Stat 430 Homework 3

due: November 6, 2009 at 5pm

1. Access the Shakespeare word usage data. Some researchers have claimed they can identify gender of modern writers from as few as 300 words. Rather than focus on the obvious words (e.g. "princess" or "fight"), which can depend on the topic of the article, the researchers focus on seemingly unsubstantial words. In particular, men are supposed to prefer words communicating quantities, like "more" and "some," while women are supposed to prefer personal pronouns and possessives, like "you" and "your." Shakespeare was a man, by all accounts, but he wrote the words of many female characters. You will test if he was able to recreate these subtle differences in word usage of males vs. females.
   (a) Use the data to test if there is a difference in Shakespeare's gender word usage for the more active words [love (v), hate (v), love (n), kill (v)].
   (b) Test if there is a gender difference in the usage of the words "more" and "some." Repeat for "you" and "yours."
   (c) You may combine words that do not have usage differences from part b. Now choose one or a group of words to "represent" each gender. Is there a difference in Shakespeare's gender-based use of these words (or sets of words)?

2. Access the virus data. To test virus fitness (as in, Darwin's survival of the fittest), live virus are added to cell cultures and the amount of virus is measured after a fixed period of time. Since growth conditions may vary tremendously, the virus is added with a standard virus and the log ratio of the numbers of the viruses is reported (a kind of paired sample, where the number reported is a measure of the difference between the members of pair). In this question, you will use ANOVA to detect if there is a fitness difference in three different EIA viruses, HIV-like viruses that infect horses.
   (a) There are three named viruses in the file, PND1, PND4, and PND5, that differ in their envelop protein (the protein that is responsible for binding and infecting target cells in the host and the part of the virus most susceptible to attack by the immune system). Do they differ in fitness as measured by a significant different log ratio relative to the standard virus?
   (b) You will notice there is substantially more data for virus PND5. It is because that virus was tested with three different initial conditions, indicated in the second column. Do these initial conditions affect the measurement of viral fitness?
   (c) Evaluate the ANOVA assumptions and whether they apply to this data.

3. [For practice; not to be turned in.] Repeat question 2 using non-parametric tests. Do you reach the same conclusions?