STAT 802 Applied Multivariate Analysis Sept. 26, 2005

Today:

1. Simulating and Understanding the Multivariate Normal.
2. The sum of squared statistical distances for the p-variate Normal (Result 4.7)
3. Sampling Distributions of the sample mean and sample covariance from p-Normal
4. Testing for Multivariate Normality (chi-square plots)
5. Transformations to Normality (Section 4.8)

Bivariate Normal ($\rho \neq 0$ say)

How do you simulate it?
One way: generate 3 IID N(0,1) variates z_1 z_2 z_3
Let $x=z_1+a\, z_3$ and $y=z_2 + a\, z_3$ where $a=(\rho/(1-\rho))^{1/2}$

Note also $\rho = a^2/(1+ a^2)$ so we have, for example:

a =	0	1	2	3
ρ =	0	0.5	0.8	0.9

```
> bvrnorm
function (n=25, corr=.8)
{
# program produces simulated normal data from bivariate standard normal bivariate
population with means 0, sds=1 and correlation ="corr")
        a=(corr/(1-corr))^.5
        z=rnorm(n)
        x=(rnorm(n)+a*z)/(1+a^2)^.5
        y=(rnorm(n)+a*z)/(1+a^2)^.5
        plot(x,y)
        invisible(list(x=x,y=y))
}
```

so you use
a=bvrnorm(25,.8)
and the output is of the form
a[[1]] and a[[2]] (or a$x and a$y)

Note mvrnorm does this for any p-norm including p=2, and also can set the means and sds of the population sampled as well. Note the input is the population mean vector and the population covariance matrix. (Note also that if the mean, sd, and corr are to be reproduced exactly, one would need to specify the parameter "empirical=T".) To replicate the above (bvrnorm) with mvrnorm, use

a=mvrnorm(n=25,mean=c(0,0), sigma= matrix(c(1,.8,.8,1),ncol=2))
and the output will be of the form a[,1] and a[,2].

2. The sum of squared statistical distances for the p-variate Normal (Result 4.7)
It is chi-square on p degrees of freedom. Why?
3. Sampling Distribution of Sample Mean and Sample Covariance from p-Normal.

Mle of μ and Σ - see Result 4.11 p 171 (note the n version of S is used)

Sampling Distribution of $\bar{X}$ and S - Box p 174

Note assumption of p-normality for above. But for large samples,

CLT p 177

i) $\bar{X}$ is approx $N_p(\mu, (1/n)\Sigma)$
ii) $n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu)$ is approx chi-square on p df.

Can we check this with simulation? Note: The details of a simulation often clarify the result.

With the program below, check that

i) $\bar{X}$ is approx bivariate normal when p=2. Use both estimated Σ and known Σ.
ii) $n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu)$ is approx chi-square on p df. Try p=2, chiplot=T, cutoff=6 and p=3, chiplot=T, cutoff=7.8.

```
>clt.ex
function (m=250,n=25,p=2,corr=.5,chiplot=F,cutoff=6)
{
        cov=matrix(ncol=p,nrow=p) #we specify the cov matrix that we want to generate.
        for (i in 1:p){
                for (j in 1:p){
                        cov[i,j]=corr}
                        cov[i,i]=1
                        }
        diff=matrix(nrow=m,ncol=p)   # initializes diff
        for (i in 1:m) {
        x=as.data.frame(mvrnorm(n,c(rep(0,p)),cov)) # This command uses
        #mvrnorm from MASS package.  Use help(mvrnorm) for syntax.
        diff[i,]=(n^.5)*t(mean(x))
        d[i]=n*t(mean(x))%*%solve(cov(x))%*%mean(x)
                        }
        plot(as.data.frame(diff))    # need df so will do matrix plot if p>2
        index=1:m
        index=index[d>cutoff]  # 6 is chisq (.95) for 2 df
        print(c("Number outside cutoff contour"))
        print(length(index))
        pts=diff[index,]
```

```
        points(pts,col="red",cex=2)
        if (chiplot==T) {quartz();
                                my.dotplot(d)
        print(c("Mean and SD of ChiSq Stat (p df)"))
        print(mean(d))
        print(sd(d))}
        invisible(list(diff,d))
}
```

4. Testing for Multivariate Normality (chi-square plots)

What if data is not normal?

Try replace x by $x+(x^2)/10$ in simulation program.
Repeat look at distribution. Inconclusive.

What about using a chi-square plot ?

```
chi.plot
function (x,df)
{
        x=sort(x)
        l=length(x)
        a=(1:l-0.5)/l
        q=qchisq(a,df)
        plot(q,x)
}
```

Try it with Normal and non-normal data.

5. Transformations to Normality. (p 194)

Box p 194.

In practice, try a few transformations, rather than estimate optimal λ in Box-Cox.

Next time: Ch 5 Hotellings T^2