Chapter 2: Inference about the mean vector(s)

Ana-Maria Staicu
SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu
In this chapter we will discuss inference about a single population mean vector as well as compare two or more population mean vectors. We discuss one way MANOVA and two-way MANOVA. We will consider typical component-by-component comparison as well as profile analysis. Recall by inference we mean reaching conclusions concerning a population parameter using information from data.

Let $\mu_1$ and $\mu_2$ be two population mean vectors. In this chapter we will study

- how to carry out inference about $\mu_1$ (hypothesis testing and confidence intervals)
- how to formally assess a hypothesis of the form $H_0 : \mu_1 = \mu_2$
- how to formally assess a hypothesis of the form $H_0 : \text{the change in the components of } \mu_1 \text{ is the same as the change in the components of } \mu_2$
- how to construct confidence intervals for the mean difference $\mu_1 - \mu_2$

in a variety of scenarios.

**Motivating example**: Fisher’s or Anderson’s Iris data.


*Data set*: Available in R as ‘iris’

This data set consists of the measurements of the variables sepal length and width, and petal length and width (in centimeters), respectively, for 50 flowers from each of 3 species (setosa, versicolor and virginica) of iris. The first few rows of the data matrix is shown below.

<table>
<thead>
<tr>
<th>Id</th>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>2</td>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>3</td>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>4</td>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>6</td>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td>setosa</td>
</tr>
</tbody>
</table>

**Question 1**: Estimate the mean of (SL, SW, PL, PW) for setosa flowers.

**Question 2**: Provide confidence intervals.

**Question 3**: Compare the mean of these characteristics across species.
1 Inference for a single population: Confidence intervals. Hypothesis testing.

**Review of the univariate case:** Let $X_1, X_2, \ldots, X_n$ be a sample from a normal distribution with mean $\mu$ and unknown variance. Let $\alpha \in (0, 1)$.

Constructing confidence intervals / bands for $\mu$ as well as hypothesis testing procedures mainly rely on the result:

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where $t_{n-1}$ is the Student $t$ distribution with $n - 1$ degrees of freedom.

Then we can construct confidence intervals for $\mu$ by using an estimate for the standard error and appropriate critical value. Specifically let $t_{n-1}(\alpha/2)$ be the upper tail probability corresponding to the $t_{n-1}$ distribution defined as $P(t_{n-1} > t_{n-1}(\alpha/2)) = \alpha/2$. Then the 100(1 $-$ $\alpha$)% confidence interval (CI) for $\mu$ is

$$\left( \bar{X} - t_{n-1}(\alpha/2) \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}(\alpha/2) \frac{S}{\sqrt{n}} \right).$$

The 100(1 $-$ $\alpha$)% CI for $\mu$ can be written as $|\sqrt{n}(\bar{X} - \mu)/S| \leq t_{n-1}(\alpha/2)$ or equivalently

$$n(\bar{X} - \mu)^2/S^2 \leq F_{1,n-1}(\alpha)$$

since $t_{n-1}(\alpha/2)^2 = F_{1,n-1}(\alpha)$ where $F_{1,n-1}(\alpha)$ is the upper tail probability corresponding to the $F_{1,n-1}$ distribution defined as $P(F_{1,n-1} > F_{1,n-1}(\alpha)) = \alpha$.

This remark allows us to extend the confidence interval to confidence region, which is more appropriate when the population parameter of interest (e.g. mean vector) has a higher dimension than one.

**Definition:** When the parameter, denoted generically by $\theta$, has dimension $p \geq 2$, the 100(1 $-$ $\alpha$)% confidence region based on same sample $X$, denoted by $R(X)$, is defined as

$$P(R(X) \text{ will cover the true } \theta) = 1 - \alpha,$$

where the probability is calculated using the true parameter value $\theta$. 
CASE 1: Normal parent distribution.

Assume \( \mu \) is the mean \( p \)-dimensional vector parameter of interest. Let \( X_1, X_2, \ldots, X_n \) be a random sample from a \( p \)-variate normal parent distribution with mean \( \mu \) and unknown covariance matrix \( \Sigma \).

Let \( \bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \) be the sample mean and \( S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \) be the sample covariance matrix. Most of the inferential procedures that we study use the test statistic (called Hotelling’s \( T^2 \))

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

and rely on the result that

\[
T^2 \sim \frac{p(n-1)}{n-p} F_{p,n-p}.
\]

It follows that \( P\{T^2 \leq \frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)\} = 1 - \alpha \) for critical value \( F_{p,n-p}(\alpha) \).

Confidence region for \( \mu \).

Let \( x_1, x_2, \ldots, x_n \) be the observed data. The 100(1 - \( \alpha \))% confidence region for \( \mu \) is the ellipsoid determined by all \( \mu \) such that

\[
n(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha);
\]

here \( \bar{x} \) and \( S \) are the sample mean and sample covariance for the particular observed data.

Remarks:

- The confidence region is exact for any value of \( n \)
- For large \( n \) the limiting distribution \( \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \approx \chi^2_p \), i.e. the two variables on each side of \( \approx \) have approximately the same distribution

Example: In class demonstration using iris data.
While the (ellipsoidal) region \( n(\bar{x} - \mu)^T S^{-1}(\bar{x} - \mu) \leq c \) for some \( c > 0 \) assesses joint knowledge concerning plausible values of \( \mu \), it is common to summarize the conclusions for the individual components of \( \mu \) separately. In doing so, we adopt the principle that the individual intervals hold simultaneously with a specified high probability.

**Definition:** Suppose \( \mu = (\mu_1, \ldots, \mu_p)^T \). The intervals \( I_1, \ldots, I_p \) are simultaneous 100(1 - \( \alpha \))% confidence intervals for \( \mu_1, \ldots, \mu_p \) if

\[
P(I_k \text{ will contain true } \mu_k \text{ simultaneously for every } k = 1, \ldots, p) = 1 - \alpha.
\]

**Key ideas:** Consider linear combinations \( a^T \mu \); the 100(1 - \( \alpha \))% Cl for \( a^T \mu \) is

\[
\left| \frac{\sqrt{n}(a^T \bar{X} - a^T \mu)}{\sqrt{a^T S a}} \right| \leq t_{n-1}(\alpha/2).
\]

By conveniently choosing \( a \) we can obtain individual confidence intervals for the components of \( \mu \). However these confidence intervals are not simultaneous; the confidence associated with all the statements taken together is smaller than \( (1 - \alpha) \).

In order to obtain simultaneous confidence intervals we need to evaluate

\[
\max_a \left| \frac{\sqrt{n}(a^T \bar{X} - a^T \mu)}{\sqrt{a^T S a}} \right|
\]

which is an equivalent problem to evaluating

\[
\max_a \left| \frac{n(a^T \bar{X} - a^T \mu)^2}{a^T S a} \right| = n(\bar{X} - \mu)^T S^{-1}(\bar{X} - \mu);
\]

the right hand side terms is \( T^2 \), and its sampling distribution is \( \frac{(n-1)p}{(n-p)} F_{p,n-p} \).

Thus, simultaneously for all \( p \)-dimensional vectors \( a \), the interval

\[
a^T \bar{X} \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{kk}}{n}}, \quad k = 1, \ldots, p,
\]

will contain \( a^T \mu \) with probability \( 1 - \alpha \).

**Simultaneous confidence intervals for the components of \( \mu \)**

Let \( \bar{X} = (\bar{X}_1, \ldots, \bar{X}_p) \). The simultaneous 100(1 - \( \alpha \))% confidence intervals for \( \mu_k \), \( k = 1, \ldots, p \) are

\[
\bar{X}_k \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{S_{kk}/n}, \quad k = 1, \ldots, p,
\]

where \( S_{kk} \) is the \( k \)th element of the diagonal of the sample covariance \( S \).
We can compare these simultaneous confidence intervals to the individual intervals (one-at-a time) for $\mu_k, k = 1, \ldots, p$. Assume we are concerned with a smaller number of parameters; these could be individual components of the population mean vector, or more generally linear combinations of the population mean vector $a_j^T \mu$ for $j = 1, \ldots, m$ for some $m$. In these situations, it may be possible to do better than the simultaneous confidence intervals. Specifically it may be possible to find a critical value that yields shorter intervals while still preserving the correct level of confidence; the method is commonly referred to as Bonferroni correction.

**Key idea:** Key ideas: Let $C_j$ denote a confidence statement about the value of $a_j^T \mu$ with level of confidence $(1 - \alpha')$. Then $P(C_j \text{ true}) = 1 - \alpha'$. It follows that:

$$P(\text{all } C_j \text{ true}) = 1 - P(\text{at least one } C_j \text{ false}) \geq \ldots = 1 - m\alpha'.$$

One can preserve an overall level of confidence of $1 - \alpha$ by choosing the individual level of confidence $1 - \alpha'$ such that $\alpha' = \alpha/m$.

**Bonferroni-corrected confidence intervals for the components of $\mu$.**

The simultaneous $100(1 - \alpha)\%$ confidence intervals for $\mu_k, k = 1, \ldots, p$ are

$$\bar{X}_k \pm t_{n-1} \left( \frac{\alpha}{2p} \right) \sqrt{S_{kk}/n}, \quad k = 1, \ldots, p$$

where $S_{kk}$ is the $k$th element of the diagonal of the sample covariance $S$. 
Hypothesis testing \( H_0 : \mu = \mu_0 \) (using Hotelling’s \( T^2 \)).

We want to test the above null hypothesis versus the alternative that \( \mu \neq \mu_0 \) using significance level \( \alpha \). Let \( x_1, x_2, \ldots, x_n \) be the observed data. For this we will use the test statistic \( T^2 = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0) \). Recall that when the null hypothesis is true \((\mu = \mu_0)\) then \( T^2 \sim (n - 1)p F_{p,n-p} \).

So we would reject \( H_0 \) at level \( \alpha \) if

\[
n(\bar{x} - \mu_0)^T S^{-1}(\bar{x} - \mu_0) > \frac{(n - 1)p}{n - p} F_{p,n-p}(\alpha).
\]

There is a general principle used to develop powerful testing procedure, and this is called likelihood ratio (LR) method. Intuitively, the test derived by this principle is equal to the ratio between the maximum likelihood under the null hypothesis \((\mu = \mu_0)\) and the maximum likelihood under the the alternative hypothesis; the LR testing procedures have several optimal properties for large samples. A LR test rejects \( H_0 \) in favor of the alternative if the LR is smaller than some suitably chosen constant.

Hypothesis testing \( H_0 : \mu = \mu_0 \) (using Wilks’s lambda).

The likelihood ratio test for testing \( H_0 : \mu = \mu_0 \) against \( H_1 : \mu \neq \mu_0 \) is

\[
\Lambda := \left( \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} \right)^{n/2}
\]

where \( \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \) and \( \hat{\Sigma}_0 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)(x_i - \mu_0)^T \) are the estimated covariance under the alternative \( H_1 \) and null \( H_0 \) hypotheses respectively. Note that \( \hat{\Sigma} \) was previously denoted by \( S \); the new notation is required by the set up. Also \( |A| \) denotes the determinant of the square matrix \( A \).

The equivalent statistic \( \Lambda^{2/n} = |\hat{\Sigma}|/|\hat{\Sigma}_0| \) is also known as the Wilks’s lambda statistic.

It turns out that \( \Lambda^{2/n} = \{1 + T^2/(n - 1)\}^{-1} \), where \( T^2 \) is the Hotelling’s statistic. So we would reject \( H_0 \) at level \( \alpha \) if

\[
\Lambda < c_\alpha.
\]

where \( c_\alpha \) is the lower 100\( \alpha \)% percentile of the distribution of \( \Lambda \).
CASE 2: Unknown parent distribution. Large sample size $n$.

Assume $\mu$ is the mean $p$-dimensional vector parameter of interest. Let $X_1, X_2, \ldots, X_n$ be a sample from a $p$-multivariate parent distribution with mean $\mu$ and some unknown covariance matrix $\Sigma$. When the sample size $n$ is large, the inference about $\mu$ is done as above, with the difference that the distribution of the test statistics is approximate and thus the confidence intervals are approximately accurate.

When the sample size is large, the limiting distribution, \( \frac{(n-1)p}{n-p} F_{p,n-p} \), converges to chi-square distribution $\chi^2_p$. Thus

- the hypothesis testing is based on the $\chi^2_p$ asymptotic null distribution of the $T^2$ test;
- the confidence intervals/region use the critical value from $\chi^2_p$.

In the lack of information about the latent parent distribution (that is multivariate normal) one needs to check the multivariate normality assumption.

**Assessing normality.** Most of the techniques that we study in this course assume that the parent distribution is multivariate normal or that the sample size sufficiently large (in which case the normality assumption is less crucial). However the quality of the inferences relies on how close the parent distribution is to the multivariate normal. Thus it is important to validate the normality assumption.

Some diagnostic plots:

- Assess whether the individual components (characteristics) come from univariate normal distribution. For this use: quantile-quantile plot (\texttt{qqplot()}), goodness-of-fit test, Shapiro-Wilk test

- Assess whether pairs of components (pairs of characteristics) have ellipsoidal appearance. Contour and perspective plots for assessing bivariate normality

- Chisquare Q-Q plot to check the multivariate normality.

  The variable $d^2 = (X - \mu)^T \Sigma^{-1} (X - \mu)$ has a chi-square distribution with $p$ degrees of freedom, and for large samples the observed Mahalanobis distances have an approximate chi-square distribution. This result can be used to evaluate (subjectively) whether a data point may be an outlier and whether observed data may have a multivariate normal distribution.

  A qq-plot can be used to plot the Mahalanobis distance for the sample vs the theoretical quantile of a chi-square distribution with $p$ degrees of freedom. This should resemble a straight-line for data from a multivariate normal distribution. Outliers will show up as points on the upper right side of the plot for which the Mahalanobis distance is notably greater than the chi-square quantile value.

- Other testing procedures to check for normality exist. See the R code part.

Exercise: Application to the iris data.
2 Treatment mean comparison, when treatments arise from factors

Dogs anesthetics example (Johnson and Wichern, 2007). Improved anesthetics are often developed by first studying their effects on animals. In one study 19 dogs were initially given the drug pentobarbitol. Each dog was then administered carbon dioxide $\text{CO}_2$ at each of two pressure levels. Next, halothane (H) was added and then administration of $\text{CO}_2$ was repeated. The response, milliseconds between heartbeats was measured for the four treatment combinations. The data are included in the folder called "Data" that is available on the course website.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>CO$_2$ pressure</th>
<th>Halothane</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high</td>
<td>present</td>
</tr>
<tr>
<td>2</td>
<td>low</td>
<td>present</td>
</tr>
<tr>
<td>3</td>
<td>high</td>
<td>absent</td>
</tr>
<tr>
<td>4</td>
<td>low</td>
<td>absent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Id</th>
<th>trt 1</th>
<th>trt 2</th>
<th>trt 3</th>
<th>trt 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>426</td>
<td>609</td>
<td>556</td>
<td>600</td>
</tr>
<tr>
<td>2</td>
<td>253</td>
<td>236</td>
<td>392</td>
<td>395</td>
</tr>
<tr>
<td>3</td>
<td>359</td>
<td>433</td>
<td>349</td>
<td>357</td>
</tr>
<tr>
<td>4</td>
<td>432</td>
<td>431</td>
<td>522</td>
<td>600</td>
</tr>
<tr>
<td>5</td>
<td>405</td>
<td>426</td>
<td>513</td>
<td>513</td>
</tr>
<tr>
<td>6</td>
<td>324</td>
<td>438</td>
<td>507</td>
<td>539</td>
</tr>
<tr>
<td>7</td>
<td>310</td>
<td>312</td>
<td>410</td>
<td>456</td>
</tr>
<tr>
<td>8</td>
<td>326</td>
<td>326</td>
<td>350</td>
<td>504</td>
</tr>
<tr>
<td>9</td>
<td>375</td>
<td>447</td>
<td>547</td>
<td>548</td>
</tr>
<tr>
<td>10</td>
<td>286</td>
<td>286</td>
<td>403</td>
<td>422</td>
</tr>
<tr>
<td>11</td>
<td>349</td>
<td>382</td>
<td>473</td>
<td>497</td>
</tr>
<tr>
<td>12</td>
<td>429</td>
<td>410</td>
<td>488</td>
<td>547</td>
</tr>
<tr>
<td>13</td>
<td>348</td>
<td>377</td>
<td>447</td>
<td>514</td>
</tr>
<tr>
<td>14</td>
<td>412</td>
<td>473</td>
<td>472</td>
<td>446</td>
</tr>
<tr>
<td>15</td>
<td>347</td>
<td>326</td>
<td>455</td>
<td>468</td>
</tr>
<tr>
<td>16</td>
<td>434</td>
<td>458</td>
<td>637</td>
<td>524</td>
</tr>
<tr>
<td>17</td>
<td>364</td>
<td>367</td>
<td>432</td>
<td>469</td>
</tr>
<tr>
<td>18</td>
<td>420</td>
<td>395</td>
<td>508</td>
<td>531</td>
</tr>
<tr>
<td>19</td>
<td>397</td>
<td>556</td>
<td>645</td>
<td>625</td>
</tr>
</tbody>
</table>
Denote the mean response under treatment $j$ as $\mu_j$, $j = 1, ..., 4$.

**Hypotheses:**

1. Null effect: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

2. Main effect of halothane. $H_0 : (\mu_1 + \mu_2) - (\mu_3 + \mu_4) = 0$

3. Main effect of CO$_2$. $H_0 : (\mu_1 + \mu_3) - (\mu_2 + \mu_4)$

4. Interaction of CO$_2$ and halothane. $H_0 : (\mu_1 + \mu_4) - (\mu_2 + \mu_3) = 0$

5. In general, no treatment effect can be written as $H_0 : C\mu = 0$ for a $q \times 4$ matrix $C$

In class: Write down the matrix $C$ for points 2–4 above.
In general, assume that multiple competing treatments are assigned to the same experimental unit for many units and of interest is to formally assess whether there is significant difference between the two corresponding mean responses. Specifically, for the $i$th experimental unit denote by $X_{i1}$ the scalar response for the first treatment, $X_{i2}$ the response for the second treatment, and so on for $p$ competing treatments. Here $i$ indexes the experimental unit, $i = 1, \ldots, n$.

Let $X_i = (X_{i1}, \ldots, X_{ip})^T$ be the vector of responses for the $i$th experimental unit. It is assumed that the sample $\{X_1, \ldots, X_n\}$ is IID from a $p$-variate normal distribution with some unknown mean vector $\mu$ and unknown covariance matrix. The parameter of interest is $C\mu$, for some appropriately defined $q \times p$ matrix $C$.

### A 100(1 - $\alpha$)% confidence region for $C\mu$:

\[
    n(C\bar{X} - C\mu)^T(CSC^T)^{-1}(C\bar{X} - C\mu) \leq \frac{(n - 1)q}{n - q} F_{q,n-q}(\alpha);
\]

here $\bar{X}$ and $S$ are the sample mean and covariance of $X_i$'s.

### 100(1 - $\alpha$)% simultaneous confidence intervals are

\[
    c^T\bar{X} \pm \sqrt{\frac{(n - 1)q}{n - q} F_{q,n-q}(\alpha)} \sqrt{c^TSc/n}
\]

will contain $c^T\mu$ simultaneously for all the columns $c$ of the matrix $C$ with probability $1 - \alpha$.

### Testing $H_0 : C\mu = 0$

is done using the test statistic:

\[
    T^2 := n(C\bar{X})^T(CSC^T)^{-1}C\bar{X}.
\]

If the null hypothesis is true then $T^2 \sim \frac{(n-1)q}{n-q} F_{q,n-q}$. Thus we reject $H_0$ at level $\alpha$ if the observed value of $T^2$ exceeds the critical value $\frac{(n-1)q}{n-q} F_{q,n-q}(\alpha)$.

Remark: The above results are presented for the case that the matrix $C$ has $q$ rows. In the previous data example, the matrix has $p - 1$ rows (each row corresponding to one contrast). In that case, the analogous results are as the above with $q$ replaced by $p - 1$. 
3 Inference for two populations: Confidence intervals. Hypothesis testing.

The ideas of the previous part can be used to develop procedures for comparison of two or more population mean vectors. We will first study such comparisons when the data arise from a paired design (measurements before/after), then when the data arises from multiple treatments, and finally when we have multiple independent populations.

3.1 Paired comparison

Wastewater monitoring example (Johnson and Wichern, 2007). Municipal wastewater treatment plants are required by law to monitor their discharge into rivers and streams on a regular basis. Concerns about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to labs for testing. One half of each sample was sent to the Wisconsin State Laboratory of Hygiene and one half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for \( n = 11 \) sample splits from the two laboratories.

Goal: compare commercial lab and state lab.

<table>
<thead>
<tr>
<th>Commercial BOD</th>
<th>Commercial SS</th>
<th>StateLabHygiene BOD</th>
<th>StateLabHygiene SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>27</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>18</td>
<td>64</td>
<td>36</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>34</td>
<td>75</td>
<td>44</td>
<td>64</td>
</tr>
<tr>
<td>28</td>
<td>26</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>71</td>
<td>124</td>
<td>54</td>
<td>64</td>
</tr>
<tr>
<td>43</td>
<td>54</td>
<td>34</td>
<td>56</td>
</tr>
<tr>
<td>33</td>
<td>30</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>39</td>
<td>21</td>
</tr>
</tbody>
</table>
Assume two competing treatments are assigned to the same experimental unit for many units and of interest is to formally assess whether there is significant difference between the two corresponding mean responses.

Specifically, for the $i^{th}$ experimental unit denote by $X_{1i}$ the response for the first treatment and $X_{2i}$ the response for the second treatment, where $i = 1, \ldots, n$. Denote by $\mu_1 = E[X_{1i}]$ and let $\mu_2 = E[X_{2i}]$. We want to draw inferences about $\mu_1 - \mu_2 = E[X_{1i} - X_{2i}]$.

Approach: calculate pairwise differences $d_i = X_{1i} - X_{2i}$. Assume that $d_1, \ldots, d_n$ is an IID sample from the multivariate normal distribution (or that the sample size is very large) with mean $\delta$ and some unknown covariance $\Sigma$. Of interest is to draw inference about $\delta = \mu_1 - \mu_2$.

<table>
<thead>
<tr>
<th>Paired test for $H_0 : \mu_1 = \mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>To test the null hypothesis $H_0 : \delta = 0_p$ vs $H_1 : \delta \neq 0_p$, we calculate the $T^2$ test statistic</td>
</tr>
<tr>
<td>$T^2 := n\bar{d}^T S_d^{-1} \bar{d}$;</td>
</tr>
<tr>
<td>here $\bar{d}$ and $S_d$ are the sample mean and sample covariance of $d_i$’s. If the null hypothesis is true then $T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$. Thus we reject $H_0$ at level $\alpha$ if the observed value of $T^2$ exceeds the critical value $\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Confidence region for the mean difference, $\mu_1 - \mu_2$, is</th>
</tr>
</thead>
<tbody>
<tr>
<td>${ \text{all } \delta \in \mathbb{R}^p \text{ such that } n(\bar{d} - \delta)^T S_d^{-1}(\bar{d} - \delta) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) }$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simultaneous confidence intervals for the components of $\mu_1 - \mu_2$ are:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{d}<em>k \pm \sqrt{\frac{(n-1)p}{n-p} F</em>{p,n-p}(\alpha)} \sqrt{s_{d,kk}/n}, \quad k = 1, \ldots, p$;</td>
</tr>
<tr>
<td>here $d_k$ is the $k^{th}$ element of the mean difference vector $\bar{d}$ and $s_{d,kk}$ is the $k^{th}$ element of the diagonal of $S_d$.</td>
</tr>
</tbody>
</table>

| When the sample size is large, then $\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$ is replaced by $\chi^2_p(\alpha)$. |

<table>
<thead>
<tr>
<th>Bonferroni-corrected confidence intervals for the components of $\mu_1 - \mu_2$ are:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{d}<em>k \pm t</em>{n-1} \left( \frac{\alpha}{2p} \right) \sqrt{s_{d,kk}/n}, \quad k = 1, \ldots, p$,</td>
</tr>
<tr>
<td>Example: in class demonstration using the wastewater data.</td>
</tr>
</tbody>
</table>
3.2 Independent populations comparison

**Electricity consumption example** (Johnson and Wichern, 2007). A total of $n_1 = 45$ and $n_2 = 55$ homes with and without air conditioning respectively are measured in Wisconsin. Two measurements of electrical usage (in kilowatt hours) were considered. The first is a measure of total on-peak consumption ($X_1$) during July and the second is a measure of total off-peak consumption ($X_2$) during July. The resulting summaries are:

\[
\bar{x}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix} \quad S_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} \\
\bar{x}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix} \quad S_1 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}
\]

**Goal:** inference on the difference of electricity consumption with and without air conditioning.

<table>
<thead>
<tr>
<th>Review of univariate two sample t test:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Setup: $X_{11}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$; $X_{21}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$</td>
</tr>
<tr>
<td>• $H_0: \mu_1 - \mu_2 = \delta_0$</td>
</tr>
<tr>
<td>• Assumptions: independent populations; equal variance $\sigma_1^2 = \sigma_2^2$; normal distributions</td>
</tr>
<tr>
<td>• Test statistic:</td>
</tr>
<tr>
<td>$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}}$ where $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$</td>
</tr>
<tr>
<td>where $S_1^2$ and $S_2^2$ are sample variances for the two samples.</td>
</tr>
<tr>
<td>• Distribution under $H_0$: $t \sim t_{n_1+n_2-2}$</td>
</tr>
</tbody>
</table>

The multivariate procedures rely on the direct extension of this result to higher dimensions.
Hotelling’s $T^2$ to test $H_0 : \mu_1 = \mu_2$

- Setup: $X_{11}, \ldots, X_{1n_1} \sim N_p(\mu_1, \Sigma_1)$; $X_{21}, \ldots, X_{2n_2} \sim N_p(\mu_2, \Sigma_2)$
- $H_0 : \mu_1 - \mu_2 = \delta_0$ versus $H_1 : \mu_1 - \mu_2 \neq \delta_0$
- Assumptions: independent populations; equal covariance $\Sigma_1 = \Sigma_2$; normal distributions
- Test statistic:
  $$T^2 = \left\{ \bar{X}_1 - \bar{X}_2 \right\}^T S_p \left\{ \left( \frac{1}{n_1} \right) \left( \frac{1}{n_2} \right) \right\}^{-1} \left\{ \bar{X}_1 - \bar{X}_2 \right\}$$
  where $S_p := \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$
- Distribution under $H_0$:
  $$T^2 \sim \chi^2_p$$

What if $\Sigma_1 \neq \Sigma_2$?

- When $n_1 - p$ and $n_2 - p$ are both large, we define the test statistic as
  $$T^2 = \left\{ \bar{X}_1 - \bar{X}_2 \right\}^T \left\{ \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right\}^{-1} \left\{ \bar{X}_1 - \bar{X}_2 \right\}$$
  Under $H_0$ we have $T^2 \sim \chi^2_p$ approximately.

- When sample sizes are not large, but the two populations are multivariate normal, then
  $$T^2 \sim \frac{vp}{v-p+1} F_{p,v-p+1}$$
  approximately, where $v$ is
  $$v = \sum_{i=1}^2 \frac{1}{n_i} \left( tr\left[ \left( \frac{1}{n_i} S_i \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right)^2 \right] + tr\left[ \frac{1}{n_i} S_i \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right]^2 \right)^2.$$  

Some comments:

- If the sample size is moderate to large, Hotelling’s $T^2$ is is still quite robust even if there are slight departures from normality and/or a few outliers are present.
3.3 Independent populations comparison (for variables that have similar values). Profile Analysis

High School data example. School A and School B are two rival traditional middle schools in San Francisco. A researcher wants to get a sense of the level of preparation of the 7th graders in the two schools. She collects a random sample of students from each school and records the students' scores in the four core subjects: Math, ELA, Science and Social studies. The researcher is interested in the following questions:

• Is the students' typical performance similar across the two schools?

• Is it true that the students perform at the same level for all the subjects on average, in the two schools?

• If the previous question is correct is the typical performance the same in the two schools?

<table>
<thead>
<tr>
<th>ID</th>
<th>Math</th>
<th>ELA</th>
<th>Science</th>
<th>SocialStudies</th>
<th>school</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78.80</td>
<td>81.26</td>
<td>77.81</td>
<td>91.92</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>96.64</td>
<td>90.50</td>
<td>87.93</td>
<td>100.00</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>79.43</td>
<td>79.09</td>
<td>84.19</td>
<td>87.05</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>85.19</td>
<td>80.48</td>
<td>85.48</td>
<td>89.87</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>73.96</td>
<td>81.50</td>
<td>76.13</td>
<td>91.08</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>79.42</td>
<td>77.55</td>
<td>78.05</td>
<td>88.53</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>86.54</td>
<td>84.08</td>
<td>87.67</td>
<td>92.36</td>
<td>B</td>
</tr>
<tr>
<td>102</td>
<td>79.77</td>
<td>82.25</td>
<td>78.14</td>
<td>93.04</td>
<td>B</td>
</tr>
<tr>
<td>103</td>
<td>80.40</td>
<td>83.84</td>
<td>80.43</td>
<td>93.20</td>
<td>B</td>
</tr>
<tr>
<td>104</td>
<td>87.76</td>
<td>83.38</td>
<td>80.93</td>
<td>95.66</td>
<td>B</td>
</tr>
<tr>
<td>105</td>
<td>87.39</td>
<td>83.38</td>
<td>82.07</td>
<td>94.82</td>
<td>B</td>
</tr>
<tr>
<td>106</td>
<td>89.29</td>
<td>82.61</td>
<td>82.68</td>
<td>94.56</td>
<td>B</td>
</tr>
</tbody>
</table>

In class: Formulate mathematically the goal of the problem.
General setup:

- $p$ variables (whose values are expressed in similar units, such as, questions, tests administered) given to two groups/populations
- $X_{1j} = (X_{1j1}, \ldots, X_{1jp})^T$ - the response for the $j$th unit in the first group/population
- $X_{2k} = (X_{2k1}, \ldots, X_{2kp})^T$ - the response for the $k$th unit in the second group/population
- Assumption: independent samples, $X_{1j} \sim N_p(\mu_1, \Sigma)$, $X_{2k} \sim N_p(\mu_2, \Sigma)$
- The sample sizes for the groups are $n_1$ and $n_2$, respectively
- Two group mean vectors $\mu_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1p})^T$ and similarly define $\mu_2 = (\mu_{21}, \mu_{22}, \ldots, \mu_{2p})^T$

Questions of interest:

- Stage 1: are the two profiles parallel (change between the components of the mean vector is similaat across the groups)?
  \[ \mu_{1i} - \mu_{1,i-1} = \mu_{2i} - \mu_{2,i-1}, \quad i = 2, 3, \ldots, p \]

- Stage 2: If parallel, are the two profiles coincident?
  \[ \mu_{1i} = \mu_{2i}, \quad i = 1, 2, \ldots, p \]

- Stage 3: If coincident, are the means equal?
  \[ \mu_{11} = \ldots = \mu_{1p} = \mu_{21} = \ldots = \mu_{2p} \]

In class: Express (the null hypothesis implied by) each of these questions in form of $C \mu_1$ and $C \mu_2$. 

17
Stage 1 (parallel profiles):

- $H_0 : C\mu_1 = C\mu_2$ for appropriate $(p - 1) \times p$ dimensional matrix $C$
- $T^2 = \{C(\bar{x}_1 - \bar{x}_2)\}^T \left\{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)CS_{pool}C^T\right\}^{-1} C(\bar{x}_1 - \bar{x}_2)$
- Under $H_0$, the test statistic follows $\frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p} F_{p-1,n_1+n_2-p}$ distribution

Hint: use the result for the equality of two mean vectors from independent populations $CX_{1j}$ and $CX_{2k}$.

Stage 2 (coincident profiles, given they are parallel):

- $H_0 : \Gamma^T \mu_1 = \Gamma^T \mu_2$, where $\Gamma = (1, \ldots, 1)^T$
- $T^2 = \{\Gamma(\bar{x}_1 - \bar{x}_2)\}^T \left\{(\frac{1}{n_1} + \frac{1}{n_2})\Gamma S_{pool}\Gamma\right\}^{-1} \Gamma(\bar{x}_1 - \bar{x}_2)$
- Under $H_0$, the test statistic follows $F_{1,n_1+n_2-2}$ distribution

Stage 3 (constant profiles, given they are coincident):

- Combine two populations together, assume common mean $\mu$
- $H_0 : C\mu = 0$
- $T^2 = (C\bar{x})^T \left\{\frac{1}{n_1+n_2} CSC^T\right\}^{-1} C\bar{x}$, where $S$ is the covariance matrix based on all $n_1 + n_2$ observations
- Under $H_0$, the test statistic follows $\frac{(n_1+n_2-1)(p-1)}{n_1+n_2-p+1} F_{p-1,n_1+n_2-p+1}$ distribution

Hint: When the means of the two groups are the same, then we can use the overall sample mean as an estimator for the common mean; the pooled sample covariance becomes the sample covariance of the full sample (obtained by combining the two groups together).

Note that Stage 3 is appropriate only if all the questions are measured on the same scale.

In class: R code to analyze the motivating data application.