amLab7

Pradeep Paladugula

4/7/2020

Inference for numerical data

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.

```
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"
```

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

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

```
dim(nc)
```

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

```
summary(nc)
```

```
##       fage            mage            mature        weeks
premie
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00   full
term:846
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00   premie
:152
##  Median :30.00   Median :27                     Median :39.00   NA's     :
2
##  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
##      visits            marital        gained          weight
##  Min.   : 0.0   married    :386   Min.   : 0.00   Min.   : 1.000
##  1st Qu.:10.0   not married:613   1st Qu.:20.00   1st Qu.: 6.380
##  Median :12.0   NA's       :  1   Median :30.00   Median : 7.310
##  Mean   :12.1                     Mean   :30.33   Mean   : 7.101
```

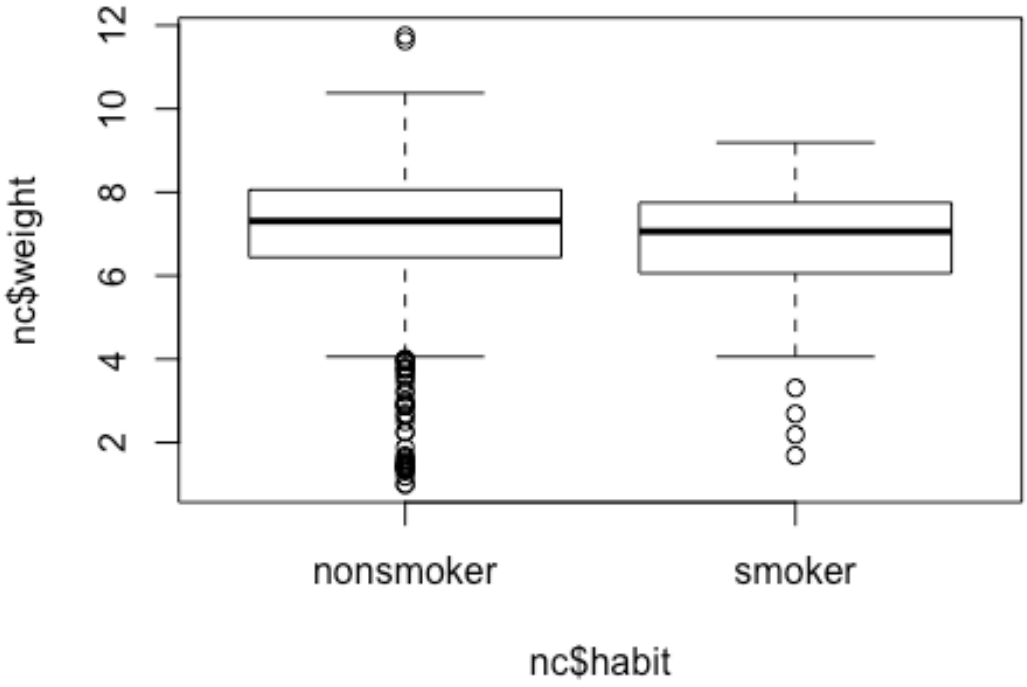
```
##  3rd Qu.:15.0                    3rd Qu.:38.00   3rd Qu.: 8.060
##  Max.   :30.0                    Max.   :85.00   Max.   :11.750
##  NA's   :9                       NA's   :27
##  lowbirthweight    gender          habit          whitemom
##  low    :111    female:503   nonsmoker:873   not white:284
##  not low:889    male  :497   smoker   :126   white    :714
##                              NA's     :  1   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.

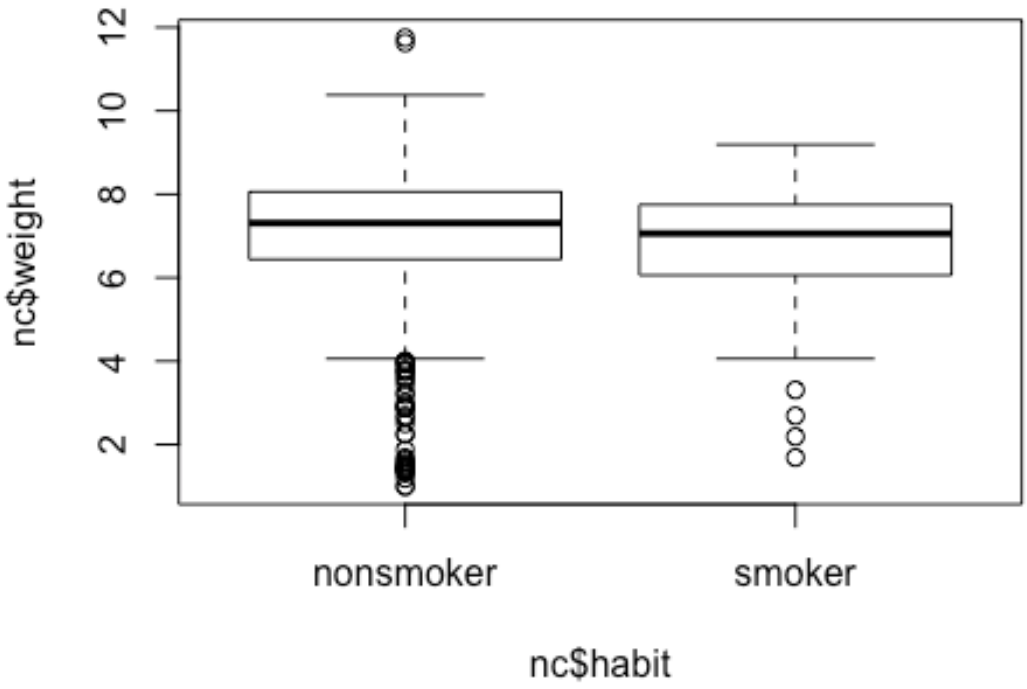
```
plot(nc$weight ~ nc$habit)
```



Exercise 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)
```



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

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
```

```
table(nc$habit)
```

```
##
## nonsmoker    smoker
##       873       126
```

Exercise 4

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

Null Hypothesis Ho: Their is no difference in wieghts of the babies of two groupd of smokers

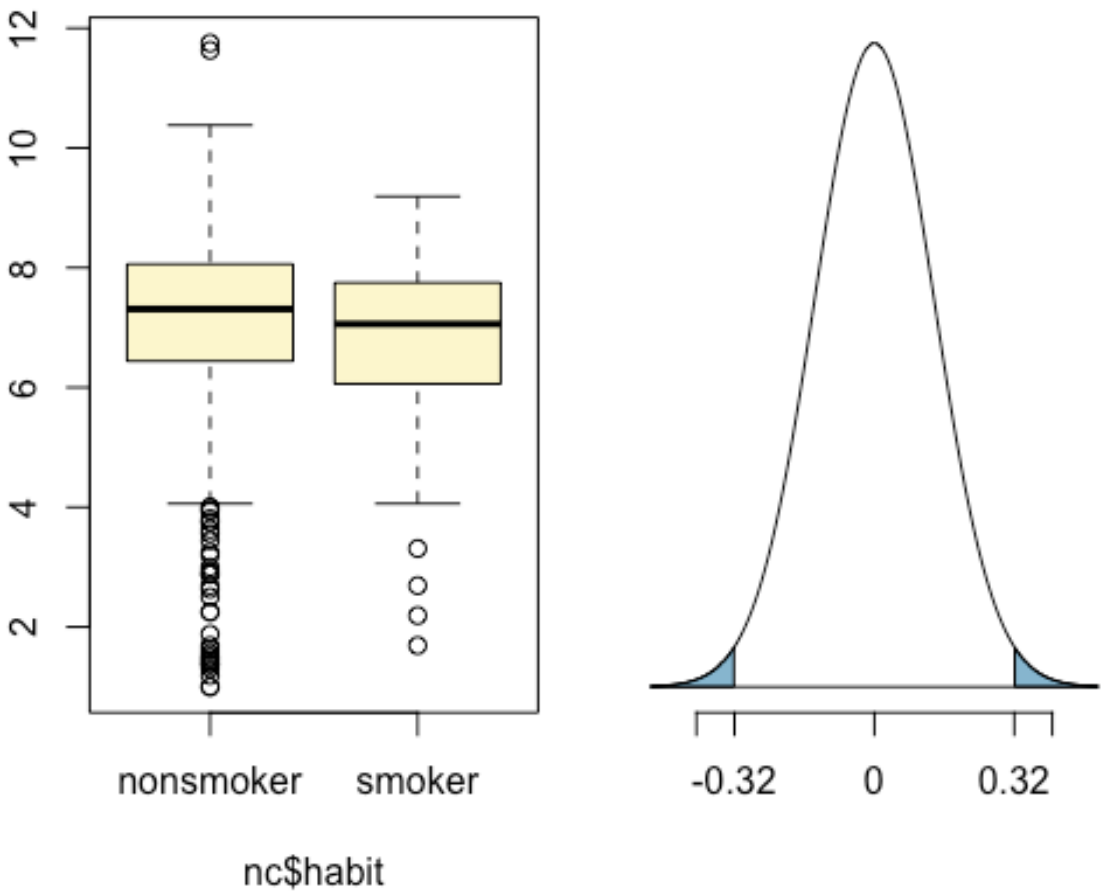
Alt Hypothesis Ha: Their is a difference in wieghts of the babies of two groupd of smokers

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
```



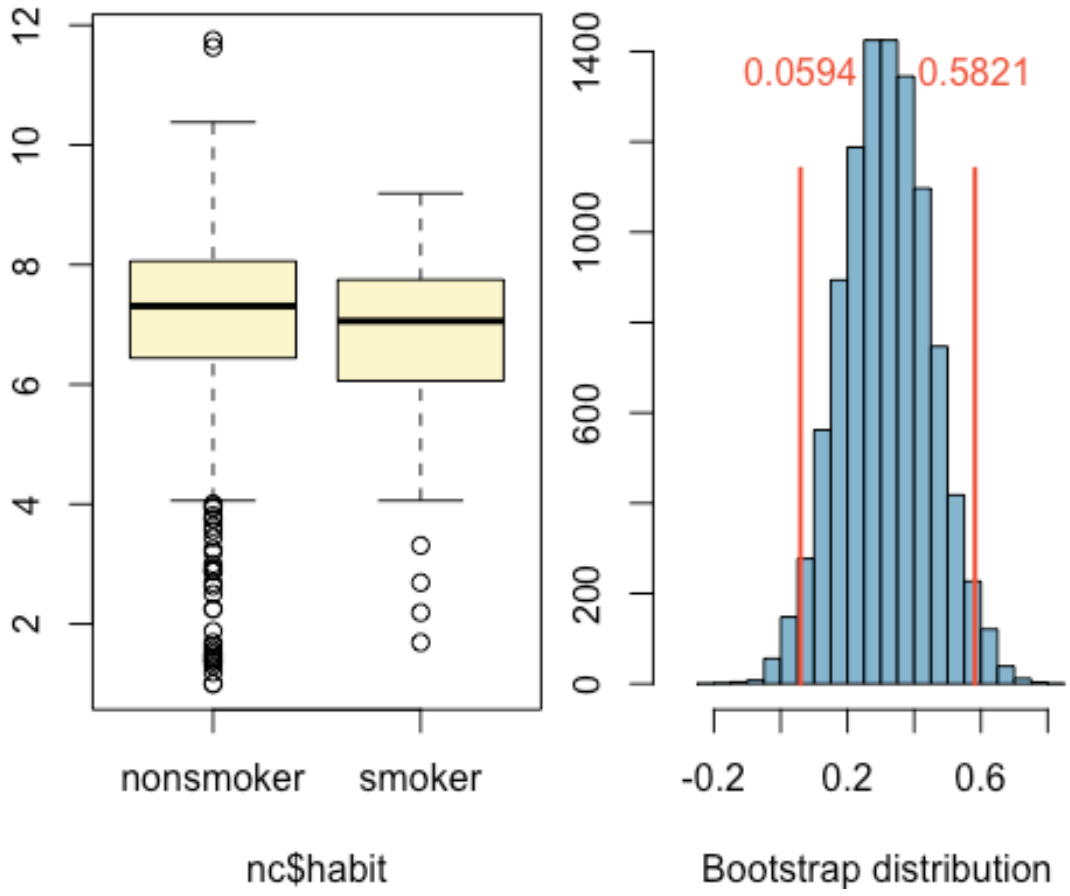
Let's pause for a moment to go through the arguments of this custom function. # The first argument is y, which is the response variable that we are interested in: nc$weight. # The second argument is the explanatory variable, x, which is the variable that splits the data into two groups, # smokers and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

Exercise5: 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 = "simulation")

## 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
```



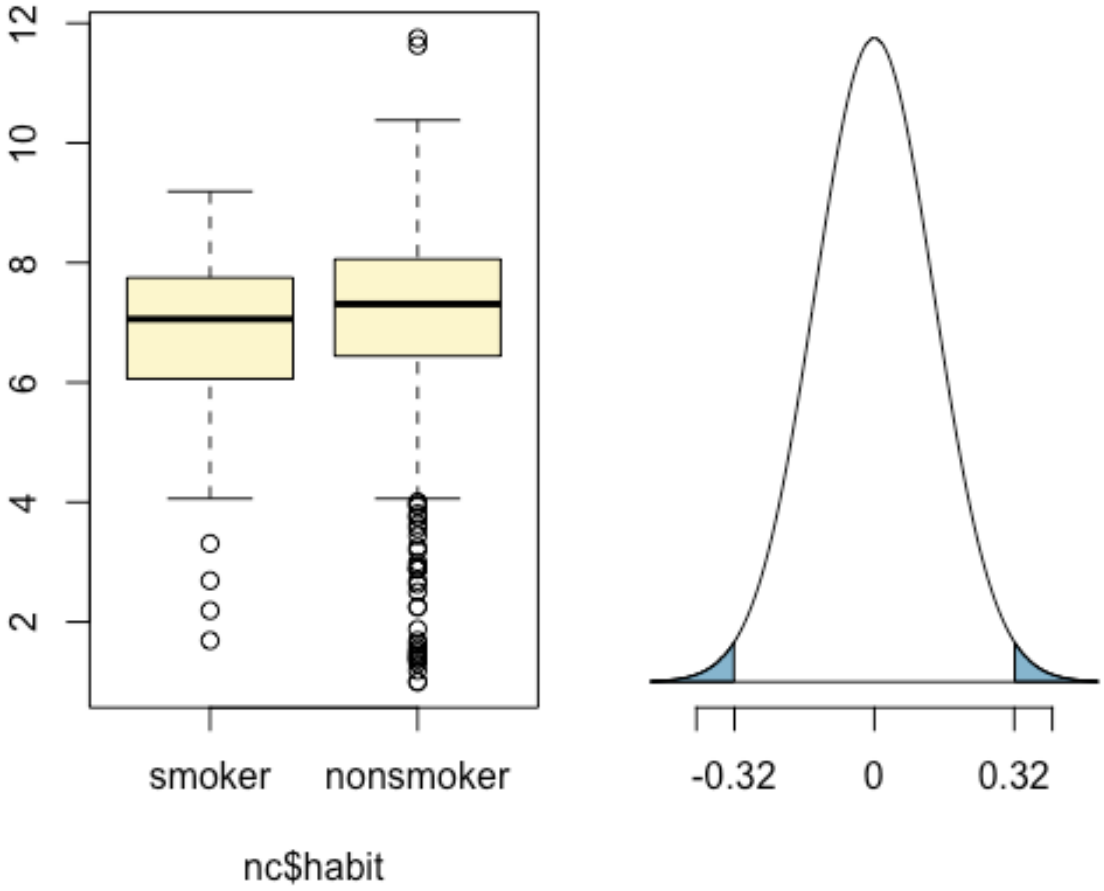

```
## 95 % Bootstrap interval = ( 0.0594 , 0.5821 )
```

By default the function reports an interval for (?? nonsmoker ????? smoker

We can easily change this order by using the order argument:

```
  inference(y = nc$weight, x = nc$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 = 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
##
## H0: mu_smoker - mu_nonsmoker = 0
## HA: mu_smoker - mu_nonsmoker != 0
## Standard error = 0.134
## Test statistic: Z =  -2.359
## p-value =  0.0184
```

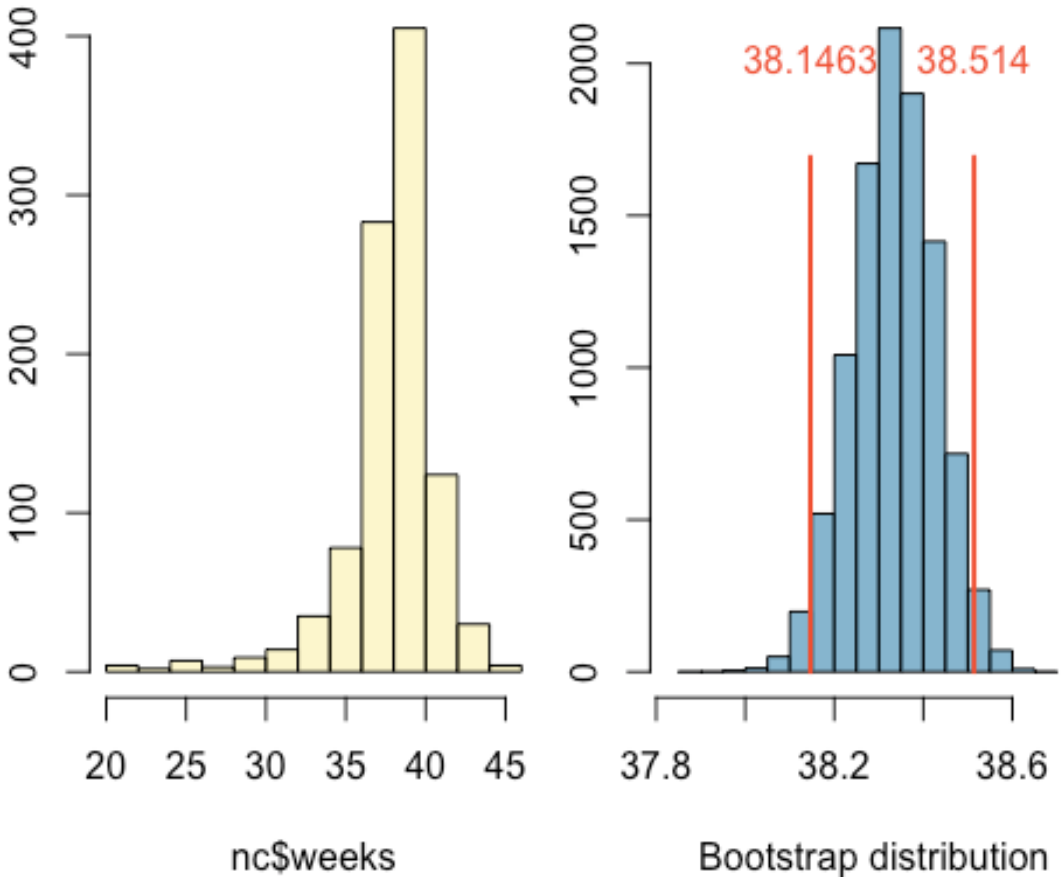


#On your own

#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", null = 0,
            alternative = "twosided", method = "simulation")

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

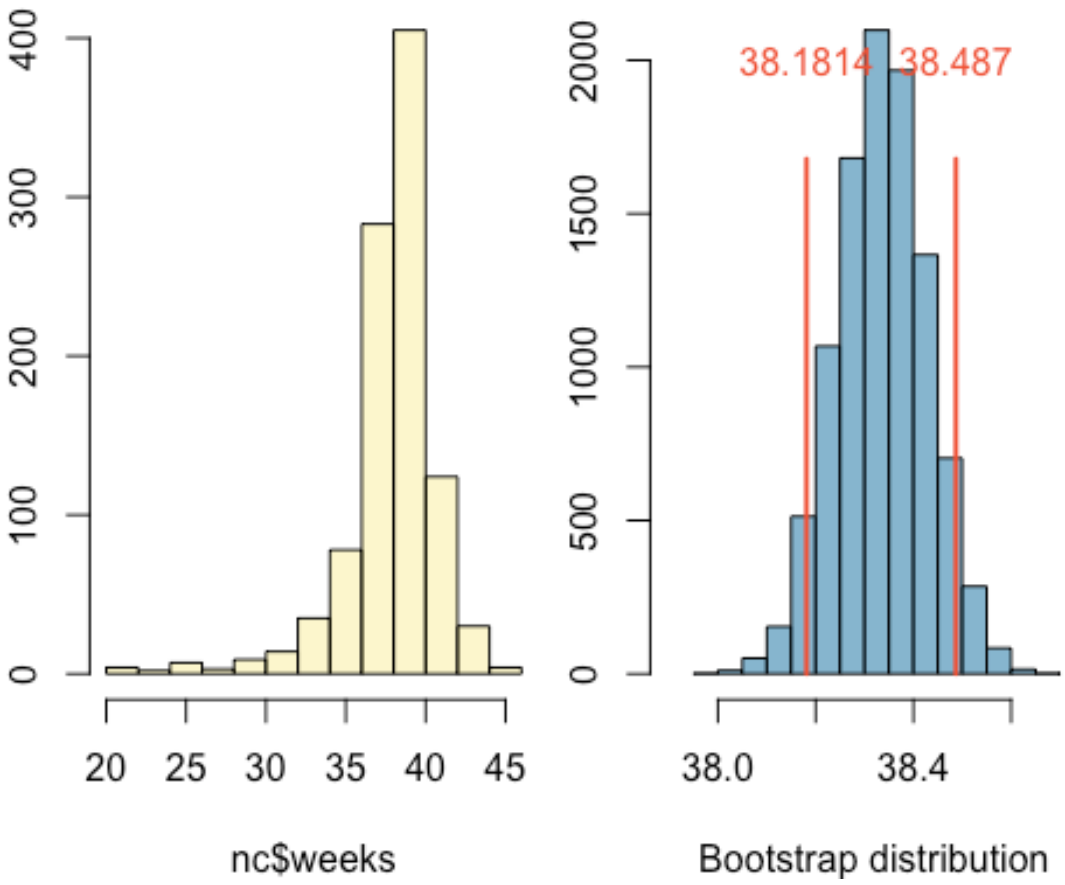
```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
```




```
## 95 % Bootstrap interval = ( 38.1463 , 38.514 )
```

#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.

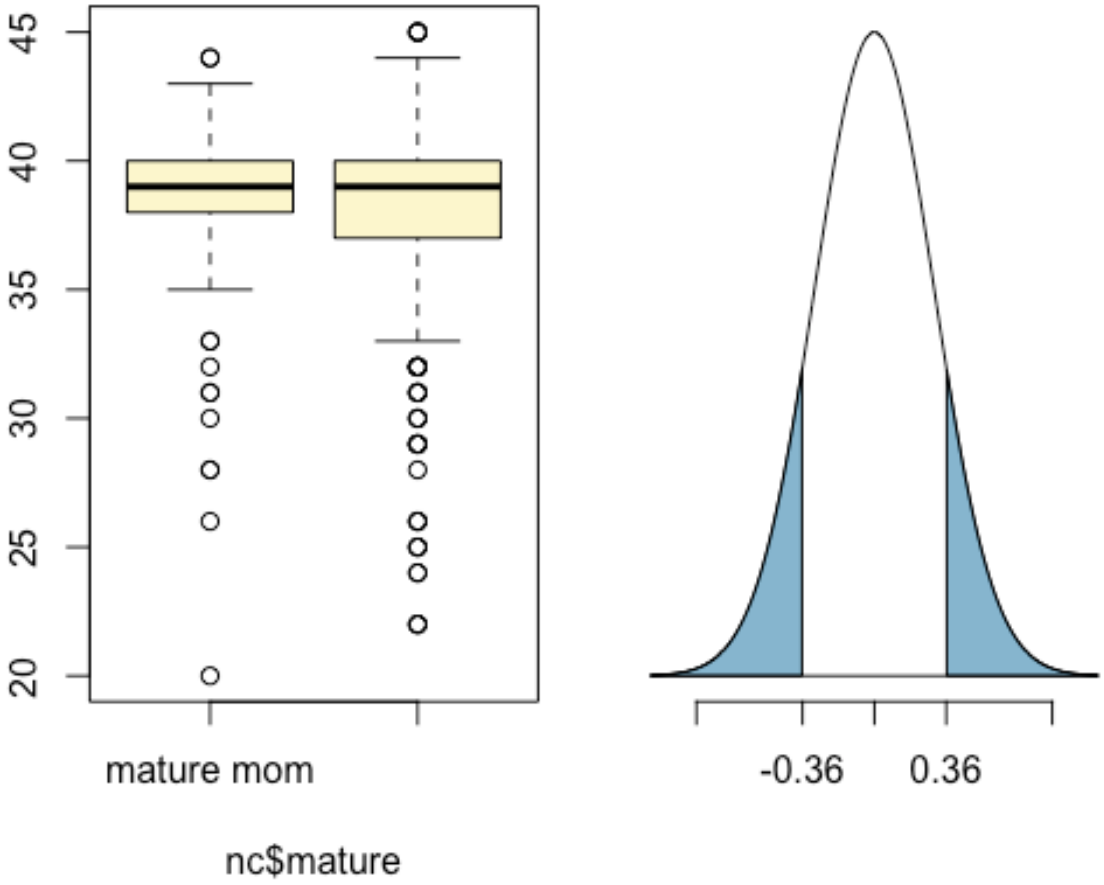
```
inference(y = nc$weeks, est = "mean", type = "ci", null = 0,
            alternative = "twosided", method = "simulation", conflevel =
0.90)

## Single mean
## Summary statistics:

## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
```




```
## 90 % Bootstrap interval = ( 38.1814 , 38.487 )
```

#3.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$weeks, x = nc$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 = 132, mean_mature mom = 38.0227, sd_mature mom = 3.2184
## n_younger mom = 866, mean_younger mom = 38.3822, sd_younger mom = 2.8844

## Observed difference between means (mature mom-younger mom) = -0.3595
##
## H0: mu_mature mom - mu_younger mom = 0
## HA: mu_mature mom - mu_younger mom != 0
## Standard error = 0.297
## Test statistic: Z =  -1.211
## p-value =  0.2258
```



#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. The cuttoff age is in years, either the maximum age of younger momor minium age age of mature mom leads to the results.

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

```
## [1] 34
```

```
min(nc$mage[nc$mature == "mature mom"])
```

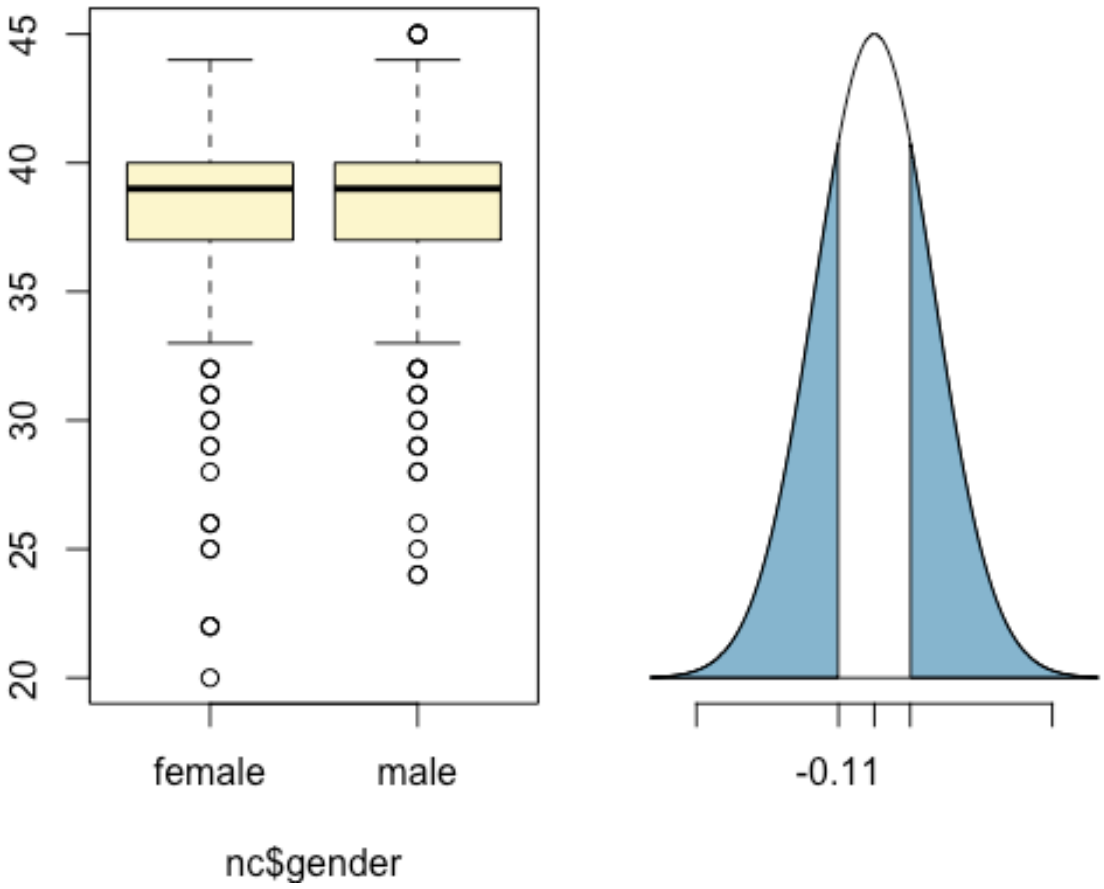
```
## [1] 35
```

#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.

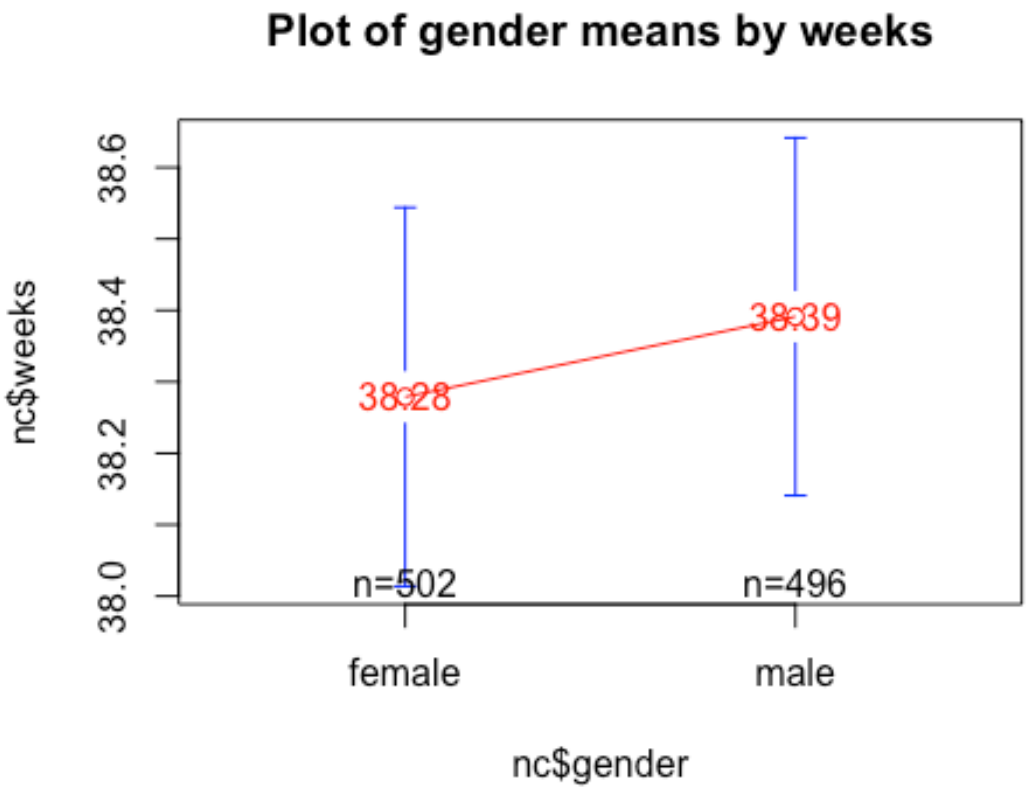
```
inference(y = nc$weeks, x = nc$gender, est = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

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

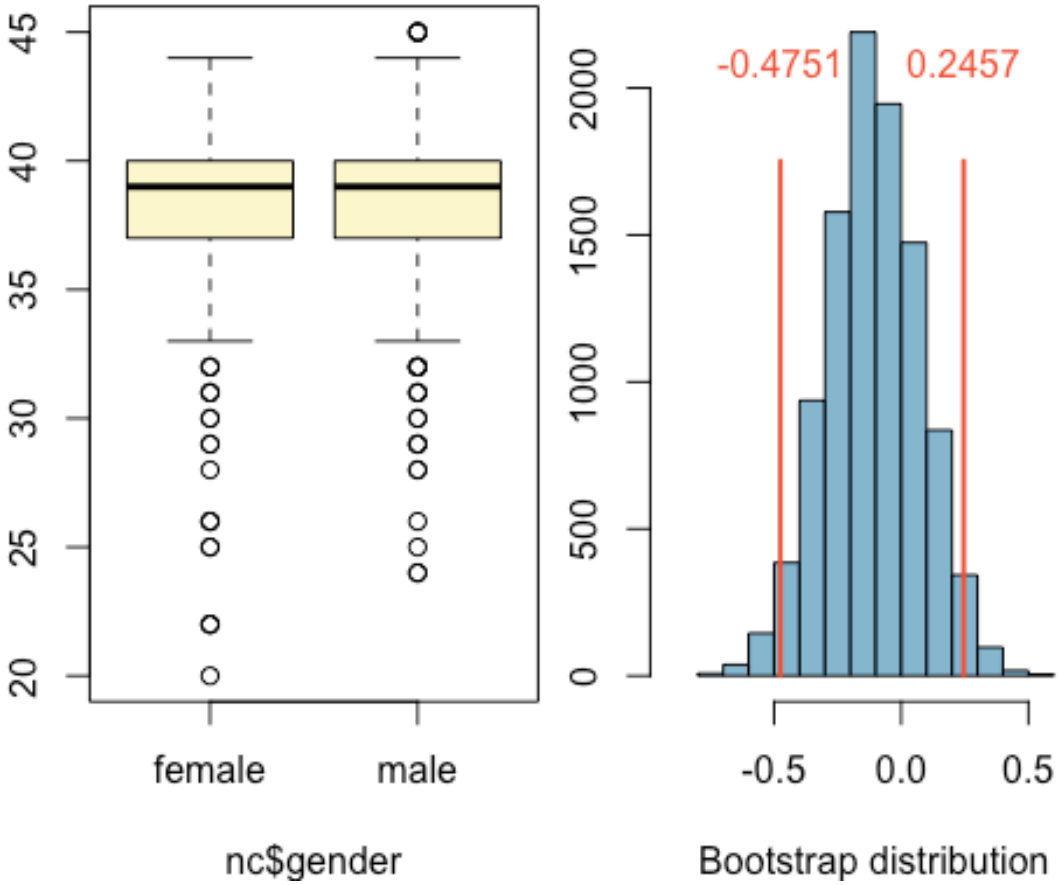
```
## Summary statistics:
## n_female = 502, mean_female = 38.2789, sd_female = 3.0232
## n_male = 496, mean_male = 38.3911, sd_male = 2.8377

## Observed difference between means (female-male) = -0.1122
##
## H0: mu_female - mu_male = 0
## HA: mu_female - mu_male != 0
## Standard error = 0.186
## Test statistic: Z =  -0.605
## p-value =  0.5454
```




```
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

plotmeans(nc$weeks ~ nc$gender, digits=2, col="red", mean.labels=T,
main="Plot of gender means by weeks")
```




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

## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_female = 502, mean_female = 38.2789, sd_female = 3.0232
## n_male = 496, mean_male = 38.3911, sd_male = 2.8377

## Observed difference between means (female-male) = -0.1122
```




```
## 95 % Bootstrap interval = ( -0.4751 , 0.2457 )

summary(aov(nc$weeks ~ nc$gender))

##              Df Sum Sq Mean Sq F value Pr(>F)
## nc$gender     1      3   3.143   0.366  0.546
## Residuals   996   8565   8.599
## 2 observations deleted due to missingness
```

From the box plots it is not clearly seen the variations in the means of male and female. So the plotmeans graph clearly shows us that their is a variation in meand of the male and female. there fore we can not reject the null hypothesis. The F value is very less,which is kind of bad, the p-value is 0.546 whihc is greater than the 0.056 (as suggested by normal scientific standard) from the ANOVA summary. Hence we can conclude that for our confidence interval we accept the alternative hypothesis H1 that there is no significant relationship between weeks and gender.