Lab 5: Inference for numerical data

North Carolina births

In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are 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

Let's load the `nc` data set into our workspace.

```
download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")
load("nc.RData")
```

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

`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.
`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`.
`marital`	whether mother is `married` or `not married` at birth.
`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?

Before we begin our analysis let's take a look at features of all the variables in the data set by getting a summary of each.

```
summary(nc)
```

We will first tackle the relationship between a mother's smoking habit and the weight of her baby. Exploratory analysis is a useful first step when examining data because it helps us notice trends and develop research questions. By now you have had practice using R commands to summarize and visualize data.

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

The side-by-side box plots show how the medians of the two distributions compare, and we can also compare the means of the distributions. The following command gives the mean weights of babies born to smoker and non-smoker mothers.

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

There is clearly an observed difference, but is this difference statistically significant? In order to answer this question we need to conduct a hypothesis test.

Exercise 3 Check if the conditions necessary for inference are satisfied? Note that you will need to obtain sample sizes to check the conditions.

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

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

```
inference(data = nc$weight, group = nc$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 `data`, this is the response variable that we are interested in: `weight`
- The second argument is the grouping variable, `group`, this is the variable that we use to split the data into two groups, smokers and nonsmokers: `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 doing 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`, `twosided`.
- Lastly, the `method` of inference can be `theoretical` or `simulation` based.

Exercise 5 Change the `type` argument to `ci` construct a confidence interval for the difference between the weights of babies born to smoker and non-smoker mothers.

By default the function reports an interval for $(\mu_{nonsmoker} - \mu_{smoker})$, we can easily change this order:

```
inference(data = nc$weight, group = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

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 grouping variable, so you can omit the `group` variable from the function.
2. 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`.
3. Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.
4. Now, a non-inference task: Determine the age cutoff for younger and mature mothers. You can use any method you like for answering this question, but make sure to explain your method in your write up.
5. Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Your research question should be a question that you can answer using a hypothesis test and/or a confidence interval. Write up an answer for your research question using the `inference`.