Hotelling's T² Test for Two Populations with Large Samples

As described by AR p. 136, Hotelling's T²,Φ,v distribution has two parameters: p = number of variables (columns in the dataset) and v which adjusts for sample size (rows). For one and two sample Hotelling's tests, v degrees of freedom are the same as the corresponding t-tests; v=(n-1) for one sample, and v=(n₁+n₂-2) for two samples. If parameter p=1 then the values of T² equal the square of the t-distribution (T²=t²), and Hotelling's T² statistic reduces to the square of Student's t-statistic. As sample size increases (i.e., as v increases towards infinity) Hotelling's T²,Φ,v distribution approaches the χ²,Φ,v distribution, and sample covariance matrix S (or S_pooled) approaches population covariance matrix Σ. With increasing numbers of variables p, T² approaches χ² 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:

\[
\begin{array}{c|c|c}
 p & n & \\
\hline
 1 & 30 & \\
 5 & 100 & \\
 10 & 200 & \\
\end{array}
\]

Clearly, these are only ballpark figures, with user discretion strongly advised. For a particular study, one can always run Hotelling's T² both ways, and compare results. The example below, showing calculations for the "large" sample χ² 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:

\[
M := \text{READPRN}("c:\DATA\Multivariate\JW-T6-7-longform.txt")
\]

\[
X := \text{submatrix}(M, 1, 20, 2, 3)
\]

\[
Y := \text{submatrix}(M, 21, 60, 2, 3)
\]

\[
n_1 := \text{rows}(X) \quad n_1 = 20 \quad n_2 := \text{rows}(Y) \quad n_2 = 40
\]

\[
p := \text{cols}(X) \quad p = 2
\]

\[
X := \ln(X)
\]

\[
Y := \ln(Y)
\]

Summary Statistics:

\[
i := 1..n_1 \quad ii := 1..n_2 \quad j := 1..p
\]

\[
l_1 := \text{identity}(n_1) \quad l_2 := \text{identity}(n_2) \quad l_{n_1} := 1 \quad l_{n_2} := 1
\]

Mean Vectors:

\[
X_{\text{bar}} := \frac{1}{n_1} \cdot X^T \cdot l_{n_1} \quad X_{\text{bar}} = \begin{pmatrix} 2.24 \\ 4.394 \end{pmatrix}
\]

\[
Y_{\text{bar}} := \frac{1}{n_2} \cdot Y^T \cdot l_{n_2} \quad Y_{\text{bar}} = \begin{pmatrix} 2.368 \\ 4.308 \end{pmatrix}
\]

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

\[ S_1 := \frac{1}{n_1 - 1} \cdot X^T \left( I_1 - \frac{1}{n_1} \cdot I_{n_1} \right) X \quad S_1 = \begin{pmatrix} 0.35304883 & 0.09416815 \\ 0.09416815 & 0.02595325 \end{pmatrix} \]

\[ S_2 := \frac{1}{n_2 - 1} \cdot Y^T \left( I_2 - \frac{1}{n_2} \cdot I_{n_2} \right) Y \quad S_2 = \begin{pmatrix} 0.5068444 & 0.14539213 \\ 0.14539213 & 0.04255351 \end{pmatrix} \]

Observed Difference Vector:

\[ d := X_{\bar{\text{bar}}} - Y_{\bar{\text{bar}}} \quad d = \begin{pmatrix} -0.128 \\ 0.086 \end{pmatrix} \]

Pooled Covariance Matrix:

\[ S_{\text{pooled}} := \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \quad S_{\text{pooled}} = \begin{pmatrix} 0.03032355 & 0.00834321 \\ 0.00834321 & 0.0023615 \end{pmatrix} \]

Assumptions:

No other assumptions required; Underlying population distributions need not be Normally Distributed. Covariance matrices \( \Sigma_1 \) and \( \Sigma_2 \) need not be equivalent.

Hypotheses:

Specify a test vector \( \mu_0 \):

\[ \delta_0 := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \]

\( H_0: \delta_0 = \mu_1 - \mu_2 = \delta_0 \) \quad < \delta_0 (0,0) or any value of difference between X & Y is a specified value for \( \delta \)

\( H_1: \delta_0 = \mu_1 - \mu_2 \neq \delta_0 \) \quad < 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:

\[ T_{sq} := (d - \delta_0)^T \left( S_{\text{pooled}} \right)^{-1} (d - \delta_0) \quad T_{sq} = (224.79604) \]

Wilk's Lambda:

\[ \Lambda := \frac{1}{T_{sq} + \frac{1}{n_1 + n_2 - 2}} \quad \Lambda = (0.2050948) \quad \text{See AR Eq. 5-17 p. 147} \]

Sampling Distribution:

If Assumptions hold and \( H_0 \) is true, then \( T_{sq} \sim T^2_{(n-1)} = \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{(p,n_1+n_2-p-1)} \)
Critical Value of the Test:

\[ \alpha := 0.05 \quad < \text{Probability of Type I error must be explicitly set} \]

\[ C := \text{qchisq}(1 - \alpha, p) \quad C = 5.991465 \quad < C \text{ gives the boundary for the confidence ellipsoid.} \]

^ NOTE: qchisq(1-\(\alpha\)) is used in function.

Decision Rule:

Reject \( H_0 \) if \( T_{sq} > C \)

\[ T_{sq} = (224.796) \quad C = 5.991465 \]

Decision := if \( T_{sq} > C, 1, 0 \)

Decision = 1 < 0 = Do not reject \( H_0 \)

1 = Reject \( H_0 \)

^ Therefore REJECT \( H_0 \)

Probability Value:

\[ P := 1 - \text{pchisq}(T_{sq}, p) \quad P = (0) \]

Discriminant Coefficient Vector:

\[ S_{\text{pooled}} - 1 \cdot d = \begin{pmatrix} -511.5796304 \\ 1843.9770584 \end{pmatrix} \quad \text{Coefficients proportional to vector direction most responsible for rejection of } H_0. \quad \text{See JW Remark p. 288} \]

The Multivariate Confidence Ellipsoid:

\[ i := 1 .. p \]

\[ \lambda := \text{reverse}\left( \text{sort}\left( \text{eigvals}\left( S_{\text{pooled}} \right) \right) \right) \]

\[ \epsilon^{(i)} := \text{eigenvc}\left( S_{\text{pooled}}, \lambda_i \right) \]

\[ \lambda = \begin{pmatrix} 0.03262 \\ 0.00006 \end{pmatrix} \]

\[ \epsilon = \begin{pmatrix} 0.964033 & -0.265781 \\ 0.265781 & 0.964033 \end{pmatrix} \]

< Coordinates of each column vector of \( \epsilon \) gives the directions of confidence ellipsoid

\[ L_i := \sqrt{C \cdot \lambda_i} \]

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

\[ d = \begin{pmatrix} -0.128 \\ 0.086 \end{pmatrix} \quad < \text{Center of ellipsoid} \]

\[ L = \begin{pmatrix} 0.442113 \\ 0.019165 \end{pmatrix} \]

< \( L \) are half-lengths of the axes of the confidence ellipsoid for \( m \) in the directions of \( \epsilon \)
Simultaneous $\chi^2$ Confidence Intervals:

$$\text{CI}_{\text{lower}} := d_i - \sqrt{C_i} \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$$\text{CI}_{\text{upper}} := d_i + \sqrt{C_i} \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$\text{CI} := \text{augment}(\text{CI}_{\text{lower}}, \text{CI}_{\text{upper}})$

**Simultaneous Confidence Intervals:**

$$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix} < \text{Mean values}$$

$$\text{CI} = \begin{pmatrix} -0.55446 & 0.29802 \\ -0.03261 & 0.20528 \end{pmatrix}$$

$< \chi^2$ 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:**

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

$$c_i_{\text{lower}} := d_i - c_t \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$$c_i_{\text{upper}} := d_i + c_t \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$ci := \text{augment}(c_i_{\text{lower}}, c_i_{\text{upper}})$

**Univariate t Intervals:**

$$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix} < \text{Mean values}$$

$$ci = \begin{pmatrix} -0.47679 & 0.22035 \\ -0.01094 & 0.18361 \end{pmatrix}$$

$< \text{Univariate t confidence intervals}$

**Bonferroni Simultaneous Confidence Intervals:**

$$c_b := \text{qt}\left(1 - \frac{\alpha}{2p}, n_1 + n_2 - 2\right)$$

$$c_i_{\text{lower}} := d_i - c_b \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$$c_i_{\text{upper}} := d_i + c_b \sqrt{\left(\frac{S_{\text{pooled}}}{i_i}\right)_{i,i}}$$

$ci := \text{augment}(c_i_{\text{lower}}, c_i_{\text{upper}})$

**Bonferroni Intervals:**

$$d = \begin{pmatrix} -0.12822 \\ 0.08634 \end{pmatrix} < \text{Mean values}$$

$$ci = \begin{pmatrix} -0.528925 & 0.272481 \\ -0.025486 & 0.198157 \end{pmatrix}$$

$< \text{Bonferroni confidence intervals}$
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

 # ICSNP
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")
```
> 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.