Lab Exercise 2 - Random Error

There is an answer sheet attached to this lab (last page). CAREFULLY remove it, put your name on it, and write your answers in the spaces provided. Be prepared to hand it in on Thursday (24th). Yes, I will collect it.

For this exercise, to keep everyone on the same page until we get used to R, and in case anyone doesn't yet have access to it, we will use the web interface to R. Go to my website for this course, scroll down, and find the link to Rweb in the box labeled R Stuff For People Who Want It. When the Rweb website comes up, you will be asked which version of Rweb you want. Choose the simplest one. (I've had problems with the others, but this one always works.)

R Stuff For People Who Want It

- Link to the R Project home page (get R here) (External link)
- Download page for RStudio (optional) (External)
- Rweb (online interface to R) (External)
- R-Tutorials
- Three YouTube Videos About R (External)
  - What's Up With R
  - Reasons To Learn the R Programming Language
  - Obtaining and Installing R

Other Stuff For People Who Want It

Answer this one! Which Version Do You Want?

Rweb
A simple web interface to R that works on most!
Rweb only requires a browser that can handle it

JavaScript Version of Rweb
This is the same Rweb as above but with a few d

Rweb modules
These modules are designed as a point and click
analysis, and the options for the analysis. The cut
graphs which can be cut and pasted into project

That will bring up the webpage you see illustrated on the next page of this handout. Refer to that illustration now. To get Rweb ready for this lab, do the following. Type the following URL into the box labeled Enter a dataset URL.

http://ww2.coastal.edu/kingw/psyc480/data/marijuana.txt

It will have to be typed exactly as you see it here, of course. This will load the marijuana data into a data frame called X. To check that everything is working, go to the big empty box where you type commands and type summary(X). Then click the Submit button.

That will bring up another webpage with the results of your analysis on it. To get back to the Rweb page, click the back button on your browser. These steps are illustrated on the next page.
To run Rweb just type the R (or Splus) code you want to execute into the text window and then click on the submit button.

Below the submit button is a text area where you can enter the URL for a Web accessible dataset and a browser title. The URL for the dataset is "http://www.coastal.edu/king/psych480/data/marijuana.txt". The data frame, X, will then be attached so you can use the variables.

If you use the back button on your browser to come back to this page you can modify your old code and then re-run it.

The output is shown on the next page. Refer to that now as you read the following.

You should see the usual opening screen for R. Below that, the command prompt Rweb : > appears. The first two commands read in the dataset from my website and store it in a data frame called X. The names of the variables are then displayed as a summary. Finally, after a couple blank command prompts, your summary(X) command appears, followed by the output. Once you're happy this is working, click the back arrow next to the address box (Firefox is illustrated but similar in other browsers) to get back to the Rweb page.
Results from Rweb

You are using Rweb 1.03 on the server at rweb.stat.umn.edu

Reading data from URL http://ww2.coastal.edu/kingw/psyc480/data/marijuana.txt

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Rweb: > png(file = "/tmp/Rout.18100+%03d.png")
Rweb: > X <- read.table("/tmp/Rdata.18100.data", header=T)
Rweb: > attach(X)
Rweb: > names(X)
[1] "group" "score"
Rweb: >
Rweb: > summary(X)
group    score
      nonsmoker:10  Min. :13.00
       smoker :10  1st Qu.:17.00
               Median :19.00
               Mean :18.65
               3rd Qu.:21.00
               Max. :23.00
Rweb: >
Rweb: >

Pretty cool, right? Okay, maybe that isn't an entirely appropriate use of the word "cool," but it's at least convenient. Rweb can do just about anything that R can do, but going back and forth between the command window and output page can be a bit tedious. The best way to use Rweb is by scripting. Here is an example. Erase the summary command (DO NOT click the Erase Everything button!), and type these commands into the command box. (You can copy and paste from this handout if you have it up in your browser.) Then click submit.

tapply(score, group, mean) # group means
tapply(score, group, sd) # group sds
t.test(score ~ group, var.eq=T) # pooled variance t-test
aov(score ~ group) # partitioning the variability
You should get an output page with the results of all those commands on it. Notice the first thing Rweb did was to read in the dataset again. Unlike desktop R, in Rweb your workspace is temporary. Every time you submit commands to it, it will have to read in your data again. So make sure it's available.

Copy the group means onto the answer sheet (Question 1). Then click the back arrow to return to the Rweb input page. Erase the commands in the command box by highlighting them with your mouse and pressing Delete on your keyboard. If you get too rambunctious and press the Erase Everything button, that's not a disaster, but you will have to retype the dataset URL into the dataset URL box before you can continue.

Question 2. What is the difference between the group means. (You can do this in Rweb if you have to, or use a calculator, or just subtract the means in your head.)

Why does this difference exist? Anytime you see a difference in ANY statistic (not just the mean) between (or among) groups in a sample, there are always three possible reasons for it. (I will ask you this over and over again, so memorize it!)

1) The difference could be a real effect of the independent variable.
2) The difference could be an effect of a confound that the experimenter failed to control for.
3) It could be nothing more than random error.

We are dealing with a sample. Always remember and don't ever forget! Samples are noisy! That is, they contain random error. The mean of the smoking group is probably not the true mean for all people who smoke marijuana (the population mean), the mean of the nonsmoking group is probably not the true mean for all people who do not smoke marijuana, and the difference between the sample means probably isn't right either. All of those statistics are noisy. Think of them as being fuzzy estimates of what is "really true." One of our jobs as statisticians is to figure out how fuzzy those estimates are. That is, how much random error is there likely to be in these statistics.

Suppose we are sampling (at random) from a population of marijuana nonsmokers, and suppose the population has a mean digit span score of 20 with a standard deviation of 2.5. We are taking a sample of size n = 10. What sample mean are we going to get? We can simulate this in R. To the command box! (Remember to erase the previous commands unless you want them to be executed again.)

```r
my.sample = rnorm(n=10, mean=20, sd=2.5)
mean(my.sample)
```

Question 3. What sample mean did you get? Write out ALL the decimal places that R gives you. WARNING! The function rnorm() generates a random sample of size n from a normal distribution with the mean and sd we have specified. The key word in that warning is random. Your answer will not be the same as anyone else's. I will check!

Okay, do it again. Go back to the input page and click the Submit button again.

Question 4. What sample mean did you get this time? Write all decimal places, then do it a third time.

Question 5. What sample mean did you get the third time? (Write all decimal places.)
Your three answers should be different. You are sampling from the same population with the same mean but getting different answers. (No guarantees, but it would be unlikely that you got two sample means the same. If you did, write it down that way.) That's noise, more correctly called random error or random sampling error. All samples have it! It's not hard to imagine why. You're taking ten scores at random from a hat full of thousands or millions of scores. You will get different scores each time and, therefore, different means. It's kind of like the lottery.

How different from the true value of 20 is it possible for these sample means to get? Well, let's do 1000 samples and see what they look like. Anybody up for that? Very fortunately, you do not have to click that Submit button a thousand times and copy out all your answers. You're about to become an R programmer! Erase the command box and type the following.

```r
repeats = 1000
my.sample.means = numeric(repeats)
for (i in 1:repeats) {
  my.sample = rnorm(n=10, mean=20, sd=2.5)
  my.sample.means[i] = mean(my.sample)
}
summary(my.sample.means)
sd(my.sample.means)
```

What does it mean? First of all, here's something I was unaware of until now. You cannot use the word "repeat" as an R object (variable). It has to be "repeats." I have no idea why. Anyway, we want to "repeats" this sampling procedure 1000 times. We create a numeric data object (called a vector in R’speak) in which to store the means of our samples. Then we start the sampling with a "for loop." This loop will execute all the commands between the curly braces until it runs out of repeats, and each time it will store the sample mean in place i in the my.sample.means vector. Don't get it? Don't worry. It'll work. Once we're done with all that, we do a summary of the sample means and also calculate the standard deviation of them.

**Question 6.** What does the summary tell you about your sample means? (Once again, random!)

**Question 7.** What is the standard deviation of your sample means? (All decimal places.)

The mean and median of your sample means should be very close to 20, but the individual sample means will not all be close to 20. I got means as low as 17.77 and as high as 22.91. Your answers will be different, of course. In principle, any result is possible. If scores on the digit span test range from 0 to 30, then we could get means as low as 0 and as high as 30. Obviously, that's very unlikely, but in principle it's possible. (Here's something to think about, if you want a puzzle. How did I get a mean of 17.77 if I had a sample of 10 whole numbers? It's a good practice to get in the habit of making sure your answers make sense!)

The standard deviation of a bunch of sample statistics from repeated samples is called the estimated standard error of that statistic. The sd we calculated above is an estimate of the standard error of the mean. In my case, it came out to be 0.8120614. Your answer will again probably be different, but probably not too different. The true standard error of the mean is the population standard deviation (2.5 in this case) divided by the square root of the sample size (10). I'm guessing based on my estimate from the simulation that the true standard error of the mean (i.e., in the population) is about 0.8.
**Question 8.** What is the true standard error of the sample mean?

The *standard error* is a statistic that gives us an idea of the amount of random error (noise or fuzziness) that is likely to occur in a sample statistic. What is the true mean of the population of marijuana nonsmokers? We don't actually know. However, if we are sampling from a normal distribution (or something reasonably close to that) and the sampling is adequately random, and the sample is large enough, then the true population mean will probably be within a standard error of the sample mean, and will almost certainly be within two standard errors. The sample mean of the nonsmokers in Scott Keat's marijuana data was 20.1, and if we estimate the standard error at 0.8, then, provided the sampling was done with adequate randomness, the true sample mean is probably between 20.1 - 0.8 and 20.1 + 0.8, or between 19.3 and 20.9. Can you think of three reasons why the true population mean may NOT be between these two values?

**Question 9.** Give three reasons why the true population mean for Scott's marijuana nonsmokers may not be between 19.3 and 20.9. (If you're confused by this question, reread the previous paragraph.)

Anything calculated from a sample is a *statistic*. Scott took a sample of people, divided them into smokers and nonsmokers, gave them the digit span test, and then calculated from his sample that the difference between the two group means was, well, look at your answer to question 2. Thus, the difference between the two sample means is also a statistic and, as all statistics do, it has a standard error. You learned how to calculate that standard error in the previous lab.

In general (whether the design is balanced or not), it is:

$$\text{pooled standard deviation} \times \sqrt{\frac{2}{\text{effective group size}}}$$

In the case of a balanced design, the effective group size is just the group size (10 in this case). The pooled standard deviation can be obtained most easily by using the `aov()` function. Back to the Rweb command box we go. Erase our beautiful R program and do this.

```r
aov(score ~ group)
```

In the output, the number labeled "Residual standard error" is the pooled standard deviation.

**Question 10.** What is it? (All decimal places please.)

**Question 11.** So what is the standard error of the difference between the means? (Three places is fine.)

The difference between the means from Scott's data is not the true difference between the means, it is the difference between the sample means, and samples are ...? Right, noisy! But it's probably within one standard error of the true difference, and almost certainly with two standard errors, if the sample is large enough. This sample is not large enough (only 20), so the actual multiplier we need to use in this case is not 2 but 2.1. Thus, it is almost certainly true that the true difference between the means (in the population or general case, that is) is somewhere within these bounds.

$$\text{lower limit} = \text{sample difference} - 2.1 \times \text{standard error of difference}$$

$$\text{upper limit} = \text{sample difference} + 2.1 \times \text{standard error of difference}$$
Question 12. What are those limits? (You'll have to do this one by hand.)

Question 13. Where have we seen these limits before (to within reasonable rounding error)?

Here's a hint to help you with question 13, if you need it.

\[ \text{t.test}(\text{score} \sim \text{group}, \text{var.eq}=\text{T}) \]

Now for the $64K$ question: What does a t-test do? The t-test often seems very complex to people who first encounter it, but what it does is actually very simple.

\[
\begin{align*}
\text{what it seems like} & \quad \text{what is actually true} \\
\frac{\bar{X} - \mu}{s_{M}} & \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_{p} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - (\mu_1 - \mu_2)
\end{align*}
\]

The t-test takes a difference between a sample statistic (M) and the value the null hypothesis says it should have, and then divides that by the estimated standard error of the sample statistic. Why all this fuss about the *null hypothesis*? Isn't it the *alternative (or experimental) hypothesis* we are really interested in? Of the two, the null hypothesis is the only one that makes a precise mathematical statement. It says the difference between the group means, in the case of two groups, is zero. (Technical note: it doesn't actually have to say zero, but it does have to say some exact value.) The alternative hypothesis says, basically, no it isn't, which is not a precise statement. Therefore, it's the null hypothesis that will be used in our statistical calculations.

In other words, a t-statistic is a ratio of how far off our sample result is from what the null says it should be, to the size of the (estimated) standard error or fuzz or noise in the sample. If the difference in the numerator is considerably larger than the noise in the denominator, then we conclude that the difference is probably not entirely due to random error (noise). That is what we mean, and it is ALL we mean, when we say the difference is statistically significant. We are not saying the difference is important, or even that it is large. We are saying ONLY that it is probably not entirely due to noise.

Go to the Rweb command box, erase whatever is there, type X, and click Submit. Rweb will print out the data frame for you. In the column labeled score we see the digit span scores. The first 10 are from the smokers, and the second 10 are from the nonsmokers. We know from answering question 2 that the nonsmokers' scores are, on average, a little larger than the smokers' scores. The null hypothesis says essentially that these scores have nothing to do with whether or not the person smokes marijuana, and if marijuana smoking is the only difference between these two groups (no confounds), then the difference in the means should be zero. But the difference is not zero. It's whatever your answer to question 2 is. Yeah, says the null, that's just an accident of random sampling. It's just dumb luck that
we're seeing this difference. It has nothing to do with smoking. These particular subjects would have gotten these scores whether or not they were smokers.

Really? Says the alternative hypothesis. Then you won't mind if we change history and make some of the smokers into nonsmokers and some of the nonsmokers into smokers. They'll still get the same scores, right? Of course, says the null. And, says the alternative, if we do that, it should be fairly easy to get a difference between the groups like the one we are seeing here. It's just random noise, so you say, and there's always random noise, so a difference this large is no big deal. Correct, says the null. Then let's do that, says the alternative, and see what happens.

We're going to do it by just randomly changing people from nonsmokers to smokers and vice versa. We'll keep the sample at 10 each. Go to the Rweb command box, erase whatever is there, and type:

```
sample(group, size=20, replace=F)
```

Click Submit and see what happens. This is what happened for me. Your result will be different.

```
Rweb:> sample(group, size=20, replace=F)
[1] smoker    nonsmoker smoker    nonsmoker nonsmoker smoker    nonsmoker
[8] nonsmoker nonsmoker smoker    smoker    smoker    nonsmoker smoker
[15] smoker    nonsmoker nonsmoker smoker    smoker    smoker    nonsmoker
Levels: nonsmoker smoker
```

We have taken a new random sample of size 20 from the `group` variable, a sample that includes everyone in `group` (without replacement). This procedure is called resampling. Essentially what we've done is turned some of the smokers into nonsmokers and some of the nonsmokers into smokers. We have not altered their digit span scores. Let's do that again, and then use or new grouping variable to calculate group means, and see what we get.

```
new.groups = sample(group, size=20, replace=F)
tapply(score, new.groups, mean)
```

**Question 14.** What were the means for your new groups?

**Question 15.** What is the difference between those means?

My results were 19.9 for the nonsmokers, 17.4 for the smokers, for a mean difference of 2.5. (Your results will be different.) That's a difference of a pretty respectable size. It's not quite as big as the original difference, but still, it suggests that differences that large can occur by just random error (simulated by randomly remixing the groups). I'd be willing to bet your new mean difference was smaller than that, perhaps a lot smaller. Guess what we have to do now. You got it!

```
repeats = 1000
my.mean.diffs = numeric(repeats)
for (i in 1:repeats) {
  new.groups = sample(group, size=20, replace=F)
  new.means = tapply(score, new.groups, mean)
  my.mean.diffs[i] = diff(new.means)
}
summary(my.mean.diffs)
quantile(my.mean.diffs, probs=c(.025, .975))
```
Question 16. What were your results? (Record them on the answer sheet.)

My results (need I even say it--yours will be different!) were:

```
Rweb:> summary(my.mean.diffs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -3.7000  -0.7500   0.1000   0.0146   0.9000   3.9000
Rweb:> quantile(my.mean.diffs, probs=c(.025, .975))
     2.5%    97.5%
     -2.5     2.3
```

What does it mean? This procedure is called a resampling test or, if you want a cooler name for it, it's called jackknife resampling. It tells us what we can expect from random remixing of the group labels. I got new mean differences as large as -3.7 and +3.9. Those are more extreme than the original mean difference of (see your answer to question 2), telling us that the original mean difference certainly could have been nothing but random error. But look at the quartiles. 75% of the results fall between the quartiles, so 75% of my new mean differences were between -0.75 and +0.9, which are nowhere near the size of the original mean difference (question 2). So although we CAN get a difference that large, it doesn't appear to be likely. The quantile command asks for the middle 95% of the new mean differences, leaving out only the bottom 2.5% and the top 2.5%. Those limits tell me that 95% of the time the new mean difference was between -2.5 and +2.3, neither one of which is as large as the original mean difference. So it appears to be quite difficult to get a mean difference as large as the original one, if we are relying on nothing but random error to get it. Something else is very probably going on here. Or to put it another way, the original mean difference was probably not entirely due to random error. Or to put it yet another way:

Question 17. What two words could now be used to describe the original mean difference in the sample?

You may not have found the same thing. So:

Question 18. Summarize your own results.

Well hey! That seems much easier to understand than a t-test. Why don't we always do it that way? The t-test was invented in 1908.

Question 19. Were there computers in 1908?

If someone had wanted to do this in 1908, they would have been doing it by hand. Anybody up for that? I didn't think so. Hence, the t-test was invented. The t-test was essentially a way to calculate in a few minutes what would have taken several days to do by resampling. You can think of the t-test as a mathematical approximation to jackknife resampling. There is a second type of resampling called bootstrap resampling, which is used more often these days. We'll talk about the difference at another time.

Question 20. With the pooled variance t-test, we found a p-value of 0.016. In other words, p < .05. Would it be reasonable to say from YOUR resampling results that p < .05? Briefly explain.
Question 1. Group means for the marijuana data: non-smoker ________  smoker ________

Question 2. Difference between the means: __________

Question 3. What sample mean did you get? __________

Question 4. What sample mean did you get this time? __________

Question 5. What sample mean did you get the third time? __________

Question 6. What does the summary tell you about your 1000 sample means?

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>__________</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
</tr>
</tbody>
</table>

Question 7. What is the standard deviation of your 1000 sample means? __________

Question 8. What is the true standard error of the sample mean? __________

Question 9. Give three reasons why the true population mean for Scott's marijuana nonsmokers may not be between 19.3 and 20.9.

a) _______________________________________________________________________

b) _______________________________________________________________________

c) _______________________________________________________________________

Question 10. What is the pooled standard deviation in the marijuana data? __________

Question 11. What is the standard error of the difference between the means? __________

Question 12. Within what limits will we be very likely to find the true mean difference?

  lower limit = ________  upper limit = ________

Question 13. Where have we seen these limits before?

Question 14. What were the means for your new groups? non-smoker ________  smoker ________

Question 15. What is the difference between those means? __________
**Question 16.** What were your results from the resampling procedure?

```r
Rweb:> summary(my.mean.diffs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   _______ _______ _______ _______ _______ _______
```

```r
Rweb:> quantile(my.mean.diffs, probs=c(.025, .975))
          2.5%       97.5%
   _______ _______
```

**Question 17.** The original mean difference of ________ (copy your answer from question 2 into this blank) was ____________________  ____________________ from 0 (two words), according to Dr. King's resampling results.

**Question 18.** My jackknife resampling results show that the original mean difference of ________ (copy your answer from question 2 into this blank) was / was not   (circle one of those choices) ____________________  ____________________ from 0 (two words).

**Question 19.** Were there computers in 1908? __________

**Question 20.** With the pooled variance t-test, we found a p-value of 0.016. In other words, p < .05. Would it be reasonable to say from YOUR resampling results that p < .05? Briefly explain.