Statistics and Hypothesis Testing
Due Date: Thursday 9/14 @ 11:59 PM

1. Define a p-value.
This is the value that describes the probability that if the null hypothesis is true, the likelihood of seeing results at least as extreme as the alternative hypothesis.

2. What is the p-value in the rigged coin example (from the stats slides), and what does this p-value indicate?
P-value is 0.0059 which is less than the 0.05 cut off we are using. This p-value also means that assuming the coin was fair, the results would only happen 0.59% of the time. This means that it is likely the coin was biased. However, we cannot say for sure and are only 99.5% confident that it is biased.

3. Suppose we are examining the mutation status of the gene *MSH2* between two groups: old and young colorectal cancer patients.

a. Define the null hypothesis, H_0.
The mutation status of the gene MSH2 for old colorectal cancer patients and young colorectal cancer patients is the same.

b. Define the alternative hypothesis, H_a.
The mutation status of the gene MSH2 for old colorectal cancer patients and young colorectal cancer patients is different.

c. Suppose we find that *MSH2* is mutated more frequently in younger patients than in old patients, with a p-value of 0.07. From this, what can we conclude? What if the p-value was 0.03? Be specific.
If using a 95% confidence, we can conclude that there is not a significant difference in mutated MSH2 gene among younger patients than in older patients. If the p-value was 0.03, we would be able to conclude that MSH2 is significantly mutated more frequently in younger patients than older patients.

d. In a typical biological analysis, we look at many thousands of genes. Suppose we have 20,000 genes in our data set. Is there an issue if we set our p-value threshold at 0.05 as usual and examine each gene one at a time? (Hint: think about false positives.)
The issue with this method is that there is a chance that we will miss a lot of genes due to the high number of genes. Setting our p-value to 0.05 means that there is still a 5% chance that we are wrong. When we work though 20,000 genes individually in this manner, it means we would have missed 1000 genes.

4. **Two students, Alex and Jamie, are looking at the incidence of seizures in a set of glioblastoma patients in a clinical study. They find that seizures concurrently occur with**

usage of a certain hypothetical drug, Placebomab, with a p-value of 0.04. Determine whether the following statements are true or false. If false, explain why.

- Alex: “Because the p-value of 0.04 is less than 0.05, the drug Placebomab is most likely a cause of seizures in this group of patients.” (TRUE/FALSE)

Just because the p-value is less than 0.05 does not mean we know for a fact that Placebomab is the reason for seizures. We are only able to conclude that Placebomab is very likely to be associated with the seizures among the patients.

- Jamie: “Even though the p-value of 0.04 is less than 0.05, we can only say that Placebomab and seizures are significantly correlated in glioblastoma patients.” (TRUE/FALSE)