STAT 621 HOMEWORK 1
Due: Friday September 6

1. Two-Sample Kolmogorov-Smirnov test: The data below are measurements of human plasma growth hormone after a certain treatment for both coronary-prone subjects (Type A) and coronary resistant subjects (Type B).

```
Type A:   3.6   2.6  4.7   8.0  3.1  8.8   4.6   5.8   4.0  4.6
Type B:  16.2  17.4  8.5  15.6  5.4  9.8  14.9  16.6  15.9  5.3  10.5
```

   (a) Plot the empirical distribution functions (EDFs) on the same set of axes. Comment.

   (b) Use the KS test to test whether there is a difference in probability distributions of hormone levels between Type A and Type B subjects. State your hypotheses, report the test statistic, p-value and conclusion.

2. The one-sample Kolmogorov-Smirnov test is somewhat impractical because it requires that the null distribution is specified exactly. For example, I can't just test whether my data were generated from some Normal distribution, I need to specify the mean and standard deviation. Typically these parameters are unknown. A solution might just be to substitue estimates computed from the sample. But is that really valid? Couldn't this affect the sampling distribution of the test statistic under the null?.

   In this problem you'll investigate this question using *Monte Carlo simulation* (much more about this technique soon). Basically we will simulate many, many data sets and run the KS test (1) assuming the true parameters are known and (2) substituting estimated values of the parameters. Each time we will save the value of the test statistic. This results in random samples of the KS test statistics under both scenarios. We then use these sets of test statistics to approximate the true sampling distributions for the two cases.

   (a) Open the file `KSsim.txt` on Blackboard. Copy the funciton `KSdist.sim` and paste it into your R console. Look over the function and see if you can tell what it's doing.

   (b) Use the function to estimate the sampling distribuiton of the KS test statistic using both known and estimated parameters (code below). We will suppose that data are Normal with mean $\mu = 3$ and standard deviation $\sigma = 4$, and that we have a sample size of $n = 50$. We will base our estimates on `nsim=5000` simulated data sets. Print out the first few lines of the output and examine them. You should have a data set with two columns – these are your samples of KS test statistics for the known and estimated parameter cases.

```
a=KSdist.sim(n=50, mu=3, sd=4, nreps=5000)
head(a)
```

(c) Make a density plot (basically a smoothed histogram) of the two sampling distributions. You may need to adjust the axes so they both fit.

```
plot(density(a$D.est))
lines(density(a$D.true), lty=3)
legend(.15,20, c("est","true"),lty=c(1,3))
```

(d) Comment. Does it appear that estimating the mean and standard deviation in the KS is a wise idea? What does this tell you about the reported p-value of the KS test if you substitute estimated parameters? Will you be more or less likely to reject your null?

(e) Does sample size affect these results in general? Repeat parts b and c using a sample size of 5,000. Comment.

3. The following data come from a study of the effect of aspirin on bleeding time. Bleeding time from a small incision was measured immediately after the ingestion of a 600-mg dose of aspirin (X), and again 2 hours after administration of the aspirin (Y). Data are saved on blackboard as `bleeding.txt`. (Data and example from Hollander, Wolfe and Chicken, Chapter 3).

(a) Use the sign test to determine whether bleeding time is typically longer immediately after the ingestion of aspirin. State your hypotheses, test statistic, p-value and conclusion.

(b) Repeat the hypothesis test, this time using the large-sample approximation discussed in class. Compute the test statistic and find the p-value. Do you reach the same conclusion? How do the p-values compare?

(c) Another option for testing this hypothesis is to use a paired t-test. This assumes that the differences are normally distributed. Evaluate this assumption (nothing fancy, plot a histogram or normal probability plot). Perform the test (you can use the `t.test` function in R). Report you p-value and conclusion. Any advantages to making the stronger assumption of Normality? Are there any disadvantages?