Test of Hypothesis for Multivariate Mean: One- and Two-sample

Tanweer Shapla

2023-10-04

Hotelling’s test is a multivariate statistical test used to see if p popultion means have particular values, or to compare the means of two groups. It is an extension of the univariate t-test to multiple dimensions. Hotelling’s test is particularly useful in cases where you have multiple dependent variables and you want to determine if there are significant differences between the means of these variables across two groups.

Here are the key steps for performing Hotelling’s test:

Assumptions:

1. Multivariate Normality: The data should follow a multivariate normal distribution.
2. Homogeneity of Covariance: The variances of the groups should be equal across all outcome variables.
3. Random Sampling: The data should be obtained through random sampling.

## Hotelling’s T2 test for One-sample mean

Here the null hypothesis is , and the alternative hypothesis is .

Below we work on swiss data considering 3 variables, namely, Agriculture, Examination, and Education. We use Hotelling’s test statistic to test one sample multivariate mean.

.

Under the null hypothesis, follows an distribution with numerator degrees of freedom and denominator degrees of freedom .

We wish to test vs . We use HotellingsT2() function available in ICSNP package from R.

#downloading necessary packages below  
#install.packages("ICSNP")  
#ICSNP depends on mvtnorm and ICS packages  
library(mvtnorm)

Warning: package 'mvtnorm' was built under R version 4.3.1

library(ICS)

Warning: package 'ICS' was built under R version 4.3.1

library(ICSNP) #needed to use HotellingsT2()

Warning: package 'ICSNP' was built under R version 4.3.1

newswiss=subset(swiss, select=c("Agriculture", "Examination","Education"))  
colMeans(newswiss) #calculating mean for each variable

Agriculture Examination Education   
 50.65957 16.48936 10.97872

HotellingsT2(newswiss, mu=c(51,16,11), test="f")

Hotelling's one sample T2-test  
  
data: newswiss  
T.2 = 0.1211, df1 = 3, df2 = 44, p-value = 0.9472  
alternative hypothesis: true location is not equal to c(51,16,11)

Based on the p-value, we fail to reject the null hypothesis at level, and conclude that the true population means are not significantly different than .

To test , we use

HotellingsT2(newswiss, mu=NULL, test=“f”).

If we want to use the chi-squared approximation of Hotelling’s test statistic to test vs , we use test=“chi” option:

HotellingsT2(newswiss, mu=c(51,16,11), test="chi")

Hotelling's one sample T2-test  
  
data: newswiss  
T.2 = 0.37981, df = 3, p-value = 0.9444  
alternative hypothesis: true location is not equal to c(51,16,11)

## Hotelling’s T2 test for Two-sample means

Here we wish to test if the two population mean difference is equal to zero or some specific number.

Thus, for testing mean difference is zero or not, the null and alternative hypotheses are

vs

For testing mean difference equal to some constant or not, the null and alternative hypotheses are

vs where is a vector of constants .

## Example 2

In this example, we consider two sets of data, where each set is generated from a 2-variate normal variables with specific mean vector and covariance matrix . First and second set have respectively 10 and 20 random observations.

set.seed(12)  
x=rmvnorm(10, mean=c(3, 2), sigma=matrix(c(2,.5,.5,1), nrow=2,ncol=2))  
#print(x)  
y=rmvnorm(20, mean=c(5, 6), sigma=matrix(c(2,.5,.5,1), nrow=2,ncol=2))  
#print(y)

Now we perform Hotelling’s T2 test to see if the mean difference is (-2,-4), that is,

vs

HotellingsT2(x, y, mu=c(-2,-4)) # testing if mean difference is (-2, -4)

Hotelling's two sample T2-test  
  
data: x and y  
T.2 = 1.7866, df1 = 2, df2 = 27, p-value = 0.1868  
alternative hypothesis: true location difference is not equal to c(-2,-4)

Based on the p-valie or 0.1868, we fail to reject the null hypothesis as expected.

For the same data, we now test if the mean difference is (0,0), that is,

vs

HotellingsT2(x, y, mu=NULL)

Hotelling's two sample T2-test  
  
data: x and y  
T.2 = 66.946, df1 = 2, df2 = 27, p-value = 3.429e-11  
alternative hypothesis: true location difference is not equal to c(0,0)

Once is rejected, we would like to see which variable mean difference is significantly different from the hypothesized value. This can be achieved by forming Bonferroni confidence intervals for each mean difference.

We find Bonferroni CI for each mean using t.test() function:

t.test(x[,1],y[,1], mu=0, alt="t", conf.level = 1-0.05/2)

Welch Two Sample t-test  
  
data: x[, 1] and y[, 1]  
t = -6.0647, df = 13.321, p-value = 3.606e-05  
alternative hypothesis: true difference in means is not equal to 0  
97.5 percent confidence interval:  
 -3.926062 -1.617985  
sample estimates:  
mean of x mean of y   
 2.281958 5.053981

t.test(x[,2],y[,2], mu=0, alt="t", conf.level = 1-0.05/2)

Welch Two Sample t-test  
  
data: x[, 2] and y[, 2]  
t = -12.233, df = 22.743, p-value = 1.753e-11  
alternative hypothesis: true difference in means is not equal to 0  
97.5 percent confidence interval:  
 -5.122080 -3.442055  
sample estimates:  
mean of x mean of y   
 1.726078 6.008145

## Example 3: Psychological data

In this example, we will use Psychological Test data that is available in Canvas in Data unit. Download data from canvas, save it in a folder and use setwd() function to provide the location of the data.

setwd("C:/Users/tshapla/Desktop/COURSES/Stat 577/RPrograms")#change this path   
  
PT=read.table("PsychologicalTests.DAT", header=F)  
colnames(PT)=c("gender", "t1","t2", "t3", "t4")  
library(ICSNP)  
gender=as.factor(PT$gender)  
with(PT,HotellingsT2(cbind(t1,t2,t3,t4)~gender))

Hotelling's two sample T2-test  
  
data: cbind(t1, t2, t3, t4) by gender  
T.2 = 23.22, df1 = 4, df2 = 59, p-value = 1.464e-11  
alternative hypothesis: true location difference is not equal to c(0,0,0,0)

## Finding Bonferroni Confidence Intervals

Since we reject the null hypothesis above, we now wish to investigate which mean differences are significantly different that zero. We do so by constructing Bonferroni confidence intervals for each variable.

t.test(t1~gender, data=PT, conf.level = 1-0.05/4, alt="t")

Welch Two Sample t-test  
  
data: t1 by gender  
t = 5.4173, df = 57.634, p-value = 1.234e-06  
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
98.75 percent confidence interval:  
 1.899573 5.350427  
sample estimates:  
mean in group 1 mean in group 2   
 15.96875 12.34375

t.test(t2~gender, data=PT,conf.level = 1-0.05/4, alt="t")

Welch Two Sample t-test  
  
data: t2 by gender  
t = 2.0066, df = 60.249, p-value = 0.04928  
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
98.75 percent confidence interval:  
 -0.5663741 4.5663741  
sample estimates:  
mean in group 1 mean in group 2   
 15.90625 13.90625

t.test(t3~gender, data=PT,conf.level = 1-0.05/4, alt="t")

Welch Two Sample t-test  
  
data: t3 by gender  
t = 7.7748, df = 61.966, p-value = 9.765e-11  
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
98.75 percent confidence interval:  
 7.046424 14.016076  
sample estimates:  
mean in group 1 mean in group 2   
 27.18750 16.65625

t.test(t4~gender, data=PT, conf.level = 1-0.05/4, alt="t")

Welch Two Sample t-test  
  
data: t4 by gender  
t = 0.68791, df = 58.235, p-value = 0.4942  
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
98.75 percent confidence interval:  
 -2.23201 3.85701  
sample estimates:  
mean in group 1 mean in group 2   
 22.7500 21.9375

Or, alternately:

ci=array(0,c(4,2))  
for (i in 2:5){  
 result=t.test(PT[,i]~gender,conf.level =1-0.05/4, alt="t")  
 ci[i-1,]=result$conf.int  
}  
print(ci)

[,1] [,2]  
[1,] 1.8995727 5.350427  
[2,] -0.5663741 4.566374  
[3,] 7.0464238 14.016076  
[4,] -2.2320096 3.857010