Lab 4: Inference for numerical data

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform. Note that the order of the choices in multiple choice questions may be different on the Duke Coursera platform than the order in this document.

Part A: 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(url("http://bit.ly/dasi_nc"))
```

If the above shortened link doesn't work for you, try http://d396qusza40orc.cloudfront.net/statistics/lab_resources/nc.RData (http://d396qusza40orc.cloudfront.net/statistics/lab_resources/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` .

Question 1 There are 1,000 cases in this data set, what do the cases represent?

A) The hospitals where the births took place
B) The fathers of the children

C) The days of the births
D) The births

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:** Using visualization and summary statistics, describe the distribution of weight gained by mothers during pregnancy.

Question 2 How many mothers are we missing weight gain data from?

A) 0
B) 13
C) 27
D) 31

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 few steps 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:

- Step 1: Take a bootstrap sample (a random sample with replacement of size equal to the original sample size) from the original sample.
- Step 2: Record the mean of this bootstrap sample.
- Step 3: Repeat steps (1) and (2) many times to build a bootstrap distribution.

- Step 4: Calculate the XX% interval using the percentile or the standard error method.

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)
}
```

Once again you used a for loop, but this time the objective is different.

Question 3 True / False: The sampling distribution is calculated by resampling from the population, the bootstrap distribution is calculated by resampling from the sample.

A) True
B) False

- **Exercise:** Make a histogram of the bootstrap distribution (`boot_means`).

Question 4 True / False: To construct the 95% bootstrap confidence interval using the percentile method, we estimate the values of the 5^{th} and 95^{th} percentiles of the bootstrap distribution.

A) True
B) False

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

The `inference` function

Next we'll introduce a new function that you'll be seeing a lot more of in the upcoming labs – a custom 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://bit.ly/dasi_inference")
```

If the above shortened link doesn't work for you, try http://d396qusza40orc.cloudfront.net/statistics/lab_resources/inference.R (http://d396qusza40orc.cloudfront.net/statistics/lab_resources/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, using the percentile method.

```
inference(nc$gained, type = "ci", method = "simulation", conflevel = 0.90, est =
"mean", boot_method = "perc")
```

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", boot_method = "perc")
```

We can easily use the standard error method by changing the `boot_method`:

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

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

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

Question 5 True / False: The bootstrap distribution of the median weight gain is a smooth, unimodal, symmetric distribution that yields a reliable confidence interval estimate.

A) True
B) False

- **Exercise:** Create a 95% bootstrap confidence interval for the mean age of fathers at the birth of their child, `nc$fage`, using the standard error method. Interpret the interval within the context of the data.

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:** What type of variables are `habit` and `weight` (numerical/categorical)? Make an

appropriate plot that visualizes the relationship between these variables.

Question 6 Based on the plot from the previous exercise, which of the following is false about the relationship between `habit` and `weight`?

A) Median birth weight of babies born to non-smoker mothers is slightly higher than that of babies born to smoker mothers.
B) Range of birth weights of babies born to non-smoker mothers is greater than that of babies born to smoker mothers.
C) Both distributions are slightly right skewed.
D) The IQRs of the distributions are roughly equal.

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.

Hypothesis tests and confidence intervals

- **Exercise:** 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`.

Next, we'll use the `inference` for evaluating whether there is a difference between the average birth weights of babies born to smoker and non-smoker mothers.

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0, altern
ative = "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.

By default the function sets the parameter of interest to be ($\mu_{nonsmoker} - \mu_{smoker}$). We can easily change this order by using the `order` argument:

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0, altern
ative = "twosided", method = "theoretical", order = c("smoker","nonsmoker"))
```

Question 7 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. Which of the following is the **best** interpretation of the interval?

A) We are 95% confident that babies born to nonsmoker mothers are on average 0.05 to 0.58 pounds heavier at birth than babies born to smoker mothers.
B) We are 95% confident that babies born to nonsmoker mothers are on average 0.05 to 0.58 pounds lighter at birth than babies born to smoker mothers.
C) We are 95% confident that the difference in average weights of babies whose moms are smokers and nonsmokers is between 0.05 to 0.58 pounds.
D) We are 95% confident that the difference in average weights of babies in this sample whose moms are smokers and nonsmokers is between 0.05 to 0.58 pounds.

Question 8 Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice. What is the maximum age of a younger mom and the minimum age of a mature mom, according to the data?

A) The maximum age of younger moms is 32 and minimum age of mature moms is 33.
B) The maximum age of younger moms is 33 and minimum age of mature moms is 34.
C) The maximum age of younger moms is 34 and minimum age of mature moms is 35.
D) The maximum age of younger moms is 35 and minimum age of mature moms is 36.

- **Exercise:** 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.

Part B: The General Social Survey

The General Social Survey (http://www3.norc.org/GSS+Website/) (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. The survey asks many questions, two of which we will focus on for the next few exercises: `wordsum` (vocabulary test scores) and `class` (self-reported social class).

`wordsum` ranges between 0 and 10, and is calculated as the number of vocabulary questions (out of 10) that the respondent answered correctly.

Let's load that dataset and take a look at these variables:

```
load(url("http://bit.ly/dasi_gss_ws_cl"))
```

If the above shortened link doesn't work for you, try http://d396qusza40orc.cloudfront.net/statistics/lab_resources/gss.RData (http://d396qusza40orc.cloudfront.net/statistics/lab_resources/gss.RData).

- **Exercise:** Create numerical and visual summaries of the two variables individually to better understand their distribution. Also compute summary statistics and visualizations that display the relationship between the two variables.

Question 9 Which of the following methods is appropriate for testing for a difference between the average vocabulary test scores among the various social classes?

A) Z test
B) T test
C) ANOVA
D) χ^2 test

Now let's run the test:

```
inference(y = gss$wordsum, x = gss$class, est = "mean", type = "ht", alternative =
"greater", method = "theoretical")
```

Let's examine the output: First, we have a numerical response variable (score on vocabulary test) and a categorical explanatory variable (class). Since `class` has four levels, comparing average scores across the levels of the class variable requires ANOVA. Before we get to the ANOVA output we are presented with a series of summary statistics and the associated hypotheses. Since the p-value is low, we reject H_0 and conclude that there is evidence that at least one pair of means are different.

The last part of the ANOVA output displays results of the pairwise tests.

Question 10 Calculate the modified α (α^*) to be used for these tests.

A) $\alpha^* = \alpha = 0.05$

B) $\alpha^* = \alpha/2 = 0.025$
C) $\alpha^* = \alpha/4 = 0.0125$
D) $\alpha^* = \alpha/6 = 0.0083$

Question 11

View the p-values of the pairwise tests from the ANOVA output. Which of the following pairs of means are concluded to be different at the modified significance level?

A) Middle and lower
B) Working and lower
C) Middle and upper
D) Working and upper
E) Upper and lower