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.
`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?

The cases are for all the births population in North Carolina in 2004. Each case has the relevant information about the parents of the child, and some child's information as well. The sample has 1000 cases.

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  
##                 
##                 
##                 
##                 
```

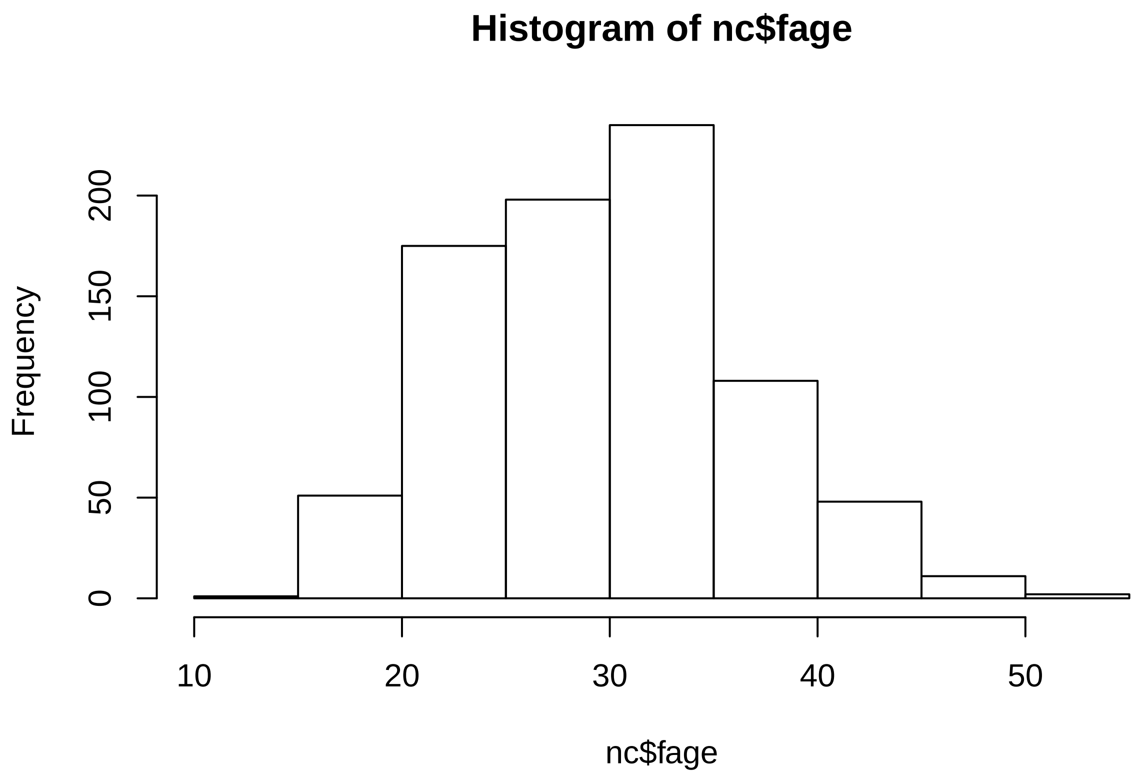
As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

Categorical variables are: mature premie marital low birth weight gender habit whitemom Numerical variables are: fage mage weeks visits gained weight

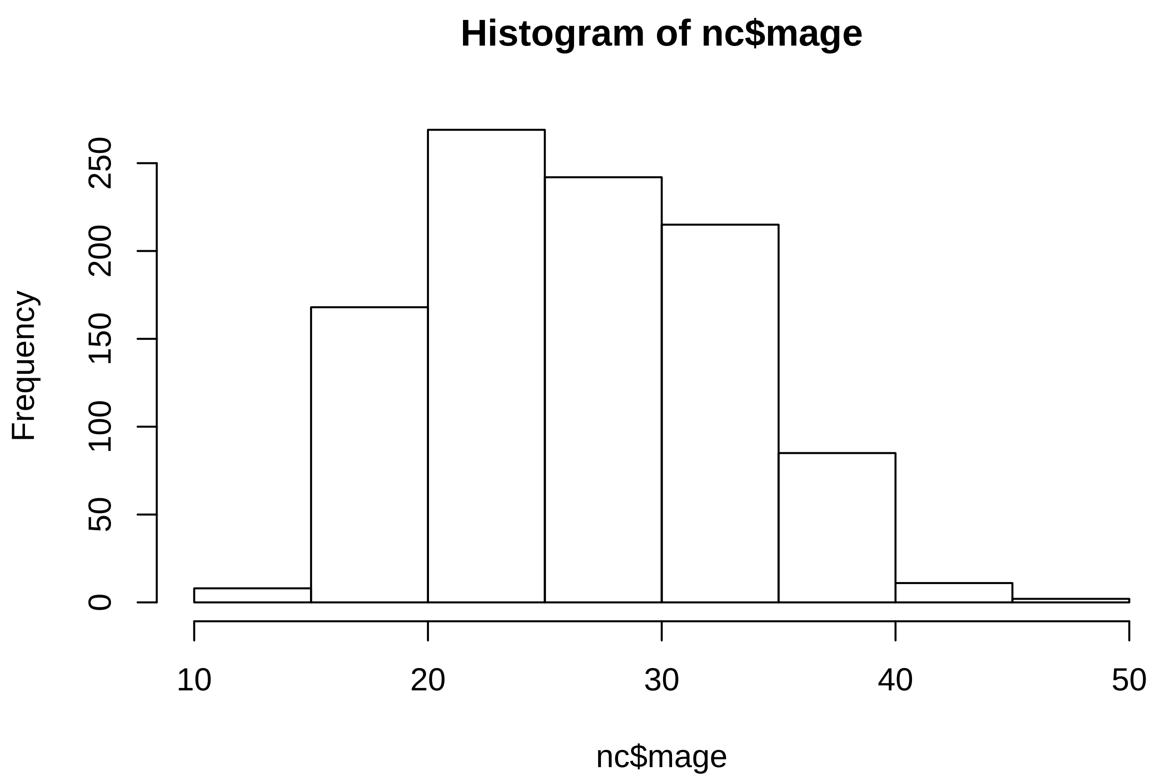
So out of total 13 variables, 7 are categorical and 6 are numerical.

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.

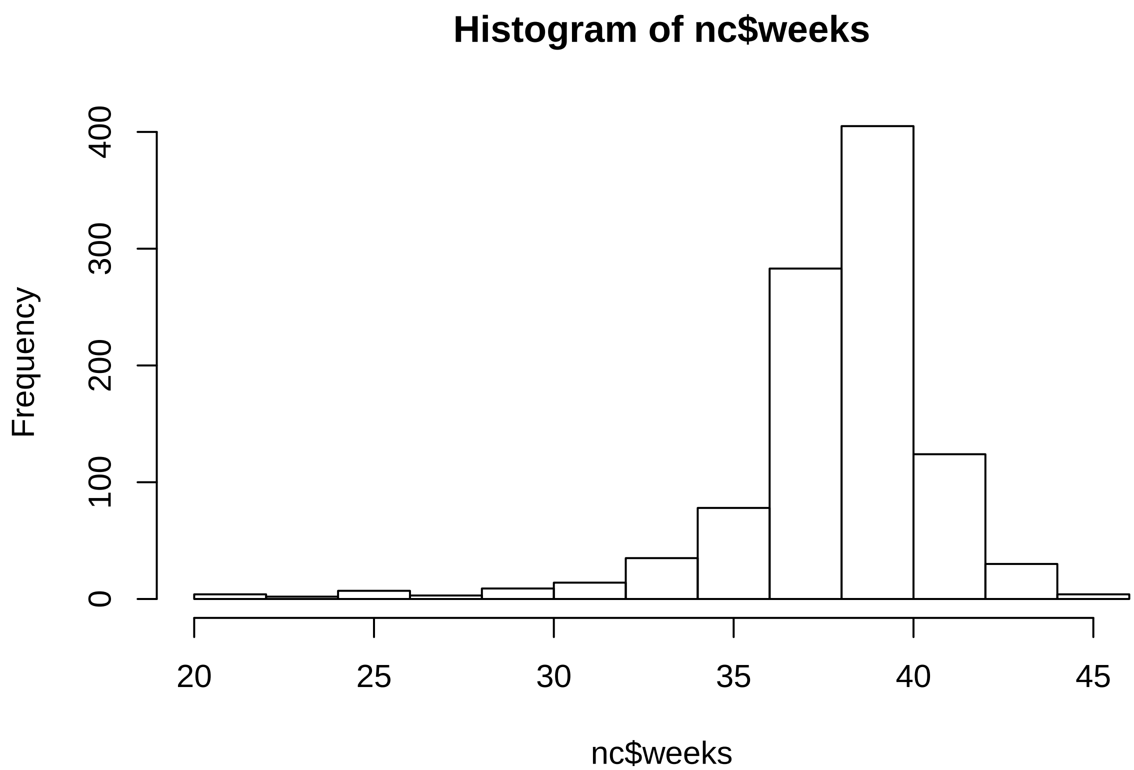
```r
hist(nc$fage)
```


Histogram of nc$fage
Frequency
0
50
100
150
200
10
20
30
40
50
nc$fage

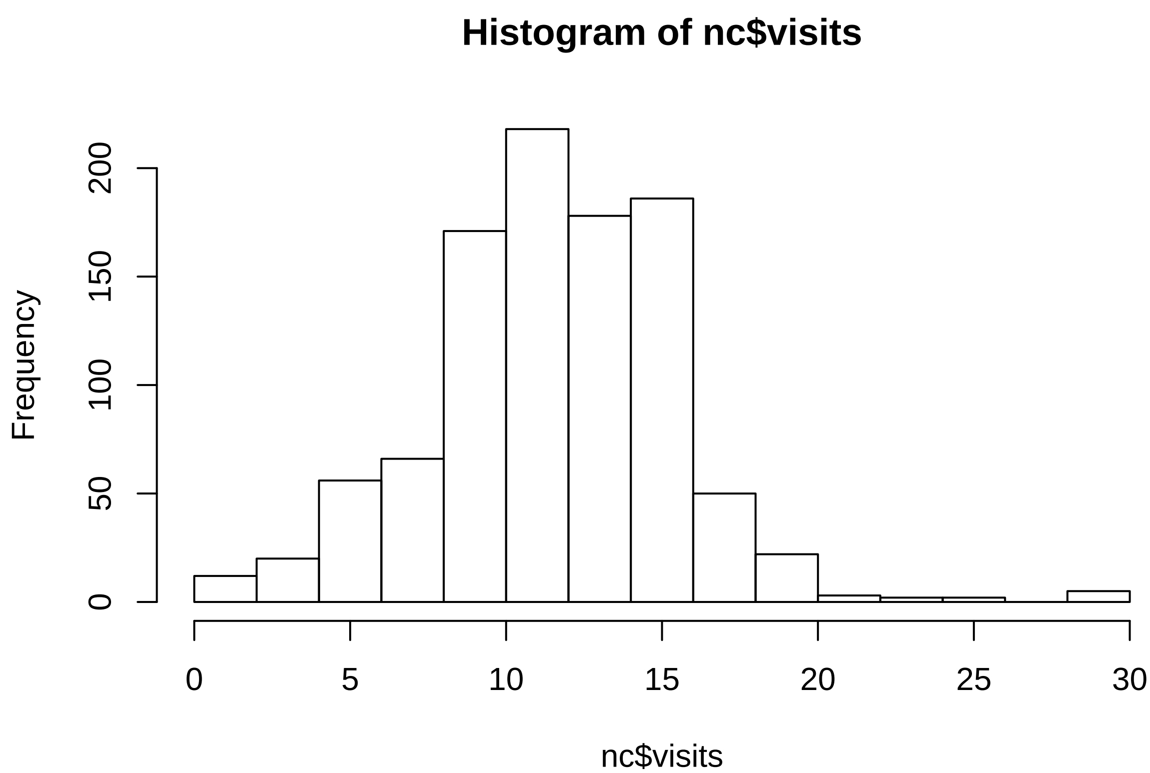
```
hist(nc$mage)
```


Histogram of nc$mage
Frequency
0
50
100
150
200
250
10
20
30
40
50
nc$mage

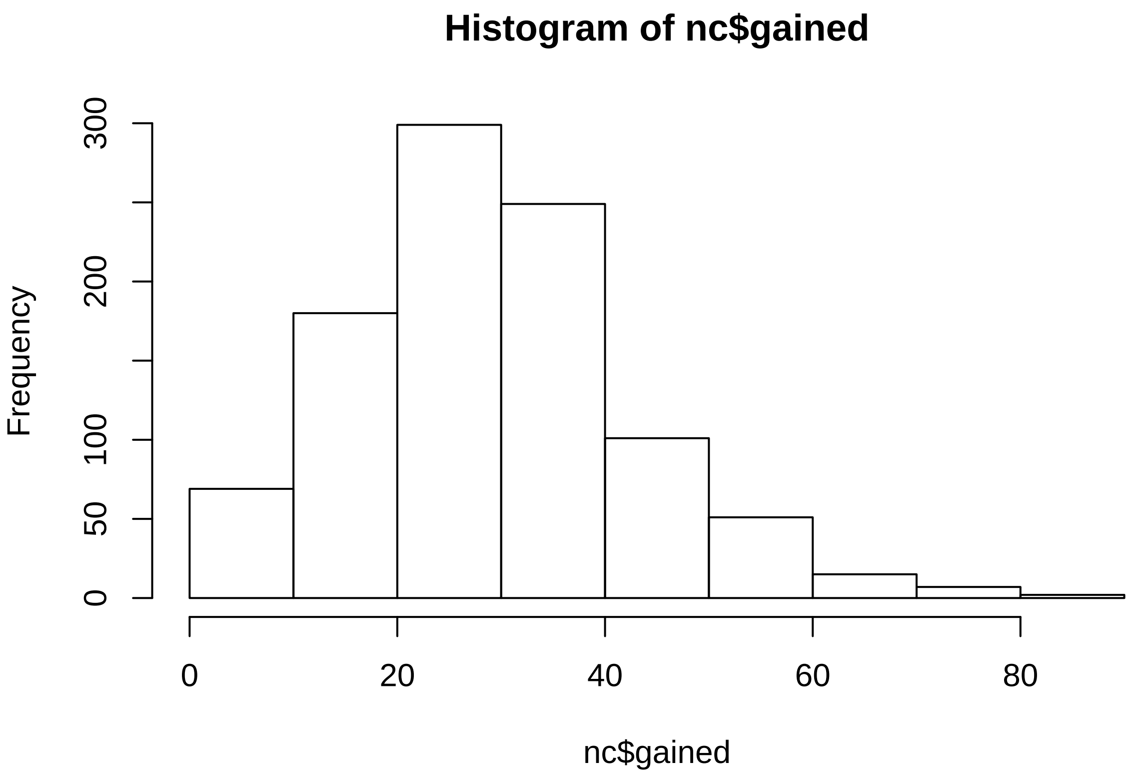
```r
hist(nc$weeks)
```



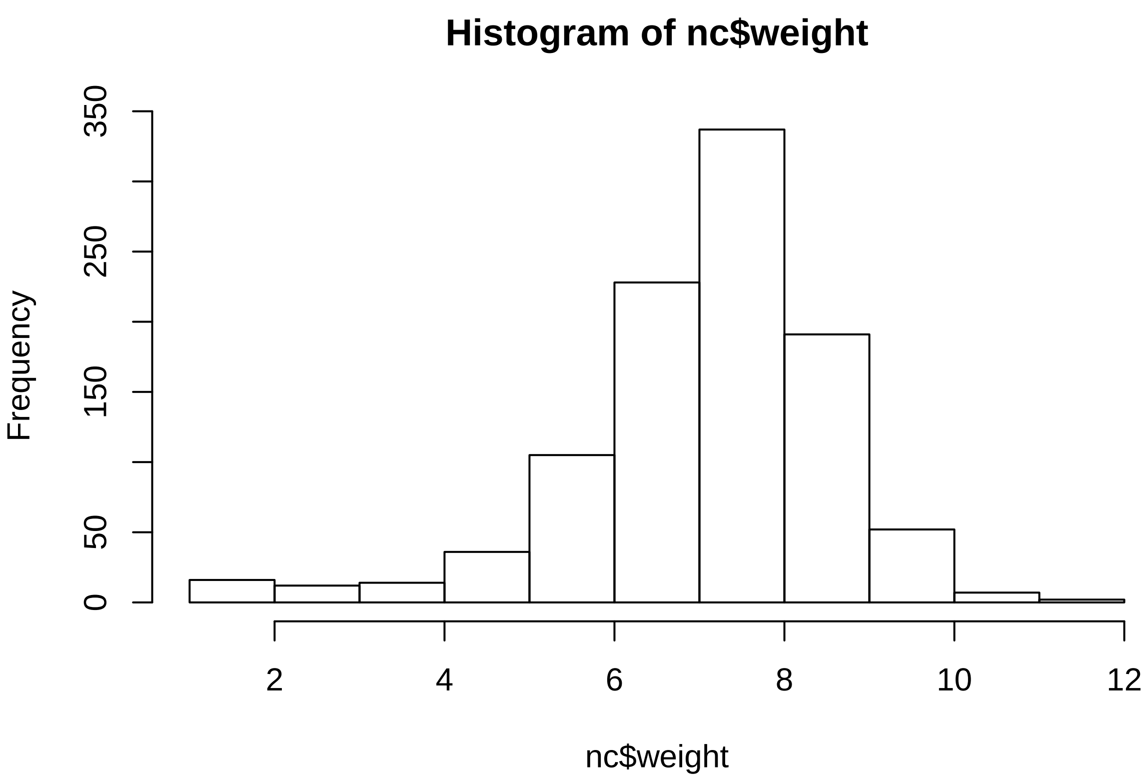

```r
hist(nc$visits)
```



```r
hist(nc$gained)
```



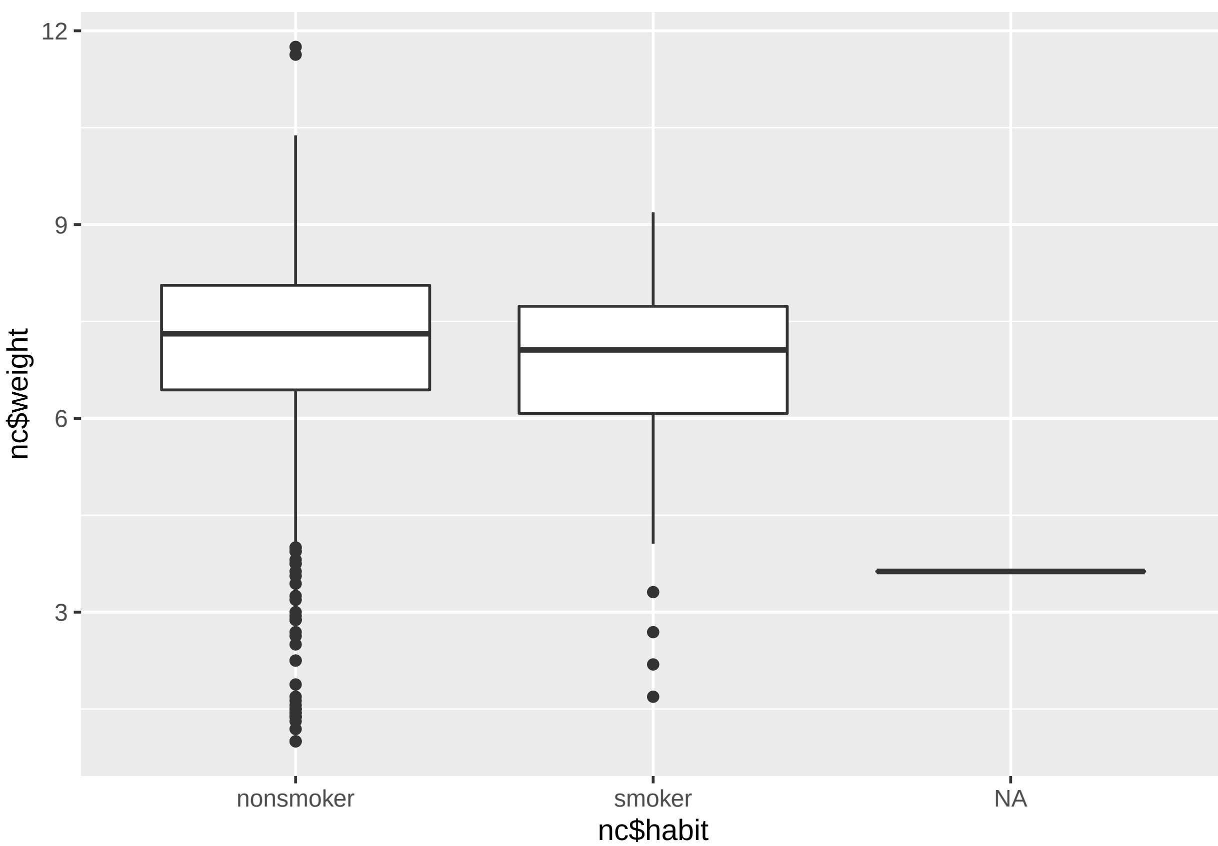

```r
hist(nc$weight)
```



From the histogram plots, we will deduce the following: 1) Father's age - no outliers as per the histogram. 2) Mother's age - no outliers as per the histogram. 3) Lengths of pregeneancy is highly shewed on the lefyt. That clearly shows that there are outliers in this case. 4) Number of hospital visits is having outliers on the right, that means more number of visits. 5) Weight gained also has a strong right skew, and there are outliers on the right side. 6) Weight of the baby has strong skew on the left. It means there are some observations where new born babies have very less weights.

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

```
library(ggplot2)
qplot(nc$habit, nc$weight, geom = "boxplot", na.rm = TRUE)
```



The median weight of the new brons from the mothers who smoke is less than the same fromt eh mother who does not smoke. Even though there are many outliers on the lower side in case of the non-smoking mothers, but the general range of weight for the non-smoker mothers is slightly more than the same for the 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

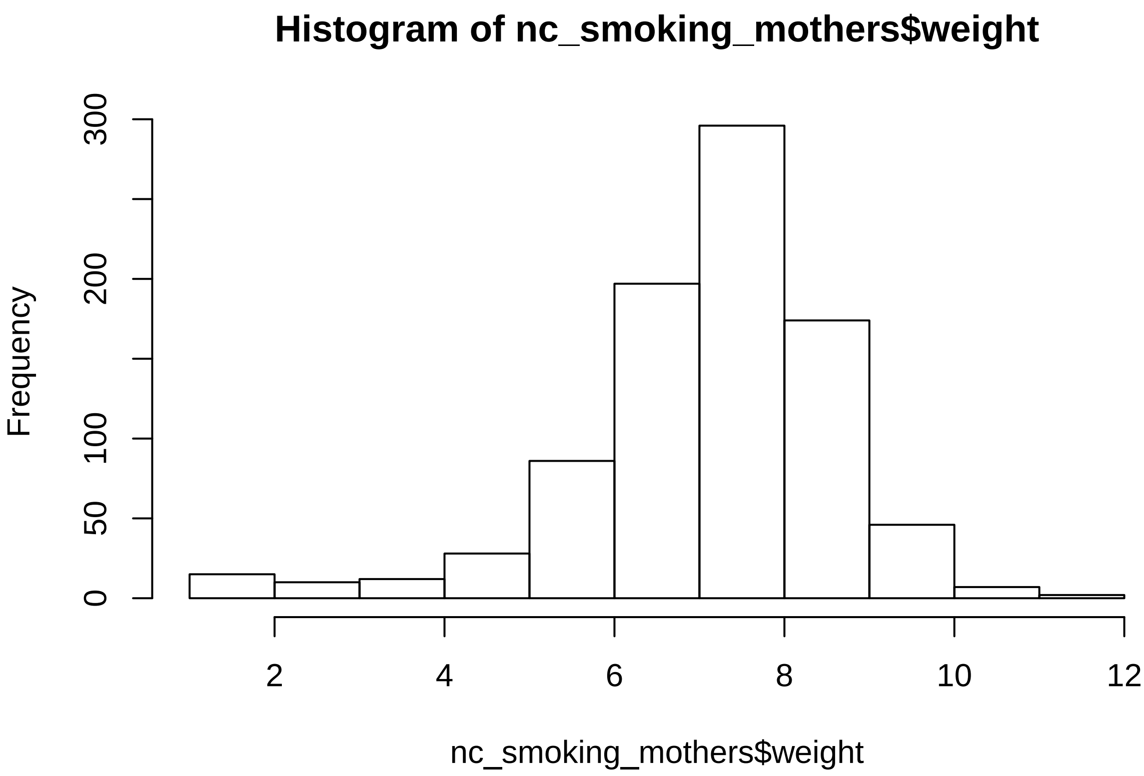
3. Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same `by` command above but replacing `mean` with `length`.

we are dealing with the weight of the new borns under the 2 categories - smoking mothers and non-smoking mothers. We will reate the histograms of the 2 categories separately:

```r
nc_smoking_mothers <- nc[nc$habit == "nonsmoker",]
nrow(nc_smoking_mothers)
```

```
## [1] 874
```

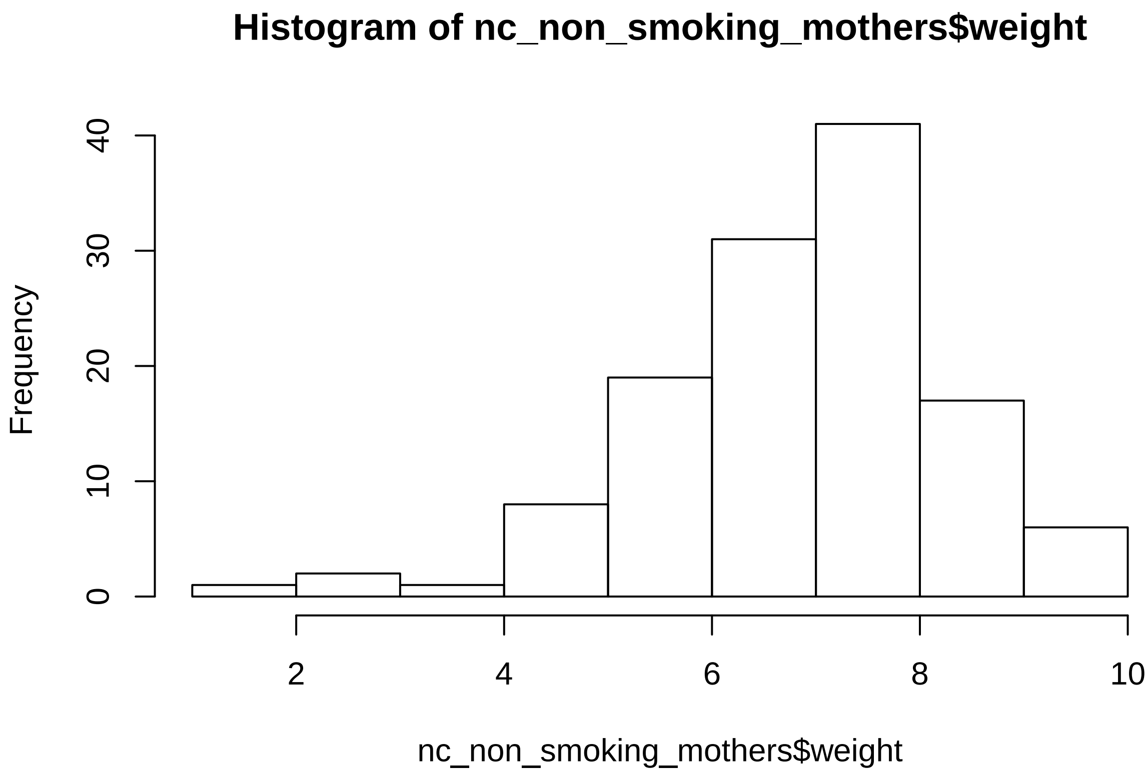
```r
hist(nc_smoking_mothers$weight)
```




```r
nc_non_smoking_mothers <- nc[nc$habit == "smoker",]
nrow(nc_non_smoking_mothers)
```

```
## [1] 127
```

```r
hist(nc_non_smoking_mothers$weight)
```



Both data smoking and non smoking are independent, and sample size is less than 10% of the total population

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

H0: There is no difference in the average weights of babies born to smoking and non-smoking mothers.

HA: There is a difference in the average weights of babies born to smoking and on-smoking mothers.

The mean weight of babies of smoking mothers: 6.82873. The mean weight of babies of nonsmoking mothers: 7.144273 smoker - nonsmoking: -0.3155425

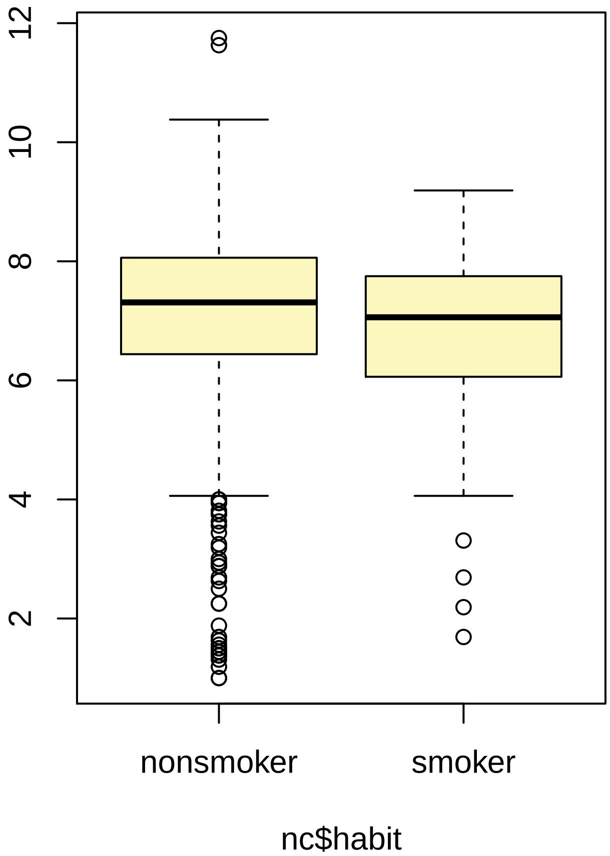
Next, we introduce a new function, `inference`, that we will use for conducting hypothesis tests and constructing confidence intervals.

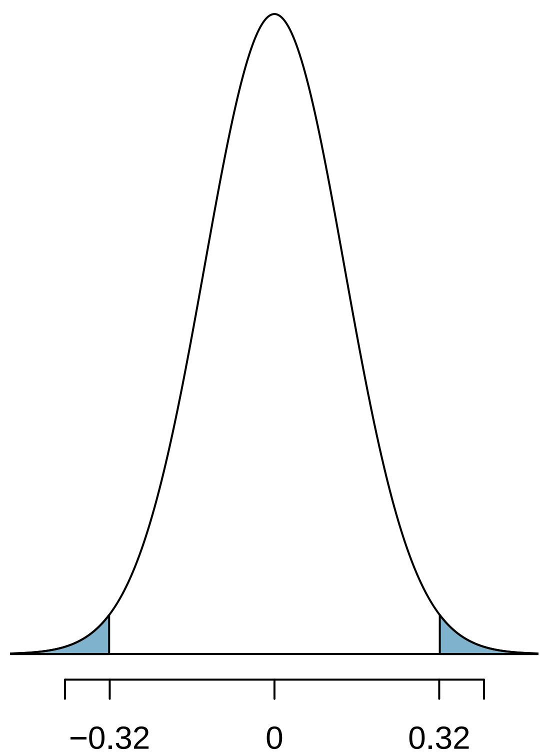
```r
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

```
## 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
```



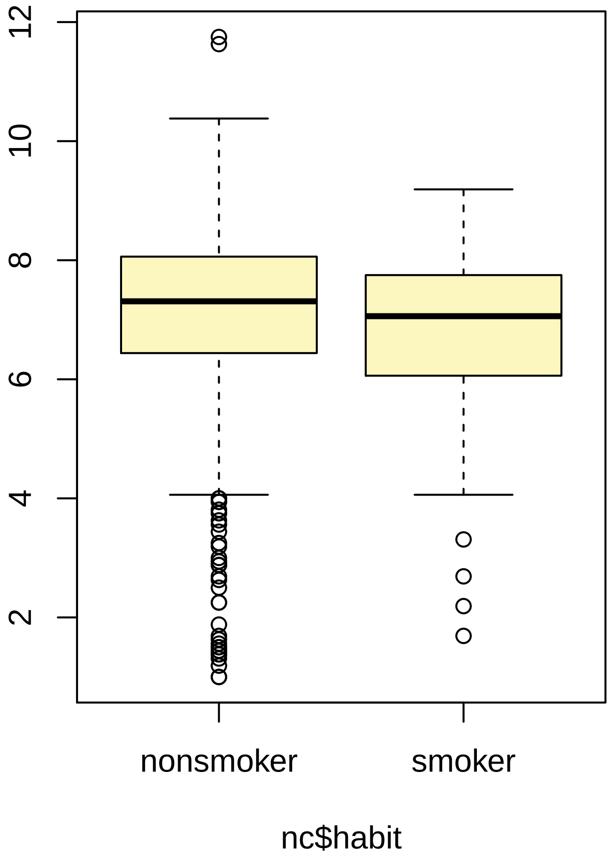


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

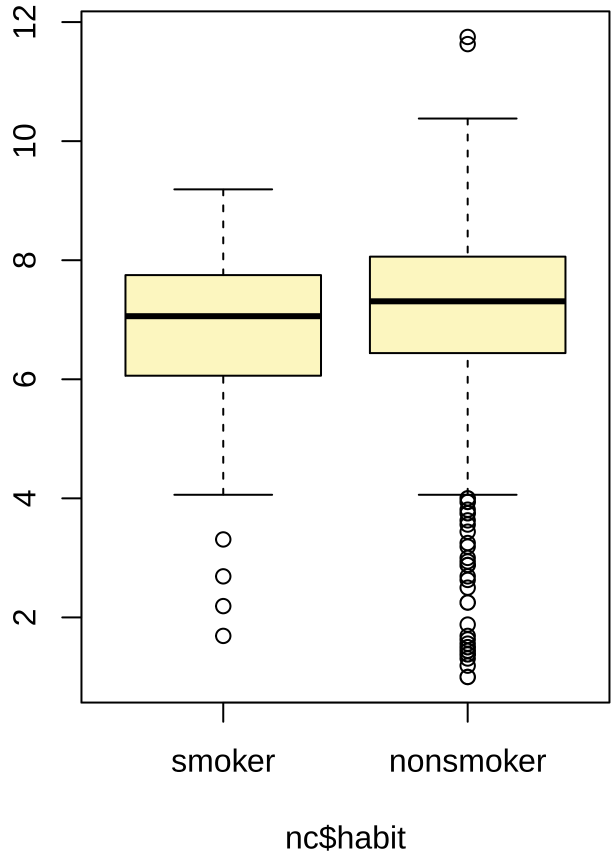



```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

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
```



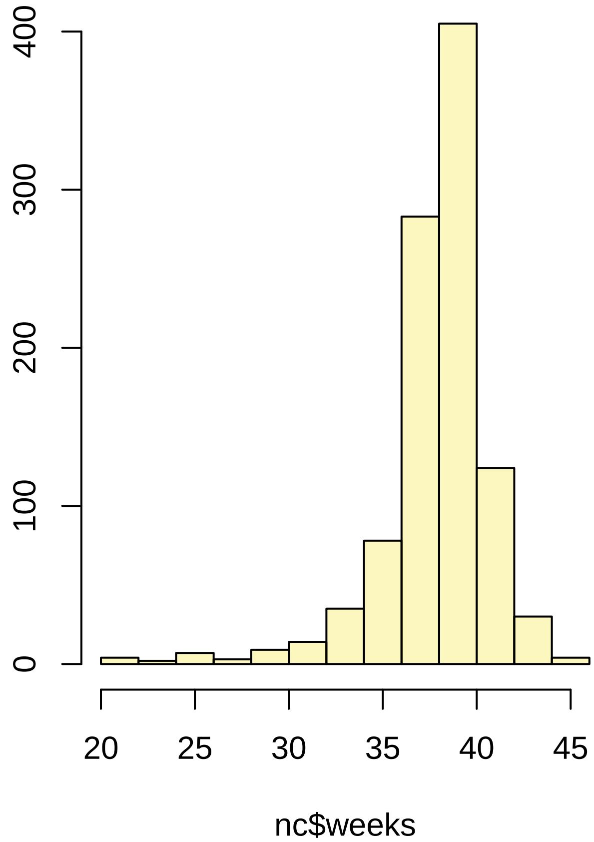

```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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:
```

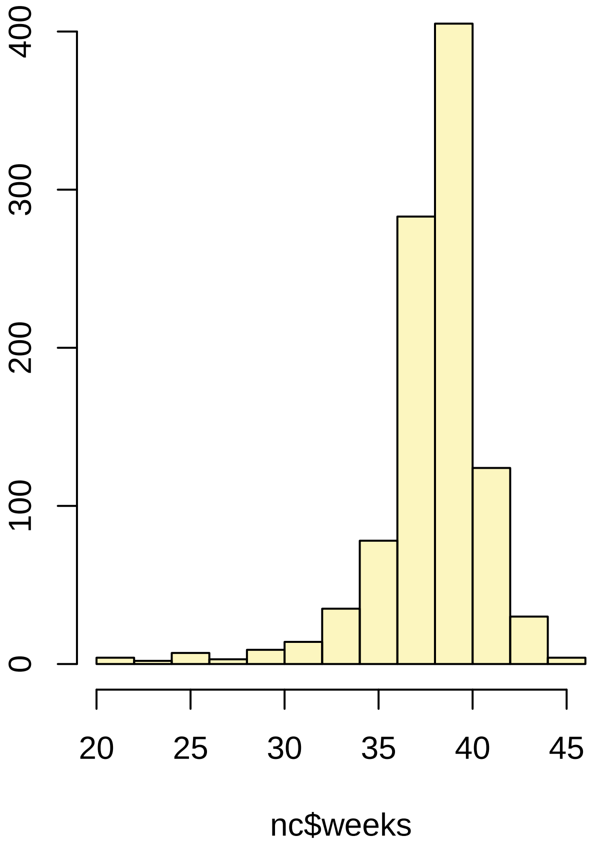



```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

- 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 = .9)
```

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



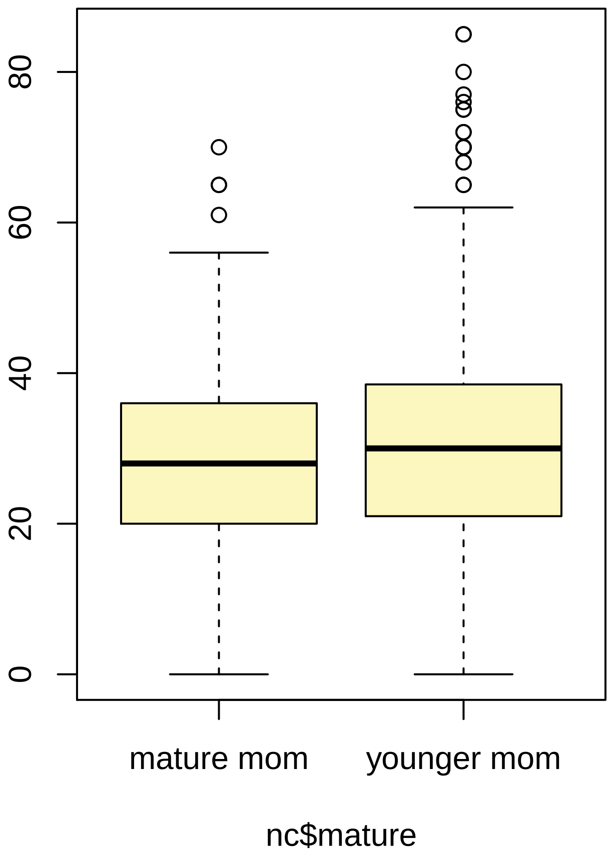

```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 90 % Confidence interval = ( 38.182 , 38.4873 )
```

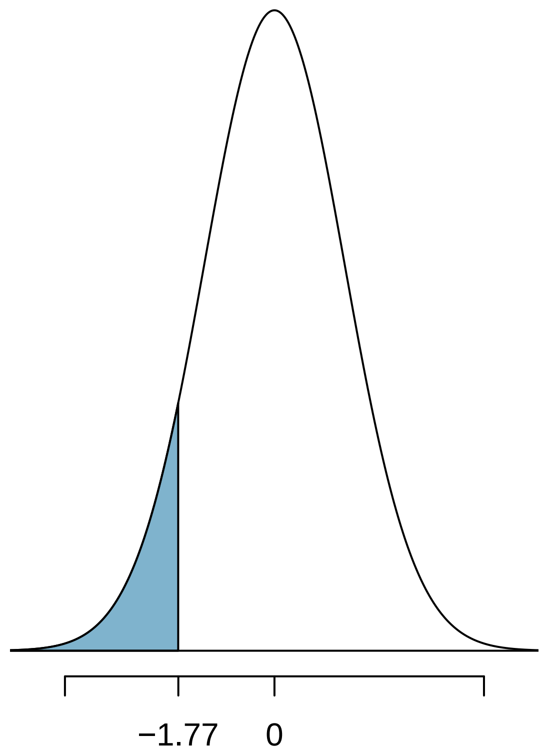
- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
inference(y = nc$gained, x = nc$mature, est = "mean", type = "ht", null = 0,
          alternative = "less", 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 =  0.0843
```



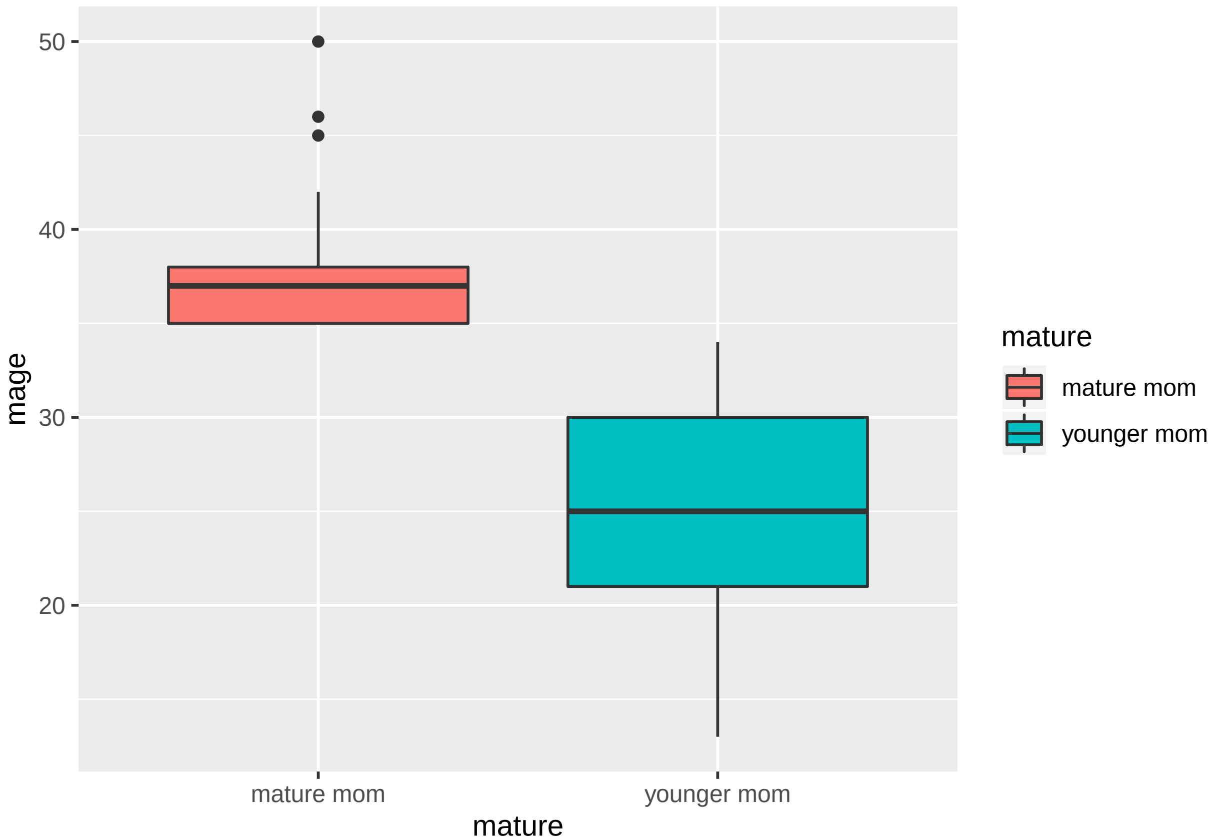


Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

```r
by(nc$mage, nc$mature, summary)
```

```
## nc$mature: mature mom
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   35.00   37.00   37.18   38.00   50.00
## --------------------------------------------------------
## nc$mature: younger mom
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   21.00   25.00   25.44   30.00   34.00
```

```r
ggplot(nc, aes(x = mature, y = mage, fill = mature)) +
  geom_boxplot()
```



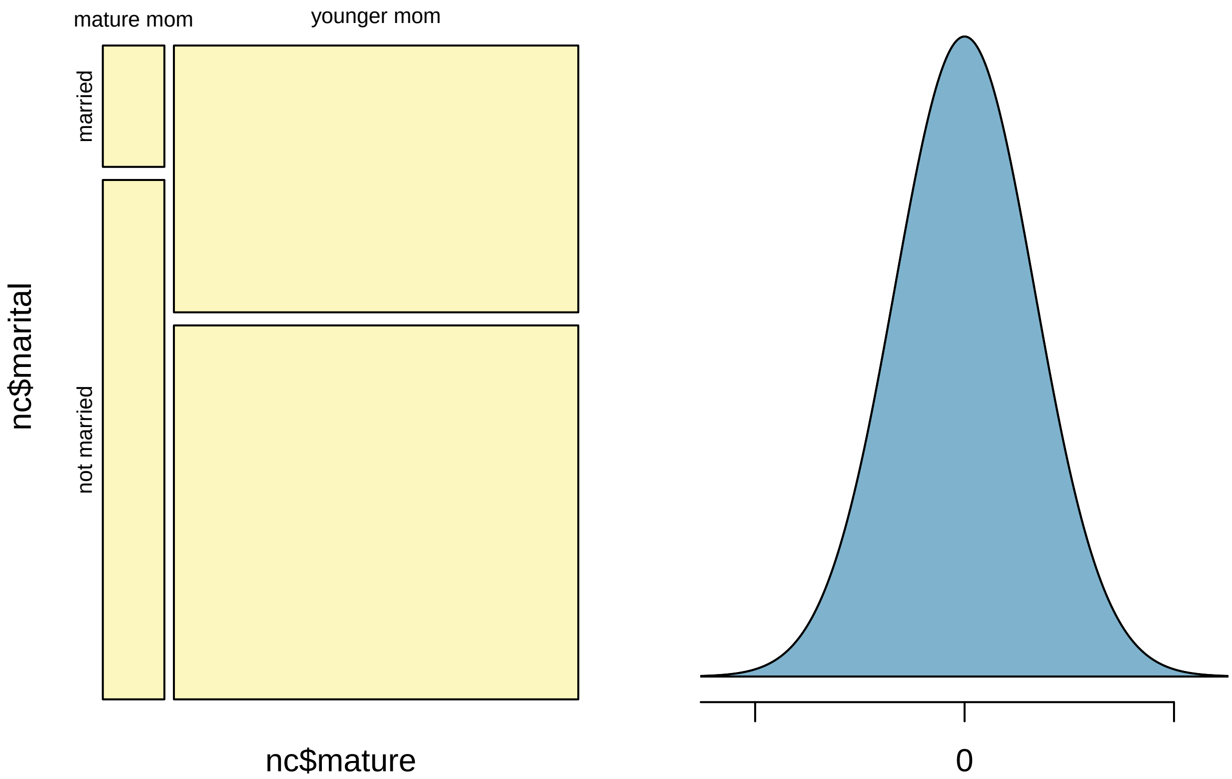
Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the inference function, report the statistical results, and also provide an explanation in plain language. H0: The ratio of mature mothers who are married is not different than the ones who are younger HA: The ratio of mature mothers who are married is greater than the ones who are younger

```r
inference(y = nc$marital, x = nc$mature, est = "proportion", type = "ht", null = 0,
          alternative = "greater", method = "theoretical", success = "married")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: married
## Summary statistics:
##                x
## y               mature mom younger mom Sum
##   married               25         361 386
##   not married          107         506 613
##   Sum                  132         867 999

## Observed difference between proportions (mature mom-younger mom) = -0.227
##
## H0: p_mature mom - p_younger mom = 0
## HA: p_mature mom - p_younger mom > 0
## Pooled proportion = 0.3864
## Check conditions:
##    mature mom : number of expected successes = 51 ; number of expected failures = 81
```

```
##    younger mom : number of expected successes = 335 ; number of expected failures = 532
## Standard error = 0.045
## Test statistic: Z =  -4.989
## p-value =  1
```



- 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.

Let us consider mother's marital status and number of hispital visits per pregnancy and see if there is any difference between married and not married mothers when it comes to the average number of hospital visits.

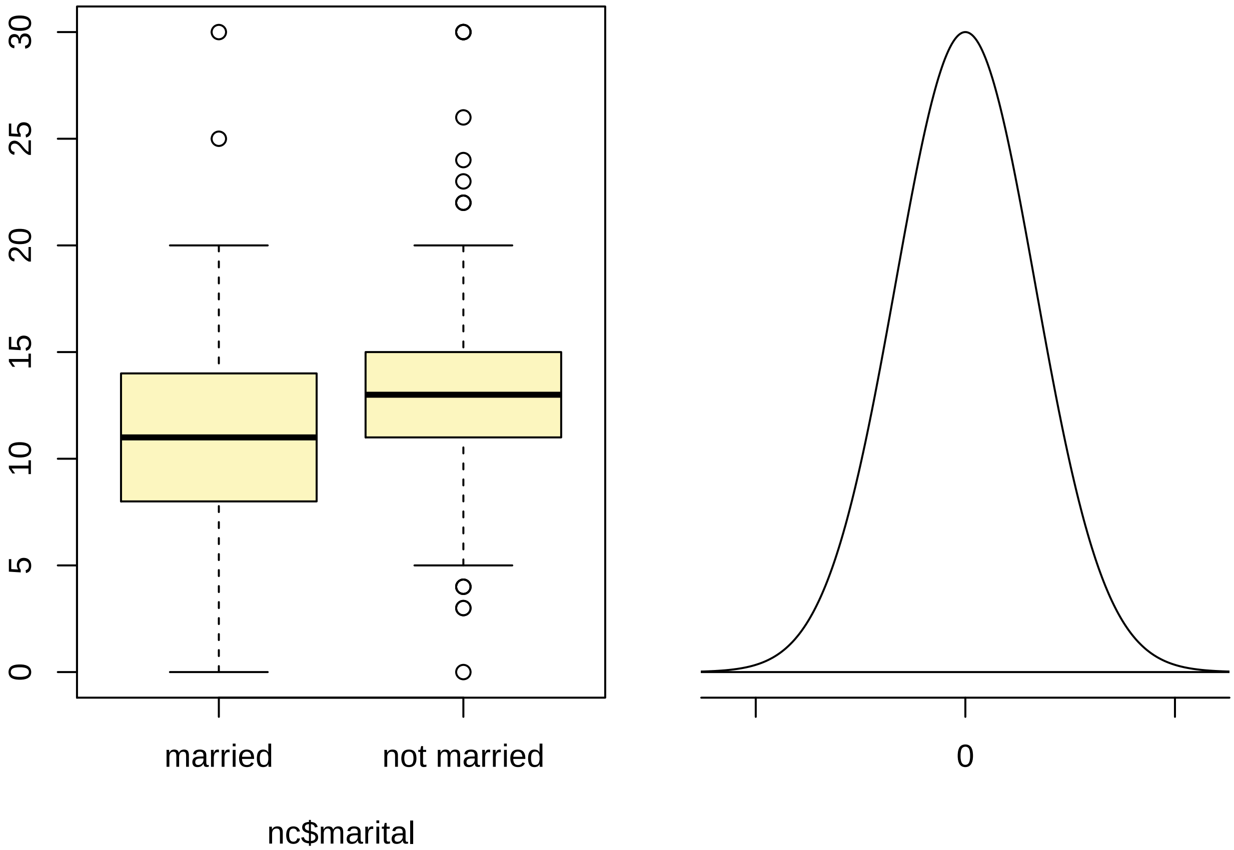
H0 : Average numbers of visits are the same for married mothers and not married mothers HA : Average numbers of visits are different.

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

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_married = 380, mean_married = 10.9553, sd_married = 4.2408
## n_not married = 611, mean_not married = 12.82, sd_not married = 3.5883

## Observed difference between means (married-not married) = -1.8647
##
## H0: mu_married - mu_not married = 0
```

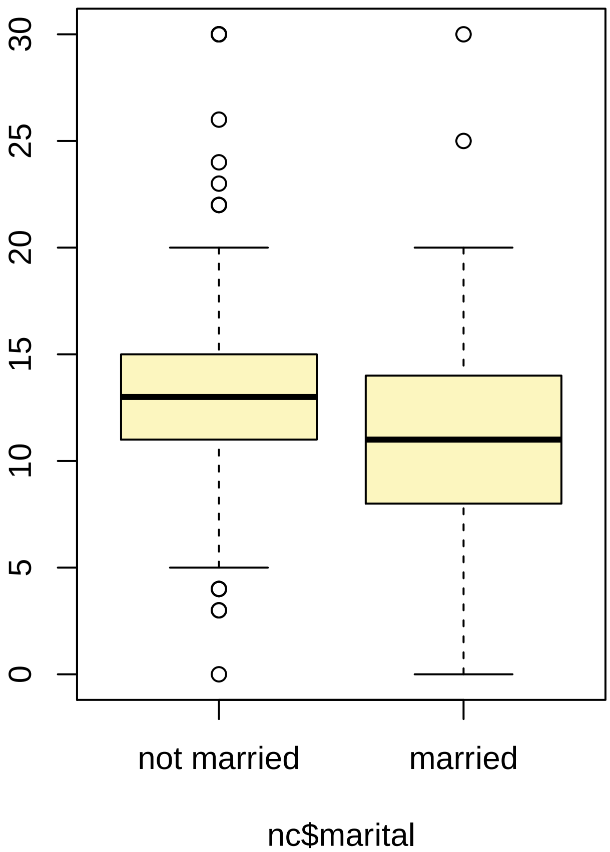
```
## HA: mu_married - mu_not married != 0
## Standard error = 0.262
## Test statistic: Z =  -7.13
## p-value =  0
```



The p-value is practically 0, so we reject the null hypothesis. The difference in number of visits between married and not married mothers is not due to chance.

```
inference(y = nc$visits, x = nc$marital, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("not married", "married"))
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_not married = 611, mean_not married = 12.82, sd_not married = 3.5883
## n_married = 380, mean_married = 10.9553, sd_married = 4.2408
```




```
## Observed difference between means (not married-married) = 1.8647
##
## Standard error = 0.2615
## 95 % Confidence interval = ( 1.3521 , 2.3773 )
```

We are 95% confident that the population average difference between number of hospital visits for married mothers and not married mothers is between 1.3521 and 2.3773 visits. Perhaps, having extra support from a spouse at home lowers the need for hospital visits by about 2 visits on average, but there may be other explanations.