Inference for numerical data

Jose Mawyin

10/27/2019

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.
`marital`	whether mother is `married` or `not married` at birth.
`gained`	weight gained by mother during pregnancy in pounds.

variable	description
`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?

```r
dim(nc)
```

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

The cases in this data set are the 1000 births from which the 13 variables above were collected.

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

```r
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.

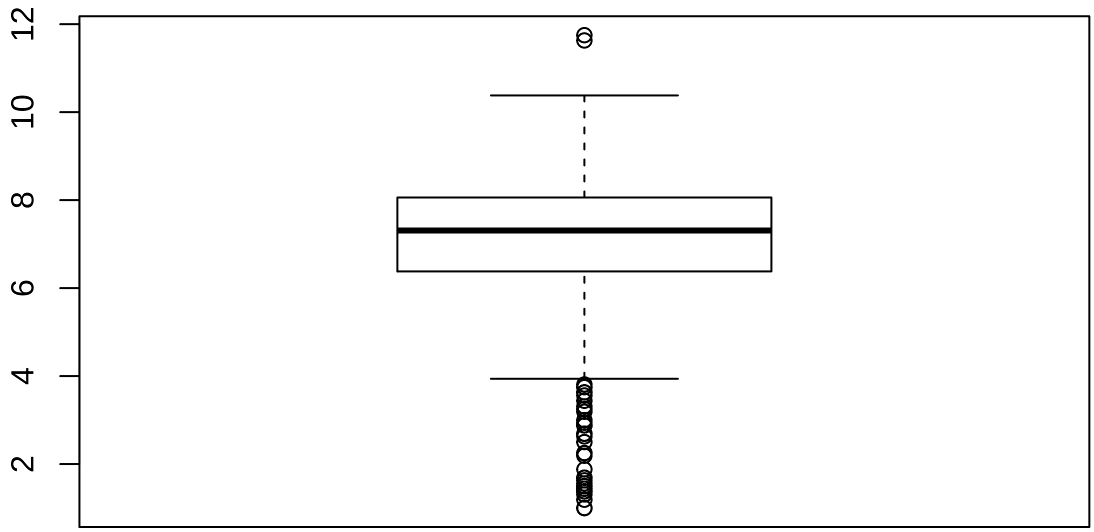
Consider the possible relationship between a mother's smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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

```r
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 ...
```

```r
boxplot(nc$weight)
```



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`.

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

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

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

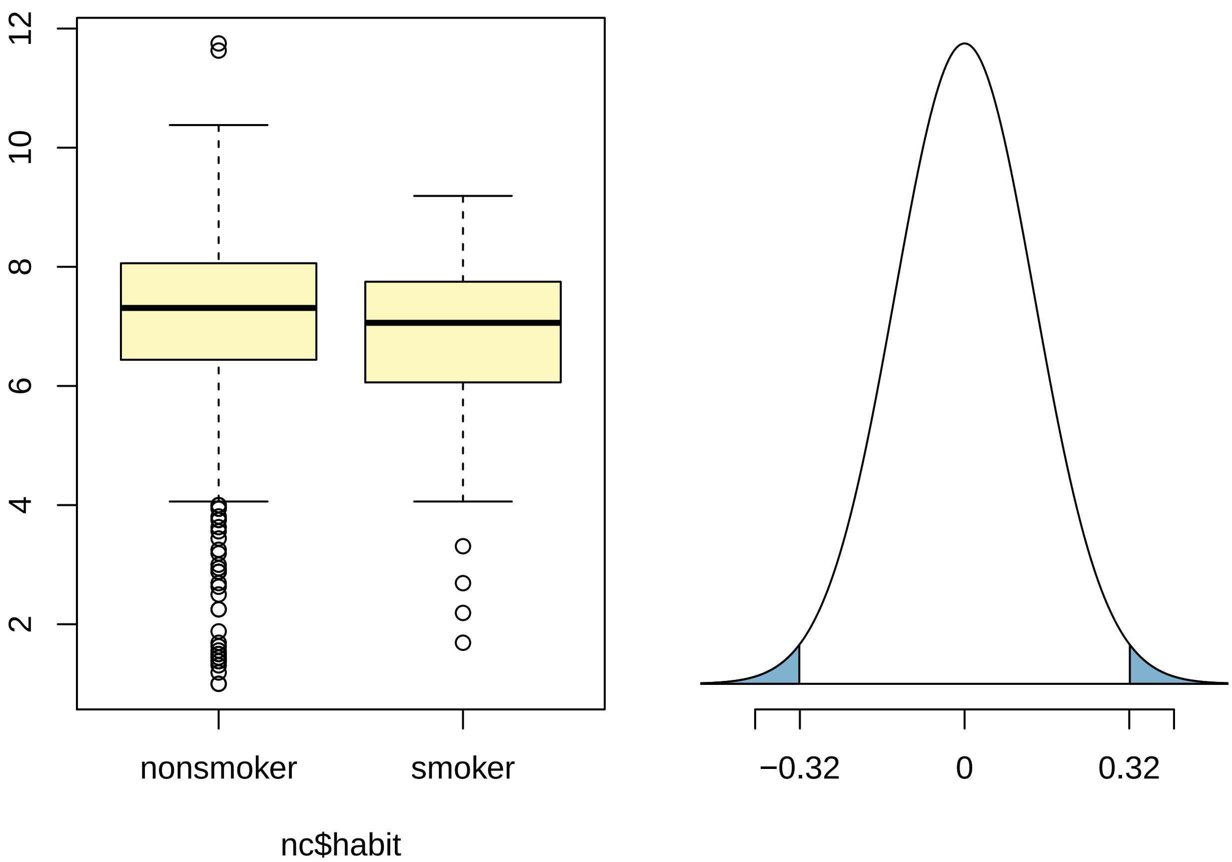
The Null Hypotheses is that mean weight of the smoking and non-smoking mothers is the same.

The Alternative Hypotheses is that the mean weight of the babies is influenced by the mother's smoking.

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
```



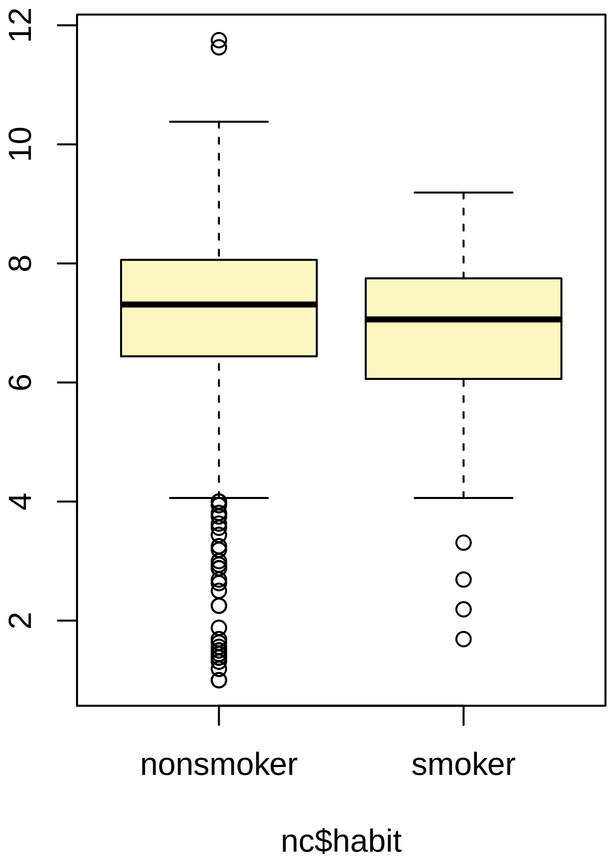
Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

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",
          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
```

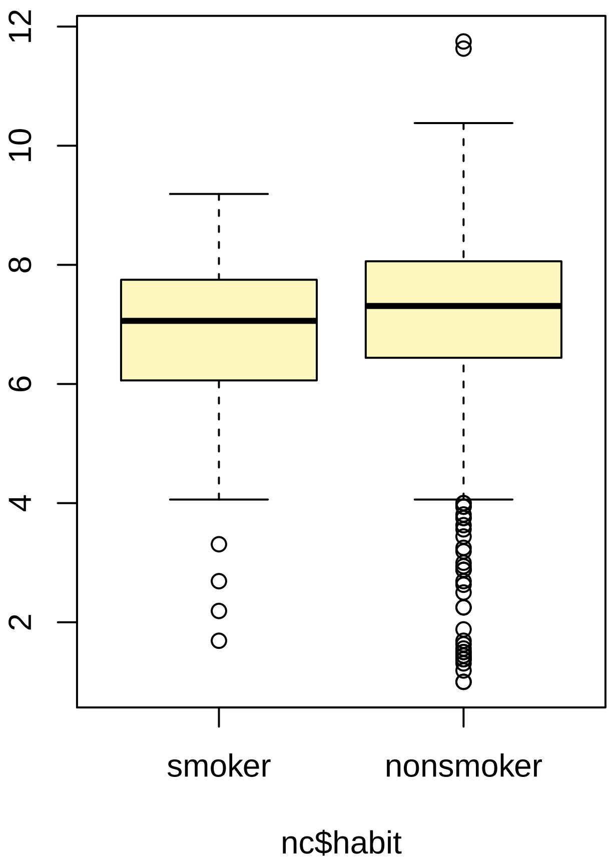



```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

By default the function reports an interval for $(\mu_{nonsmoker} - \mu_{smoker})$. We can easily change this order by using the `order` argument:

```r
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
```



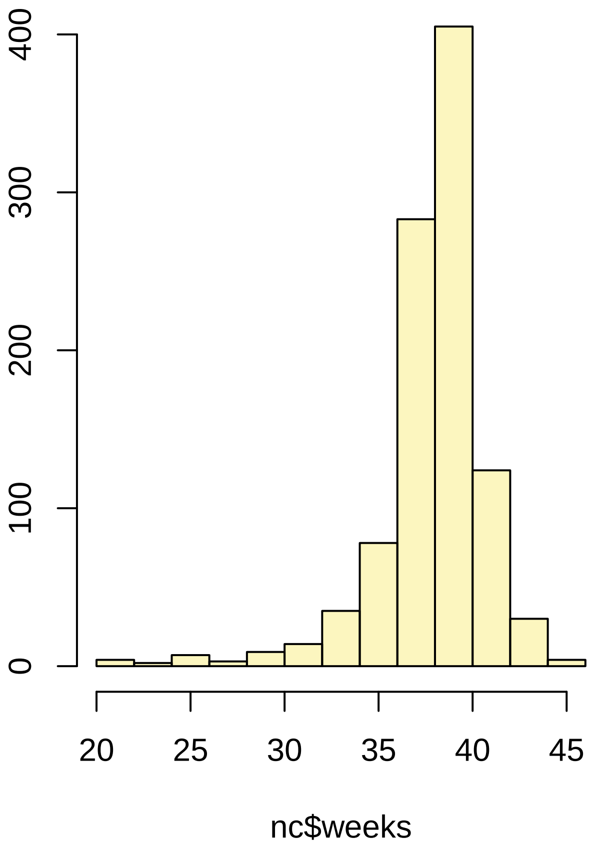

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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.

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

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



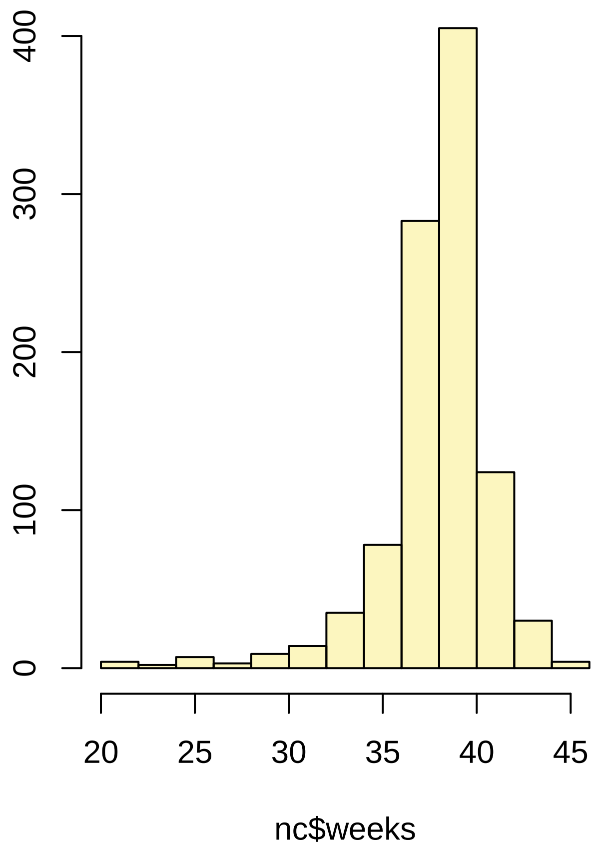

```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

The average length of pregnancy is 38.3347 weeks. This length has a low spread at 95 % Confidence interval of (38.1528 , 38.5165)

- 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`.

```r
inference(y = nc$weeks, est = "mean", type = "ci",
          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 )
```

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

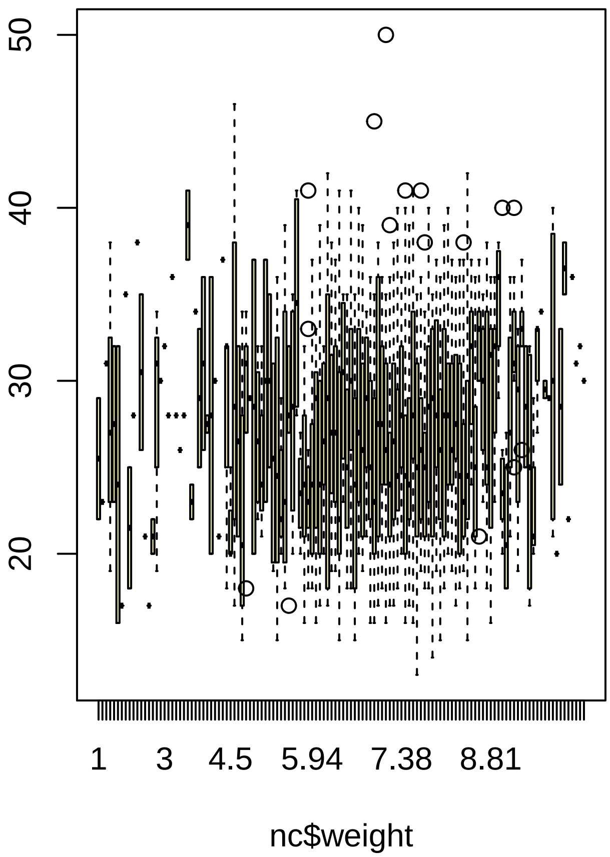
```
inference(y = nc$mage, x = nc$weight, est = "mean", type = "ht", null = 0,
          alternative = "greater", method = "theoretical")
```

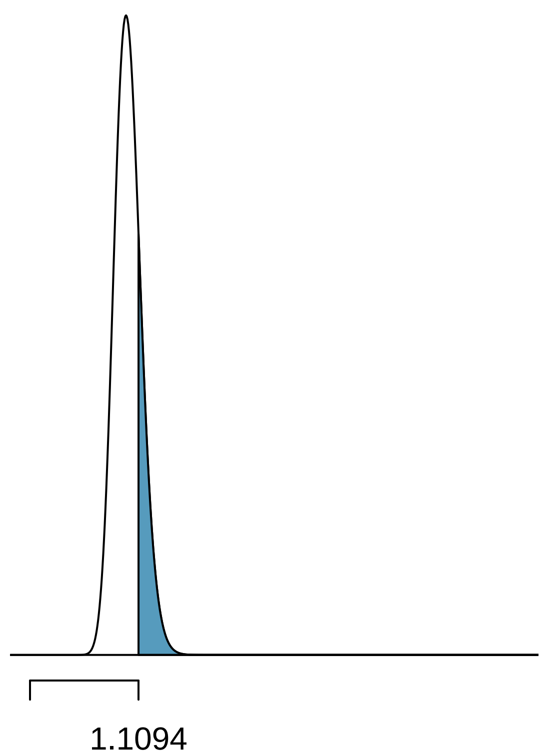
```
## Warning: Explanatory variable was numerical, it has been converted to
## categorical. In order to avoid this warning, first convert your explanatory
## variable to a categorical variable using the as.factor() function.

## Response variable: numerical, Explanatory variable: categorical

## Warning: Ignoring null value since it's undefined for ANOVA.

## ANOVA
##
## Summary statistics:
## n_1 = 2, mean_1 = 25.5, sd_1 = 4.9497
## n_1.19 = 1, mean_1.19 = 23, sd_1.19 = NA
## n_1.31 = 1, mean_1.31 = 31, sd_1.31 = NA
## n_1.38 = 3, mean_1.38 = 28, sd_1.38 = 9.5394
```

```
## n_1.44 = 2, mean_1.44 = 27.5, sd_1.44 = 6.364
## n_1.5 = 2, mean_1.5 = 24, sd_1.5 = 11.3137
## n_1.56 = 1, mean_1.56 = 17, sd_1.56 = NA
## n_1.63 = 1, mean_1.63 = 35, sd_1.63 = NA
## n_1.69 = 2, mean_1.69 = 21.5, sd_1.69 = 4.9497
## n_1.88 = 1, mean_1.88 = 28, sd_1.88 = NA
## n_2.19 = 1, mean_2.19 = 38, sd_2.19 = NA
## n_2.25 = 2, mean_2.25 = 30.5, sd_2.25 = 6.364
## n_2.5 = 1, mean_2.5 = 21, sd_2.5 = NA
## n_2.63 = 1, mean_2.63 = 17, sd_2.63 = NA
## n_2.69 = 2, mean_2.69 = 21, sd_2.69 = 1.4142
## n_2.88 = 3, mean_2.88 = 28, sd_2.88 = 7.9373
## n_2.94 = 1, mean_2.94 = 30, sd_2.94 = NA
## n_3 = 1, mean_3 = 32, sd_3 = NA
## n_3.19 = 1, mean_3.19 = 28, sd_3.19 = NA
## n_3.25 = 1, mean_3.25 = 36, sd_3.25 = NA
## n_3.31 = 1, mean_3.31 = 28, sd_3.31 = NA
## n_3.44 = 1, mean_3.44 = 26, sd_3.44 = NA
## n_3.56 = 1, mean_3.56 = 28, sd_3.56 = NA
## n_3.63 = 2, mean_3.63 = 39, sd_3.63 = 2.8284
## n_3.75 = 2, mean_3.75 = 23, sd_3.75 = 1.4142
## n_3.81 = 1, mean_3.81 = 34, sd_3.81 = NA
## n_3.94 = 2, mean_3.94 = 29, sd_3.94 = 5.6569
## n_4 = 2, mean_4 = 31, sd_4 = 7.0711
## n_4.06 = 2, mean_4.06 = 27.5, sd_4.06 = 0.7071
## n_4.13 = 2, mean_4.13 = 28, sd_4.13 = 11.3137
## n_4.19 = 1, mean_4.19 = 30, sd_4.19 = NA
## n_4.25 = 1, mean_4.25 = 21, sd_4.25 = NA
## n_4.31 = 1, mean_4.31 = 37, sd_4.31 = NA
## n_4.44 = 3, mean_4.44 = 27.3333, sd_4.44 = 8.0829
## n_4.5 = 3, mean_4.5 = 21.6667, sd_4.5 = 2.8868
## n_4.56 = 4, mean_4.56 = 30, sd_4.56 = 12.0277
## n_4.63 = 2, mean_4.63 = 26.5, sd_4.63 = 7.7782
## n_4.69 = 4, mean_4.69 = 22.5, sd_4.69 = 8.1854
## n_4.75 = 6, mean_4.75 = 28.8333, sd_4.75 = 5.7764
## n_4.88 = 1, mean_4.88 = 29, sd_4.88 = NA
## n_4.94 = 2, mean_4.94 = 28.5, sd_4.94 = 12.0208
## n_5 = 4, mean_5 = 26.75, sd_5 = 4.5735
## n_5.06 = 3, mean_5.06 = 25.6667, sd_5.06 = 5.6862
## n_5.13 = 2, mean_5.13 = 30, sd_5.13 = 9.8995
## n_5.19 = 2, mean_5.19 = 30, sd_5.19 = 7.0711
## n_5.25 = 4, mean_5.25 = 25.25, sd_5.25 = 6.6521
## n_5.38 = 8, mean_5.38 = 25.5, sd_5.38 = 7.7644
## n_5.44 = 7, mean_5.44 = 23.5714, sd_5.44 = 3.8235
## n_5.5 = 7, mean_5.5 = 26.7143, sd_5.5 = 9.0132
## n_5.56 = 5, mean_5.56 = 27.2, sd_5.56 = 6.14
## n_5.63 = 8, mean_5.63 = 28.125, sd_5.63 = 6.1514
## n_5.69 = 4, mean_5.69 = 34.5, sd_5.69 = 6.9522
## n_5.75 = 4, mean_5.75 = 23.5, sd_5.75 = 2.8868
## n_5.81 = 9, mean_5.81 = 24, sd_5.81 = 5.3619
## n_5.88 = 12, mean_5.88 = 24.5833, sd_5.88 = 6.4025
## n_5.94 = 15, mean_5.94 = 24.3333, sd_5.94 = 5.5506
## n_6 = 15, mean_6 = 25.7333, sd_6 = 5.8975
## n_6.06 = 10, mean_6.06 = 25.9, sd_6.06 = 7.5491
```

```
## n_6.13 = 6, mean_6.13 = 26.6667, sd_6.13 = 4.5461
## n_6.19 = 11, mean_6.19 = 27.2727, sd_6.19 = 9.4349
## n_6.25 = 16, mean_6.25 = 27.75, sd_6.25 = 5.7096
## n_6.31 = 15, mean_6.31 = 27.2, sd_6.31 = 5.7719
## n_6.38 = 16, mean_6.38 = 25.3125, sd_6.38 = 7.3186
## n_6.44 = 8, mean_6.44 = 29.875, sd_6.44 = 5.0832
## n_6.5 = 15, mean_6.5 = 25.4, sd_6.5 = 5.0822
## n_6.56 = 13, mean_6.56 = 29.2308, sd_6.56 = 5.5701
## n_6.63 = 9, mean_6.63 = 24.3333, sd_6.63 = 6.7082
## n_6.69 = 13, mean_6.69 = 28.2308, sd_6.69 = 6.7596
## n_6.75 = 23, mean_6.75 = 26.4348, sd_6.75 = 6.0289
## n_6.81 = 12, mean_6.81 = 28.5, sd_6.81 = 4.6221
## n_6.88 = 25, mean_6.88 = 26.28, sd_6.88 = 5.8132
## n_6.94 = 14, mean_6.94 = 25.4286, sd_6.94 = 7.9199
## n_7 = 22, mean_7 = 27.5, sd_7 = 7.3209
## n_7.06 = 20, mean_7.06 = 27.75, sd_7.06 = 5.7754
## n_7.13 = 24, mean_7.13 = 27.625, sd_7.13 = 6.7037
## n_7.19 = 21, mean_7.19 = 25.2381, sd_7.19 = 5.9237
## n_7.25 = 22, mean_7.25 = 26.6818, sd_7.25 = 6.2212
## n_7.31 = 24, mean_7.31 = 26, sd_7.31 = 5.6492
## n_7.38 = 21, mean_7.38 = 28.0476, sd_7.38 = 4.4663
## n_7.44 = 30, mean_7.44 = 24.8, sd_7.44 = 6.5779
## n_7.5 = 24, mean_7.5 = 25.6667, sd_7.5 = 5.8508
## n_7.56 = 19, mean_7.56 = 28.7368, sd_7.56 = 6.8786
## n_7.63 = 18, mean_7.63 = 25.8889, sd_7.63 = 6.2862
## n_7.69 = 17, mean_7.69 = 26.7647, sd_7.69 = 6.0055
## n_7.75 = 17, mean_7.75 = 25.3529, sd_7.75 = 5.1349
## n_7.81 = 20, mean_7.81 = 28.4, sd_7.81 = 6.2694
## n_7.88 = 25, mean_7.88 = 27.16, sd_7.88 = 6.6187
## n_7.94 = 16, mean_7.94 = 28.6875, sd_7.94 = 5.2627
## n_8 = 19, mean_8 = 25.8421, sd_8 = 5.5203
## n_8.06 = 13, mean_8.06 = 27.6154, sd_8.06 = 6.4748
## n_8.13 = 17, mean_8.13 = 28.5882, sd_8.13 = 5.5233
## n_8.19 = 17, mean_8.19 = 27.3529, sd_8.19 = 5.3379
## n_8.25 = 16, mean_8.25 = 27.75, sd_8.25 = 5.2599
## n_8.31 = 12, mean_8.31 = 25.5, sd_8.31 = 6.4315
## n_8.38 = 20, mean_8.38 = 24.7, sd_8.38 = 5.4299
## n_8.44 = 14, mean_8.44 = 26.4286, sd_8.44 = 7.2718
## n_8.5 = 15, mean_8.5 = 30.7333, sd_8.5 = 4.6054
## n_8.56 = 11, mean_8.56 = 25.6364, sd_8.56 = 6.0708
## n_8.63 = 5, mean_8.63 = 31, sd_8.63 = 6.1237
## n_8.69 = 4, mean_8.69 = 29.5, sd_8.69 = 5
## n_8.75 = 14, mean_8.75 = 27.4286, sd_8.75 = 6.1858
## n_8.81 = 12, mean_8.81 = 28.25, sd_8.81 = 7.3747
## n_8.88 = 9, mean_8.88 = 30.1111, sd_8.88 = 4.4845
## n_8.94 = 4, mean_8.94 = 34.75, sd_8.94 = 4.0311
## n_9 = 8, mean_9 = 25.25, sd_9 = 6.2507
## n_9.06 = 6, mean_9.06 = 21.5, sd_9.06 = 3.7283
## n_9.13 = 7, mean_9.13 = 28.4286, sd_9.13 = 5.3807
## n_9.19 = 7, mean_9.19 = 32.1429, sd_9.19 = 4.7409
## n_9.25 = 6, mean_9.25 = 27.6667, sd_9.25 = 5.5737
## n_9.31 = 5, mean_9.31 = 32.4, sd_9.31 = 4.0373
## n_9.38 = 2, mean_9.38 = 28.5, sd_9.38 = 4.9497
## n_9.5 = 4, mean_9.5 = 24.75, sd_9.5 = 7.8475
```

```
## n_9.56 = 3, mean_9.56 = 23.3333, sd_9.56 = 4.9329
## n_9.63 = 3, mean_9.63 = 31, sd_9.63 = 3.4641
## n_9.69 = 1, mean_9.69 = 34, sd_9.69 = NA
## n_9.75 = 2, mean_9.75 = 29.5, sd_9.75 = 0.7071
## n_9.81 = 1, mean_9.81 = 29, sd_9.81 = NA
## n_9.88 = 4, mean_9.88 = 30.25, sd_9.88 = 9.6393
## n_9.94 = 1, mean_9.94 = 20, sd_9.94 = NA
## n_10.06 = 2, mean_10.06 = 28.5, sd_10.06 = 6.364
## n_10.13 = 2, mean_10.13 = 36.5, sd_10.13 = 2.1213
## n_10.19 = 1, mean_10.19 = 22, sd_10.19 = NA
## n_10.25 = 1, mean_10.25 = 36, sd_10.25 = NA
## n_10.38 = 1, mean_10.38 = 31, sd_10.38 = NA
## n_11.63 = 1, mean_11.63 = 32, sd_11.63 = NA
## n_11.75 = 1, mean_11.75 = 30, sd_11.75 = NA

## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value Pr(>F)
## x         125   5282  42.255  1.1094 0.2087
## Residuals 874  33288  38.087
```





- 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.

One approach if rather arbitrary is to find the mean age of mothers and arbitrarily use as mean + one standard deviation (33.2 years) as a cutoff between younger and mature mother. The resulting cutoff is not that far off from the one used (34 years) in the study as seen below.

```
mean.mother <- mean(nc$mage)
sd.mother <- sd(nc$mage)
mean.mother + sd.mother
```

```
## [1] 33.21358
```

```
table(nc$mage, nc$mature)
```

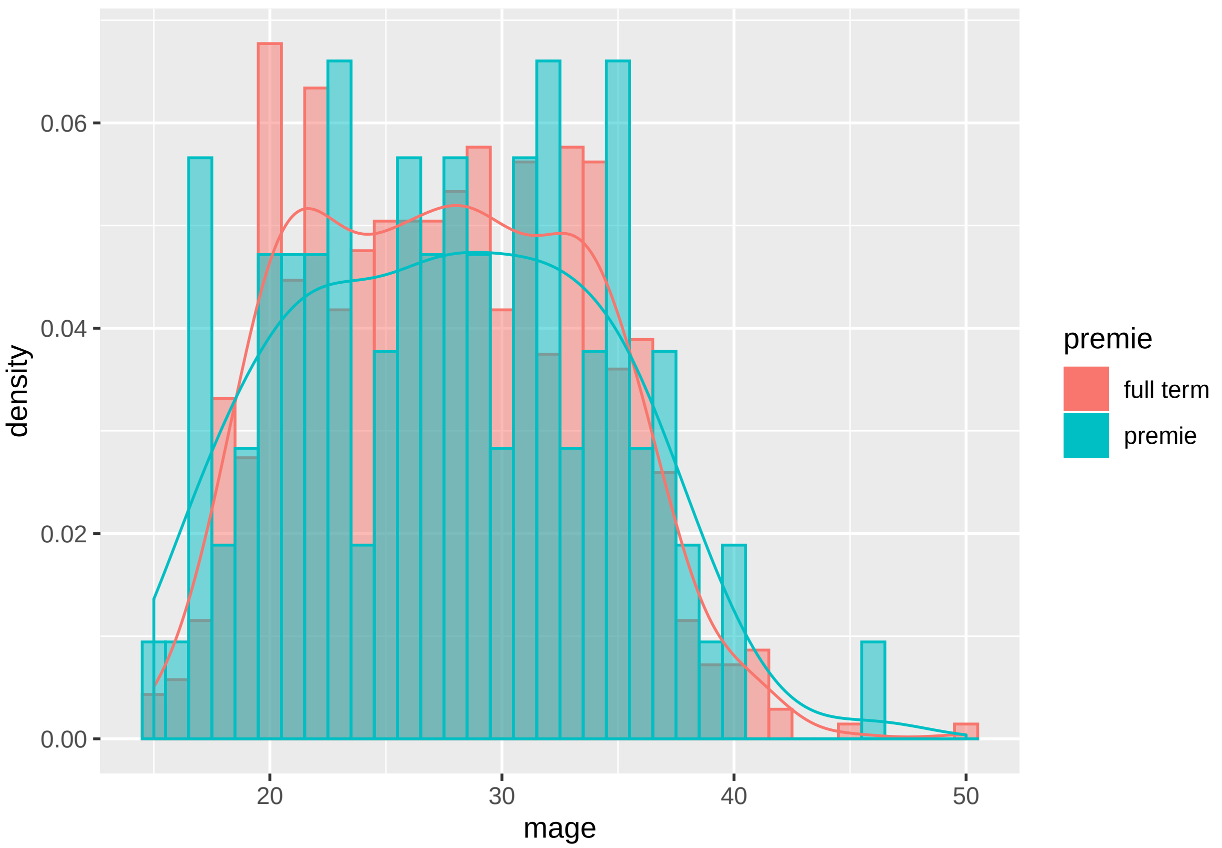
```
##
##      mature mom younger mom
##   13          0           1
##   14          0           1
##   15          0           6
##   16          0          10
##   17          0          19
##   18          0          38
##   19          0          35
##   20          0          66
##   21          0          51
##   22          0          60
##   23          0          51
##   24          0          53
##   25          0          54
##   26          0          51
##   27          0          47
##   28          0          53
##   29          0          52
##   30          0          39
##   31          0          52
##   32          0          38
##   33          0          45
##   34          0          45
##   35         35           0
##   36         31           0
##   37         26           0
##   38         12           0
##   39          7           0
##   40          9           0
##   41          8           0
##   42          2           0
##   45          1           0
##   46          1           0
##   50          1           0
```

I tried to find a correlation between women age and wether the pregnancy would be full term or not. As seen below there is no obvious correlation between mother age and full term or not.

```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

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

```r
ggplot(na.omit(nc), aes(mage, color=premie)) +
  geom_histogram(position="identity", binwidth=1, aes(y=..density.., fill=premie),  alpha=0.5) +
  geom_density()
```



- 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.

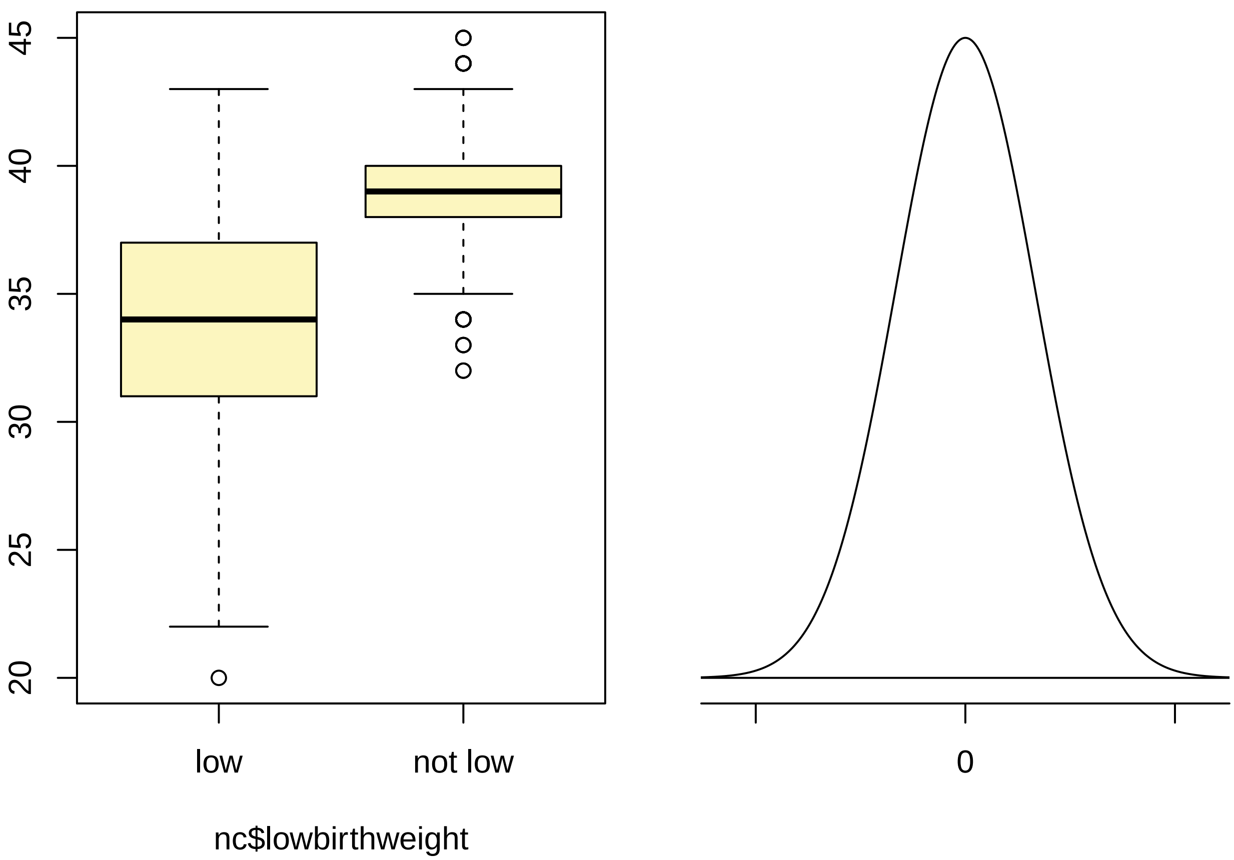
What is the correlation between the length of the pregnancy in weeks and the birth weight?

The Null Hypotheses is that mean length of pregnancy of the low birthweight and non low birthweight is the same.

The Alternative Hypotheses is that the mean length of pregnancy influences the birthweight of the baby.

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_low = 110, mean_low = 33.4273, sd_low = 4.6991
## n_not low = 888, mean_not low = 38.9426, sd_not low = 1.8947

## Observed difference between means (low-not low) = -5.5153
##
## H0: mu_low - mu_not low = 0
## HA: mu_low - mu_not low != 0
## Standard error = 0.453
## Test statistic: Z =  -12.188
## p-value =  0
```

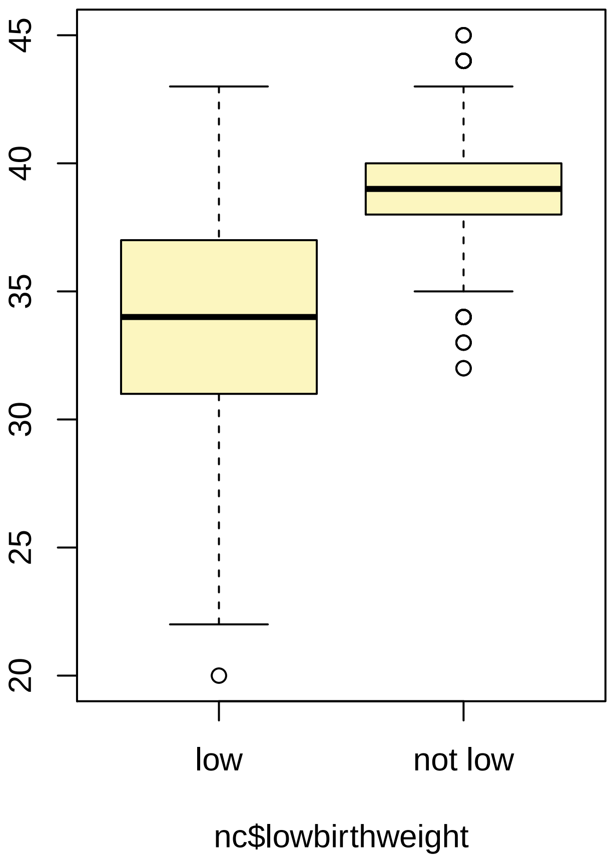


This one was easy. There is a clear correlation between birthweight and length of pregnancy as we see ont he hypotheses test above. Let's also do a 95 confidence interval.

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
```

```
## Summary statistics:
## n_low = 110, mean_low = 33.4273, sd_low = 4.6991
## n_not low = 888, mean_not low = 38.9426, sd_not low = 1.8947
```




```
## Observed difference between means (low-not low) = -5.5153
##
## Standard error = 0.4525
## 95 % Confidence interval = ( -6.4022 , -4.6283 )
```

The standard error (0.4525 weeeks)is much less than the observed difference (5.5153 weeks) between length of pregnancy of low and non-low birthweight.