QERM 598 - HW 1
Due January 16, 2008
Eli Gurarie

Two sample comparisons and the T-distibution

1. First, a little math. Equations 1 and 2 in the lecture presentation should be sufficient to demonstrate that for $(X_1, X_2, ...X_n)$ where if $X \sim \mathrm{N}\{\mu, \sigma^2\}$, the following statistic:

$$t_0 = \sqrt{n\,(n-1)}\frac{\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \tag{1}$$

has a Student's-T distribution with $n-1$ degrees of freedom under the null hypothesis that $\mu = 0$. You don't need to know the exact form of the distribution, all you need to know is that if $Y \sim \mathrm{N}\{0, 1\}$ and $Z \sim$ Chi-squared$\{n\}$ then:

$$\frac{Y}{\sqrt{Z/n}} \sim T\{n\} \tag{2}$$

This is a useful exercise in manipulating distributions and should give you an algebraic feel for where that one degree of freedom is lost. Confirm and illustrate this result using simulations in R. Note that this result is most significant for low values of n. Make sure you download the most recent version of the lecture!

2. Note: (For the following problem, you will need to analyze a dataset. You can use the ant data that we looked at in class (posted on the website), some data that you perhaps are working on yourself, or anything else that you can dig up of interest on the internets. You will need to compare two sets of measurements on different populations. The dataset should have at least 100 measurements per population. Non-normal looking distributions are welcome!)

   There are cases where comparing means is not the best way to answer the question "Is A bigger than B?". As an extreme example, the sequence A $= (1, 1, 1, 1, 1, 1, 1, 100)$ is consistently smaller than B $= (4, 4, 4, 4, 4, 4, 4)$, although $\bar{A} > \bar{B}$. The problem arises when outliers have disproportionate effect on the mean. There exist robust test statistics that deal with this problem and have the general advantage that they do not rely on any parameter assumptions. One of these is the Wilcox ranked sum test, the two-sample version of which is called the Mann-Whitney test.

   (a) Read the first half of Mike Keim's notes on non-parametric tests (posted on the website) and generate a function in R that performs the Mann-Whitney two-sample test, that is: it returns the value of the test statistic and the p-value against the so-called Wilcox null-distribtion (see the `dwilcox()`function in R). Apply the test to a subset of your data $(A_1...A_n)$ and $(B_1...B_n)$ where $n < 10$ and see if it gives the same results as the `wilcox.text()`function. Illustrate the null-distribution of your test with the observed statistic.

(b) Subsample your data a bunch of times for $n = 10$ and perform t-tests and Mann-Whitney tests on each subsample. Collect the p-values from each test. Do they give the same p-value in the long run?

(c) Choose a test of your choice and perform an experiment similar to the one above, but sample at a range of n values between 5 and 100. How does the value of the p-value depend on n? Comment.