ORIGIN ≡ 1

Hotelling's T^2 Test for a Single Population

Hotelling's T^2 tests are multivariate analogs to univariate Student's t-tests. As such, they have similar formal structure regarding null H_0 and alternative hypotheses H_1, rejection (or not) of the null by consulting a critical value CV or probability P based upon the sampling distribution of the test statistic assuming H_0 to be true, and construction of confidence intervals for location of the mean that essentially restate the rejection criterion. Student's t test statistic for a single population:

$$t = \frac{(X_{bar} - \mu)}{\frac{\sigma}{\sqrt{n}}} \quad \text{or alternatively in squared form:} \quad t^2 = n \cdot \frac{(X_{bar} - \mu)^2}{\sigma^2}$$

measures "statistical distance", i.e., the difference (X_{bar} - μ) corrected for population standard deviation ($\sigma/\sqrt{n}$) or variance $(\sigma/\sqrt{n})^2$ *of the mean*, estimated by sample standard deviation ($s/\sqrt{n}$) or variance $(s/\sqrt{n})^2$ of the mean respectively. There is only one value of X for each object (= case) in a dataset.

Hotelling's T^2 adopts the squared form of statistical distance above and extends the problem to cases where each object has p observations (X_1, X_2, X_3 ... X_p). Here population and sample means become mean vectors (μ_1,μ_2,μ_3 ...μ_p) and ($X1_{bar}$,$X2_{bar}$,$X3_{bar}$ ... Xp_{bar}), and σ^2 becomes the population covariance matrix Σ estimated by sample covariance matrix S. In strict analogy to Student's t (squared), the test statistic now becomes:

$$T^2 = n \cdot (X_{bar} - \mu)^T \cdot \Sigma^{-1} \cdot (X_{bar} - \mu) \quad \text{or:} \quad n \cdot \Delta^2$$

where Δ^2 is the squared Mahalanobis distance, and the factor (n) accounts for distance between *mean* vectors. Although numerical values for the T^2 distribution can be found in tables, in actual practice one rarely uses them because T^2 can be converted to values of more widely used F distribution, shown below, or with large sample size values of the χ^2 distribution.

Interpretation of multivariate results such as Hotelling T^2 is typically more involved than that of its univariate analog, and is further complicated by variant approaches and terminology. Notable among these is the use of Likelihood Ratio test statistics such as Wilks lambda shown below and reported by most professional software. Thus greater sophistication is required, but interpetation is greatly facilitated by intuition gained from the geometric perspective of points or vectors in multidimensional space. Shown here is the one population/sample situation that tests whether the mean vector μ from a population (of which we have a sample with mean vector X_{bar}) is sufficiently close in the space of p-dimensions to some specified vector μ_0. The example is drawn from RA Johnson & DW Wichern (JW) *Applied Multivariate Statistical Analysis 4th Edition* 1998. Useful discussion may also be found in AC. Rencher (AR) *Methods of Multivariate Analysis* 1995.

Read Data:

Perspiration in n = 20 healthy females (JW p. 214):
Columns:

X_1 = sweat rate
X_2 = Na content
X_3 = K content

X := READPRN("c:\DATA\Multivariate\T5-1.DAT")

n := rows(X)   p := cols(X)

Summary Statistics:

$i := 1..n \qquad j := 1..p$

$l_{n_i} := 1 \qquad I := \text{identity}(n)$

Mean Vector:

$$X_{bar} := \frac{1}{n} \cdot X^T \cdot l_n \qquad X_{bar} = \begin{pmatrix} 4.64 \\ 45.4 \\ 9.965 \end{pmatrix}$$

Covariance Matrix:

$$S := \frac{1}{n-1} \cdot X^T \cdot \left(I - \frac{1}{n} \cdot l_n \cdot l_n^T \right) \cdot X \qquad S = \begin{pmatrix} 2.87937 & 10.01 & -1.80905 \\ 10.01 & 199.78842 & -5.64 \\ -1.80905 & -5.64 & 3.62766 \end{pmatrix}$$

X =

	1	2	3
1	3.7	48.5	9.3
2	5.7	65.1	8
3	3.8	47.2	10.9
4	3.2	53.2	12
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Assumptions:

- Observed values $X_1, X_2, X_3, \ldots X_n$ are a random sample from $\sim N_p(\mu,\Sigma)$.

- Covaraince Matrix Σ of the population is *unknown* and estimated by S.

Hypotheses:

^ Note: this test is reasonably robust for deviations from $\sim N_p(\mu,\Sigma)$ with sufficient sample size

Specify a test vector μ_0:

$$\mu_o := \begin{pmatrix} 4 \\ 50 \\ 10 \end{pmatrix}$$

H_0: $\mu = \mu_0$ **< μ_0 is a specified value for μ**

H_1: $\mu \neq \mu_0$ **< 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:

$$T_{sq} := n \cdot (X_{bar} - \mu_o)^T \cdot S^{-1} \cdot (X_{bar} - \mu_o) \qquad T_{sq} = (9.738773)$$

^ T^2 is the normalized distance between vectors X_{bar} and μ_0

Likelihood Ratio Test Statistic:

Maximum likelihood estimates of μ & Σ (JW Result 4.11 p. 171) and Σ_0:

$$\mu_{hat} := X_{bar}$$

$$\Sigma_{hat} := \frac{(n-1)}{n} \cdot S \qquad \left|\Sigma_{hat}\right| = 1.014 \times 10^3$$

$$\Sigma_{0.hat} := \sum_i \left[\left(X^T\right)^{\langle i \rangle} - \mu_o \right] \cdot \left[\left(X^T\right)^{\langle i \rangle} - \mu_o \right]^T \qquad \left|\Sigma_{0.hat}\right| = 1.226 \times 10^7$$

Likelihood Ratio & Wilks' lambda (Λ) (JW Eq. 5-13 p. 217):

$$\Lambda := \left(\frac{|n \cdot \Sigma_{hat}|}{|\Sigma_{0.hat}|}\right)^{\frac{n}{2}} \qquad \Lambda = 0.01595342 \qquad \Lambda^{\frac{2}{n}} = 0.661128$$

< Λ = Wilks' lambda - value is 1.0 when $X_{bar} = \mu_0$ but decreases as μ_0 increases in distance from the sample mean.

Converting to Hotellings T^2:

$$\left[1 + \frac{T_{sq}}{(n-1)}\right]^{-1} = (0.661128)$$

< Equivalent value in terms of Hotellings T^2 (jw Result 5.1 p. 218)

$$\left[\Lambda^{-\left(\frac{2}{n}\right)} - 1\right] \cdot (n-1) = 9.739 \qquad T_{sq} = (9.739)$$

< Solving for T^2 (same as T^2 above)

Sampling Distribution:

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

Critical Value of the Test:

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

$$C := \frac{(n-1) \cdot p}{(n-p)} \cdot qF(1-\alpha, p, n-p) \qquad C = 10.719$$

< JW eq. 5-6 p. 212

NOTE: qF(1-α) is used in function qF.

Decision Rule:

Reject H_0 if $T_{sq} > C$ $\qquad T_{sq} = (9.739) \qquad C = 10.7186$

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

< 0 = Do not reject H_0
1 = Reject H_0

^ Therefore DO NOT Reject H_0

Probability Value:

$$P := 1 - pF\left[T_{sq} \cdot \frac{(n-p)}{(n-1) \cdot p}, p, n-p\right] \qquad P = (0.06492834)$$

Confidence Intervals in Multivariate Analysis:

Because the problem of "confidence intervals" involves simultaneous estimation across p variables (X_1, X_2, X_3 ... X_p) that may be correlated, one "natural" discription involves use of ellipses (for p=2), ellipsoids (p=3) or "hyper-ellipsoids" (p >3) analogous to those described in worksheet MTB 070. Here, however, because the confidence intervals involve behavior with repeated sampling of *mean vectors* X_{bar}, the ellipsoids will be smaller than corresponding ellipsoids for data points at the same (1-α) level. Shown below are four common representations:

Multivariate Confidence Ellipsoid - **This is an actual description of the rejection/acceptance boundary in Hotelling's T^2 test taking into account all covariances seen in S. Constructing a confidence ellipsoid involves calculation of eigenvalues and eigenvectors of covariance matrix S, and results in a p-dimensional description of the ellipsoid in terms of eigenvector directions. All points enclosed by the ellipsoid reside within the (1-α) confidence limit for a specified α. When p > 2, results are often graphed as projections (="shadows") onto a specific 2-dimensional planes of variables two at a time. When p is large, the confidence ellipsoid is very hard to visualize.**

Simultaneous T^2 Confidence Intervals - This represents a projection of the extreme limits of the T^2 confidence ellipsoid onto a range for each variable (X_1, X_2, X_3 ... X_p). The result is a hyper-rectangular block in p-dimensional space that is considerably larger in volume than the original T^2 confidence ellipsoid. Thus, it is a highly "conservative" estimate because it may include vectors that don't actually belong. However, T^2 confidence intervals are easily tabulated, and all statements of means and all linear combinations are covered simultaneously by family-wise α. Thus all forms of post *hoc analysis* are allowed (see *Biostatistics* worksheet 250 for a discussion). Simultaneous T^2 confidence intervals, are the widest among possibilities considered.

Univariate t Confidence Intervals - Here, confidence intervals are calculated for each variable (X_1, X_2, X_3 ... X_p) independently, just as one would in unvariate t-tests. This approach ignores all covariances between the variables and may seriously inflate family-wise α. However, the results produce the narrowest invervals among possibilities considered.

Bonferroni Simultaneous Confidence Intervals - This represents a Bonferroni style compromise in which (b) comparisons are specified "in advance of data collection" and then (b) is used to modify α into a family-wise α. *Post-hoc* "data-snooping" is not allowed. The results produce intervals that are intermediate between simultaneous T^2 confidence intervals and univaraiate t confidence intervals.

So which tabulated interval should one use? Obviously, if one's results are clear than use of the most conservative T^2 approach would be completely unproblematic. However, these intervals may be too wide to be of practical use in many cases. AR cites simulation studies suggesting preference (most powerful and also maintaining family-wise α) for the univariate t intervals *following rejection* of the null hypothesis using a T^2 test. JW (p. 231) acknowledge this argument, but sound less than enthusiastic ("some researchers think..."). The Bonferroni "compromise" is also worth considering, but here one may inevitably run into controversy surrounding when the researcher conceived of certain questions relative to data collection (see *Biostatistics* worksheet 250).

The Multivariate Confidence Ellipsoid:

$\lambda := \text{reverse}(\text{sort}(\text{eigenvals}(S)))$

$\varepsilon^{\langle j \rangle} := \text{eigenvec}(S, \lambda_j)$

$$\lambda = \begin{pmatrix} 200.46246 \\ 4.53159 \\ 1.30139 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} 0.05084 & -0.5737 & 0.81748 \\ 0.99828 & 0.05302 & -0.02488 \\ -0.02907 & 0.81735 & 0.57541 \end{pmatrix}$$

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

$$CT := \sqrt{\frac{p\cdot(n-1)}{n\cdot(n-p)}\cdot qF(1-\alpha, p, n-p)} \qquad CT = 0.7320726$$

< CT gives the boundary for the confidence ellipsoid for μ - see JW eq. 5-18 p. 221

$i := 1 .. p$

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

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

$$X_{bar} = \begin{pmatrix} 4.64 \\ 45.4 \\ 9.965 \end{pmatrix}$$

< Center of ellipsoid

$$L = \begin{pmatrix} 10.36503 \\ 1.5584 \\ 0.83514 \end{pmatrix}$$

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

Simultaneous T^2 Confidence Intervals:

< JW eq. 5-24 p. 225 -slightly modified:

$$CT = \sqrt{\frac{C}{n}}$$

$$CI_{lower_i} := X_{bar_i} - CT \cdot \sqrt{S_{i,i}} \qquad CI_{upper_i} := X_{bar_i} + CT \cdot \sqrt{S_{i,i}}$$

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

Simultaneous Confidence Intervals:

$$X_{bar} = \begin{pmatrix} 4.64 \\ 45.4 \\ 9.965 \end{pmatrix}$$ **< Mean values**

$$CI = \begin{pmatrix} 3.39777 & 5.88223 \\ 35.05241 & 55.74759 \\ 8.57066 & 11.35934 \end{pmatrix}$$ **< T^2 confidence intervals**

Univariate t Confidence Intervals:

$$ct := qt\left(1 - \frac{\alpha}{2}, n - 1\right) \qquad ct = 2.093024$$

< Critical value ct based on t distribution without correction

$$ci_{lower_i} := X_{bar_i} - ct \cdot \sqrt{\frac{S_{i,i}}{n}} \qquad ci_{upper_i} := X_{bar_i} + ct \cdot \sqrt{\frac{S_{i,i}}{n}}$$

< jw eq. 5-29 p. 232

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

Univariate t Intervals:

$$X_{bar} = \begin{pmatrix} 4.64 \\ 45.4 \\ 9.965 \end{pmatrix}$$ **< Mean values**

$$ci = \begin{pmatrix} 3.84584 & 5.43416 \\ 38.78478 & 52.01522 \\ 9.0736 & 10.8564 \end{pmatrix}$$ **< Univariate t confidence intervals**

Bonferroni Simultaneous Confidence Intervals:

$$cb := qt\left(1 - \frac{\alpha}{2 \cdot p}, n - 1\right) \qquad cb = 2.625106$$

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

$$ci_{lower_i} := X_{bar_i} - cb \cdot \sqrt{\frac{S_{i,i}}{n}} \qquad ci_{upper_i} := X_{bar_i} + cb \cdot \sqrt{\frac{S_{i,i}}{n}}$$

< jw eq. 5-29 p. 232

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

Bonferroni Intervals:

$$X_{bar} = \begin{pmatrix} 4.64 \\ 45.4 \\ 9.965 \end{pmatrix}$$ **< Mean values**

$$ci = \begin{pmatrix} 3.643952 & 5.636048 \\ 37.103078 & 53.696922 \\ 8.846992 & 11.083008 \end{pmatrix}$$ **< Bonferroni confidence intervals**

Prototype in R:

```
#LOAD DATA & HYPOTHESIS:
X=read.table("c:/DATA/Multivariate\\/T5-1.DAT")
X          #DATA TABLE

#FUNCTION FOR ONE SAMPLE HOTELLING'S T2 TEST:
#X     = dataset
#mu0   = hypothesis vector
#alpha = alpha of the test
```

... function OS.Hotelling.T2() body is in R script ...

```
OS.Hotelling.T2(X,mu0=c(4,50,10),alpha=0.05)
```

```
> X
    V1   V2   V3
1  3.7 48.5  9.3
2  5.7 65.1  8.0
3  3.8 47.2 10.9
4  3.2 53.2 12.0
5  3.1 55.5  9.7
6  4.6 36.1  7.9
7  2.4 24.8 14.0
8  7.2 33.1  7.6
9  6.7 47.4  8.5
10 5.4 54.1 11.3
11 3.9 36.9 12.7
12 4.5 58.8 12.3
13 3.5 27.8  9.8
14 4.5 40.2  8.4
15 1.5 13.5 10.1
16 8.5 56.4  7.1
17 4.5 71.6  8.2
18 6.5 52.8 10.9
19 4.1 44.1 11.2
20 5.5 40.9  9.4
```

All values verified above and in JW >

```
> RES=OS.Hotelling.T2(X,mu0=c(4,50,10),alpha=0.05)

 One Sample Hotelling's T2

 Hypothesis Vector:        ( 4 50 10 )
 T2 Ellipsoid half lengths: ( 10.36503 1.558402 0.835138 )
 Hotelling's T2 Statistic:  9.738773
 Equivalent F Statistic :   2.904546
 F degrees of freedom:      ( 3 17 )
 Wilks's Lambda:            0.01595342
 alpha:                     0.05
 Critical Value:            10.7186
 Probability:               0.06492834

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

    T2.lower   B.lower   t.lower   Mean  t.upper   B.upper  T2.upper
V1  3.397768  3.643952  3.845840  4.640  5.43416  5.636048  5.882232
V2 35.052408 37.103078 38.784779 45.400 52.01522 53.696922 55.747592
V3  8.570664  8.846992  9.073601  9.965 10.85640 11.083008 11.359336
```

```
library(ICSNP)
HotellingsT2(X,mu=c(4,50,10))
```

```
> HotellingsT2(X,mu=c(4,50,10))

        Hotelling's one sample T2-test

data:  X
T.2 = 2.9045, df1 = 3, df2 = 17, p-value = 0.06493
alternative hypothesis: true location is not equal to c(4,50,10)
```

Test provides the equivalent F statistic > P values match.

```
library(rrcov)
T2.test(X,mu=c(4,50,10),method="c")
```

```
> T2.test(X,mu=c(4,50,10),method="c")

        One-sample Hotelling test

data:  X
T^2 = 2.9045, df1 = 3, df2 = 17, p-value = 0.06493
alternative hypothesis: true mean vector is not equal to (4, 50, 10)'

sample estimates:
                V1   V2    V3
mean x-vector 4.64 45.4 9.965
```

Test provides the equivalent F statistic > P values match.

```
plot.CI <- function(M,R,X,Y)
--- body of function in R script ...
plot.CI(X,RES,1,2)
plot.CI(X,RES,1,3)
plot.CI(X,RES,2,3)
```

Simultaneous T^2 intervals (solid red) are the widest, univariate t intervals (dashed) are the narrowest, with Bonferroni intervals (dotted) intermediate. Blue cross is the mean vector, red triangle is the hypothesis vector: