18.650 Problem Set 1 Spring 2017
Statistics for Applications
Due Date: Wed 2/15/2017, prior to 3:00pm
Where: Problem Set Box (outside 4-174)
(or electronically to Stellar website)

1. Let $X_1, X_2, \ldots, X_n$ be a random sample of known size n from a Normal distribution with unknown mean $\mu \in (-\infty, +\infty)$ and unknown variance $\sigma^2 > 0$. The sample variance is given by

   $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

   where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

   4 Points/Part

   (a) By Theorem 6.3 B, the distribution of $U = (n-1)S^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom. Using the known density function for U:

   $$f_U(u) = \frac{1}{2^{(n-1)/2}\Gamma[(n-1)/2]} u^{(n-1)/2-1} e^{-u/2},\ u > 0.$$

   derive the density function for the distribution of $Y = S^2$.

   (b) Compute the variance of the sample variance: $Var(S^2)$

   (c) Consider the scaled transformation of S^2:

   $$Z = \frac{S^2 - \sigma^2}{\sqrt{2\sigma^4/(n-1)}} = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\sigma^2} - 1\right).$$

   Compute the expected value and variance of Z.

   (d) For fixed sample size $n = 20$, find the constants C_{Lower} and C_{Upper} such that

   $$P(\frac{S^2}{\sigma^2} < C_{Lower}) = 0.25$$
   $$P(\frac{S^2}{\sigma^2} < C_{Upper}) = 0.75$$

   Hint: Use the quantiles ($p = .25, .75$) of the chi-square distribution with 19 degrees of freedom

   ```
   > qchisq(p=c(.25,.75),df=20-1)
   [1] 14.56200 22.71781
   >
   ```

   (e) With the values determined from part (c), it follows that

   $$P(C_{Lower} < \frac{S^2}{\sigma^2} < C_{Upper}) = 0.5$$

   This statement can be re-expressed as:

   $$P(\frac{S^2}{C_{Upper}} < \sigma^2 < \frac{S^2}{C_{Lower}}) = 0.50.$$

   Given that σ^2 is a fixed constant, explain how to interpret this probability statement.

2. Let X and Y be independent $N(0,1)$ random variables.

   5 Points/Part:

   (a) Compute the moment generating function for the distribution of
   $$W = X^2 + Y^2, \qquad M_W(t) = E[e^{tW}].$$

   (b) By Property A of Rice, Section 4.5, the moment-generating function uniquely determines the probability distribution when the function exists for argument t in an open interval containing zero. Identify the moment-generating function in (a) with: (i) the mgf of a $Gamma(\alpha, \lambda)$ distribution (See Appendix A2 of Rice) for specific values of α, and λ (give these values); and (ii) the mgf of an $Exponential(\lambda)$ distribution (give the λ value).

   (c) Derive the density function for the (joint) conditional distribution of (X, Y) given $W = w$.

   (d) Using R, generate 1000 replicates of (X, Y, W, V), where $X \sim N(0,1)$, independent of $Y \sim N(0,1)$, and
   $$\begin{aligned} W &= X^2 + Y^2 \\ V &= arctan(Y/X) \end{aligned}$$
   Plot the histograms of each variable alone. Describe the shape of each histogram and comment whether it conforms with the shape of the density for the true distribution (name it explicitly) of each variable.

   The following R code should be sufficient:

```
> X=rnorm(1000)
> Y=rnorm(1000)
> W=X^2 + Y^2
> V=atan(Y/X)
> par(mfcol=c(2,2))
> hist(X,breaks=100)
> hist(Y,breaks=100)
> hist(W,breaks=100)
> hist(V,breaks=100)
```

   Note: To complete this problem, create an R script file with the R commands and create an HTML file of the commands/output using the Compile-Report button in the RStudio file editor:

   - In RStudio, create a new script file (e.g., named "ProblemSet1_2_d.r") which contains these R commands. From the top left menu of RStudio, select File/New File / R Script to open a new script file. Type these commands in the file. Save the file: from the top-left menu of RStudio, select File/Save As, giving the name of your file.
   - Test the script file by hitting the "Source" button at the top-right side of the file editor window (top-left panel of RStudio). Keep fixing/re-testing until the commands execute without error.

- Create an html file with the commands/output of the script file by hitting the "Compile Report" button in the top-middle of the file editor window. (It looks like a spiral notebook).
- Turn in your HTML file as part of the homework.

3. Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$.

10 Points for Part (a) and 5 Points each for Parts(b),(c) and (d):

(a) (Problem 6.4.2) Prove **Proposition 6.2 B**: The density function of W is given by

$$f(w) = \frac{\Gamma[(m+n)2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}, \quad w \geq 0.$$

(b) Prove that for $n > 2$, $E[W] = n/(n-2)$

(c) Using R, generate a 10000-vector called *wvec*, with independent, random deviates (i.e., i.i.d. pseudo-random realizations) from the $F(m = 3, n = 3)$ distribution. Using *wvec*, compute the cumulative mean vector *wvec.cummeans* (of length 10000), where

$$wvec.cummeans[j] = (\textstyle\sum_{i=1}^{j} wvec[j])/j,\ j = 1, \ldots, 10000.$$

Construct the index plot of *wvec.cummeans*. Does the cumulative mean vector appear to have a limiting value? If so, what is it? If not, explain why.

The following R code should work:

```
> set.seed(1)
> wvec=rf(10000,df1=3,df2=3)
> wvec.cummeans=cumsum(wvec)/c(1:length(wvec))
> plot(wvec.cummeans)
> abline(h=3/(3-2))
```

(d) Repeat (c) with a 100000-vector of deviates (i.e., i.i.d. pseudo-random observations) from the $F(m = 3, n = 2)$ distribution.

The following R code should work:

```
> set.seed(1)
> wvec=rf(100000,df1=3,df2=2)
> wvec.cummeans=cumsum(wvec)/c(1:length(wvec))
> plot(wvec.cummeans)
```

4. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. (independent and identically distributed) $N(0,1)$ random variables, i.e., a simple random sample of size n from the $N(0,1)$ distribution. State the explicit distribution for each of the following statistics; if it exists state the expectation of the statistic. Explain your reasoning.

   10 Points: 1 Point/Part

   (a) $T_a = \sum_{i=1}^{n} X_i$
   (b) $T_b = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
   (c) $T_c = X_1^2$
   (d) $T_d = X_1^2 + X_2^2$
   (e) $T_e = \sum_{i=1}^{n} X_i^2$
   (f) $T_f = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
   (g) $T_g = t_n = \overline{X}/(S/\sqrt{n})$
   (h) $T_h = X_1/\sqrt{X_2^2}$
   (i) $T_i = X_1^2/X_2^2$
   (j) $T_j = (X_1^2 + X_2^2 + X_3^2)/(X_4^2 + X_5^2 + X_6^2)$