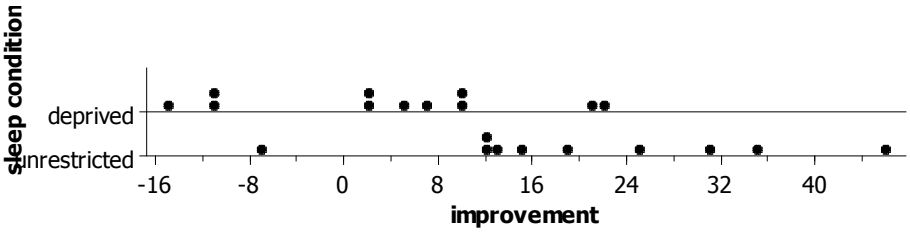
Unit 3 Assignment: Randomization methods for comparing two groups (quantitative dependent variable)

Source: Allan Rossman: http://statweb.calpoly.edu/arossman/stat325/notes.html

Researchers have established that sleep deprivation has a harmful effect on visual learning, but is it possible to "make up" for sleep deprivation by getting a full night's sleep in subsequent nights? A recent study (Stickgold, James, and Hobson, 2000) investigated this question by randomly assigning 21 subjects (volunteers between the ages of 18 and 25) to one of two groups: one group was deprived of sleep on the night following training and pre-testing with a visual discrimination task, and the other group was permitted unrestricted sleep on that first night. Both groups were then allowed as much sleep as they wanted on the following two nights. All subjects were then re-tested on the third day. Subjects' performance on the test was recorded as the minimum time (in milli-seconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen. The sorted data and dotplots presented here are the improvements in those reporting times between the pre-test and post-test (a negative value indicates a decrease in performance):

Sleep deprivation (n = 11): -14.7, -10.7, -10.7, 2.2, 2.4, 4.5, 7.2, 9.6, 10.0, 21.3, 21.8
Unrestricted sleep (n = 10): -7.0, 11.6, 12.1, 12.6, 14.5, 18.6, 25.2, 30.5, 34.5, 45.6



1. Based on the dotplot, does it appear as though the subjects who got unrestricted sleep on the first night tended to have higher improvement scores than subjects who were deprived on the first night? Briefly explain your answer.

2. Calculate the median improvement score for each group. Do you believe the median improvement score is *significantly* higher for those who got unrestricted sleep?

3. Notice that 9 of the 10 lowest improvement scores belong to subjects who were sleep deprived. Also, the mean improvement score for the sleep-deprived group was 15.92 ms smaller than the unrestricted group mean (3.90 ms vs. 19.82 ms). Keep in mind this information, along with the dotplot and medians you calculated earlier as you answer this question: Is it possible that there is really no harmful effect of sleep deprivation and random chance alone produced the observed differences between these two groups?

The key question is how likely would it be for random chance to produce experimental data that favor the unrestricted group by at least as much as the data we observed do.

We will try to answer this question using the same simulation analysis we used in analyzing yawns and dolphins.

1. Randomize. We will assume sleep deprivation has no negative effect (the null model) and replicate the random assignment of the 21 subjects (and their improvement scores) between the two groups.
2. Repeat. We will repeat this random assignment a large number of times and calculate a measure of how different the groups are in order to get a sense for what is expected and what is surprising.
3. Reject? If the results we observed in our study are in the tail of the null model's distribution, we will reject that null model.

In this analysis, we will calculate the difference in mean improvement scores between the two groups after each new random assignment. If we do this a large number of times, we will have a good sense for whether the difference in group means that the researcher observed is surprising under the null model of no real difference between the two groups. Note that we could just as easily use the medians instead of the means, which is a nice feature of this analysis strategy.

4. To manually simulate this experiment, take 21 index cards and write one of the observed improvement scores on each card. Next, shuffle the cards and randomly deal 11 cards for the sleep deprivation group. The remaining 10 cards will represent the unrestricted sleep group.

   Run this simulation and write your results in the table below (on the lines under the Replication 1 column). Shuffle the cards again and run the simulation 4 more times to complete the table.

Replication	**1**	**2**	**3**	**4**	**5**
Unrestricted sleep mean	_____				
Sleep deprivation mean	_____				
Difference in group means	_____				

5. Combine your results (the differences in group means) with those of your classmates. How many positive differences did everyone get? How many negative differences? Did you get any differences of exactly zero? Produce a dotplot below to display everyone's results.

6. Did you get any results as extreme as the researcher's actual results (difference of 15.92)? Do the results from you and your classmates suggest there is a statistically significant difference between the two groups? Explain.

7. Let's use technology to more efficiently simulate 1000 replications of this experiment. We will get the computer to:
   - Randomly assign groups to the 21 improvement scores (11 sleep deprived and 10 unrestricted sleep)
   - Calculate the difference in group means
   - Store that difference
   - Repeat 1000 times
   - Produce a display of the results and calculate the proportion of results that are at least as extreme as what was observed

   Open the *Randomization Tests* applet at http://www.rossmanchance.com/applets/randomization20/Randomization.html and notice that the experimental data from this study already appears. Notice, once again, the observed difference between group means was 15.92.

   Click on *RE-RANDOMIZE* to randomize the 21 improvement scores between the two groups. Notice the new difference between group means (calculated from this randomization). Click RE-RANDOMIZE one more time to ensure you get a different difference between group means (although it *is* possible for you to get the same difference twice).

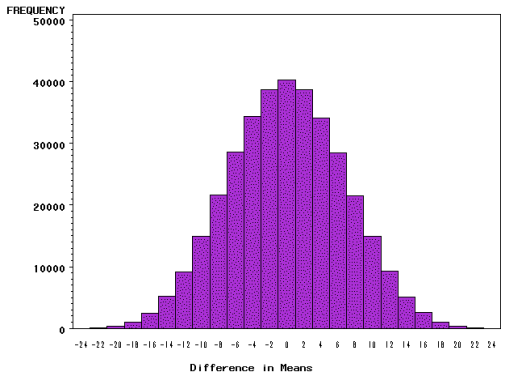
   Now un-check the *ANIMATE* feature and run 998 more randomizations. Look at the distribution of the 1000 simulated differences in group means. Is the center where you would expect? Does the shape have a recognizable pattern? Explain.

8. Count how many of your 1000 simulated differences in group means are as extreme (or more extreme) than what the researchers actually observed (greater than 15.92). To do this, you can enter 15.92 in the *COUNT SAMPLES BEYOND* field and click *GO*. What approximate p-value did you obtain from your simulation?

9. Do these simulation analyses reveal that the researchers' data provide strong evidence that sleep deprivation has harmful effects three days later? Explain.

10. Even if you found a statistically significant difference between the mean improvement scores of the two groups, is it legitimate to draw a cause-and-effect conclusion between sleep deprivation and lower improvement scores? Explain. Hint: Ask yourself whether this was a randomized experiment or an observational study.

If you are wondering whether an exact mathematical calculation of the p-value is possible here, the answer is yes. But the calculation is more difficult than with yes/no variables. It turns out that there are 352,716 different ways to assign 21 subjects into one group of 11 and another group of 10. Of these 352,716 possible randomizations, it turns out that 2533 of them produce a difference in group means (favoring the unrestricted sleep group) of at least 15.92. The exact p-value is therefore 2533/352,716 ≈ .0072. The exact randomization distribution is shown here:



Extra credit assignment: Re-do this simulation analysis using the difference in group *medians* rather than means. Don't do this by hand; use the website applet. Report the p-value and summarize your conclusion. Describe how your conclusion differs (if at all) from the analysis we conducted based on group means.

11. Let's investigate one more way in which we might analyze this data using randomizations. Most introductory statistics students, if asked to analyze this data, would run an independent samples t-test. Here's the syntax used to run a t-test in Stata along with the output:

Syntax:
```
input improve group
-14.7 1
... (All 20 observations are entered. Let's skip to the last one…)
45.6 0
end

ttest improve, by(group)
```

Let's see if we can make sense of the syntax. The first part (which we don't really need to use) inputs the data into Stata. The first line tells Stata we will have two variables: *improve* and *group*. The *improve* variable contains all the improvement scores. The *group* variable is either 0 (for unrestricted sleep) or 1 (for sleep deprived). Note that I could have chosen any names for the variables or values for the groups. The *end* command tells Stata I have finished entering data.

The last line of syntax runs the t-test. The general form for this syntax is: ttest *DependentVariable*, by(*IndependentVariable*).

The output I got from this syntax is:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      10       19.82    4.656556    14.72532    9.286139    30.35386
       1 |      11         3.9    3.669952    12.17185   -4.277162    12.07716
---------+--------------------------------------------------------------------
combined |      21    11.48095    3.366725    15.42827    4.458087    18.50382
---------+--------------------------------------------------------------------
    diff |            15.92    5.873229                 3.62719    28.21281
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.7106
Ho: diff = 0                                     degrees of freedom =       19

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9931         Pr(|T| > |t|) = 0.0139          Pr(T > t) = 0.0069
```

I want to be clear – you should not run a t-test like this without investigating assumptions and visual displays of the data! See if you can make sense of this output. The **bold text** displays our alternate hypothesis and p-value obtained from this study.

As practice, go ahead and calculate an independent-samples t-test by hand for this data. Looking at the Stata output, you can find means, standard deviations, and sample sizes to speed up your calculations. You can also find the standard error in there if you know where to look. Use the space below to show that you are able to calculate the t-statistic to be 2.7106.

12. The t-statistic of 2.7106 and the p-value of 0.0069 were calculated from the observed data. Let's see if we can support the results of this study through the use of randomizations.

    The following Stata code runs 1,000 randomizations of the t-test. That is, for every randomization, the program calculates a t-statistic.

    Syntax: `permute group t=r(t), reps(1000) nodots: ttest improve, by(group)`

    The following output was obtained:

```
Monte Carlo permutation results                 Number of obs = 21

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   2.710604      17    1000  0.0170  0.0041  .0099335   .0270795
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```

    N=1000 replications

    2.7106 is the t-statistic we observed in our study.

    C = 17. This tells us 17 of the 1000 replications were greater than 2.7106. So, 1.7% of our replications gave us results as or more extreme than what we observed.

    What do these results tell us? Do they support the conclusions we would have made from our t-test?

EXTRA CREDIT OPPORTUNITY

Example 2: Age Discrimination?

Robert Martin turned 55 in 1991. Earlier in that same year, the Westvaco Corporation, which makes paper products, decided to downsize. They ended up laying off roughly half of the 50 employees in the engineering department where Martin worked, including Martin. Later that year, Martin went to court, claiming that he had been fired because of his age. A major piece of evidence in Martin's case was based on a statistical analysis of the relationship between the ages of the workers and whether they lost their jobs.

Part of the data analysis presented at his trial concerned the ten hourly workers who were at risk of layoff in the second of five rounds of reductions. At the beginning of Round 2, there were ten employees in this group. Their ages were 25, 33, 35, 38, 48, 55, 55, 55, 56, 64. Three were chosen for layoff: the two 55-year-olds (including Martin) and the 64-year old.

What to make of these data requires balancing two points of view, one that favors Martin, and the other that favors Westvaco. To understand the two views, imagine a dialog between two people, one representing Martin, the other, Westvaco:

Martin: The pattern in the data is very striking. Of the five people under age 50, all five kept their jobs. Of the five over age 50, only two kept their jobs. The average age of those chosen to lose their jobs is 58 years; that's way above the average for the whole group. The pattern is clear evidence of discrimination.

Westvaco: Not so fast! Your sample is way too small to be evidence of anything. There are only ten people in all, and only three in the group that got fired. How can you expect me to take your patterns seriously when just a small change will destroy the pattern? Look how different things would be if we just switch the 64-year old and the 25-year old:

Actual data:	25 33 35 38 48 **55** **55** 55 56 **64**	Avg age: 58
"What-if" data:	**25** 33 35 38 48 **55** **55** 55 56 64	Avg age: 45

Martin: But look at what you did: You deliberately choose the oldest one fired and switched him with the youngest one not fired. Of all the possible choices, you picked the most extreme. Why not compare what actually happened with *all* the possible choices?

Westvaco: What do you mean?

Martin: Start with the ten workers, and treat them all alike. Let random chance decide which three get chosen for layoff. Repeat the same process over and over, to see what typically happens. Then compare the actual result with what typically happens.

The issue here is whether the data suggest that Westvaco was not acting in an age-neutral manner. One way to address that is to consider randomness as an age-neutral device for deciding which employees to lay off.

1) **Write out the null hypothesis (or null model) for this situation.**

2) **Conduct (or explain how you would conduct) a simulation analysis to investigate whether or not Westvaco engaged in age discrimination. If you conduct the simulation, calculate your p-value and explain what it represents. If you explain how you would conduct the simulation, write out some syntax or instructions for conducting the simulation (possibly using index cards) and describe how you would interpret the results.**