Goal

This is designed to help you gain familiarity with R as you will need it for this course. You will get practice uploading data, subsetting data, calculating statistics, generating randomization distributions, and calculating p-values from randomization distributions. This will also get you thinking about concepts related to randomized experiments and statistics!

As much as you can, try to make your code generic. The goal is to be able to copy much of this code and use it in the future - you don't want to edit variable names and numbers throughout each time you do this! If you are relatively new to R, here's a basic guide.

Data

Today we look at data from a paper[1] studying the effect that light at night has on weight gain and other variables in mice. According to the paper, "The global increase in the prevalence of obesity and metabolic disorders coincides with the increase of exposure to light at night and shift work", and the goal of this study was to determine whether light at night may play a *causal* role in the obesity epidemic.

The study took $n = 30$ mice and randomized them to three different treatment groups. All mice spent 16 hours in light, and the explanatory variable was the level of light during the remaining 8 hours. Some of the mice were randomized to darkness during those 8 hours (as is typical for regular mice), some were randomized to a dim light (equivalent to a TV on in the room for humans), and the remaining mice were exposed to bright light for all 24 hours. Mice are nocturnal, and typically most of their activity and eating happen at night. The hypothesis of this study was that having a light on at night may alter mouse eating habits and/or metabolism, and so increase body mass. The paper is available here.

The data are available here. There are some missing values because one mouse died, and one mouse did not receive the full glucose injection for it's glucose tolerance test.

Variable Name	Variable Description
Light	Treatment group (dark, dim, bright)
BodyMass0	Body mass at the beginning of the experiment (week 0)
BodyMass8	Body mass after 8 weeks of the experiment
Corticosterone	Blood corticosterene level (a measure of stress)
DayPct	Percentage of calories consumed during the day
Consumption	Average daily consumption (in grams)
GlucoseInt	Glucose intolerant at end of study? (Yes or No)
GTT15	Glucose level in the blood 15 minutes after a glucose injection
GTT120	Glucose level in the blood 120 minutes after a glucose injection
Activity	A measure of physical activity level

[1] Fonken, LK, Workman, JL, Waltona JC, Weila ZM, Morris, JS, Haim, A, and Nelson, RJ (2010). "Light at night increases body mass by shifting the time of food intake," *Proceedings of the National Academy of Sciences*, Vol 107, No 43, pp 1866418669.

1. Load the data into R. (You can do this any way you know how, but one possibility is downloading it as a .csv file, then using `read.csv`. The function `file.choose()` can be helpful for finding a particular file.)

2. We are interested in the causal effect of light at night on body mass. What should the outcome variable be? Plot the outcome by treatment group.

3. Here we will compare the mice exposed to darkness to the mice exposed to bright light overnight (once you have the code it is easy to rerun the analysis for the dim light group, if you are interested). Subset the data to only consider these two groups.

4. Set up the data such that everything you will need has generic names (such as `Y.obs` or whatever you want to call them). Everything specific to the context of your data (variable names, sample sizes) should only be in your R Script here. Everything else should be generic so you can copy/paste it for later use. What quantities will you need? (you can ignore any covariates for now)

5. Suppose we want the statistic to be the difference in means between the two treatment groups. Calculate T^{obs}. (Hint: you have to deal with the NA value, either now or when setting up your data).

6. How many different possibilities are there for W? Enumerate all of these possibilities in a matrix. (Hint: it's probably easiest to first install the `perm` package, and then the function `chooseMatrix` may come in handy.)

7. Calculate the test statistic under one of these possibilities for W (the first one), under the sharp null hypothesis of no difference.

8. Generate the exact randomization distribution for T, under the sharp null hypothesis of no difference.

9. Plot this distribution, and mark the observed test statistic (`abline` may be useful).

10. Calculate the exact p-value, based on this distribution. Make a conclusion in context.

11. Simulate one randomization, and calculate T for this randomization assuming the null. (Hint: `sample` may be useful)

12. Generate an approximate randomization distribution for T under the null, by simulating randomizations. Plot this distribution.

13. Calculate an approximate p-value, based on this distribution[2]. How does this compare to the exact p-value?

[2]Sometimes, the p-value is so small that none of the simulated randomizations will yield test statistics as extreme as the observed test statistic. If this is the case, increase the number of simulated randomizations until you get a non-zero p-value, or else you can just say it is less than 1/(number of simulated randomizations).

Congratulations - you've done all the hard coding! Now you can use the code you've already written to answer some interesting questions.

14. The researchers were also interested in whether having a light on at night can cause diabetes, so gave the mice a glucose tolerance test (GTT). Test whether light at night causes glucose intolerance in mice (use the categorical outcome). Look at a two-way table of the data, choose a test statistic, calculate T^{obs}, calculate the p-value (exact or approximate) via a randomization test, and make a conclusion in context.

15. Should `BodyMass0` be significantly different on average between the two treatment groups? Why or why not?

16. Plot `BodyMass0` across treatment groups and test whether initial body mass differs significantly between the two treatment groups (dark and bright light). At a significance level of $\alpha = 0.05$, are you making either a Type I or a Type II error? If so, which type?

17. The paper is titled "Light at night increases body mass by shifting the time of food intake." Their explanation for the increase in body weight under exposure to light at night is that a higher percentage of calories are consumed during the day, while mice are typically sleeping. Create the relevant plot, and test whether the percentage of calories consumed during the day is significantly higher on average in mice exposed to bright light at night, as opposed to darkness.

18. This difference only really matters if the percentage of calories consumed during the day is correlated with change in body mass. Plot the relationship between `DayPct` and body mass change, and test whether the association is significant. (Hint: to generate randomizations under the null, "break" the correlation by randomly permuting one of the variables.)

19. Might the fact that mice exposed to light at night gain significantly more weight be caused by a shift in when calories are consumed? Based on the data we have, can we establish with certainty whether this is true? Why or why not? (The researchers conducted additional studies involving limiting when the mice had access to food, in order to further address this question. If you are interested, read the paper!)

20. The paper claims "Mice housed in either bright or dim light at night have significantly increased body mass and reduced glucose tolerance compared with mice in a standard light/dark cycle, despite equivalent levels of caloric intake and total daily activity output". We have established the first part of the claim (at least for bright light). Do you agree with the second part of the claim? Why or why not?