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? –vb– Cases are birth of child. there are 1000 in this data set.

As a first step in the analysis, we should consider summaries of the data. This can be done using the `summary` command:

```
summary(nc)
```

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.

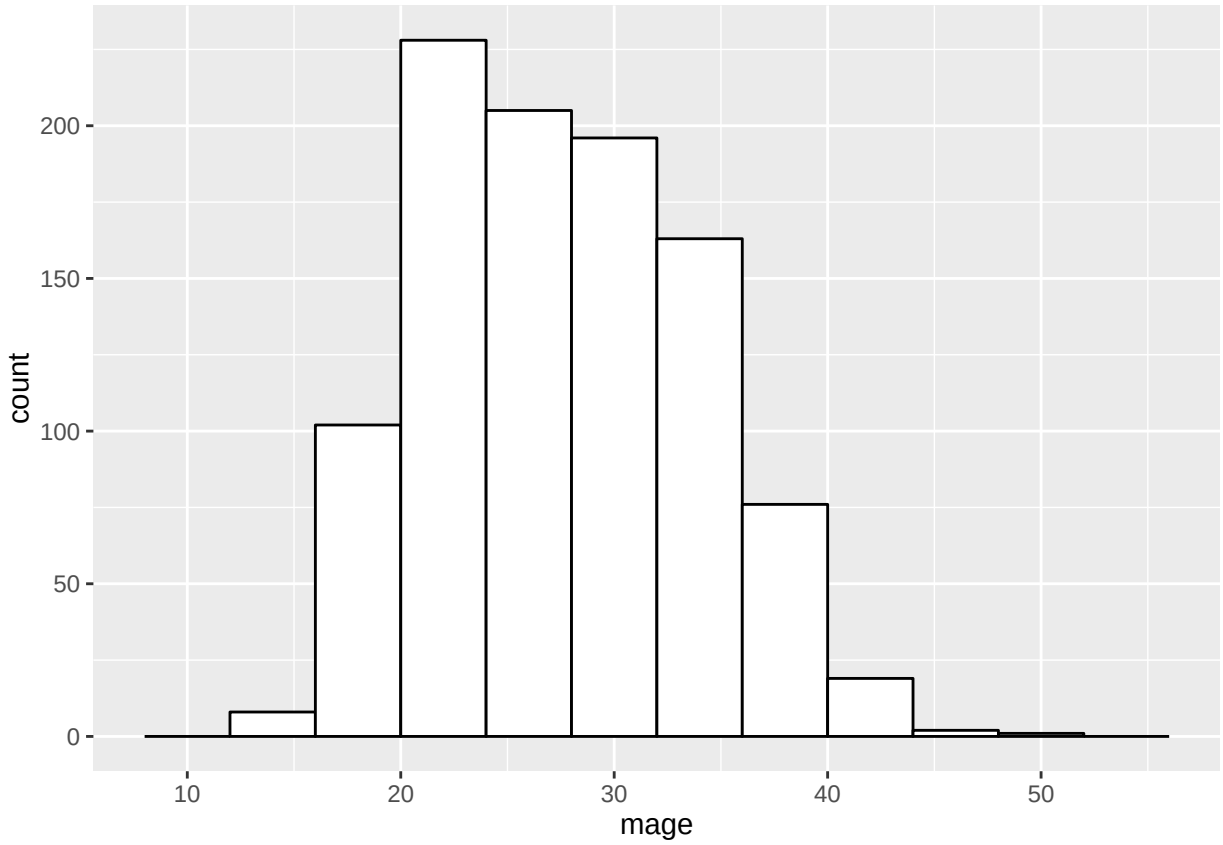
—vb variables:

variable	type
fage	numerical

variable	type
mage	numerical
mature	categorical
weeks	numerical
premie	categorical
visits	numerical, discreet
marital	categorical
gained	numerical
weight	numerical
lowbirthweight	categorical
gender	categorical
habit	categorical
whitemon	categorical

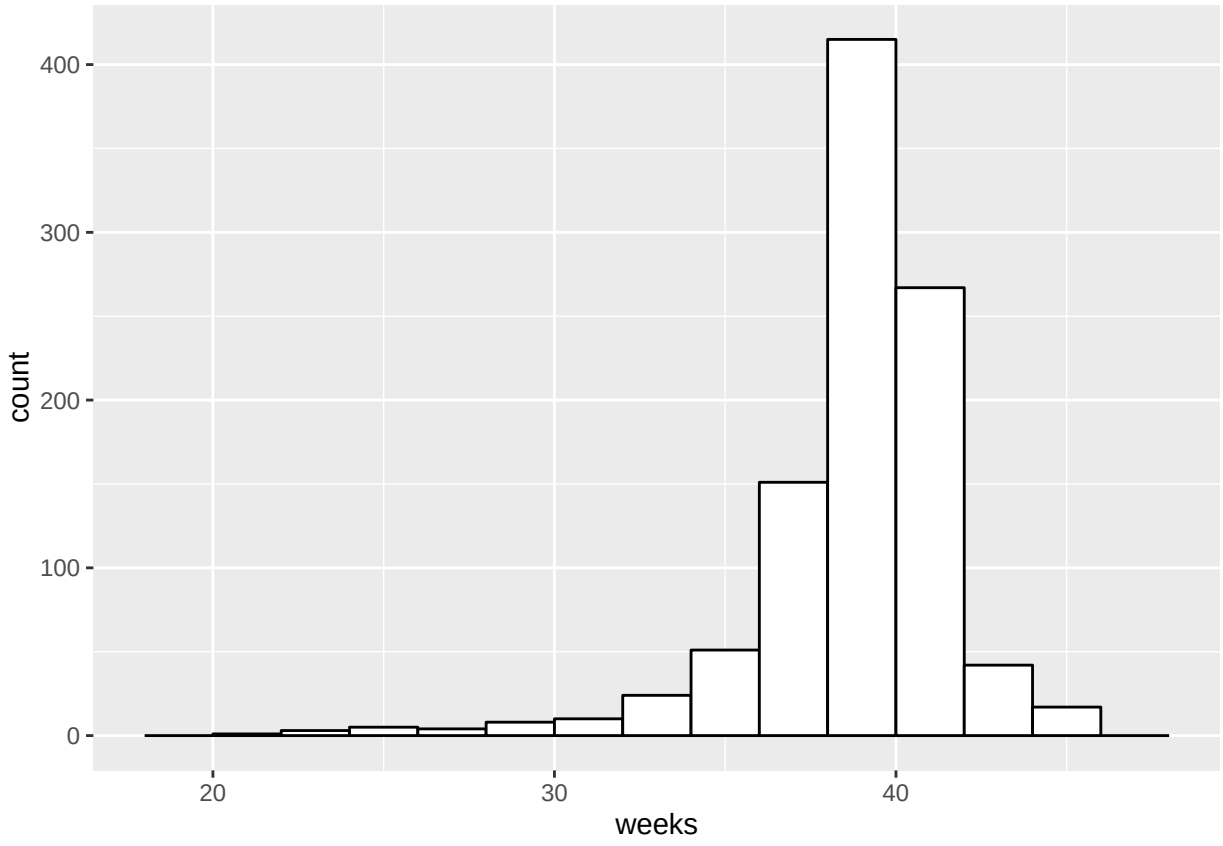
From the summary information, it appears that weeks and weight may have some outliers.
To confirm, we will draw histogram for following variables: fage, weeks, and weight.

```
library(ggplot2)
ggplot(nc, aes(x=mage)) + geom_histogram(binwidth = 4, fill="white", colour = "black")
```

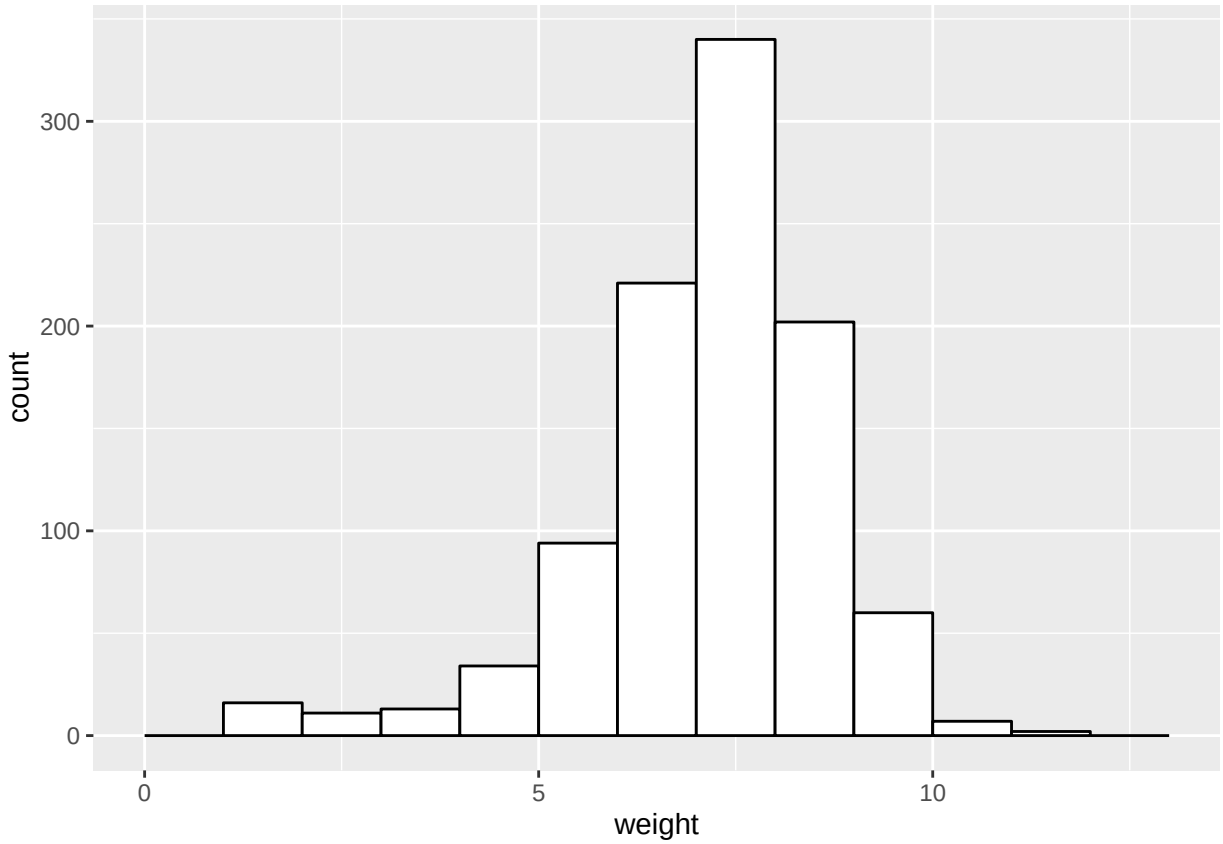



```
ggplot(nc, aes(x=weeks)) + geom_histogram(binwidth = 2, fill="white", colour = "black")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```




```
ggplot(nc, aes(x=weight)) + geom_histogram(binwidth = 1, fill="white", colour = "black")
```



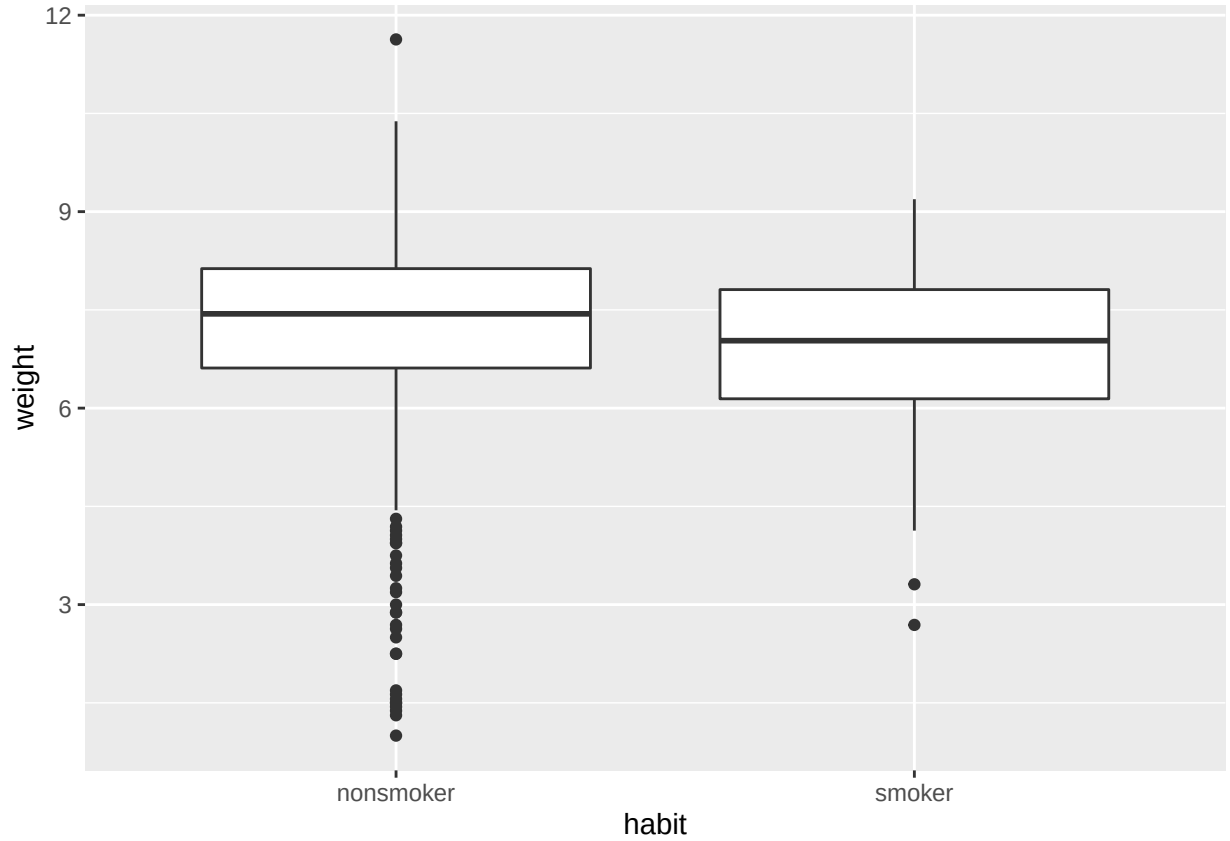
From the histograms, it appears that the distribution for variable weight and weeks are skewed to the left. They may be some outliers with very low and also very weights (above 10 less than 2). The distribution for "mage" is fairly normal.

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?

–vb– From Summary function, we could see that we have 1 case where value for habit was not specified. We are ommiting this case.

```
ggplot(nc[complete.cases(nc), ], aes(x=habit, y=weight)) + geom_boxplot()
```



From the plots, we can tell that the mean for the population of non-smokers has a value of median higher than for the smokers population. Also, they are more possible ouliers for the non-smokers population. Finally, the IQR for non-smokers is slightly less 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.

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

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`.

–vb–

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

```
## nc$habit: nonsmoker
## [1] 873
```

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

- The sample was randomly selected
- we can assume that the sample (1000 cases) represent less than 10% of population we can assume independence of cases
- Although there is a skewedness for the variable weight, the sample size is quite large (over 100 for each set somkers / non-smokers)

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

–vb– H0 = There is no difference in the average birth weight in new borns from mothers who did not smoke and from mothers who did. ($\mu_{ns} - \mu_s = 0$) Ha = There is a difference in the average birth weight in new borns from mothers who smoke and from mothers who did not smoke. ($\mu_{ns} - \mu_s \neq 0$)

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")
```

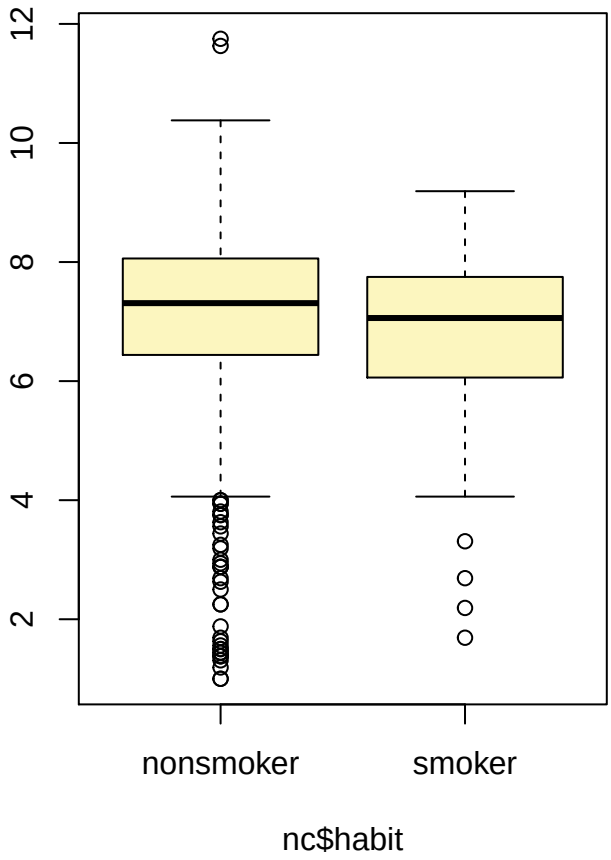
Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

```
## Warning: package 'BHH2' was built under R version 3.2.4
```

```
## 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"))
```

On your own

when looking at the summary statistics, there are NA's in some of the variables of interest for rest of analysis. First we will take a look at all the rows with at least one na.

```
nc_na <- nc[!complete.cases(nc), ]
dim(nc_na)
```

```
## [1] 200  13
```

By the results, it is clear that if we indescriminately remove all "incomplete" rows we would remove 200 row (out of 1000).
This is too large a size and may impact some of the results. The variable "fage"" is the one with the most na's (171). Since we are not using this variable in the rest of the analysis, we will drop it from our data set. Once this is done, we will filter out any row with a value of na in any of the remaining variables (38 out of 1000). We will perform the analysis this new data set.

```
nc_fage_drop <- nc[, 2:13]
str(nc_fage_drop)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ mage          : int  13 14 15 15 15 15 15 15 16 16 ...
##  $ mature        : Factor w/ 2 levels "mature mom","younger mom": 2 2 2 2 2 2 2 2 2 2 ...
##  $ weeks         : int  39 42 37 41 39 38 37 35 38 37 ...
##  $ premie        : Factor w/ 2 levels "full term","premie": 1 1 1 1 1 1 1 2 1 1 ...
##  $ visits        : int  10 15 11 6 9 19 12 5 9 13 ...
##  $ marital       : Factor w/ 2 levels "married","not married": 1 1 1 1 1 1 1 1 1 1 ...
##  $ gained        : int  38 20 38 34 27 22 76 15 NA 52 ...
##  $ weight        : num  7.63 7.88 6.63 8 6.38 5.38 8.44 4.69 8.81 6.94 ...
##  $ lowbirthweight: Factor w/ 2 levels "low","not low": 2 2 2 2 2 1 2 1 2 2 ...
##  $ gender        : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ habit         : Factor w/ 2 levels "nonsmoker","smoker": 1 1 1 1 1 1 1 1 1 1 ...
##  $ whitemom      : Factor w/ 2 levels "not white","white": 1 1 2 2 1 1 1 1 2 2 ...
```

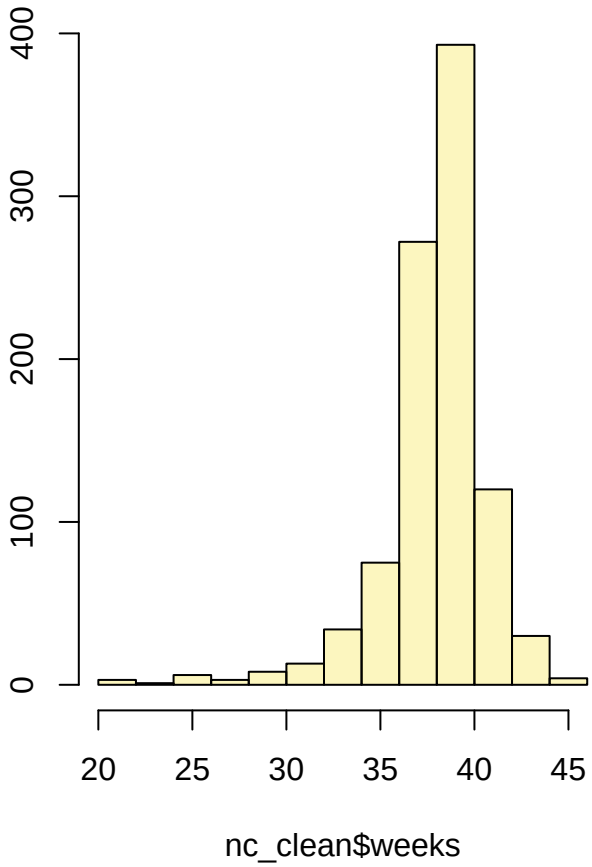
```
# Check for na on new data frame
nc_clean <- nc_fage_drop[complete.cases(nc_fage_drop), ]
dim(nc_clean)
```

```
## [1] 962  12
```

- 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_clean$weeks, est="mean", type = "ci", conflevel=0.95, method="theoretical")
```

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



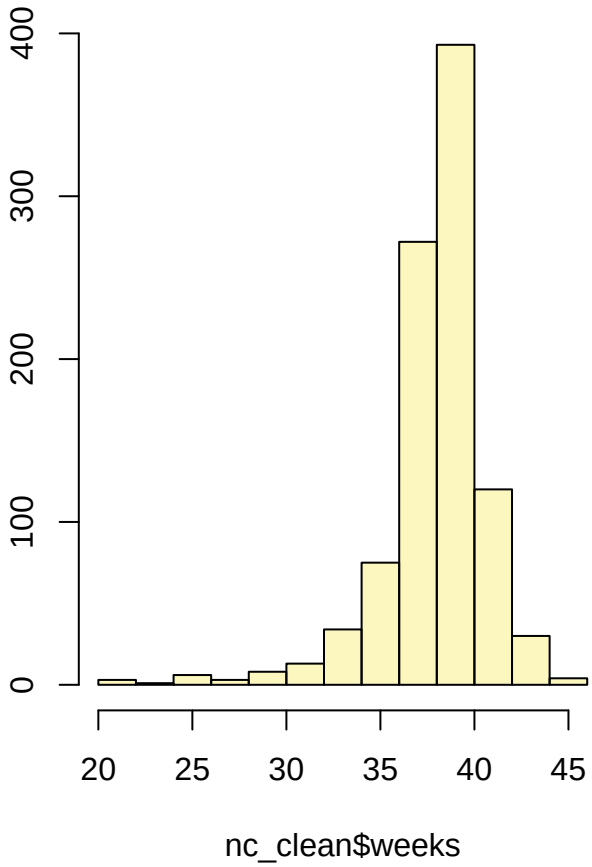

```
## mean = 38.395 ;  sd = 2.8295 ;  n = 962
## Standard error = 0.0912
## 95 % Confidence interval = ( 38.2162 , 38.5738 )
```

We are 95% confident that the average length of pregnancy (in weeks) is between 38.2162 and 38.5738.

- 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_clean$weeks, est="mean", type = "ci", conflevel=0.90, method="theoretical")
```

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

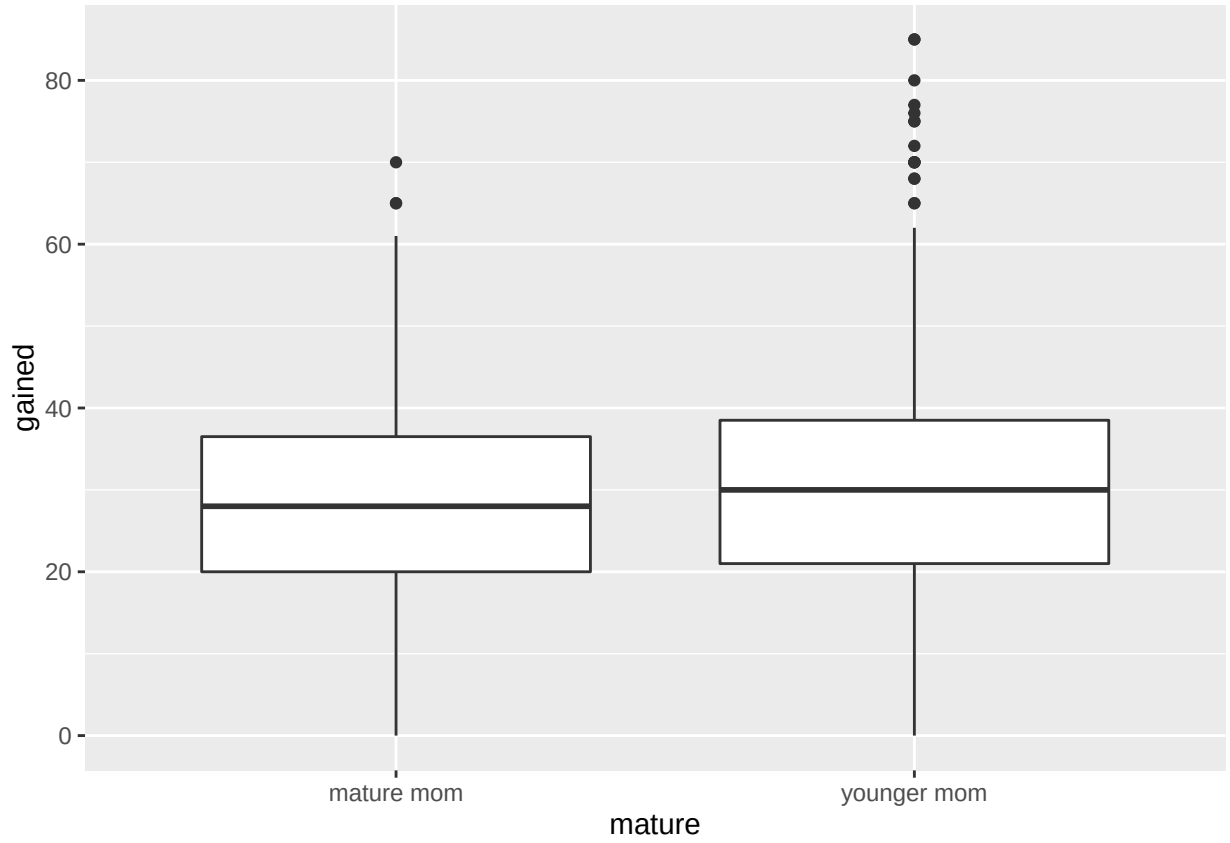



```
## mean = 38.395 ;  sd = 2.8295 ;  n = 962
## Standard error = 0.0912
## 90 % Confidence interval = ( 38.245 , 38.5451 )
```

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

H0: There is no difference between the average weight gain of young mothers and the average weight gain of mature mothers.
Ha: There is a difference between the average weight gain of young mothers and the average weight gain of mature mothers.

```
ggplot(nc_clean, aes(x=mature, y=gained)) + geom_boxplot()
```



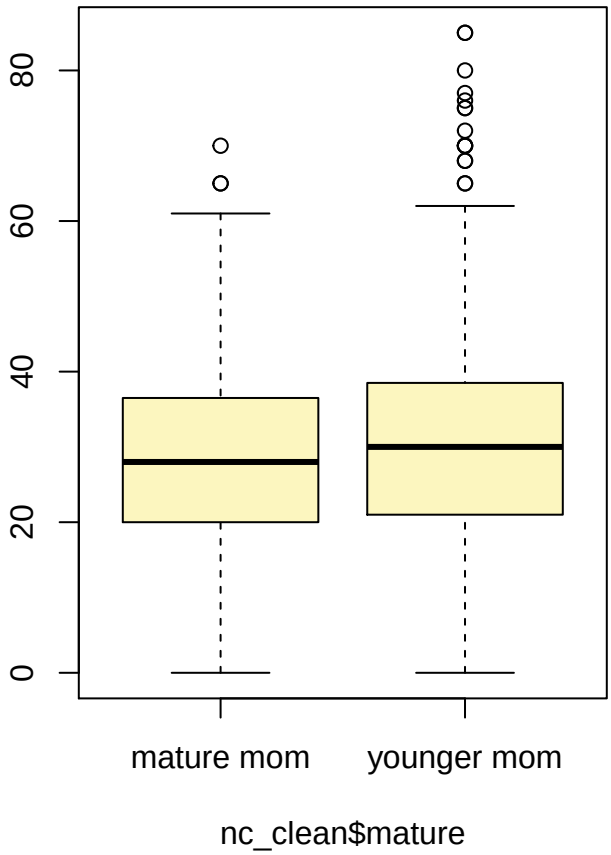

```
by(nc_clean$gained, nc_clean$mature, mean)
```

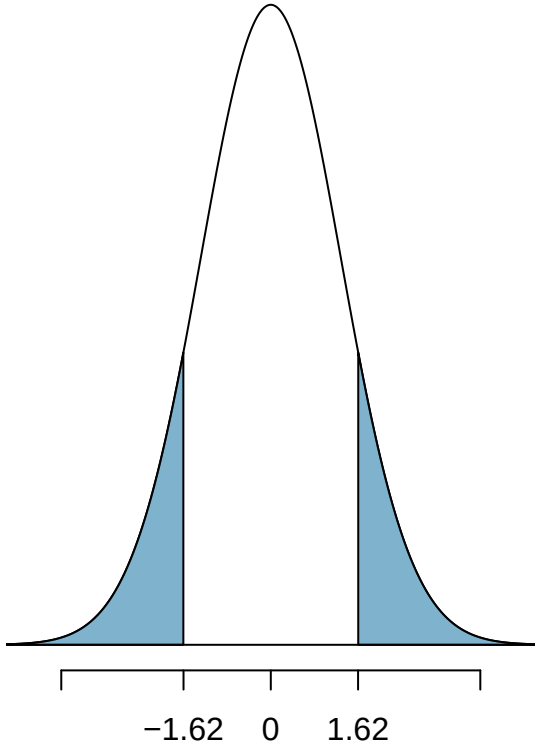
```
## nc_clean$mature: mature mom
## [1] 28.93701
## ------------------------------------------------------------
## nc_clean$mature: younger mom
## [1] 30.55569
```

```
inference(y = nc_clean$gained, x = nc_clean$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 = 127, mean_mature mom = 28.937, sd_mature mom = 13.4847
## n_younger mom = 835, mean_younger mom = 30.5557, sd_younger mom = 14.3162

## Observed difference between means (mature mom-younger mom) = -1.6187
##
## H0: mu_mature mom - mu_younger mom = 0
## HA: mu_mature mom - mu_younger mom != 0
## Standard error = 1.295
## Test statistic: Z =  -1.25
## p-value =  0.2114
```





- 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.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
nc_clean %>% group_by(mature) %>%
       summarise(min(mage), max(mage))
```

```
## Source: local data frame [2 x 3]
##
```

```
##        mature min(mage) max(mage)
##        (fctr)     (int)     (int)
## 1  mature mom        35        50
## 2 younger mom        13        34
```

Age cut-off for mature mon = 35 Age cut-off for younger mon = 34

The function above will first group "mage"" by variable "mature"" then within each group, the minimum and maximum will be found.
we are expecting that there is no overlap between the 2. Hence, maximum of younger mothers and minimum of mature mothers are the cut-off...

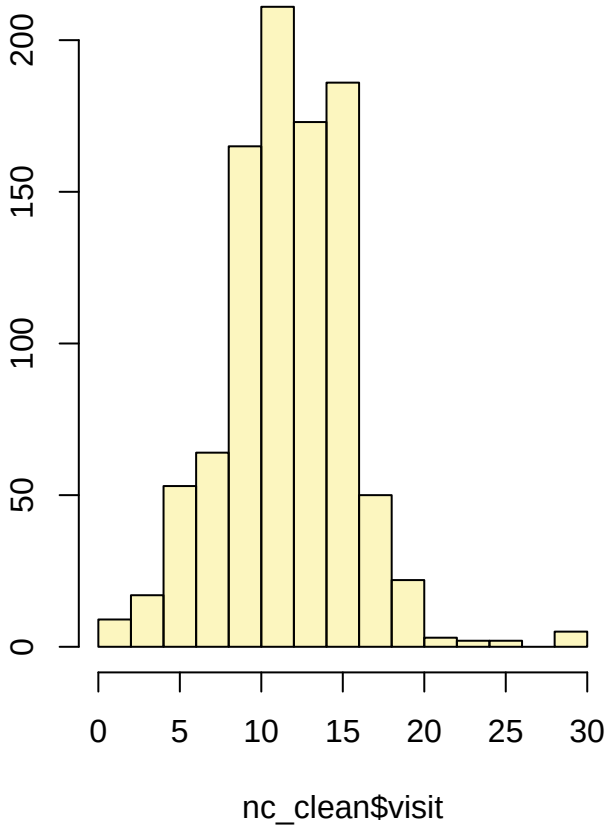
- Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the `inference` function, report the statistical results, and also provide an explanation in plain language.

Does average number of hospital visits during pregnancy varies whether the mother is white or non-white?

First we will look at a 95% confidence interval for the variable "visits":

```
inference(y=nc_clean$visit, est="mean", type = "ci", conflevel=0.95, method="theoretical")
```

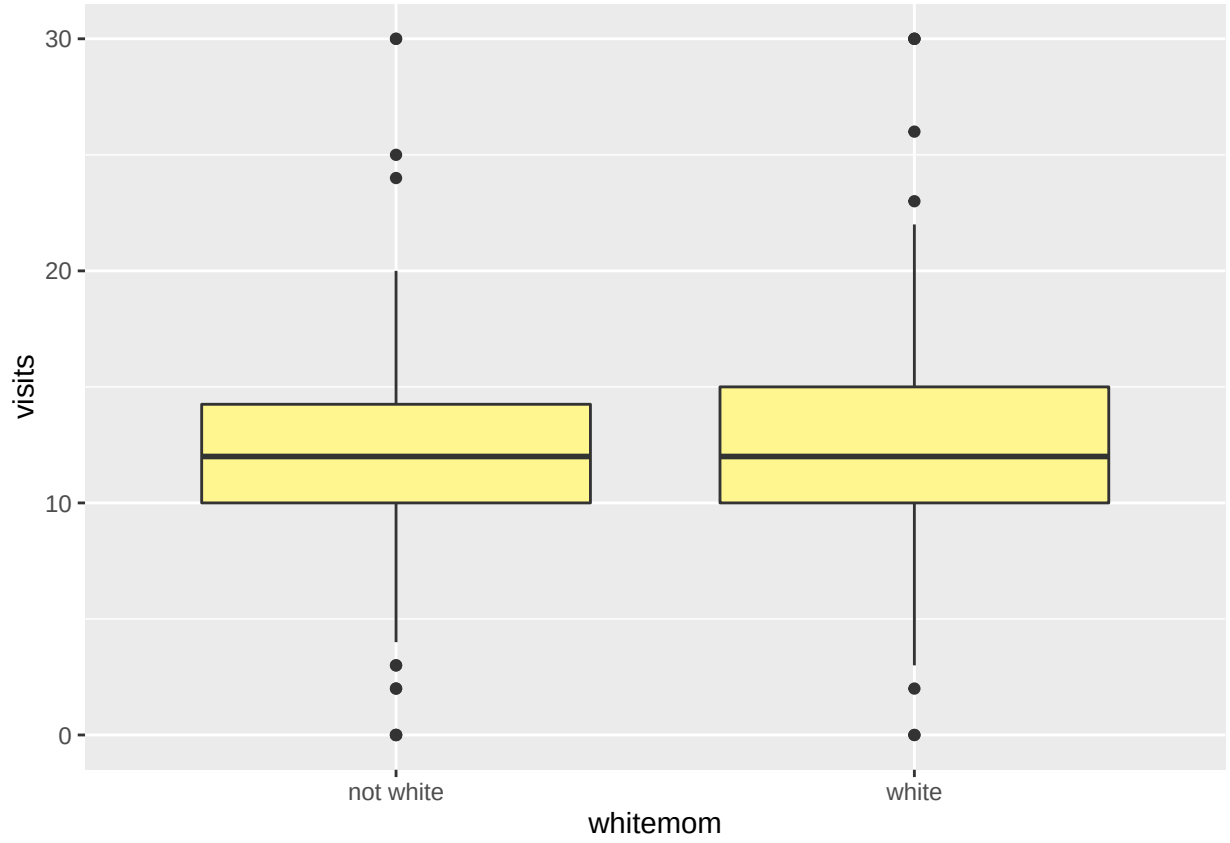
```
## Single mean
## Summary statistics:
```



```
## mean = 12.2048 ;  sd = 3.9088 ;  n = 962
## Standard error = 0.126
## 95 % Confidence interval = ( 11.9578 , 12.4518 )
```

H0: There is no difference between the average number of hospital visits for white mothers and non-white mothers Ha: There is a difference between the average number of hostpital visits for white mothers and non-white mothers

```
ggplot(nc_clean, aes(x=whitemom, y=visits)) + geom_boxplot(fill = "khaki1")
```



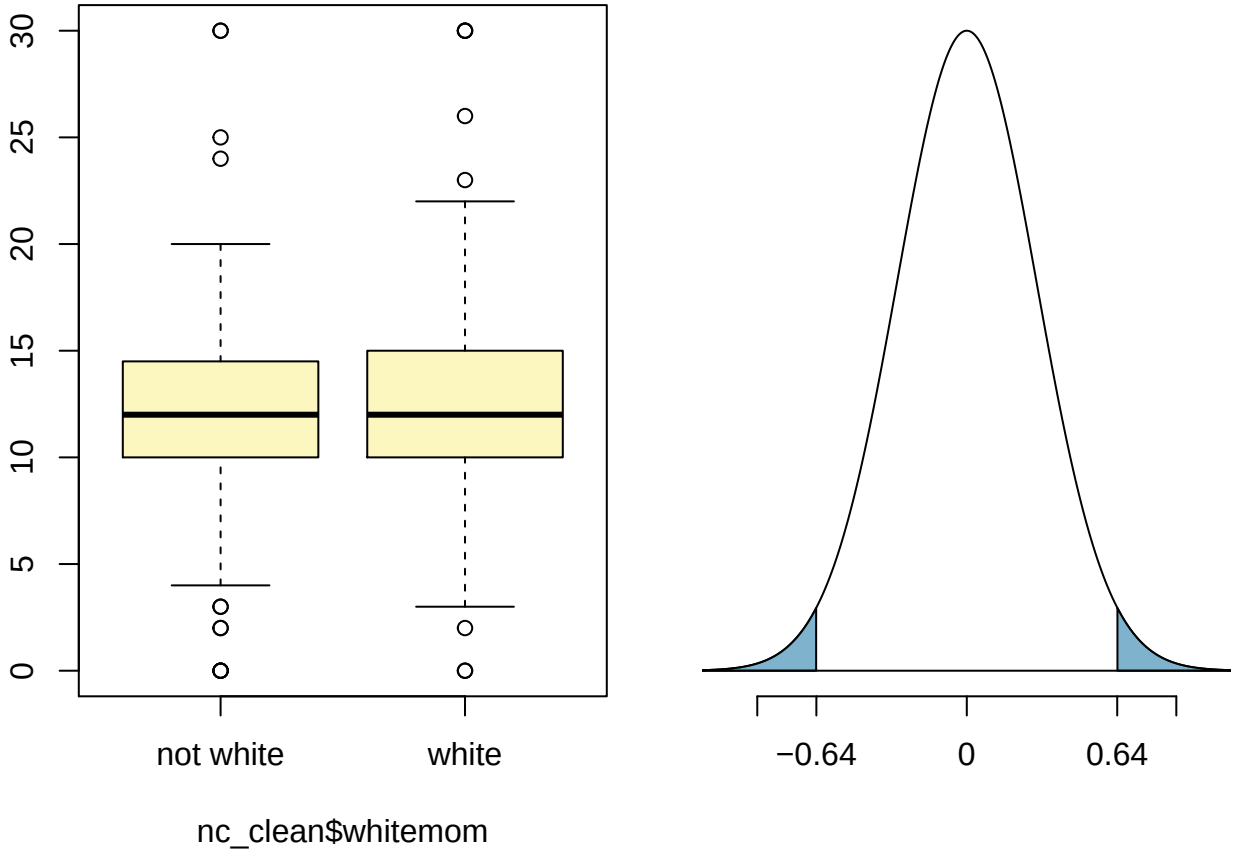

```
by(nc_clean$visits, nc_clean$whitemom, mean)
```

```
## nc_clean$whitemom: not white
## [1] 11.74632
## ------------------------------------------------------------
## nc_clean$whitemom: white
## [1] 12.38551
```

```
inference(y = nc_clean$visits, x = nc_clean$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 = 272, mean_not white = 11.7463, sd_not white = 4.2969
## n_white = 690, mean_white = 12.3855, sd_white = 3.7326
```

```
## Observed difference between means (not white-white) = -0.6392
##
## H0: mu_not white - mu_white = 0
## HA: mu_not white - mu_white != 0
## Standard error = 0.297
## Test statistic: Z =  -2.154
## p-value =  0.0312
```



The p-value = 0.0312, is less that the significance level of $\alpha = 0.05$ and we reject the H0 hypothesis in favor of Ha hypothesis.