Multivariate Analysis with R

Testing of Hypothesis (Hotelling T^2 Statistics)
Tests for Mean Vector in One and Two Samples

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): One Sample

Suppose the null hypothesis is $H_0: \underline{\mu} = \underline{\mu}_0$ where $\underline{\mu}_0$ is $(p \times 1)$ vector of known values.

Multivariate analogue of $T_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is

$$T^2 = n(\underline{\bar{x}} - \underline{\mu}_0)' S^{-1} (\underline{\bar{x}} - \underline{\mu}_0)$$

T^2 is called as Hotelling T^2 - statistics which follows

$$\frac{n-k+1}{k} \frac{T^2}{n} \sim F_{(k, n-k+1)}$$

under $H_0: \underline{\mu} = \underline{\mu}_0$. It is termed as T^2 - distribution with *n* degrees of freedom.

Confidence Region for $\underline{\mu}$ (Σ unknown): One Sample

The confidence region for $\underline{\mu}$ with confidence level $(1-\alpha)$ is the set of vectors $\underline{\mu}$

$$n(\underline{\bar{x}}-\underline{\mu})'S^{-1}(\underline{\bar{x}}-\underline{\mu}) \leq T^2_{k,n-1}$$

or

$$n(\underline{\bar{x}}-\underline{\mu})'S^{-1}(\underline{\bar{x}}-\underline{\mu}) \leq \frac{(n-1)k}{n-k} F_{(k,n-k+1)}\,(\alpha)$$

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): One Sample- R command

The test requires the package `MVTests`

```
install.packages("MVTests")

library(MVTests)
```

Description

`OneSampleHT2 computes one sample Hotelling T^2 statistics and gives confidence intervals`

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): One Sample - R command

Usage `OneSampleHT2(data, mu0, alpha = 0.05)`

`Arguments`

`data` a matrix or data frame.

`mu0` mean vector that is used to test whether population mean parameter is equal to it.

`alpha` Significance Level that will be used for confidence intervals. default alpha=0.05.

Details

This function computes one sample Hotelling T^2 statistics that is used to test whether population mean vector is equal to a vector given by a user. When H0 is rejected, this function computes confidence intervals for all variables.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): One Sample - R command

Value : a list with 7 elements:

`HT2` The value of Hotelling T^2 Test Statistic

`F` The value of F Statistic

`df` The F statistic's degree of freedom

`p.value` p value

`CI` The lower and upper limits of confidence intervals obtained for all variables

`alpha` The alpha value using in confidence intervals

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): One Sample - Example

Suppose the data on 5 persons is obtained on three variables - ages (in years), weights (in Kgs.) and hemoglobin (hb) levels - as follows:

Persons	1	2	3	4	5
age	10	15	20	25	30
weight	35	40	45	50	55
hb	11	12	13	14	15

Suppose we want to test $H_0: \underline{\mu} = \underline{\mu}_0$ at α = 5% level of significance where $\underline{\mu}_0 = (25, 45, 13)'$ and Σ is unknown.

Hypothesis Testing for $\underline{\mu}$ (Σ known): One Sample - Example

```
age = c(10, 15, 20, 25, 30)

weight = c(35, 40, 45, 50, 55)

hb = c(11, 12, 13, 14, 15)

datam = matrix(nrow=5, ncol=3, data=c(age, weight, hb),
byrow = T)

mu0 = matrix(nrow=3, ncol=1, data= c(25, 45, 13))

OneSampleHT2(datam, mu0, alpha = 0.05)
```

Hypothesis Testing for $\underline{\mu}$ (Σ known): One Sample – Example

```
> OneSampleHT2(datam, mu0, alpha = 0.05)
```

```
$HT2
         [,1]
[1,] 3773.118

$F
         [,1]
[1,] 628.8531

$df
[1] 3 2

$p.value
            [,1]
[1,] 0.001588092
```

```
$CI
      Lower     Upper Mu0 Important Variables?
1 -62.04702 119.24702  25                FALSE
2 -45.74529  91.74529  45                FALSE
3 -49.77205 102.57205  13                FALSE

$alpha
[1] 0.05

$Descriptive
          [,1]     [,2]     [,3]
N      5.00000  5.00000  5.00000
Means 28.60000 23.00000 26.40000
Sd    18.90238 14.33527 15.88395
```

Hypothesis Testing for $\underline{\mu}$ (Σ known): One Sample – Example

```
R Console
> library(MVTests)
> OneSampleHT2(datam, mu0, alpha = 0.05)
$HT2
         [,1]
[1,] 3773.118

$F
         [,1]
[1,] 628.8531

$df
[1] 3 2

$p.value
            [,1]
[1,] 0.001588092

$CI
      Lower     Upper Mu0 Important Variables?
1 -62.04702 119.24702  25                FALSE
2 -45.74529  91.74529  45                FALSE
3 -49.77205 102.57205  13                FALSE

$alpha
[1] 0.05

$Descriptive
          [,1]     [,2]     [,3]
N      5.00000  5.00000  5.00000
Means 28.60000 23.00000 26.40000
Sd    18.90238 14.33527 15.88395

$Test
[1] "OneSampleHT2"

attr(,"class")
[1] "MVTests" "list"
> |
```

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples

Suppose that

- $\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, \dots, \underline{x}_{n_1}^{(1)}$ is a random sample of size n_1 from a multivariate normal distribution with parameters mean vector $\underline{\mu}^{(1)}$ and unknown covariance matrix $\Sigma > 0$.
- $\underline{x}_1^{(2)}, \underline{x}_2^{(2)}, \dots, \underline{x}_{n_2}^{(2)}$ is a random sample of size n_2 from a multivariate normal distribution with parameters mean vector$\underline{\mu}^{(2)}$and unknown covariance matrix $\Sigma > 0$.

Suppose the null hypothesis is $H_0 : \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$ where $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ are $(p \times 1)$ vectors.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples

1. Based on the sample $\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, \dots, \underline{x}_{n_1}^{(1)}$
   - $\underline{\mu}^{(1)}$ is estimated as $\underline{\bar{x}}_1 = \frac{1}{n}\sum_{\alpha=1}^{n_1} \underline{x}_\alpha^{(1)}$
   - Σ is estimated as $S_1 = \frac{1}{n_1-1}\sum_{\alpha=1}^{n_1}(\underline{x}_\alpha^{(1)} - \underline{\bar{x}}_1)(\underline{x}_\alpha^{(1)} - \underline{\bar{x}}_1)'$
2. Based on the sample $\underline{x}_1^{(2)}, \underline{x}_2^{(2)}, \dots, \underline{x}_{n_2}^{(2)}$
   - $\underline{\mu}^{(2)}$ is estimated as $\underline{\bar{x}}_2 = \frac{1}{n}\sum_{\alpha=1}^{n_2} \underline{x}_\alpha^{(2)}$
   - Σ is estimated as $S_2 = \frac{1}{n_2-1}\sum_{\alpha=1}^{n_2}(\underline{x}_\alpha^{(2)} - \underline{\bar{x}}_2)(\underline{x}_\alpha^{(2)} - \underline{\bar{x}}_2)'$
3. Find pooled covariance matrix $S = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples

The Hotelling T^2 statistic for testing $H_0: \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$ is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)})' S^{-1} (\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)})$$

which follows

$$\frac{n_1 + n_2 - k + 1}{k} \frac{T^2}{n_1 + n_2} \sim F_{(k, n_1 + n_2 - k + 1)}$$

under $H_0: \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$. It is termed as T^2 - distribution with $(n_1 + n_2)$ degrees of freedom.

Confidence Region for $\underline{\mu}$ (Σ unknown): Two Samples

The confidence region for $\underline{\mu}^{(1)} = \underline{\mu}^{(2)}$ with confidence level $(1-\alpha)$ is the set of vectors $\underline{m}$

$$\left(\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)} - \underline{m}\right)' S^{-1}\left(\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)} - \underline{m}\right) \leq \frac{n_1 n_2}{n_1 + n_2} T^2_{k, n_1+n_2-2}$$

or

$$\left(\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)} - \underline{m}\right)' S^{-1}\left(\underline{\overline{x}}^{(1)} - \underline{\overline{x}}^{(2)} - \underline{m}\right)$$

$$\leq \left(\frac{n_1 n_2}{n_1 + n_2}\right) \frac{(n_1 + n_2 - 2)}{n_1 + n_2 - k + 1} F_{(k, n_1+n_2-k+1)}(\alpha)$$

Test Decisons using Confidence Intervals:

Interesting and useful relationship between confidence intervals and hypothesis tests

If H_0 is rejected at the significance level α, then there exists a 100(1- α)% confidence interval which yields the same conclusion as the test.

Suppose H_0: $\theta = \theta_0$

If the confidence interval does not contain the value θ_0, then H_0 is rejected.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples- R command

The test requires the package `MVTests`

```
install.packages("MVTests")

library(MVTests)
```

Description

`TwoSamplesHT2` function computes Hotelling T^2 statistic for two independent samples and gives confidence intervals.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples- R command

Usage `TwoSamplesHT2(data, group, alpha = 0.05, Homogenity = TRUE)`

Arguments

`data` a data frame.

`group` a group vector consisting of 1 and 2 values.

`alpha` Significance Level that will be used for confidence intervals. default=0.05

`Homogenity` a logical argument. If sample covariance matrices are homogeneity, then `Homogenity=TRUE`. Otherwise `Homogenity=FALSE` The homogeneity of covariance matrices can be investigated with `BoxM` function.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples- R command

Details

This function computes two independent samples Hotelling T^2 statistics that is used to test whether two population mean vectors are equal to each other. When H0 is rejected, this function computes confidence intervals for all variables to determine variable(s) affecting on rejection decision. Moreover, when covariance matrices are not homogeneity, the approach proposed by D. G. Nel and V. D. Merwe (1986) is used.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples- R command

Value : a list with 8 elements:

`HT2` The value of Hotelling T^2 Test Statistic

`F` The value of F Statistic

`df` The F statistic's degree of freedom

`p.value` p value

`CI` The lower and upper limits of confidence intervals obtained for all variables

`alpha` The alpha value using in confidence intervals

`Descriptive1` Descriptive Statistics for the first group

`Descriptive2` Descriptive Statistics for the second group

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples - Example

Suppose the data on 10 persons in two groups is obtained on three variables - ages (in years), weights (in Kgs.) and hemoglobin (hb) levels - as follows:

Group 1 ~ N($\underline{\mu}^{(1)}$, Σ)						Group 2 ~ N($\underline{\mu}^{(2)}$, Σ)					
Persons	1	2	3	4	5	Persons	1	2	3	4	5
age	10	15	20	25	30	age	12	16	21	26	32
weight	35	40	45	50	55	weight	33	42	48	52	54
hb	11	12	13	14	15	hb	10	11	13	13	14

Suppose we want to test $H_0: \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$ at α = 5% level of significance and Σ is unknown.

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples - Example

```
age12 = c(10, 15, 20, 25, 30, 12, 16, 21, 26, 32)

weight12 = c(35, 40, 45, 50, 55, 33, 42, 48, 52, 54)

hb12 = c(11, 12, 13, 14, 15, 10, 11, 13, 13, 14)

datam = matrix(nrow=10, ncol=3, data=c(age12, weight12,
hb12), byrow = F)

G = c(rep(1,times=5), rep(2,times=5))

library(MVTests)

TwoSamplesHT2(datam, group = G, alpha = 0.05, Homogenity
= TRUE)
```

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples - Example

```
> TwoSamplesHT2(datam, group = G, alpha = 0.05, Homogenity = TRUE)
```

```
$HT2
         [,1]
[1,] 38.66005

$F
         [,1]
[1,] 9.665012

$df
[1] 3 6

$p.value
           [,1]
[1,] 0.01029002
```

```
$CI
       Lower     Upper Important Variables?
1 -23.236849 20.436849                FALSE
2 -23.441137 21.841137                FALSE
3  -3.648526  5.248526                FALSE

$alpha
[1] 0.05

$Descriptive1
           [,1]      [,2]      [,3]
Means 20.000000 45.000000 13.000000
Sd     7.905694  7.905694  1.581139

$Descriptive2
           [,1]      [,2]      [,3]
Means 21.400000 45.800000 12.200000
Sd     7.924645  8.497058  1.643168
```

Hypothesis Testing for $\underline{\mu}$ (Σ unknown): Two Samples - Example

R Console

```
> TwoSamplesHT2(datam, group=G, alpha = 0.05, Homogenity = TRUE)
$HT2
         [,1]
[1,] 38.66005

$F
         [,1]
[1,] 9.665012

$df
[1] 3 6

$p.value
           [,1]
[1,] 0.01029002

$CI
       Lower     Upper Important Variables?
1 -23.236849 20.436849                FALSE
2 -23.441137 21.841137                FALSE
3  -3.648526  5.248526                FALSE

$alpha
[1] 0.05

$Descriptive1
           [,1]      [,2]      [,3]
Means 20.000000 45.000000 13.000000
Sd     7.905694  7.905694  1.581139

$Descriptive2
           [,1]      [,2]      [,3]
Means 21.400000 45.800000 12.200000
Sd     7.924645  8.497058  1.643168

$Test
[1] "TwoSamplesHT2"

attr(,"class")
[1] "MVTests" "list"
> 
```