Inference for numerical data

Jai Jeffryes

10/25/2019

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.

```r
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?

ANSWER

There's a little subtlety to this question. The cases are births in North Carolina. At first, I wanted to say the cases were mothers. Then I realized there were data about the babies, so I wanted to say the babies were the case. However, there were those variables about the mothers I saw first, and now I saw there were data about the fathers and the relationship between the parents. Therefore, really the observations are about everyone party to marginal population increase! The cases are births in NC.

There are 1000 observations 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:

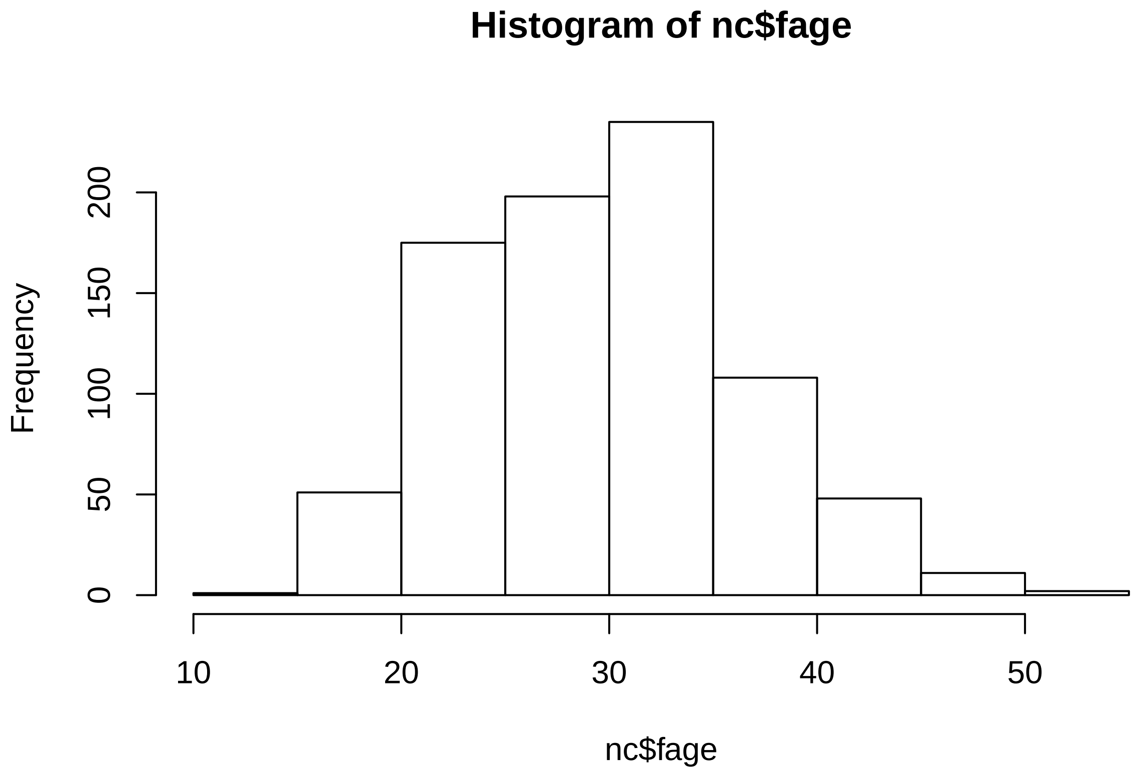
```
summary(nc)
```

```
##       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
##
##
##
##
```

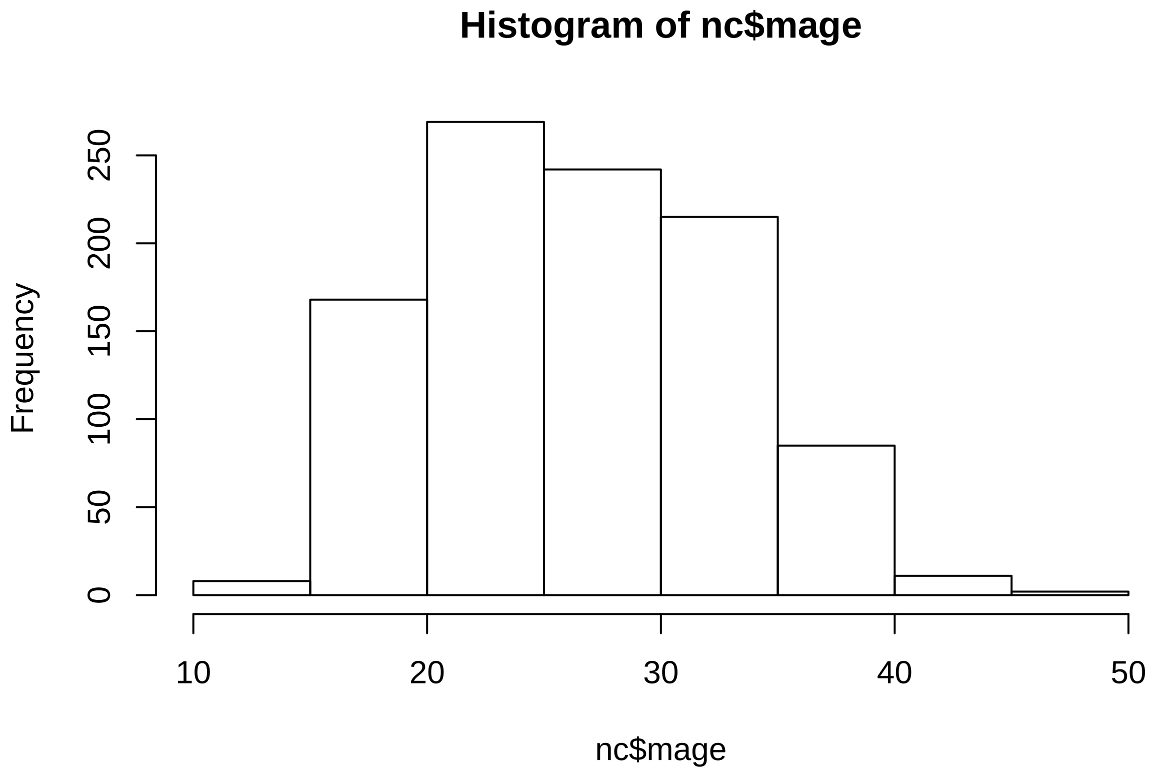
As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

ANSWER

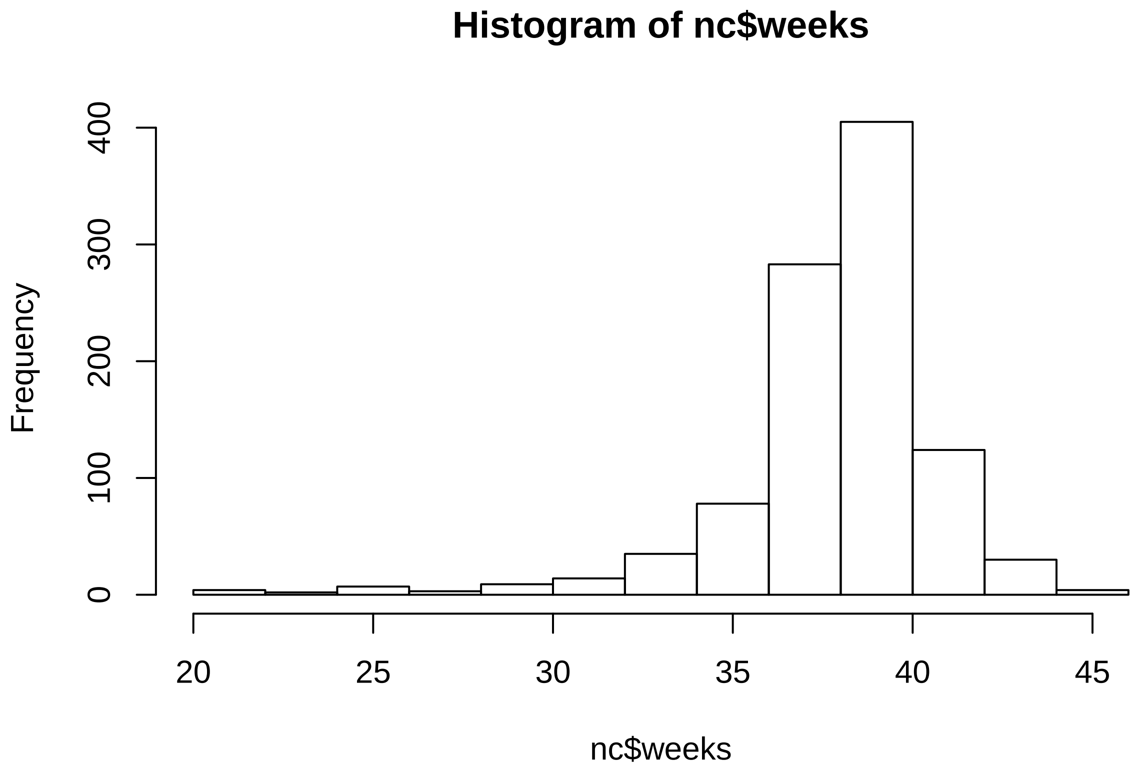
```
hist(nc$fage)
```


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

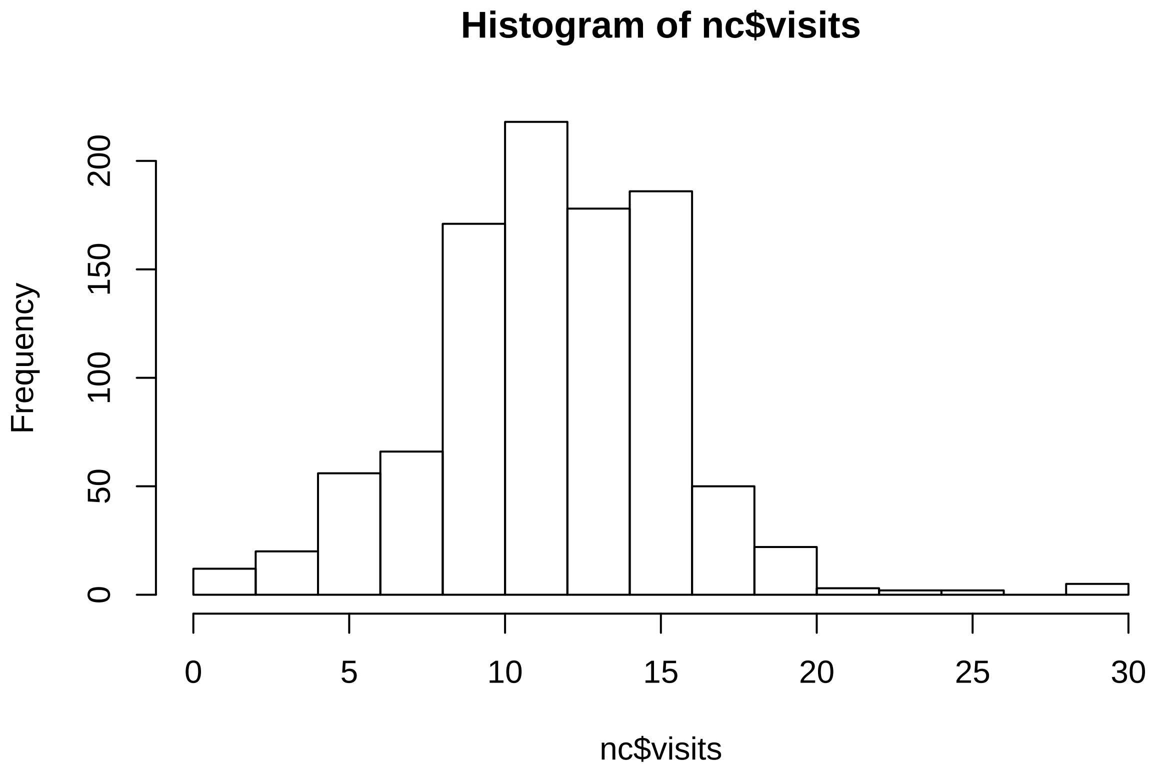
```
hist(nc$mage)
```



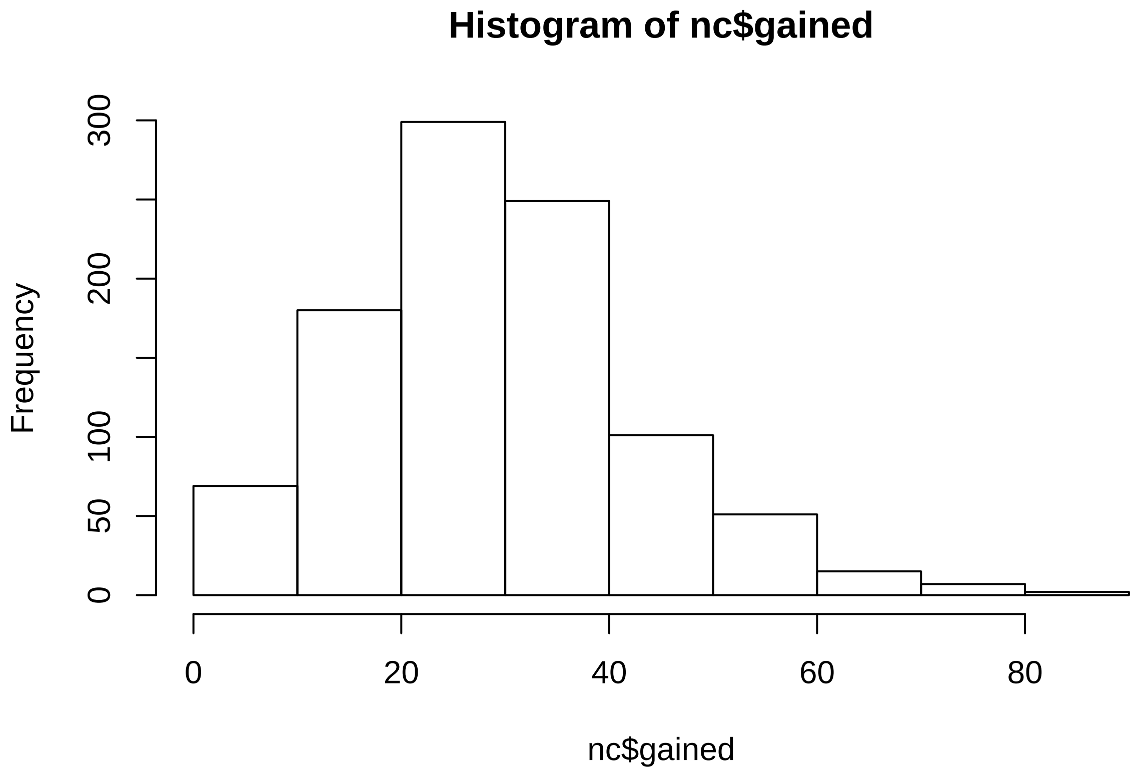
```r
hist(nc$weeks)
```



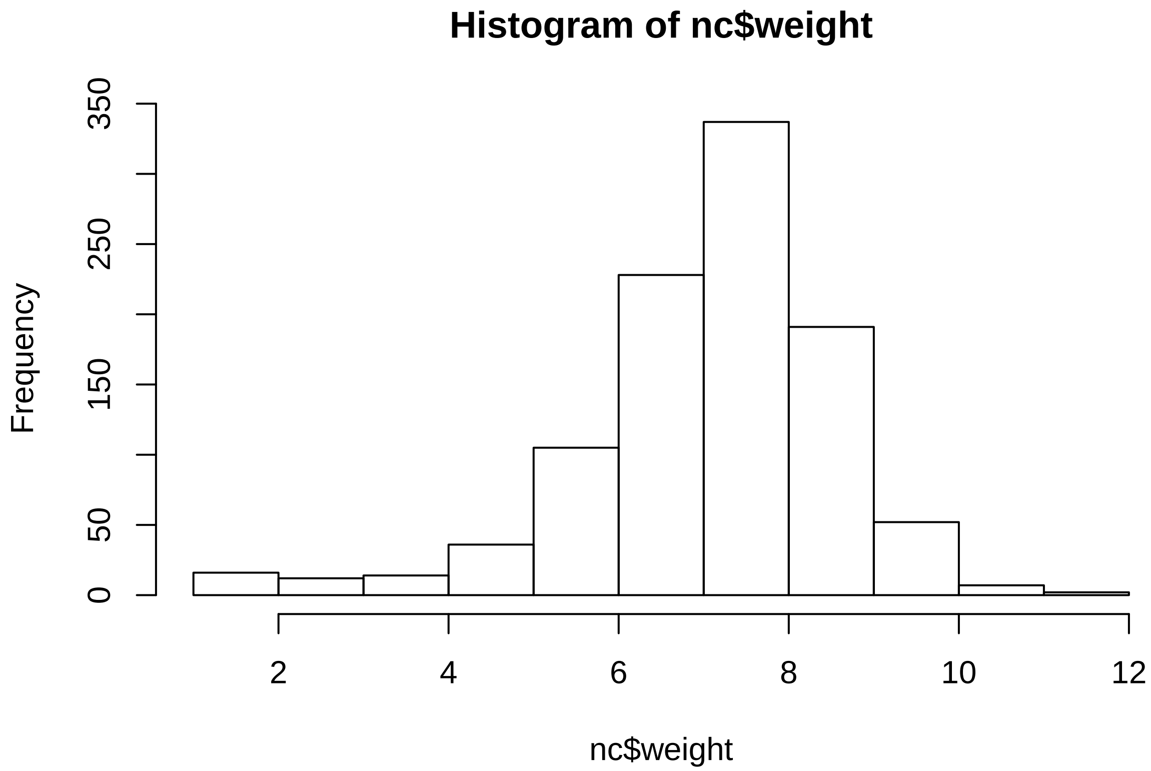
```
hist(nc$visits)
```



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



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



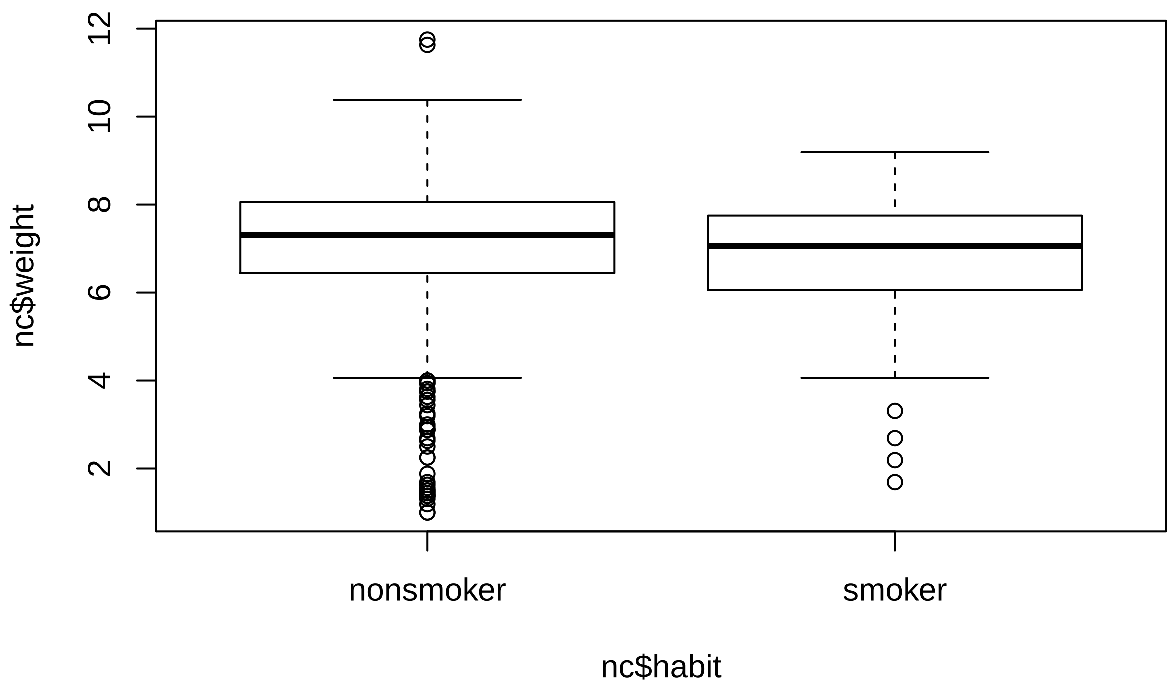
One numerical variable, visits, has some upper outliers.

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?

ANSWER

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



Weight may be a dependent variable. Smoking may be an explanatory variable

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`.

ANSWER

We want at least 30 samples in each group. The more above 30 we have, the more forgiving we can be for outliers. The variable, weight, is skewed to the left, but it appears okay since even in the smoker group we have 4 times the samples we need.

```
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.

ANSWER

H_0: The mean weight of babies born to mothers who smoke is the same as the mean weight of babies born to mothers who do not smoke.

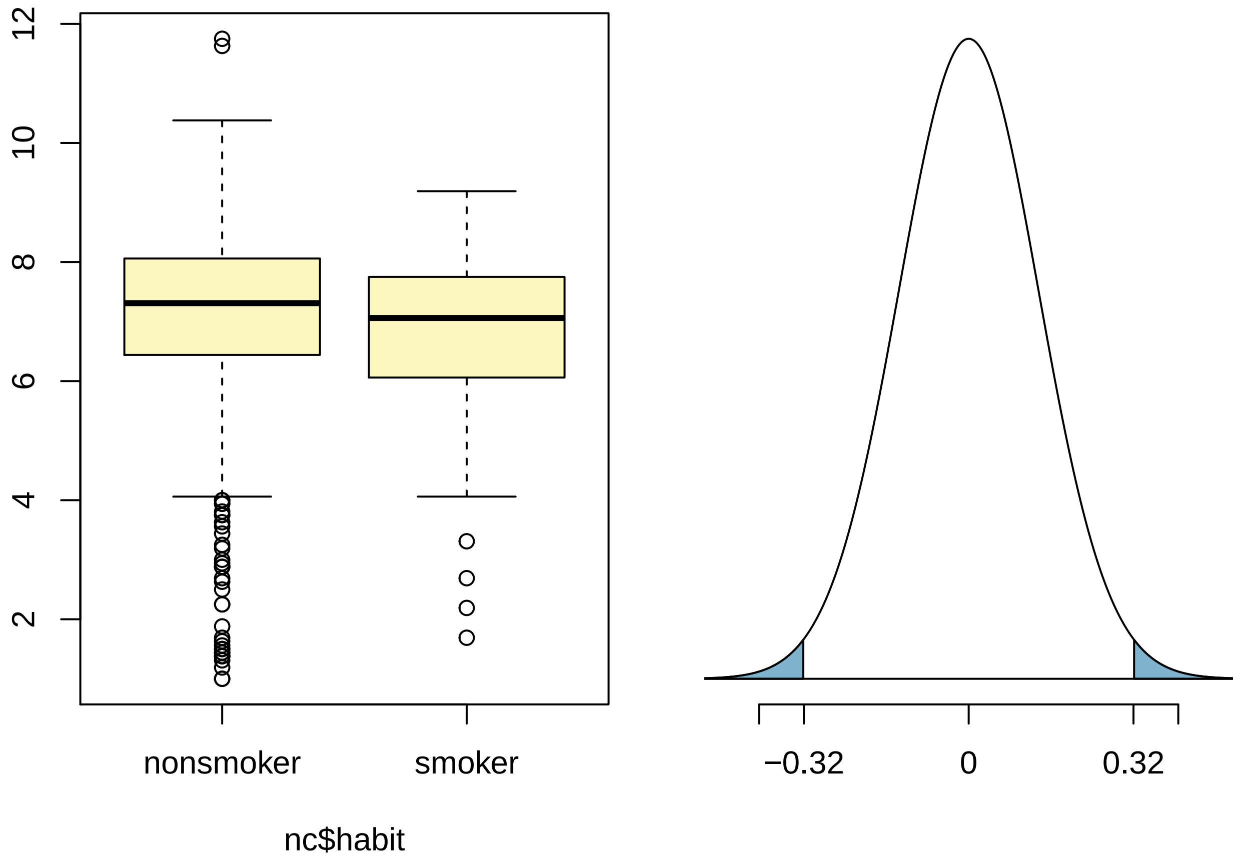
H_A: The mean weight of babies born to mothers who smoke differs from the mean weight of babies born to mothers who do not smoke.

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
```

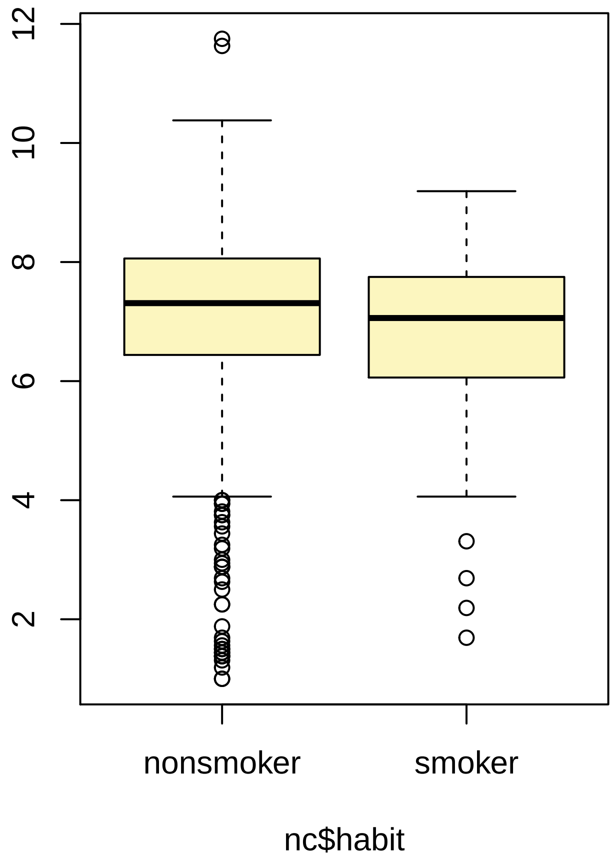


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

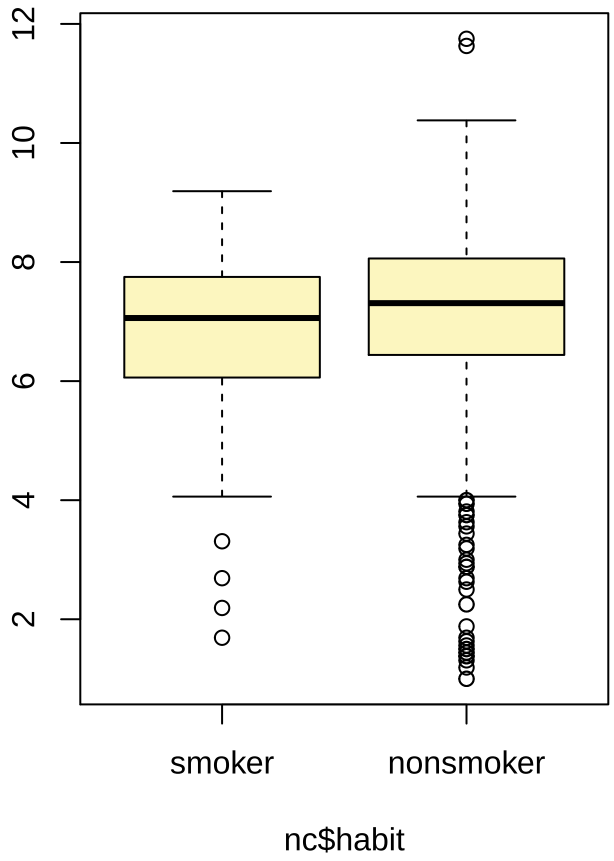


```
## Observed difference between means (nonsmoker-smoker) = 0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( 0.0534 , 0.5777 )
```

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
```



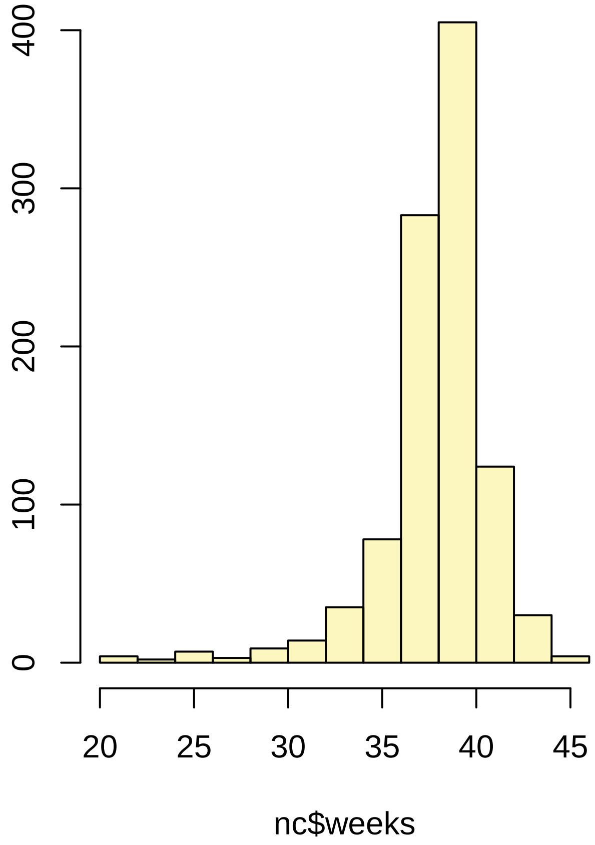
```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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 )
```

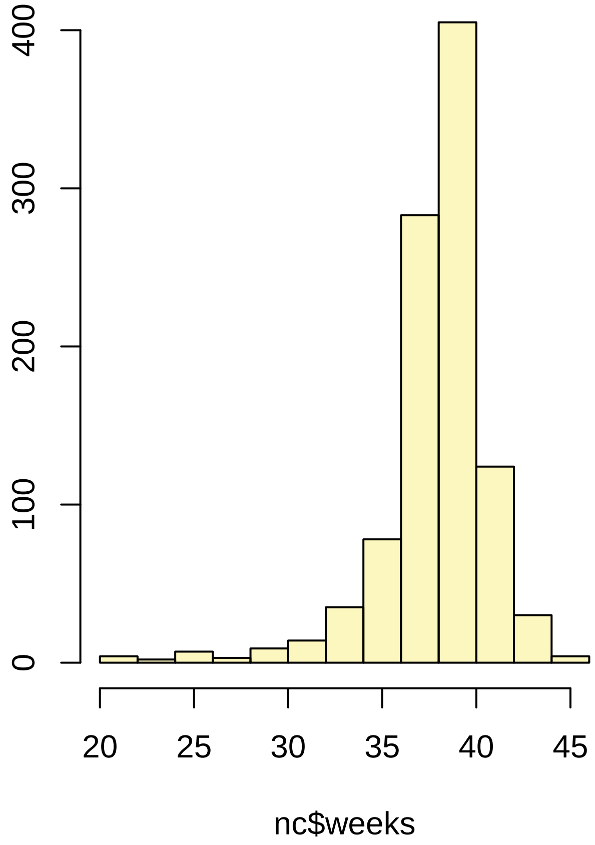
CONCLUSION

We are 95% confident that the population mean of pregnancy length lies within the 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 = 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

At a lower level of confidence, now the confidence interval is narrower, (38.182 , 38.4873). I think it is interesting to observe that the interval isn't all that different, which reflects the narrow spread of this distribution.

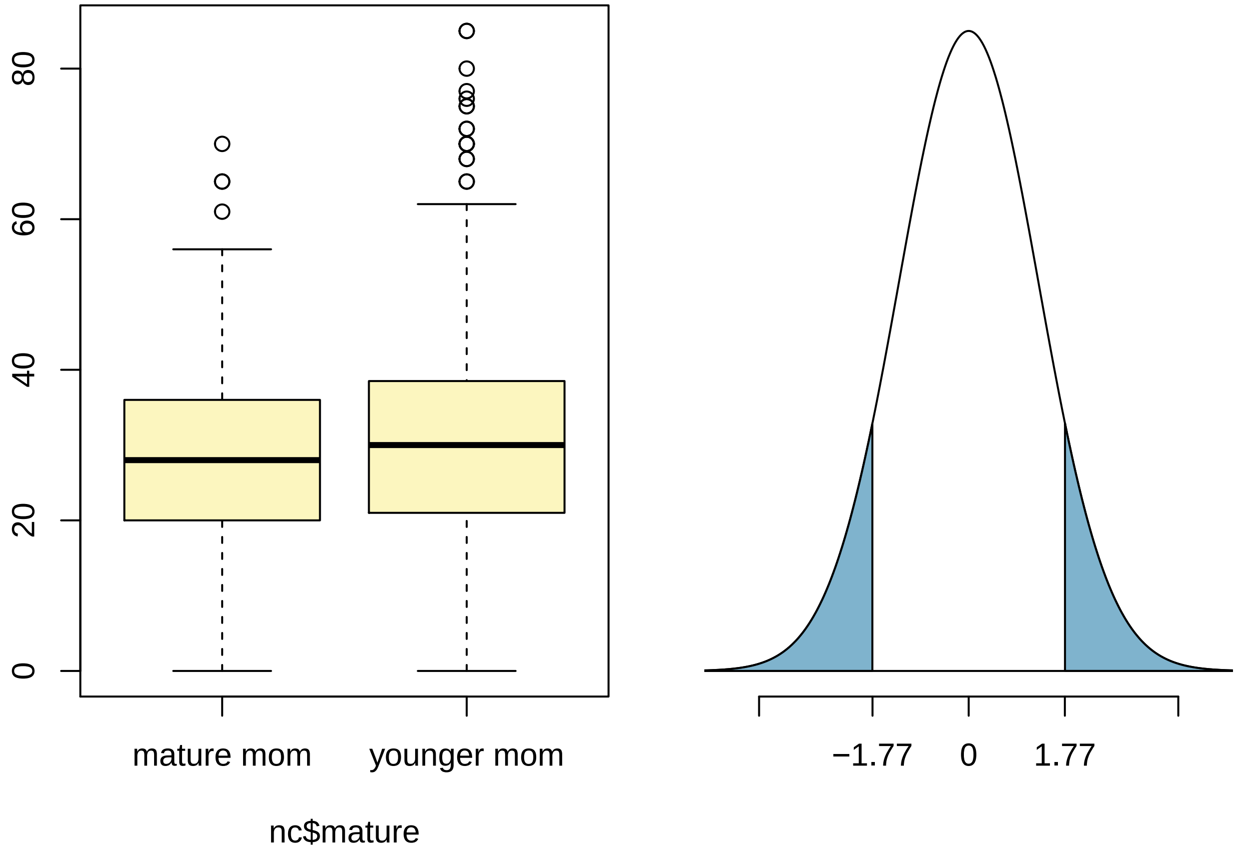
- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
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
```



CONCLUSION

As the p value, at 0.1686, exceeds 0.05, we do not reject the null hypothesis. We lack evidence that weight gain of mothers is different for young vs. mature mothers. I really like the plot highlighting the regions for p. It makes it easy to see how often the results could have been due to chance under H_0. I'm going to keep using this package.

- 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.

ANSWER: I love the convenience of that `by()` function. That's a new one on me. Cool take-away from this assignment. The problem question is about the mother's age. I want to partition it on the categorical variable, mature. The statistics I want are the minimum and maximum ages across each partition. The `range()` function buys me both of those at once.

```
by(nc$mage, nc$mature, range)
```

```
## nc$mature: mature mom
## [1] 35 50
## --------------------------------------------------------
## nc$mature: younger mom
## [1] 13 34
```

- 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.

ANSWER

Research question: Does the age of mothers depend on marital status? Really, I'm thinking about it the other way around. I want to go back and read up thoughtfully again on the connection between explanatory and response variables. I'm reversing the two in this case in order to make an inference.

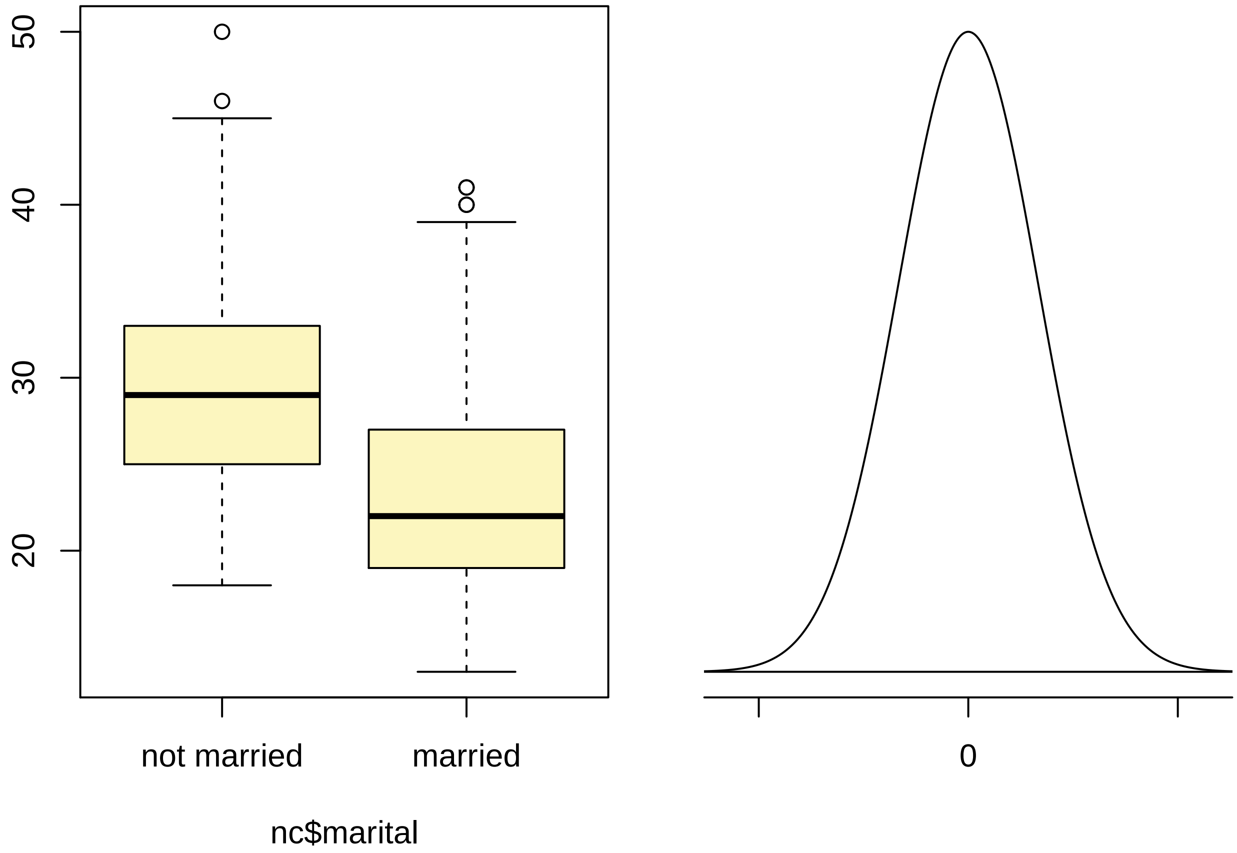
H_0: Mothers' ages do not differ between marital status.

H_A: Mothers' ages do differ between marital status.

```
inference(y = nc$mage, x = nc$marital, est = "mean", type = "ht", 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 = 613, mean_not married = 29.1419, sd_not married = 5.524
## n_married = 386, mean_married = 23.5622, sd_married = 5.658

## Observed difference between means (not married-married) = 5.5797
##
## H0: mu_not married - mu_married = 0
## HA: mu_not married - mu_married != 0
## Standard error = 0.364
## Test statistic: Z =  15.316
## p-value =  0
```



CONCLUSION

The p value is so low that it calculates to 0 within our level of precision. We conclude that mothers' ages differ by marital status. I'm surprised that the older mothers appear to be the ones more likely to be unwed.