Lab Exercise 4 - The Bootstrap

Answer sheet attached as usual. Due Thursday.

We've been studying the t-test, and the t-test is a fine statistical technique that is nearly ubiquitous in statistics. That means it ain't going away. We'll see it again when we talk about ANOVA, and yet again when we talk about regression analysis. Let's see an example. Fire up R and do the following.

```r
> ls()   # examine your workspace
[1] "Gender"  "Married"  "MJ"     "Occup"    "occupations" "TR"
[7] "Unio"    "Wages"    "WG"     "WG.bak"
> rm(list=ls())   # clean it up if necessary
> RR = read.table(header=T, file="http://ww2.coastal.edu/kingw/psyc480/data/birds.txt")
> # get the data
> summary(RR)   # summarize
  constr food.eaten
no   26    Min.   : 2.000
yes  45   1st Qu.: 6.652
NA's  4    Median :10.115
       Mean  :10.289
       3rd Qu.:13.330
       Max.  :22.830
       NA's  :7
```

The data we are going to use is from a file called "birds.txt". As always (and as I explained last time, but some of you apparently didn't listen), you can go to the website and under COURSE MATERIALS you'll find a box labeled Datasets, and inside that box you'll find links to the datasets we are using. If you click one of those links, you will be able to examine the data as well as read a description of it. Briefly, these are data from Ryan Russ's Psyc 497 project (Fall 1996), in which he collected data from a bird feeder in his back yard, recording how much food was eaten from the feeder in cc/hr. He also recorded whether or not there was activity at the construction site next door where an apartment building was being put up. His hypothesis was based on something he learned in my Animal Behavior class: activity at the construction site should disrupt bird feeding. A more complete description is available at the website.

There's a story behind these data, and it is worth hearing. Ryan's original intent was to compare bird feeding at the feeder in his back yard to bird feeding at a feeder in his girlfriend's back yard, where there was no construction next door.

**Question 1.** Give two possible confounds that might have been created by Ryan's original idea.

Unfortunately for Ryan, at the conclusion of the data collection period, which lasted over two months, he discovered his girlfriend was not very reliable about collecting data from her feeder, nor was she reliable about refilling it. So after two and half months of data collection, Ryan had no useable data from his control feeder. Fortunately, Ryan recorded EVERYTHING that happened at his feeder, including whether or not there was actually construction activity next door. So we decided to compare feeding on construction days to feeding on nonconstruction days from the one feeder. Moral of this story: record everything, whether you think you're going to use it or not!

In these data, the "constr" variable records whether or not there was construction activity (coded "no"
and "yes"), and the "food.eaten" variable records the amount of food eaten during the day in cubic centimeters of food per hour. Both variables are summarized. Look at the bottom of those columns in the summary table. What do you see?

**Question 2.** NA's? What the heck does that mean?

Cases in which there are NAs are not useful to us, so we're going to throw them out. (That is not always the correct thing to do, but here it will harm nothing.)

```
> RR = na.omit(RR)   # omit cases in which there are NAs
> summary(RR)

  constr     food.eaten
     no :25   Min.   : 2.000
    yes:43   1st Qu.: 6.652
          Median :10.115
             Mean :10.289
             3rd Qu.:13.330
              Max. :22.830
```

**Question 3.** How many observation days do we have complete data for?

Let's take a quick look at a graph to see what we can see.

```
> boxplot(food.eaten ~ constr, data=RR)   # not show here; look at your screen
```

**Question 4.** Looking at these boxplots, would you say these distributions are normal, positively skewed (long-tailed at the top), or negatively skewed (long-tailed at the bottom)? If they are skewed, does this appear to be a serious problem? (I.e., are they a lot skewed, or just a little bit skewed?)

**Question 5.** Does one group appear to be more variable than the other? If so, which one?

**Question 6.** Which group appears to be consuming more bird seed, the "no construction" group or the "construction" group?

Leave that graphics window open. We'll learn a few tricks with it in a moment. Let's take a look at the means and standard deviations of the groups.

```
> tapply(RR$food.eaten, RR$constr, mean)
     no       yes
   12.076400  9.249767
> tapply(RR$food.eaten, RR$constr, sd)
     no       yes
    5.591712  4.122538
```

Do those statistics confirm the answers you gave above? Let's pretty up the boxplots a bit. (If you closed the graphics window, you'll have to redraw the boxplots.) First, let's add the means to the boxplots. That's done with a function called points(), which plots points on a pre-existing graph. We have to give it the X and Y coordinates of the points we want to plot, and we have to tell it what kind of point we want.

```
> points(x=c(1,2), y=c(12.076,9.25), pch=16)   # plot two points; solid black dots
```

In this case, the means and the medians are fairly close together, indicating that our "problem" with
skewness is minimal. The means are a little higher than the medians in both groups, indicating that there is some skewness, however.

**Question 7.** When the means are higher than the medians, this indicates the presence of ____ skew.

Those dots would have to be explained in the figure caption, as they are not a standard part of boxplots. Now let's add some titles and axis labels to the graph. This is done with a function called `title()`.

```r
> title(main="Bird Feeding vs. Construction Activity")   # main title
> title(ylab="Bird Seed Consumed in cc/hr")   # a label for the y-axis
> title(xlab="Construction Activity")   # a label for the x-axis
```

There, now you have a graph that you shouldn't be ashamed to show someone. By the way, in R, if you make a mistake in the middle of drawing a graph, there is no undoing. You close the graphics window and start over again from scratch.

When you are finished admiring your graph, you can close the graphics window. We're done with it. Let's get to the t-test. I don't think the variances are so disparate that we need to involve ourselves with the Welch correction. We will do the good old-fashioned pooled-variance t-test.

```r
> t.test(food.eaten ~ constr, data=RR, var.eq=T)

Two Sample t-test

data:  food.eaten by constr
t = 2.3861, df = 66, p-value = 0.0199
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.4614517 5.1918134
sample estimates:
mean in group no mean in group yes
 12.076400  9.249767
```

**Question 8.** Why 66 degrees of freedom? What's the rule?

**Question 9.** We might phrase the null hypothesis here as construction activity does not disrupt bird feeding behavior. Are we rejecting that or retaining it? On what basis?

**Question 10.** Would you say this is a big effect? Estimate Cohen's d to support your answer.

**Question 11.** In a sentence, what is the confidence interval telling us?

Parametric statistical tests are ones that make a lot of assumptions about the parent distributions (i.e., the distributions of the populations we are sampling from). The t-test not only makes assumptions about the shape of the distributions (normal) but also about the variances (same in both groups). Nonparametric tests ease off on those assumptions a bit. For example, the Wilcoxon-Mann-Whitney test does not make the normality assumption, although it does assume that both groups will be sampled from populations that are similarly distributed. In R, this test uses the same syntax as the t-test.

The test is shown on the next page because I ran out of space on this one.

**Question 12.** Interpret.
wilcox.test(food.eaten ~ constr, data=RR)

Wilcoxon rank sum test with continuity correction

data:  food.eaten by constr
W = 692.5, p-value = 0.04937
alternative hypothesis: true location shift is not equal to 0

The warning message is not shown, but given how close we are to the alpha level with this result, an inaccurate p-value could be a problem! This illustrates the problem with most nonparametric tests: they are not as powerful as their parametric counterparts.

Question 13. What does that mean, not as powerful?

Bootstrap resampling is a method that can be used in place of almost any parametric procedure and is often almost as powerful as the parametric test. We did jackknife resampling in a previous lab. Bootstrap resampling is very similar. The difference is, in jackknife resampling we resample without replacement (scores are not returned to the "hat" after we draw them), which means all the scores in the original samples will be included in each round of resampling, just not in the same groups. Bootstrap resampling is done with replacement (scores are returned to the "hat" after we draw them), which means all the scores won't be used in each round of resampling while other scores will be used more than once. Here's a simple example. Suppose we have the following two groups

    group A: 12 18 15 13
    group B: 18 22 25 17

A jackknife resampling might look like this.

    group A: 17 15 13 25
    group B: 12 22 18 18

Notice that all the original values are used, just not in the same group. In the bootstrap, values usually stay in their own group, but not all of the values will be used, and some will be used more than once. It might look like this.

    group A: 12 18 18 12
    group B: 18 25 22 22

Why would anyone do such a goofy thing? The theory behind bootstrap resampling is that ALL the information we have about a population is what we see in the sample. For all we know, the population from which group A was sampled has only four values in it, and the distribution is somewhat positively skewed. By resampling with replacement, we are effectively creating an infinitely large population that has exactly the same characteristics as the sample.

The ideal way to see if two groups are different is to redo the experiment with new samples. All this statistical mumbo jumbo is fine as far as it goes, but if we really want to know if two populations are different (without looking at the entire population), we need to run the identical experiment a few hundred or a thousand times with new samples and see if we keep getting the same result. Who's up for that? More to the point, what funding agency is going to give you a grant to do that? Exactly!
And so we do the next best thing—statistics! Bootstrap resampling attempts to simulate redoing the experiment hundreds or thousands of times using everything we know about the population that we are sampling from to create new samples. And everything we know about the population is in that first sample. So we simulate the population by duplicating the sample an infinite number of times—i.e., we resample with replacement.

Let's pull the scores from the "yes" days and the "no" days out and put them into separate vectors in the workspace.

```r
> constr.yes = RR$food.eaten[RR$constr=="yes"]
> constr.no = RR$food.eaten[RR$constr=="no"]
> ls()
[1] "constr.no"  "constr.yes" "RR"
```

Those commands say, "give me the food eaten scores for days when construction is yes and days when construction is no." Just to make sure we did that right (even I'm not 100% sure on this one!), let's look at the means and standard deviations again.

```r
> mean(constr.yes)
[1] 9.249767
> mean(constr.no)
[1] 12.0764
> sd(constr.yes)
[1] 4.122538
> sd(constr.no)
[1] 5.591712
```

That looks right. Good for me. (By the way, if you can't get this to work, make sure you notice the double equal signs in the above syntax. Single equal signs are used for assignment only in R.)

```r
> constr.no
```

There are the food.eaten scores from days on which there was no construction. Let's do a bootstrap resampling of them.

```r
> length(constr.no)  # make sure we know how many there are
[1] 25
> sample(constr.no, size=25, replace=T)  # take a sample of 25 from that vector +  # with replacement
```

That would do us a lot more good if we actually saved it so we can calculate the mean and SD.

```r
> new.sample = sample(constr.no, size=25, replace=T)
> mean(new.sample)
[1] 11.1488
> sd(new.sample)
[1] 5.428522
```

**Question 14.** What mean and SD did you get? Hint: they won't be the same as mine, and they won't be the same as anyone else's.
That's the basis of bootstrap resampling. We just have to do that 1000 times or so with both groups and then compare the two new samples at each step of the way. But before we get that flashy, let's play with this one vector a little more. Open a script window and type the following script into it.

```r
R = 1000
result = numeric(R)
for (i in 1:R) {
    new.sample = sample(constr.no, size=25, replace=T)
    result[i] = mean(new.sample)
}
quantile(result, probs=c(.025,.975))
```

Do you understand what this script does? It's very similar to one we used in the lab on random error. The first line sets the number of resamples (or repeats) we are going to do. The second line creates a vector in which to store the results. Then there is a for loop that does the 1000 resamples, calculates the mean of each one, and stores the result in the result vector at place i. We exit the for loop, and then the last line finds the value that leaves 2.5% of the (sorted) values below it and the value that leaves 97.5% below it, or 2.5% above it. In other words, it finds the two resampled means that have 95% of the resampled means between them. That's a bootstrapped 95% confidence interval.

Execute the script.

```
2.5% 9.87954
97.5% 14.27140
```

**Question 15.** What result did you get? (It will be different from mine.)

**Question 16.** Now here's a question that will separate those who have been paying attention from those who haven't been. Calculate a 95% confidence interval around the mean analytically from the original data in constr.no. What is it? (Okay, hint. You need the standard deviation, which we calculated above. You need to divide that by the square root of the sample size to get an estimated standard error of the mean. Then you need to multiply that by 2.064 (from the t table for df=24). Finally, subtract that number from the sample mean to get the lower limit of the confidence interval, and add it to the sample mean to get the upper limit of the confidence interval.

My bootstrapped confidence limits were pretty close to the analytical ones. How did you do?

**Question 17.** What does in mean? In one sentence, explain what your bootstrapped confidence interval tells you.

Time for the main event. We are now going to bootstrap a t-test, or more correctly, a replacement for the t-test. The actual difference between the two sample means was:

```
> mean(constr.no) - mean(constr.yes)
[1] 2.826633
```

That is the statistic that we want to bootstrap. What we want to find out is, when we do the bootstrap resampling, what proportion of the time will we get a mean difference that large or larger?

Let's begin by cleaning up a little bit.
Also, erase the script from the script window. Here's the new script. See if you can follow what it's doing.

```
# set up the number of reps and the results vector
R = 1000
result = numeric(R)
# now do the actual resampling and store the mean differences
for (i in 1:R) {
  new.sample.no = sample(constr.no, size=25, replace=T)
  new.sample.yes = sample(constr.yes, size=43, replace=T)
  result[i] = mean(new.sample.no) - mean(new.sample.yes)
}
# now we summarize the results
summary(result)
quantile(result, probs=c(.025,.975))
sum(result<=0)
```

Here is the output I got when I executed the script. Yours will be different.

```
> summary(result)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.564  1.977   2.840   2.854   3.657   7.156
> quantile(result, probs=c(.025,.975))
                 2.5%       97.5% 
0.4168365       5.2999663
> sum(result<=0)
[1] 12
```

**Questions 18-24.** Go to the answer sheet and fill in the blanks.
Question 1. Give two possible confounds that might have been created by Ryan's original idea.

(a) ________________________________________________________________

(b) ________________________________________________________________

Question 2. NA's? What the heck does that mean? (NO MORE than two words here!)

_____________  _______________

Question 3. How many observation days do we have complete data for? __________

Question 4. Looking at these boxplots, would you say these distributions are normal, positively skewed (long-tailed at the top), or negatively skewed (long-tailed at the bottom)? If they are skewed, does this appear to be a serious problem?

circle: normal positively skewed negatively skewed

circle: it's not too bad it's really bad!

Question 5. Does one group appear to be more variable than the other? If so, which one?

The ____________________ group is the more variable of the two groups.

Question 6. Which group appears to be consuming more bird seed, the "no construction" group or the "construction" group?

Question 7. When the means are higher than the medians, this indicates the presence of

positive / negative (circle one) skew.

Question 8. Why 66 degrees of freedom? What's the rule?

Question 9. We might phrase the null hypothesis here as construction activity does not disrupt bird feeding behavior. Are we rejecting that or retaining it? On what basis?

Question 10. Would you say this is a big effect? Estimate Cohen's d to support your answer.
Question 11. In a sentence, what is the confidence interval telling us?

Question 12. Interpret the result of the Wilcoxon-Mann-Whitney test.

Question 13. What does that mean, not as powerful?

Question 14. What mean and SD did you get? mean = ____________  SD = ____________

Question 15. What result did you get?

2.5%       97.5%

________    _______

Question 16. Calculate a 95% confidence interval around the mean analytically from the original data in constr.no. What is it?

lower limit = ____________  upper limit = ____________

Question 17. What does in mean? In one sentence, explain what your bootstrapped confidence interval tells you.


I found that 75% of my bootstrapped mean differences fell between _________ and __________, and that the bootstrapped 95% confidence interval for the mean difference was _________ to __________. This indicates that the sample mean difference of 2.826633 was _______________________________ from zero at the alpha = .05 level. In fact, only _________ times out of 1000 was the bootstrapped mean for the days when there was construction greater than or equal to the mean of the days when there was no construction. That indicates a bootstrapped p-value of _________. (You can get it. Just think about it!)