STAT 802 Applied Multivariate Analysis Sept. 28, 2005

Today: Ch 5: Inferences about Mutivariate Means

We have already done most of the theory for this, but it only worked approximately for large n (CLT). But Hotelling's T^2 statistic is just the statistical distance of the sample mean from the hypothesized mean, the same as we used in the CLT when we had to estimate the covariance. (Compare (4-28) on p 177 in Ch 4 with 5-4 on p 211 in Ch 5). Now we estimate the variance with the sample variance and we postulate the value of the population mean. The statistical distance of the sample mean from the population mean is our statistic. It turns out to have an F-distribution. (box, p 212). Of course we assume a p-Normal population here. We need to think about robustness to non-Normality when n is small: in the univariate case, Student's t is very robust.

Here is an extension of our program from last day that simulates this F distribution in a particular situation:

m is the number of sample means we simulate
n is the sample size underlying each sample mean and sample covariance
p is the number of variables in the multivariate data set (or the dimension of the p-Norm)
corr is the off-diagonal covariance (correlation here) of the multivariate data we simulate
means is the value of the common mean (for all p components) that we are testing
    (Note the simulated data has means=0)
alpha is the tail probability of the T^2 statistic if the hypothesis is true

```
> tsq.ex
function (m=250,n=25,p=2,corr=.5,means=0,alpha=.10)
{
        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))
        hypmean=c(rep(means,p))
        d[i]=n*t(mean(x-hypmean))%*%solve(cov(x))%*%mean(x-hypmean)
                        }
        plot(as.data.frame(diff))
        index=1:m
        cutoff=((n-1)*p/(n-p))*qf(1-alpha,p,n-p)
        index=index[d>cutoff]  # formula p 212
        print(c("Percent outside cutoff contour=alpha?"))
        pct=100*(length(index))/m
        print(round(pct,2))
        pts=diff[index,]
        points(pts,col="red",cex=2)
        print(c("Mean statistical distance from Hypoth Mean"))
        print(round(mean(d),2))
        print(c("Exp Value Under Hypoth"))
        print(round(((n*p-p)/(n-p))*((n-p)/(n-p-2)),2))
        print(c("SD of statistical distance from Hypoth Mean"))
        print(round(sd(d),2))
        print(c("SQRT of Exp Value of Var Under Hypoth"))
        print(round(((n*p-p)/(n-p))*((2*(n-p)^2)*(n-2)/(p*((n-p-2)^2)*(n-p-4)))^.5,2))
        invisible(list(diff,d))
}
```

Note the ugly formulae for the mean and sd of the T^2 statistic. This is based on the mean and var of the F statistics in some texts.

The mean of $F_{n,m}$ is m/(m-2) and the var is $2m^2(m+n-2)/n(m-2)^2(m-4)$

You can check what happens when you let n be large (so you should get chisq means and var).

If you put
n -> p
m -> n-p
and take account of the multiple (n-1)p/(n-p) that occurs in the formula p 212.,
you get the expressions in the program for the mean and sd of the "stat distance" statistic.

Section 5.3 Likelihood Ratio Test

It turns out to be another derivation of T^2. The formula on top p 218 shows equivalence of Wilks Λ and T^2. We will not pursue this except to say that the likelihood ratio method is a general method for producing tests, and is useful in theoretical developments.

Section 5.4 Confidence Regions

P 221 box shows same expression we have been using for (estimated) statistical distance of a sample mean from a hypothesized population mean. If we consider the locus of hypothesized means that satisfy this expression, we have a "Confidence Region" for the population mean.

Note we can tell whether a particular mean is in the CR with a program like tsq.ex:
We put m=1, and see if the cutoff is exceeded. Of course, in practical use we would replace the simulated data by the real data. For p=2, the elliptical CR can be drawn as

In Fig 5.1 p 223. The directions and lengths of the axes are given by the eigenanalysis on p 221.
Simultaneous Confidence Regions (Result 5.3 p 225)

All linear combinations simultaneously included in corresponding CR with prob 1-α. Produces, in particular, CIs for each component, that apply simultaneously. These would be wider than univariate CI with prob 1-α. But they allow other comparisons to be made (other linear combinations) without adjusting the 1-α.. Note that the effective coverage of p ordinary univariate 1-α.% CIs is less than 1-α.%.

An option for using the ordinary CIs is to adjust the indivisual "1-α." to a very large value so that the simultaneous coverage is at least 1-α. This does a little better than the above simultaneous CIs since it allows focusing on a few linear combos (the one using the line combos like (1,0,0,…) See Fig 5.4 p 233.

Section 5.6 Multivariate Control Charts

Univariate Control Charts – identify unusual sources of variation and fix
(3 sigma limits, mgt by exception)
Result is reduced variation and improved control
Less waste and higher profits

Multivariate Control Chart Use T^2 But remember
univariate outlier ≠ multivariate outlier

Exercise for Wed, Oct 5

1. Use the program tsq.ex() to explore the relationship of the power of the T2 test in detecting increasing "means" from 0, to the common correlation value "corr". Summarize in words what you uncover from this exercise.

2. Ex. 5.9 on p 262.