Multivariate hypothesis testing

- Univariate tests
  - t-test, ANOVA

- Multivariate tests
  - Hotelling’s T²
  - MANOVA
  - Data reduction and MANOVA

Can’t I just run multiple univariate tests to test my hypothesis?

Example:
- Compare 2 groups that do not differ
- $\alpha$ set to 0.05
  (i.e., the probability of a correct interpretation for each comparison is 0.95)
- 10 variables

\[
\alpha_n = 1 - 0.95^n
\]

Probability of finding significant difference between 2 groups = 40%

Multivariate vs. univariate responses

- Testing single variables separately may not reveal multivariate differences
- Response variables may be correlated with one another
- Interactions may exist among response variables
Univariate situation: \( t \)-test

Compares 2 means

\[
t = \frac{|\overline{x}_1 - \overline{x}_2|}{s_{\overline{x}_1 - \overline{x}_2}}
\]

\[
s_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \text{pooled variance}
\]

Univariate situation: ANOVA

Compares > 2 means

\[
F\text{-statistic} = \frac{\text{Among groups MS}}{\text{Within groups MS}}
\]

Assumptions of univariate tests

- Data normally distributed
  - fairly robust to deviations as long as single peak
- Observations independent
- Variances homogeneous
  - tests most sensitive to deviations here
### Corresponding multivariate tests

<table>
<thead>
<tr>
<th># Groups</th>
<th>Univariate</th>
<th>Multivariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>t-test</td>
<td>Hotelling’s $T^2$</td>
</tr>
<tr>
<td>$&gt;2$</td>
<td>ANOVA</td>
<td>MANOVA</td>
</tr>
</tbody>
</table>

### Assumptions of multivariate tests

- Multivariate normality
  - difficult to test
- Observations independent
- Variances homogeneous

### Hotelling’s $T^2$ is the multivariate equivalent to the t-test

- Compares 2 groups of multiple dependent variables
- Adjusts for multicolinearity
- The $T^2$ statistic is the ratio of the difference between the sample mean vectors and the dispersion matrix
Sample mean vectors define the magnitude of difference between the treatments

\[
\bar{X}_1 = \begin{bmatrix}
\bar{x}_{11} \\
\bar{x}_{12} \\
\bar{x}_{13}
\end{bmatrix}
\quad \text{and}
\quad
\bar{X}_2 = \begin{bmatrix}
\bar{x}_{21} \\
\bar{x}_{22} \\
\bar{x}_{23}
\end{bmatrix}
\]

\( j = \) sample  
\( k = \) variable

\[\begin{array}{ccc}
\text{Var1} & \text{Var2} & \text{Var3} \\
\text{Treat1} & 2 & 18 & 5 \\
\text{Var1} & 3 & 15 & 8 \\
\text{Treat1} & 7 & 23 & 12 \\
\text{Treat2} & 3 & 17 & 3 \\
\text{Treat2} & 3 & 29 & 17 \\
\text{Treat2} & 21 & 73 & 4 \\
\end{array}\]

The dispersion matrix is a pooled sample variance matrix

Sample variance matrix for population \( i \)

\[
S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_i)(X_{ij} - \bar{x}_i)'
\]

Pooled sample variance matrix

\[
S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}
\]

Hotelling’s \( T^2 \) is the ratio of the difference between the sample mean vectors and the pooled variance matrix

\[
T^2 = (\bar{X}_1 - \bar{X}_2)' \left( S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right)^{-1} (\bar{X}_1 - \bar{X}_2)
\]

Must have more observations \((n)\) than variables \((p)\)

\[
n_1 + n_2 - 2 > p
\]

\( df \)
**Multivariate analysis of variance: MANOVA**

- An extension of ANOVA to multiple dependent variables
- Compares > 2 independent variables
- Tests whether mean differences among groups for a combination of dependent variables likely occurred by chance
- Can accommodate factorial designs

**Significance is determined by the ratio of among group vs. within group variation**

- Among group MS:
  \[ H = \sum_{i=1}^{k} \left( \bar{x}_i - \bar{x} \right)^T \left( \bar{x}_i - \bar{x} \right) \]

- Within group MS:
  \[ E = \sum_{j=1}^{n} \sum_{i=1}^{k} \left( x_{ij} - \bar{x}_j \right)^T \left( x_{ij} - \bar{x}_j \right) \]

**Four test statistics: ratios of among group variation to within group variation**

- Wilk’s Lambda (exact)
- Pillai’s Trace
- Lawley-Hotelling Trace
- Roy’s Greatest Root (upper bound)
Data reduction and MANOVA

- A PCA may be conducted to reduce the dimensionality of the dataset into fewer variables
- PCA scores can be used as summary variables in a MANOVA
- Beneficial in testing hypotheses in high variable number datasets with presumed high redundancy