BMI713 Problem Set 2

Instructions:

Please submit this problem set before class on Tuesday, November 7. Problem sets may be submitted within a week past the due date at a 20% penalty per day; each person is allowed to submit one problem late (within a week) without penalty. It is required that you comment your code. Missing comments will not allow you to receive full credit .

If you have any questions, please post on the piazza site. This problem set was prepared by Jacob Luber and Eric Bartell, so they will be most prepared to answer questions.

1. Hypothesis Testing (19 Points Total)

You are testing a new cancer therapeutic approach that combines chemotherapy and radiation in mouse models with human tumor xenografts. Assume that you have the ability to obtain an arbitrary number of identical mice. Your new approach seems to shrink the tumors about half of the time. For this question, conduct your tests in R.

(3 points)

You apply the treatment to 5 mice, observing 2 with smaller tumors. Test whether the approach shrinks tumors in half of mice, i.e. $H_0 : p = 0.5$. Subsequently, answer whether or not the null hypothesis can be rejected.

(4 points)

You apply treatment to 10 mice and observe 4 with smaller tumors. Test whether the approach shrinks tumors in half of mice, i.e. $H_0 : p = 0.5$. Subsequently, answer whether or not the null hypothesis can be rejected.

(4 points)

You apply treatment to 100 mice and observe 40 with tumor shrinkage. Test whether the approach shrinks tumors in half of mice, i.e. $H_0 : p = 0.5$. Subsequently, answer whether or not the null hypothesis can be rejected.

(4 points)

You repeat this test by applying treatment to 1000 mice and observe 400 with smaller tumors. Test whether the approach shrinks tumors in half of mice, i.e. $H_0 : p = 0.5$. Subsequently, answer whether or not the null hypothesis can be rejected.

(4 points)

What do you notice happening as we increase the total number of mice treated? Explain why what you are observing is occurring.

To get full credit on this problem provide code and a correct statement about rejecting your null hypothesis for parts 1.1-1.4 (may be different for 1.3 depending on which built in function used). For 1.5 state that an increase in sample size reduces sample variance.

2. Group Comparison (25 Points Total)

You are working on a microbial genetics project and are testing a new supplement that you add to your basic agar culturing plates that a collaborator mentioned greatly increases the observable number of CFUs (colony forming units) for the strain that you are studying on the plates *24 hours after incubation.* You plan to culture 10 of both your treatment and control plates and enlist a laboratory technician to help you.

Unfortunately, they accidentally forget to put 1 of the treatment plates and 2 of the control plates in the incubator, leaving you with the following results 24 hours later:

Group	CFUs	Sample Size	Mean ($\bar{x}$)
Treatment	10 12 8 16 22 4 7 2 9	9	10
Control	1 30 45 20 12 20 32 40	8	25

Stating Your Hypothesis (5 points)

Prior to performing the experiment, what is hypothesis H_0? What is hypothesis H_1?

H_0 is that we expect an equal number of CFUs, H_1 is that we do not expect an equal number of CFUs.

Performing the Statistical Test?

Assume that we have $Y_{ij} \sim N(\mu_i, \sigma^2)$ where i denotes treatment group and j denotes sample. μ_i is the mean CFU for group i, and σ^2 is the unknown variance. We want to test whether $\mu_1 = \mu_2$.

(5 points)

What statistical test would you utilize to compare these two groups?

Unpaired t-test

(5 points)

What exactly is being tested when you use the test from your previous answer? Specifically, write this by relating H_0 and H_1 to the assumptions made in this subsection.

H_1 indicates that adding supplement changes CFU count. H_0 indicates that supplement has no effect on CFU count.

(5 points)

Write the formula of the test statistic that you would use.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(5 points)

State why you would not want to use this test on this data if the assumptions are not true.

Test will be skewed if assumption of normality is false.

3. More Group Comparisons (35 Points Total)

For this problem, you are considering the same experimental results as the last problem. However, assume that the data being observed is not normally distributed. Further, note that the sample size is not large enough to treat it as asymptotically normal.

(5 points)

Propose a test to compare the same two groups. Is this test the same test utilized in problem 2? If this test is not the same test, then state what the differences are between the two tests in terms of what they are assuming.

Permutation test. This test is different because it makes no assumption about the distribution of the data.

(5 points)

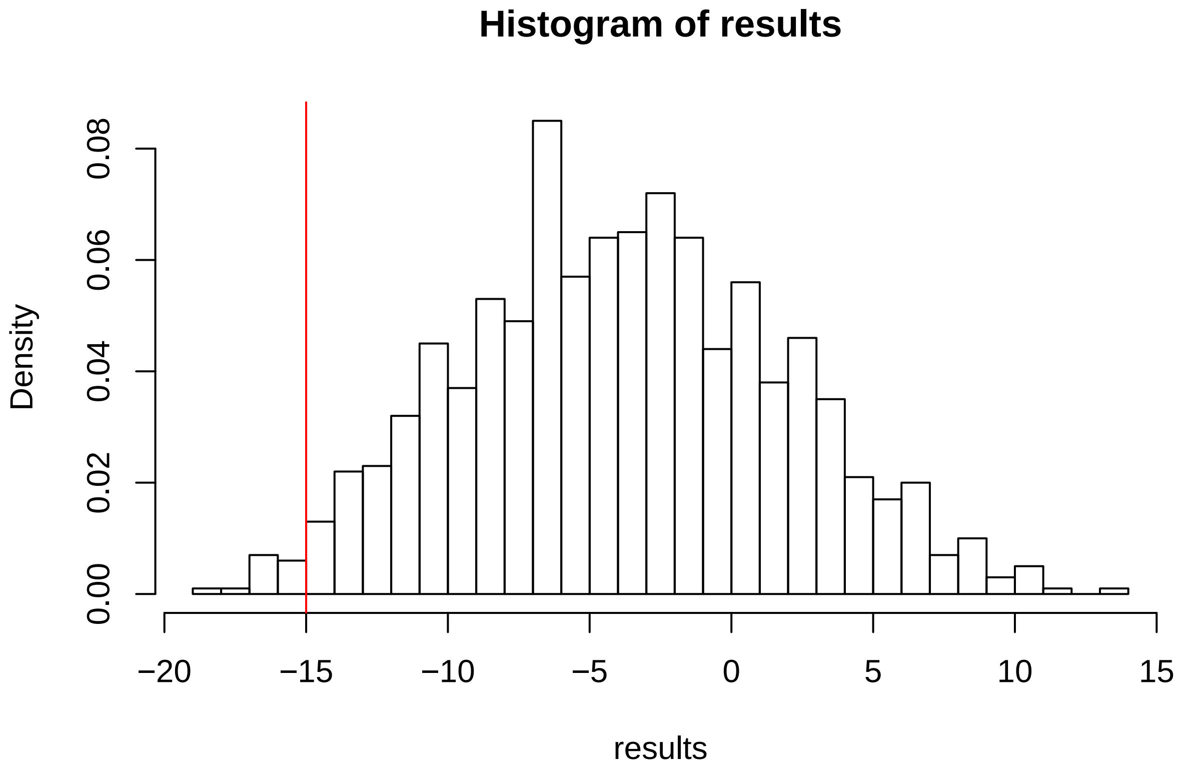
If we are determining significance via observation of the permutation distribution of a test statistic $T = \bar{Y}_1 - \bar{Y}_2$ (in terms of notation, $\bar{A}$ means the mean of A), please answer: What assumptions are you making about your data in relation to your hypotheses? What assumptions are you making about the distribution of your data?

Samples are independent of both each other and treatment assignment and that labels are exchangeable.

(20 points)

In R, generate a p-value for this test by calculating the permutation distribution of the test statistic $T = \bar{Y}_1 - \bar{Y}_2$. First, write a function that implements the test statistic (do not use packages such as exactRankTest or coin). What is the total number of permutations possible? Computing the test statistic for all of the permutations can be time consuming when there are a large number. Thus, generate a set of 1000 unique permutations of the data. Finally, plot the permutation distribution and add a vertical line representing the observed test statistic in addition to reporting the p-value.

```
test_code<- function(size_sample_1,size_sample_2,data,num_perms){
treatments<- rep(c("Treatment","Control"),c(size_sample_1,size_sample_2))
names(data) <- treatments
perms <- combn(1:(size_sample_1+size_sample_2),size_sample_1)
output_val <- apply(perms[,1:num_perms],2,function(x) {mean(data[x]) - mean(data[-x])})
return(output_val)
}
data<- c(10,12,8,16,22,4,7,2,9,1,30,45,20,12,20,32,40)
results <- test_code(8,9,data,1000)
hist(results,freq=FALSE,breaks=30)
abline(v=-15,col="red")
```



State that there are 24310 permutations possible.

(5 points)

What are your conclusions? Can we reject the null hypothesis?

It depends on the 1000 permutations selected, as long as your results were consistent with your code full credit was given.

4. Thinking About The Data (10 Points Total)

(5 points)

State another test statistic that gives the same result as the one utilized in problem 3. (hint: the formula for this test statistic does not consider $\bar{Y_2}$). Explain why this is the case.

$$t = \sum y_1$$

(5 points)

Explain why the central limit theorem/law of large numbers can play a role in determining what statistical test we use for group comparisons.

If sample size is large enough, normality can be assumed.

5. Proportion Tests (10 Points Total)

You have gotten peer reviews back from your mouse xenograft work combining your new cancer therapeutic and radiation. The reviewer asks you to consider only the new drug without radiation and look at survival at the end of one month (i.e. comparing mice that did and did not receive the drug). You proceed to conduct follow up experiments. In 134 mice (with xenografts) that did not recieve the drug, 25 were alive after a month. In 80 mice (with xenografts) that did receive the drug, 34 were alive after a month. Can the differences in survival rates in the follow up experiments be attributed to chance?

No.

6. Time spent? (1 Point)

How long did you spend working on this problem set?

7. Extra Credit | *Challenging* (5 Points)

Note: This question is intended to let you challenge yourself beyond the course material explicitly covered. We will not spend time addressing the extra credit on piazza or in office hours until after the problem set is due.

Your lab has developed a new therapy and are testing it in patients recovering from surgery. You are blinded as to which group received treatment and which received a placebo.

Two groups of patients have the following recovery times in days from surgery:

group 1: 15,20,21,26,28

group 2: 23,26,34,39

Can we reject the null hypothesis of no difference in recovery times between the two groups in favor of a two-sided alternative hypothesis at the 0.05 level using a two-sided permutation test? Subsequently, find a 95% 2-sided confidence interval for this (hint: you will need to create a range of datasets and perform many tests). Is it possible to obtain 1-sided p values from the datasets generated to obtain the confidence interval? If so, report the 99% 1-sided upper bound.

> Create a set of 40 new data sets where each set is assigned a value from a range -25 to 10. For each new data set, subtract this assigned value from each observation and perform a two sided permutation test on each new data set. To consider rejecting null hypothesis at 0.05 level, look at the two sided p-value when the assigned value in the range is 0. To find the conference interval, look for the minimum value in the range such that $p > 0.025$ and the maximum value such that $p > 0.025$. A 1 sided value cannot be calculated due to lack of symmetry in the null distribution of the test statistic. Bootstrap methods may also be used to solve this problem.