Stat 407 Lab 6 Multivariate Normal Distribution Fall 2012
Due: Wednesday, September 26 in class Hand in one solution per group.

Purpose: In this lab we will learn simulate samples from a multivariate normal distribution. Simulation is a very useful technique in statistics. We might use the technique to compare our data with what we'd expect to get from a multivariate normal distribution. Or we might use use this technique to check how the value of a statistic might change with different samples.

1. Use the MVN sample generator, `rmvnorm` in the mvtnorm package, to generate the samples from a population with the following characteristics:

   (a) $n = 30, p = 2$ standard normal

   (b) $n = 100, p = 2$ standard normal

   (c) $n = 500, p = 2$ standard normal

   (d) $n = 500, p = 4$ standard normal

   (e) $n = 500, p = 4, \boldsymbol{\mu} = \begin{bmatrix} -2 \\ 10 \\ 5 \\ -3 \end{bmatrix}, \Sigma = \begin{bmatrix} 100 & 0 & -10 & 10 \\ 0 & 50 & 30 & 0 \\ -10 & 30 & 200 & 0 \\ 10 & 0 & 0 & 80 \end{bmatrix}$

   For each sample, compute the summary statistics, and make a scatterplot matrix. Hand these in, and answer these questions:

   - Why are the means for the sample different from the means used to generate the sample?
   - Are the means for the sample closer to that used to generate the sample, when the sample size is larger? Why?
   - Which samples have a strong correlation between variables in the scatterplot matrix?

2. Use the `rmvt` function to generate a sample of size 100, dimension 2, from a multivariate t distribution. Calculate the summary statistics, and make a scatterplot matrix. Describe how this differs from a bivariate standard normal sample.

3. Check the multivariate normality of eac of these data sets:

   - The livability data, used in Lab 5, just the variables Climate, Education, Economic and Diversions.
   - The bodyfat data, used in Lab 3, just the variables NeckCircm, ChestCircm, AbdomenCircm.
   - The tips data, just the tips and total bill variables.

   You'll need to do something like this:

   (a) Compute summary statistics.

   (b) Produce summary plots: histograms, normal probability plots, scatterplot matrix, watch the data in a tour, and distances QQ plot.

   (c) Generate samples with the same mean and variance-covariance, as comparison data sets with the real data.

   If at any point you see something that proves that the sample data cannot be from a multivariate normal population, then stop there, and explain why it is not. You don't need to comprehensively do weverything, unless the data looks really close to normal, and then you have to check everything.

Notes:

- To generate a sample from a multivariate normal, use the `mvtnorm` package, and the function `rmvnorm`:

```
> library(mvtnorm)
> ?rmvnorm
> # Generate a sample of size 30, dimension 2
> x<-rmvnorm(30, mean = rep(0, 2))
> # Compute summary statistics
> summary(x)
> var(x)
> cor(x)
> # Plot using the base graphics scatterplot matrix
> ggpairs(x)
> # Plot using the the YaleToolkit graphics scatterplot matrix
> library(YaleToolkit)
> gpairs(x, pch=16)
> Use a non-standard variance-covariance matrix
> myvc<-matrix(c(100, 0, -10, 10, 0, 50, 30, 0, -10, 30, 200, 0, 10, 0, 0, 80),ncol=4,byrow=T
> myvc
> Use a non-standard variance-covariance matrix
> mymean<-c(-2, 10, 5, -2)
> mymean
> x<-rmvnorm(30, mean = mymean, sigma = myvc)
> # Draw the normal probability plot for the first variable
> qqnorm(x[,1])
> # Add the X=Y guideline
> qqline(x[,1])
> # Use the function below to make the distances qq-plot
> f.mv.distQQ(x)
```

- This function produces the χ^2 distances QQ-plot:

```
f.mv.distQQ<-function(x){
  n<-dim(x)[1]
  p<-dim(x)[2]
  mn<-apply(x,2,mean)
  vc<-var(x)
  ev<-eigen(vc)
  vcinv<-ev$vectors%*%diag(1/ev$values)%*%t(ev$vectors)
  x<-as.matrix(x-matrix(rep(mn,n),ncol=p,byrow=T))
  dx<-diag(x%*%vcinv%*%t(x))
  par(pty="s",mar=c(4,4,1,1))
  qqplot(dx,qchisq(((1:n)-0.5)/n,p),ylab="Chisq quantiles",pch=16,main="")
  lines(c(0:round(max(dx))),c(0:round(max(dx))),col="gray80")
}
```

  The function works by:

  - Computing the Mahalanobis distances between each point and the sample mean.
  - Plotting these distances against theoretical values from a χ^2 distribution.