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` .

Exercise 1 What are the cases in this data set? How many cases are there in our sample?

Answer:

```
nrow(nc)
```

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

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

summary(nc)

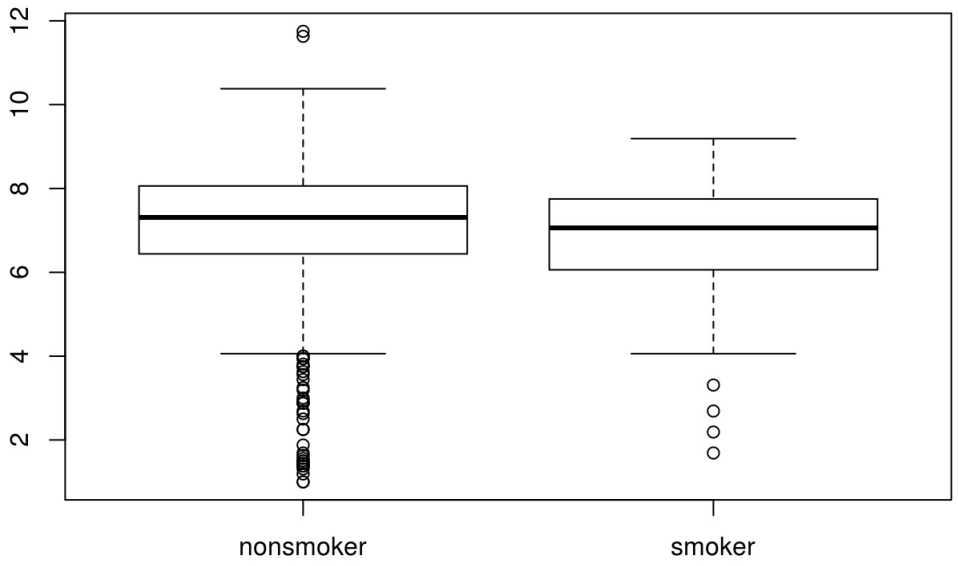
As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

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.

Exercise 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)
```



The two variables have almost the same median and that much more mothers who didn't smoke their baby births

have plenty of outliers less than 4 pounds. Mothers who smoked have less outliers and both distributions are slightly normal.

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

Exercise 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:

The observations are idependent of one another; a mother giving birth shouldn't reflect another mother giving birth.

The distribution of mothers who smoked and didn't smoked are nearly normal.

we'll use a sample size of say 30

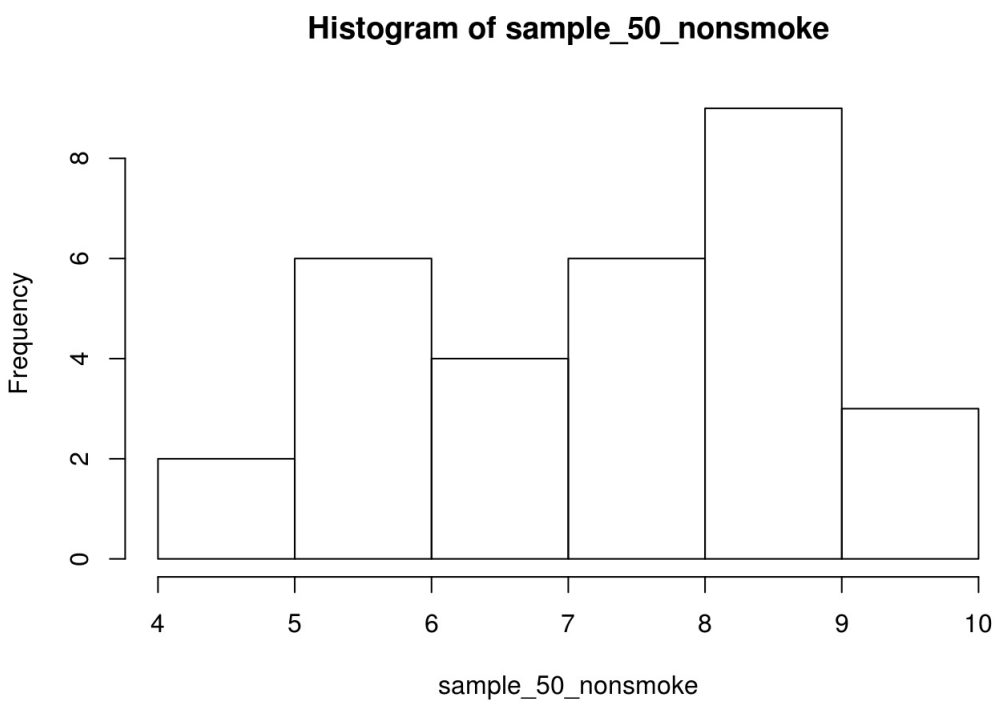
```
by(nc$weight, nc$habit, mean)
```

```
## nc$habit: nonsmoker
## [1] 7.144273
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 6.82873
```

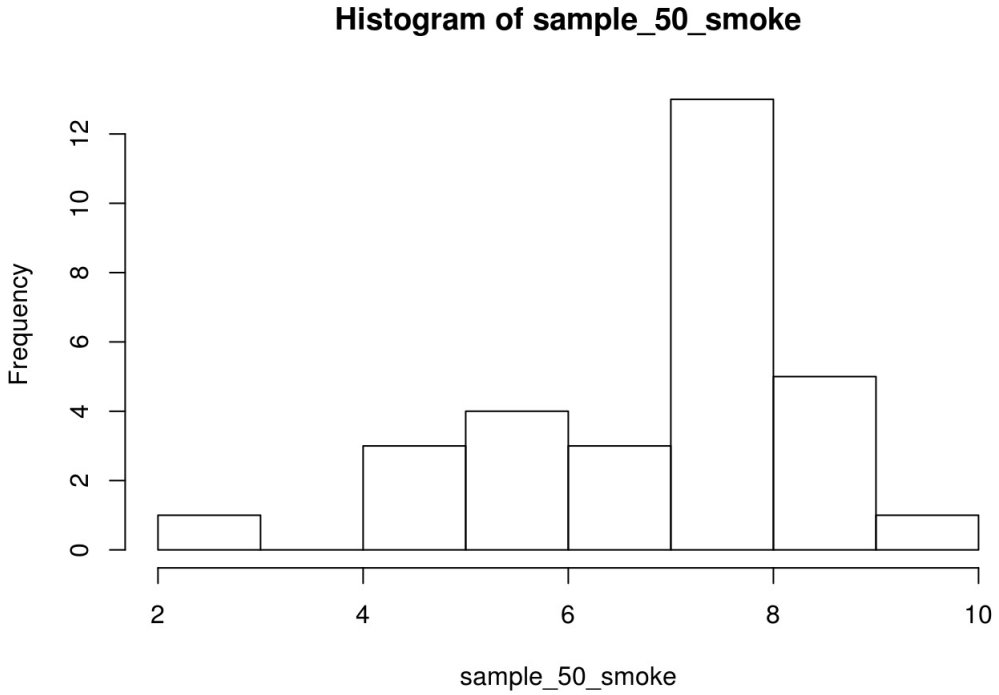
```
mother_smoke <- subset(nc, habit == 'smoker', select = weight)
mother_nonsmoke <- subset(nc, habit == 'nonsmoker', select = weight)

sample_50_smoke <- sample(unlist(mother_smoke), size = 30)
sample_50_nonsmoke <- sample(unlist(mother_nonsmoke), size = 30)

hist(sample_50_nonsmoke)
```




```
hist(sample_50_smoke)
```



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

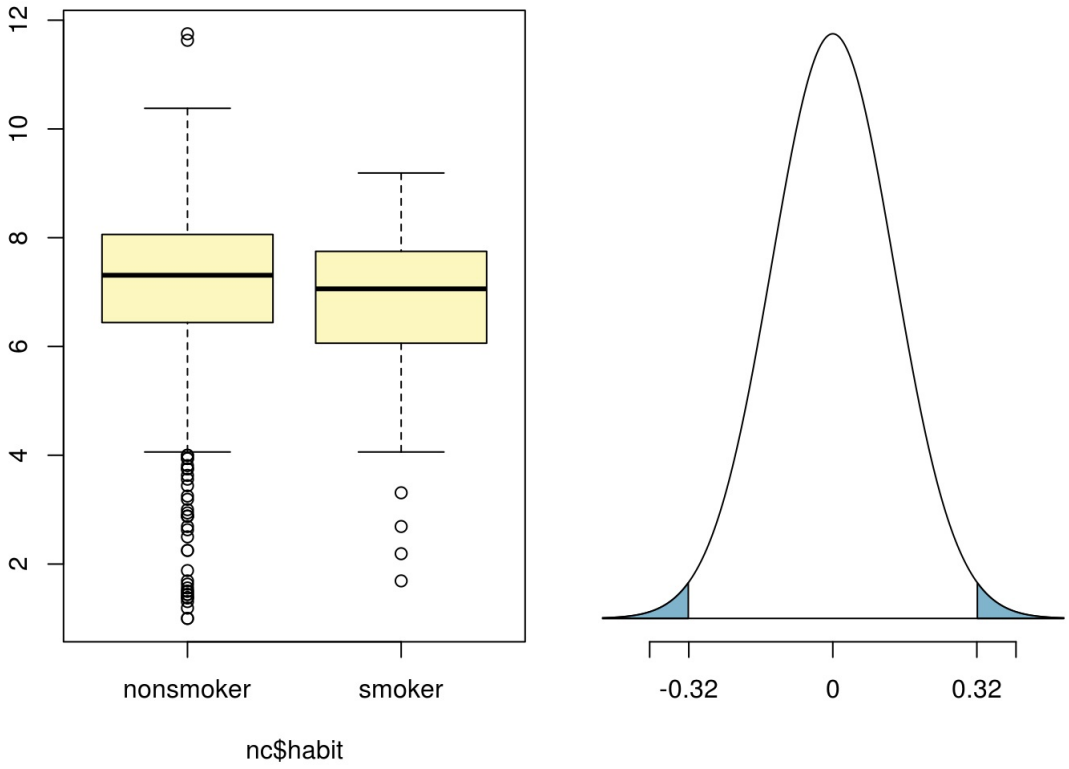
Answer: let H_0 be the null hypothesis that is $\mu_{nonsmoker}$ - μ_{smoker} = 0 let H_A be the alternative hypothesis that is $\mu_{nonsmoker}$ - μ_{smoker} != 0

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")
```

```
## 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
```



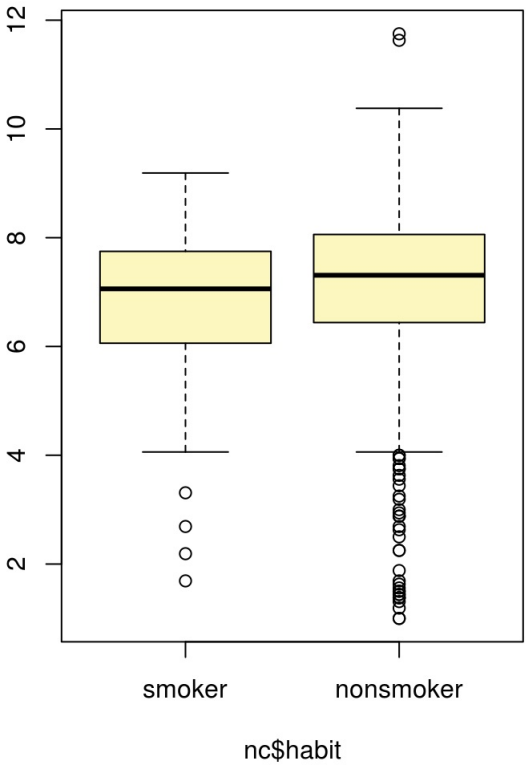
Let's pause for a moment to go through the arguments of this custom function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and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

Exercise 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.

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 )
```

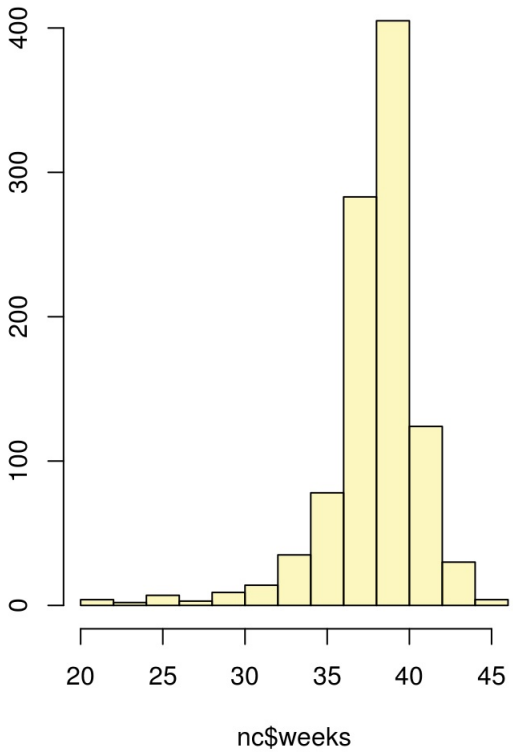
On your own

1. 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.

Answer:

```
inference(y = nc$weeks, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

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




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

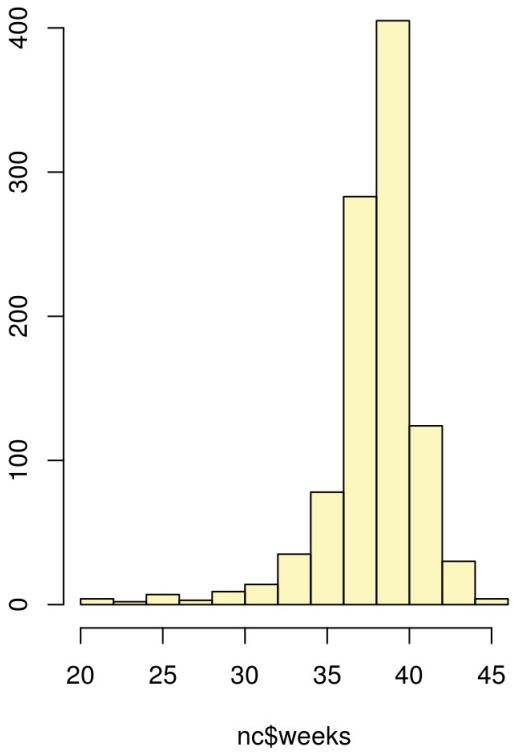
We are 95% confident that the true population mean of pregancy lengths is between

(38.1528 , 38.5165)

1. 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",
          order = c("smoker","nonsmoker"), 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 )
```

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

Answer:

let H_0 = $\mu_{younger}$ - μ_{mature} = 0 let H_A = $\mu_{younger}$ - μ_{mature} != 0

1. 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 will use the minimum age in mature mothers and the max age in younger mothers as I believe that is the max age

a mother should be young is right before she is classified as mature and the min age of a mature mother.

```
min_mature <- nc[nc$mature == 'mature mom','mage']
min_younger <- nc[nc$mature == 'younger mom', 'mage']

# get the extrema (min/max)
max(min_younger)
```

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

```
min(min_mature)
```

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

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.

Answer: my research question is “Does the race of the parent affect the weight of the newborn child?”

Variables used are the ‘whitemom’ and the ‘weight’

let H_0 = μ_{white} - $\mu_{notwhite}$ = 0 that is is there any difference in average baby weights in the two races let H_A = μ_{white} - $\mu_{notwhite}$!= 0 that is there is a difference in average baby weights in if the parent is white or not.

```
inference(y = nc$weight,x = nc$whitemom, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("white","not white"), conflevel = 0.90)
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_white = 714, mean_white = 7.2505, sd_white = 1.4333
## n_not white = 284, mean_not white = 6.7195, sd_not white = 1.6207
```

```
## Observed difference between means (white-not white) = 0.5309
##
## Standard error = 0.1101
## 90 % Confidence interval = ( 0.3498 , 0.712 )
```

Looking at the average differences using a 90% confidence interval, we see that the estimated true population mean

falls within the confidence interval and we have significant evidence to reject the null hypothesis and

there is a possible difference in the weights of babies of white and non white mothers and the child’s weight makes a difference if the mother is caucausian or not.