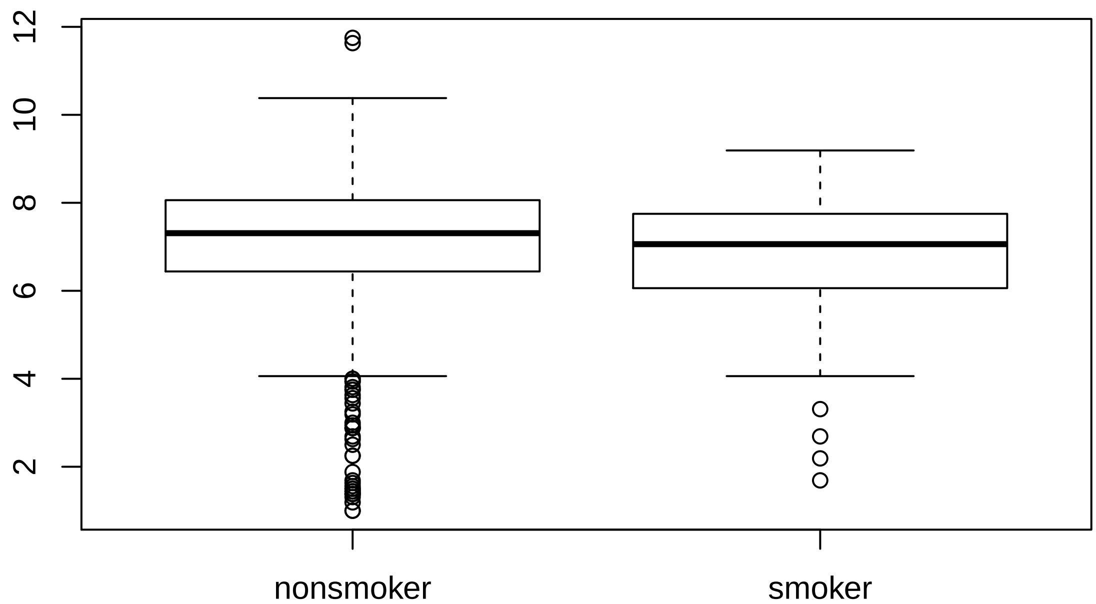
```
##                  Max.   :30.0                       Max.   :85.00
##                  NA's   :9                          NA's   :27
##       weight       lowbirthweight    gender          habit
##  Min.   : 1.000   low    :111    female:503   nonsmoker:873
##  1st Qu.: 6.380   not low:889    male  :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.

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

```r
boxplot(weight~habit, data=nc)
```



Answer : Plot highlights that non smoker mothers have high rate of healthy babies (babies with above normal weight) compared to mothers who smokes. We can see that median weight for babies with smoker mothers is lesser than average weight of the babies for non smoker mothers. We can also see that there are more observations beyond third quartile for non smoker monthers. This indicates that non smoker monthers have higher chance to give birth to above normal weight babies compared to smoker mothers

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
```

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test .

Inference

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`.

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

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

Answer: Sample sizes are good enough. We can proceed with the hypothesis testing.

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

Answer - We need to write null hypothesis and alternate hypothesis. We can write it as below

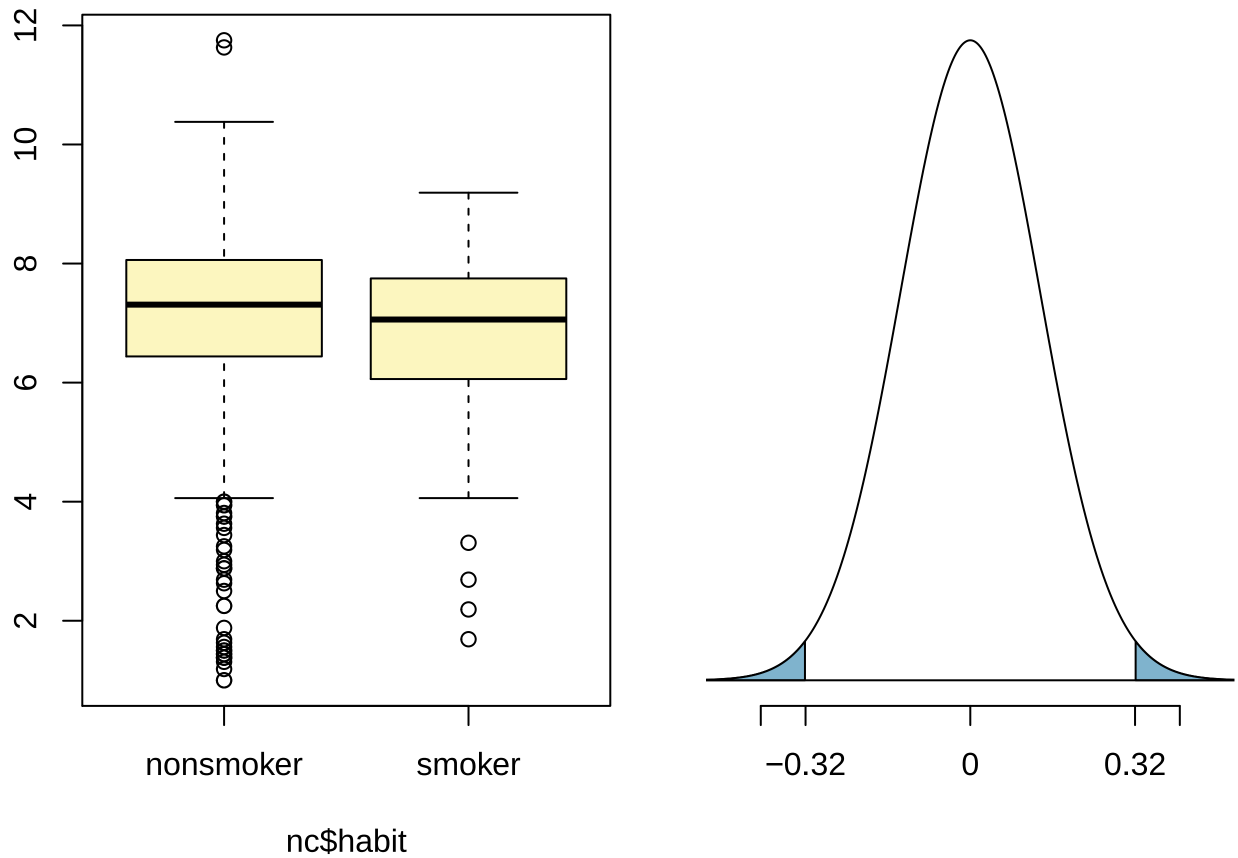
- NULL Hypothesis - Mean weight for babies born to smoking and non smoking mothers is same
- Alternate Hypotiesis - Mean weight for babies born to smoking and non smoking mothers is different

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

```r
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
```

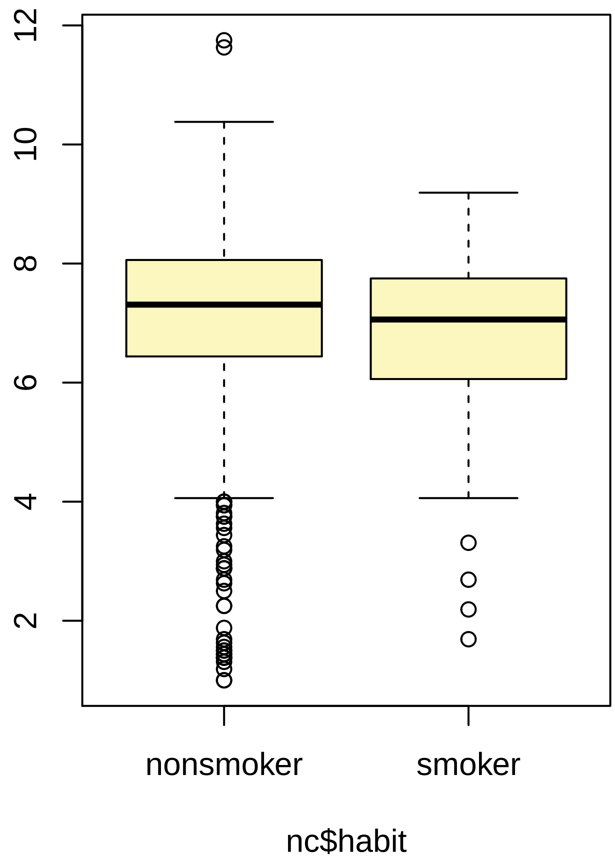


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and non-smokers: `nc$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.

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
```

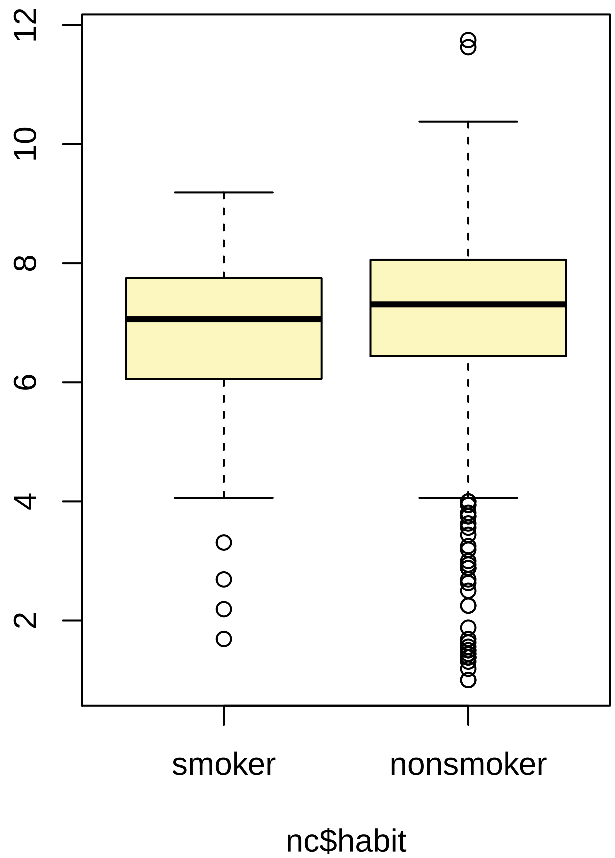



```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

By default the function reports an interval for $(\mu_{nonsmoker} - \mu_{smoker})$. 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
```



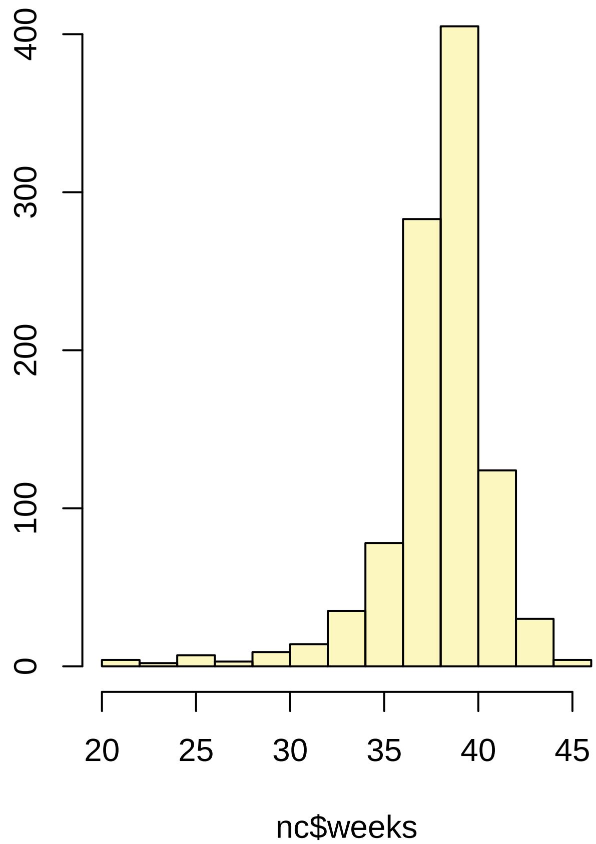

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

On your own

- 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", null = 0,
          alternative = "twosided", method = "theoretical")
```

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



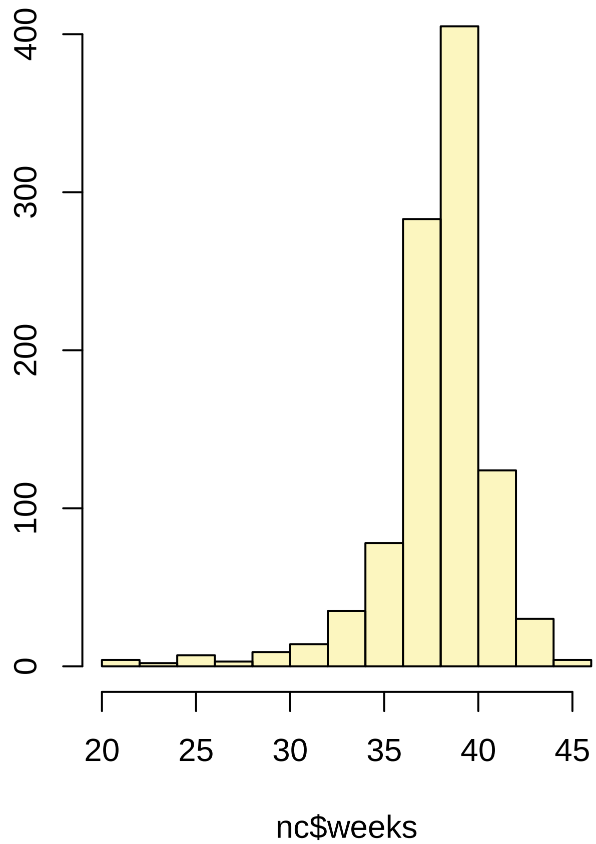

```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

Answer : 95% confidence interval for avg length of pregnancies in weeks falls between 38.15 weeks to 38.51 weeks

- 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", null = 0,
          alternative = "twosided", 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 )
```

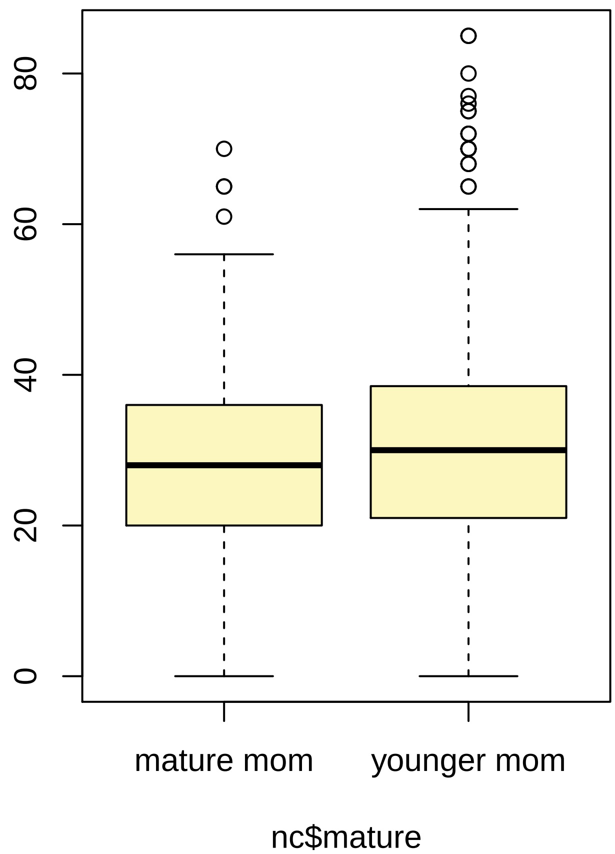
Answer : 90% confidence interval for avg length of pregnancies in weeks falls between 38.18 weeks to 38.48 weeks

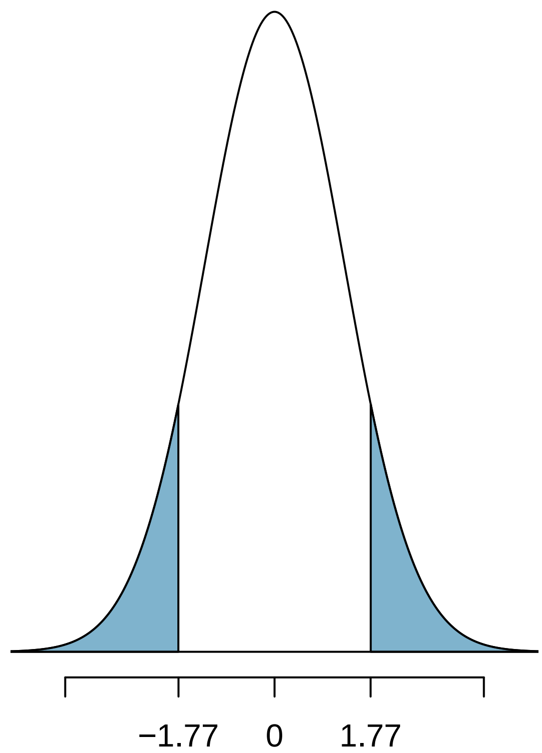
- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```r
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
```

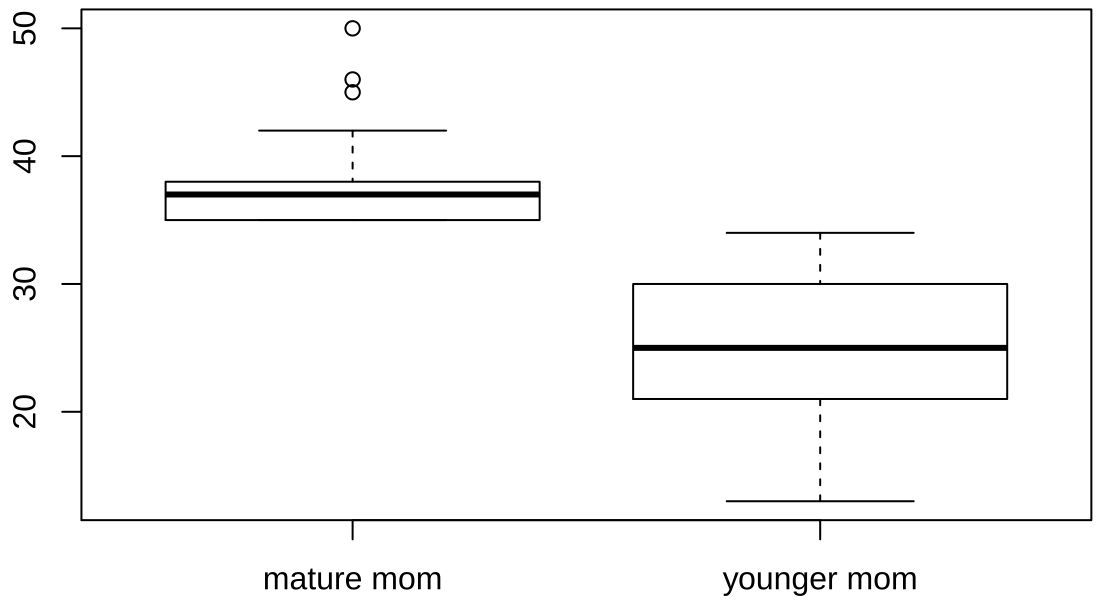




Answer : The Z stats and p-value indicates that there is no difference in mean weight gain by younger mothers and mature mothers

- 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.

```
boxplot(mage~mature, data=nc)
```



Answer : Side by side boxplot for mothers age based on their maturity provides us a min and max range for all the mothers in respective category. We can see from the box plot above that younger mom falls under the age range of 16 to 35 and mature moms age ranges from 35 to 42. From the box plot above we can say that 35 is the age cutoff for younger moms and mature moms

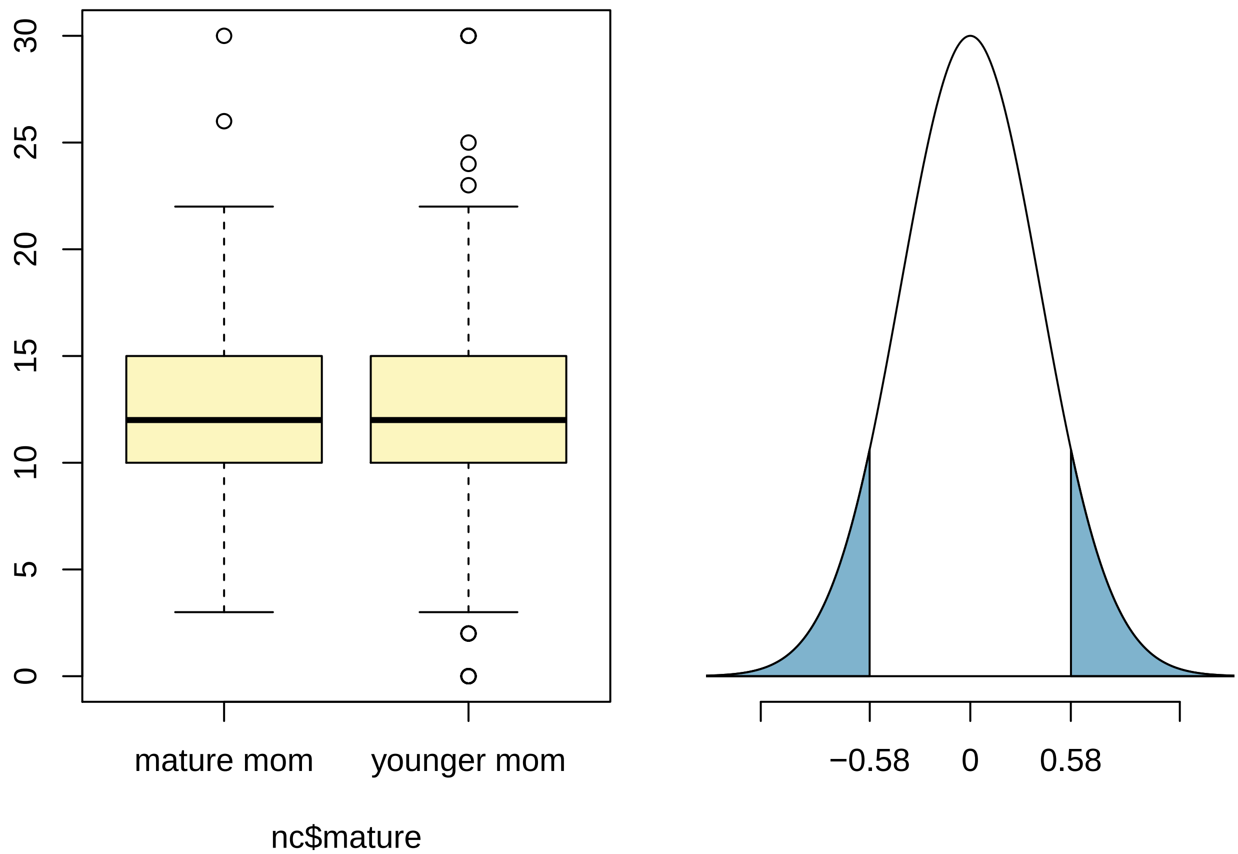
- Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the `inference` function,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

Question - Check if there is difference in the mean avg visits by young mothers and mature mothers

```r
inference(y = nc$visits, x = nc$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 = 131, mean_mature mom = 12.6107, sd_mature mom = 4.3793
## n_younger mom = 860, mean_younger mom = 12.0279, sd_younger mom = 3.8832

## Observed difference between means (mature mom-younger mom) = 0.5828
##
## H0: mu_mature mom - mu_younger mom = 0
## HA: mu_mature mom - mu_younger mom != 0
## Standard error = 0.405
## Test statistic: Z =  1.439
## p-value =  0.15
```



Answer : From the Z statistics and p-value we can conclude that there is no difference in the mean visits to the hospital by younger mom and matuare mom. We can conclude that both mature mom and younger mom visits hospital with same frequency during their pregnancy