ORIGIN ≡ 1

Hotelling's T^2 Test for Two Small Populations with Equal Covariance Matrices

Hotelling's T^2 test for two populations X & Y that are not matched (paired) is analogous to the univariate Student's t-test in structure, except that p observations are collected for both X $[X_1, X_2, \dots, X_p]$ and for Y $[Y_1, Y_2, \dots, Y_p]$. As in the univariate test, equivalence in variability is an issue. However the multivariate case, the requirement is more complex and stringent because it involves all variances and covariances in covariance matrices Σ_1 & Σ_2 (estimated by S_1 & X_2) respectively for populations X & Y. In addition, when sample sizes are "small" multivariate normality of both X and Y must be assumed, as well as equivalence of covariance, that is $\Sigma_1 = \Sigma_2$. However, when sizes are "large", neither assumption is necessary. Just what constitutes "small" versus "large" is clearly a judgement call as is, to a degree, assessment of equivalence of Σ_1 & Σ_2. JW suggest "without much factual support" that ratios in variance of S_1 versus S_2 (along the main diagonal, (i.e $S_{1(i,i)}$ /$S_{2(i,i)}$) greater than 4 should be considered "serious". This example is drawn from Table 5.1 in AC. Rencher (AR) *Methods of Multivariate Analysis* 1995, with significant clarification and testing of formulas and R script based on Example 6.4 in RA Johnson & DW Wichern (JW) *Applied Multivariate Statistical Analysis 4th Edition* 1998.

Read Data:

Psychological Test scores 1=males, 2=females AR Table 5.1..

M := READPRN("c:\DATA\Multivariate\AR-T5-1-longform.txt")

$X := \text{submatrix}(M, 1, 32, 3, 6)$

$Y := \text{submatrix}(M, 33, 64, 3, 6)$

$n_1 := \text{rows}(X)$ $n_1 = 32$ $n_2 := \text{rows}(Y)$ $n_2 = 32$

$p := \text{cols}(X)$ $p = 4$

M =

	1	2	3	4	5	6
1	1	1	15	17	24	14
2	2	1	17	15	32	26
3	3	1	15	14	29	23
4	4	1	13	12	10	16
5	5	1	20	17	26	28
6	6	1	15	21	26	21
7	7	1	15	13	26	22
8	8	1	13	5	22	22
9	9	1	14	7	30	17
10	10	1	17	15	30	27
11	11	1	17	17	26	20
12	12	1	17	20	28	24
13	13	1	15	15	29	24
14	14	1	18	19	32	28
15	15	1	18	18	31	27
16	16	1	15	14	26	21

Summary Statistics:

$i := 1..n_1$ $ii := 1..n_2$ $j := 1..p$

$I_1 := \text{identity}(n_1)$ $I_2 := \text{identity}(n_2)$ $l_{n1_i} := 1$ $l_{n2_i} := 1$

Mean Vectors:

$$X_{bar} := \frac{1}{n_1} \cdot X^T \cdot l_{n1} \qquad X_{bar} = \begin{pmatrix} 15.969 \\ 15.906 \\ 27.188 \\ 22.75 \end{pmatrix}$$

$$Y_{bar} := \frac{1}{n_2} \cdot Y^T \cdot l_{n2} \qquad Y_{bar} = \begin{pmatrix} 12.344 \\ 13.906 \\ 16.656 \\ 21.938 \end{pmatrix}$$

Covariance Matrices:

$$S_1 := \frac{1}{n_1 - 1} \cdot X^T \cdot \left(I_1 - \frac{1}{n_1} \cdot l_{n1} \cdot l_{n1}^T \right) \cdot X \qquad S_1 = \begin{pmatrix} 5.19254 & 4.54536 & 6.52218 & 5.25 \\ 4.54536 & 13.18448 & 6.76008 & 6.26613 \\ 6.52218 & 6.76008 & 28.67339 & 14.46774 \\ 5.25 & 6.26613 & 14.46774 & 16.64516 \end{pmatrix}$$

$$S_2 := \frac{1}{n_2 - 1} \cdot Y^T \cdot \left(I_2 - \frac{1}{n_2} \cdot l_{n2} \cdot l_{n2}^T \right) \cdot Y \qquad S_2 = \begin{pmatrix} 9.13609 & 7.5494 & 4.86391 & 4.15121 \\ 7.5494 & 18.60383 & 10.2248 & 5.44556 \\ 4.86391 & 10.2248 & 30.03931 & 13.49395 \\ 4.15121 & 5.44556 & 13.49395 & 27.99597 \end{pmatrix}$$

Observed Difference Vector:

$d := X_{bar} - Y_{bar}$

$$d = \begin{pmatrix} 3.625 \\ 2 \\ 10.531 \\ 0.813 \end{pmatrix}$$

Pooled Covariance Matrix:

$$S_{pooled} := \frac{(n_1 - 1)\cdot S_1 + (n_2 - 1)\cdot S_2}{(n_1 + n_2 - 2)}$$

$$S_{pooled} = \begin{pmatrix} 7.164315 & 6.047379 & 5.693044 & 4.700605 \\ 6.047379 & 15.894153 & 8.49244 & 5.855847 \\ 5.693044 & 8.49244 & 29.356351 & 13.980847 \\ 4.700605 & 5.855847 & 13.980847 & 22.320565 \end{pmatrix}$$

Assumptions:

X $[X_1, X_2, \ldots, X_p]$ rs $N_p(\mu_1, \Sigma_1)$
Y $[Y_1, Y_2, \ldots, Y_p]$ rs $N_p(\mu_2, \Sigma_2)$
X independent of Y
Because n_1 & n_2 are small $\Sigma_1 = \Sigma_2$

This test is designed to be used for two samples (n_1 & n_2) of limited size. Once samples reach sufficiently large size, the assmuption of Normality of the two populations and equal covariance matrices is no longer necessary. In this case, the T^2 sampling distribution approaches χ^2 (see MHT 040).

Hypotheses:

Specify a test vector μ_0:

$$\delta_o := \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

H_0: $\delta_0 = \mu_1 - \mu_2 = \delta_0$ **< δ_0 (0,0) or any value of difference between X & Yis a specified value for δ**

H_1: $\delta_0 = \mu_1 - \mu_2 \neq \delta_0$ **< Note: Unlike univariate t-tests, All Hotelling T^2 tests are two-sided since the T^2 distribution, or equivalent F distribution, is asymmetric.**

Hotelling's Test Statistic:

$$T_{sq} := (d - \delta_o)^T \cdot \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\cdot S_{pooled}\right]^{-1} \cdot (d - \delta_o) \qquad T_{sq} = (97.6015)$$

Wilk's Lambda:

$$\Lambda := \frac{1}{\frac{T_{sq}}{(n_1 + n_2 - 2)} + 1} \qquad \Lambda = (0.3884675)$$

See AR Eq. 5-17 p. 147

Sampling Distribution:

If Assumptions hold and H_0 is true, then $T_{sq} \sim T^2_{(n-1)} = [(n_1+n_2-2)\cdot p)/(n_1+n_2-p-1)]\ F_{(p,n1+n2-p-1)}$

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$$C := \frac{(n_1 + n_2 - 2)\cdot p}{(n_1 + n_2 - p - 1)} \cdot qF(1 - \alpha, p, n_1 + n_2 - p - 1) \qquad C = 10.625777$$

^ NOTE: qF(1-α) is used in function qF.

< C^2 gives the boundary for the confidence ellipsoid. See JW eq. 5-18 p. 221

Decision Rule:

Reject H_0 if $T_{sq} > C$

$T_{sq} = (97.601)$ $C = 10.625777$

$Decision := if\left(T_{sq_1} > C, 1, 0\right)$ $Decision = 1$

< 0 = Do not reject H_0
1 = Reject H_0

^ Therefore REJECT H_0

Probability Value:

$$P := 1 - pF\left[T_{sq}\cdot\left[\frac{(n_1 + n_2 - p - 1)}{[(n_1 + n_2 - 2)\cdot p]}\right], p, n_1 + n_2 - p - 1\right]$$

$$P = \left(1.4643175561 \times 10^{-11}\right)$$

Discriminant Coefficient Vector:

$$S_{pooled}^{-1}\cdot d = \begin{pmatrix} 0.5104227 \\ -0.2032933 \\ 0.466042 \\ -0.3096697 \end{pmatrix}$$

Coeffieints proportional to vector direction most responsible for rejection of H_o. See JW Remark p. 288

The Multivariate Confidence Ellipsoid:

$i := 1 .. p$

$\lambda := reverse(sort(eigenvals(S_{pooled})))$

$\varepsilon^{\langle i \rangle} := eigenvec(S_{pooled}, \lambda_i)$

$$\lambda = \begin{pmatrix} 46.12466 \\ 13.35619 \\ 11.34474 \\ 3.90979 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} 0.2268 & -0.3189 & -0.1282 & -0.9113 \\ 0.3545 & -0.8275 & -0.1684 & 0.4015 \\ 0.7184 & 0.1813 & 0.6713 & 0.021 \\ 0.5539 & 0.425 & -0.7103 & 0.0891 \end{pmatrix}$$

< Coordinates of each column vector of ε gives the directions of confidence elipsoid

$$L_i := \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\cdot C\cdot\lambda_i}$$

Multivariate confidence ellipsoid (JW Eq. 5-18 p. 221):

$$d = \begin{pmatrix} 3.625 \\ 2 \\ 10.531 \\ 0.813 \end{pmatrix}$$

< Center of ellipsoid

$$L = \begin{pmatrix} 5.534609 \\ 2.978254 \\ 2.744843 \\ 1.611377 \end{pmatrix}$$

< L are half-lengths of the axes of the confidence ellipsoid for m in the directions of ε

Simultaneous T^2 Confidence Intervals:

$$CI_{lower_i} := d_i - \sqrt{C}\cdot\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot\left(S_{pooled}\right)_{i,i}} \qquad CI_{upper_i} := d_i + \sqrt{C}\cdot\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot S_{pooled_{i,i}}}$$

$$CI := augment\left(CI_{lower}, CI_{upper}\right)$$

Simultaneous Confidence Intervals:

$$d = \begin{pmatrix} 3.625 \\ 2 \\ 10.53125 \\ 0.8125 \end{pmatrix}$$ **< Mean values**

$$CI = \begin{pmatrix} 1.44374 & 5.80626 \\ -1.24892 & 5.24892 \\ 6.11584 & 14.94666 \\ -3.03761 & 4.66261 \end{pmatrix}$$ **< T^2 confidence intervals See JW Result 6.3 p. 287**

Univariate t Confidence Intervals:

$$ct := qt\left(1-\frac{\alpha}{2}, n_1+n_2-2\right) \qquad ct = 1.998972$$

< Critical value ct based on t distribution without correction

$$ci_{lower_i} := d_i - ct\cdot\sqrt{\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot S_{pooled}\right]_{i,i}} \qquad ci_{upper_i} := d_i + ct\cdot\sqrt{\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot S_{pooled}\right]_{i,i}}$$

$$ci := augment\left(ci_{lower}, ci_{upper}\right)$$

Univariate t Intervals:

$$d = \begin{pmatrix} 3.625 \\ 2 \\ 10.53125 \\ 0.8125 \end{pmatrix}$$ **< Mean values**

$$ci = \begin{pmatrix} 2.28738 & 4.96262 \\ 0.00765 & 3.99235 \\ 7.82357 & 13.23893 \\ -1.54852 & 3.17352 \end{pmatrix}$$ **< Univariate t confidence intervals**

Bonferroni Simultaneous Confidence Intervals:

$$cb := qt\left(1-\frac{\alpha}{2\cdot p}, n_1+n_2-2\right) \qquad cb = 2.572664$$

< Critical value cb based on t distribution with Bonferroni correction factor p. See JW p 290.

$$ci_{lower_i} := d_i - cb\cdot\sqrt{\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot S_{pooled}\right]_{i,i}} \qquad ci_{upper_i} := d_i + cb\cdot\sqrt{\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot S_{pooled}\right]_{i,i}}$$

$$ci := augment\left(ci_{lower}, ci_{upper}\right)$$

Bonferroni Intervals:

$$d = \begin{pmatrix} 3.625 \\ 2 \\ 10.53125 \\ 0.8125 \end{pmatrix}$$ **< Mean values**

$$ci = \begin{pmatrix} 1.903487 & 5.346513 \\ -0.56414 & 4.56414 \\ 7.04648 & 14.01602 \\ -2.226115 & 3.851115 \end{pmatrix}$$ **< Bonferroni confidence intervals**

Prototype in R:

#LOAD DATA:
M=read.table("c:/DATA/Multivariate/AR-T5-1-longform.txt")
M #DATA TABLE
X=M[(M$sex=='1'),2:5]
Y=M[(M$sex=='2'),2:5]
X
Y

#FUNCTION FOR HOTELLING'S T2 TEST
FOR TWO POPULATIONS EQUAL WITH VARIANCE:
#X = dataset 1
#Y = dataset 2
#d0 = hypothesis vector
#alpha = alpha of the test

TwoPop.Hotelling.T2 <- function(X,Y,d0,alpha)
...function body in R script ...

RES=TwoPop.Hotelling.T2(X,Y,d0=c(0,0,0,0),alpha=0.05)

RES$Cov.Pooled

> M #DATA TABLE

	sex	y1	y2	y3	y4
1	1	15	17	24	14
2	1	17	15	32	26
3	1	15	14	29	23
4	1	13	12	10	16
5	1	20	17	26	28
6	1	15	21	26	21
7	1	15	13	26	22
8	1	13	5	22	22
9	1	14	7	30	17
10	1	17	15	30	27
11	1	17	17	26	20
12	1	17	20	28	24
13	1	15	15	29	24
14	1	18	19	32	28
15	1	18	18	31	27
16	1	15	14	26	21
17	1	18	17	33	26
18	1	10	14	19	17
19	1	18	21	30	29
20	1	18	21	34	26
21	1	13	17	30	24
22	1	16	16	16	16
23	1	11	15	25	23
24	1	16	13	26	16
25	1	16	13	23	21
26	1	18	18	34	24
27	1	16	15	28	27
28	1	15	16	29	24
29	1	18	19	32	23
30	1	18	16	33	23
31	1	17	20	21	21
32	1	19	19	30	28
33	2	13	14	12	21
34	2	14	12	14	26
35	2	12	19	21	21
36	2	12	13	10	16
37	2	11	20	16	16
38	2	12	9	14	18
39	2	10	13	18	24
40	2	10	8	13	23
41	2	12	20	19	23
42	2	11	10	11	27
43	2	12	18	25	25
44	2	14	18	13	26
45	2	14	10	25	28
46	2	13	16	8	14
47	2	14	8	13	25
48	2	13	16	23	28
49	2	16	21	26	26
50	2	14	17	14	14
51	2	16	16	15	23
52	2	13	16	23	24
53	2	2	6	16	21
54	2	14	16	22	26
55	2	17	17	22	28
56	2	16	13	16	14
57	2	15	14	20	26
58	2	12	10	12	9
59	2	14	17	24	23
60	2	13	15	18	20
61	2	11	16	18	28
62	2	7	7	19	18
63	2	12	15	7	28
64	2	6	5	6	13

> X

	y1	y2	y3	y4
1	15	17	24	14
2	17	15	32	26
3	15	14	29	23
4	13	12	10	16
5	20	17	26	28
6	15	21	26	21
7	15	13	26	22
8	13	5	22	22
9	14	7	30	17
10	17	15	30	27
11	17	17	26	20
12	17	20	28	24
13	15	15	29	24
14	18	19	32	28
15	18	18	31	27
16	15	14	26	21
17	18	17	33	26
18	10	14	19	17
19	18	21	30	29
20	18	21	34	26
21	13	17	30	24
22	16	16	16	16
23	11	15	25	23
24	16	13	26	16
25	16	13	23	21
26	18	18	34	24
27	16	15	28	27
28	15	16	29	24
29	18	19	32	23
30	18	16	33	23
31	17	20	21	21
32	19	19	30	28

> Y

	y1	y2	y3	y4
33	13	14	12	21
34	14	12	14	26
35	12	19	21	21
36	12	13	10	16
37	11	20	16	16
38	12	9	14	18
39	10	13	18	24
40	10	8	13	23
41	12	20	19	23
42	11	10	11	27
43	12	18	25	25
44	14	18	13	26
45	14	10	25	28
46	13	16	8	14
47	14	8	13	25
48	13	16	23	28
49	16	21	26	26
50	14	17	14	14
51	16	16	15	23
52	13	16	23	24
53	2	6	16	21
54	14	16	22	26
55	17	17	22	28
56	16	13	16	14
57	15	14	20	26
58	12	10	12	9
59	14	17	24	23
60	13	15	18	20
61	11	16	18	28
62	7	7	19	18
63	12	15	7	28
64	6	5	6	13

> RES=TwoPop.Hotelling.T2(X,Y,d0=c(0,0,0,0),alpha=0.05)

```
 Two Population Hotelling's T2 with Equal Covariance Matrices
                    for small sample sizes

 Hypothesis Vector:          ( 0 0 0 0 )
 Difference vector X-Y:      ( 3.625 2 10.53125 0.8125 )
 Discriminant vector  :      ( 0.5104227 -0.2032933 0.466042 -0.3096697 )
 T2 Ellipsoid half lengths:  ( 5.534609 2.978254 2.744843 1.611377 )
 Hotelling's T2 Statistic:     97.6015
 Equivalent F Statistic :      23.21971
 F degrees of freedom:       ( 4 59 )
 Wilks's Lambda:               0.3884675
 alpha:                        0.05
 Critical Value:               10.62578
 Probability:                  1.464318e-11

 Confidence Intervals: T2 - Bonferroni - t - Mean - t - Bonferroni - T2

    T2.lower   B.lower      t.lower     Mean   t.upper   B.upper  T2.upper
y1  1.443739  1.903487  2.287376358  3.62500  4.962624  5.346513  5.806261
y2 -1.248920 -0.564140  0.007651476  2.00000  3.992349  4.564140  5.248920
y3  6.115836  7.046480  7.823568055 10.53125 13.238932 14.016020 14.946664
y4 -3.037608 -2.226115 -1.548517477  0.81250  3.173517  3.851115  4.662608
```

> RES$Cov.Pooled

```
         y1        y2        y3        y4
y1 7.164315  6.047379  5.693044  4.700605
y2 6.047379 15.894153  8.492440  5.855847
y3 5.693044  8.492440 29.356351 13.980847
y4 4.700605  5.855847 13.980847 22.320565
```

< pooled covariance matrix

```
library(ICSNP)
HotellingsT2(X,Y,mu=c(0,0,0,0),test="f")
```

```
> HotellingsT2(X,Y,mu=c(0,0,0,0),test="f")

        Hotelling's two sample T2-test

data:  X and Y
T.2 = 23.2197, df1 = 4, df2 = 59, p-value = 1.464e-11
alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

```
library(rrcov)
T2.test(X,Y,mu=c(0,0,0,0),conf.level=0.95,method="c")
```

```
> T2.test(X,Y,mu=c(0,0,0,0),conf.level=0.95,method="c")

        Two-sample Hotelling test

data:  X and Y
T^2 = 23.2197, df1 = 4, df2 = 59, p-value = 1.464e-11
alternative hypothesis: true difference in mean vectors is not equal to
(0,0,0,0)
sample estimates:
                     y1       y2       y3      y4
mean x-vector 15.96875 15.90625 27.18750 22.7500
mean y-vector 12.34375 13.90625 16.65625 21.9375
```

```
attach(M)
YY=cbind(y1,y2,y3,y4)
detach(M)
```

```
FIT=manova(YY~sex,data=M)
summary(FIT,test='Wilks')
```

```
> summary(FIT,test='Wilks')
          Df   Wilks approx F num Df den Df    Pr(>F)
sex        1 0.38847    23.22      4     59 1.464e-11 ***
Residuals 62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```