Inference for numerical data

Ilya Kats

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.

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?

```r
dim(nc)
```

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

A case is a single birth in the state of North Caroline. There are 1,000 cases.

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                                              
```

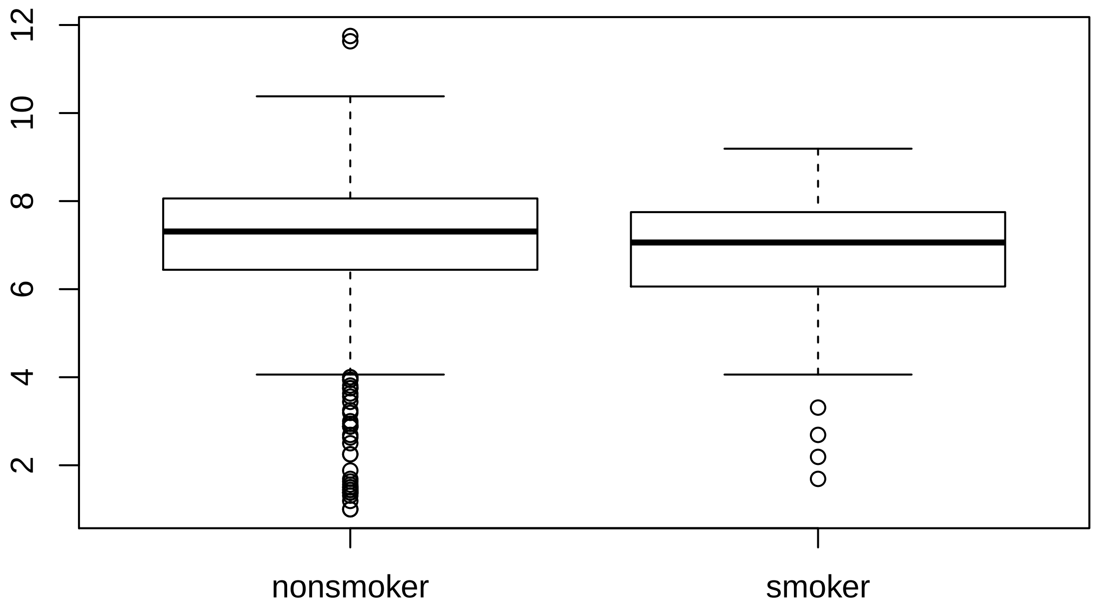
```
##
##        whitemom
##  not white:284
##  white    :714
##  NA's     :  2
##
##
##
##
```

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



The median and IQR are lower for smokers. Even though the lower whisker is at about the same value for both groups (about 4 pounds), the upper whisker is noticably lower for smokers. There are more outliers for non-smokers, but that can probably be explained by the fact that the sample has a lot more data points for non-smokers than for smokers.

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.

```r
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

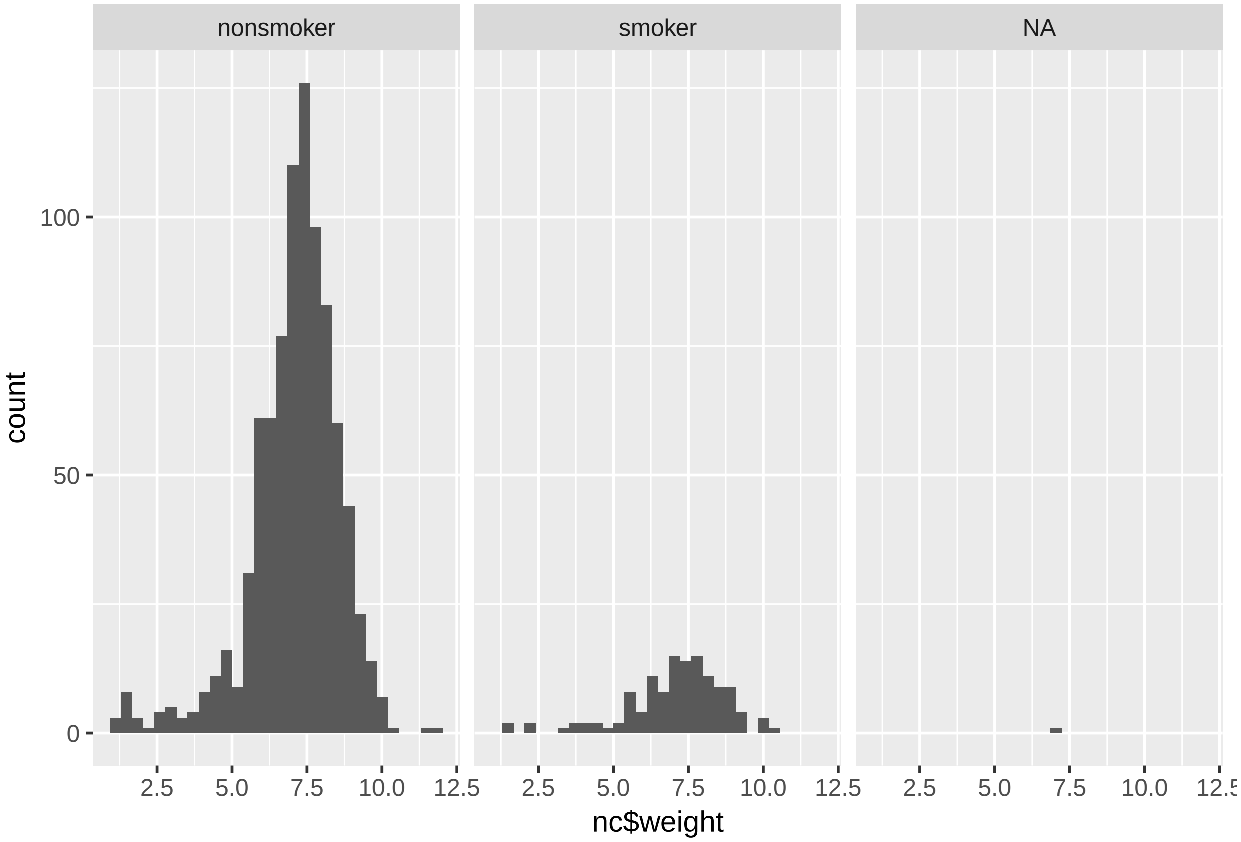
3. Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same `by` command above but replacing `mean` with `length`.

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

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

```r
library(ggplot2)
ggplot(nc,aes(x=nc$weight)) +
  geom_histogram() +
  facet_grid(~nc$habit)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is little information given about how multiple births from the same couple are treated in the data or in the sample. Births from the same couple may not be independent. However, given that the sample is random and that it is safe to assume that it repsents less than 10% of all births in North Carolina, we can consider observations within each group to be independent. Similarly, observations between each group can be considered independent; although, there is an interesting question of a mother who smoked during the first pregnancy and who quit smoking during second one. In each group we have more than 30 observations. And also there is some left skew in both groups, given a fairly large sample size, we can assume normal distribution.

---

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

$H_0 : \mu_{smoker} - \mu_{non-smoker} = 0$ (Average weights of babies born to smoking and non-smoking mothers are the same.)

$H_A : \mu_{smoker} - \mu_{non-smoker} \neq 0$ (Average weights of babies born to smoking and non-smoking mothers are different.)

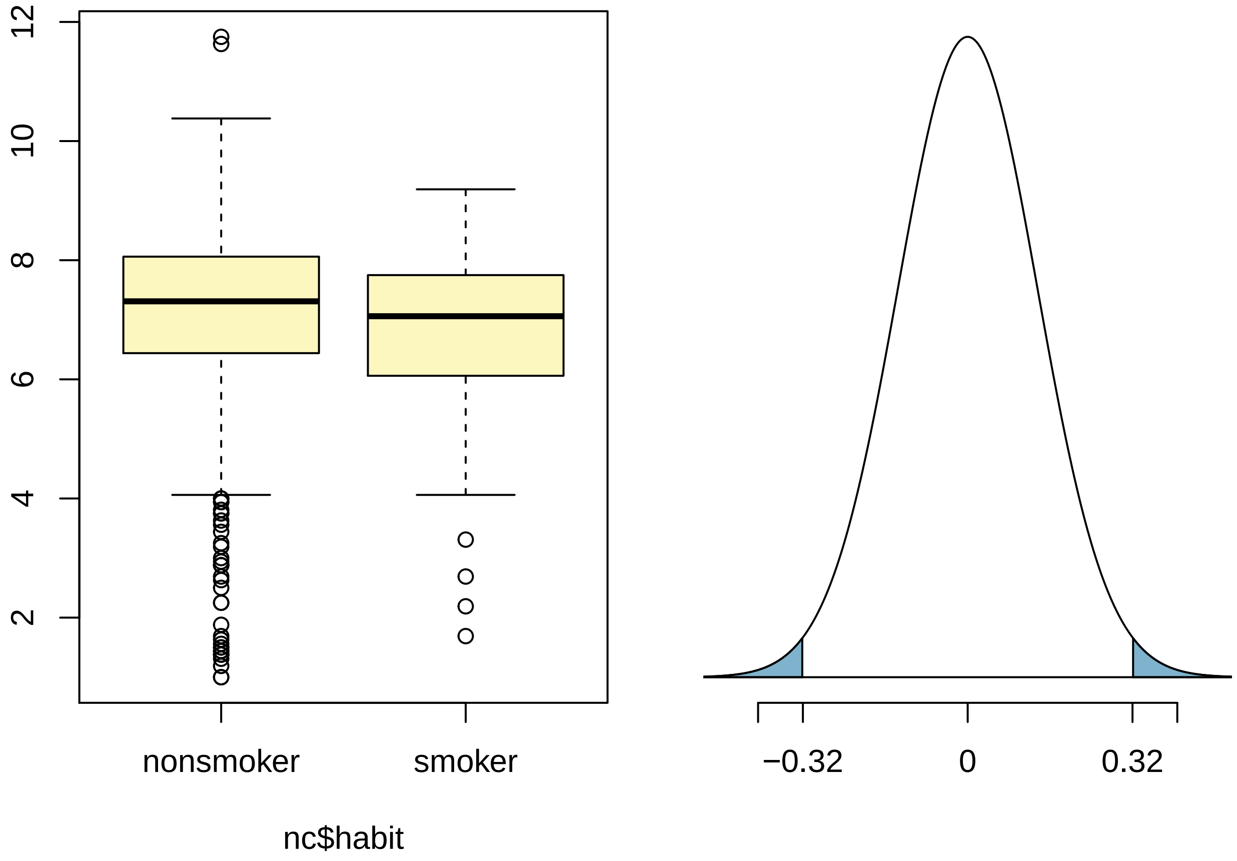
Next, we introduce a new function, `inference`, that we will use for conducting hypothesis tests and constructing confidence intervals.

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

```
## Warning: package 'BHH2' was built under R version 3.3.3

## 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
```

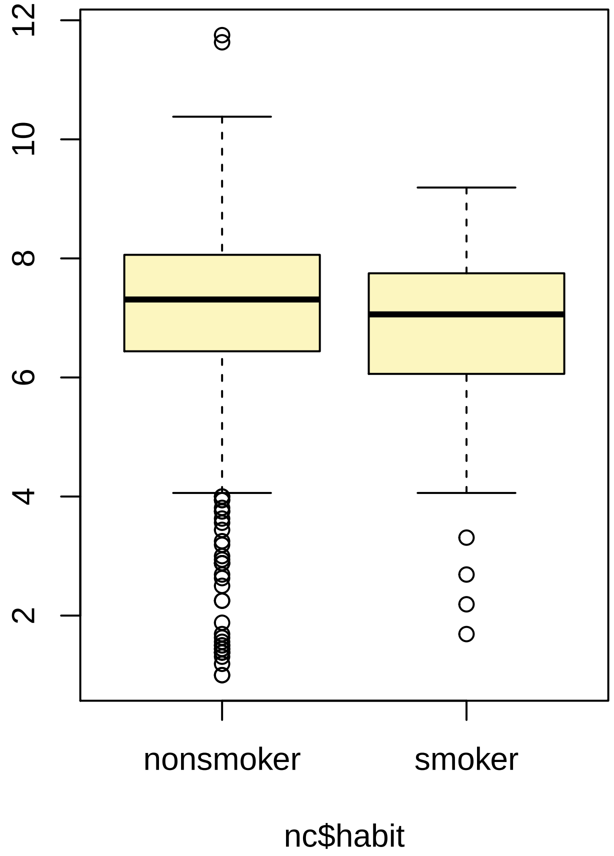


Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

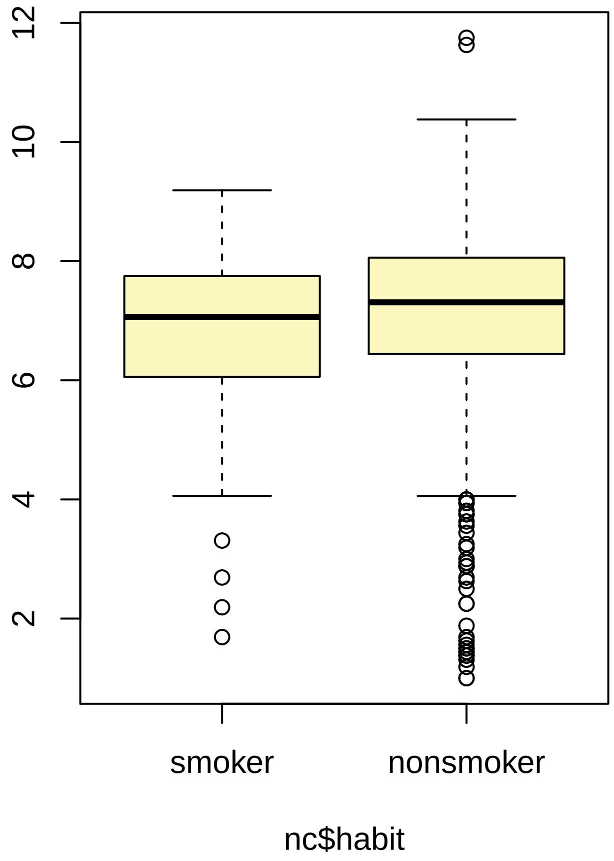
The 95% Confidence interval is (0.0534, 0.5777).

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
```



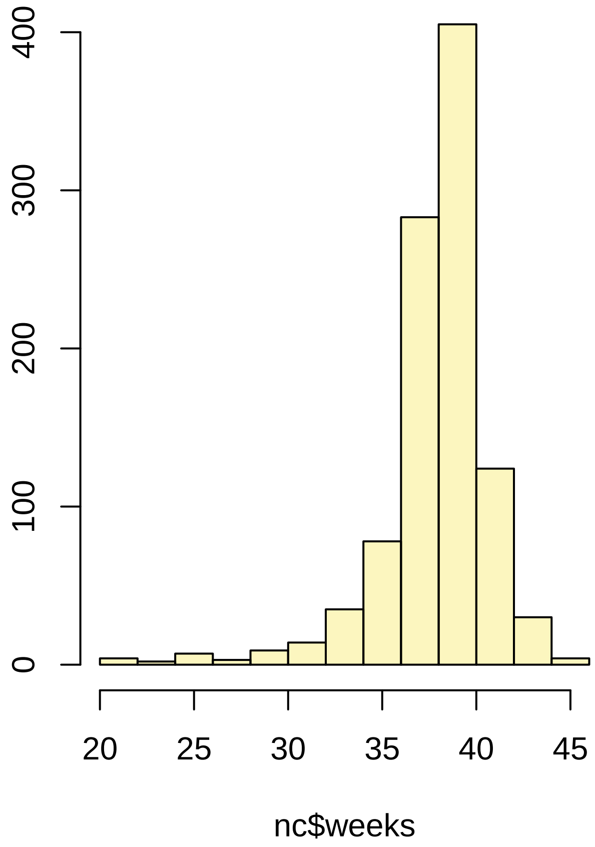
```
## Observed difference between means (smoker-nonsmoker) = -0.3155
##
## Standard error = 0.1338
## 95 % Confidence interval = ( -0.5777 , -0.0534 )
```

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:
```



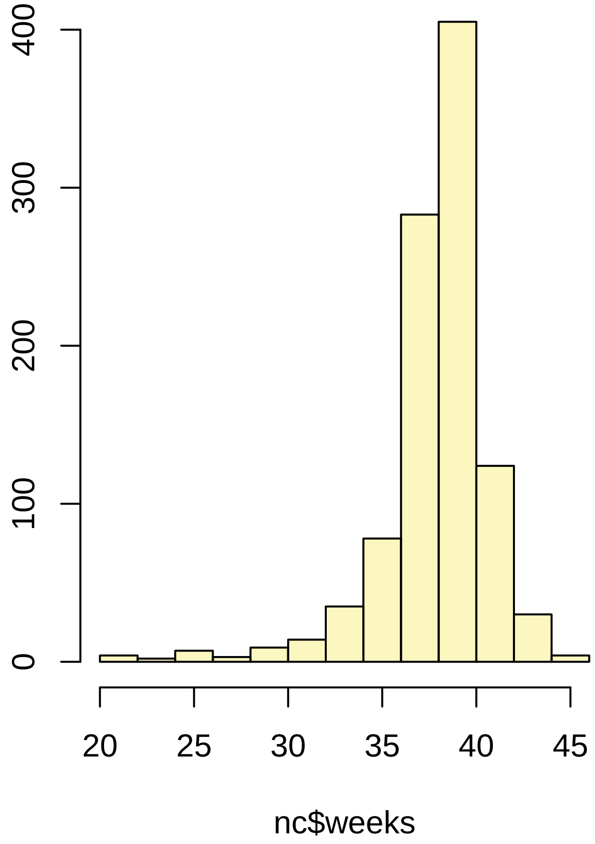
```
## mean = 38.3347 ;  sd = 2.9316 ;  n = 998
## Standard error = 0.0928
## 95 % Confidence interval = ( 38.1528 , 38.5165 )
```

We are 95% confident that the average length of pregnancies for the population is 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", null = 0,
          alternative = "twosided", method = "theoretical",
          conflevel = 0.95)
```

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



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

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

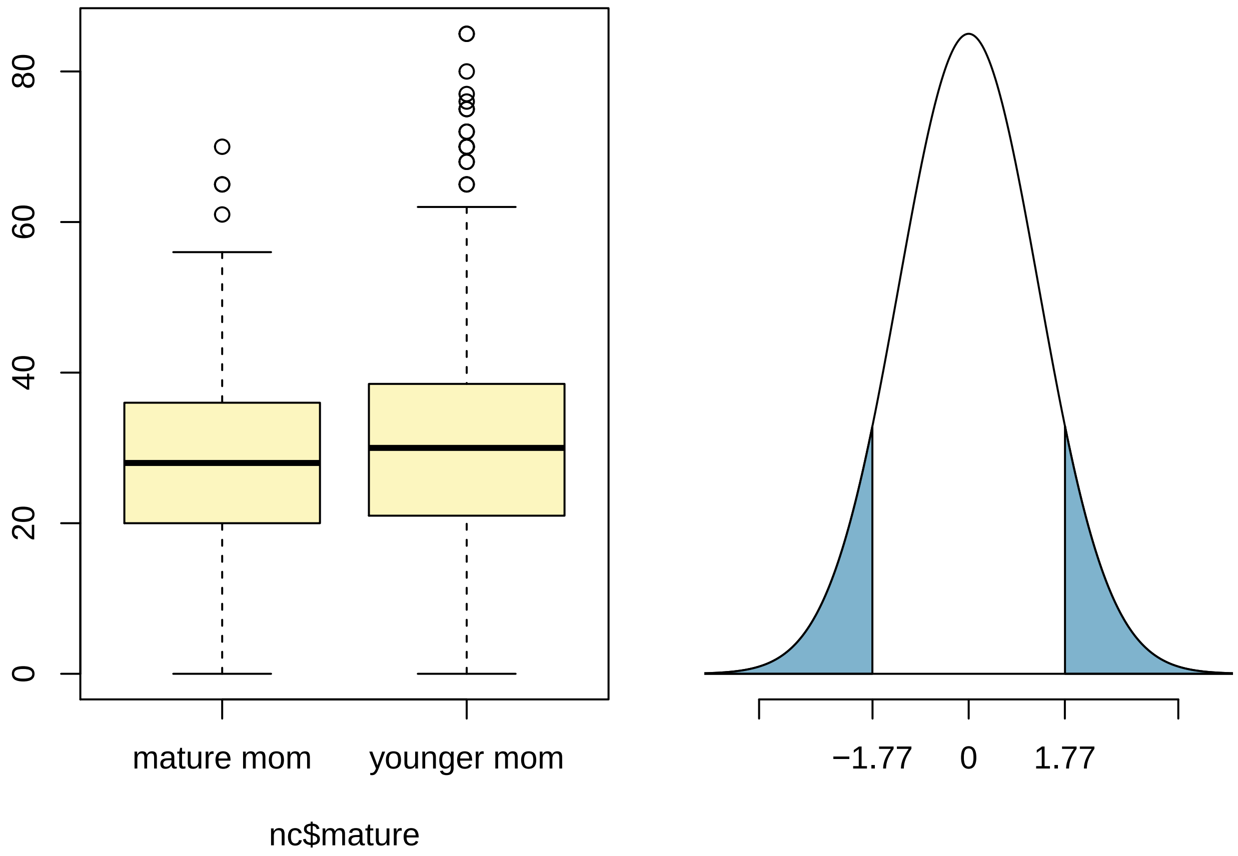
$H_0 : \mu_{mature} - \mu_{younger} = 0$ (Average weights gained by mature mothers and younger mothers are the same.)

$H_A : \mu_{mature} - \mu_{younger} \neq 0$ (Average weights are different.)

```
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
```



Assuming 95% confidence interval, since p-value = 0.1686 > 0.05, we fail to reject H_0.

- 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.

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

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

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

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

Getting minimum and maximum values of mother's age by maturity category, we can see that younger mothers are between 13 and 34 years old while mature mothers are between 35 and 50. The cutoff age is 35.

- 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 hospital 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.

$H_0 : \mu_{married} - \mu_{not\ married} = 0$ (Average numbers of visits are the same for married mothers and not married mothers.)

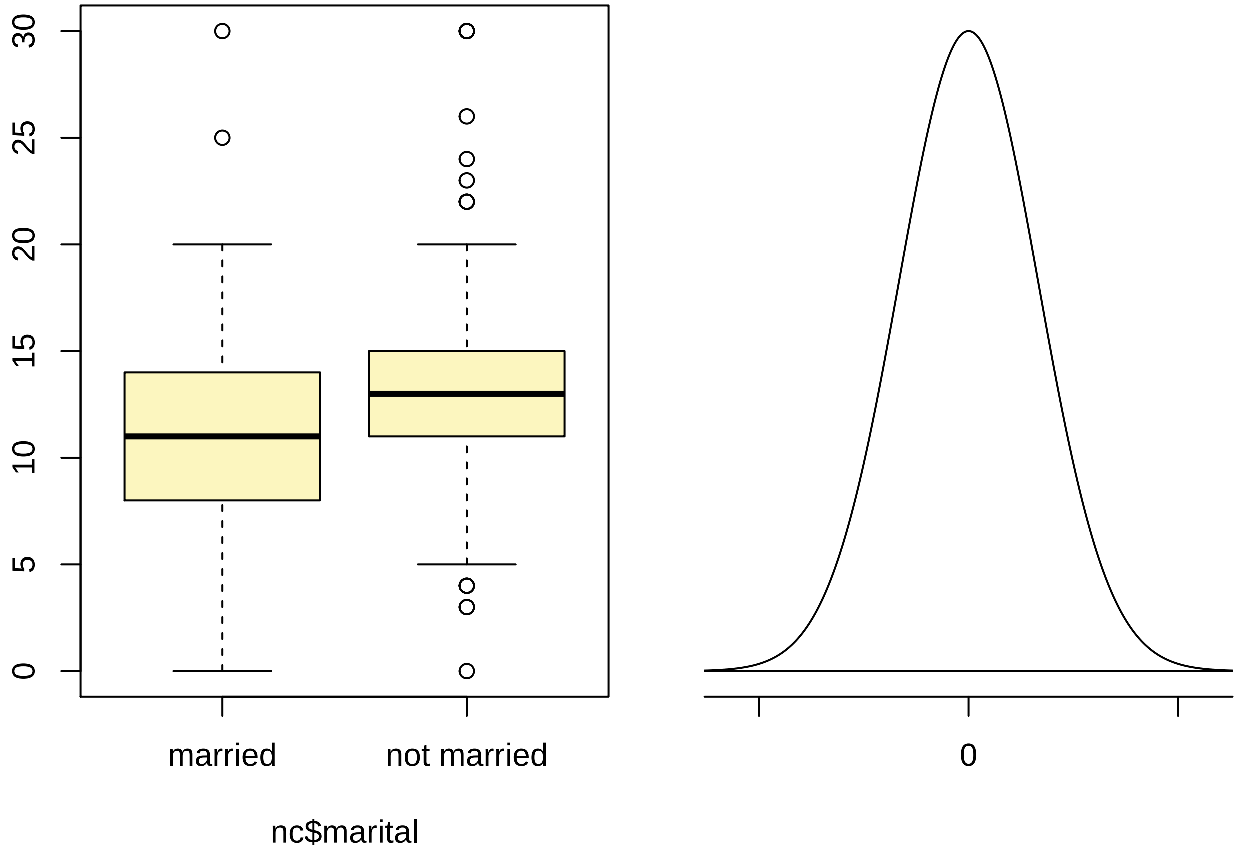
$H_A : \mu_{married} - \mu_{not\ married} \neq 0$ (Average numbers of visits are different.)

Using `inference` function for hypothesis testing.

```
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
```

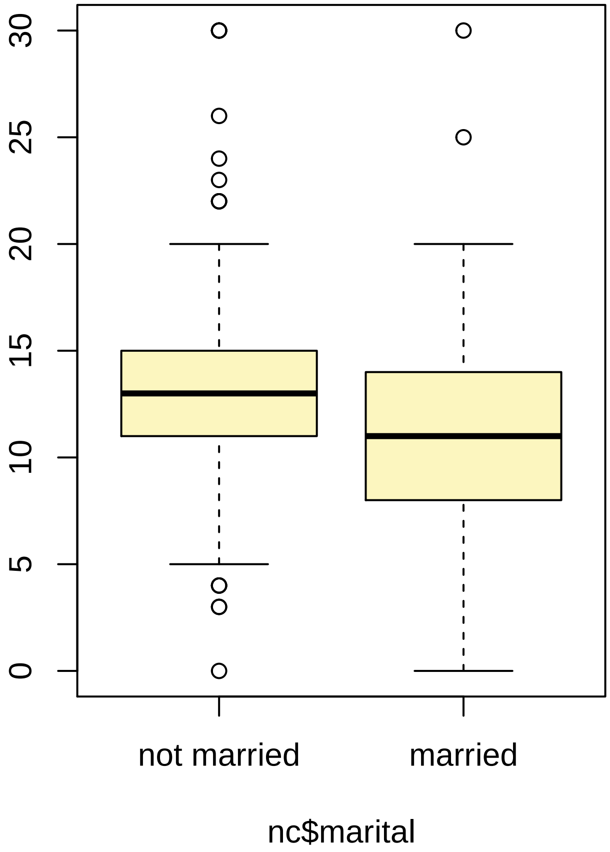


The p-value is practically 0, so we reject the null hypothesis. The difference in number of

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.