Hotelling T^2 Test and MANOVA using R

Quick Review:

Tests regarding population mean vector under multivariate Normal setup:

Let $\mathbf{X}_1$,$\mathbf{X}_2$,$\cdots$,$\mathbf{X}_n$ be i.i.d according to $N_p(\mu,\Sigma)$ where both μ and Σ are unknown

Consider the problem of testing

H_0: $\mu=\mu_0$ vs H_1: $\mu\neq\mu_0$

where μ_0 is specified

The statistic defined as

$$T_p^2=n\left(\bar{\mathbf{X}} - \mu_0\right)^T S^{-1} \left(\bar{\mathbf{X}} - \mu_0\right)$$

is known as the Hotelling T^2 Test Statistic where $\bar{\mathbf{X}}$ is the sample mean vector and S is the sample var-cov matrix based on n observations

Also $F=\frac{(n-p)}{p} \frac{{T_p}^2}{(n-1)} \sim F_{p,n-p}$ under H_0

So the test rule for the testing of H_0 against H_1 is given by:

Reject H_0 if $F > F_{p\,,\,n-p\,;\,\alpha}$, where $F_{p\,,\,n-p\,;\,\alpha}$ is the upper α^{th} quantile of the $F_{p,n-p}$ distribution

Confidence region for population mean vector μ:

From the test above we have,

$$\begin{aligned}
& P_{H_0}\left[\, F > F_{p\,,\,n-p\,;\,\alpha}\right] = \alpha \\
\iff \quad & P_{H_0}\left[\, F \leq F_{p\,,\,n-p\,;\,\alpha}\right] = 1-\alpha \\
\iff \quad & P_{H_0}\left[\, \frac{(n-p)}{p}\frac{{T_p}^2}{(n-1)} \leq F_{p\,,\,n-p\,;\,\alpha}\right] = 1-\alpha \\
\iff \quad & P\left[\, \left(\mu - \bar{\mathbf{X}}\right)^T S^{-1}\left(\mu - \bar{\mathbf{X}}\right) \leq c^2\right] = 1-\alpha
\end{aligned}$$

$$\text{where } \; c^2 = \frac{(n-1)p}{(n-p)} F_{p\,,\,n-p\,;\,\alpha}$$

Thus $\left[\left(\mu - \bar{\mathbf{X}}\right)^{\mathrm{T}} \mathrm{S}^{-1} \left(\mu - \bar{\mathbf{X}}\right) \leq c^2 \right]$ gives the 100(1-α)% confidence region for μ

NOTE:

Sometimes the mean μ may denote univariate mean and sometimes it may denote population mean vector and the distinction may be understood from the context

Example:

Consider the data on on length, width and height of turtles

Length	Width	Height
98	81	38
103	84	38
103	86	42
106	86	42
109	86	42
123	88	42
123	93	44
133	95	50
133	99	46
133	102	51
134	102	51
136	100	51
138	102	48
138	98	49
141	99	51
147	105	51
149	108	53
153	107	57
155	107	55
155	115	56
158	117	63
159	115	60
162	118	62
177	124	63
97	82	37
99	82	37
104	86	39
103	87	41
106	88	42

Here let X_1 denote the Length, X_2 denote the Width and X_3 denote the Height of turtles.
Consider the problem of testing:

$$H_0 : \mu = (\mu_1, \mu_2, \mu_3)=(149,100,53) \quad \text{vs} \quad H_1 : \text{Not } H_0$$

Here we carry out the test procedure using the Hotelling T^2 Test Statistic.
For this we first need to check whether the data comes from a tri-variate Normal distribution or not, since the Hotelling T^2 test holds true only when the data is Normally distributed with some mean and variance.

We perform the multivariate Shapiro-Wilks test to test for normality of the data. In R, the multivariate Shapiro-Wilks test is under the library "**mvnormtest**" e read the data in R and install the library **mvnormtest** using the following commands. The dataset can be found as .csv file along with this module or can simply be read directly into the R environment.

```
data<-read.csv(file.choose(),header=TRUE)
x1=data$Length
x2=data$Width
x3=data$Height
x=cbind(x1,x2,x3)

install.packages("mvnormtest")
library(mvnormtest)
```

The R command for multivariate Shapiro-Wilks test is **mshapiro.test()**
For our given data the Shapiro-Wilks test gives

```
> mshapiro.test(t(x))

        Shapiro-Wilk normality test

data:  Z
W = 0.94811, p-value = 0.6462
```

As the p-value for the test is 0.6462, we may claim that the data supports normality

Back to the testing problem:

Method 1: From the first principles

We find the sample mean vector and sample variance-covariance matrix

```
> x.bar=apply(x,MARGIN=2,FUN=mean)
> x.bar
     x1        x2        x3
130.17241  98.00000  48.31034

> A=var(x)    ##Sample Var-Cov Matrix
> A

      x1        x2        x3
x1 544.3621 281.92857 181.80172
x2 281.9286 153.85714  97.78571
x3 181.8017  97.78571  65.29310


> mu=c(149,100,53)          ### Specified value of mean under null is(149,100,53)
> n=nrow(x)                 ### Number of observations
> p=ncol(x)                 ### Number of variables

> Tp.2=as.numeric(n*(t((x.bar-mu))%*%solve(A)%*%(x.bar-mu)))

> Test.statistic=(Tp.2*(n-p))/(p*(n-1))
> Test.statistic
[1] 86.83216

> qf(0.95,p,n-p)
[1] 2.975154
```

Thus we see that as Test Statistic F $> F_{p\,,\,n-p\,;\,\alpha}$ H_0 is rejected at 5% level of significance

Method 2: Alternatively one can perform the Hotelling T^2 test using a built-in R function. The built-in R function is **HotellingsT2()** under the library "**ICSNP**". We install the library and perform the test as follows:

```
install.packages("ICSNP")
library(ICSNP)

>test=HotellingsT2(x,mu=mu,test="f")
> print(test)

        Hotelling's one sample T2-test

data:  x
T.2 = 86.832, df1 = 3, df2 = 26, p-value = 1.134e-13
alternative hypothesis: true location is not equal to c(149,100,53)
```

Thus we see that as the p-value for the test is very small H_0 is rejected at 5% level of significance

Let us find the confidence ellipsoid for μ_1 and μ_2 along with the confidence intervals:

Let X$\sim N(\mu, \sigma^2)$ where both μ and σ^2 are unknown.
Then the 100(1-α)% confidence interval for μ is given by

$$\left[\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{n}}\, t_{\frac{\alpha}{2}\,;\,n-1}\,, \hat{\mu} + \frac{\hat{\sigma}}{\sqrt{n}}\, t_{\frac{\alpha}{2}\,;\,n-1}\right]$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimators of μ and σ^2 respectively and
$t_{\frac{\alpha}{2}\,;\,n-1}$ is the upper $\frac{\alpha}{2}^{\text{th}}$ quantile of the t_{n-1} distribution.

In R we compute the confidence intervals μ_1 and μ_2 respectively from the first principls. Note that here the previous notation holds true i.e. μ_1 = population mean length of turtles and μ_2 = population mean width of turtles

```
> ci.x1=c(mean(x1)-(sqrt(var(x1))*qt(0.975,df=n-1)/sqrt(n)),
          mean(x1)+(sqrt(var(x1))*qt(0.975,df=n-1)/sqrt(n)))
> ci.x2=c(mean(x2)-(sqrt(var(x2))*qt(0.975,df=n-1)/sqrt(n)),
          mean(x2)+(sqrt(var(x2))*qt(0.975,df=n-1)/sqrt(n)))

> ci.x1
[1] 121.2976 139.0473
> ci.x2
[1]  93.2818 102.7182
```

The confidence ellipsoid for $\mu^* = (\mu_1, \mu_2)$ is given by the quadratic form

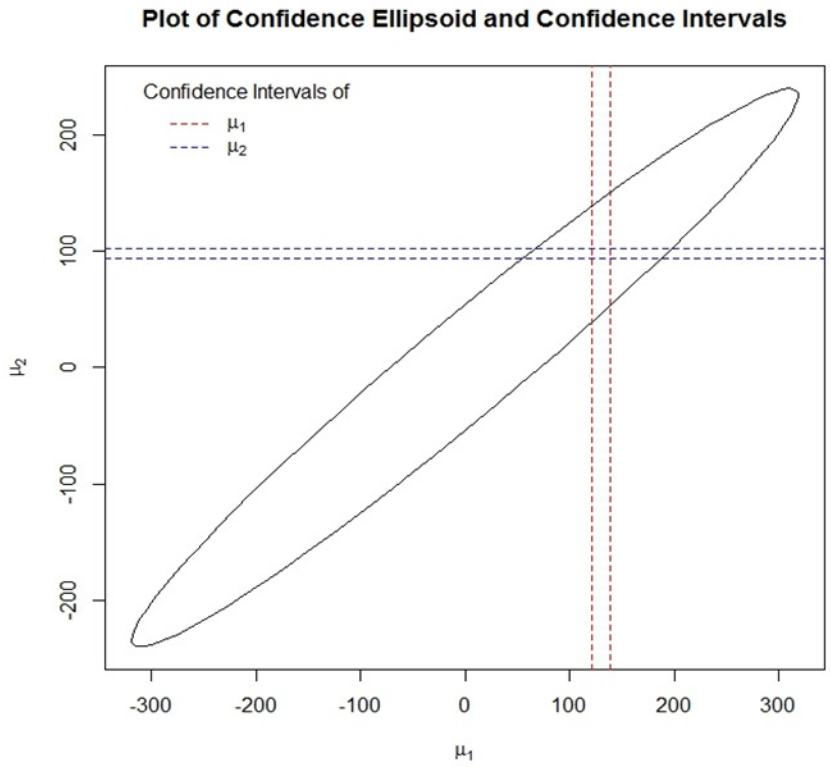
$$\left[\ \left(\mu^* - \bar{\mathbf{X}}^*\right)^{\mathrm{T}}\ \mathrm{S}^{*-1}\left(\mu^* - \bar{\mathbf{X}}^*\right) \leq c^{*2}\right]$$

where $\bar{\mathbf{X}}^* = (\bar{X}_1, \bar{X}_2)$ is the sample mean vector of $\mathbf{X}^*$= (X_1,X_2) and S^* is the sample variance covariance matrix of $\mathbf{X}^*$= (X_1,X_2)
The R command for calculating confidence ellipse is **ellipse()**.
We plot the confidence ellipse centered about the mean of variables X_1 and X_2

Plot of the confidence intervals and confidence ellipsoid:

```
plot(ellipse(cor(x1,x2),c(mean(x1),mean(x2))),type="l",
main="Plot of Confidence Ellipsoid and Confidence Intervals",
xlab=expression(paste(mu)[1]),ylab=expression(paste(mu)[2]))
abline(v=ci.x1,lty=2,col="red")
abline(h=ci.x2,lty=2,col="blue")

legend("topleft","Confidence Intervals of",bty="n")
legend(300,230
,c(expression(paste(mu)[1]),expression(paste(mu)[2])),l
ty=2,col=c("red","blue"),bty="n")
```



Note that the confidence intervals for μ_1 and μ_2 are narrower than the confidence ellipsoid for (μ_1,μ_2). This is due the fact that the variation in the data corresponding to the variables X_1 and X_2 is low.

Consider further the following testing problem:

(i) $H_0 : \mu_1 = \frac{2}{3}\mu_2 + \frac{1}{3}\mu_3$

(ii) $H_0 : \mu_1 = \frac{2}{3}\mu_2 + \frac{1}{3}\mu_3 \quad , \quad \mu_2 = \mu_3$

Result:

If $\mathbf{X} \sim N_p(\mu, \Sigma)$
Then $\mathbf{Y} = \mathbf{C}\mathbf{X} \sim N_q(\mathbf{C}\mu, \mathbf{C}\Sigma\mathbf{C}^{\mathrm{T}})$, where C is a q×p matrix having rank q

Note that for for the above two cases we have

$$(i)\ \mathbf{C} = \left(1, -\frac{2}{3}, -\frac{1}{3}\right) \text{ having rank 1}$$

$$(ii)\ \mathbf{C} = \begin{pmatrix} 1 & -\frac{2}{3} & -\frac{1}{3} \\ 0 & 1 & -1 \end{pmatrix} \text{ having rank 2}$$

Hence for (i) we have $\mathbf{Y} = \mathbf{C}\mathbf{X} \sim N(\mu^* = \mu_1 - \frac{2}{3}\mu_2 - \frac{1}{3}\mu_3, {\sigma^*}^2)$

Hence the testing problem reduces to a univariate test of a normal mean

i.e. $H_0 : \mu^* = 0$

So we first get the transformed data and perform the univariate t test for means

```
##To get the transformed data

> C=c(1,-2/3,-1/3)
>Y=0
>for(i in 1:length(x1))
+{
+Y[i]=C%*%x[i,]
+}
>Y

[1] 31.33333 34.33333 31.66667 34.66667 37.66667 50.33333 46.33333 53.00000
[9] 51.66667 48.00000 49.00000 52.33333 54.00000 56.33333 58.00000 60.00000
[17] 59.33333 62.66667 65.33333 59.66667 59.00000 62.33333 62.66667 73.33333
[25] 30.00000 32.00000 33.66667 31.33333 33.33333

> t.test(Y,alt="two.sided")


One Sample t-test

data:  Y
t = 20.348, df = 28, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 43.82959 53.64168
sample estimates:
mean of x
 48.73563
```

Since p-value < 0.05, $H_0 : \mu^*=0$ is rejected at 5% level of significance

For (ii) we have $\mathbf{Y}$=$\mathbf{CX}\sim N_2\left(\mu^* = \begin{bmatrix} \mu_1 & -\frac{2}{3}\mu_2 & -\frac{1}{3}\mu_3 \\ 0 & \mu_2 & -\mu_3 \end{bmatrix}, \Sigma^*\right)$

Similarly the problem of testing reduces to H_0 : μ^*=0 based on the transformed data $\mathbf{Y}$. So we can use the Hotelling T^2 test for testing H_0 : μ^*=0

As before we first get the transformed data $\mathbf{Y}$ and then perform the Hotelling T^2 test

```
##Getting the transformed data Y

> C=matrix(c(1,-2/3,-1/3,0,1,-1),nrow=2,byrow=TRUE)
> Y=matrix(0,ncol=2,nrow=nrow(x))

> for(i in 1:nrow(x))
+ {
+ Y[i,1]=C[1,]%*%x[i,]
+ Y[i,2]=C[2,]%*%x[i,]
+ }

> mu0=c(0,0)              ##Hypothised mean
> test2=HotellingsT2(Y,mu=mu0,test="f")
> print(test2)

        Hotelling's one sample T2-test

data:  Y
T.2 = 2453, df1 = 2, df2 = 27, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(0,0)
```

Since p-value < 0.05, H_0 : μ^*=0 is rejected at 5% level of significance

Two sample Hotelling T^2 test:

An example:

A pharmaceutical company had a drug which they want to test for effectiveness in reducing some topical diseases symptoms. A random sample of 20 people with the disease was given the drug. Based on this data and the data for the control group of size 18,we wish determine whether there is a significant difference between the drug and placebo in reducing the symptoms.

The data is as follows:

	Drug				Placebo	
Fever	Pressure	Aches		Fever	Pressure	Aches
36.5	72	18		40.9	54	14
36.6	84	16		39.5	75	18
38.2	60	29		39.4	57	24
37.6	82	13		38.2	71	24
37	68	25		39.7	65	22
37.9	54	27		38.9	49	30
37.4	80	25		38.6	58	25
35.2	99	8		39.9	52	17
38.2	65	21		41.3	62	18
37.5	55	11		38.1	57	20
35.8	70	16		39.6	78	19
37.4	76	13		37.1	92	15
37.2	49	29		39.5	63	13
36.5	59	24		40.3	52	25
38.3	77	12		41.5	46	27
37.5	66	19		39.3	56	14
36	79	14		37.6	86	16
36.9	67	12		40.6	48	21
39.3	53	7				
38.8	67	13				

So here we assume
$\mathbf{X}$=(Fever, Pressure, Aches) corresponding to the Drug group with mean μ_1
$\mathbf{Y}$=(Fever, Pressure, Aches) corresponding to the Placebo group with mean μ_2

So the testing problem here is

$H_0 : \mu_1 = \mu_2 \quad \text{ag} \quad H_1 : \mu_1 \neq \mu_2$

First we check for the normality of the data and as before we perform the multivariate Shapiro-Wilks test for normality

```
##Test for Multivariate Normality

>library(mvnormtest)

>mshapiro.test(t(X))

Shapiro-Wilk normality test

data:  Z
W = 0.90157, p-value = 0.04414


> mshapiro.test(t(Y))

Shapiro-Wilk normality test

data:  Z
W = 0.94289, p-value = 0.3245
```

Thus we see that $\mathbf{Y}$ supports normality. However, while testing for normality of $\mathbf{X}$, the p-value of the test turns out to be $0.04414 < 0.05$ whereby rejecting the claim that the data is normally distributed.

It may be however assumed with certain relaxation that $\mathbf{X}$ is also normally distributed, since the p-value of the test is not much less than the significance level 0.05

So now we perform the two sample Hotelling T^2 test using R. The R command is the same **HotellingsT2()** under "**ICSNP**" package. The only difference is that here along with the first sample we also provide the second sample

```
> HotellingsT2(X,Y)

        Hotelling's two sample T2-test

data:  X and Y
T.2 = 14.115, df1 = 3, df2 = 34, p-value = 3.857e-06
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

Since p-value < 0.05, $H_0 : \mu_1 = \mu_2$ is rejected at 5% level of significance

MANOVA:

Consider the skulls dataset from the R library(HSAUR), which gives the measurement of various dimensions of skulls of Egyptians over the years

epoch	mb	bh	bl	nh
c4000BC	131	138	89	49
c4000BC	125	131	92	48
c4000BC	131	132	99	50
c4000BC	119	132	96	44
c4000BC	136	143	100	54
c4000BC	138	137	89	56
.....	...	...	...	...
cAD150	129	128	81	52
cAD150	140	135	103	48
cAD150	147	129	87	48
cAD150	136	133	97	51

Here the variable of interest is $\mathbf{Y}$= (mb , bh , bl , nh)
And we wish to test whether the measurement of dimensions of skull vary over the years.
Here we have 5 years, namely: 4000BC 3300BC 1850BC 200BC AD150

Let $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ be the means corresponding to the different years and we wish to test the following hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ ag H_1: Not H_0

First we read the data into the R environment and as usual we test for multivariate normality of the data

```
> data<-read.csv(file.choose())
> names(data)
[1] "X"     "epoch" "mb"    "bh"    "bl"    "nh
> attach(data)
> Y1=cbind(mb,bh,bl,nh)
> library(mvnormtest)
> mshapiro.test(t(Y1))

        Shapiro-Wilk normality test

data:  Z
W = 0.98687, p-value = 0.1685
```

As p-value = 0.1685 > 0.05, the data is Multivariate Normal

MANOVA is a generalization of Hotelling T^2 test. When we have more than 2 multivariate normal populations and we are interested in testing for the equality of the population means we perform MANOVA or Multivariate Analysis of Variance.

The R code for performing MANOVA test is given by **manova()**.

There are several MANOVA test criterion like "Pillai's trace", "Wilks Λ", "Roy's Union-Intersection" etc. These can be specified by the 'test' argument within the R function **summary()** when the summary is taken over the output from **manova()**.

```
> manova1<-manova(Y1~as.factor(epoch))
> summary(manova1,test="Pillai")

                  Df  Pillai approx F num Df den Df    Pr(>F)
as.factor(epoch)   4 0.35331    3.512     16    580 4.675e-06 ***
Residuals        145
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Thus we see as p-value < 0.05, the mean skull measurements over the years are different

Similarly we can also get the other MANOVA tests simply by changing the argument 'test'

```
> summary(manova1,test="Wilks")
> summary(manova1,test="Roy")
> summary(manova1,test="Hotelling-Lawley")
```

So our next natural question is: Which variables are significantly different over the years?

Such a question may be answered if we test for the equality of the means over the years corresponding to the marginal variables Y_1=mb,Y_2=bh,Y_3=bl and Y_4=nh

In R this test for equality of marginal means over the years may be performed very simply with the help of the command **summary.aov()** and the argument for this function is the manova() output

```
> summary.aov(manova1)

 Response mb :
                  Df  Sum Sq Mean Sq F value    Pr(>F)
as.factor(epoch)   4  502.83 125.707  5.9546 0.0001826 ***
Residuals        145 3061.07  21.111
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 Response bh :
                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(epoch)   4  229.9  57.477  2.4474 0.04897 *
Residuals        145 3405.3  23.485
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 Response bl :
                  Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(epoch)   4  803.3 200.823  8.3057 4.636e-06 ***
Residuals        145 3506.0  24.179
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 Response nh :
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(epoch)   4   61.2  15.300   1.507 0.2032
Residuals        145 1472.1  10.153
```

Thus we see that the variable Y_2=bh is **Marginally Significant** over the years as the p-value for the equality of the means over time for variable Y_2=bh is 0.04897 is marginally smaller than 0.05

And the variable Y_4=nh is **Not Significant** over the years as the p-value for the equality of the means over time for variable Y_4=nh is 0.2032

The other two variables however, are significant over the years as can be understood from the small p-values

SUMMARY

- Hotelling T^2 test is the multivariate analogue of the univariate t-test for testing any hypothesis regarding the population vector, when samples are drawn from a multivariate normal distribution with unknown mean and unknown dispersion matrix
- Like the two sample t-test, the multivariate analogue is the two sample Hotelling T^2 test. Here samples are drawn from two independent multivariate normal distributions with unknown yet equal dispersion matrices and we are testing for equality of the mean vectors
- When there are more than two normal populations with unknown but equal variance, ANOVA is used to test hypothesis regarding the equality of the population means. The multivariate version of ANOVA is known as MANOVA
- For the Hotelling T^2 test using R, we can either go for testing using first principles or use the **HotellingsT2()** function under the library **ICSNP**
- For the two sample Hotelling T^2 test using R, we can go for testing using first principles. However it becomes a bit cumbersome. The same R function **HotellingsT2()** also performs the two sample Hotelling T^2 test
- MANOVA in R can be performed very easily using the R command **manova()**. There are several other R functions which aid in the analysis of further questions which may arise when the test for equality of population means rejects the claim of equality
- In all cases the very first step in analysis is checking for the normality of the data, as all the topics discussed like Hotelling T^2 test or MANOVA etc. hold true only when the data comes from normal population or populations