Stats questions: Please turn in answers to these questions by Wednesday, January 19, 8:00am by email to Nicole (blackn@usc.edu) with the nano worksheet!

1. Suppose we are comparing the survival times of two groups of colon cancer patients. The first group of patients have a mutation in the *APC* gene, while the second does not have a mutation in this gene. It turns out after performing survival analysis (you'll learn how to do so in a few weeks), the group with a mutation has significantly worse survival![1] For this observation:
    a. Define the null hypothesis, H_0. The survival times of the two groups is the same
    b. Define the alternative hypothesis, H_1. there is a difference in survival times for the two groups
    c. What assumption do we make when we calculate the p-value? we assume the survuval times are the same
    d. What can we conclude? Be careful with your wording! we conclude that there is a correlation between having the mutation and survival time

2. Now, suppose we examine the transcriptomes (RNA expression) of colon cancer patients. We notice that in young patients, the gene *PD1* is expressed at significantly higher levels than in old patients.[2] In terms of examining gene expression in this context, repeat the previous analysis, namely:
    a. Define the null hypothesis, H_0. young and old patients have the same expression levels of the gene
    b. Define the alternative hypothesis, H_1. the two groups have difference expression levels
    c. What assumption do we make when we calculate the p-value? we assume that the expression levels is the same
    d. What can we conclude? Be careful with your wording! there is a correlation between age and expression levels
    e. (Bonus) In a typical RNA-seq 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: what does the p-value assume? Think about false positives.)
    many genes are expressed at super high levels natrually, we need to compare one gene to the same gene somewhere else

[1] I didn't actually perform this analysis. You should go back in a few weeks and see if this claim is actually true! *APC* is a colon cancer-specific oncogene though.

[2] Again, I didn't do any analysis to see if this was true or not. *PD1* does suppress $CD8^+$ T-cell activation and is a target for many therapeutic approaches.