Session 6: Single Sample and Two Sample Multivariate Hotelling's $T^2$ Test

Lecture Notes by: J. René van Dorp

www.seas.gwu.edu/~dorpjr

---

1 Department of Engineering Management and Systems Engineering, School of Engineering and Applied Science, The George Washington University, 1776 G Street, N.W., Suite 110, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.
Recall Scatter Plot of Height and Weight of 20 Individuals
• Perhaps the joint normal distribution is a good model to describe the joint variation in height and weight of our multivariate sample of height and weight combinations of 20 individuals.
**MULTIVARIATE INFERENCE**  
Joint Normal Distribution

- **Probability density function** of a bivariate normal distribution:

\[
X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim MN(\mu, \Sigma), \text{ Mean Vector: } \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},
\]

**Covariance Matrix**: \(\Sigma = \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{pmatrix}\)

\[
f(x, y) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[ (x - \mu)^T \Sigma^{-1} (x - \mu) \right]
\]

- **Definition Covariance Matrix**:

\[
\Sigma = E[(X - \mu)(X - \mu)^T] \ (\text{Recall: } Var(X) = E[(X - \mu)^2], \mu = E[X])
\]

\[(n \times 1\text{-matrix}) \cdot (1 \times n\text{-matrix}) = (n \times n\text{-matrix})\]

- **How do we estimate the mean vector** \(\mu\), and the variance covariance matrix \(\Sigma\)
**MULTIVARIATE INFERENCES**  
**Point Estimation**

- Let \( \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \) be a random sample from a joint distribution (with dimension \( m \)) with mean vector \( \mu \) and covariance matrix \( \Sigma \), where the \( m \)-dimensional vectors \( \mathbf{X}_i, i = 1, \ldots, n \) are independent, the RV elements within each random vector are not. Then

\[
\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i
\]

is an unbiased estimator for the mean vector \( \mu \) and its variance-covariance matrix is \( \frac{1}{n} \Sigma \). In other words: \( E[\bar{\mathbf{X}}] = \mu \). Also:

\[
Cov[\bar{\mathbf{X}}] = E[(\bar{\mathbf{X}} - \mu)(\bar{\mathbf{X}} - \mu)^T] = \frac{1}{n} \Sigma
\]

For convenience we shall from hereon write vectors only in bold font and not underline them anymore.

- Recal, the univariate case where \( E[\bar{X}] = \mu \) and \( V[\bar{X}] = \sigma^2/n \) (given a random sample \((X_1, \ldots, X_n), X_i \sim X \) with \( E[X] = \mu \) and \( V[X] = \sigma^2 \)).
MULTIVARIATE INFERENCE

Point Estimation

- Let \( X_1, X_2, \ldots, X_n \) be a random sample from a joint distribution with dimension \( m, (m \times 1) \) mean vector \( \mu \) and \( (m \times m) \) covariance matrix \( \Sigma \), where the \( (m \times 1) \) vectors \( X_i, i = 1, \ldots, n \) are independent. Then

\[
S = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T
\]

is an unbiased estimator for the variance covariance matrix \( \Sigma \), i.e. \( E[S] = \Sigma \), where the expectatios are taken elements wise.

- Recall the unbiased estimator for \( \sigma^2 \) in the univariate case

\[
S^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

Example: Height-Weight data

\[
\bar{X} = \begin{pmatrix} 62.85 \\ 123.60 \end{pmatrix}, \quad S = \begin{pmatrix} 10.87 & 44.52 \\ 44.52 & 242.46 \end{pmatrix}
\]
MULTIVARIATE INFEERENCE

Matrix Determinant

- When we are considering more than two variables it is useful to obtain a single measure of linear dependence between them. The most common measure for this purpose is the matrix determinant $|\Sigma|$ of the variance-covariance matrix $\Sigma$. Considering the two-dimensional case:

$$\Sigma = \begin{pmatrix} V[X_1] & Cov[X_1, X_2] \\ Cov[X_1, X_2] & V[X_2] \end{pmatrix} \text{ then } |\Sigma| = V[X_1]V[X_2] - Cov^2[X_1, X_2]$$

- Thus, in the 2-dimensional case this dependence measure is reminiscent of the $Cov[X_1, X_2]$ between $X_1$ and $X_2$. Note that:

$$Cov[X_1, X_2] = 0 \iff |\Sigma| = V[X_1]V[X_2]$$

The highest value of $|\Sigma|$ is attained when $X_1$ and $X_2$ are uncorrelated and the smallest when $X_2 = aX_1 + b$, where $a$ and $b$ are constants. Hence, the higher the value of $|\Sigma|$ the less linearly dependent $X_1$ and $X_2$ are. This interpretation carries over to more than two variables. The measure $|\Sigma|$ is called the Generalized Variance (since its value is affected by the variances of the random variables, just like $Cov[X_1, X_2]$).
Example: Height-Weight data

\[
S = \begin{pmatrix}
10.87 & 44.52 \\
44.52 & 242.46
\end{pmatrix},
\]

\[
|S| = (10.87 \times 242.46) - (44.52 \times 44.52) \approx 654.17
\]
The EXCEL function MDETERM calculates $|\Sigma|$ for more than 2 variables.

The value of $|\Sigma|$ may be dominated by one variable due to a difference in scale between the two variables (leading to a comparatively large variance of one of the variables). The same is true for the value of $\text{Cov}(X_1, X_2)$.

To resolve this issue it is not uncommon to first standardize the data and next calculate the variance-covariance matrix $R$ of the standardized data set and its determinant $|R|$. We denote this matrix by $R$ since it is the correlation matrix of the original data set. Similarly, we look at $\rho(X_1, X_2)$ (i.e. the standardized covariance) to obtain a measure of linear dependence.

Example: Height-Weight data

$$R = \begin{pmatrix} 1 & 0.867 \\ 0.867 & 1 \end{pmatrix},$$

$$|R| = 1 - (0.867)^2 \approx 0.248$$
MULTIVARIATE INFERENCE

Geometric Interpretation

- It can be shown that:

\[ |S| = \left\{ \prod_{i=1}^{n} s_{ii} \right\} |R| = \left\{ \prod_{i=1}^{n} Var[X_i] \right\} |R| \]

(which is why they named $|S|$ the generalized variance.)

Example: Height-Weight data

\[ |S| = (10.87 \times 242.46), \quad |R| = (10.87 \times 242.46) \times 0.248 \approx 654.17 \]
MULTIVARIATE INference  Geometric Interpretation

• Note that:

\[
R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \iff |R| = 1 \iff \rho(X_1, X_2) = 0
\]

The highest value of \(|R|\) is attained when \(X_1\) and \(X_2\) are uncorrelated and the smallest when \(X_2 = aX_1 + b\), where \(a\) and \(b\) are constants. Hence, the higher the value of \(|R|\) the less linearly dependent \(X_1\) and \(X_2\) are. The above interpretation of \(|R|\) carries over to more than two dimensions.

• BE CAREFULL! The closer the value of \(|R|\) to 0 the higher the degree of linear dependence (referred to as multi-collinearity in dimensions higher than 2). When the value of \(|R| = 1\) there is no colinearity present within the data.

• Observe that "the direction" of \(|R|\) is opposite to that of the correlation coefficient \(\rho(X_1, X_2)\).

\[
\rho(X_1, X_2) = \pm 1 \iff X_2 = aX_1 + b \text{ for some } a, b \in \mathbb{R}
\]

• Also \(|R| > 0\), whereas \(-1 \leq \rho(X_1, X_2) \leq 1\).
Recall univariate $t$-hypothesis test

$H_0 : \mu = \mu_0, \ H_1 : \mu \neq \mu_0 \Rightarrow \text{Reject } H_0 \text{ when } |T| \geq t_{n-1,1-\frac{\alpha}{2}}$
MULTIVARIATE INFERENCE

Hotelling $T^2$ Test

Reject $H_0$ when $|T| \geq t_{n-1,1-\alpha/2} \iff$ Reject $H_0$ when $T^2 \geq t_{n-1,1-\alpha/2}^2$

\[ T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}}{s} (\bar{x} - \mu_0) \iff T^2 = n(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0) \]

- Let $X_1, \ldots, X_n$ now be an i.i.d. \textbf{$p$-dimensional} sample from a $MVN(\mu, \Sigma)$. Note that the random vectors are \textit{i.i.d.} but the elements of each random vector are not. Then with

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \]

we can now define the Hotelling $T^2$-Statistic/Estimator:

\[ T^2 = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0) \]

which is a \textbf{direct generalization of the $T$-statistic for a univariate normal i.i.d. sample.}
MULTIVARIATE INference

Hotelling $T^2$ Test

Hotelling showed that:

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

$$H_0 : \mu = \mu_0, \ H_1 : \mu \neq \mu_0 \Rightarrow \text{Reject } H_0 \text{ when } \frac{n-p}{(n-1)p} \times t^2 \geq F_{p,n-p,1-\alpha}$$

Example: Height-Weight data

$$\bar{X} = \begin{pmatrix} 62.85 \\ 123.60 \end{pmatrix}, \ S = \begin{pmatrix} 10.87 & 44.52 \\ 44.52 & 242.46 \end{pmatrix}, \ \mu_0 = \begin{pmatrix} 62 \\ 120 \end{pmatrix}, \ p = 2, n = 20$$

$$t^2 = 20(.85 \ 3.60)\begin{pmatrix} 10.87 & 44.52 \\ 44.52 & 242.46 \end{pmatrix}^{-1}\begin{pmatrix} .85 \\ 3.60 \end{pmatrix} \approx 1.33$$

Significance Level: $\alpha = 10\%$, $\frac{18}{19 \times 2} t^2 \approx 0.63 < F_{2,18,0.9} \approx 2.624$

$\Rightarrow$ Fail to Reject $H_0: \mu_0 = \begin{pmatrix} 62 \\ 120 \end{pmatrix}$
MULTIVARIATE INFERENCE

Two Sample Mean Test

The two-sample \textit{t} test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value:
\[
t_0 = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}\]

Degrees of freedom $\nu$:
\[
\nu = \frac{\left[ \frac{s_1^2}{n} + \frac{s_2^2}{m} \right]^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_1^2/m)^2}{m-1}}
\]

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Rejection Regions for significance $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1 : \mu_1 - \mu_2 &gt; \Delta_0$</td>
<td>$t_0 &gt; t_{\nu,1-\alpha}$ (upper-tailed)</td>
</tr>
<tr>
<td>$H_1 : \mu_1 - \mu_2 &lt; \Delta_0$</td>
<td>$t_0 &lt; -t_{\nu,1-\alpha}$ (lower-tailed)</td>
</tr>
<tr>
<td>$H_1 : \mu_1 - \mu_2 \neq \Delta_0$</td>
<td>$t_0 &gt; t_{\nu,1-\alpha/2}$ or $t_0 &lt; -t_{\nu,1-\alpha/2}$ (two-tailed)</td>
</tr>
</tbody>
</table>

$p$-values can be constructed in a similar fashion as before.
MULTIVARIATE INFERENCE Two Sample Mean Test

Let $X_{11}, \ldots, X_{1n}$ now be an i.i.d. $p$-dimensional sample from population 1:
$$MVN(\mu_1, \Sigma_1).$$

Let $X_{21}, \ldots, X_{2m}$ now be an i.i.d. $p$-dimensional sample from population 2:
$$MVN(\mu_2, \Sigma_2).$$

Then with
$$\bar{X}_1 = \frac{1}{n} \sum_{i=1}^{n} X_{1i}, \quad S_1 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)^T,$$
$$\bar{X}_2 = \frac{1}{m} \sum_{i=1}^{m} X_{2i}, \quad S_2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_{2i} - \bar{X}_2)(X_{2i} - \bar{X}_2)^T,$$

(The first subscript 1 or 2 denotes the population number)

we can define another $T^2$-statistic if $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2m}$ are independent too.
MULTIVARIATE INFERENCE  Two Sample Mean Test

The two-multivariate $T^2$ test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ is as follows:

$$T^2 = [\overline{X}_1 - \overline{X}_2 - \Delta_0]^T \left[\left(\frac{1}{n} + \frac{1}{m}\right) S_{pooled}\right]^{-1} [\overline{X}_1 - \overline{X}_2 - \Delta_0] ,$$

where

$$S_{pooled} = \frac{(n-1)S_1 + (m-1)S_2}{n + m - 2} ,$$

which is a direct generalization of the two sample $T$-statistic for a univariate normal $i.i.d.$ sample. If sample sizes $n$ and $m$ are small, one additional assumption is needed to be able to use $T^2$ to conduct hypothesis tests.

That assumption is that the variance-covariance matrices of both populations are the same: $\Sigma_1 = \Sigma_2$. In that case:

$$\frac{(n + m - p - 1)}{(n + m - 2)p} \times T^2 \sim F_p, n + m - p - 1$$
Example Wisconsin Power Data:
Samples of sizes $n = 45$ and $m = 55$ were taken of Wisconsin homeowners with and without 

**airconditioning**, respectively. (Data courtesy of Statistical Laboratory, University of Wisconsin). Two measurements of electrical usage (in kilowatt hours) were considered. The first is a measure of total on-peak consumptions ($X_1$) during July 1977 and the second is a measure of total off-peak consumption ($X_2$) during July 1977. (The off-peak consumption is higher than the on-peak consumption because there are more off-peak hours in a month). The resulting summary statistics are:

With AirCo: $\bar{x}_1 = \left(\begin{array}{c} 204.4 \\ 556.6 \end{array}\right)$, $S_1 = \left(\begin{array}{cc} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{array}\right)$, $n = 45$

Without AirCo: $\bar{x}_2 = \left(\begin{array}{c} 130.0 \\ 355.0 \end{array}\right)$, $S_2 = \left(\begin{array}{cc} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{array}\right)$, $m = 55$

$$S_{pooled} = \frac{(n - 1)S_1 + (m - 1)S_2}{n + m - 2} = \left(\begin{array}{cc} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{array}\right)$$
MULTIVARIATE INFERENCE  Two Sample Mean Test

- We want to test: \( H_0 : (\mu_1 - \mu_2)^T = (0 \quad 0)^T = \Delta_0^T \)

\[
t^2 = [\bar{x}_1 - \bar{x}_2 - \Delta_0]^T \left( \frac{1}{n} + \frac{1}{m} \right) S_{\text{pooled}}^{-1} [\bar{x}_1 - \bar{x}_2 - \Delta_0]
\]

\[
= \begin{pmatrix} 204.4 - 130.0 \\ 556.6 - 355.0 \end{pmatrix}^T \begin{pmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{pmatrix}^{-1} \begin{pmatrix} 204.4 - 130.0 \\ 556.6 - 355.0 \end{pmatrix}
\]

\[
= \begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix}^T \begin{pmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{pmatrix}^{-1} \begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \approx 16.07
\]

\[
\frac{(n + m - p - 1)}{(n + m - 2)p} t^2 = \frac{(45 + 55 - 3)}{(45 + 55 - 2)^2} 16.07 \approx 7.95 > F_{2,97,0.95} \approx 3.09
\]

Conclusion: Reject \( H_0 \) (i.e. there is a difference between airconditioning and no airconditioning consumption).
MULTIVARIATE INFERENCE

Two Sample Mean Test

• In the case that sample sizes are large, the assumption \( \Sigma_1 = \Sigma_2 \) may be relaxed to allow for \( \Sigma_1 \neq \Sigma_2 \). In that case:

\[
T^2 = \left[ \overline{X}_1 - \overline{X}_2 - \Delta_0 \right]^T \left[ \left( \frac{S_1}{n} + \frac{S_2}{m} \right) \right]^{-1} \left[ \overline{X}_1 - \overline{X}_2 - \Delta_0 \right]
\]

and

\[
T^2 \sim \chi_p^2
\]

Example Wisconsin Power Data:

\[
\overline{x}_1 = \begin{pmatrix} 204.4 \\ 556.6 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{pmatrix}, \quad n = 45
\]

\[
\overline{x}_2 = \begin{pmatrix} 130.0 \\ 355.0 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{pmatrix}, \quad m = 55
\]

\[
\frac{S_1}{n} + \frac{S_2}{m} = \begin{pmatrix} 443.0 & 868.9 \\ 868.9 & 2572.2 \end{pmatrix}
\]
Two Sample Mean Test

\[ T^2 = \left[ \overline{X}_1 - \overline{X}_2 - \Delta_0 \right]^T \left[ \left( \frac{1}{n} + \frac{1}{m} \right) S_{pooled} \right]^{-1} [ \overline{X}_1 - \overline{X}_2 - \Delta_0 ] \]

\[ = \begin{pmatrix} 204.4 - 130.0 \\ 556.6 - 355.0 \end{pmatrix}^T \begin{pmatrix} 443.0 & 868.9 \\ 868.9 & 2572.2 \end{pmatrix}^{-1} \begin{pmatrix} 204.4 - 130.0 \\ 556.6 - 355.0 \end{pmatrix} \]

\[ = \begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix}^T (10^{-4}) \begin{pmatrix} 59.874 & -20.080 \\ -20.080 & 10.519 \end{pmatrix} \begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \approx 15.66 \]

\[ T^2 \approx 15.66 > \chi^2_{2,0.95} \approx 5.99 \]

Conclusion: Reject \( H_0 \) (i.e. there is a difference between airconditioning and no airconditioning consumption).