Lab 6: Statistical inference for small samples

The Data

The data come from an article by Fearnbach, Durban, Ellifrit, and Balcomb called “Size and long-term growth trends of Endangered fish-eating killer whales”. You can find the complete paper here.

Below is an excerpt from the abstract of this paper:

> “The Endangered southern resident population of killer whales *Orcinus orca* has been shown to be food-limited, and the availability of their primary prey, Chinook salmon *Oncorhynchus tshawytscha*, has been identified as a key covariate for the whales individual survival and reproduction. We collected aerial photogrammetry data on individual whale size, which will help to better inform energetic calculations of food requirements, and we compared size-at-age data to make inferences about long-term growth trends. A helicopter was used to conduct 10 flights in September 2008, resulting in 2803 images from which useable measurements were possible for 66 individually identifiable whales, representing more than three-quarters of the population.”

And given below is another excerpt from the body of the paper giving further information on the data:

> “Whales were typically measured more than once (median: 7 surfacing sequences, range: 1 to 38). Variability within estimates of the same whale was likely due to a foreshortening effect of whales not being directly underneath the photographer and surfacing whales not being at their most elongated body position at the time of the photograph. The main bias was therefore likely to be negative, resulting in underestimates of length, and we thus chose to use the maximum estimate to be the best (least biased) for each whale. It should be noted, however, that even the maximum estimate may still have been negatively biased for full body length, and simply represented the longest body position measured for that whale. To reduce this effect, we only considered estimates to be reliable if measurements had been obtained from $\geq$5 different images. All further analysis was based solely on the 46 whales for which this was the case.”

Exercise 1 In your own words, describe why the researchers did not use data from the entire sample of 66 whales?

In this lab we will analyze data from the 46 whales that the researchers used in their paper. ~~Click on the link below to go to the Google Spreadsheet where the data are stored.~~

~~*https://docs.google.com/spreadsheet/ccc?key=0AkY2lFgS9uiDdHZlWHYzSk9HMnFWal9leGdLaU1LaVE*~~

~~Then go back to RStudio, and under Import Dataset choose `whales` and click on Import. You should see the data added to your workspace.~~

As of March 25, 2012 RStudio no longer has Google spreadsheet data import capability. If you’re working on this lab after this date and need to get the data, use the following command:

```
whales = read.csv("http://stat.duke.edu/courses/Spring12/sta101.1/labs/whales.csv")
```

Exercise 2 What are the cases and the variables? How many of each are there?

Before we proceed let's also reload the `inference()` function, it has some updates that will be useful for today's lab as well as your projects.

```
source("http://stat.duke.edu/courses/Spring12/sta101.1/labs/inference.R")
```

Calculating average whale size

There are two variables in the dataset that give information on the sizes of whales, `minlen` (minimum observed length) and `maxlen` (maximum observed length). Both are measured in meters. We will first create a new variable called `avglen` and use this variable in the rest of our analysis.

> **Exercise 3** Create a new variable in the dataset called `avglen` that is the mean of `minlen` and `maxlen` variables. If you're not sure how to do this, ask your teammates first, and then your TA if need be.

Sizes of male whales

The Killer whales article on Wikipedia states "Males typically range from 6 to 8 metres (20-26 ft) long and weigh in excess of 6 tonnes." Let's see how the sizes of male whales in this dataset compare to the estimate given in this article.

We will first crate a new dataset `male` that only has data for male whales using subletting.

```
male = whales[whales$sex == "M", ]
```

Exercise 4 This is a multi-part exercise:

(a) Describe the distribution of sizes of male whales using both visualization (plots) and summary statistic.

(b) Conduct a hypothesis test to see if the average length of a whale is different than 7 meters (the midpoint of the range given in the Wikipedia article). Use the `inference` function and a theoretical method.

(c) Using hand calculations verify the T statistic and the p-value given by the `inference` function. *Hint: You will need to use summary statistics such as the sample mean, standard deviation, and sample size. The `summary` function in R is useful for that.*

(d) Why did we use a T instead of a Z distribution for this hypothesis test?

(e) Now, re-do the hypothesis test using a simulation method. Do the results of the two methods agree?

Comparing sizes of male and female whales

In this section you once again have a multi-part exercise that takes you through an analysis of comparison of sizes of male and female whales.

Exercise 5 This is a multi-part exercise:

(a) Create a new dataset called `female` that only has data for female whales.

(b) Describe the distribution of sizes of female whales using both visualization (plots) and summary statistic.

(c) Write the hypotheses for testing if the average lengths of male and female values differ.

(d) Conduct this hypothesis test using the `inference` function and a theoretical method.

(e) Using hand calculations verify the T statistic and the p-value given by the `inference` function.

(f) Why did we use a T instead of a Z distribution for this hypothesis test?

(g) Now, re-do the hypothesis test using a simulation method. Do the results of the two methods agree?

On Your Own

Note that this lab is not due this Thursday, but instead the following Monday, March 26 at noon along with HW6.

The answers to each question should, at a minimum, include the output of the `inference` function including any plots and interpretation of results in context of the question.

1. Calculate a 95% confidence interval for the average length of female whales.
   (a) Do this using a theoretical approach, and interpret your interval in context.
   (b) The Wikipedia article also states "Females are smaller, generally ranging from 5 to 7 metres (1623 ft) and weighing about 3 to 4 tonnes." Does your interval provide support for this statement on the lengths of female whales?
   (c) Re-do the interval using a simulation method. Add an additional argument to the `inference` function in order to use only 100 bootstrap samples (instead of the default 10,000). You can do this by including the argument `nsim = 100` in the inference function. In your own words, describe how bootstrapping works, and what each dot on the resulting plot means.
2. Using the entire dataset, conduct a hypothesis test to see if the average number of measurements taken on males and females are different. You can use either a theoretical or a simulation approach.
3. Using the entire dataset, create a new variable called `gt10`, which stands for "greater than 10 years old". This variable should be "yes" if the age of the whale is greater than or equal to 10 and "no" if the age is less than 10. For this question, simply include the code you used in your answer. *Hint: Referring to the previous lab where we recoded the type of school as rural and not rural might be helpful.*
4. Construct a 95% confidence interval for the proportion of whales that are greater than 10 years old. Interpret this interval in context.
5. Conduct a hypothesis test to see if the proportions of male and female whales who are greater than 10 years old are different. Make sure to interpret your conclusion in context of the question.