ORIGIN ≡ 1

Hotelling's T^2 Test for Two Populations with Large Samples

As described by AR p. 136, Hotelling's $T^2_{p,\nu}$ distribution has two parameters: p = number of variables (columns in the dataset) and ν which adjusts for sample size (rows). For one and two sample Hotelling's tests, ν degrees of freedom are the same as the corresponding t-tests; $\nu=(n-1)$ for one sample, and $\nu=(n_1+n_2-2)$ for two samples. If parameter p=1 then the values of T^2 equal the square of the t-distribution ($T^2= t^2$), and Hotelling's T^2 statistic reduces to the square of Student's t-statistic. As sample size increases (i.e., as ν increases towards infinity) Hotelling's $T^2_{p,\nu}$ distribution approaches the χ^2_ν, distibution, and sample covariance matrix S (or S_{pooled}) approaches population covariance matrix Σ. With increasing numbers of variables p, T^2 approaches χ^2 at a slower rate - meaning that as one increases variables in a multivariate study, correspondingly greater sample size is required for the "large" sample approximation. AR gives an approximate scaling of "sufficient" sample size:

p=1 n=30
p=5 n=100
p=10 n=200

Clearly, these are only ballpark figures, with user discretion strongly advised. For a particular study, one can always run Hotelling's T^2 both ways, and compare results. The example below, showing calculations for the "large" sample χ^2 approximation, is drawn from Example 6.15 in RA Johnson & DW Wichern (JW) *Applied Multivariate Statistical Analysis 4th Edition* 1998. For an unknown reason, results here approximate, but do not exactly match, reported for confidence intervals in JW, although formulas for CI's in Example 6.4 (for the "small" sample case) do. Also acknowledged is the very helpful commentary in AC. Rencher (AR) *Methods of Multivariate Analysis* 1995.

Read Data:

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

M := READPRN("c:\DATA\Multivariate\JW-T6-7-longform.txt")

X := submatrix(M, 1, 20, 2, 3)

Y := submatrix(M, 21, 60, 2, 3)

$n_1 := \text{rows}(X)$ $n_1 = 20$ $n_2 := \text{rows}(Y)$ $n_2 = 40$

$p := \text{cols}(X)$ $p = 2$

$X := \overrightarrow{\ln(X)}$

$Y := \overrightarrow{\ln(Y)}$

M =

	1	2	3	4
1	1	7.513	74	0
2	2	5.032	69.5	0
3	3	5.867	72	0
4	4	11.088	80	0
5	5	2.419	56	0
6	6	13.61	94	0
7	7	18.247	95.5	0
8	8	16.832	99.5	0
9	9	15.91	97	0
10	10	17.035	90.5	0
11	11	16.526	91	0
12	12	4.53	67	0
13	13	7.23	75	0
14	14	5.2	69.5	0
15	15	13.45	91.5	0
16	16	14.08	91	0
17	17	14.665	90	0

Summary Statistics:

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

$I_1 := \text{identity}(n_1)$ $I_2 := \text{identity}(n_2)$ $l_{n1_i} := 1$ $l_{n2_{ii}} := 1$

Mean Vectors:

$X_{bar} := \frac{1}{n_1} \cdot X^T \cdot l_{n1}$ $X_{bar} = \begin{pmatrix} 2.24 \\ 4.394 \end{pmatrix}$

$Y_{bar} := \frac{1}{n_2} \cdot Y^T \cdot l_{n2}$ $Y_{bar} = \begin{pmatrix} 2.368 \\ 4.308 \end{pmatrix}$

Covariance Matrices:

$$S_1 := \frac{1}{n_1 - 1} \cdot X^T \cdot \left(I_1 - \frac{1}{n_1} \cdot l_{n1} \cdot l_{n1}^T \right) \cdot X \qquad S_1 = \begin{pmatrix} 0.35304883 & 0.09416815 \\ 0.09416815 & 0.02595325 \end{pmatrix}$$

$$S_2 := \frac{1}{n_2 - 1} \cdot Y^T \cdot \left(I_2 - \frac{1}{n_2} \cdot l_{n2} \cdot l_{n2}^T \right) \cdot Y \qquad S_2 = \begin{pmatrix} 0.5068444 & 0.14539213 \\ 0.14539213 & 0.04255351 \end{pmatrix}$$

Observed Difference Vector:

$$d := X_{bar} - Y_{bar} \qquad d = \begin{pmatrix} -0.128 \\ 0.086 \end{pmatrix}$$

Pooled Covariance Matrix:

$$S_{pooled} := \frac{1}{n_1} \cdot S_1 + \frac{1}{n_2} \cdot S_2 \qquad S_{pooled} = \begin{pmatrix} 0.03032355 & 0.00834321 \\ 0.00834321 & 0.0023615 \end{pmatrix}$$

Assumptions:

X independent of Y

No other assumptions required; Underlying population distributions need not be Normally Distributed. Covariance matrices Σ_1 and Σ_2 need not be equivalent.

Hypotheses:

Specify a test vector μ_0:

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

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:

$$T_{sq} := (d - \delta_o)^T \cdot (S_{pooled})^{-1} \cdot (d - \delta_o) \qquad T_{sq} = (224.79604)$$

Wilk's Lambda:

$$\Lambda := \frac{1}{\frac{T_{sq}}{(n_1 + n_2 - 2)} + 1} \qquad \Lambda = (0.2050948)$$

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**

$C := \text{qchisq}(1-\alpha, p)$ $C = 5.991465$ < **C gives the boundary for the confidence ellipsoid. See JW Result 6.4 p. 291**

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

Decision Rule:

Reject H_0 if $T_{sq} > C$ $T_{sq} = (224.796)$ $C = 5.991465$

$\text{Decision} := \text{if}\left(T_{sq_1} > C, 1, 0\right)$ $\text{Decision} = 1$ < **0 = Do not reject H_0**
1 = Reject H_0

^ **Therefore REJECT H_0**

Probability Value:

$P := 1 - \text{pchisq}\left(T_{sq}, p\right)$ $P = (0)$

Discriminant Coefficient Vector:

$$S_{pooled}^{-1} \cdot d = \begin{pmatrix} -511.5796304 \\ 1843.9770584 \end{pmatrix}$$

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

The Multivariate Confidence Ellipsoid:

$i := 1..p$

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

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

$$\lambda = \begin{pmatrix} 0.03262 \\ 0.00006 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} 0.964033 & -0.265781 \\ 0.265781 & 0.964033 \end{pmatrix}$$

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

$L_i := \sqrt{C \cdot \lambda_i}$

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

$$d = \begin{pmatrix} -0.128 \\ 0.086 \end{pmatrix}$$

< **Center of ellipsoid**

$$L = \begin{pmatrix} 0.442113 \\ 0.019165 \end{pmatrix}$$

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

Simultaneous χ^2 Confidence Intervals:

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

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

Simultaneous Confidence Intervals:

$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix}$ **< Mean values** $\qquad$ $CI = \begin{pmatrix} -0.55446 & 0.29802 \\ -0.03261 & 0.20528 \end{pmatrix}$

< χ^2 confidence intervals in JW p. 329 are close, but do not exactly match. However, method here follows calculations in JW Example 6.5 p. 292 & verified exactly.

Univariate t Confidence Intervals:

$ct := qt\left(1 - \frac{\alpha}{2}, n_1 + n_2 - 2\right)$ $\qquad$ $ct = 2.001717$

< Critical value ct based on t distribution without correction

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

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

Univariate t Intervals:

$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix}$ **< Mean values** $\qquad$ $ci = \begin{pmatrix} -0.47679 & 0.22035 \\ -0.01094 & 0.18361 \end{pmatrix}$ **< Univariate t confidence intervals**

Bonferroni Simultaneous Confidence Intervals:

$cb := qt\left(1 - \frac{\alpha}{2\cdot p}, n_1 + n_2 - 2\right)$ $\qquad$ $cb = 2.301084$

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

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

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

Bonferroni Intervals:

$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix}$ **< Mean values** $\qquad$ $ci = \begin{pmatrix} -0.528925 & 0.272481 \\ -0.025486 & 0.198157 \end{pmatrix}$ **< Bonferroni confidence intervals**

Prototype in R:

```
#HOTELLING'S T2 TEST FOR TWO POPULATIONS
#WITH LARGE SAMPLE SIZE

#LOAD DATA:
M=read.table("c:/DATA/Multivariate/JW-T6-7-longform.txt")
M   #DATA TABLE

X=M[(M$Genus=='0'),1:2]
Y=M[(M$Genus=='1'),1:2]
X
Y
X=log(X)
Y=log(Y)


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

TwoPopLS.Hotelling.T2 <- function(X,Y,d0,alpha)
... body of function in R script ...
RES=TwoPopLS.Hotelling.T2(X,Y,d0=c(0,0),alpha=0.05)
```

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

 Two Population Hotelling's T2 for Large Sample Sizes
 Hypothesis Vector:            ( 0 0 )
 Difference vector X-Y:        ( -0.1282218 0.08633533 )
 Discriminant vector  :        ( -511.5796 1843.977 )
 Chisq Ellipsoid half lengths: ( 0.4421132 0.01916479 )
 Hotelling's T2 Statistic:       224.796
 Wilks's Lambda:                 0.2050948
 alpha:                          0.05
 Critical Value:      d           5.991465
 Probability:                   0

 Confidence Intervals: Chi - Bonferroni - t - Mean - t - Bonferroni - Chi

        Chi.lower     B.lower      t.lower         Mean    t.upper    B.upper Chi.upper
Mass -0.55446412 -0.52892469 -0.47679409 -0.12822184 0.2203504 0.2724810 0.2980204
SVL  -0.03261359 -0.02548645 -0.01093867  0.08633533 0.1836093 0.1981571 0.2052842
```

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

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

#USING MANOVA:
 attach(M)
YY=cbind(log(Mass),log(SVL))
detach(M)
FIT=manova(YY~Genus,data=M)
summary(FIT,test='Wilks')
```

```
> M   #DATA TABLE
     Mass   SVL Genus
1   7.513  74.0     0
2   5.032  69.5     0
3   5.867  72.0     0
4  11.088  80.0     0
5   2.419  56.0     0
6  13.610  94.0     0
7  18.247  95.5     0
8  16.832  99.5     0
9  15.910  97.0     0
10 17.035  90.5     0
11 16.526  91.0     0
12  4.530  67.0     0
13  7.230  75.0     0
14  5.200  69.5     0
15 13.450  91.5     0
16 14.080  91.0     0
17 14.665  90.0     0
18  6.092  73.0     0
19  5.264  69.5     0
20 16.902  94.0     0
21 13.911  77.0     1
22  5.236  62.0     1
23 37.331 108.0     1
24 41.781 115.0     1
25 31.995 106.0     1
26  3.962  56.0     1
27  4.367  60.5     1
28  3.048  52.0     1
29  4.838  60.0     1
30  6.525  64.0     1
31 22.610  96.0     1
32 13.342  79.5     1
33  4.109  55.5     1
34 12.369  75.0     1
35  7.120  64.5     1
36 21.077  87.5     1
37 42.989 109.0     1
38 27.201  96.0     1
39 38.901 111.0     1
40 19.747  84.5     1
41 14.666  80.0     1
42  4.790  62.0     1
43  5.020  61.5     1
44  5.220  62.0     1
45  5.690  64.0     1
46  6.763  63.0     1
47  9.977  71.0     1
48  8.831  69.5     1
49  9.493  67.5     1
50  7.811  66.0     1
51  6.685  64.5     1
52 11.980  79.0     1
53 16.520  84.0     1
54 13.630  81.0     1
55 13.700  82.5     1
56 10.350  74.0     1
57  7.900  68.5     1
58  9.103  70.0     1
59 13.216  77.5     1
60  9.787  70.0     1
```

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

```
        Hotelling's two sample T2-test

data:  X and Y
T.2 = 228.1898, df = 2, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0,0)
```

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

```
        Two-sample Hotelling test

data:  X and Y
T^2 = 112.1278, df1 = 2, df2 = 57, p-value < 2.2e-16
alternative hypothesis: true difference in mean vectors is not equal to (0,0)
sample estimates:
                  Mass      SVL
mean x-vector 2.239919 4.394427
mean y-vector 2.368140 4.308091
```

> FIT=manova(YY~Genus,data=M)
> summary(FIT,test='Wilks')

```
          Df   Wilks approx F num Df den Df    Pr(>F)    
Genus      1 0.20266   112.13      2     57 < 2.2e-16 ***
Residuals 58                                             
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#RE-RUNNING THE PROBLEM AS T2 DISTRIBUTION FOR SMALL
#SAMPLE SIZES:
#USING SCRIPT FROM MHT 030
RES=TwoPop.Hotelling.T2(X,Y,d0=c(0,0),alpha=0.05)

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

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

 Hypothesis Vector:          ( 0 0 )
 Difference vector X-Y:      ( -0.1282218 0.08633533 )
 Discriminant vector  :      ( -39.57378 139.4564 )
 T2 Ellipsoid half lengths: ( 0.4874232 0.01980395 )
 Hotelling's T2 Statistic:    228.1898
 Equivalent F Statistic :     112.1278
 F degrees of freedom:       ( 2 57 )
 Wilks's Lambda:              0.2026627
 alpha:                       0.05
 Critical Value:              6.428522
 Probability:                 0

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

        T2.lower     B.lower     t.lower        Mean   t.upper   B.upper  T2.upper
Mass -0.59734733 -0.55398282 -0.49859222 -0.12822184 0.2421485 0.2975392 0.3409037
SVL  -0.04743624 -0.03507081 -0.01927613  0.08633533 0.1919468 0.2077415 0.2201069
```

Difference in approaches appear minimal for this dataset. Online R procedures found so far, seem to reflect only the two-sample approach for small sample sizes.