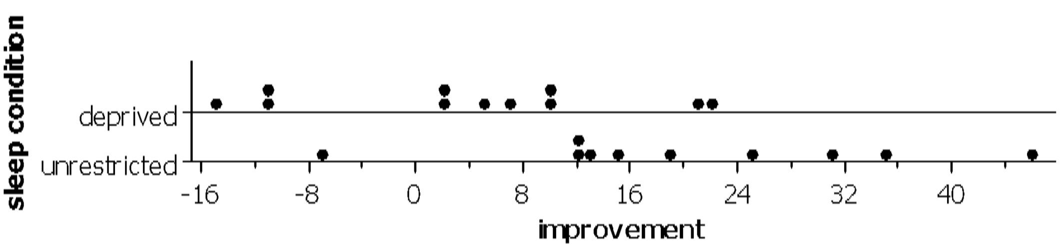
Lingering Effects of Sleep Deprivation (randomization test for comparing means)
Researchers have established that sleep deprivation has a harmful effect on visual learning. But do these effects linger for several days, or can a person "make up" for sleep deprivation by getting a full night's sleep in 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.

a) Identify the explanatory and response variables in this study. Also classify them as either categorical or quantitative.

b) Was this an experiment or an observational study? Explain how you are deciding.

The sorted data and dotplots presented here are the improvements in those reaction times (in milliseconds) 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



c) Does it appear that subjects who got unrestricted sleep on the first night tended to have higher improvement scores than subjects who were sleep deprived on the first night? Explain briefly.

d) Calculate the median of the improvement scores for each group. Is the median improvement higher for those who got unrestricted sleep? By a lot?

e) What are two possible explanations for the tendency for higher improvement groups for the unrestricted group?

The dotplots and medians provide at least some support for the researchers' conjecture that sleep deprivation still has harmful effects three days later. Nine of the ten lowest improvement scores belong to subjects who were sleep deprived, and the median improvement score was more than 12 milli-seconds better in the unrestricted sleep group (16.55 ms vs. 4.50 ms). The average (mean) improvement scores reveal an even larger advantage for the unrestricted sleep group (19.82 ms vs. 3.90 ms). But before we conclude that sleep deprivation is harmful three days later, we should consider once again this question:

f) Is it possible that there's really no harmful effect of sleep deprivation, and random chance alone produced the observed differences between these two groups?

You will notice that this question is very similar to questions asked of the helper/hinderer toy study. Once again, the answer is yes, this is indeed possible. Also once again, the key question is how likely it would be for random chance to produce experimental data that favor the unrestricted sleep group by as much as the observed data do.

g) In words, state the null and the alternative hypotheses to investigate whether sleep deprivation has a negative effect on improvement in performance on visual discrimination tasks. (*Hint*: For the alternative hypothesis: Do you expect the people to do better or worse when sleep deprived? Based on your answer, what sign/direction should you choose for the alternative hypothesis?)

h) Describe how you might go about deciding whether the observed difference between the two sample medians in this randomized experiment is statistically significant.

We will answer that question using the same 3S simulation strategy that we used earlier:

1. Statistic: We need a measure of the difference in the centers of the two groups, such as the difference in medians (or means).

2. Simulate: We will assume that the null hypothesis is true and replicate the random assignment process a large number of times, in order to get a sense for what's expected and what's surprising for values of the statistic under the null model.

3. Strength of evidence? If the result observed by the researcher is in the tail of the null model's "what if" distribution, we will reject that null model.

Because the null hypothesis asserts that improvement score is not associated with sleep condition, we will assume that the 21 subjects would have had exactly the same improvement

scores as they did, *regardless* of which sleep condition group (unrestricted or deprived) the subject had been assigned.
But there's an important difference here as opposed to those earlier studies: the data that the researchers recorded on each subject are not yes/no responses (survival or not). In this experiment the data recorded on each subject are numerical measurements: improvements in reaction times between pre-test and post-test. So, the complication this time is that after we do the random assignment, we must do more than just count yes/no responses in the groups.

What we will do instead is, after each new random assignment, calculate the median improvement in each group and determine the difference between them. After we do this a large number of times, we will have a good sense for whether the difference in group medians actually observed by the researchers is surprising under the null model of no real difference between the two groups (no treatment effect). Note that we could just as easily use the means instead of the medians, which is a very nice feature of this analysis strategy.

One way to implement the simulated random assignment is to use 21 index cards. On each card, we would write one subject's improvement score. Then shuffle the cards and randomly deal out 11 for the sleep deprivation group, with the remaining 10 for the unrestricted sleep group.

This time will turn to technology right away:
i) Go to the **Two Means** applet, change the statistic to **Difference in Medians**, check the **Show Shuffle Options** box, select the **Plot** display, and press **Shuffle Responses**. What is the applet doing?

j) Press Shuffle Responses again to repeat the re-randomization process. Did you obtain the same result as before?

k) Now ask for 998 more re-randomizations (for 1000 total). Look at the distribution of the 1000 simulated differences in group medians. Is the center where you would expect? Does the shape have a recognizable pattern? Explain.

l) Count how many of your 1000 simulated differences in group medians are as extreme (or more extreme), in the direction favoring the unrestricted sleep group, as the researchers' actual result (which was?) and determine the approximate p-value by calculating the proportion of your simulated differences that are at least this extreme.

m) Do these simulation analyses reveal that the researchers' data provide strong evidence that sleep deprivation has harmful effects three days later? Explain the reasoning process by which your conclusion follows from the simulation analyses.

n) Based on your analysis and the study design, is it legitimate to draw a cause-and-effect conclusion between sleep deprivation and lower improvement scores? To what population are you willing to generalize these results? Explain.

Possible Assessment/Follow-up Questions:

o) Redo this simulation analysis, using the difference in group means rather than medians. When you re-run the simulation, you will have to tell the computer or calculator to calculate means rather than medians. Report the approximate p-value, and summarize your conclusion. Also describe how your conclusion differs (if at all) from this analysis as compared to the analysis based on group medians.

p) What does one dot on the dotplot represent? (*Hint*: Think about what you would have to do to put another dot on the graph.)

q) How would you *interpret* the p-value (*Hint*: Describe the process you used to obtain it.)

r) Use the 2SD Method to approximate a 95% confidence interval for the difference in long-run mean improvement score for subjects who get unrestricted sleep minus the long-run mean improvement score for subjects who are sleep deprived. (*Hints*: Remember the observed value of the difference in group means, and obtain the SD of the difference in group means from the applet's simulation results. The interval should be observed difference in means ± 2SD, where SD represents the standard deviation of the null distribution of the difference in group means.) Interpret what this confidence interval reveals, paying particular attention to whether the interval is entirely positive, entirely negative, or contains zero. (*Hint*: Be sure to mention "direction" in your interpretation by saying how much larger improvement scores are on average for the treatment you find to have the larger long-run mean: I'm 95% confident that the long-run mean improvement score is _______ to _____________ higher with the _______________ treatment as opposed to the _____________ treatment.)

s) Looking back and ahead: Did anything about the design and conclusions of this study concern you?
What should the researchers' next steps be to fix the limitations or build on this knowledge?

t) Carry out a two-sample *t*-test for these data. Are the p-values and conclusions from these two approaches (simulation of randomization test vs. conventional test) similar? Discuss advantages and disadvantages of these two approaches to significance testing.

As we discussed previously, statistical significance is distinct from practical significance. One way to measure practical significance is to standardize the difference in means, but without factoring in the sample sizes:

$$effect\ size = \frac{group\ 1\ mean - group\ 2\ mean}{pooled\ SD}$$

u) Calculate and interpret the effect size for this study. Does it seem impressive to you? How are you deciding?

Additional resources:

- Beth Chances online lab assignments
  - html lab
  - jotform
  - Runestone