Lab 5: Inference for numerical data

Template for lab report

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Compile often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/courses/Fall13/sta101/labs/lab5.Rmd", destfile = "lab5.Rmd")
```

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.

```
download.file("http://stat.duke.edu/~mc301/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.
`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?

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.

We will first start with analyzing the weight gained by mothers throughout the pregnancy: `gained`.

> **Exercise 2** Using visualization and summary statistics, describe the distribution of weight gained by mothers during pregnancy. Also comment on how many mothers we're missing weight gain data from.

Since we have some missing data on weight gain, we'll first create a cleaned-up version of the weight gain variable, and use this variable in the next portion of the analysis. There are many ways of accomplishing this task in R, we'll do it using the `na.omit` function:

```
gained_clean = na.omit(nc$gained)
```

We'll also store the sample size of the new variable (which should be less than 1000 since we dropped the observations with NAs) in order to be able to use this value in the next portion of the analysis as well. We'll use the `length` function for this:

```
n = length(gained_clean)
```

Quick check: Double check that `n` is what it's expected to be based on the number of NAs present in the original weight gain variable.

The bootstrap

Using this sample we would like to construct a bootstrap confidence interval for the average weight gained by *all* mothers during pregnancy. Below is a quick reminder of how bootstrapping works:

(1) Take a bootstrap sample (a random sample with replacement of size equal to the original sample size) from the original sample.

(2) Record the mean of this bootstrap sample.

(3) Repeat steps (1) and (2) many times to build a bootstrap distribution.

(4) Calculate the XX% interval using the percentile or the standard error method.

Since were going to do some random sampling, lets start by setting a seed. As usual, type

```
set.seed(xxx)
```

and replace `xxx` with a number of your choosing. Make sure to include this piece of code in your report as well, so that a seed is set in the report workspace.

Now let's take 100 bootstrap samples (i.e. with replacement), and record their means in a new object called `boot_means`. Before we take the samples, we start with creating a new object called `boot_means` where we can store the bootstrap means as we collect them.

```
boot_means = rep(NA, 100)
for(i in 1:100){
  boot_sample = sample(gained_clean, n, replace = TRUE)
  boot_means[i] = mean(boot_sample)
}
```

Exercise 3 Make a dot plot of the bootstrap distribution, and estimate a 90% confidence interval using the percentile method for the average weight gained by mothers during pregnancy, explain briefly how you estimated the interval, and interpret this interval in context of the data. (Note, the `dotPlot` function is in a library called `BHH2`, so we need to first load that library.)

```
library("BHH2")
dotPlot(boot_means)
```

Exercise 4 Next, calculate the bootstrap standard error. Note that this is basically the standard deviation of the bootstrap means stored in `boot_means`. Using this value, calculate a 90% confidence interval for the same parameter of intertest. Are the two intervals approximately equal?

Exercise 5 **Connect:** Briefly describe (1-2 sentences) the code above for bootstrapping (what's happening in the for loop) relates to the class activity on bootstrapping. Make a deliberate connection between the two methods.

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.

The `inference` function

Next we'll introduce a new function that you'll be seeing a lot more of in the upcoming labs – a function that allows you to apply any statistical inference method that you'll be learning in this course. Since this is a custom function, we need to first go and download it from the course website.

```
source("http://stat.duke.edu/~mc301/R/inference.R")
```

Writing a for-loop every time you want to calculate a bootstrap interval or run a randomization test is cumbersome. This function automates the process.

By default the function takes 10,000 bootstrap samples (instead of the 100 you've taken above), creates a bootstrap distribution, and calculates the confidence interval.

```
inference(nc$gained, type = "ci", method = "simulation", conflevel = 0.9, est = "mean")
```

We can easily change the confidence level to 95% by changing the `conflevel`:

```
inference(nc$gained, type = "ci", method = "simulation", conflevel = 0.95, est = "mean")
```

Or create an interval for the median instead of the mean:

```
inference(nc$gained, type = "ci", method = "simulation", conflevel = 0.95, est = "median")
```

Exercise 6 Create a 95% confidence interval for the median age of fathers at the birth of their child: (`nc$mage`. Interpret the interval within the context of the data. Comment on the reliability of this estimate.

Evaluating relationships between two variables

When the response variable is numerical and the explanatory variable is categorical, we can evaluate the relationship between the two variables by comparing means (or medians, or other measures) of the numerical response variable across the levels of the explanatory categorical variable.

Exercise 7 What type of variables are `habit` and `weight` (numerical/categorical)? Make an appropriate plot that visualizes the relationship between these variables. What does the plot highlight about this relationship?

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

Exercise 8 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`.

Exercise 9 Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

Next, we'll use the `inference` for evaluating these hypotheses.

```
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 grouping variable, `x`, which is the explanatory variable – the grouping variable accross the levels of which we're comparing the average value for the response variable, 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 10 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. Then, interpret the interval in context of the data.

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"))
```

Exercise 11 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.

Exercise 12 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.

Exercise 13 **Reflect:** Which concepts does this lab help you understand better? Which (if any) concepts do you feel that you need more practice in?