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.

```
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.
`weight`	weight of the baby at birth in pounds.

variable	description
`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?

```
nrow(nc)
```

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

The cases in the data set are information on births within the State of North Carolina for the year 2004. The output of R's **nrow** function (above) shows that there are 1000 cases in the 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)
```

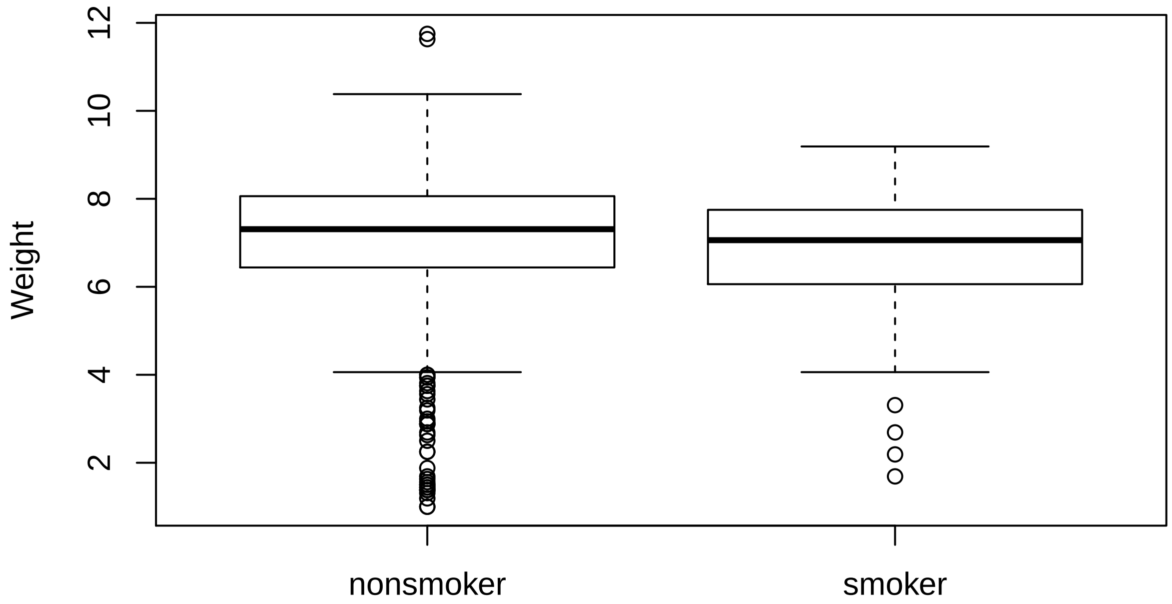
```
##       fage            mage            mature        weeks
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00
##  Median :30.00   Median :27                     Median :39.00
##  Mean   :30.26   Mean   :27                     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(nc$weight ~ nc$habit, ylab = "Weight")
```



The box plots show that the median weight of a smoker is a bit less than that of a non-smoker.

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`.

- **Samples must be independent events:** TRUE
- **Samples must be randomly selected:** TRUE
- **Sample size must be less than 10% of the population:** We'll check this using the 'by' command:

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

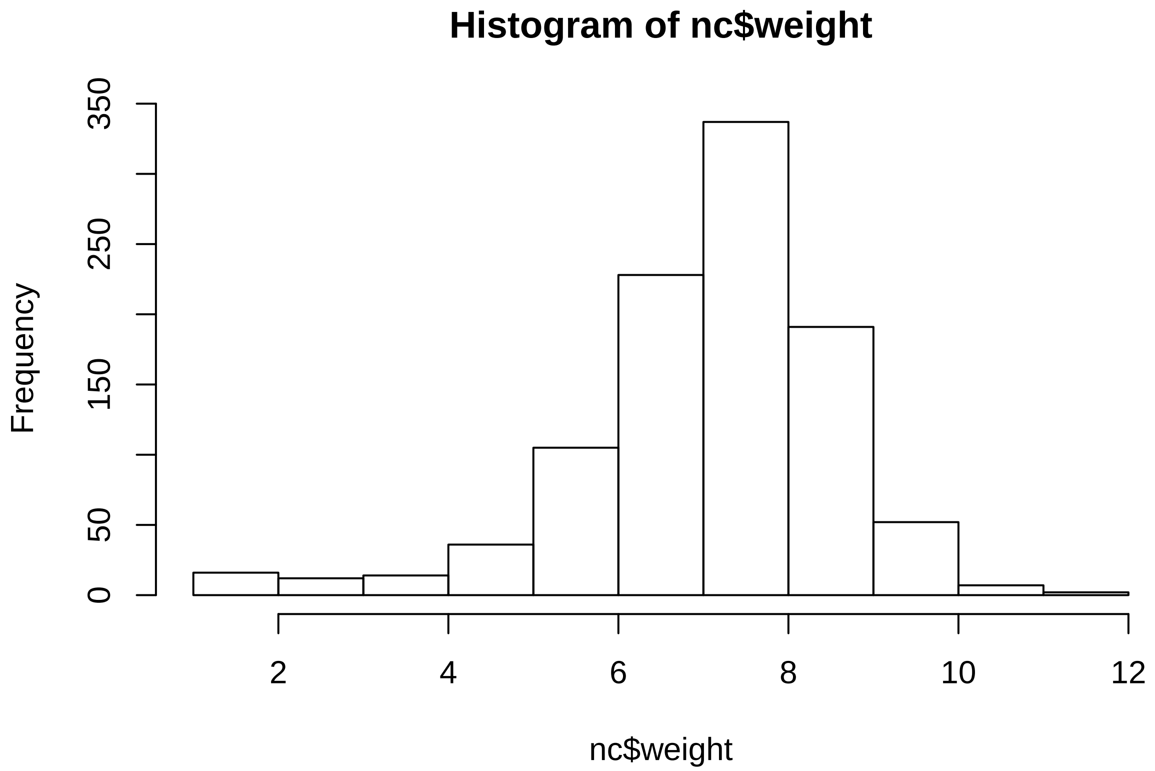
```
## nc$habit: nonsmoker
## [1] 873
## --------------------------------------------------------
## nc$habit: smoker
## [1] 126
```

TRUE (assuming North Carolina had at least 8,730 birthmothers who did not smoke in 2004 and 1,260 birthmothers who did smoke in 2004).

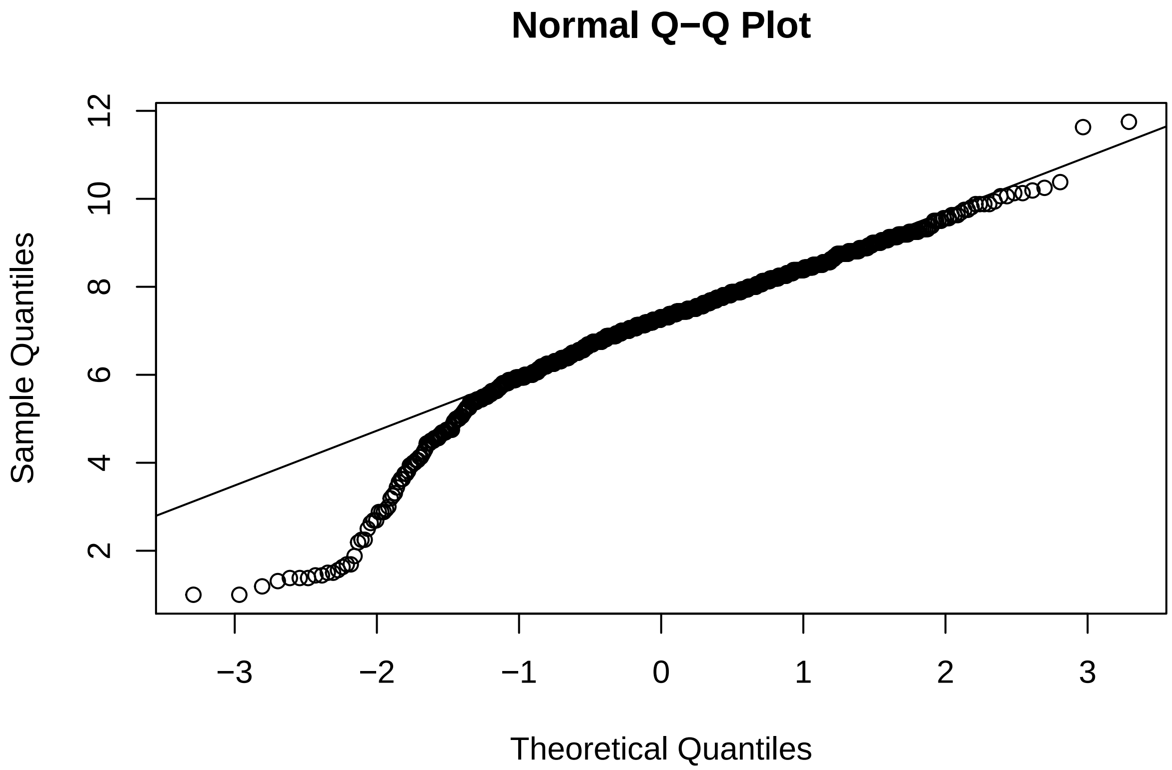
- **The distribution of original population cannot be heavily skewed.**

TRUE. As shown in the following plots, the distribution is left skewed. However, the sample size of 1000 is sufficiently large to allow us to use a normal approximation.

```
hist(nc$weight)
```



```
qqnorm(nc$weight)
qqline(nc$weight)
```



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

H0: mu(weight of baby born to smoking mother) - mu(weight of baby born to nonsmoking mother) = 0
There is no difference in average birth weight for newborns from mothers who did and did not smoke.

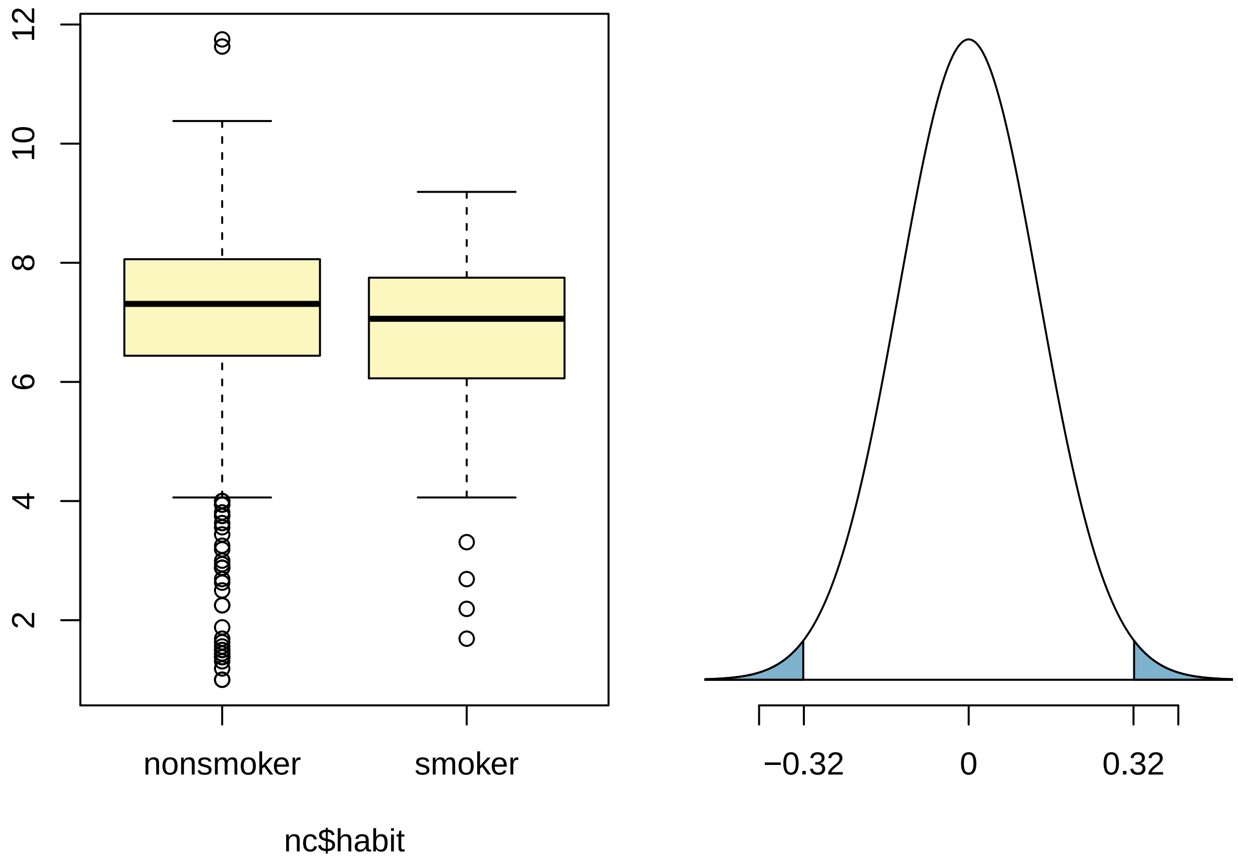
HA: mu(weight of baby born to smoking mother) - mu(weight of baby born to nonsmoking mother) ! = 0
There is a difference in average birth weight for newborns from mothers who did and did not smoke.

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
```

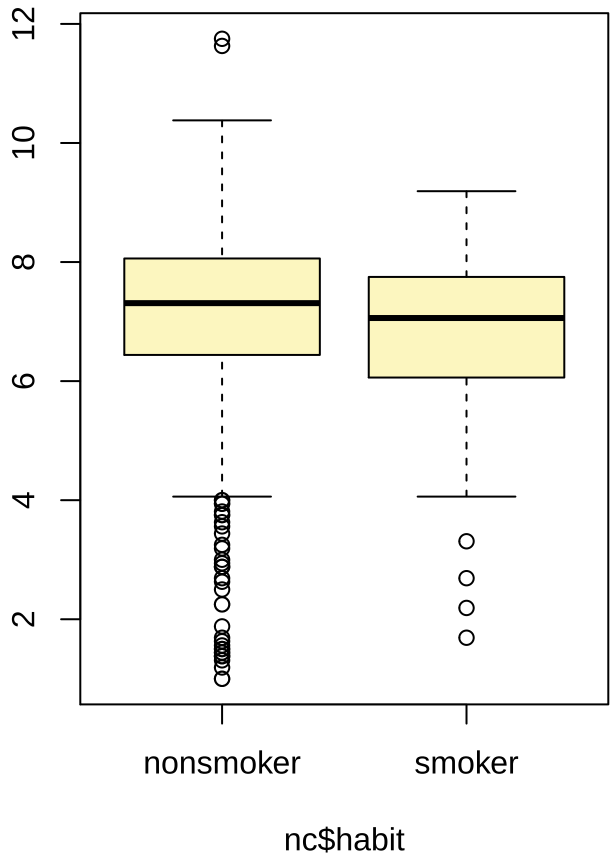


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

```r
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
```



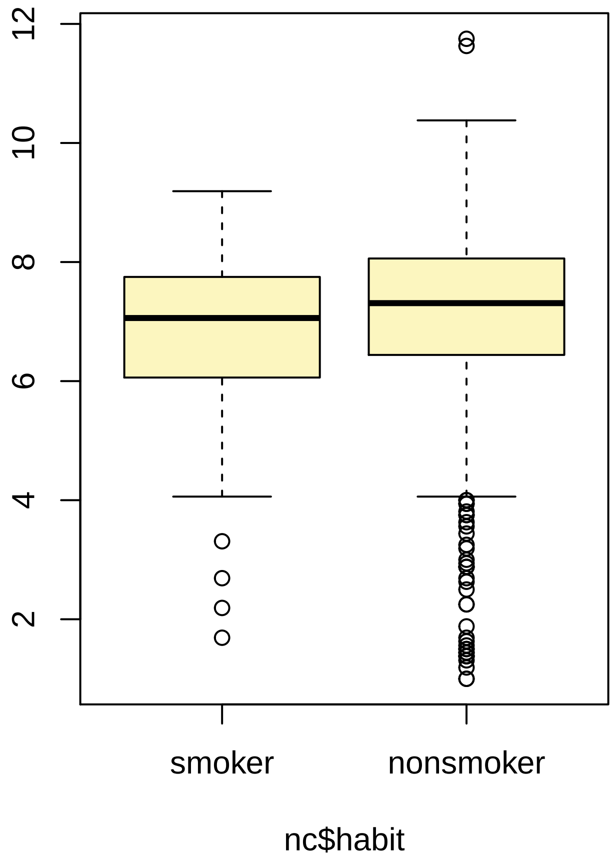
```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

The 95% confidence interval is (0.0534, 0.5777) for the difference between the weights of babies born to smoking and non-smoking mothers. It tells us that we are 95% confident that the actual difference in average birth weights between smoking and non-smoking mothers within the total population should be between 0.0534 and 0.5777 pounds.

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
```



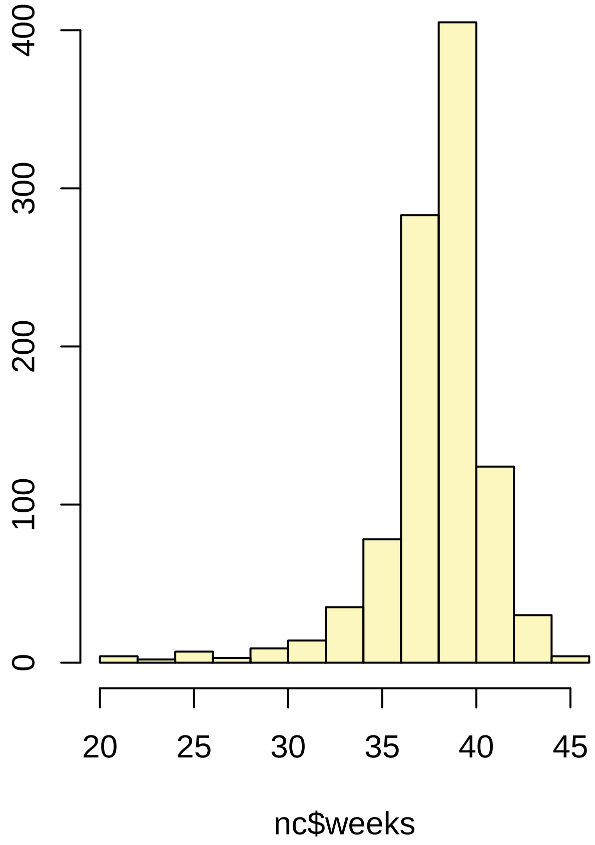
```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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", 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 )
```

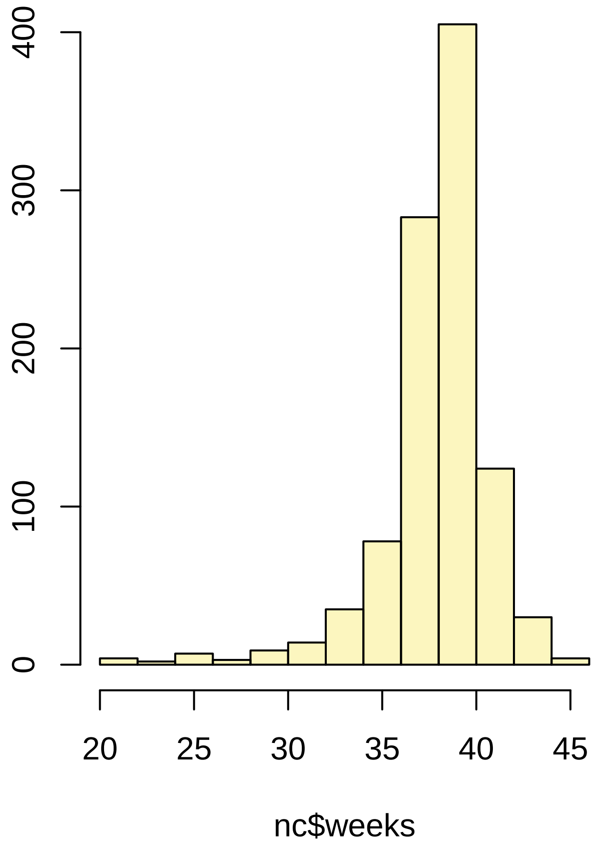
The confidence interval (38.1528, 38.5165) tells us that 95% of the sample means for the “weeks” variable should fall between 38.1528 and 38.5165.

It also tells us that we are 95% confident that the average number of weeks for the population should be between 38.1528 and 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`.

```
inference(y = nc$weeks, est = "mean", type = "ci", method = "theoretical", conflevel = 0.90)
```

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



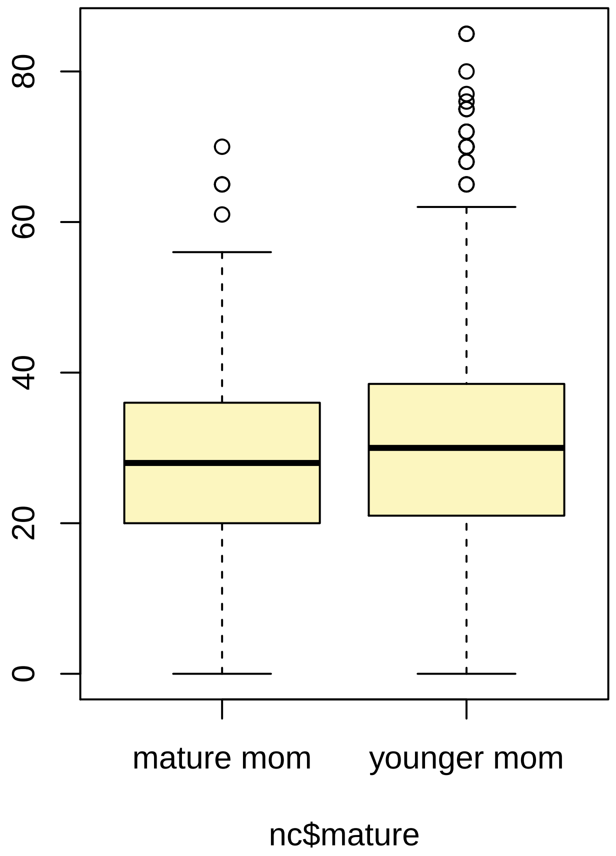
```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 90 % Confidence interval = ( 38.182 , 38.4873 )
```

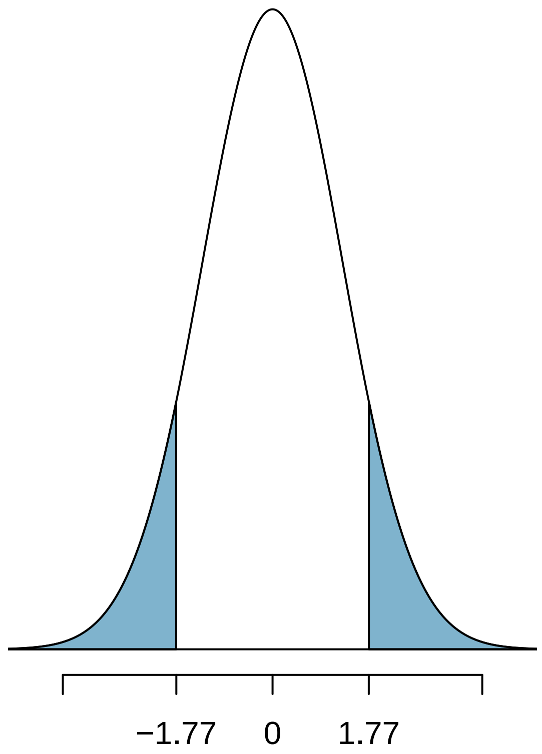
- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
inference(y = nc$gained, x = nc$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
```





Since the P value of .1686 is greater than .05, we fail to reject the null hypothesis that the mean weights of mature and younger mothers are equivalent.

- 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.

To find the maximum age of younger mothers, we make use of R's **subset** and **max** functions, applying a boolean search limiter to the nc$mature variable:

```
maxAgeYm <- max(subset(nc, mature == 'younger mom' & !is.na(mage))$mage)
maxAgeYm
```

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

This tells us that the maximum age of a younger mom is 34.

Now that we have the maximum age of a younger mom, we can perform a similar search using 'mature mom' as the limiting search boolean to check whether the minimum age of a mature mom is greater than the maximum age of a younger mom:

```
minAgeMm <- min(subset(nc, mature == 'mature mom' & !is.na(mage))$mage)
minAgeMm
```

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

We see that the minimum age of a mature mom is 35, which conforms with our earlier results for the maximum age of a younger mom.

- 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.

Research Question: Is the average weeks of pregnancy for white mothers equivalent to the average weeks of pregnancy for non-white mothers?

Our hypothesis test is stated as:

H0: mu(white mother) - mu(non-white mother) = 0
There is no difference between the average weeks of pregnancy for white mothers and non-white mothers

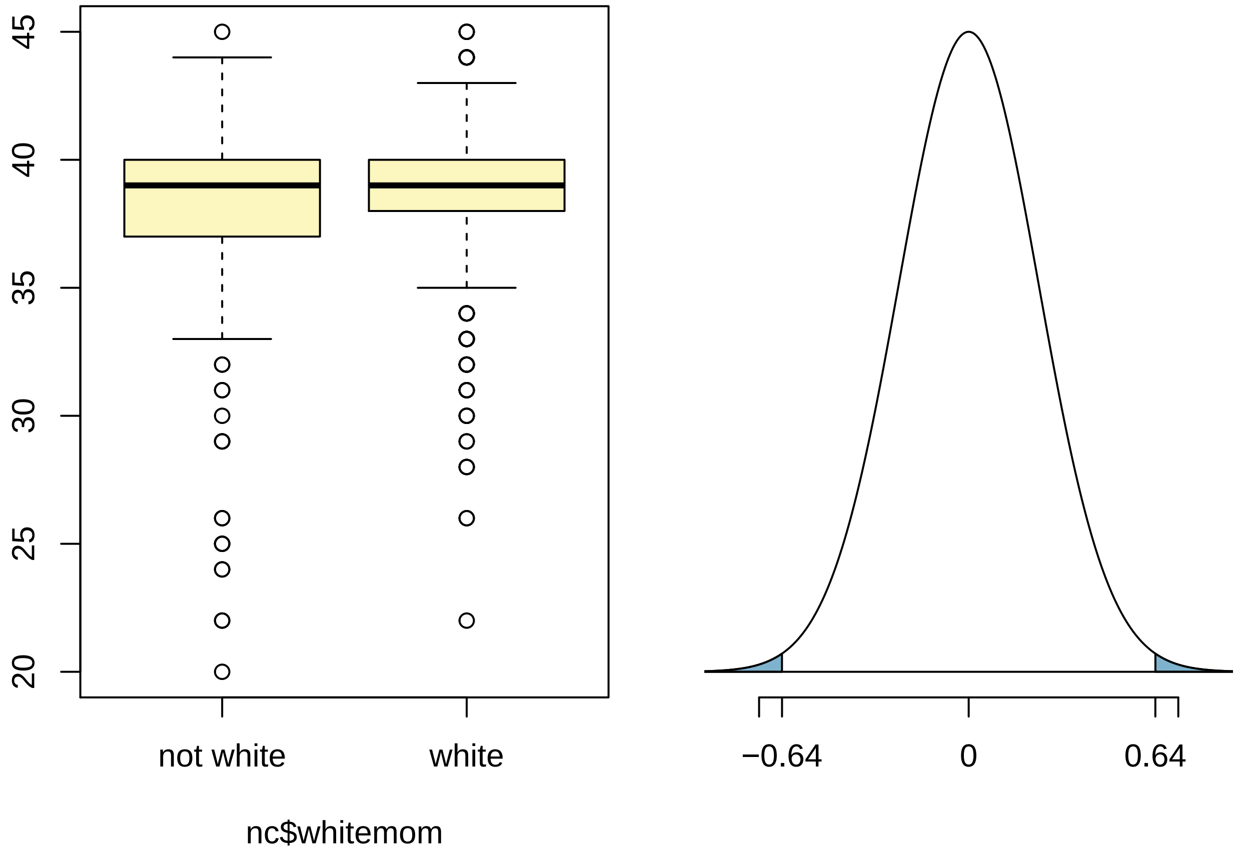
HA: mu(white mother) - mu(non-white mother) != 0
There is a difference in the average weeks of pregnancy for white mothers and non-white mothers.

Now we apply the hypothesis test at a 95% confidence interval:

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_not white = 284, mean_not white = 37.8768, sd_not white = 3.6705
## n_white = 712, mean_white = 38.5126, sd_white = 2.5617

## Observed difference between means (not white-white) = -0.6359
##
## H0: mu_not white - mu_white = 0
## HA: mu_not white - mu_white != 0
## Standard error = 0.238
## Test statistic: Z =  -2.672
## p-value =  0.0076
```

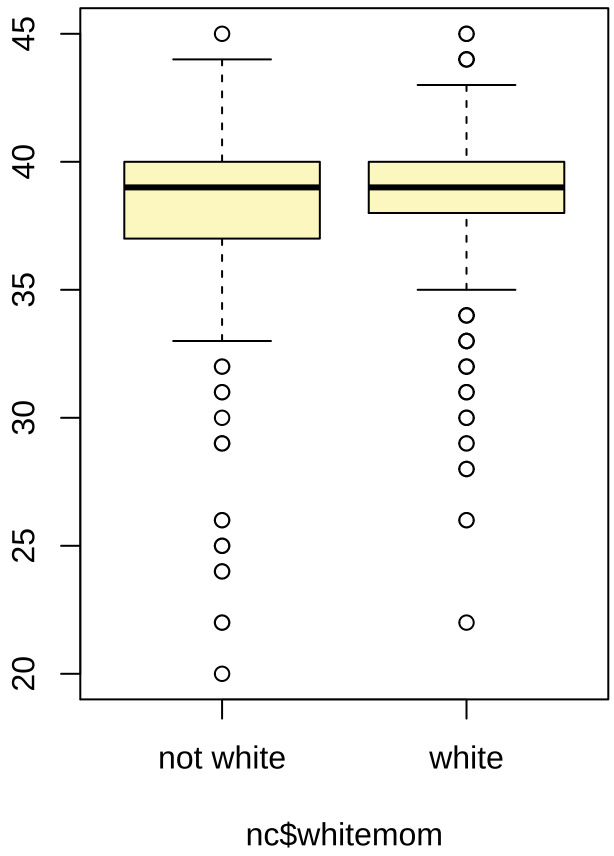


The results of the hypothesis test show a Z score of -2.672 and a p-value of .0076. The P value of .0076 is less than the .05 value we compare against for a 95% confidence interval. As such, we reject the null hypothesis and conclude that the average number of weeks of pregnancy for white mothers is not equal to the average weeks of pregnancy for non-white mothers. This result is confirmed via an examination of the side-by-side boxplots generated by the **inference** function. It clearly shows that non-white mothers' appear to experience fewer weeks of pregnancy than do white mothers.

Now let's check to see if these results are confirmed by a confidence interval test:

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_not white = 284, mean_not white = 37.8768, sd_not white = 3.6705
## n_white = 712, mean_white = 38.5126, sd_white = 2.5617
```



```
## Observed difference between means (not white-white) = -0.6359
##
## Standard error = 0.238
## 95 % Confidence interval = ( -1.1024 , -0.1694 )
```

The 95% confidence interval is (-1.1024, -0.1694). That is, we are 95% confident that the real difference in average weeks of pregnancy between non-white and white mothers is between -1.1 and -0.1694 weeks. Our null hypothesis **H0** presumed a real difference of 0 weeks, a value which clearly falls outside of the 95% condifence interval we've derived here. Therefore, we should reject the **H0**, and the result of our confidence interval test has confirmed the results of our p-value test