Notes On
Applied
Multivariate
Analysis
Introduction

“Multivariate” refers to the presence of multiple random variables.

So Multivariate Data are data that are thought of as the realizations of several random variables,

and Multivariate Analysis can be defined broadly as an inquiry into the structure of interrelationships among multiple random variables.

Haven’t we already studied that?
(e.g., Regression, Anova)
Regression:

Simple: explain variability in $Y$ from $X$.

Multiple: explain variability in $Y$ from $X_1, X_2, \ldots, X_r$.

Anova:

One-way: explain variability in $Y$ based on $X$, where $X$ is a factor

These examples are not usually considered to be multivariate analyses.

Why?

We distinguish between the Response Variable ($Y$) and Explanatory Variables ($X$’s) and in the analysis we treat the $X$’s as fixed.

$\Rightarrow$ only 1 random variable in each situation.

(multiple regression $\neq$ multivariate regression)
One type of analysis that is a good example of a multivariate method is **correlation analysis**.

Recall:

**Correlation Analysis:** measure linear association between $X, Y$ without distinguishing response versus explanatory.

Other examples:

**M’variate 1 & 2 Samp. Inference:** Hypothesis tests and confidence intervals for the mean vector(s) from one (or two) group(s). Procedures analogous to $t$ and $Z$ in univariate setting.

**MANOVA:** Acronym for multivariate analysis of variance. Inference for mean vectors from several populations.

**Multivariate Linear Regression:** Explain variability in response variables $Y_1, Y_2, \ldots, Y_m$ based on explanatory variables $X_1, \ldots, X_r$. 
Principal Components Analysis: Reduce the dimension of a data set by examining a small number of linear combinations of the original variables that explain most of the variability among the original variables.

Factor Analysis: Similar to P.C.A. Describe the variance/covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors.

Canonical Correlation Analysis: Determine the linear association between two sets of random variables. Find a linear combination of several $X$’s and a linear combination of several $Y$’s such that the two linear combinations have maximum correlation.

Classification and Discrimination: Methods for separating distinct sets of objects (experimental units) and with allocating (classifying) objects to previously defined groups.
Cluster Analysis: Methods for grouping objects based on distances (which objects are most similar to which others). Groups are not predefined and no assumptions are made concerning the number of groups or the group structure.

Variable- vs. Individual-Directed Methods


Individual-directed Techniques: primarily concerned with relationships among the among the experimental units or individuals being measured. E.g., Discrim. Anal. Clustering, MANOVA.

- Some methods can be used as either variable-directed or individual-directed. E.g., clustering can be done to group individuals or to group variables.

- Multivariate methods also differ with respect to how informal or exploratory their typical uses are. E.g., cluster analysis, P.C.A. are often preliminary or exploratory tools used as part of a larger analysis.
Theory vs. Methods:

(1) problem: estimate population mean $\mu$
(given $x_1, \ldots, x_n$)

model: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
$\mu, \sigma^2$ unknown

theory: $\bar{X} \sim N(\mu, \sigma^2/n),$ 
$(n - 1)S^2 \sim \sigma^2 \chi^2(n - 1),$ 
$\bar{X}, S^2$ independent.  
$\Rightarrow \sqrt{n}(\bar{X} - \mu)/S \sim t(n - 1)$

method: $\bar{x} \pm t_{1-\alpha/2}(n - 1)s/\sqrt{n}$ forms a
$100(1 - \alpha)\%$ C.I. for $\mu$

application: $n = 20,$ $\alpha = 0.10$
$\bar{x} = 49.70,$ $s = 0.32$
$t_{.95}(19) = 1.729$
$90\%$ CI $= 49.70 \pm 1.729(0.32/\sqrt{20})$
$= (49.58, 49.82)$

(2) (problem $\rightarrow$ method $\rightarrow$ application) not suffi-
cient.
(3) validity of method depends on the validity of model. Is model valid? (need to know something about $\sim iid N(\mu, \sigma^2)$) If not valid, how is method affected? Alternatives?

(4) How do we interpret the results of our application? (need to understand what model means)
Preliminaries

1. Vector and matrix algebra.

2. Multivariate versions of basic statistical concepts.

3. The multivariate normal distribution.

Vector and Matrix Algebra

(See section 1.4 and chapter 2 of Johnson and Wichern.)

Definitions:

Matrix: a rectangular array of numbers with $n$ rows and $p$ columns

\[
A = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1p} \\
a_{21} & a_{22} & \cdots & a_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{np}
\end{pmatrix} = (a_{ij}) \quad \begin{cases} 
i = 1, \ldots, n \\
j = 1, \ldots, p
\end{cases}
\]

9
e.g.,

\[
A = \begin{pmatrix}
2 & -1 \\
1 & 0 \\
2 & 3
\end{pmatrix}. \quad (n = 3, p = 2)
\]

**Dimension of a Matrix:** the ordered pair \((n, p)\).
We say that \(A\) is \(n \times p\).

**Vector:** a matrix with \(p = 1\) column.

\[
x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix} \quad \text{e.g.,} \quad x = \begin{pmatrix}
2 \\
1 \\
2
\end{pmatrix}.
\]

Note: \(A = (a_1, a_2, \cdots, a_p)\). A "vector" in this course will always refer to a column array of numbers. Some authors define row vectors and column vectors separately.

**Scalar:** a real valued variable, or a \(1 \times 1\) matrix.
We will typically denote matrices with bold capitals, vectors with bold lower case, and scalars with regular type lower case.
Basic Operations:

**Transpose:** If \( A \) is \( n \times p \) then \( A' \) is the \( p \times n \) matrix by interchanging the rows and columns of \( A \). E.g.,

\[
A' = \begin{pmatrix}
2 & 1 & 2 \\
-1 & 0 & 3
\end{pmatrix}.
\]

If we want a row vector version of \( x \) then we use \( x' \) (= \((2 \ 1 \ 2)\) in the example).

A matrix with the property \( A = A' \) is called **symmetric**.

**Scalar Multiplication:** A scalar, \( c \), times a matrix \( A \) is given by

\[
cA = (ca_{ij}) \quad \text{e.g.,} \quad 2A = \begin{pmatrix}
4 & -2 \\
2 & 0 \\
4 & 6
\end{pmatrix}
\]
**Addition:** \( \mathbf{A} = (a_{ij}) \) added to \( \mathbf{B} = (b_{ij}) \) is given by

\[
\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})
\]

\( \mathbf{A} \) and \( \mathbf{B} \) must have the same dimension.
E.g., for \( \mathbf{A} \) as before and \( \mathbf{B} \) given by

\[
\mathbf{B} = \begin{pmatrix}
1 & 2 \\
-3 & 1 \\
0 & -2
\end{pmatrix}
\]
we get \( \mathbf{A} + \mathbf{B} = \begin{pmatrix}
3 & 1 \\
-2 & 1 \\
2 & 1
\end{pmatrix}\)

**Subtraction:** \( \mathbf{A} - \mathbf{B} = \mathbf{A} + (-1)\mathbf{B} \) (\( \mathbf{A}, \mathbf{B} \) of same dimension). E.g.,

\[
\mathbf{A} - \mathbf{B} = \begin{pmatrix}
1 & -3 \\
4 & -1 \\
2 & 5
\end{pmatrix}
\]

**Matrix Multiplication:** \( \mathbf{AB} \) is defined if and only if \( \mathbf{A} \) has the same number of columns as \( \mathbf{B} \) has rows; that is, if and only if \( \mathbf{A} \) and \( \mathbf{B} \) are **conformable.** \( \mathbf{AB} \) will have as many rows as \( \mathbf{A} \) and as many columns as \( \mathbf{B} \).
If $A_{m \times n} = (a_{ij})$ and $B_{n \times p} = (b_{ij})$ are conformable then the $m \times p$ matrix product $AB$ has $(i, j)^{th}$ element

$$\sum_{k=1}^{n} a_{ik} b_{kj}.$$

E.g., take $A_{3 \times 2}$ as before and let $B$ be given by

$$B_{2 \times 4} = \begin{pmatrix} 2 & 0 & 1 & -1 \\ -1 & 3 & 1 & 2 \end{pmatrix}$$

Then $AB$ is a $3 \times 4$ matrix given by

$$AB = \begin{pmatrix} 2 & -1 \\ 1 & 0 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 & -1 \\ -1 & 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & -3 & 1 & -4 \\ 2 & 0 & 1 & -1 \\ 1 & 9 & 5 & 4 \end{pmatrix}$$
Special cases:

1. Matrix times a vector:

\[ \mathbf{Ax} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}_{n \times p} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}_{p \times 1} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{np}x_p \end{pmatrix}_{n \times 1} = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \cdots + x_p \mathbf{a}_p = \sum_{i=1}^{p} x_i \mathbf{a}_i \]

a linear combination of the columns of \( \mathbf{A} \).
2. Dot product or inner product:

\[ x' y = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \sum_{i=1}^{p} x_i y_i \]

Sometimes written \( x \cdot y \) or \( < x, y > \). Note: \( x \cdot x = \sum_{i=1}^{n} x_i^2 \) (sum of squared elements).

3. Column Vector times row vector:

\[ xy' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \cdots & y_p \end{pmatrix} = (x_i y_j)_{p \times p} \]

3.a. Vector times its transpose:

\[ xx' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix} = (x_i x_j)_{p \times p} \]
Properties:

1. \( A + B = B + A. \)
2. \( (A + B) + C = A + (B + C). \)
3. \( c(A + B) = cA + cB. \)
4. \( (c + d)A = cA + dA. \)
5. \( (cd)A = c(dA). \)
6. \( c(AB) = (cA)B. \)
7. \( A(BC) = (AB)C. \)
8. \( A(B + C) = AB + AC. \)
9. \( (A + B)C = AC + BC. \)
10. \( (A + B)' = A' + B'. \)
11. \( (cA)' = cA'. \)
12. \( (AB)' = B'A'. \)
13. \( AB \neq BA. \) E.g.,

\[
A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}
\]
\[
AB = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}, \quad BA = \begin{pmatrix} 0 & 2 \\ 1 & 5 \end{pmatrix}
\]

\( BA \) may not be defined; if it is, \( AB \) and \( BA \) may not have the same dimension; if \( \dim(AB) = \dim(BA) \), doesn’t imply that \( AB = BA. \)
Geometrical Concepts:

**Vector:** a line segment from the origin (0, all coordinates equals to zero) to the point indicated by the coordinates of the (algebraic) vector.

\[ \mathbf{x} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} \]

**Vector Addition:**
Scalar Multiplication: Multiplication by a scalar scales a vector by shrinking or extending the vector in the same direction.

Length (Norm) of Vector: $L_x = (x'x)^{1/2}$ is the length, or norm of $x$. Sometimes written as $||x||$. $L_x^{-1}x$ is the vector of length 1 which lies in the direction of $x$.

Angle Between Vectors: If $\theta$ is the angle between two vectors $x$ and $y$, then

$$\cos(\theta) = \frac{x'y}{L_xL_y}.$$
Orthogonal (Perpendicular) Vectors: Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to one another if and only if $\mathbf{x}'\mathbf{y} = 0$ ($\Rightarrow$ $\cos(\theta) = 0 \Rightarrow \theta = \pi/2$).

Projection: The projection of $\mathbf{x}$ on $\mathbf{y}$ is the vector in the direction of $\mathbf{y}$ that is closest to $\mathbf{x}$ and is given by

$$\frac{x'y}{y'y}y = \frac{x'y}{L_y}L_y^{-1}y.$$
**Euclidean Distance:** Since a vector $\mathbf{x}$ has the interpretation of a directed line segment from the origin $(0)$ to the point $(x_1, x_2, \ldots, x_p)$, the Euclidean (straight-line) distance between $0$ and $\mathbf{x}$ is clearly the length of the vector, $L_\mathbf{x}$:

$$d(\mathbf{x}, 0) = L_\mathbf{x} = \sqrt{x_1^2 + \cdots + x_p^2}.$$ 

More generally, the Euclidean distance between two arbitrary points $\mathbf{x}$ and $\mathbf{y}$ is

$$d(\mathbf{x}, \mathbf{y}) = L_{\mathbf{x}-\mathbf{y}} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}.$$ 

• Many ways to define distance. Euclidean distance is just one (familiar) example.
• Properties of a distance measure $d$:

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$$

$$d(\mathbf{x}, \mathbf{y}) \geq 0, \text{ with equality when } \mathbf{x} = \mathbf{y}$$

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}) \text{ (triangle inequality)}$$

• We’ll consider another distance measure, statistical distance, later.
More Matrix Functions and Definitions:

**Determinant:** The determinant of a square matrix $A$ (same number of rows as columns) with dimension $k \times k$ is the scalar denoted as $|A|$ given by

\[
|A| = \begin{cases} 
  a_{11} & \text{if } k = 1 \\
  \sum_{j=1}^{k} a_{1j}|A_{1j}|(-1)^{1+j} & \text{if } k > 1,
\end{cases}
\]

where $A_{1j}$ is the $(k - 1) \times (k - 1)$ matrix obtained by deleting the first row and the $j^{th}$ column of $A$. Note that $|A|$ can be defined as $\sum_{j=1}^{k} a_{ij}|A_{ij}|(-1)^{i+j}$ using the $i^{th}$ row in place of the first.

Special Cases:

$k = 2$:

\[
\begin{vmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22}
\end{vmatrix} = a_{11}a_{22}(-1)^2 + a_{12}a_{21}(-1)^3 = a_{11}a_{22} - a_{12}a_{21}
\]

E.g.,

\[
\begin{vmatrix}
  3 & 1 \\
  -2 & 4
\end{vmatrix} = 12 - (-2) = 14
\]
\( k = 3: \)

\[
\begin{vmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{vmatrix} = +a_{11}
\begin{vmatrix}
  a_{22} & a_{23} \\
  a_{32} & a_{33}
\end{vmatrix} - a_{12}
\begin{vmatrix}
  a_{21} & a_{23} \\
  a_{31} & a_{33}
\end{vmatrix} \\
  \quad + a_{13}
\begin{vmatrix}
  a_{21} & a_{22} \\
  a_{31} & a_{32}
\end{vmatrix}
\]

\[= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})\]

Properties:

1. \(|cA| = c^k|A|\), for \(A_{k \times k}\).
2. \(|A'| = |A|\).
3. \(|AB| = |A||B|\).
4. For a diagonal matrix \(D_{k \times k}\) of the form
   \[
   D = \begin{pmatrix}
   d_{11} & 0 & \cdots & 0 \\
   0 & d_{22} & \cdots & 0 \\
   \vdots & \vdots & \ddots & \vdots \\
   0 & 0 & \cdots & d_{kk}
   \end{pmatrix},
   \]
   \[|D| = d_{11}d_{22} \cdots d_{kk}|.
Linear Dependence: A set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is said to be linearly dependent if there exist constants $c_1, \ldots, c_k$ not all zero such that

$$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_k \mathbf{x}_k = \mathbf{0}$$

- Recall collinearity/multicollinearity complications occurred in regression when the explanatory variables were linearly dependent; that is, when the columns of the $\mathbf{X}$ matrix were linearly dependent.
- A set of vectors which is not linearly dependent is said to be **linearly independent**.

$|\mathbf{A}| = 0$ if and only if there exists a nonzero vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{0}$; that is, if and only if the columns of $\mathbf{A}$ are linearly dependent.
Identity Matrix: A matrix with ones on the diagonal and zeros elsewhere is called an identity matrix and is denoted by $I_{k \times k}$ or $I$ when the dimension is clear from the context. E.g.,

$$I_{3 \times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

$I$ is the “identity” matrix because

$$AI = IA = A$$

Matrix Inverse: Matrix division is defined as multiplication by the matrix inverse. Just as $a/b$ can be defined as $ab^{-1}$ for scalars $a$ and $b$, we will think of $A$ divided by $B$ as $AB^{-1}$ where $B^{-1}$ is that matrix such that

$$BB^{-1} = I$$

(just as $b^{-1} = 1/b$ is that scalar such that $bb^{-1} = 1$).

- The matrix inverse is only defined for square matrices that have nonzero determinants.
- If $A$ (square) has $|A| = 0$ then its columns are linearly dependent and it has no inverse. Such a matrix is called singular.
Computation of Matrix Inverse: $A^{-1}$ has $(i, j)^{th}$ element $(-1)^{i+j} |A_{ji}| / |A|$, where $A_{ji}$ is the matrix obtained by deleting the $j^{th}$ row and $i^{th}$ column of $A$. E.g.,

$$
\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}
$$

We will rely on the computer to compute inverses for matrices with dimension $> 2$.

Properties:
1. $(A^{-1})' = (A')^{-1}$.
2. $(AB)^{-1} = B^{-1}A^{-1}$.
3. $|A^{-1}| = 1/|A|$. 
Eigenvalues and Eigenvectors: A square $k \times k$ matrix $A$ is said to have an eigenvalue $\lambda$, with a corresponding eigenvector $x \neq 0$ if

$$Ax = \lambda x \quad (1)$$

Since (1) can be written

$$(A - \lambda I)x = 0$$

for some nonzero vector $x$, the eigenvalues of $A$ are the solutions of

$$|A - \lambda I| = 0$$

If we look at how a determinant is calculated it is not difficult to see that $|A - \lambda I|$ will be a polynomial in $\lambda$ of order $k$ so there will be $k$ (not necessarily real) eigenvalues (solutions).

1. if $|A| = 0$ then $Ax = 0$ for some nonzero $x$. That is, if the columns of $A$ are linearly dependent then $\lambda = 0$ is an eigenvalue of $A$. 
2. The eigenvector associated with a particular eigenvalue is unique only up to a scale factor. That is, if \( A\mathbf{x} = \lambda \mathbf{x} \), then \( A(c\mathbf{x}) = \lambda (c\mathbf{x}) \) so \( \mathbf{x} \) and \( c\mathbf{x} \) are both eigenvectors for \( A \) corresponding to \( \lambda \). We typically normalize eigenvectors to have length 1 (choose the \( \mathbf{x} \) that has the property \( \mathbf{x}'\mathbf{x} = 1 \)).

3. If \( \lambda_i \) and \( \lambda_j \) are distinct eigenvalues of the symmetric matrix \( A \) then their associated eigenvectors \( \mathbf{x}_i, \mathbf{x}_j \), are orthogonal.
4. Computation: By computer typically, but for 2 \times 2 case we have

\[
\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \Rightarrow \quad \mathbf{A} - \lambda \mathbf{I} = \begin{pmatrix} a - \lambda & b \\ c & d - \lambda \end{pmatrix}
\]

\[
|\mathbf{A} - \lambda \mathbf{I}| = 0 \quad \Rightarrow \quad (a - \lambda)(d - \lambda) - bc = \lambda^2 - (a + d)\lambda + (ad - bc) = 0
\]

\[
\Rightarrow \quad \lambda = \frac{(a + d) \pm \sqrt{(a + d)^2 - 4(ad - bc)}}{2}
\]

To obtain associated eigenvector solve \( \mathbf{A}\mathbf{x} = \lambda\mathbf{x} \). There are infinitely many solutions for \( \mathbf{x} \) so choose one by setting \( x_1 = 1 \), say, and solve for \( x_2 \). Normalize to have length one by computing \( L_x^{-1}\mathbf{x} \).
**Orthogonal Matrices:** We say that $\mathbf{A}$ is orthogonal if $\mathbf{A}' = \mathbf{A}^{-1}$ or, equivalently, $\mathbf{A} \mathbf{A}' = \mathbf{A}' \mathbf{A} = \mathbf{I}$, so that the columns (and rows) of $\mathbf{A}$ all have length 1 and are mutually orthogonal.

**Spectral Decomposition:** If $\mathbf{A}_{k \times k}$ is a symmetric matrix (i.e., $\mathbf{A} = \mathbf{A}'$) then it can be written (decomposed) as follows:

$$
\mathbf{E}_{k \times k} \Lambda_{k \times k} \mathbf{E}'_{k \times k},
$$

where

$$
\Lambda = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_k
\end{pmatrix}
$$

and $\mathbf{E}$ is an orthogonal matrix with columns $e_1, e_2, \ldots, e_k$ the eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$.

We haven’t yet talked about how to interpret eigenvalues and eigenvectors, but the spectral decomposition says something about the significance of these quantities (why we’re interested in them): the “information” in $\mathbf{A}$ can be broken apart into its eigenvalues and a set of orthogonal eigenvectors.
Results:

1. For $A_{k \times k}$ symmetric,

$$|A| = |E| \Lambda |E'| = |EE'| \Lambda = |I| \Lambda = \prod_{i=1}^{k} \lambda_i$$

I.e., the determinant of a symmetric matrix is the product of its eigenvalues.

2. If $A$ has eigenvalues $\lambda_1, \ldots, \lambda_k$, then $A^{-1}$ has the same associated eigenvectors and eigenvalues $1/\lambda_1, \ldots, 1/\lambda_k$ since

$$A^{-1} = (E \Lambda E')^{-1} = E \Lambda^{-1} E'.$$
3. Similarly, if \( A \) has the additional property that it is \textbf{positive definite} (defined below) a \textbf{square root matrix}, \( A^{1/2} \), can be obtained with the property \( A^{1/2}A^{1/2} = A \):

\[
A^{1/2} = EE^{1/2}E', \text{ where }
\]

\[
\Lambda^{1/2} = \begin{pmatrix}
\sqrt{\lambda_1} & 0 & \cdots & 0 \\
0 & \sqrt{\lambda_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sqrt{\lambda_k}
\end{pmatrix}
\]

\( A^{1/2} \) is symmetric, has eigenvalues that are the square roots of the eigenvalues of \( A \), and has the same associated eigenvectors as \( A \).

**Quadratic Forms:** For a symmetric matrix \( A_{k \times k} \), a quadratic form in \( x_{k \times 1} \) is defined by

\[
x'Ax = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i x_j.
\]

(\( x'Ax \) is a sum in squared (quadratic) terms, \( x_i^2 \), and \( x_i x_j \) terms.)
Positive Definite Matrices: \( A \) is positive definite (p.d.) if \( \mathbf{x}'A\mathbf{x} > 0 \) for all \( \mathbf{x} \neq \mathbf{0} \). If \( \mathbf{x}'A\mathbf{x} \geq 0 \) for all nonzero \( \mathbf{x} \) then \( A \) is positive semi-definite (p.s.d.) or, sometimes, non-negative definite.

- If \( A_{k \times k} \) is p.d. then for \( i = 1, \ldots, k \), \( \mathbf{e}_i'A\mathbf{e}_i = \lambda_i(\mathbf{e}_i'|\mathbf{e}_i) > 0 \);

- i.e., the eigenvalues of \( A \) are all positive.

- It can also be shown that if the eigenvalues of \( A \) are all positive then \( A \) is p.d.
Statistical Distance:

Consider points $\mathbf{P}$ and $\mathbf{Q}$ in the above scatter-plot.

- Euclidean Distance: $d(\mathbf{0}, \mathbf{P}) > d(\mathbf{0}, \mathbf{Q})$.

But suppose $X_1$ and $X_2$ represented random variables each with mean 0, but with different variances ($\text{var}(X_1) > \text{var}(X_2)$). If we are interested in how unusual the points $\mathbf{P}$ and $\mathbf{Q}$ are then we want to know how far from the average (0) these points are. For this purpose it does not make sense that $d(\mathbf{0}, \mathbf{P}) > d(\mathbf{0}, \mathbf{Q})$.

- Euclidean distance does not account for variability inherent in the random variables representing the coordinates.
• Need a distance measure where coordinates are weighted differentially and distance is relative to the variability of the data.

One solution, **Karl Pearson Distance**: Distance of a point \( \mathbf{P} = (x_1, x_2) \) from the origin \( \mathbf{0} = (0, 0) \) takes into account how many \( (X_1) \) standard deviations \( x_1 \) is from 0 and how many \( (X_2) \) standard deviations \( x_2 \) is from 0:

\[
d(0, \mathbf{P}) = \sqrt{\left( \frac{x_1}{s_1} \right)^2 + \left( \frac{x_2}{s_2} \right)^2},
\]

where \( s_1 \) is the standard deviation of \( X_1 \) and \( s_2 \) is the standard deviation of \( X_2 \). The generalization to \( k \)-dimensional space is straightforward:

\[
d(0, \mathbf{P}) = \sqrt{\left( \frac{x_1}{s_1} \right)^2 + \cdots + \left( \frac{x_k}{s_k} \right)^2}.
\]

• A problem: Karl Pearson distance doesn’t account for correlation among the coordinates.
An alternative: *Statistical Distance*: Karl Pearson Distance would be o.k. if it were computed with respect to the rotated axes $\tilde{X}_1$ and $\tilde{X}_2$. So, in two dimensions, we can define statistical distance as

$$d(0, P) = \sqrt{\left(\frac{\tilde{x}_1}{\tilde{s}_1}\right)^2 + \left(\frac{\tilde{x}_2}{\tilde{s}_2}\right)^2},$$

where

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$

$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta),$$

$\tilde{s}_1$, $\tilde{s}_2$ are standard deviations of $\tilde{X}_1$ and $\tilde{X}_2$, and $\theta$ is the angle of rotation.
With respect to the original coordinate system we have
\[
d(0, \mathbf{P}) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2},
\]
where \(a_{11}, a_{12}, \text{ and } a_{22}\) are coefficients depending on \(s_1, s_2, \text{cov}(X_1, X_2)\) and trigonometric functions of \(\theta\) (not important now).

Notice that \(d(0, \mathbf{P})\) as defined above can be written as
\[
d(0, \mathbf{P})^2 = (x_1 \quad x_2) \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{x}' \mathbf{A} \mathbf{x},
\]
a quadratic form in \(\mathbf{x}\) involving the symmetric matrix \(\mathbf{A}\).

- In fact, any positive definite quadratic form can be thought of as a squared statistical distance. The positive definiteness of the matrix \(\mathbf{A}\) is required to ensure the properties of a distance metric (non-negativity, triangle inequality, etc.).

**\(\mathbf{x}' \mathbf{A} \mathbf{x}\): Squared distance from \(\mathbf{x}\) to \(0\),

\[(\mathbf{x} - \mathbf{y})' \mathbf{A} (\mathbf{x} - \mathbf{y})\): Squared distance from \(\mathbf{x}\) to \(\mathbf{y}\).
Geometric Interpretation to Spectral Decomposition: Consider the set of points a constant (statistical) distance $c$ away from the origin. These points satisfy

$$x'Ax = c^2$$

$$\Rightarrow x'E\Lambda E'x = y'\Lambda y = \sum_i \lambda_i y_i^2 = c^2,$$

which is the equation of an ellipsoid in the transformed variable, $y = E'x$.

A transformation of this form, $y = E'x$, involving an orthogonal matrix $E$, defines a rotation from the $x$ coordinate system into the $y$ coordinate system which is rigid, in the sense that the orthogonality of the axes are maintained. The new axes are in the direction of the eigenvectors, $e_1, e_2, \ldots$. 
In two dimensions we have

$$
\lambda_1 y_1^2 + \lambda_2 y_2^2 = c^2,
$$

the equation of an ellipse, with \( y_1 = x'e_1 \) and \( y_2 = x'e_2 \). If we let \( x = c\lambda_1^{-1/2} e_1 \) we see that this value satisfies \( x'Ax = c^2 \):

$$
x'Ax = \lambda_1 (x'e_1)^2 + \lambda_2 (x'e_2)^2 \\
= \lambda_1 (c\lambda_1^{-1/2} e_1')^2 + \lambda_2 (c\lambda_1^{-1/2} e_2')^2 = c^2.
$$

Therefore, \( ||c\lambda_1^{-1/2} e_1|| = c/\sqrt{\lambda_1} \) gives the length of the ellipse along \( e_1 \). Similarly, \( c/\sqrt{\lambda_2} \) gives the length of the ellipse in the perpendicular direction along \( e_2 \).
One More Matrix Function:

**Trace of a Matrix:** Defined only for square matrices. Let $A_{k \times k} = (a_{ij})$ be a square matrix. The trace of $A$ is the sum of its diagonal elements:

$$\text{tr}(A) = \sum_{i=1}^{k} a_{ii}$$

**Properties:**

1. $\text{tr}(cA) = c\text{tr}(A)$.
2. $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$.
3. $\text{tr}(ABC) = \text{tr}(CAB)$ as long as the matrix products are defined (matrices are conformable) and $ABC$ and $CAB$ are square. In particular,

$$\text{tr}(B^{-1}AB) = \text{tr}(A)$$

and

$$\text{tr}(Ax x') = \text{tr}(\underbrace{x'Ax}_{\text{a scalar}}) = x'Ax.$$

4. $\text{tr}(AA') = \sum_i \sum_j a_{ij}^2$.
5. If $A_{k \times k}$ is symmetric with eigenvalues $\lambda_1, \ldots, \lambda_k$, then $\text{tr}(A) = \sum_{i=1}^{k} \lambda_i$. 

39
Basic Multivariate Statistical Concepts

(See sections 2.5, 2.6 and chapter 3 of Johnson and Wichern.)

Definitions:

**Random Vector**: A vector whose elements are random variables. E.g.,

\[ \mathbf{x}_{p \times 1} = (x_1 \ x_2 \ \cdots \ x_p)', \]

where \( x_1, \ldots, x_p \) are each random variables.

**Random Matrix**: A matrix whose elements are random variables. E.g., \( \mathbf{X}_{n \times p} = (x_{ij}) \), where \( x_{11}, x_{12}, \ldots, x_{np} \) are each random variables.

**Expected Value**: The expected value (population mean) of a random matrix (vector) is the matrix (vector) of expected values. For \( \mathbf{X}_{n \times p} \),

\[
\mathbf{E}(\mathbf{X}) = \begin{pmatrix}
E(x_{11}) & E(x_{12}) & \cdots & E(x_{1p}) \\
\vdots & \vdots & \ddots & \vdots \\
E(x_{n1}) & E(x_{n2}) & \cdots & E(x_{np})
\end{pmatrix} \equiv \mathbf{\mu}.
\]
• Recall, for a univariate random variable \( X \),

\[
E(X) = \begin{cases} 
\int_{-\infty}^{\infty} xf_X(x) \, dx & \text{if } X \text{ is continuous;} \\
\sum_{\text{all } x} x p_X(x) & \text{if } X \text{ is discrete.}
\end{cases}
\]

Here, \( f_X(x) \) is the probability density function of \( X \) in the continuous case, \( p_X(x) \) is the probability function of \( X \) in the discrete case.

(\textbf{Population) Variance-Covariance Matrix:} For a random vector \( \mathbf{x}_{p \times 1} = (x_1, x_2, \ldots, x_p)' \), the matrix

\[
\begin{pmatrix}
\text{var}(x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_p) \\
\text{cov}(x_2, x_1) & \text{var}(x_2) & \cdots & \text{cov}(x_2, x_p) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(x_p, x_1) & \text{cov}(x_p, x_2) & \cdots & \text{var}(x_p)
\end{pmatrix}
\]

\[
\equiv \begin{pmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}
\end{pmatrix}
\]

is called the variance-covariance matrix of \( \mathbf{x} \) and is denoted \( \text{var}(\mathbf{x}) \).
• Recall: for a univariate random variable \( x_i \) with expected value \( \mu_i \),

\[
\sigma_{ii} = \text{var}(x_i) = \mathbb{E}[(x_i - \mu_i)^2]
\]

• Recall: for univariate random variables \( x_i \) and \( x_j \),

\[
\sigma_{ij} = \text{cov}(x_i, x_j) = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]
\]

• \( \text{var}(\mathbf{x}) \) is symmetric because \( \sigma_{ij} = \sigma_{ji} \).

• In terms of vector/matrix algebra, \( \text{var}(\mathbf{x}) \) has formula

\[
\text{var}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'].
\]

• If the random variables \( x_1, \ldots, x_p \) in \( \mathbf{x} \) are mutually independent, then \( \text{cov}(x_i, x_j) = 0 \), when \( i \neq j \), and \( \text{var}(\mathbf{x}) \) is \textbf{diagonal} with \( (\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})' \) along the diagonal and zeros elsewhere.
(Population) Correlation Matrix: For a random variable $x_{p \times 1}$, the population correlation matrix is the matrix of correlations among the elements of $x$:

$$
corr(x) = \begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1p} \\
\rho_{21} & 1 & \cdots & \rho_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p1} & \rho_{p2} & \cdots & 1
\end{pmatrix},
$$

where $\rho_{ij} = corr(x_i, x_j)$.

- Recall: for random variables $x_i$ and $x_j$,

$$
\rho_{ij} = corr(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}}
$$

measures the amount of linear association between $x_i$ and $x_j$.

- Correlation matrices are symmetric.

- For a random vector $x_{p \times 1}$, let $\rho = corr(x)$, $\Sigma = \text{var}(x)$ and $V = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$. The relationship between $\rho$ and $\Sigma$ is

$$
\Sigma = V^{1/2} \rho V^{1/2}
$$

$$
\rho = (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}
$$
Properties:

Let $\mathbf{X}$, $\mathbf{Y}$ be random matrices of the same dimension and $\mathbf{A}$, $\mathbf{B}$ be matrices of constants such that $\mathbf{AXB}$ is defined

1. $\mathbf{E}(\mathbf{X} + \mathbf{Y}) = \mathbf{E}(\mathbf{X}) + \mathbf{E}(\mathbf{Y})$.
2. $\mathbf{E}(\mathbf{AXB}) = \mathbf{A}\mathbf{E}(\mathbf{X})\mathbf{B}$.

Now let $\mathbf{x}_{p \times 1}$ be a random vector with mean $\mu_\mathbf{x}$ and variance $\Sigma_\mathbf{x}$. Let $\mathbf{c}_{p \times 1}$ be a vector of constants, and $\mathbf{C}_{k \times p}$ a matrix of constants.

For the linear combination $\mathbf{c}'\mathbf{x} = c_1 x_1 + \cdots + c_p x_p$,

3. $\mathbf{E}(\mathbf{c}'\mathbf{x}) = \mathbf{c}'\mu$, and
4. $\text{var}(\mathbf{c}'\mathbf{x}) = \mathbf{c}'\Sigma_\mathbf{x}\mathbf{c}$.

For the $k$-dimensional vector of linear combinations, $\mathbf{Z}_{k \times 1} = \mathbf{C}\mathbf{x}$,

5. $\mu_\mathbf{Z} = \mathbf{E}(\mathbf{Z}) = \mathbf{E}(\mathbf{C}\mathbf{x}) = \mathbf{C}\mathbf{E}(\mathbf{x}) = \mathbf{C}\mu_\mathbf{x}$ and
6. $\Sigma_\mathbf{Z} = \text{var}(\mathbf{Z}) = \text{var}(\mathbf{C}\mathbf{x}) = \mathbf{C}\Sigma_\mathbf{x}\mathbf{C}'$.  

44
All of the definitions and properties of the last 5 pages have been defined in terms of population quantities. There are corresponding definitions of sample quantities for which the same properties hold:

Examples (sample size $n$):

- Sample mean: \( \bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)' \), where \( \bar{x}_i = n^{-1} \sum_{k=1}^{n} x_{ki} \). (\( x_{ki} \) is the \( k^{th} \) observation of random variable \( x_i \)).
- Sample variance-covariance matrix:

\[
S_x = \begin{pmatrix}
s_{11} & \cdots & s_{1p} \\
\vdots & \ddots & \vdots \\
s_{p1} & \cdots & s_{pp}
\end{pmatrix} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})',
\]

where

\[
s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j).
\]
Sample correlation matrix:

\[
\mathbf{R}_X = \begin{pmatrix}
1 & r_{12} & \cdots & r_{1p} \\
r_{21} & 1 & \cdots & r_{2p} \\
& & \ddots & \vdots \\
r_{p1} & r_{p2} & \cdots & 1
\end{pmatrix},
\]

where \( r_{ij} \) is the sample correlation computed on the \( n \) sample pairs of observations \((x_{1i}, x_{1j}), \ldots, (x_{ni}, x_{nj})\):

\[
r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}} \sqrt{s_{jj}}},
\]
In the univariate setting we typically have a random variable $X$, say, from which we observe a random sample. That is, we assume that $X$ has some distribution that is partially or completely unknown and we observe $n$ independent copies $X_1, \ldots, X_n$ of $X$. The observed values of these copies we typically denote with lower case letters: $x_1, \ldots, x_n$, and we infer features of the true distribution from the sample.

In the multivariate setting we are interested in the true (multivariate) distribution of a random vector $\mathbf{x}_{p \times 1}$. Typically the random variables that make up $\mathbf{x}$, $x_1, \ldots, x_p$ are not independent and that dependence (correlation) among $x_1, \ldots, x_p$ is the reason for our interest. A random sample here consists of $n$ copies of $\mathbf{x}$ that we can arrange in an $n \times p$ data matrix:

$$
\mathbf{X} = 
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np} 
\end{pmatrix}
$$

(I’m using lower case here, no longer distinguishing between the random variable and its observed value.)
• Rows are assumed independent \((n\) independent copies).

• Dependence typically among columns.

The sample mean and variance can be computed from the data matrix \(X\) as

\[
\bar{x} = \frac{1}{n}X'1
\]

\[
S = \frac{1}{n-1}X'(I - \frac{1}{n}11')X
\]

where \(1\) is an \(n\)-dimensional vector of ones.

**Generalized Sample Variance:** \(S\) summarizes the variability in each of \(x_1, \ldots, x_p\) and the covariability among these variables. \(S\) involves \(p\) variances and \(p(p - 1)/2\) covariances, though. A more succinct summary is provided by the generalized sample variance, \(|S|\).
• Sample variance-covariance matrices with quite different patterns can have the same generalized variances. E.g.,

\[
\begin{pmatrix}
5 & 4 \\
4 & 5
\end{pmatrix}, \quad \begin{pmatrix}
5 & -4 \\
-4 & 5
\end{pmatrix},
\]

\[
\begin{pmatrix}
3 & 0 \\
0 & 3
\end{pmatrix}
\]

have values of \( r_{12} \) equal to 0.8, -0.8 and 0.0, respectively. However,

\[
|S_1| = |S_2| = |S_3| = 9.
\]

• |\( S \)| is sensitive to large values of the sample variance, \( s_{ii} \), for one or a few of the \( p \) variables in \( X \). In some cases this may simply be an effect of scale. In such cases it is of interest to compute the generalized variance on the standardized data matrix where \( x_{ij} \) has been replaced by \((x_{ij} - \bar{x}_i) / \sqrt{s_{ii}}\). This procedure is equivalent to computing

\[
|R| = |S| / (s_{11}s_{22} \cdots s_{pp}),
\]

the determinant of the sample correlation matrix.
• $|S|$ is only one way to quantify the total variability and covariability in $S$. Another possibility that ignores the covariance structure is the total variance:

$$
\sum_{i=1}^{p} s_{ii}.
$$
**Example, Kite Data:** Consider the bivariate data of Table 5.11 in the text. A wildlife ecologist measured $x_1 =$ tail length (mm) and $x_2 =$ wing length (mm) for a sample of $n = 45$ female hook-billed kites. The handout consists of a printout of a SAS program, kite1.sas, and the associated output, kite1.lst.
• PROC PRINT prints out a listing of the data on p. 1 of kite1.lst.

• PROC CORR computes the sample correlation matrix $R$ for the variables on the VAR statement. In addition the COV option on the PROC CORR line requests the sample covariance matrix $S$ and, by default, PROC CORR computes summary statistics ($\bar{x}_1$, $\bar{x}_2$, etc.). All of these results appear on p. 2 of kite1.lst.

• PROC PLOT prints a scatter plot of winglen versus taillen on p. 3 of kite1.lst.

• The generalized sample variance is

$$|S| = 120.69(208.54) - 122.35^2 = 10201.24$$

and we have

$$|R| = \frac{10201.24}{120.69(208.54)} = 1.0(1.0) - 0.77^2 = 0.41.$$  

• The total variance is $120.69 + 208.54 = 329.24$. 

52
Suppose we decide that we are interested in the size and shape of the birds in a way that is not captured by wing length ($x_1$) and tail length ($x_2$). Specifically, we might look at $y_1 = x_1 + x_2$ (size) and $y_2 = x_1 - x_2$ (shape).

To compute summary statistics on $y = (y_1, y_2)'$, we can use our methods for linear combinations of $x = (x_1, x_2)'$. Notice,

$$y = Ax,$$

where

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

so

$$\bar{y} = A\bar{x} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 193.62 \\ 279.78 \end{pmatrix} = \begin{pmatrix} 473.40 \\ -86.16 \end{pmatrix}.$$
In addition,

\[ S_y = A S_x A' \]

\[
= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 120.69 & 122.35 \\ 122.35 & 208.54 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 243.04 & -1.65 \\ 330.89 & -86.19 \end{pmatrix} \\
= \begin{pmatrix} 573.93 & -87.85 \\ -87.85 & 84.54 \end{pmatrix}
\]
The Multivariate Normal Distribution
(See chapter 4 of Johnson and Wichern.)

• Recall the central role that the (univariate) normal distribution played in the development of univariate statistical methods and why:

  a. For many continuous random variables the assumption of normality is appropriate.

  b. numerous statistics, particularly those that can be expressed as sums or averages, have normal sampling distributions regardless of the underlying population distribution.

  c. the normal distribution is easy to work with mathematically and many useful results are available.

• Items (a) and (b) above are due to the Central Limit Theorem.
Recall:

**Central Limit Theorem.** If $Y_1, \ldots, Y_n$ is a sequence of $n$ i.i.d. random variables each with mean $\mu$ and variance $\sigma^2$ (both finite), then

$$Z_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma \sqrt{n}} \sim N(0, 1).$$

That is,

$$\lim_{n \to \infty} \frac{F_{Z_n}(z)}{\Phi(z)} \to 1,$$

where $F_{Z_n}(z)$ is the c.d.f. of $Z_n$ and $\Phi(z)$ is the c.d.f. of a standard normal r.v.

- How large must $n$ be? Depends on the problem (on the distribution of the $Y$’s).

- CLT directly implies (b) but also implies (a) because the measured realization of a random variable can often be thought of as including the sum of numerous random quantities (e.g., reaction time).
• The multivariate normal distribution plays an even more central role in multivariate analysis. The justification/motivation is the same, plus

d. There are few alternatives in the multivariate setting (few tractable multivariate distributions as alternatives). There are alternatives in Discrete Multivariate Analysis, see Bishop, et al., STAT8620.

Recall the **univariate normal distribution**: we write $X \sim N(\mu, \sigma^2)$ to signify that the univariate r.v. $X$ has the normal distribution with mean $\mu$ and variance $\sigma^2$ which means that $X$ has probability density function (p.d.f.)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

Meaning: for two values $x_1 < x_2$ the area under the graph of the p.d.f. between $f_X(x_1)$ and $f_X(x_2)$ gives $\Pr(x_1 < X < x_2)$. 

57
• The cumulative distribution function (c.d.f.) of \( X \) is

\[
F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.
\]

\( F_X \) is not available in closed-form but the standard normal version (with \( \mu = 0, \sigma^2 = 1 \)) is extensively tabulated.

The **multivariate normal distribution**: for the vector \( \mathbf{x} \), we write \( \mathbf{x} \sim \mathcal{N}_p(\mathbf{\mu}, \Sigma) \) to signify that \( \mathbf{x} \) is a \( p \)-dimensional random vector with mean \( \mathbf{\mu} \) and variance \( \Sigma \). \( \mathbf{x} \) has p.d.f.

\[
f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mathbf{\mu})' \Sigma^{-1} (\mathbf{x} - \mathbf{\mu}) \right],
\]

and c.d.f.

\[
F(\mathbf{x}) = \Pr(x_1 \leq x_1, x_2 \leq x_2, \ldots, x_p \leq x_p)
= \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(t)dt.
\]
• We’ve replaced a univariate distance from $x$ to $\mu$ with a multivariate statistical distance,

$$(x - \mu)'\Sigma^{-1}(x - \mu),$$

from $x$ to $\mu$.

• It is now the volume under the p.d.f. which represents the joint probability that $x_1, \ldots, x_p$ fall within given intervals.
• Contours of constant density for the \( p \)-dimensional normal distribution are ellipsoids defined by \( \mathbf{x} \) such that
\[
(\mathbf{x} - \mathbf{\mu})'\Sigma^{-1}(\mathbf{x} - \mathbf{\mu}) = c^2.
\]
These ellipsoids are centered at \( \mathbf{\mu} \) and have axes \( \pm c\sqrt{\lambda_i}\mathbf{e}_i \), for eigen-pairs \( \lambda_i, \mathbf{e}_i \) of \( \Sigma \).

• We will show that for \( \mathbf{x} \sim N_p(\mathbf{\mu}, \Sigma) \) with \( |\Sigma| > 0 \)
\[
(\mathbf{x} - \mathbf{\mu})'\Sigma^{-1}(\mathbf{x} - \mathbf{\mu}) \sim \chi^2(p).
\]
This result implies that the solid ellipsoid of \( \mathbf{x} \) values satisfying
\[
(\mathbf{x} - \mathbf{\mu})'\Sigma^{-1}(\mathbf{x} - \mathbf{\mu}) \leq \chi^2(\alpha)(p)
\]
has probability \( 1 - \alpha \).
Additional Properties:

1. If \( x \sim N_p(\mu, \Sigma) \) then any linear combination of variables \( a'x = a_1x_1 + a_2x_2 + \cdots + a_px_p \) is distributed as \( N(a'\mu, a'\Sigma a) \).

2. If \( x \sim N_p(\mu, \Sigma) \), the \( q \) linear combinations

\[
A_{q \times p}x = \\
\begin{pmatrix}
a_{11} x_1 + \cdots + a_{1p} x_p \\
a_{21} x_1 + \cdots + a_{2p} x_p \\
\vdots \\
a_{q1} x_1 + \cdots + a_{qp} x_p
\end{pmatrix}
\]

are distributed as \( N_q(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}') \).

3. For \( x \sim N_p(\mu, \Sigma) \) and \( d_{p \times 1} \) a vector of constants, \( x + d \) is distributed as \( N_p(\mu + d, \Sigma) \).

4. For \( p \)-dimensional random vector \( x \), if \( a'x \sim N(a'\mu, a'\Sigma a) \) for every \( p \)-vector of constants \( a \) then \( x \sim N_p(\mu, \Sigma) \).
5. For $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, all subsets of $\mathbf{x}$ are (multi-
variate) normally distributed. If we partition $\mathbf{x}$, and $\boldsymbol{\mu}$ and $\Sigma$ accordingly as

$$
\mathbf{x} \quad (p \times 1) = \begin{pmatrix}
\mathbf{x}_1 \\
(\times 1)
\vdots \\
\mathbf{x}_2 \\
((p-q) \times 1)
\end{pmatrix}
= \begin{pmatrix}
\mu_1 \\
(\times 1)
\vdots \\
\mu_2 \\
((p-q) \times 1)
\end{pmatrix}
$$

$$
\Sigma \quad (p \times p) = \begin{pmatrix}
\Sigma_{11} & & \\
(\times q) & \cdots \\
& \Sigma_{12} \\
\vdots & \cdots & \cdots & \vdots \\
\Sigma_{21} & & \\
((p-q) \times q) & \cdots & \cdots & \\
& \Sigma_{22} \\
((p-q) \times (p-q)) & & \\
& & \cdots & \\
& & & \cdots
\end{pmatrix}
$$

then $\mathbf{x}_1$ is distributed as $N_q(\boldsymbol{\mu}_1, \Sigma_{11})$. 

62
6.

a. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent $q_1$ and $q_2$ dimensional vectors, respectively, then regardless of the distributions of $\mathbf{x}_1$ and $\mathbf{x}_2$, $\text{cov}(\mathbf{x}_1, \mathbf{x}_2) = 0_{q_1 \times q_2}$.

b. If $\mathbf{x}_1$, $\mathbf{x}_2$ are as before and are each multivariate normally distributed, then $\text{cov}(\mathbf{x}_1, \mathbf{x}_2) = 0$ implies that $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent.

c. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent with $N_{q_1}(\mu_1, \Sigma_{11})$ and $N_{q_2}(\mu_2, \Sigma_{22})$ distributions, respectively, then

$$
\begin{pmatrix}
\mathbf{x}_1 \\
\vdots \\
\mathbf{x}_2
\end{pmatrix}
\sim N_{q_1 + q_2}
\begin{pmatrix}
\mu_1 \\
\vdots \\
\mu_2
\end{pmatrix},
\begin{bmatrix}
\Sigma_{11} & 0 \\
0 & \Sigma_{22}
\end{bmatrix}.
$$

If we combine (multivariate) normally distributed components into a single vector, the result is not necessarily multivariate normal. However, if the components are (multivariate) normal and independent then the result will be multivariate normal.
7. If \( \mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \) where

\[
\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_2 \end{pmatrix}
\]

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \ldots & \Sigma_{12} \\ \vdots & \ddots & \vdots \\ \Sigma_{21} & \ldots & \Sigma_{22} \end{pmatrix}
\]

with \( |\Sigma_{22}| > 0 \), then the conditional distribution of \( \mathbf{x}_1 \) given \( \mathbf{x}_2 = \mathbf{x}_2 \), is (multivariate) normal with

\[
\text{Mean} = \boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)
\]

and

\[
\text{Covariance} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\]

Notice that the conditional covariance does not depend on the value of the conditioning variable.
8. For $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ with $|\Sigma| > 0$,

$$(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(p)$$

where $\chi^2(p)$ denotes the chi-square distribution with $p$ degrees of freedom.

Proof: we can write the quadratic form as

$$(\mathbf{x} - \boldsymbol{\mu})' \mathbf{E} \Lambda^{-1/2} \Lambda^{-1/2} \mathbf{E}' (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{Z}' \mathbf{Z},$$

where $\mathbf{Z} = \Lambda^{-1/2} \mathbf{E}' (\mathbf{x} - \boldsymbol{\mu})$ is multivariate normal with mean

$$\mathbf{E}(\mathbf{Z}) = \Lambda^{-1/2} \mathbf{E}' \mathbf{E} (\mathbf{x} - \boldsymbol{\mu}) = 0$$

and variance

$$\text{var}(\mathbf{Z}) = \Lambda^{-1/2} \mathbf{E}' \Sigma \mathbf{E} \Lambda^{-1/2}$$

$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = \mathbf{I}$$

so $\mathbf{Z}' \mathbf{Z}$ is the sum of $p$ squared independent $\mathcal{N}(0, 1)$ r.v.s $\Rightarrow \mathbf{Z}' \mathbf{Z} \sim \chi^2(p)$. 

65
Multivariate Sampling Distributions:

In this course we will most often consider data that are or are assumed to be a random sample from a multivariate population. That is, we will have \( n \) independent copies, \( x_1, x_2, \ldots, x_n \), of a \( p \)-dimensional random variable which we arrange in a \( n \times p \) data matrix:

\[
X = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
x'_1 \\
x'_2 \\
\vdots \\
x'_n
\end{pmatrix}
\]

- Rows of \( X \) are independent.
- Columns of \( X \) are dependent and this dependence is of interest.
First we consider the general case where \( x_1, \ldots, x_n \) have arbitrary (not necessarily normal) common distribution with mean \( \mu \) and variance \( \Sigma \):

1. The sample mean \( \bar{x} \) is unbiased for \( \mu \); i.e.,

\[
E(\bar{x}) = \mu.
\]

This is easily seen by writing \( \bar{x} \) as \( \bar{x} = \frac{1}{n}X'1 \).

It follows that

\[
E(\bar{x}) = \frac{1}{n}E(X')1
\]

\[
= \frac{1}{n} \left( \mu, \mu, \ldots, \mu \right)1 = \mu.
\]
2. \( \text{var}(\bar{x}) = \frac{1}{n} \Sigma \). We can write \( \bar{x} \) as

\[
\bar{x} = \left( \frac{1}{n} I, \frac{1}{n} I, \ldots, \frac{1}{n} I \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}
\]

Therefore, \( \text{var}(\bar{x}) \) equals

\[
\left( \frac{1}{n} I, \frac{1}{n} I, \ldots, \frac{1}{n} I \right) \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{pmatrix} \begin{pmatrix} \frac{1}{n} I \\ \frac{1}{n} I \\ \vdots \\ \frac{1}{n} I \end{pmatrix}
\]

\[
= \left( \frac{1}{n} I, \frac{1}{n} I, \ldots, \frac{1}{n} I \right) \begin{pmatrix} \frac{1}{n} \Sigma \\ \vdots \\ \frac{1}{n} \Sigma \end{pmatrix}
\]

\[
= n \left( \frac{1}{n^2} \Sigma \right) = \frac{1}{n} \Sigma
\]
3. The sample variance matrix \( \mathbf{S} \) is unbiased for \( \Sigma \); i.e., \( \mathbb{E}(\mathbf{S}) = \Sigma \). Proof:

\[
\mathbb{E}(\mathbf{S}) = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})']
\]

\[
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i') - n\mathbb{E}(\bar{\mathbf{x}}\bar{\mathbf{x}}') \right].
\]

As a general result,

\[
\Sigma_{\mathbf{y}} = \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})']
\]

\[
= \mathbb{E}(\mathbf{y}\mathbf{y}') - \boldsymbol{\mu}_{\mathbf{y}}\boldsymbol{\mu}_{\mathbf{y}}'
\]

\[
\Rightarrow \quad \mathbb{E}(\mathbf{y}\mathbf{y}') = \Sigma_{\mathbf{y}} + \boldsymbol{\mu}_{\mathbf{y}}\boldsymbol{\mu}_{\mathbf{y}}'.
\]

So,

\[
\mathbb{E}(\mathbf{S}) = \frac{1}{n-1} \left[ n(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}') - n(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}') \right] = \Sigma.
\]
Large Sample Results (general case):

1. $\bar{x} \xrightarrow{p} \mu$, $\mathbf{S} \xrightarrow{p} \Sigma$.

2. $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N_p(0, \Sigma)$.

3. Combining (1) and (2) we get
   $$\sqrt{n}\mathbf{S}^{-1/2}(\bar{x} - \mu) \xrightarrow{d} N_p(0, \mathbf{I}).$$

4. (3) implies
   $$n(\bar{x} - \mu)'\mathbf{S}^{-1}(\bar{x} - \mu) \xrightarrow{d} \chi^2(p).$$

- $a \xrightarrow{p} b$ means as the sample size gets large, $a$ converges to $b$ in the sense that the probability that $a$ is “close” to $b$ goes to 1.

- statistic $\xrightarrow{d}$ distribution means that as the sample size gets large the distribution of the statistic on the left becomes approximately equal to the distribution on the right.
Next we consider the special case where we assume that the common distribution of our random sample \( x_1, \ldots, x_n \) is \( N_p(\mu, \Sigma) \). Our additional assumption of normality buys us some stronger results:

1. 
\[
\bar{x} \sim N_p(\mu, n^{-1}\Sigma) \\
n(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu) \sim \chi^2(p)
\]

Unlike the general case, these results are exact and hold for any sample size.

2. The random \( p \times p \) matrix \((n - 1)S\) has a distribution known as the **Wishart distribution**. The Wishart distribution has two parameters, the degrees of freedom \((n - 1 \text{ in this case})\) and the scale matrix \((\Sigma \text{ in this case})\). We write \((n - 1)S \sim W_p(\Sigma, n - 1)\).

3. \( \bar{x} \) and \( S \) are independent.
Properties of the Wishart Distribution:

1. $W_p(\Sigma, m)$ is the distribution of $\sum_{i=1}^m z_i z_i'$ where $z_1, \ldots, z_m \overset{iid}{\sim} N_p(0, \Sigma)$ which implies that

$$\sum_{i=1}^n (x_i - \mu)(x_i - \mu)' = (n-1)S \sim W_p(\Sigma, n-1)$$

for a multivariate normal sample $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \Sigma)$ with sample variance-covariance matrix $S$.

2. If $A_i \overset{ind}{\sim} W_p(\Sigma, m_i)$ then

$$\sum_i A_i \sim W_p \left( \Sigma, \sum_i m_i \right).$$

3. If $A \sim W_p(\Sigma, m)$ then for $C_{q \times p}$,

$$CAC' \sim W_q(\bar{C}\Sigma C', m).$$
Assessment of Normality:

- We have seen that $x_{p \times 1}$ is multivariate normal if and only if every linear combination of $x_1, \ldots, x_p$ is normal. Therefore, one check is whether selected linear combinations are (univariate) normal.

- As a special case of the above we can check whether each component of $x$ is (univariate) normal.

- Other selected linear combinations of $x_1, \ldots, x_p$ worth checking are $\hat{e}_1' x$ and $\hat{e}_p' x$, where $\hat{e}_1$ and $\hat{e}_p$ are the eigenvectors associated with $\hat{\lambda}_1$ and $\hat{\lambda}_p$, respectively, the largest and smallest eigenvalues of $S$. 
How do we assess univariate normality?

One method is the quantile-quantile (Q-Q) plot (also known as a probability plot).

• Idea is to plot the sample quantiles (percentiles) for the variable in question versus the quantiles one would expect to observe assuming exact normality.

• If the data are normal, plot should follow a straight line. Departures from a straight line may indicate the type of non-normality (e.g., heavy tails, skewness).

• Q-Q plots can be misleading for small sample sizes \( (n < 20, \text{ say}) \). Small samples of truly normal data can look crooked.
Example:
• A formal test for normality is based on $r_q$ the correlation coefficient of the observed and expected quantiles. The hypothesis of normality is rejected for values of $r_q$ below the tabulated value in Table 4.2 on p.193 of the text.

• The test statistic based on $r_q$ is approximately equivalent to the Shapiro-Wilk test computed by PROC UNIVARIATE and in, general, these two tests give very similar results. Either may be used.

**Example – Sparrow Data:**

Recall: $x_1 =$ total length, $x_2 =$ alar extent, $x_3 =$ length of beak and head, $x_4 =$ length of humerus, $x_5 =$ length of keel of sternum.

• See sparrow2.sas and sparrow2.lst. Here we just consider $x_1, x_2, x_3$ to keep the size of the example manageable. In addition, we consider only the birds who died (birds 1–21).
• Notice that the Q-Q plots produced by PROC UNIVARIATE have such small scales as to make them practically worthless.

• The Shapiro-Wilk tests for \( x_1 - x_3 \) all give \( p \)-values > .05, so there is not strong evidence in any of the individual variables to indicate non-multivariate normality.

• E.g., the Shapiro-Wilk test statistic for \( x_1 \), the total length of the bird, is \( W = 0.934 \) with \( p \)-value 0.165 (see p.1 of sparrow2.lst).

• However, we also should check the univariate normality of \( \hat{e}'x \) and \( \hat{e}'x \). \( \hat{e}'x \) and \( \hat{e}'x \) are computed in PROC IML and printed on p.9.

• The S-W test for \( \hat{e}'x \) is given by PROC UNIVARIATE on p.11 as \( W = 0.861 \) with \( p \)-value 0.0067. Therefore, we reject the null hypothesis of normality for \( e'_x \) at the 0.05 level ⇒ we reject the hypotheses that \( x \) is trivariate normal.
• As an alternative to the Shapiro-Wilk test, \( r_q = 0.920 \) is computed on p.16. Since 0.920 falls below the critical value for 0.05 and a sample size of 20 (0.9269 from Table 4.2 in the text) we reject the null hypothesis of normality.

• The tests agree.

• Note that the procedure here deviates from the usual one by which we reject when the test statistic exceeds the critical value.

• Notice that as an alternative to letting PROC UNIVARIATE automatically produce the Q-Q plot, it is fairly easy to compute the quantiles to be plotted in a Q-Q plot and then to plot these quantiles with PROC PLOT. This I have done for both \( \hat{e}_1'x \) and \( \hat{e}_3'x \) to produce the plots on p.15 and p.17. These are obviously much easier to read than the corresponding plots produced by PROC UNIVARIATE on pp. 12 and 14, respectively.
A graphical procedure to examine multivariate normality more directly is the **Chi-square plot** or **gamma plot**.

- When the sample size is moderate to large, the squared generalized distances

\[ d_j^2 = (x_j - \bar{x})' S^{-1} (x_j - \bar{x}), \quad j = 1, \ldots, n, \]

should be approximately distributed as $\chi^2(p)$ random variables.

- The chi-square plot is a Q-Q plot of $d_1^2, \ldots, d_n^2$ versus the quantiles one would expect to observe based on the $\chi^2(p)$ distribution.

- A straight line shape indicates multivariate normality.
Sparrow Example (Continued):

- The Q-Q plot for $d^2$ appears on p.4 of sparrow3.lst.

- Except for one point, this plot looks pretty close to a straight line. The unusual point, or outlier, is bird #8. This outlier was also seen in the Q-Q plot of $\hat{e}_1'x$ and seems to be the main cause of non-multivariate normality in these data.

- It would be interesting to eliminate bird #8 and then repeat our examination of the multivariate normality of this data set. Perhaps bird #8’s data was incorrectly recorded, or perhaps bird #8 happened to die for some reason other than exposure to the severe storm.

- If a good reason can be found that this bird’s data do not belong in the data set, then this outlier should be removed. Otherwise, a good solution would be to analyze the data both with and without the outlier and to report both sets of results.
Suggested Procedure:

1. Examine Q-Q plots for all univariate components of $x$.

2. Examine Q-Q plots for $\hat{e}_1'x$ and $\hat{e}_p'x$.

3. Examine Q-Q plot for $d_j^2$’s.

- While one bad plot is enough to invalidate the multivariate normality assumption, it is not enough that plots on all examined linear combinations look good to confirm (with certainty) the normality assumption.

- However, we can only go so far in trying to confirm multivariate normality. If all plots look good, we will consider the multivariate normality assumption to be satisfied.

- Other “goodness of fit” tests of normality exist (chi-square, Kolmogorov-Smirnov), but these suffer serious drawbacks. They have low power to detect nonnormality in small samples and oversensitivity in large samples.
Handling Non-normality – Transformations

• Basic idea: if $x$ is not normal, find a function of $x$ (square root, logarithm, etc.) that is.

• Most of the useful transformations are contained in the family of power transformations:

$$f(x) = \begin{cases} 
    x^\lambda, & \text{if } \lambda \neq 0 \\
    \log(x), & \text{corresponds to } \lambda = 0.
\end{cases}$$

• Although there are analytic methods for picking the $\lambda$ that gives the best power transformations, in practice, trial and error is an effective and acceptable approach.

• Trial and error is aided by the knowledge that power transformations with $\lambda < 1$ ($x^{1/2} = \sqrt{x}$, $x^0 = \log(x)$, $x^{-1} = 1/x$, etc.) shrink large values of $x$. Power transformations with $\lambda > 1$ ($x^2$, $x^3$, etc.) increase large values of $x$. 
• Power transformations are defined for positive valued random variables. However they can be applied to variables with negative values by first adding a large value to each observation (to make all values positive) before taking the transformation.

• For certain special types of data, specific transformations are available that often work well:

<table>
<thead>
<tr>
<th>Original Scale</th>
<th>Transformed Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts, $y$</td>
<td>$\sqrt{y}$</td>
</tr>
<tr>
<td>Proportions, $p$</td>
<td>$\frac{1}{2}\logit(p) = \frac{1}{2} \log\left(\frac{p}{1-p}\right)$</td>
</tr>
<tr>
<td>Correlations, $r$</td>
<td>Fisher’s $z(r) = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right)$</td>
</tr>
</tbody>
</table>

• Theory may also suggest an appropriate transformation.

• Often transforming the univariate components that are non-normal will be sufficient. Always re-check the multivariate normality of the transformed vectors.
Inference about a Mean Vector
(See chapter 5 of Johnson and Wichern.)

Finally, some data analysis!

Recall the univariate situation: $t$–tests and $Z$–tests for a single mean.

The simplest scenario is one in which we want to know whether the true mean of a population, $\mu$, is equal to some specified value, $\mu_0$. We frame the problem as a statistical test of competing hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

where $\mu$ is the mean of the population from which we have drawn a sample $x_1, \ldots, x_n \sim iid N(\mu, \sigma^2)$.

$\sigma^2$ known: If the population variance is known then

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1),$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{(sample mean)}.$$
$\sigma^2$ unknown: (usual case) if the population variance is unknown we have to estimate it with the sample variance $s^2$ and the substitution of $s$ for $\sigma$ in our test statistic changes its distribution (test statistic is more variable now because of errors in estimating $\sigma^2$):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n - 1) \quad t \text{ dist'\n with } n - 1 \text{ d.f.}$$

where

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (\text{sample variance}).$$

We reject $H_0$ in favor of $H_1$ at the $\alpha$–level if the observed $|t|$ exceeds a critical value:

$$\text{Reject } H_0 \text{ if } |t| > t_{\alpha/2}(n - 1).$$

We can equivalently form a rejection rule based on $t^2$: Reject $H_0$ if

$$t^2 = n(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0) > t_{\alpha/2}^2(n - 1).$$

(looks sort of like a quadratic form (statistical distance) doesn’t it?)
As an alternative to looking at the problem from a testing perspective, we can focus on estimation.

- The acceptance region of our two sided test (those values of $\bar{x}$ for which we reject $H_0$ in favor of $H_1$) at level $\alpha$, corresponds to a $100(1 - \alpha)$% confidence interval for $\mu$.

Why?

Because

$$
\Pr \left( \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \leq t_{\alpha/2}(n - 1) \right)
= \Pr \left( |\bar{x} - \mu_0| \leq t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \right)
= \Pr \left( - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \leq \bar{x} - \mu_0 \leq t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \right)
= \Pr \left( \bar{x} - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \right)
$$
Extension to Multivariate Case:

From a $N_p(\mu, \Sigma)$ population we draw a sample

\[ x_1, \ldots, x_n \overset{iid}{\sim} N_p(\mu, \Sigma). \]

The corresponding hypotheses in the multivariate case are

\[ H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0, \]

where

\[ \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \mu_0 = \begin{pmatrix} \mu_{10} \\ \vdots \\ \mu_{p0} \end{pmatrix}. \]

Note that $H_0$ requires all component means to be equal to the specified values, $\mu_{10}, \ldots, \mu_{p0}$ and $H_A$ holds if any (one or more) $\mu_i \neq \mu_{i0}$.

Test statistic (Hotelling’s $T^2$):

\[ T^2 = n(\bar{x} - \mu_0)'S^{-1}(\bar{x} - \mu_0). \]

- We reject $H_0$ in favor of $H_1$ if $T^2$ is large.
How large? What critical value do we compare the observed $T^2$ to?

- We need to know the distribution of $T^2$ to answer these questions.

- $t$ in the univariate case had a student’s $t$ distribution because it $t$ is of the form

$$
\frac{N(0, 1)}{\sqrt{\chi^2(\text{d.f.})/\text{d.f.}}}
$$

- Random variables which are of the form

$$
\frac{\chi^2(\text{d.f.}_1)/\text{d.f.}_1}{\chi^2(\text{d.f.}_2)/\text{d.f.}_2}
$$

are known to have an $F(\text{d.f.}_1, \text{d.f.}_2)$ distribution.
• So, since
\[ t^2 = \sqrt{n}(\bar{x} - \mu_0)(s^2)^{-1}\sqrt{n}(\bar{x} - \mu_0) \]
is of the form
\[
(N(0, 1)) \left( \chi^2(\text{d.f.}_2)/\text{d.f.}_2 \right)^{-1} (N(0, 1))
\]
\[
= \frac{\chi^2(1)/1}{\chi^2(\text{d.f.}_2)/\text{d.f.}_2},
\]
it follows that \( t^2 \sim F(1, \text{d.f.}_2) \).

• In the multivariate case
\[ T^2 = \sqrt{n}(\bar{x} - \mu_0)'S^{-1}\sqrt{n}(\bar{x} - \mu_0) \]
is of the form
\[
(N_p)(W_p/\text{d.f.})^{-1}(N_p).
\]
Since the multivariate normal is a generalization of the normal and the Wishart is a generalization of the \( \chi^2 \), it is not very surprising that it turns out that \( T^2 \) has a (scaled) \( F \) distribution:
\[ T^2 \sim \frac{(n - 1)p}{n - p}F(p, n - p). \]
• We reject $H_0$ in favor of $H_1$ if

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

where $F_\alpha(\text{d.f.}_1, \text{d.f.}_2)$ is the upper $(100\alpha)^{th}$ percentile of the $F(\text{d.f.}_1, \text{d.f.}_2)$ distribution.

• Notice that $T^2$ is the statistical distance between $\bar{x}$ and $\mu_0$.

**Example – Kite Data:**

$p = 2$ variables ($x_1 =$wing length, $x_2 =$tail length) measured on a sample of $n = 45$ birds. Suppose we want to test

$$H_0 : \mu = \begin{pmatrix} 190 \\ 275 \end{pmatrix} \quad \text{versus} \quad H_1 : \mu \neq \begin{pmatrix} 190 \\ 275 \end{pmatrix}$$
Recall
\[
\bar{x} = \begin{pmatrix} 193.62 \\ 279.78 \end{pmatrix} \quad S = \begin{pmatrix} 120.69 & 122.35 \\ 122.35 & 208.54 \end{pmatrix}
\]
so \( S^{-1} \) equals
\[
\frac{1}{120.69(208.54) - (122.35)^2} \begin{pmatrix} 208.54 & -122.35 \\ -122.35 & 120.69 \end{pmatrix}.
\]
So
\[
T^2 = 45 \left(193.62 - 190 \quad 279.78 - 275\right) \frac{1}{10201.24} \\
\times \begin{pmatrix} 208.54 & -122.35 \\ -122.35 & 120.69 \end{pmatrix} \begin{pmatrix} 193.62 - 190 \\ 279.78 - 275 \end{pmatrix}
\]
\[
= 5.543
\]
For a significance level of \( \alpha = 0.05 \), the critical value is
\[
\frac{(n - 1)p}{n - p} F_{\alpha}(p, n - p) = \frac{2(44)}{43} F_{0.05}(2, 43)
\]
\[
= (2.047)(3.214) = 6.578
\]
so we do not reject \( H_0 \) and we conclude that \( \mu' = (190, 275) \) is plausible.
• Hotelling’s $T^2$ test is invariant under transformations of the form

$$y_{p\times1} = C_{p\times p}x_{p\times1} + d_{p\times1}$$

where $C$ is nonsingular.

**Kite Example (Continued):**

So, for example, if we wanted to test the corresponding hypothesis on size ($y_1 = x_1 + x_2$) and shape ($y_2 = x_1 - x_2$) as we defined them before:

$$y = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} x,$$

we would test

$$H_0 : \mu_y = \begin{pmatrix} 190 + 275 \\ 190 - 275 \end{pmatrix} = \begin{pmatrix} 465 \\ -85 \end{pmatrix}$$

For this problem, $T^2 = 5.543$ as before with the same critical value, 6.578, so we do not reject $H_0$.

Note that

$$\begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} \neq 0 \text{ (nonsingular transformation).}$$
Why not just do $p$ separate (univariate) $t$ tests for each component of $x$?

- Does not take into account interrelationships (correlations) among the components of $x$.

- The multivariate test controls the **type I error rate**, the probability of rejecting $H_0$ when $H_0$ is true.
  
  - When doing multiple univariate tests the probability that at least one univariate null hypothesis will be rejected *by chance alone* increases with the number of tests being performed.

  - Procedures to account for this problem exist (e.g., Bonferroni) but generally such procedures are not as attractive as performing a single multivariate test.
$T^2$ as a LRT:

- So far the use of $T^2$ as a test statistic has been motivated by analogy to the univariate case.

Could we do better? (Is $T^2$ optimal? E.g., in terms of power?)

- It turns out that $T^2$ does have several optimality properties because $T^2$ can also be derived as a **likelihood ratio test** (LRT). The theory of LRTs (don’t worry) establishes a number of optimality properties that $T^2$ inherits.
Likelihood Ratio Tests:

- The **likelihood function** is just the density function, but thought of as a function of the parameters rather than of the data.

For Example, the univariate normal distribution based on a sample $x_1, \ldots, x_n$:

- We think of the density

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \times \exp \left[ -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

as a function of the data (the $x_i$’s) for given values of the parameters ($\mu, \sigma^2$) giving, in some sense, the probability of observing $x_1, \ldots, x_n$ for given $\mu, \sigma^2$. Hence, the argument of $f$ is $x_1, \ldots, x_n$.

- Actually, the density function involves both the data and the parameters.
– Once we have the data, we want to do inference concerning the parameters. Therefore, we can think of the density as a function of the parameters \((\mu, \sigma^2)\) given the data (the \(x_i\)’s).

– To avoid confusion when we think of the density this way we give it a new name, the likelihood function, and write it as a function of the parameters:

\[
L(\mu, \sigma^2) = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left[ -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

– Likelihood function gives, in some sense, the probability (or likelihood) of the parameters given the data.
• The **maximum likelihood** approach to estimation says to estimate the parameters with those values that maximize the likelihood.

  - For the univariate normal distribution, with some calculus we can establish that the maximizers of \( L(\mu, \sigma^2) \) (the maximum likelihood estimates or MLEs) are

    \[
    \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \\
    \hat{\sigma}^2 = \frac{n-1}{n} s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.
    \]

  - For the multivariate normal distribution, the likelihood function is

    \[
    L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \times \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \right].
    \]
– Again some calculus lead to the MLEs:

\[ \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \]

\[ \hat{\Sigma} = \frac{n-1}{n} S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'. \]

– The idea is that \( \hat{\mu} \) and \( \hat{\Sigma} \) are those choices for \( \mu \), \( \Sigma \) that make observing the data actually obtained, most likely.

• Idea behind the LRT for some hypothesis \( H_0 \):

– Compare the maximum of the likelihood function assuming \( H_0 \) is true (the constrained maximum likelihood) and the maximum of the likelihood without assuming \( H_0 \) is true (the unconstrained maximum likelihood).

– If the constrained ML is “significantly” smaller then we conclude that assuming \( H_0 \) makes our data much less likely than not assuming \( H_0 \); \( \Rightarrow \) we should reject \( H_0 \).
Application to $H_0 : \mu = \mu_0$ for multivariate normal data:

- The unconstrained ML is the likelihood evaluated at the MLEs:

$$\max_{\mu, \Sigma} L(\mu, \Sigma) = L(\hat{\mu}, \hat{\Sigma})$$

$$= \frac{1}{(2\pi)^{np/2} |\hat{\Sigma}|^{n/2}}$$

$$\times \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \hat{\mu})' \hat{\Sigma}^{-1} (x_i - \hat{\mu}) \right]$$

$$= \frac{1}{(2\pi)^{np/2} |\hat{\Sigma}|^{n/2}} \exp \left( -\frac{np}{2} \right)$$

after some matrix algebra.
• The constrained ML is the maximum of the likelihood function when we assume $H_0$ is true; that is, when we assume that $\mu = \mu_0$.

  – Thus the constrained ML is the likelihood evaluated at the constrained MLE, the estimators of $\mu$, $\Sigma$ that assumes $\mu = \mu_0$.

  – The constrained MLEs are $\mu_0$ and the $\Sigma$ that maximizes $L(\mu_0, \Sigma)$, which turns out to be (calculus)

  $\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)(x_i - \mu_0)'$.

  – Therefore the constrained ML is

    $\max_\Sigma L(\mu_0, \Sigma) = L(\mu_0, \hat{\Sigma}_0)$

    $= \frac{1}{(2\pi)^{np/2} |\hat{\Sigma}_0|^{n/2}}$

    $\times \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu_0)'\hat{\Sigma}_0^{-1}(x_i - \mu_0) \right]$

    $= \frac{1}{(2\pi)^{np/2} |\hat{\Sigma}_0|^{n/2}} \exp \left( -\frac{np}{2} \right)$
• We compare constrained ML and unconstrained ML by looking at their ratio:

\[ \Lambda = \frac{\text{constrained ML}}{\text{unconstrained ML}} \]
\[ = \frac{1}{(2\pi)^{np/2} |\hat{\Sigma}_0|^n/2} \exp\left(-\frac{np}{2}\right) \]
\[ = \left( \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} \right)^{n/2}. \]

• We reject $H_0$ for small values of $\Lambda$.

*How is this equivalent to $T^2$?*

• With some algebra we can establish that

\[ \Lambda = \left( 1 + \frac{T^2}{n - 1} \right)^{-n/2} \]

a **monotone** function of $T^2$. That is, increases in $T^2$ correspond to decreases in $\Lambda$. $\Rightarrow$ rejecting for large values of $T^2$ is equivalent to rejecting for small values of $\Lambda$. 

101
• Also equivalent to rejecting for small values of

\[ \Lambda^{2/n} = \frac{|\Sigma|}{|\hat{\Sigma}_0|} \]

which is known as Wilk’s Lambda.

• We can also express \( T^2 \) in terms of \( \Lambda \):

\[ T^2 = (n - 1)(\Lambda^{-2/n} - 1) \]

\[ = \frac{(n - 1)|\hat{\Sigma}_0|}{|\hat{\Sigma}|} - (n - 1) \]

• \( \Rightarrow \) Hotelling’s \( T^2 \), Wilk’s Lambda and the likelihood ratio test are equivalent. Essentially, Hotelling’s \( T^2 \) and Wilk’s Lambda are the LRT.
Example – Kite Data:

\[
\hat{\Sigma} = \frac{44}{45} \mathbf{S} = \begin{pmatrix} 118.01 & 119.63 \\ 119.63 & 203.91 \end{pmatrix}
\]

\[\Rightarrow |\hat{\Sigma}| = 9752.89. \text{ From the formula}
\]

\[
\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)'
\]

with \(\mu_0 = (190, 275)'\), we can obtain

\[
\hat{\Sigma}_0 = \begin{pmatrix} 131.13 & 136.93 \\ 136.93 & 226.73 \end{pmatrix} \Rightarrow |\hat{\Sigma}_0| = 10,981.56.
\]

It follows that

\[
\Lambda = \left( \frac{9752.89}{10981.56} \right)^{45/2} = .8881^{45/2} = 0.069.
\]

The critical value can be obtained from the critical value for \(T^2\) (remember that for \(\alpha = 0.05\) it was 6.578):

\[
\Lambda_{\text{crit}} = \left( 1 + \frac{6.578}{44} \right)^{-45/2} = .8699^{45/2} = 0.043.
\]

So the result agrees, 0.069 > 0.043 so, again, we do not reject \(H_0 : \mu = (190, 275)\).
• In this special case (testing $H_0 : \mu = \mu_0$) the LRT happen to be the Hotelling’s $T^2$ and we know the sampling distribution of $T^2$. Therefore, we don’t need the sampling distribution of the likelihood ratio test statistic, $\Lambda$.

• Although there is no general result about the exact distribution of the likelihood ratio test statistic, we have the following approximate result for large sample sizes:

\textbf{Theorem.} Under certain (technical) regularity conditions, for large samples

$$-2 \log(\Lambda) = -2 \log \left( \frac{\text{constrained ML}}{\text{unconstrained ML}} \right) \sim \chi^2(r)$$

where $r =$ the number of restrictions involved in the null hypothesis.
Kite Example (Continued):

The number of restrictions is \( r = 2 \) because the null hypothesis involves restrictions on two means:

Test Statistic: \(-2 \log(\Lambda) = -2 \log(0.069) = 5.339\);

Critical Value: \( \chi^2_{0.05}(2) = 5.991 \);

\( p \)-value: \( p = .069 \)

Compare with the exact result:

Test Statistic: \( T^2 = 5.543 \);

Critical Value: \( (2(44)/43)F_{0.05}(2, 43) = 6.578 \);

\( p \)-value: \( p = .078 \).
Estimation of a Multivariate Mean:

Univariate Case: Since $\sqrt{n}(\bar{x} - \mu)/s \sim t(n - 1)$,

$$1 - \alpha = \Pr \left( \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| \leq t_{\alpha/2}(n - 1) \right)$$

$$= \Pr \left( \bar{x} - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \leq \mu \right)$$

$$\leq \bar{x} + t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}$$

so $\bar{x} \pm t_{\alpha/2}(n - 1)s/\sqrt{n}$ covers $\mu$ with probability $1 - \alpha$.

- Equivalently,

$$\left\{ \mu : \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| < t_{\alpha/2}(n - 1) \right\}$$

is a 100(1 - $\alpha$)% confidence set (in this case an interval) for $\mu$. 

106
Multivariate Case: Similarly, the fact that
\[ n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) \sim \frac{p(n - 1)}{n - p} F(p, n - p) \]
implies
\[ \left\{ \mu : n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) < \frac{p(n - 1)}{n - p} F_\alpha(p, n - p) \right\} \]
is a 100(1 - \alpha)\% confidence region for \( \mu \).

- The equation
  \[ (\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) = \frac{p(n - 1)}{n(n - p)} F_\alpha(p, n - p) \]
defines an ellipsoid
  
  i. centered at \( \bar{x} \);
  ii. with axes along \( e_1, \ldots, e_p \) the eigenvectors of \( S \);
  iii. with the length along \( e_i \) equal to
  \[ \sqrt{\lambda_i} \sqrt{\frac{p(n - 1)}{n(n - p)}} F_\alpha(p, n - p) \]
  where \( \lambda_1, \ldots, \lambda_p \) are the eigenvalues of \( S \).

- Confidence region is an ellipsoid.
Example – Kite Data:

Recall $\bar{x} = (193.62, 279.78)'$ will be the center of confidence ellipsoid.

$$S = \begin{pmatrix} 120.69 & 122.35 \\ 122.35 & 208.54 \end{pmatrix}$$

has eigenvalues

$$\lambda = \frac{(a + c) \pm \sqrt{(a - c)^2 + 4b^2}}{2}$$

$$= \frac{329.24 \pm \sqrt{(-87.85)^2 + 4(122.35)^2}}{2} = 294.61, 34.63$$

To get $e_1, e_2$: for $i = 1, 2$ solve $Sx = \lambda_i x$ or, equivalently, the simultaneous equations

$$120.69x_1 + 122.35x_2 = \lambda_i x_1$$
$$122.35x_1 + 208.54x_2 = \lambda_i x_2$$

Setting $x_1 = 1,$

$$x_2 = \frac{\lambda_i - 120.69}{122.35} = 1.421$$

for $\lambda_1 = 294.61$. Normalizing, we get

$$e_1 = \frac{1}{\sqrt{1^2 + (1.421)^2}} \begin{pmatrix} 1 \\ 1.421 \end{pmatrix} = \begin{pmatrix} 0.575 \\ 0.818 \end{pmatrix}$$
Similarly, we get \( \mathbf{e}_2 = (0.818, -0.575)' \). Since \( F_{0.05}(2, 43) = 3.214 \), the lengths along \( \mathbf{e}_1, \mathbf{e}_2 \) are

\[
\sqrt{294.61} \sqrt{\frac{2(44)}{45(43)}} (3.214) = 6.563 \quad \text{and} \\
\sqrt{34.63} \sqrt{\frac{2(44)}{45(43)}} (3.214) = 2.250
\]

Now we can sketch the confidence region:
Drawbacks:

1. Difficult or impossible to draw for $p \geq 3$.

2. Knowing whether a particular value is inside the confidence region is not so simple as in the univariate case. “Is $\mu_0$ in the confidence region?” necessitates testing $H_0 : \mu = \mu_0$.

3. The region is a joint confidence region. Often we want to be able to make confidence statements about individual components as well.

An Alternative – Separate Univariate CIs:

Recall that for each component $x_i$ of $\mathbf{x}$, the univariate $100(1 - \alpha)\%$ confidence interval for $\mu_i$ is given by

$$\bar{x} \pm t_{\alpha/2}(n - 1) \sqrt{\frac{s_{ii}}{n}}.$$
• If we combine the $p$ $100(1 - \alpha)\%$ univariate CIs, we do not get a $100(1 - \alpha)\%$ joint CI for $\mu$!

joint cov. prob. $= \Pr(\text{each int covers resp. } \mu_i) \\
= 1 - \Pr \left( \bigcup_{i=1}^{p} \left\{ \mu_i \notin \bar{x}_i \pm t_{\alpha/2}(n - 1)\sqrt{\frac{s_{ii}}{n}} \right\} \right) \\
\geq 1 - \sum_{i=1}^{p} \alpha$

• (This result is known as the Bonferroni Inequality.)
• For example, suppose we have a sample from a bivariate normal population and we are interested in estimating the mean $\mu = (\mu_1, \mu_2)'$. If we form a joint confidence region for $\mu$ by combining the univariate 95% confidence intervals, we will get a rectangular joint confidence region:

Assuming $x_1, x_2$ are not independent, the only confidence statement that we can make about the joint region is that the probability that both $\mu_1$ and $\mu_2$ are in the region is at least 90%.
• **Bonferroni Approach:** Use a significance level of $\alpha/p$ for each of the $p$ separate univariate CIs. I.e., use

$$
\bar{x}_i \pm t_{\alpha/(2p)}(n-1) \sqrt{\frac{s_{ii}}{n}}
$$

as the CI for each $\mu_i$.

  – Bonferroni approach ensures that

$$
joint \ cov. \ prob. \geq 1 - \sum_{i=1}^{p} \alpha/p = 1 - \alpha.
$$

  – Drawback: If $p$ is large each component interval will be very wide (too conservative).
In addition to desiring CIs for each of the components of \( \mu \), we may be interested in estimating (and forming a CI around) linear combinations of the \( \mu_i \)'s of the form \( a' \mu \).

- E.g., for the kite data, we may want to estimate the mean difference between wing length and tail length; the mean of \( y = x_1 - x_2 = a'x \), where \( a' = (1, -1) \).

From our sample \( x_1, \ldots, x_n \) we have a sample of the linear combinations: \( y_1, \ldots, y_n \), where \( y_i = a'x_i \). From what we know about multivariate normal distributions:

\[
y_1, \ldots, y_n \overset{iid}{\sim} N(a'\mu, a'\Sigma a), \\
\bar{y} = a'\bar{x}, \quad \text{and} \quad s^2_y = a'Sa.
\]

Therefore,

\[
\frac{a'(\bar{x} - \mu)}{\sqrt{a'Sa/n}} \sim t(n - 1)
\]

and

\[
a'\bar{x} \pm t_{\alpha/2}(n - 1)\sqrt{a'Sa/n} \quad (*)
\]

is a \( 100(1 - \alpha)\% \) CI for \( a'\mu \).
• **Maximal $t$–statistic Approach:** the probability statement underlying (*) is
\[
\Pr \left( \frac{|a'(\bar{x} - \mu)|}{\sqrt{a'Sa/n}} < t_{\alpha/2}(n-1) \right) = 1 - \alpha.
\]
Question: is there a value $c$ such that
\[
\Pr \left( \bigcap_{j=1}^{\infty} \frac{|a'_j(\bar{x} - \mu)|}{\sqrt{a'_jSa_j/n}} < c \right) \geq 1 - \alpha
\]
for all possible linear combinations (all $a_1, a_2, \ldots$)?

• If there were, then we could estimate any and all linear combinations of the component means, even those that were suggested by the data, without our joint coverage probability dropping below $1 - \alpha$.

• There is such a $c$, the maximum over all possible $a$ of $|a'(\bar{x} - \mu)|/\sqrt{a'Sa/n}$. (Why?)

• Doing the maximization we get
\[
\max_a \frac{|a'(\bar{x} - \mu)|}{\sqrt{a'Sa/n}} = \sqrt{n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)} = \sqrt{T^2}.
\]
• So we can use the square root of the critical value of $T^2$,

$$\sqrt{\frac{p(n - 1)}{n - p} F_{\alpha}(p, n - p)}$$

to obtain simultaneous confidence intervals for any and all linear combinations $a' \mu$:

$$a' \bar{x} \pm \sqrt{\frac{p(n - 1)}{n - p} F_{\alpha}(p, n - p)} \sqrt{a'Sa/n}$$

• In particular, for a component mean $\mu_i$ this procedure gives

$$\bar{x}_i \pm \sqrt{\frac{p(n - 1)}{n - p} F_{\alpha}(p, n - p)} \sqrt{S_{ii}/n}$$

• Typically, when interested in only estimating component means, Bonferroni method is better. To allow data snooping or when the number of intervals/tests is large, should use Max $t$ approach.
Example – Reaction Time:

Reaction times to visual stimuli were obtained from $n = 20$ normal young men under three conditions, $A$, $B$, $C$, of stimulus display. The sample mean vector and variance-covariance matrix for these data are

$$
\bar{x} = \begin{pmatrix} 21.05 \\ 21.65 \\ 28.95 \end{pmatrix}, \quad S = \begin{pmatrix} 2.2605 & 2.1763 & 1.6342 \\ 2.1763 & 2.6605 & 1.8237 \\ 1.6342 & 1.8237 & 2.4710 \end{pmatrix}
$$

Suppose we want to

1. obtain individual CIs for $\mu_1 = \text{E}(x_1)$, $\mu_2 = \text{E}(x_2)$, and $\mu_3 = \text{E}(x_3)$;
2. compare conditions $A$ and $B$ with condition $C$ (control, perhaps);
3. and after seeing the data suppose that we see a possible unexpected difference between groups $A$ and $B$. We’d like a post hoc CI for $\mu_A - \mu_B$;

all without exceeding a joint significance level of $\alpha = 0.05$. 

117
Max-\( t \) approach:

\[
\frac{p(n - 1)}{n - p} F_{\alpha} (p, n - p) = \frac{3(19)}{17} F_{0.05} (3, 17)
\]

\[= (3.353)(3.20) = 10.730\]

1. Individual CIs:

\[
21.05 \pm \sqrt{10.730 \cdot 2.2605/20} = (19.95, 22.15)
\]

\[
21.65 \pm \sqrt{10.730 \cdot 2.6605/20} = (20.46, 22.84)
\]

\[
28.95 \pm \sqrt{10.730 \cdot 2.4710/20} = (27.80, 30.10)
\]

2. Estimate \( \frac{1}{2}(\mu_A + \mu_B) - \mu_C = (\frac{1}{2}, \frac{1}{2}, -1) \mu \):

\[
\left( \frac{1}{2}, \frac{1}{2}, -1 \right) \bar{x} = \frac{1}{2} (21.05) + \frac{1}{2} (21.65) - 28.95
\]

\[= -7.60\]

\[
\left( \frac{1}{2}, \frac{1}{2}, -1 \right) S \left( \begin{array}{c} 1/2 \\ 1/2 \\ -1 \end{array} \right) = \left( \frac{1}{2}, \frac{1}{2}, -1 \right) \left( \begin{array}{c} 0.5842 \\ 0.5947 \\ -0.7421 \end{array} \right)
\]

\[= 1.3315\]
so our CI is

\[-7.60 \pm \sqrt{10.730} \sqrt{1.3315/20} = (-8.45, -6.75).\]

3. Estimate \(\mu_A - \mu_B = (1, -1, 0)\mu:\)

\[(1, -1, 0)\bar{x} = 21.05 - 21.65 = -0.60\]

\[(1, -1, 0)S \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = (1, -1, 0) \begin{pmatrix} 0.0842 \\ -0.4842 \\ -0.1895 \end{pmatrix} \]

\[= 0.5684\]

so our CI is

\[-0.60 \pm \sqrt{10.730} \sqrt{0.5684/20} = (-1.15, -0.048).\]
Large Sample Result:

- The CIs and tests we have considered thus far rely on the assumption of multivariate normality. Under normality, confidence levels and $p$-values are exact, because the distribution results are exact.

- Recall from an earlier result, for large samples

$$n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) \sim \chi^2(p)$$

regardless of the distribution of $\mathbf{x}$.

- Therefore, nonnormality can be overcome by a large sample size and we can base inference on the $\chi^2$ distribution.
• An approximate $100(1-\alpha)\%$ CI for $\mathbf{a'}\mathbf{\mu}$ is given by
\[
\mathbf{a'}\bar{x} \pm \sqrt{\chi_{\alpha}^2(p)} \sqrt{\mathbf{a'Sa}/n}
\]
or, for $\mu_i$,
\[
\bar{x}_i \pm \sqrt{\chi_{\alpha}^2(p)} \sqrt{s_{ii}/n}
\]

• An approximate $\alpha$–level large-sample test of $H_0 : \mathbf{\mu} = \mathbf{\mu}_0$ is reject if
\[
-2 \log(\Lambda) > \chi_{\alpha}^2(p)
\]
$(-2 \log(\Lambda)$ is approximately equal to $T^2$ in large samples).
Inference about a Pop. Cov. Matrix

Model: (same as before) $x_1, \ldots, x_n \sim N_p(\mu, \Sigma)$, $\mu$, $\Sigma$ unknown.

- Its unusual that we would be interested in testing $H_0 : \Sigma = \Sigma_0$. Its more likely that we want to test

$$H_0 : \Sigma = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}$$

versus

$$H_1 : \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

That is, we may want to test independence with unequal variances versus a general, unconstrained covariance structure.
• The null hypothesis here may be equivalently expressed as the hypothesis of uncorrelated components: \( H_0 : \mathbf{P} = \mathbf{I} \) where \( \mathbf{P} \) is the population correlation matrix of each \( \mathbf{x}_i \) and \( \mathbf{I} \) is the identity matrix.

• LRT: We look at the ratio of the constrained maximum likelihood versus the unconstrained maximum of the likelihood. Recall the unconstrained ML was

\[
\max_{\mu, \Sigma} L(\mu, \Sigma) = L(\hat{\mu}, \hat{\Sigma}) = \frac{1}{|\hat{\Sigma}|^{n/2}} \exp(-np/2)
\]

In this case the constrained ML is

\[
\max_{\mu, \Sigma_0} L(\mu, \Sigma_0) = \prod_{i=1}^{p} \frac{1}{\hat{\sigma}_{ii}^{n/2}} \exp(-np/2)
\]

where \( \Sigma_0 \) is the value of \( \Sigma \) under the null hypothesis.

• The ratio of these quantities simplifies to

\[
\Lambda = \frac{|S|^{n/2}}{\prod_{i=1}^{p} (s_{ii})^{n/2}} = |\mathbf{R}|^{n/2}
\]

We reject \( H_0 \) if \( |\mathbf{R}|^{n/2} \) is smaller than some critical value.
• In large samples we have the result that the LR is approximately \( \chi^2(p(p - 1)/2) \) distributed (there are \( p(p - 1)/2 \) off-diagonal elements of \( \Sigma \) which are constrained by \( H_0 \) to be equal to 0).

• Rejection rule: we reject \( H_0 \) at significance level \( \alpha \) if

\[
-2 \log(\Lambda) = -n \log |\mathbf{R}| > \chi^2_{\alpha}(\frac{p(p - 1)}{2})
\]

Example - Kite Data:

Recall \( n = 45, \ p = 2 \) variables so \( p(p - 1)/2 = 1 \) and

\[
\mathbf{R} = \begin{pmatrix} 1 & 0.771 \\ 0.771 & 1 \end{pmatrix}
\]

so

\[
-2 \log(\Lambda) = -n \log |\mathbf{R}| = -45 \log(1 - (0.771)^2) = -45 \log(0.405) = 40.641
\]

Since \( \chi^2_{0.05}(1) = 3.841 \) we reject \( H_0 \) (as you should expect).
Comparisons of 2 or More Mean Vectors
(See chapter 6 of Johnson and Wichern.)

Generalizing the paired $t$–test:

Recall the univariate paired $t$–test:

Suppose we are interested in reaction time and we have two methods of stimulus presentation that we’d like to compare. Suppose 40 subjects are available for an experiment.

- **Design 1:** Reaction time measured on 20 subjects receiving presentation method 1, reaction time measured on remaining 20 subjects who receive presentation method 2.

  $x_{1i} = \text{R.T. for } i^{th} \text{ subject in group 1, } i = 1, \ldots, 20.$

  $x_{2i} = \text{R.T. for } i^{th} \text{ subject in group 2, } i = 1, \ldots, 20.$
• **Design 2:** All 40 subjects are measured for R.T. using presentation method 1, *then* all subjects are measured again using presentation method 2.

\[ x_{1i} = \text{R.T. for } i^{\text{th}} \text{ subject when presentation method 1 is used, } i = 1, \ldots, 40. \]

\[ x_{2i} = \text{R.T. for } i^{\text{th}} \text{ subject when presentation method 2 is used, } i = 1, \ldots, 40. \]

(How could we improve on design 2?)

*Assuming there is no order effect, which design is better?*

**Answer:** design 2.

• Comparisons can be made within subjects, eliminating subject-subject differences.
• Allows for more precise estimate of treatment difference.
• Allows for more power to detect a treatment effect.
Two-sample \( t \)-test: To test \( H_0 : \mu_1 = \mu_2 \) where \( \mu_j = \mathbb{E}(x_{ji}) \) for all \( i \), we compute the test statistic

\[
t = \frac{\overline{x}_1 - \overline{x}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

and compare to its distribution, \( t(n_1 + n_2 - 2) \).

- Two-sample \( t \) test assumes:
  1. \( x_{11}, \ldots, x_{1n_1} \sim iid N(\mu_1, \sigma_1^2) \) and \( x_{21}, \ldots, x_{2n_2} \sim iid N(\mu_2, \sigma_2^2) \).
  2. \( \sigma_1^2 = \sigma_2^2 \) (\( s_P \) is the pooled estimate of the common variance)
  3. Independent samples

- Because of point 3, two sample \( t \) is appropriate for design 1 but not for design 2.
Paired \( t \)-test: We look at within subject differences in reaction time for the two presentation methods.

- Define \( d_i = x_{1i} - x_{2i}, i = 1, \ldots, n \).
- We assume \( d_1, \ldots, d_n \sim iid N(\mu_d, \sigma^2_d) \).
- We test \( H_0 : \mu_d = 0 \), the hypothesis that the mean difference between the treatments is 0, versus \( H_1 : \mu_d \neq 0 \).
- Our test statistic is now just the one-sample \( t \) (we have a single sample of differences):

\[
t = \frac{\bar{d} - \mu_{d0}}{s_d / \sqrt{n}}
\]

which we compare against its distribution, \( t(n-1) \).
- A \( 100(1 - \alpha)\% \) CI for the mean difference, \( \mu_d \), is given by

\[
\bar{d} \pm t_{\alpha/2}(n - 1)s_d / \sqrt{n}
\]

where

\[
\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i, \quad s^2_d = \frac{1}{n - 1} \sum_{i=1}^{n} (d_i - \bar{d})^2.
\]
Multivariate Generalization:

- The analogous multivariate scenario is when we have a $p$–dimensional vector measured on each of $n$ subjects in two treatment groups:

\[
\begin{align*}
x_{11i} &= \text{variable 1 for } i^{th} \text{ subj. getting trt 1} \\
x_{12i} &= \text{variable 2 for } i^{th} \text{ subj. getting trt 1} \\
&\vdots \\
x_{1pi} &= \text{variable } p \text{ for } i^{th} \text{ subj. getting trt 1} \\
x_{21i} &= \text{variable 1 for } i^{th} \text{ subj. getting trt 2} \\
x_{22i} &= \text{variable 2 for } i^{th} \text{ subj. getting trt 2} \\
&\vdots \\
x_{2pi} &= \text{variable } p \text{ for } i^{th} \text{ subj. getting trt 2}
\end{align*}
\]

- For the $i^{th}$ subject, we define a vector of differences:

\[
d_i = \begin{pmatrix} d_{1i} \\ d_{2i} \\ \vdots \\ d_{pi} \end{pmatrix} = \begin{pmatrix} x_{11i} - x_{21i} \\ x_{12i} - x_{22i} \\ \vdots \\ x_{1pi} - x_{2pi} \end{pmatrix}
\]

for $i = 1, \ldots, n$. 
• We assume that \( d_1, \ldots, d_n \overset{iid}{\sim} N_p(\mu_d, \Sigma_d) \).

• To test \( H_0 : \mu_d = 0 \) we know what to do: Hotelling’s \( T^2 \) test. Compute

\[
T^2 = n(\bar{d} - \mu_{d0})' S_d^{-1}(\bar{d} - \mu_{d0}), \text{ in general}
\]

\[
= n\bar{d}' S_d^{-1} \bar{d}, \text{ for } \mu_{d0} = 0
\]

and compare with its distribution:

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

• Once we have framed the problem as a one-sample problem (we have a single sample of \( n \) \( p \)-vectors), then there is nothing new here. Confidence regions can be formed as in chapter 5:

- A 100(1 – \( \alpha \))% confidence region for \( \mu_d \) consists of all \( \mu_d \) such that

\[
(\bar{d} - \mu_d)' S_d^{-1}(\bar{d} - \mu_d) \leq \frac{p(n - 1)}{n(n - p)} F_{\alpha}(p, n-p)
\]
- 100(1 – \( \alpha \))% max \( t \) simultaneous confidence intervals for the individual mean differences \( \mu_{d_j} \) are given by

\[
\bar{d}_j \pm \sqrt{\frac{p(n-1)}{n-p} F_{\alpha}(p, n-p)} \sqrt{\frac{s_{d_j}^2}{n}},
\]

where \( \bar{d}_j \) is the \( j^{th} \) element of \( \bar{d} \) and \( s_{d_j}^2 \) is the \( j^{th} \) diagonal element of \( S_d \).

- The Bonferroni 100(1 – \( \alpha \))% simultaneous confidence intervals for the individual mean differences \( \mu_{d_j} \) are given by

\[
\bar{d}_j \pm t_{\alpha/(2p)}(n - 1) \sqrt{\frac{s_{d_j}^2}{n}}, i = 1, \ldots, n.
\]
Example – Effluent Data:

Municipal wastewater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent water were divided and sent to two laboratories for testing. One-half of each sample was sent to the Wisconsin State Laboratory of Hygiene and one-half of each sample was sent to a private commercial laboratory. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for the $n = 11$ samples from each lab. The data are as follows:
<table>
<thead>
<tr>
<th>Sample $i$</th>
<th>Commercial Lab</th>
<th>State Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{11i}$ (BOD)</td>
<td>$x_{21i}$ (BOD)</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>54</td>
</tr>
<tr>
<td>9</td>
<td>43</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>

- Do the two labs agree? If not, how so?

- See handouts effluent.sas effluent.lst.

- From effluent.lst we see that $T^2 = 13.64$. The critical value for this test is

\[
\frac{p(n - 1)}{n - p} F_{\alpha}(p, n - p) = \frac{2(10)}{9} F_{0.05}(2, 9) = 9.47
\]

using a significance level of 0.05.
• Conclusion: there is a significant mean difference between the measurements of the two labs.

• Looking at the means, we see that the commercial lab tends to produce lower BOD measurements and higher SS measurements.

• The normality assumption should be checked for these data.

• See example 6.1 in the text for further discussion of this example.
- It is good practice to compute the differences $d_{i1}, \ldots, d_{ip}$ for each sample member $i$ and compute the paired $T^2$ test as described above. This allows checks of normality for each component of the vector of differences.

- However, the paired $T^2$ test can be computed another way by taking advantage of what we know about linear combinations:

Define $\mathbf{x}_i = (x_{11i}, x_{12i}, \ldots, x_{1pi}, x_{21i}, x_{22i}, \ldots, x_{2pi})'$. Then $\mathbf{x} = (\bar{x}_{11}, \ldots, \bar{x}_{1p}, \bar{x}_{21}, \ldots, \bar{x}_{2p})'$ and

$$
\mathbf{S} = \begin{pmatrix}
\mathbf{S}_{11} & \vdots & \mathbf{S}_{12} \\
(2p \times 2p) & \vdots & (2p \times 2p) \\
\mathbf{S}_{21} & \vdots & \mathbf{S}_{22} \\
(2p \times 2p) & \vdots & (2p \times 2p)
\end{pmatrix},
$$

where $\mathbf{S}_{11}$ contains the variances and covariances of the $p$ variables measured on treatment 1, $\mathbf{S}_{22}$ contains the variances and covariances of the $p$ variables measured on treatment 2, and $\mathbf{S}_{12} = \mathbf{S}_{21}'$ contains covariances between treatment 1 variables and treatment 2 variables.
If we define the matrix

\[ C_{p \times 2p} = (I_p | (-1)I_p) \]

(\(|\) means combine together horizontally) then it is easily established that

\[ d_i = Cx_i, \quad i = 1, \ldots, n, \]
\[ \bar{d} = C\bar{x}, \quad \text{and} \quad S_d = CSC' \]

Thus \( T^2 \) can be computed from \( \bar{x} \) and \( S \) without computing the differences:

\[ T^2 = n\bar{x}'C'(CSC')^{-1}C\bar{x} \]

Example – Effluent Data (Continued) See effluent.sas and effluent.lst.
A Simple Repeated Measures Design:

A generalization of the univariate paired $t$–test scenario occurs when a random variable is measured on each of $n$ subjects $q$ times, possibly under different experimental conditions.

In this situation we have a vector of repeated measures

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{qi} \end{pmatrix}$$

for each subject $i = 1, \ldots, n$. Here, for example, $x_{2i}$ is the response for the $i^{th}$ subject on the $2^{nd}$ occasion ($2^{nd}$ treatment, perhaps).

In the univariate paired $t$–test we had $q = 2$ and the $t$– test could be performed by defining differences

$$d_i = c'x_i = \begin{pmatrix} c_1 & c_2 \end{pmatrix} \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = x_{1i} - x_{2i}$$

based on a contrast vector, $c = (1, -1)'$.  

137
The \( d_i \)'s each have mean \( \mathbb{E}(d_i) = \mu_d = c'\mu_x \) and variance \( \text{var}(d_i) = c'\Sigma_x c \) that we estimate by \( \bar{d} = c'\bar{x} \) and \( s_d^2 = c'S_x c \), respectively. Our univariate paired \( t \)-test statistic for \( H_0 : \mu_d = 0 \) is

\[
\frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{\sqrt{n}c'\bar{x}}{\sqrt{c'S_x c}} = \sqrt{n(c'\bar{x})(c'S_x c)^{-1}(\bar{x}'c)}
\]

We generalize this procedure to our repeated measures design where \( q > 2 \) by considering contrasts of the components of \( \mu = \mathbb{E}(x_i) \). We can express the null hypothesis of equal means across time (or across treatments if the repeated measures correspond to different treatments) as \( H_0 : C\mu = 0 \) if we choose \( C \) so that

\[
\begin{pmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & -1
\end{pmatrix}

\begin{pmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_q
\end{pmatrix}

= \begin{pmatrix}
\mu_1 - \mu_2 \\
\mu_2 - \mu_3 \\
\vdots \\
\mu_{q-1} - \mu_q
\end{pmatrix}
\]
or

\[
\begin{pmatrix}
1 & -1 & 0 & \cdots & 0 \\
1 & 0 & -1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & -1
\end{pmatrix}
\begin{pmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_q
\end{pmatrix}
= \begin{pmatrix}
\mu_1 - \mu_2 \\
\mu_1 - \mu_3 \\
\vdots \\
\mu_1 - \mu_q
\end{pmatrix}
\]

C must be a **contrast matrix**, meaning that it must have \( q - 1 \) linearly independent rows each of which is (the transpose of) a contrast vector.

- A **contrast vector** is a vector which has components that sum to 0.

For \( \mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_q(\mathbf{\mu}, \Sigma) \) and \( \mathbf{C} \) a contrast matrix, an \( \alpha \)-level test of \( H_0 : \mathbf{C}\mathbf{\mu} = \mathbf{0} \) (equal means across measurement occasions) versus \( H_1 : \mathbf{C}\mathbf{\mu} \neq \mathbf{0} \) is: Reject \( H_0 \) if

\[
T^2 = n (\mathbf{C}\bar{\mathbf{x}})' (\mathbf{C}\mathbf{\Sigma}\mathbf{C}' )^{-1} \mathbf{C}\bar{\mathbf{x}} > \frac{(q - 1)(n - 1)}{n - (q - 1)} F_{\alpha}(q - 1, n - (q - 1))
\]
A $100(1 - \alpha)\%$ confidence region (ellipsoid) for a contrast $C\mu$ (a vector) is given by the set of all $C\mu$ such that

$$n(C\bar{x} - C\mu)'(CSC')^{-1}(C\bar{x} - C\mu)$$

$$\leq \frac{(q - 1)(n - 1)}{n - q + 1}F_\alpha(q - 1, n - q + 1)$$

Based on the max $t$ approach, $100(1 - \alpha)\%$ confidence intervals for single contrasts of the form $c'\mu$ (a scalar) are given by

$$c'\bar{x} \pm \sqrt{\frac{(q - 1)(n - 1)}{n - q + 1}F_\alpha(q - 1, n - q + 1)} \sqrt{\frac{c'Sc}{n}}$$
Example – Metabolism in the Brain:

Metabolic rates in the left and right thalamus and the left and right frontal eyefield regions of the brain were measured by positron emission tomography for \( n = 17 \) normal subjects. These means and covariance matrix were computed from the rates:

\[
\begin{array}{ccc|ccc}
\text{Thalamus} & & \text{Eyefield} & & \\
\text{Left} & \text{Right} & \text{Left} & \text{Right} \\
\hline
\mu' = & \mu_1 & \mu_2 & \mu_3 & \mu_4 \\
\bar{x}' = & 4.535 & 4.703 & 4.902 & 5.221 \\
S = & 2.0843 & 2.1698 & 1.9942 & 2.1361 \\
& & 2.3002 & 2.0823 & 2.2282 \\
& & & 2.1783 & 2.3279 \\
& & & & 2.5412 \\
\end{array}
\]

- We’d like to test the hypothesis that the mean metabolic rate is the same in all four locations of the brain.

- If the means differ, where are the significant differences?
Here, there are three linearly independent treatment contrasts of interest:

\[
\begin{align*}
  c_1' \mu &= (\mu_1 + \mu_2) - (\mu_3 + \mu_4) \\
  c_2' \mu &= (\mu_1 + \mu_3) - (\mu_2 + \mu_4) \\
  c_3' \mu &= (\mu_1 - \mu_2) - (\mu_3 - \mu_4)
\end{align*}
\]

Interpretations:

\textbf{c}_1: \text{ Difference between rates in the thalamus and the frontal eyefield region.}

\textbf{c}_2: \text{ Difference between rates on the left and right sides.}

\textbf{c}_3: \text{ Interaction contrast; Does left vs. right difference depend on region?}

We arrange these contrasts in a matrix

\[
C = \begin{pmatrix}
  c_1' \\
  c_2' \\
  c_3'
\end{pmatrix} = \begin{pmatrix}
  1 & 1 & -1 & -1 \\
  1 & -1 & 1 & -1 \\
  1 & -1 & -1 & 1
\end{pmatrix}
\]

and we test the hypothesis of equal means by testing \( H_0 : C \mu = 0. \)
• See handouts brain1.sas and brain1.lst.

• \( T^2 = 41.92 \) exceeds the critical value 11.46, so we reject the null hypothesis of equal mean metabolic rates in the four sites.

To determine specifically where the significant differences among the means are we form simultaneous 95% confidence intervals for \( c_k \mu \), \( i = 1, 2, 3. \)

• From brain1.lst we have the three intervals \((-1.79, 0.021), (-0.77, -0.21), (-0.11, 0.41)\) based on the max-\( t \) approach, and the three intervals \((-1.60, -0.17), (-0.71, -0.27), (-0.055, 0.36)\) based on the Bonferroni approach. Assuming that the three comparisons corresponding to \( c_1, c_2, c_3 \) were all planned comparisons, the Bonferroni approach is appropriate here and preferrable because of its narrower intervals. Based on the Bonferroni intervals, only the interval for \( c'_3 \mu \) straddles 0. Therefore, there are significant differences between the thalamus and eyefield regions and between the hemispheres, but there is not significant evidence of an interaction between region and hemisphere.
• The above analysis of repeated measures based on the $T^2$ statistic does not assume any particular structure on the covariance matrix $\Sigma$.

• If the covariance matrix can reasonably be assumed to have a particular structure, more powerful inferential techniques exist. In particular, under the assumption that $x_1, \ldots, x_n$ each have variance-covariance matrix $\Sigma$ with the \textbf{compound symmetry} structure:

\[
\Sigma = \sigma^2 \begin{pmatrix}
1 & \rho & \rho & \cdots & \rho \\
\rho & 1 & \rho & \cdots & \rho \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \rho & \cdots & 1
\end{pmatrix},
\]

the repeated measures design just discussed may be analyzed as a randomized block design, with subjects corresponding to blocks.
Comparing Mean Vectors from 2 Populations

We can generalize the two sample $t$ test for a single mean in each of two populations to the two sample $T^2$ test for comparing two mean vectors.

Setup: We have two samples of multivariate observations of size $n_1$ and $n_2$ from two populations:

$$\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$$

with sample means

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_{1i}, \quad \bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{x}_{2i}$$

and sample variance-covariance matrices

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'$$

$$\mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)'$$
Assumptions:

1. $x_{11}, \ldots, x_{1n_1} \overset{iid}{\sim} N_p(\mu_1, \Sigma_1)$.

2. $x_{21}, \ldots, x_{2n_2} \overset{iid}{\sim} N_p(\mu_2, \Sigma_2)$.

3. Independent samples ($x_{11}, \ldots, x_{1n_1}$ are independent of $x_{21}, \ldots, x_{2n_2}$).

4. Population covariance matrices are equal ($\Sigma_1 = \Sigma_2$).

- We wish to test $\mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$ or, more generally, $H_0 : \mu_1 - \mu_2 = \delta_0$ versus $H_1 : \mu_1 - \mu_2 \neq \delta_0$.

- Assumptions are analogous to the univariate assumptions of the two sample $t$-test. However, notice that assumption 4 is much stronger in the multivariate case. Assuming $\Sigma_1 = \Sigma_2$ requires that $p(p+1)/2$ pairs of variances and covariance are equal.
We know that $S_1$ estimates $\Sigma_1$ and $S_2$ estimates $\Sigma_2$. Assuming 4, the two samples have a common variance-covariance matrix $\Sigma$, say. So, $\Sigma_1 = \Sigma_2 = \Sigma$ and $S_1$ and $S_2$ are both estimates of $\Sigma$.

**Which estimate of $\Sigma$ do we use?**

Answer: We *pool* the two estimates $S_1$ and $S_2$ by using $S_P$ a weighted average of $S_1, S_2$

$$S_P = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

(a *weighted* average because $S_1$ and $S_2$ are based on different sample sizes).

The LR test statistic for $H_0 : \mu_1 - \mu_2 = \delta_0$ is

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_P \right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0)$$

and we reject $H_0$ for large values of $T^2$.

- Notice $T^2$ is a squared statistical distance between $\bar{x}_1 - \bar{x}_2$ and the null value of $\mu_1 - \mu_2$. If the distance is large, we reject.
• Notice that $T^2$ generalizes the one sample test statistic which we also called $T^2$.

How large must $T^2$ be to reject?

We can get the distribution of $T^2$ by looking at the form of the statistic.

• By assumption 3 the $x_{1i}$’s and the $x_{2i}$’s are independent so $(n_1 - 1)S_1$ and $(n_2 - 1)S_2$ are independent with Wishart distributions, $W_p(\Sigma, n_1 - 1)$ and $W_p(\Sigma, n_2 - 1)$, respectively.

• By property 2 of Wishart distributions (p. 72 of notes)

$$(n_1 - 1)S_1 + (n_2 - 1)S_2 \sim W_p(\Sigma, n_1 + n_2 - 2)$$

• We also know that $\bar{x}_1, \bar{x}_2$ are independent multivariate normals with the same variance covariance matrix, so linear combinations such as $\bar{x}_1 - \bar{x}_2$ will be multivariate normal (result 4.8, p.174 of text). Specifically,

$$\bar{x}_1 - \bar{x}_2 \sim N_p \left( \mu_1 - \mu_2, \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \Sigma \right).$$
Combining these distributional results we see that $T^2$ is of the form

$$T^2 = (N_p(0, \Sigma))'(W_p(\Sigma, \text{d.f.})/\text{d.f.})^{-1}(N_p(0, \Sigma))$$

so

$$T^2 \sim \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F(p, n_1 + n_2 - p - 1)$$

- We reject $H_0 : \mu_1 - \mu_2 = \delta_0$ at significance level $\alpha$ if

$$T^2 > \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_\alpha(p, n_1 + n_2 - p - 1)$$
Example – Aging and Intelligence

Forty-nine elderly men participating in an interdisciplinary study of human aging were classified into the diagnostic categories “senile factor present” and “no senile factor” on the basis of an intensive psychiatric examination. The Wechsler Adult Intelligent Scale (WAIS) had been administered to all subjects by an independent investigator, and certain subtests showed large differences between the groups. It was proposed that a test be made of the hypothesis that the groups arose from populations with a common mean vector, and if this hypothesis was rejected, simultaneous confidence intervals would be used to determine which individual subtest means were significantly different.

The pooled sample variance-covariance matrix in this study was

\[
S_P = \begin{pmatrix}
11.2624 & 9.4060 & 7.1550 & 3.3791 \\
13.5265 & 7.3784 & 2.5014 \\
11.5796 & 2.6167 & 5.8133
\end{pmatrix}
\]
The sample means are given in the following table

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Senile Factor</td>
</tr>
<tr>
<td></td>
<td>$n_1 = 37$</td>
</tr>
<tr>
<td>Information</td>
<td>12.57</td>
</tr>
<tr>
<td>Similarities</td>
<td>9.57</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>11.49</td>
</tr>
<tr>
<td>Picture Completion</td>
<td>7.97</td>
</tr>
</tbody>
</table>
To test \( H_0 : \mu_1 - \mu_2 = 0 \) we compute

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

\[
= \begin{pmatrix}
12.57 - 8.75 \\
9.57 - 5.33 \\
11.49 - 8.50 \\
7.97 - 4.75
\end{pmatrix}
\]

\[
\times \left[ \left( \frac{1}{37} + \frac{1}{12} \right) S_P \right]^{-1} \left( \begin{array}{c}
12.57 - 8.75 \\
9.57 - 5.33 \\
11.49 - 8.50 \\
7.97 - 4.75
\end{array} \right)
\]

\[
= 22.13
\]

The critical value for a significance level of 0.05 is

\[
p(n_1 + n_2 - 2) \\
f_{\alpha}(p, n_1 + n_2 - p - 1)
\]

\[
= \frac{4(47)}{44} F_{0.05}(4, 44) = (4.2727)(2.5837) = 11.04
\]

so we reject the null hypothesis of equal mean vectors at the .05 level.

To determine which pairs of subtest means differ we need to form simultaneous confidence intervals around \( \mu_{1j} - \mu_{2j} \), \( j = 1, \ldots, 4 \).
Confidence Regions:

We will also be interested in estimating $\mu_1 - \mu_2$ and forming confidence regions.

A $100(1 - \alpha)\%$ confidence region for $\mu_1 - \mu_2$ is given by the acceptance region of our test; i.e., the set of all $\mu_1 - \mu_2$ such that

$$(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_P \right]^{-1} \times (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)) \leq F_{\text{crit}}$$

where

$$F_{\text{crit}} = \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_\alpha(p, n_1 + n_2 - p - 1).$$

- Confidence region is an ellipsoid
  
  i. centered at $\bar{x}_1 - \bar{x}_2$;
  ii. with axes along $\mathbf{e}_1, \ldots, \mathbf{e}_p$ the eigenvectors of $S_P$;
  iii. with the length along $\mathbf{e}_i$ equal to

  $$\sqrt{\lambda_i} \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right) F_{\text{crit}}}$$

  where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $S_P$.  

153
Simultaneous Confidence Intervals:

A general result for any and all linear combinations \( \mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \) is available similar to the result we gave in the one sample case. With probability \( 1 - \alpha \),

\[
\mathbf{a}'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{F_{\text{crit}} \mathbf{a}' \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_P \mathbf{a}}
\]

will cover \( \mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \) for all \( \mathbf{a} \).

- As a special case, we have simultaneous confidence intervals for \( \boldsymbol{\mu}_{1j} - \boldsymbol{\mu}_{2j}, j = 1, \ldots, p \), given by

\[
(\bar{x}_{1j} - \bar{x}_{2j}) \pm \sqrt{F_{\text{crit}} \mathbf{a}' \left( \frac{1}{n_1} + \frac{1}{n_2} \right) s_{P, jj}}
\]

where \( s_{P, jj} \) is the \((j, j)\)th element of the pooled sample variance-covariance matrix, \( \mathbf{S}_P \).
As an alternative, we can base simultaneous confidence intervals for \( \mu_{1j} - \mu_{2j} \) on the Bonferroni method. The Bonferroni 100(1 - \( \alpha \))% CIs for the \( p \) population mean differences are given by

\[
(\bar{x}_{1j} - \bar{x}_{2j}) \pm t_{\alpha/(2p)}(n_1 + n_2 - 2) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{P,jj}}
\]

**Example – Aging (Continued):**

Based on our \( T^2 \) test, we have already concluded that the vector of WAIS subtest means differs between groups (senile factor present versus absent).

To form 95% simultaneous Bonferroni intervals for \( \mu_{1j} - \mu_{2j}, \ j = 1, \ldots, 4 \), we need

\[
t_{\alpha/(2p)}(n_1 + n_2 - 2) = t_{0.05/8}(37 + 12 - 2)
= t_{0.00625}(47) = 2.60.
\]
Our Bonferroni intervals are

\[(12.57 - 8.75) \pm 2.60 \sqrt{\left(\frac{1}{37} + \frac{1}{12}\right) 11.2624} = 3.82 \pm (2.60)(1.11) = (0.92, 6.72)\]

for the difference in mean scores on the information test, and, similarly, we obtain the intervals

\[(9.57 - 5.33) \pm (2.60)(1.22) = (1.07, 7.41)\]
\[(11.49 - 8.50) \pm (2.60)(1.13) = (0.054, 5.93)\]
\[(7.97 - 4.75) \pm (2.60)(0.80) = (1.14, 5.30)\]

for \(\mu_{12} - \mu_{22}, \mu_{13} - \mu_{23},\) and \(\mu_{14} - \mu_{24},\) respectively.

Conclusions: There are significant differences on all four subtests with the “no senile factor” group consistently outperforming the “senile factor present” group. The largest difference between the groups is on the similarities subtest.
Example – Sparrow Data:

Recall the sparrow data where, following a storm, 5 variables were measured on a sample of sparrows, and their subsequent survival was tracked. The variables were $x_1 =$ total length, $x_2 =$ alar extent, $x_3 =$ length of beak and head, $x_4 =$ length of humerus, $x_5 =$ length of keel of sternum. Birds 1–21 survived, the remaining 28 birds died.

To test $H_0 : \mu_1 = \mu_2$, we compute $T^2$. See sparrow4.sas and sparrow4.lst.

- Out test statistic $T^2 = 2.82$ which we compare against $F_{crit} = 13.29$. Clearly, $T^2$ is far from significant based on a .05 level test. We conclude that there is not a difference between the mean vectors of survivors and non-survivors.
• Simultaneous confidence intervals for the mean differences between the five components based on the max $t$ approach are

$$(-4.89, 2.80) \quad (-5.95, 4.81)$$

$$(-0.89, 0.80) \quad (-0.55, 0.65) \quad (-1.08, 1.02)$$

These intervals all cover 0, which is not surprising given the weakness of the evidence for rejecting $H_0 : \mu_1 = \mu_2$ provided by $T^2$.

• The linear combination $\hat{a}'(\bar{x}_1 - \bar{x}_2)$ with $\hat{a}$ proportional to $S_P^{-1}(\bar{x}_1 - \bar{x}_2)$ quantifies the largest population difference. Therefore, if we reject $H_0 : \mu_1 - \mu_2 = 0$ it makes sense to look at the elements of the vector $S_P^{-1}(\bar{x}_1 - \bar{x}_2)$. The size of these elements tells us the relative importance of the components of $\bar{x}_1 - \bar{x}_2$ in the linear combination of $\bar{x}_1 - \bar{x}_2$ most responsible for rejecting $H_0$. 
• In the sparrow example we did not reject $H_0$. In the aging example we did, and the value of $\hat{a}$ is proportional to

$$S_p^{-1}(\bar{x}_1 - \bar{x}_2)$$

$$= \begin{pmatrix}
11.2624 & 9.4060 & 7.1550 & 3.3791 \\
13.5265 & 7.3784 & 2.5014 & \\
11.5796 & 2.6167 & & \\
& & & 5.8133
\end{pmatrix}^{-1}$$

$$\times \begin{pmatrix}
12.57 - 8.75 \\
9.57 - 5.33 \\
11.49 - 8.50 \\
7.97 - 4.75
\end{pmatrix} = \begin{pmatrix}
0.026 \\
0.208 \\
0.009 \\
0.445
\end{pmatrix}$$

• The coefficients are all positive, indicating that all components contributed in the same direction to rejecting $H_0$.

• The first and third coefficients are relatively very small, indicating that in the linear combination most responsible for rejecting $H_0$, differences in subtests 1 and 3 played a relatively minor role.
About the Assumptions:

We assumed

- Multivariate normality of each sample.
- Equal covariance matrices.

These assumptions are unnecessary if we have large sample sizes because of the following result:

If $n_1 - p$ and $n_2 - p$ are large, the statistical distance

$$
(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1}
\times (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))
$$

is approximately distributed as $\chi^2(p)$. 
Based on this result, for moderate to large sample sizes we have the following approximate test statistic for $H_0: \mu_1 = \mu_2$:

$$T^2 = (\bar{x}_1 - \bar{x}_2)' \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} (\bar{x}_1 - \bar{x}_2)$$

which we compare to $\chi^2_{\alpha}(p)$.

Approximate $100(1 - \alpha)\%$ simultaneous CIs for all linear combinations $a' (\mu_1 - \mu_2)$ are given by

$$a' (\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi^2_{\alpha}(p)} \sqrt{a' \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right) a}$$

For such a CI on $\mu_{1j} - \mu_{2j}$, choose $a$ to have a 1 in the $j^{th}$ position and 0’s elsewhere.

- Effects of nonnormality or unequal covariance matrices are exaggerated by samples sizes $n_1$ and $n_2$ that differ greatly.
One-way MANOVA

In the univariate case we generalized the two sample $t$ test for comparing two (univariate) means to the one-way anova for testing the equality of several means.

In the multivariate case, we generalize the two-sample $T^2$ to the situation where we have several mean vectors to compare. The method of analysis is known as the multivariate analysis of variance, or MANOVA.

Recall the univariate one-way anova:

1. We have independent samples of size $n_1, n_2, \ldots, n_g$ from $g$ populations or treatments. For the $\ell^{th}$ treatment we have a sample of univariate observations $x_{\ell 1}, \ldots, x_{\ell n_\ell}$ each with mean $\mu_\ell$.

2. Each of the $g$ populations is normal.

3. All $g$ populations have common variance $\sigma^2$. 
We can express these assumption succinctly as a statistical model:

\[ x_{\ell j} = \mu_\ell + \varepsilon_{\ell j}, \ell = 1, \ldots, g, j = 1, \ldots, n_\ell \]
\[ \varepsilon_{\ell j} \sim N(0, \sigma^2) \]

This model says that the observation on the \( j^{th} \) individual in the \( \ell^{th} \) group, \( x_{\ell j} \), is equal to the group mean plus some (positive or negative) error. The errors average out to 0 and have unknown variance \( \sigma^2 \).

- Our model implies \( x_{\ell j} \sim N(\mu_\ell, \sigma^2) \).

We often **reparameterize** our model by replacing \( \mu_\ell \) with \( \mu + \tau_\ell \):

\[ x_{\ell j} = \mu + \tau_\ell + \varepsilon_{\ell j} \]

which just breaks the \( \ell^{th} \) treatment mean \( \mu_\ell \) into two parts: the grand mean across all treatments \( (\mu) \) and an effect specific to the \( \ell^{th} \) treatment \( (\tau_\ell) \).

- To uniquely define model parameters and their estimates we impose the constraint \( \sum_\ell n_\ell \tau_\ell = 0 \).
Surprisingly, perhaps, to test a hypothesis about means, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$, we compare variances.

The idea comes from the following decomposition of the total variability in the data:

$$
\sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x})^2 = \sum_{\ell=1}^g n_\ell (\bar{x}_\ell - \bar{x})^2 + \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (x_{\ell j} - \bar{x}_\ell)^2
$$

or

$$
SS_T = SS_{Trt} + SS_E
$$

In words, this decomposition says that the total variability in the data ($SS_T$ would be the sample variance if divided by $\sum_\ell n_\ell - 1$) can be divided into the variability of the treatment means around the grand mean ($SS_{Trt}$) and the variability of the observations around their respective treatment means ($SS_E$).
If the treatments means (the $\mu_\ell$’s) were all the same (equal to to the grand mean $\mu$) then the variability quantified by $SS_{\text{Ttrt}}$ would be just error variance of the same order as the residual error quantified by $SS_E$.

Hence, to test $H_0 : \mu_1 = \cdots = \mu_g$ versus $H_1 :$ at least one pair of means differs, we compare $SS_{\text{Ttrt}}$ and $SS_E$.

An $\alpha$–level test of $H_0$ is: Reject $H_0$ if

$$F = \frac{SS_{\text{Ttrt}}/(g - 1)}{SS_E/\left(\sum_\ell n_\ell - g\right)} > F_\alpha(g - 1, \sum_{\ell=1}^g n_\ell - g)$$

Equivalently, instead of rejecting for large values of $F$, we could reject for small values of

$$\frac{1}{1 + SS_{\text{Ttrt}}/SS_E} = \frac{SS_E}{SS_E + SS_{\text{Ttrt}}}$$
We usually summarize the analysis of variance in a table:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>SS_{Trt}</td>
<td>$g - 1$</td>
<td>MS_{Trt}</td>
</tr>
<tr>
<td>Error</td>
<td>SS_{E}</td>
<td>$\sum n_\ell - g$</td>
<td>MS_{E}</td>
</tr>
<tr>
<td>Total</td>
<td>SS_{T}</td>
<td>$\sum n_\ell - 1$</td>
<td></td>
</tr>
</tbody>
</table>

**MANOVA:**

1. We have independent samples of size $n_1, \ldots, n_g$ from $g$ treatments. For the $\ell^{th}$ treatment we have a sample of multivariate observations $x_{\ell 1}, \ldots, x_{\ell n_\ell}$ with population mean $\mu_\ell$.

2. Each of the $g$ samples is drawn from a multivariate normal distribution.

3. All $g$ populations have a common population variance-covariance matrix $\Sigma$. 

166
Statistical Model:

\[ x_{\ell j} = \mu + \tau_{\ell} + \varepsilon_{\ell j} \quad \begin{cases} \ell = 1, \ldots, g \\ j = 1, \ldots, n_{\ell} \end{cases} \]

\[ \varepsilon_{\ell j} \sim N_p(0, \Sigma) \]

- \( \mu \) = overall mean
- \( \tau_{\ell} \) = \( \ell \)th treatment effect
- the \( \ell \)th treatment mean is \( \mu + \tau_{\ell} = \mu_{\ell} \).
- We impose the constraint \( \sum_{\ell} n_{\ell} \tau_{\ell} = 0 \).

- Model implies \( x_{\ell j} \sim N_p(\mu_{\ell}, \Sigma) \)

- This multivariate model implies that each component of \( x_{\ell j} \) satisfies the corresponding univariate one-way anova model.

- The errors are independent across treatments, but have covariance matrix \( \Sigma \) within any given treatment.
Analogous to the univariate situation, we can decompose the total variability in the sample into a component for treatment-to-treatment differences and a components for all else (error). In the multivariate setting, though, variability is quantified by a matrix including both variances and covariances. We have

\[
\sum_{\ell=1}^{g} \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x})(x_{\ell j} - \bar{x})' \\
= \sum_{\ell=1}^{g} n_{\ell} (\bar{x}_{\ell} - \bar{x})(\bar{x}_{\ell} - \bar{x})' + \sum_{\ell=1}^{g} \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})(x_{\ell j} - \bar{x}_{\ell})'
\]

- The terms of this decomposition are no longer just sums of squares, but now they are matrix quantities, consisting of sums of squares and cross-products (SSCP’s). Therefore, we write

\[
\text{SSCP}_T = \text{SSCP}_{Trt} + \text{SSCP}_E
\]

- Sometimes \(\text{SSCP}_{Trt}\) is called the between treatments SSCP matrix, or \(\mathbf{B}\), and \(\text{SSCP}_E\) is called the within treatments SSCP matrix, or \(\mathbf{W}\).
The MANOVA generalizes the two-sample $T^2$ procedure, and $W/(N - g)$ plays the role of $S_{\text{Pooled}}$. With a little algebra we can write $W$ as

$$W = \sum_{\ell=1}^{g} \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})(x_{\ell j} - \bar{x}_{\ell})'$$

$$= (n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g$$

The primary goal of a one-way MANOVA is to test for equal treatment means ($\mu_1 = \cdots = \mu_g$). This is equivalent to testing the hypothesis of zero treatments effects:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$

versus the alternative that at least one treatment effect is nonzero (at least one pair of means are unequal).

In the univariate situation, we test the hypothesis of equal treatment means by comparing $SS_{\text{Trr}}$ with $SS_E$ via an $F$ ratio. Because $SS_{\text{Trr}}$ and $SS_E$ are scalars, there is essentially only one way to compare them, and the $F$ ratio test statistic turns out to be optimal.
In the multivariate setting, again, $H_0$ is tested by comparing $SSCP_{Trt}$ ($B$) and $SS_E$ ($W$). However, because we are dealing with matrices, there are several ways to make this comparison, leading to several test statistics:

- **Roy’s test**: based on the largest eigenvalue of $BW^{-1}$.
- **Lawley’s and Hotelling’s test**: test statistic is $\text{tr}(BW^{-1})$.
- **Pillai’s test**: test statistic is a function of $\text{tr}[B(B + W)^{-1}]$.
- **Roy’s second test**: test statistic is a function of $|B(B + W)^{-1}|$.
- **Wilks’ likelihood ratio test**: Wilks’ test statistic depends on $\Lambda^* = |W|/|B + W|$ and is equivalent to the LRT.

Unfortunately, none of these tests is “best” in all situations. However, all of these tests are approximately equivalent in large samples, and differ very little in power for small samples.

- Because LRTs have good properties in general, we will confine attention to Wilks’ test.
The formula for Wilks’ test statistic and the distribution of the test statistic depend on the number of treatment means we are comparing ($g$) and the number of variables in our response vector ($p$). Exact distributions are available for the following cases:

<table>
<thead>
<tr>
<th>Case</th>
<th>Test Statistic</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p=1$ $g\geq 2$</td>
<td>$\left( \frac{N-g}{g-1} \right) \left( \frac{1-\Lambda^<em>}{\Lambda^</em>} \right)$</td>
<td>$F(g - 1, N - g)$</td>
</tr>
<tr>
<td>$p=2$ $g\geq 2$</td>
<td>$\left( \frac{N-g-1}{g-1} \right) \left( \frac{1-\sqrt{\Lambda^<em>}}{\sqrt{\Lambda^</em>}} \right)$</td>
<td>$F(2(g - 1), 2(N - g - 1))$</td>
</tr>
<tr>
<td>$p\geq 1$ $g=2$</td>
<td>$\left( \frac{N-p-1}{p} \right) \left( \frac{1-\Lambda^<em>}{\Lambda^</em>} \right)$</td>
<td>$F(p, N - p - 1)$</td>
</tr>
<tr>
<td>$p\geq 1$ $g=3$</td>
<td>$\left( \frac{N-p-2}{p} \right) \left( \frac{1-\sqrt{\Lambda^<em>}}{\sqrt{\Lambda^</em>}} \right)$</td>
<td>$F(2p, 2(N - p - 2))$</td>
</tr>
</tbody>
</table>

where $N = \sum_\ell n_\ell$ is the total sample size.
• In other cases we rely on approximations to obtain $p$–values for Wilks’ Lambda.

• A large sample approximation due to Bartlett gives the following rejection rule to obtain an approximate $\alpha$–level test: Reject $H_0$ if

\[-\left( N - 1 - \frac{p + g}{2} \right) \log(\Lambda^*) > \chi^2_\alpha(p(g - 1)).\]

• Other approximations are available to obtain approximate $p$–values when the total sample size $N$ is small. These approximation are implemented in SAS and other computer programs and are quite good even for small sample sizes.
Example – Turkey Meat:

A food scientist was studying the effectiveness of phosphate salts in conjunction with vacuum packaging on the preservation of precooked ground turkey meat during long storage periods. In particular, the researcher wanted to compare five phosphate salt treatments to determine the particular salt treatment that would be most effective. The five phosphate salt treatments consisted of

1. a control (no phosphate salt);
2. sodium tripolyphosphate (STP) at a 0.3% level;
3. STP at a 0.5% level;
4. sodium ascorbate monophosphate (SAsMP) at a 0.3% level; and
5. SAsMP at a 0.5% level.

Samples of cooked ground turkey meat were vacuum packaged using one of the five salt treatments, frozen, and stored for 150 days at approximate 

–14°C. Five samples were assigned to each treatment.
The variables measured on each sample included cooking loss (CKG\_LOSS), pH (PH), moisture after cooking (MOIST), fat content (FAT), and hexanal content (HEX).

- See turkey1.sas and turkey1.lst.

- In turkey1.sas I compute $\mathbf{B}$, $\mathbf{W}$ and Wilks’ Lambda with PROC IML first to indicate how the calculations are done from the formulas given here and in the book.

- In practice, though, we rely on a software routine such as PROC GLM in SAS to do the computations. Notice, that the values of $\mathbf{B}$, $\mathbf{W}$, and $\Lambda^*$ agree with the PROC IML computations.

- PROC GLM gives an approximate $p-$value of 0.0325 for $H_0 : \mu_1 = \cdots = \mu_5$, here. Using the conventional significance level of 0.05, we reject the hypothesis of equal treatment means.
It is fairly obvious that a MANOVA test of $H_0 : \mu_1 = \cdots = \mu_g$ can only be the first step in any analysis.

If we reject $H_0$ we still know relatively little.

- Which treatments differ?

- For which component variables do the treatments differ?

Simultaneous CIs for Treatment Effects:

The book describes **Bonferroni’s method** for simultaneous $\alpha$–level CIs for differences in treatment effects; that is, for differences between treatment means for component variables.

- For the $\ell^{th}$ treatment, the treatment effect is $\tau_\ell$ which we estimate with $\hat{\tau}_\ell = \bar{X}_\ell - \bar{x}$.  

175
• Between the \( k^{th} \) and \( \ell^{th} \) treatment, say, the difference in treatment effects is \( \tau_k - \tau_\ell \) or

\[
\mu_k - \mu_\ell = \begin{pmatrix}
\mu_{k1} - \mu_{\ell1} \\
\vdots \\
\mu_{kp} - \mu_{\ell p}
\end{pmatrix}
\]

which we estimate with

\[
\bar{x}_k - \bar{x}_\ell = \begin{pmatrix}
\bar{x}_{k1} - \bar{x}_{\ell1} \\
\vdots \\
\bar{x}_{kp} - \bar{x}_{\ell p}
\end{pmatrix}
\]

• The standard error of an estimated mean difference \( \bar{x}_{ki} - \bar{x}_{\ell i} \) in the one-way layout is given by

\[
\sqrt{\left( \frac{1}{n_k} + \frac{1}{n_\ell} \right) \frac{w_{ii}}{N - g}}
\]

where \( w_{ii} \) is the \((i, i)^{th}\) component of \( W \).
• Notice, each difference in treatment effects has \( p \) components, one for each component variable. Since there are \( g(g - 1)/2 \) pairs of treatments that may be compared, we have a total of \( pg(g - 1)/2 \) univariate treatment effects that we may look at.

Therefore, the Bonferroni approach leads to the following result for the one-way layout:

• With probability at least \( 1 - \alpha \), intervals given by

\[
\bar{x}_{ki} - \bar{x}_{\ell i} \pm t_{\alpha/(2m)}(N - g)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right) \frac{w_{ii}}{N - g}}
\]

cover differences \( \mu_{ki} - \mu_{\ell i} \) for components \( i = 1, \ldots, p \), and treatment pairs \( (\ell, k), \ell, k = 1, \ldots, g \), where \( m \) equals the number of such quantities to be estimated \( (m = pg(g - 1)/2 \) is we’re going to look at every pair).
**Roy’s Method:** If \( m \) is quite large, the Bonferroni approach will be too conservative (CIs will be too wide). In addition, Bonferroni intervals are not appropriate for data-snooping. An alternative approach to forming simultaneous confidence intervals is due to Roy, and is similar to the max-t or Scheffe procedure for multiple comparisons.

Using this approach, we can obtain simultaneous 100(1 - \( \alpha \))% CIs for any and all contrasts among linear combinations \( a' \mu_1, \ldots, a' \mu_g \), where \( a \) is any \( p \)-dimensional vector of coefficients. That is, for contrasts of the form \( \sum_{\ell=1}^{g} a' \mu_{\ell} c_{\ell} \) where \( \sum_{\ell=1}^{g} c_{\ell} = 0 \), simultaneous confidence intervals for any \( a \) and \( c \) are given by

\[
\sum_{\ell=1}^{g} a' \bar{x}_{\ell} c_{\ell} \pm \sqrt{\frac{x_\alpha}{1 - x_\alpha}} \sqrt{ \left( \sum_{\ell} \frac{c_{\ell}^2}{n_{\ell}} \right) a' Wa },
\]

where \( x_\alpha \) is obtained from the tables of Srivistava and Carter (1983) (see handout).

- The corresponding Bonferroni interval is

\[
\sum_{\ell=1}^{g} a' \bar{x}_{\ell} c_{\ell} \pm t_{\alpha/(2m)} (N - g) \sqrt{ \left( \sum_{\ell} \frac{c_{\ell}^2}{n_{\ell}} \right) \frac{a' Wa}{N - g} },
\]

178
Specifically, we can use Roy’s procedure to form simultaneous $100(1 - \alpha)\%$ CIs for differences in treatment means:

- With probability at least $1 - \alpha$, intervals given by
  \[
  \bar{x}_{ki} - \bar{x}_{\ell i} \pm \sqrt{\frac{x_\alpha (N - g)}{1 - x_\alpha}} \sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)} \frac{w_{ii}}{N - g}
  \]
  cover differences $\mu_{ki} - \mu_{\ell i}$ for all components $i = 1, \ldots, p$, and treatment pairs $(\ell, k)$, $\ell, k = 1, \ldots, g$.

- $x_\alpha$ is tabulated against the parameters $\alpha$, $p_0$, $m_0$, and $n$. $\alpha$ is the desired significance level. The other parameters are defined as follows:
  - $p_0 = p$, the number of component r.v.s
  - $m_0 = d.f.$ for treatments ($g - 1$ in one-way layout)
  - $n = d.f.$ for error ($N - g$ in one-way layout)

- Roy’s simultaneous CIs will be narrower (and hence, preferable) if
  \[
  \sqrt{\frac{x_\alpha (d.f. \text{Error})}{1 - x_\alpha}} < t_{\alpha/(2m)}(d.f. \text{Error})
  \]
Example – Turkey Meat (Continued):

- The SAS program turkey2.sas computes both Bonferroni and Roy simultaneous 95% CIs for the differences among treatment means.

- We see that in this example, the Bonferroni simultaneous CIs are narrower because

\[
\sqrt{\frac{x_\alpha (N - g)}{1 - x_\alpha}} > t_{\alpha/(2m)}(N - g)
\]

- We postpone further discussion of this example because, in fact, this experiment was not designed and performed as a simple one-way layout.
In turkey1.sas and turkey2.sas we treated the turkey meat experiment as a one-way layout. In fact, there was an additional feature in the design of the turkey meat experiment that should be taken into account in the analysis.

When preparing turkey meat samples it was not possible to prepare all 25 samples on a single day. Instead, the experimenters prepared one complete replication of the treatment structure on each of five different days. That is, five samples were prepared on each of five days and on each day each of the five salt treatments was used on one sample.

Therefore, the design of the experiment was a randomized complete block (RCB) design, with days forming the blocks.

- Although day-to-day variability is not of interest here, it will likely be present, and should be taken into account in the statistical analysis.
**Data:** Corresponding to the $\ell^{th}$ treatment in the $j^{th}$ block, we observe a $p-$ dimensional ($p = 5$ here) random vector $\mathbf{x}_{\ell j}$. We have treatments $\ell = 1, \ldots, g$ and blocks $j = 1 \ldots, b$.

**Model:**

$$
\mathbf{x}_{\ell j} = \mu + \tau_\ell + \beta_j + \varepsilon_{\ell j} \quad \left\{ \begin{array}{ll}
\ell = 1, \ldots, g \\
j = 1, \ldots, b
\end{array} \right.
$$

$$
\varepsilon_{\ell j} \sim N_p(0, \Sigma)
$$

The MANOVA Table for the RCB design (without replications) is

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>$SSCP_{Tht}$</td>
<td>$g - 1$</td>
</tr>
<tr>
<td>Blocks</td>
<td>$SSCP_{Blocks}$</td>
<td>$b - 1$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSCP_E$</td>
<td>$(g - 1)(b - 1)$</td>
</tr>
<tr>
<td>Total</td>
<td>$SSCP_T$</td>
<td>$gb - 1$</td>
</tr>
</tbody>
</table>
Strategy for Analysis with MANOVA:

1. Conduct MANOVA test (Wilks’ Lambda) of $\mu_1 = \cdots = \mu_g$ at level $\alpha$ (.05, for example).

2. If test in step (1) is significant, conduct univariate analyses including:
   
   a. ANOVA $F$-tests of $\mu_{1i} = \cdots = \mu_{gi}$ for each $i = 1, \ldots, p$ (each variable) each at level $\alpha$.

   b. For a given variable (the $i^{th}$, say), if the test in (2a) is significant, then do planned comparisons with the LSD method (i.e., use the CONTRAST statement and compare $p$-value with $\alpha$) (this is Fisher’s protected LSD since its only done if test in (2a) is signif.).

   c. If the test in (2a) is not significant, do planned comparisons with Bonferroni. Use Roy’s method (Scheffé’s method) for unplanned comparisons (data snooping).

   (continued)

183
3. If test in step (1) is not significant, do planned comparisons with Bonferroni method. For example, test all $p$ univariate hypotheses of equal means at level $\alpha/p$. Or, if additional inferences are desired, use $\alpha/(p+\# \text{ of additional inferences})$ for each one. Do not do unplanned comparisons.

- These recommendations follow Johnson (the reserve book) and do not represent the most conservative of approaches to multiple comparisons in MANOVA analyses. For example, our text recommends using Bonferroni intervals for all inferences other than the overall Wilks’ lambda test of equal multivariate means (regardless of the significance of this test). One’s level of conservatism including one’s choice of the collection of inferences for which the probability level should be controlled at a given level $\alpha$ is largely a matter of opinion and personal preference. (In the extreme, one could choose to set a lifetime error rate of .05.)
Example – Turkey Meat (Continued):

- We analyze the data according to the RCB model in turkey3.sas. Notice all that is necessary is to add REP (which corresponds to DAY, our blocking factor) to the CLASS and MODEL statements.

- We also remove the option NOUNI from the model statement so that SAS gives us the results of the model fits separately for each component variable (the univariate anova results).

- The PRINTE option on the MANOVA statement requests that SAS print out the error SSCP matrix (the matrix we’ve called W). This appears on p.8.

- The CONTRAST statements test contrasts among the levels of TRT. These contrasts will be tested both in the univariate analyses (one variable at a time) and in the MANOVA (all variables simultaneously).
• E.g., the first one tests $H_0 : 4\mu_{1i} - \mu_{2i} - \mu_{3i} - \mu_{4i} - \mu_{5i} = 0$, for each $i = 1, \ldots, p$ ($p = 5$, here) and it tests $H_0 : 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 = 0$. These contrasts compare the control treatment mean with the average of the active treatment means ($4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 = 0$ if and only if $\mu_1 = \frac{1}{4}(\mu_2 + \mu_3 + \mu_4 + \mu_5)$).

• The ESTIMATE statements have the same syntax as CONTRAST. ESTIMATE requests an estimate and standard error for a linear combination of the (univariate) treatment means. I have specified the same coefficients on the ESTIMATE statements as I did on the CONTRAST statements, except that in a couple of cases I used the DIVISOR option. For example, I want to estimate $\frac{1}{4}(4\mu_{1i} - \mu_{2i} - \mu_{3i} - \mu_{4i} - \mu_{5i})$ not $(4\mu_{1i} - \mu_{2i} - \mu_{3i} - \mu_{4i} - \mu_{5i})$ (DIVISOR=4 gives me the $\frac{1}{4}$ multiplier).
• The univariate results of the ESTIMATE and CONTRAST statements appear on pp.13–17. Note that the $t$—tests from the ESTIMATE statements are just the square roots of the corresponding $F$—tests from the corresponding CONTRAST statements (they’re equivalent tests). The $p$—values from these tests would be compared to $\alpha/m$ where $m = \#$ inferences to be performed according to the Bonferroni approach to multiple comparisons.

• The multivariate results of the CONTRAST statements appear on pp.10–12. These tests pertain to specific comparisons (contrasts) among the groups made by treating all random variables simultaneously rather than one at a time. In some situations, such tests may be desired; but we will focus on univariate contrasts.
Results:

- Wilks’ Lambda statistic is 0.0365 (p.9 of turkey3.lst), which has an approximate $p$-value of 0.0004. Therefore, we reject $H_0$ that the treatment mean vectors are equal. Taking into account day-to-day variability has strengthened our conclusion (relative to the analysis in turkey1.sas that ignored the blocking factor) that the treatment means differ.

- After rejecting the manova null hypothesis we can look at the results of the corresponding univariate anovas, to see which variables exhibit differences in the treatment means.

- The treatments differed significantly for percentage cooking loss ($p < 0.0001$ p.2), moisture content ($p = 0.0052$ p.4), and hexanal content ($p < 0.0001$ p.6). Significant differences between treatments did not show up in the one-at-a-time analysis for pH, and fat content.
• Let’s assume that the comparisons implied by the CONTRAST statements are all planned comparisons. Since the univariate $F$—tests for equal treatment means were significant for CKG_LOSS, MOIST and HEX, the univariate CONTRAST results ($p$—values) can be compared to $\alpha = .05$ (this is Fisher’s LSD) for these three variables. Contrasts 1–3 are significant for CKG_LOSS, contrast 1 is significant for MOIST, and all contrasts are significant for HEX.

• There are four planned comparisons for each of the variables PH and FAT. We can test these 8 contrasts with Bonferroni. We can guarantee that the probability of falsely rejecting one or more of the $m = 8$ null hypothesis implicit in these comparisons is no more than $.05$ by comparing the $p$—values from the univariate CONTRAST results to $.05/8 = .00625$. By this criterion, none of the contrasts on PH or FAT are significant.
• Alternatively (and equivalently), we could compute Bonferroni joint confidence intervals for the linear combinations specified in the ESTIMATE statements for PH and FAT using \( m = 8 \) and examine which ones cover 0.

For any design (not necessarily the one-way layout), 100(1 – \( \alpha \)% simultaneous Bonferroni intervals for \( m \) contrasts of the form \( \sum_{\ell=1}^{g} a' \mu c_\ell \) where \( \sum_{\ell=1}^{g} c_\ell = 0 \), are given by

\[
\sum_{\ell=1}^{g} a' \bar{x}_\ell c_\ell \pm t_{\alpha/(2m)}(\nu) \sqrt{\left( \sum_{\ell} \frac{c_\ell^2}{n_\ell} \right) \frac{a' Wa}{\nu}},
\]

where \( \nu = \text{d.f.}_E \), the degrees of freedom for error in the model (\( \nu = (g - 1)(b - 1) \) in the RCBD).

The corresponding Roy intervals are given by

\[
\sum_{\ell=1}^{g} a' \bar{x}_\ell c_\ell \pm \sqrt{\nu x_\alpha} \sqrt{\left( \sum_{\ell} \frac{c_\ell^2}{n_\ell} \right) \frac{a' Wa}{\nu}},
\]

where \( x_\alpha \) is obtained from the tables of Srivistava and Carter (1983)
For the special case of a pairwise difference these intervals simplify as follows. For the difference $\mu_{k} - \mu_{\ell}$ between the means for the $k^{\text{th}}$ and $\ell^{\text{th}}$ treatments in the $i^{\text{th}}$ variable, the Bonferroni interval is given by

$$
\bar{x}_{ki} - \bar{x}_{\ell i} \pm t_{\alpha/(2m)}(\nu) \sqrt{\frac{2 w_{ii}}{b \nu}},
$$

where $w_{ii}$ is the $i^{\text{th}}$ diagonal element of $W$. The corresponding Roy interval is given by

$$
\bar{x}_{ki} - \bar{x}_{\ell i} \pm \sqrt{\frac{\nu x_{\alpha}}{1 - x_{\alpha}}} \sqrt{\frac{2 w_{ii}}{b \nu}}.
$$

- Bonferroni 95% CIs for the planned comparisons on PH and FAT are computed in PROC IML and are given on p.18. These results can also be obtained from the results of the ESTIMATE statements.
• For example, the Bonferroni interval for $\mu_{12} - (1/4)(\mu_{22} + \mu_{32} + \mu_{42} + \mu_{52})$ (control - average of active treatment means for PH) is given on p.18 as $(-.4551, .2441)$. On p.14, the ESTIMATE statement gives an estimate for this quantity of -.1055 with an s.e. of .1111. The Bonferroni interval is therefore,

$$- .1055 \pm t_{1-.05/(2^8)}(16)(.1111)$$

$$= -.1055 \pm 3.1458(.1111) = (-.4551, .2441)$$
Principal Components Analysis
(See chapter 8 of Johnson and Wichern.)

- PCA involves taking a set of correlated response variables and transforming them into a smaller set of uncorrelated variables without losing much information.

- General objectives:

  1. Data reduction;
  2. Data screening; and
  3. Analysis of variance-covariance relationships among original variables.

- PCA is often an initial or intermediate step in a full multivariate analysis.
Population Principal Components:

We take $p$ random variables and replace them with $k$, $k < p$, random variables called principal components that are linear combinations of the original variables such that

I. the $k$ principal components are uncorrelated;
II. the first PC accounts for as much of the variability in the data as possible; and
III. each succeeding PC accounts for as much of the remaining variability as possible.

Suppose the $p$–variate random vector $\mathbf{x} = (x_1, \ldots, x_p)'$ has population mean $\mu$ and population var-cov matrix $\Sigma$. (Not assuming M’variate normality yet.)

Consider the linear combinations

$$y_1 = \mathbf{a}_1' \mathbf{x} = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p$$
$$y_2 = \mathbf{a}_2' \mathbf{x} = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p$$
$$\vdots$$
$$y_p = \mathbf{a}_p' \mathbf{x} = a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p$$
By our rules for variances and covariances of linear combinations of random vectors we have

\[ \text{var}(y_i) = \mathbf{a}_i' \Sigma \mathbf{a}_i, \quad i = 1, \ldots, p, \]

\[ \text{cov}(y_i, y_k) = \mathbf{a}_i' \Sigma \mathbf{a}_k, \quad i, k = 1, \ldots, p. \]

By (II) above, we want to choose \( \mathbf{a}_1 \) so that \( \text{var}(y_1) = \mathbf{a}_1' \Sigma \mathbf{a}_1 \) is maximized subject to the constraint that \( \mathbf{a}_1' \mathbf{a}_1 = 1. \)

- The constraint is necessary because for any choice of \( \mathbf{a}_1 \) I can increase \( \text{var}(y_1) = \mathbf{a}_1' \Sigma \mathbf{a}_1 \) by taking instead \( 2\mathbf{a}_1. \)

By (I) and (III) above, the second PC corresponds to the choice of \( \mathbf{a}_2 \) so that \( \text{var}(y_2) = \mathbf{a}_2' \Sigma \mathbf{a}_2 \) is maximized subject to \( \mathbf{a}_2' \mathbf{a}_2 = 1 \) and \( \text{cov}(y_2, y_1) = 0. \)

The \( i^{th} \) PC corresponds to the choice of \( \mathbf{a}_i \) so that \( \text{var}(y_i) \) is maximized subject to \( \mathbf{a}_i' \mathbf{a}_i = 1 \) and \( \text{cov}(y_i, y_j) = 0 \) for \( j = 1, \ldots, i - 1. \)
First PC: need to find
\[
\max_{a' a_1 = 1} a'_1 \Sigma a_1.
\]
Using the Spectral Decomposition of \( \Sigma \) we see that
\[
a'_1 \Sigma a_1 = a'_1 EE' a_1 = w' \Lambda w = \sum_{i=1}^{p} \lambda_i w_i^2
\]
The constraint \( a'_1 a_1 = 1 \) is equivalent to \( w' w = 1 \) because
\[
w' w = a'_1 EE' a_1 = a'_1 a_1.
\]
We know that the largest eigenvalue is \( \lambda_1 \) so to maximize \( \sum_{i=1}^{p} \lambda_i w_i^2 \) subject to \( w' w = 1 \) we take \( w = (1, 0, 0, \ldots, 0)' \). Since \( a_1 = Ew \), this implies that the first PC is \( y_1 = a_1 x \) where
\[
a_1 = E \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = e_1
\]
• The first principal component of \( x \) is
\[
y_1 = e'_1 x
\]
with variance
\[
\text{var}(y_1) = \lambda_1
\]
196
Similar arguments lead to

- The second principal component of $\mathbf{x}$ is
  
  $$y_2 = \mathbf{e}_2' \mathbf{x}$$

  with variance

  $$\text{var}(y_2) = \lambda_2.$$

  Notice

  $$\text{cov}(y_2, y_1) = \mathbf{e}_2' \Sigma \mathbf{e}_1 = \mathbf{e}_2' \mathbf{\lambda}_1 \mathbf{e}_1$$

  $$= \mathbf{\lambda}_1 \mathbf{e}_2' \mathbf{e}_1 = 0$$

- The $i^{\text{th}}$ principal component of $\mathbf{x}$ is
  
  $$y_i = \mathbf{e}_i' \mathbf{x}$$

  with variance

  $$\text{var}(y_i) = \lambda_i$$
• Putting the $p$ principal components together in a single vector, $\mathbf{y} = (y_1, \ldots, y_p)'$, we see that $\mathbf{y} = \mathbf{E}'\mathbf{x}$. The geometric interpretation is that $\mathbf{y} = \mathbf{E}'\mathbf{x}$ constitutes a (rigid) rotation of the original coordinate axes to line up with the directions of maximal variability.
Recall that one way to reduce a variance-covariance matrix $\Sigma$ to a single (scalar) quantity is by computing the total variance

$$\text{tr}(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$$

We also had the result that $\text{tr}(\Sigma)$ equals the sum of the eigenvalues of $\Sigma$:

$$\text{tr}(\Sigma) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} \lambda_i$$

**Property 1:** Since $\text{var}(y_i) = \lambda_i$, the proportion of the total variance explained by the $i^{th}$ principal component is

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$
Property 2: If \( y_1 = e_1'x, y_2 = e_2'x, \ldots, y_p = e_p'x \) are the principal components of \( x \) based on the covariance matrix \( \Sigma \), then the correlation between the \( i^{\text{th}} \) principal component \( y_i \) and the \( k^{\text{th}} \) original variable \( x_k \) is given by

\[
\text{corr}(y_i, x_k) = \frac{e_{ki} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}
\]

Why?

If we take \( a = (0, \ldots, 0, 1, 0, \ldots, 0)' \) (1 in the \( k^{\text{th}} \) position), then \( x_k = a'x \). Therefore,

\[
\text{cov}(x_k, y_i) = \text{cov}(a'x, e_i'x) = a'\Sigma e_i
\]

\[
= a' \lambda_i e_i = \lambda_i a' e_i = \lambda_i e_{ki}
\]

So,

\[
\text{corr}(y_i, x_k) = \frac{\text{cov}(x_k, y_i)}{\sqrt{\text{var}(x_k)} \sqrt{\text{var}(y_i)}}
\]

\[
= \frac{\lambda_i e_{ki}}{\sqrt{\sigma_{kk}} \sqrt{\lambda_i}} = \frac{e_{ki} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}
\]

200
Additional Properties when \( \mathbf{x} \sim N_p(\mathbf{0}, \Sigma) \).

- We assume \( \mathbf{E}(\mathbf{x}) = \mathbf{0} \), but if \( \mathbf{E}(\mathbf{x}) = \mathbf{\mu} \) just substitute \( \mathbf{x} - \mathbf{\mu} \) for \( \mathbf{x} \).

**Property 3:** \( \mathbf{y} = \mathbf{E}'\mathbf{x} \sim N_p(\mathbf{0}, \Omega) \) where \( \Omega \) is a diagonal matrix with the \( \lambda_1, \ldots, \lambda_p \) on the diagonal.

- Property 3 says that the principal components are mutually independent normally distributed r.v.s.

Recall that the contours of constant density are ellipsoids defined by

\[
c^2 = \mathbf{x}'\Sigma^{-1}\mathbf{x}
\]

and since, using the spectral decomposition,

\[
\mathbf{x}'\Sigma^{-1}\mathbf{x} = \sum_{i=1}^{p} \frac{1}{\lambda_i} (\mathbf{e}_i'\mathbf{x})^2.
\]
So the contours of constant density are given by

\[ c^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2, \]

the equation of an ellipsoid with axes \( y_1, \ldots, y_p \). Therefore, we have the following result:

**Property 4:** The axes of the constant density ellipsoids for the \( p \)--variate normal random vector \( \mathbf{x} \) are the principal components.

- For example, for a bivariate normal random vector \( \mathbf{x} = (x_1, x_2)' \):
Principal Components from Standardized Variables:

- Principal components generally depend on the scale on which the variables are measured.

- If one variable has a much larger variance than the other variables, then it will dominate the first principal component regardless of the covariance structure between variables.

- A variable may have larger variance simply because it was measured on a different scale (grams rather than pounds).

In such a case we may want to put all variables on the same footing by standardizing the variables prior to computing principal components.

From each $x_i$ we compute the standized variable $z_i$: 

$$ z_i = \frac{x_i - \mu_i}{\sqrt{\sigma_{ii}}} $$

203
The vector of standardized variables is

\[ z = V^{-1/2}(x - \mu) \]

where \( V = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp}) \). Notice, the covariance matrix of \( z \) is the correlation matrix of \( x \):

\[ \text{cov}(z) = \text{corr}(x) = P \]

If \((\lambda_1, e_1), \ldots, (\lambda_p, e_p)\) are the eigen-pairs of \( P = \mathbf{E}\Lambda\mathbf{E}' \), then the principal components of \( z \) are \( y_i = e_i z, i = 1, \ldots, p \), or \( y = \mathbf{E}'z \).

- Principal component computed on the standardized variables are different from PCs computed from original variables.

Which do we use?

Answer depends on whether you want different variances among variables to be taken into account when computing PCs. Often you don’t, because differences in variances reflect only differences in measurement scale. Other times differences in variance are relevant and should be considered.
• For example, suppose that a large number of job applicants for an executive position are evaluated on 15 different variables. A goal of the analysis of the resulting data is to select six individuals to receive job offers. Consider two variables: ambition and appearance. Both variables were measured on a 10 point scale. It is likely that there will be significantly larger variance in ambition than in appearance. Ambition ratings of applicants are likely to range widely, perhaps from 0–10, but since all applicants are likely to look pretty good, appearance may range only from 7–10. The larger variance in ambition reflects that it is a better criterion by which to choose job applicants, and in this situation differences in variances should be taken into account when computing principal components. That is, PCs should be based on the original variables rather than on standardized variables.
• Since $tr(P) = 1 + 1 + \cdots + 1 = p$, Property 1 becomes: for principal components computed on standardized variables, the proportion of the total variance accounted for by the $i^{\text{th}}$ PC is $\lambda_i/p$.

• Property 2 becomes: for PCs computed on the standardized variables, the correlation between $y_i$ and $z_k$ is given by

$$\text{corr}(y_i, z_k) = e_{ki} \sqrt{\lambda_i}.$$ 

Sample Principal Components:

Typically, the population quantities $\Sigma = \text{cov}(x)$, and $\mu = \mathbb{E}(x)$ are unknown; therefore, the eigenpairs $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ are unknown and the (population) principal components $y_1, \ldots, y_p$ are unknown.

We estimate the population mean and variance by computing the corresponding quantities based on a sample $x_1, \ldots, x_n$. 

206
• We estimate $\Sigma$ with $S$, $\mu$ with $\bar{x}$.

It is natural to estimate the population PCs with the corresponding sample quantities as well.

• If $S$ is the sample variance-covariance matrix with eigen-pairs $(\hat{\lambda}_1, \hat{e}_1), \ldots, (\hat{\lambda}_p, \hat{e}_p)$, the $i^{th}$ sample principal component is given by

$$\hat{y}_i = \hat{e}_i' x, \quad i = 1, \ldots, p.$$ 

The sample PCs have properties analogous to those of the population PCs:

• The sample variance of $\hat{y}_i$ is $\hat{\lambda}_i$.

• The sample covariance between $\hat{y}_i$ and $\hat{y}_k$ is 0 $(k \neq i)$.
• The total sample variance is

\[
\sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i
\]

so the proportion of the total sample variance explained by the \(i^{th}\) sample PC is

\[
\frac{\hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i}
\]

• The sample correlation between \(\hat{y}_i\) and \(x_k\) is

\[
\frac{\hat{e}_{ki} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}
\]
Since we have a sample \( \mathbf{x}_1, \ldots, \mathbf{x}_n \) with \( n \) elements, we will have \( n \) values of each sample principal component. That is, the \( i^{\text{th}} \) sample PC will have values \( \hat{y}_{i1}, \ldots, \hat{y}_{in} \).

Usually, each observation vector is centered by subtracting off the sample mean vector, before sample PCs are computed. So, the \( i^{\text{th}} \) sample PC has values

\[
\hat{y}_{ij} = \tilde{e}'_i (\mathbf{x}_j - \bar{x}), \quad \begin{cases} 
  i = 1, \ldots, p \\
  j = 1, \ldots, n 
\end{cases}
\]

**How many Principal Components?**

One of the purposes of PCA is to reduce the dimension of the data. More accurately, we want to determine the dimensionality of the space in which the data fall.

We’d like to be able to look at only the first \( k \) PCs where \( k \) is small (certainly less than \( p \)).
Methods:

1. *A priori* we decide that we want to be able to explain some fixed percentage of the total variance, $\gamma\%$, say. Then we pick $k$ to be as small as possible subject to

$$100 \left( \frac{\lambda_1 + \cdots + \lambda_k}{\sum_{i=1}^{p} \lambda_i} \right) > \gamma$$

2. SCREE plot. Plot pairs $(1, \lambda_1), (2, \lambda_2), \ldots, (p, \lambda_p)$ and look for the point at which the $\lambda$'s level off. Pick $k$ corresponding to the smallest “large” eigenvalue. For the SCREE plot below, we would pick $k = 3$. 

210
3. When PCs are computed from \( \mathbf{R} \) rather than \( \mathbf{S} \) we can pick \( k \) to be the number of eigenvalues greater than 1. When using standardized variables, each original standardized variable contributes 1 to the total variance. Therefore, a principal component should not be considered if it cannot account for more variability than a single original variable can by itself.

In practice, one should use a combination of these methods and subjective judgement to choose \( k \).
Example – Job Applicants:

Forty-eight individuals who had applied for a job with a large firm were interviewed and rated on 15 criteria. Individuals were rated on the form of their letter of application (FL), their appearance (APP), their academic ability (AA), their likability (LA), their self-confidence (SC), their lucidity (LC), their honesty (HON), their salesman ship (SMS), their experience (EXP), their drive (DRV), their ambition (AMB), their ability to grasp concepts (GSP), their potential (POT), their keenness to join (KJ), and their suitability (SUIT). Each criterion was evaluated on a scale ranging from 0–10, with 0 being low.

- Principal components analysis may be performed with SAS’ PROC PRINCOMP. See jobapp1.sas and jobapp1.lst.

- On p.1 we have sample means and standard deviations for each original variable, \( x_1, \ldots, x_{15} \), and the sample var-cov matrix \( S_x \) is printed on pp.1–3.
• The total variance is \( \text{tr}(S_x) = 122.535 \) (p.3).

• Eigenvalues and eigenvectors are printed on p.3–5. Notice that the first three eigenvalues are 66.5, 18.2, and 10.6, indicating that the first three PCs account for 54.3\%, 14.8\%, and 8.6\% of the total variance, respectively. For example, the first PC accounts for \(100\% \times (66.5/122.5) = 54.3\%\) of the total variance. Combined, the first three PCs account for 77.8\% of the total variance.
• A SCREE plot of the eigenvalues appears below. From this plot we see that there seems to be a break between the 4th and 5th eigenvalues. Also, choosing to reduce the original 15 variables to 4 PCs allows us to explain over 80% (83.3%) of the total variance.
- The 48 values of all 15 principal components are computed and output to a data set with the \texttt{OUT=PCSCORES} option. The values of the first 4 of the PCs are printed with \texttt{PROC PRINT} on p.6.

- The $i^{th}$ PC score for the $j^{th}$ individual is given by
  \[ \hat{y}_{ij} = \hat{e}_i'(x_j - \bar{x}). \]

So the first value of the first PC is
\[
y_{11} = 0.149(6 - 6) + 0.132(7 - 7.083) + \\
\cdots + 0.274(10 - 5.958) = 4.304
\]

- The sign and magnitude of $\hat{e}_{ki}$ give some information about the dependence of the $i^{th}$ PC on the $k^{th}$ variable, but the sample correlation $\text{corr}(\hat{y}_i, x_k)$ has a more precise interpretation as a measure of association. For this reason, interpretation of the sample correlations (computed in \texttt{PROC CORR} and appearing on pp.7–8) is generally preferred to the interpretation of the values of the eigenvectors.
• The variables that exhibit a strong relationship with the first PC are LC, SMS, DRV, AMB, GSP, POT. The variables that show a strong relationship with the second PC are FL, EXP, and SUIT. LA, HON and KJ strongly influence the third PC; and AA and KJ strongly influence the fourth PC.

• Notice that all the the coefficients of the first PC have the same sign. This is a direct consequence of the fact that the correlations among the 15 variables are all positive. In this situation, the first PC measures the total score. Most other PCs have positive and negative coefficients, contrasting scores on different criteria.

• In this example, interpretations for the first four PCs are limited. This is often the case, and interpretations should not be forced. However, there are cases where the PCs have nice interpretations that are plausible and worth some attention.
Structural Relationships:

If some of the eigenvalues of $\Sigma$ or $P$ are zero, then some of the original variables are linearly dependent on others. That is, there are structural relationships among the original variables.

The number of structural relationships among the variables equals the number of 0 eigenvalues. Such relationships can cause interpretive and computational difficulties in subsequent analyses and should be removed.

- The eigenvectors that correspond to eigenvalues near zero define the structural relationships among the original variables.

- The structural relationships are only unique when there is one linear dependence (one zero eigenvalue), so only in that case are physical interpretations for the structural relationship appropriate.
Graphing the PCs:

- Can reveal outliers that may be suspicious.
- Useful as a check on normality assumption.
- Useful for suggesting clusters of items.

1. To check normality:

   a. scatterplots for pairs of the first few PCs should look approximately elliptical; and
   b. Q-Q plots from the sample values of each PC should be linear (not necessary to produce plots for all $p$ PCs, just for first and last 2 or 3 PCs).

2. To check for outliers:

   a. scatterplots for pairs of last few PCs will reveal outliers; and
   b. outliers will show up on Q-Q plots for last few PCs.
Large-Sample Inference:

The essential quantities in PCA are the eigenvalues and eigenvectors of the population covariance matrix $\Sigma$.

- The eigenvectors determine the direction of maximum variability; and
- the eigenvalues specify the variances in those directions.

Remember that we don’t know $\Sigma$. We have to estimate it with $S$ and we also estimate the population eigenvectors and eigenvalues with the sample quantities.

We can form large-sample confidence intervals for $\lambda_i$ and $e_i$ based on the asymptotic (large sample) distributions of $\hat{\lambda}_i$ and $\hat{e}_i$. 
Assumptions:

- \( x_1, \ldots, x_n \sim N_p(\mu, \Sigma) \)
- The eigenvalues of \( \Sigma \) are distinct and all positive.

An approximate (asymptotic) \( 100(1 - \alpha)\% \) confidence interval for \( \lambda_i \) is given by

\[
\left( \frac{\hat{\lambda}_i}{1 + z_{\alpha/2} \sqrt{2/n}}, \frac{\hat{\lambda}_i}{1 - z_{\alpha/2} \sqrt{2/n}} \right)
\]

where \( z_{\alpha/2} \) is the upper \( 100(\alpha/2)\)th percentile of the standard normal distribution (1.96 for a 95\% CI).

If we want \( m \) such intervals we can use Bonferroni’s method, replacing \( z_{\alpha/2} \) with \( z_{\alpha/(2m)} \).

- Other inferential methods are available. E.g., one can test \( H_0 : \lambda_{k+1} = \cdots = \lambda_p \) using a LRT. See, e.g., Karson (1982, Multivariate Statistical Methods) for details.
Example – Job Applicants:

- In jobapp1.sas Q-Q plots are produced using PROC UNIVARIATE. These plots appear on p.10, p.12, p.14, and p.16 for \( y_1, y_2, y_3, y_4 \), respectively.

- The Q-Q plot for the second PC, \( y_2 \), on p. 12 indicates that applicants 41 and 42 are outliers.

- The Q-Q plot for \( y_3 \) indicates that applicant number 11 and possibly number 10 are outliers.

- If no justification is found for deleting outliers altogether, it would be appropriate to compare analyses performed with and without some or all outliers, to assess their effects on the results.

- A 95\% CI for \( \lambda_1 \), the variance of the first population principal component is given by

\[
\left( \frac{66.54}{1 + 1.96 \sqrt{2/48}}, \frac{66.54}{1 - 1.96 \sqrt{2/48}} \right) = (47.53, 110.92)
\]
Example – Employment in European Countries:

In the SAS program employ1.sas, data on the percentage of the labor force employed in various industry sectors are analyzed using principal components. The data are from 1979 and are listed on p.1 of employ1.lst. The percentage of the labor force is given in the following exhaustive categories: AGR (agriculture), MIN (mining), MAN (manufacturing), PS (power supplies), CON (construction), SER (service industries), FIN (finance), SPS (social and personal services), TC (transport and communications).

- The first PC accounts for

\[
100\% \times \frac{\lambda_1}{p} = 100\% \times \frac{3.487}{9} = 38.7\%
\]

of the total variance.
• The first two PCs are the only ones with variances substantially greater than one, but they account for only 62.4% of the total variance. We might prefer to explain more (85.7%) by considering the first four PCs. Either choice would be supported to some extent by a SCREE plot. We will consider just two PCs here, for simplicity.

• Notice that the last eigenvalue is very close to 0. This reflects the structural relationship among the variables that the total percentage across all 9 sectors must be 100%. This linear dependence should be removed for any subsequent statistical analyses (drop one variable).

• The first PC seems to be a contrast between percentage engaged in agriculture and percentages engaged in manufacturing, power supplies, construction, service industries, social and personal services, and transport and communications.

• The second PC contrasts percentages in mining and manufacturing with percentages in service industries and finance.
A scatterplot of the first two PCs for each of the 26 countries appears below. Notice that most of the Western democracies are grouped together with low values of $y_1$ and $y_2$. Ireland, Portugal, Spain and Greece have higher values of $y_1$, while Turkey and Yugoslavia have very high values of $y_1$. Most of the communist countries are grouped together with high values of $y_2$. 
**Factor Analysis**
(See chapter 9 of Johnson and Wichern.)

- A method to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called **factors**.

- In PCA, we try to explain the variability in a set of variables via a smaller set of uncorrelated principal components. Relatively few assumptions.

- In Factor Analysis, we try to explain the covariance/correlation structure among a group of variables. We do so by assuming a particular statistical model.

- Model assumes that there are a few underlying, unobserved (or *latent*) variables called **factors**, and high correlations among certain of the observed variables arise because those observed variables are determined by the same factor.
• For example, suppose we have decathlon scores on each of the ten events from a random sample of athletes. It may be that certain of the scores are highly correlated with each other (e.g., javelin, discus, and shot put; high jump, pole vault and broad jump), but weakly correlated with others. The factor analysis model posits that the highly associated variables measure, or are determined by, the same factor (e.g., arm strength, jumping ability), and that the several factors that are represented in the data set are uncorrelated.

The goals of FA include

1. to determine whether a smaller set of uncorrelated variables exist that will explain the relationships that exist between the original variables (i.e., are the data consistent with the factor analysis model?);

2. to determine the number of underlying factors;

3. to interpret these factors;
4. to evaluate individuals or experimental units on these factors; and

5. to use these new variables in other statistical analyses of the data.

**Caveats:**

- Factor analysis attempts to create a new set of uncorrelated variables from a set of correlated variables. Therefore, there’s no point to performing a FA on uncorrelated random variables.

- Unlike most statistical techniques, FA involves some subjectivity at several stages of the analysis. Therefore, it has been criticized harshly by some statisticians.

- The results of any factor analysis should be interpreted cautiously and one should understand the method thoroughly before attempting to use it.
Suppose we have a $p$–dimensional random vector $\mathbf{x}$ with mean $\mu$ and variance-covariance matrix $\Sigma$. The FA model assumes

$$\mathbf{x} = \mu + \mathbf{LF} + \varepsilon$$

or, equivalently, the model for the centered vector is

$$\mathbf{x} - \mu = \mathbf{LF} + \varepsilon,$$

or, elementwise,

$$x_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1$$
$$x_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2$$
$$\vdots$$
$$x_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p.$$

- Here, $\ell_{ij}$ is the $(i, j)^{\text{th}}$ element of matrix $\mathbf{L}_{p \times m}$ and is called the loading of the $i^{\text{th}}$ variable on the $j^{\text{th}}$ factor. $\mathbf{L}$ is called the matrix of factor loadings. $F_1, \ldots, F_m$ are called the common factors, and $\varepsilon_1, \ldots, \varepsilon_p$ are called specific factors, or error terms.
• Notice that the FA model is essentially the multiple regression model with \( \mu \) playing the role of the vector of intercept terms, \( L \) playing the role of the parameter matrix (slope parameters), and \( F \) playing the role of the design matrix (matrix of explanatory variables).

• **With one crucial difference:** The “explanatory variables” (the factors) are unobserved random variables.

• With so many unknown quantities (everything on the right hand side of the FA model (*)), it is not possible to obtain unique estimates of the parameters of this model. However, it is possible to determine whether or not the data are consistent with the covariance structure that the model implies that the data should have.

Additional assumptions of the FA model are that

\[
\begin{align*}
E(F) &= 0, \quad \text{var}(F) = I, \\
E(\varepsilon) &= 0, \quad \text{var}(\varepsilon) = \Psi,
\end{align*}
\]

and \( F \) and \( \varepsilon \) are independent.
Here, $\Psi$ is assumed to be a diagonal matrix:

$$
\Psi = \begin{pmatrix}
\psi_1 & 0 & \cdots & 0 \\
0 & \psi_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \psi_p
\end{pmatrix}
$$

so that each variable depends upon the same set of factors, but in addition has its own additional source of variability ($\psi_i$ for $x_i$, $i = 1, \ldots, p$).

**Covariance Structure Implied by FA Model:**

The FA model implies

$$
\mathbb{E}(X) = \mu, \quad \text{var}(X) = LL' + \Psi = \Sigma
$$

or, elementwise, $\mathbb{E}(x_i) = \mu_i$ and

$$
\sigma_{ii} = \text{var}(x_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i
$$

$$
\text{cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}
$$

The variance of the $i^{th}$ variable decomposes into two parts:

$$
\sigma_{ii} = \underbrace{\ell_{i1}^2 + \cdots + \ell_{im}^2}_{{\ell^2}_{im}} + \psi_i
$$

$$
= h_i^2
$$

$$
= \text{communality} + \text{specific variance}
$$

230
• In FA, the goal is to determine whether or not the decomposition $\Sigma = LL' + \Psi$ holds when the number of factors $m$ is small relative to the number of variables $p$, and then to interpret such factors.

• Unfortunately, when $m$ is much less than $p$, there is no guarantee that such a decomposition of $\Sigma$ exists.

• In addition, even when a FA solution (numerical values of $L$ and $\Psi$ such that $\Sigma = LL' + \Psi$ holds) exists, it is typically not unique.
Non-uniqueness:

Let \( \mathbf{T} \) be an \( m \times m \) orthogonal matrix (a matrix with the property that \( \mathbf{T}\mathbf{T}' = \mathbf{I} \)).

Suppose that the (loading matrix, factor matrix) pair \( (\mathbf{L}, \mathbf{F}) \) satisfies the FA model. That is, \( \mathbf{X} - \mathbf{\mu} = \mathbf{LF} + \mathbf{\varepsilon} \) holds, \( \mathbf{E}(\mathbf{F}) = \mathbf{0} \), \( \text{var}(\mathbf{F}) = \mathbf{I} \), and \( \Sigma \) can be decomposed as \( \Sigma = \mathbf{LL}' + \mathbf{\varepsilon} \).

Then it is also true that the (loading matrix, factor matrix) pair \( (\mathbf{LT}, \mathbf{T}'\mathbf{F}) = (\mathbf{L}^*, \mathbf{F}^*) \) also satisfies the FA model since

\[
\mathbf{X} - \mathbf{\mu} = \mathbf{LF} + \mathbf{\varepsilon} = \mathbf{LT}\mathbf{T}'\mathbf{F} + \mathbf{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \mathbf{\varepsilon},
\]
\[
\mathbf{E}(\mathbf{F}^*) = \mathbf{E}(\mathbf{T}'\mathbf{F}) = \mathbf{T}'\mathbf{E}(\mathbf{F}) = \mathbf{T}'\mathbf{0} = \mathbf{0},
\]
\[
\text{var}(\mathbf{F}^*) = \text{var}(\mathbf{T}'\mathbf{F}) = \mathbf{T}'\text{var}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}, \quad \text{and}
\]
\[
\Sigma = \mathbf{LL}' + \Psi = \mathbf{LT}\mathbf{T}'\mathbf{L} + \Psi = \mathbf{L}^* (\mathbf{L}^*)' + \Psi
\]

- Therefore, \( \mathbf{L} \) and \( \mathbf{L}^* \) cannot be distinguished from one another.
• This result says that factor loadings are determined only up to multiplication by an orthogonal matrix $T$. In fact, the situation is worse than this in the sense that in many cases equivalent FA solutions exist which are not rotations of one another.

• Since such a multiplication has the interpretation of being a rotation of $L$, any FA solution is not unique and we may obtain an equivalent solution by rotating the solution by post-multiplying $L$ by an orthogonal matrix of our choosing.

• The opportunity to rotate a solution allows the analyst to choose the solution he/she finds most meaningful among many equivalent candidates.

• “Most meaningful” is subjective, and this feature of FA can be regarded as a strength of the method or as a serious weakness, depending on your point of view.
Fitting the FA Model:

The FA model is **scale invariant**. That is, if \( \mathbf{x} \) satisfies the FA model, and if we rescale \( \mathbf{x} \) by premultiplying by a diagonal matrix \( \mathbf{C} = \text{diag}(c_1, \ldots, c_p) \) of constants to form \( \mathbf{y} = \mathbf{C}\mathbf{x} \), then

\[
\mathbf{y} = \mathbf{C}(\mathbf{\mu} + \mathbf{LF} + \mathbf{\epsilon}) = \mathbf{C}\mathbf{\mu} + \mathbf{CLF} + \mathbf{C}\mathbf{\epsilon}
\]

and \( \text{var}(\mathbf{y}) = \mathbf{C}\Sigma\mathbf{C}' = \mathbf{CLL'} + \mathbf{C}\Psi\mathbf{C}' \)

\( \Rightarrow \) the rescaled response vector \( \mathbf{y} \) satisfies the FA model as well with loading matrix which is just \( \mathbf{C} \) times the loading matrix based on \( \mathbf{x} \), and rescaled specific variances \( c^2_1\psi_1, \ldots, c^2_p\psi_p \).

- FA, unlike PCA, is unaffected by re-scaling of the variables, so **according to the model**, equivalent results are obtained based on \( \Sigma \) and \( \mathbf{\rho} \).

Most methods of fitting the FA model do not have this scale invariance property. This is a serious deficiency of these methods.
However, scale invariant methods are typically more difficult computationally and require the assumption of multivariate normality, so scale invariant fitting methods persist.

- We will describe two methods for fitting the FA model: the principal component method and the maximum likelihood method. Only the latter is scale invariant.

The goal of these fitting methods is to find estimates $\hat{L}$ and $\hat{\Psi}$ such that the equation

$$ S = \hat{L}\hat{L}' + \hat{\Psi} \quad \text{or} \quad R = \hat{L}\hat{L}' + \hat{\Psi} $$

is satisfied, at least approximately.

- The resulting (non-unique) solution is then rotated to attempt to produce interpretable factors.

- For the PC method, choice between $R$ and $S$ should be made based on the same sort of considerations as applied in PCA (although in practice $R$ is almost always used). For ML estimation, it doesn’t matter whether $R$ or $S$ is used.
The Principal Component Method:

Recall that the spectral decomposition of \( \Sigma \) is given by \( \Sigma = \mathbf{E} \Lambda \mathbf{E}' \) where \( \mathbf{E} \) has columns the eigenvectors of \( \Sigma \) and \( \Lambda \) is a diagonal matrix containing the eigenvalues of \( \Sigma \).

Factor analysis is most useful when \( m \) is much smaller than \( p \). However, if \( m = p \) the covariance structure implied by the factor analysis model can be reproduced exactly by taking \( \mathbf{L} = \mathbf{E} \Lambda^{1/2} \) and \( \Psi = \mathbf{0} \):

\[
\Sigma = \mathbf{E} \Lambda \mathbf{E}' + \mathbf{0} = \underbrace{\mathbf{E} \Lambda^{1/2}}_{\mathbf{L}} \underbrace{\Lambda^{1/2} \mathbf{E}'}_{\mathbf{L}'} + \underbrace{\mathbf{0}}_{\Psi}
\]

Notice that the \( i^{th} \) column of \( \mathbf{L} \) is \( \sqrt{\lambda_i} \mathbf{e}_i \), or the coefficients of the \( i^{th} \) principal component scaled by a factor \( \sqrt{\lambda_i} \).
We would like to fit a FA model with \( m < p \). If we assume that the last \( p - m \) eigenvalues are small, then it makes sense to approximate \( \Sigma \) based on only the first \( m \) columns of \( L \): \( \Sigma \approx LL' + 0 \) where now we’ve redefined \( L \) to be

\[
L = (\sqrt{\lambda_1 e_1} | \cdots | \sqrt{\lambda_m e_m})
\]

We can improve on this by replacing \( \Psi = 0 \) with \( \Psi \) equal to the diagonal matrix that corrects the error in estimating \( \sigma_{11}, \ldots, \sigma_{pp} \) by the diagonal elements of \( LL' \) (the \( i^{th} \) of which is \( \sum_{j=1}^{m} \ell_{ij}^2 \)). That is, we take

\[
\Sigma \approx LL' + \Psi,
\]

where \( \Psi = \begin{pmatrix} \psi_1 & 0 \\ \vdots & \ddots \\ 0 & \psi_p \end{pmatrix} \),

and \( \psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2 \) \( i = 1, \ldots, p \).
Since the population variance-covariance is unavailable, we must fit the FA model to sample data using $S$ (or $R$) in place of $\Sigma$.

Thus, the principal factor solution becomes

$$S \approx \tilde{L}\tilde{L}^\prime + \tilde{\Psi},$$

where

$$\tilde{L} = (\hat{e}_1 | \cdots | \hat{e}_m) \begin{pmatrix} \sqrt{\hat{\lambda}_1} & 0 \\ & \ddots & \vdots \\ 0 & & \sqrt{\hat{\lambda}_m} \end{pmatrix}$$

where $(\hat{\lambda}_1, \hat{e}_1), \ldots, (\hat{\lambda}_m, \hat{e}_m)$ are the first $m$ ($m < p$) eigenvalue-eigenvector pairs (in decreasing order) computed from $S$, and

$$\tilde{\Psi} = \begin{pmatrix} \tilde{\psi}_1 \\ & \ddots \\ & & \tilde{\psi}_p \end{pmatrix}, \text{ where } \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2$$

and $\tilde{\ell}_{ij}$ is the $(i, j)^{th}$ element of $\tilde{L}$.

The principal factor estimates of the communalities are

$$\tilde{h}_i^2 = \sum_{j=1}^{m} \tilde{\ell}_{ij}^2 = \tilde{\ell}_{i1}^2 + \cdots + \tilde{\ell}_{im}^2$$
The Maximum Likelihood Method:

If we add to our FA model the assumption that the rows of $\mathbf{X}$ (the $p$–variate observations on each sample member) are multivariate normal, then it is possible to write down a likelihood function that is a function of $\mathbf{L}$ and $\Psi$.

Maximizing this likelihood function with respect to $\mathbf{L}$ and $\Psi$ yields the maximum likelihood estimates of these quantities.

- Note that to do this maximization, we must impose some uniqueness condition on $\mathbf{L}$. Typically, we maximize subject to the convenient uniqueness constraint $\mathbf{L}'\Psi^{-1}\mathbf{L} = \Delta$, where $\Delta$ is required to be a diagonal matrix.

- Maximum likelihood has the advantage that it is invariant to the scale of the original variables. This means that for ML, factoring $\mathbf{S}$ is equivalent to factoring $\mathbf{R}$.
• Another advantage to ML estimation is that there are several statistical methods to help determine the appropriate number of factors. These statistics are not appropriate when using other estimation methods such as the principal factor method.

• A disadvantage, of course, is the addition of the assumption of normality, which will not be justified for some data sets.
Example – Olympic Decathlon Data

Linden (1977) conducted a factor analytic study of Olympic decathlon scores since World War II. A total of 160 complete scores were available and Linden reports that each of the ten variables analyzed were normal or approximately normal. The original data are not available, but the sample correlation matrix from these data is as follows.

Following Linden, we will perform Factor Analyses using both principal component estimation and maximum likelihood estimation. We will consider both $m = 3$ and $m = 4$.

- See decath1.sas and decath1.lst.
• After reading the correlation matrix into data set deccorr (a data set of type corr) and printing it out, our first call to PROC FACTOR fits a FA model to the correlation matrix using the principal component method (option METHOD=PRIN) allowing PROC FACTOR to choose \( m \).

• The default method used by PROC FACTOR to choose \( m \) is to take \( m = \) the number of eigenvalues greater than 1. The SCREE option requests a scree plot of the eigenvalues, which appear on p.3.

• By the > 1 criterion the program chooses \( m = 3 \). The scree plot is not as clearly suggestive of a value for \( m \) as we might like, but it appears to suggests \( m = 4 \) or 5. We’ll settle on \( m = 4 \). Later, we’ll support this choice by factor interpretations and statistical evidence.

• In the second call to PROC FACTOR we again use the principal component method, but now with \( m = 4 \). The results from this call begin on p.5.
• First, the eigenvalues and other information similar to the results of PROC PRINCOMP are reported.

• Next, listed under “Factor Pattern” on p.5 are the factor loadings (the matrix $\tilde{L}$). It is not worthwhile to interpret these loadings prior to rotation.

• The “Variance explained by each factor” for each factor is the sum of the squared elements of that factor’s column of $\tilde{L}$. That is, $3.786608 = .69052^2 + \cdots + .14660^2 = \sum_{i=1}^{P} \tilde{l}_{i1}^2$.

• The “Final Communality Estimates” are the sum of the squared elements of the rows of $\tilde{L}$. That is, for RUN100M we have

$$ .837016 = .69052^2 + \cdots + .20603^2 $$

$$ = \sum_{j=1}^{m} \tilde{l}_{1j}^2 $$
• The third call to PROC FACTOR attempts to use ML to fit the \( m = 4 \) factor model. At the top of p.6, the ML method begins with the "Prior Communality Estimates" reported there as starting values and then proceeds through an iterative algorithm working toward a fitted model. However, notice that on iteration 3, the last of the ten communality estimates becomes \( > 1 \) (1.64033, to be exact). This is an impossible value for a communality, because it says that

\[
\sigma_{10,10} = 1.64033 + \psi_{10}, \quad \text{or} \quad 1 = 1.64033 + \psi_{10}, \quad \Rightarrow \quad \psi = -0.64033 \]

But since \( \psi_{10} \) is a (specific) variance, it makes no sense for it to be negative. This is what’s known as a "**Heywood case**", and the algorithm fails and prints an ERROR message.

• A simple fix is to do the following: whenever the ML algorithm encounters a communality \( > 1 \), re-assign it a value equal to 1 and proceed. This variant on the ML algorithm is obtained by using the HEYWOOD option and is implemented on the fourth call to PROC FACTOR.
• Notice (p.7) that with the Heywood option, at iteration 3, the tenth communality is set to 1.0 rather than 1.64033, and the algorithm proceeds to convergence and a solution. Notice that the factor pattern (loadings), variance explained by each factor and communalities (pp.8–9) differ substantially from the results from the principal component method. We will return to this example after we consider factor rotations and choosing $m$.

**Choosing the Number of Factors ($m$):**

• A complete factor analysis will typically involve fitting the FA model several times with more than one value of $m$ and more than one fitting method.

• A reasonable initial choice of $m$ can be made in the same way as we choose the number of principal components in a PCA: scree plots, number of eigenvalues $> 1$ (when working with R), cumulative percentage of variance explained $> $ some specified value.
To arrive at a final choice of $m$, several subjective and objective criteria should be taken into account:

**Subjective Criteria:**

1. Do not include *trivial* factors. Trivial factors being factors on which only one of the original variables has a high loading. Variables that load highly on only one factor are generally uncorrelated with the other variables and should be removed from the FA. It does not make sense to create a factor for such a variable when the variable itself can be used.

2. Many statistical packages will produce matrices of differences between the observed correlations between the variables and those that are reproduced by the FA solution (SAS option RES). If these differences are small, you might be able to reduce $m$. If some differences are large (many greater than .25 or a few greater than .40, say) then it might be necessary to increase $m$. 

246
3. Some packages also produce partial correlations among the variables after adjusting for the common factors (also produced by the RES option in SAS). If any of these partial correlations is large, then you should consider increasing $m$. If all are small, then perhaps $m$ can be reduced.

- Some researchers believe that the $\psi_j$’s should all be close to zero. Increasing $m$ will tend to do this; however, there’s nothing in the FA model that implies that the specific variances should be close to zero. Increasing $m$ to this end pushes FA toward PCA.

Objective Criteria:

If maximum likelihood estimation is used, several additional methods are available:

4. Likelihood ratio test: For any choice of $m$, a LRT is available to test the hypothesis that that number of factors is sufficient to explain the variance-covariance (or correlation) matrix being modelled.
As described when we introduced the concept of likelihood ratio testing in describing Hotelling’s $T^2$, the LRT compares the maximized value of the likelihood under the null and alternative hypothesis.

In this case, the null hypothesis is $H_0 : \Sigma = LL' + \Psi$ where $L$ is $p \times m$. Under the null hypothesis, $\Sigma$ is estimated by $\hat{L}\hat{L}' + \hat{\Psi}$ where $\hat{L}$ and $\hat{\Psi}$ are the ML estimates of these quantities. Under the alternative hypothesis, the (unconstrained) estimate of $\Sigma$ is $S_n = \{(n - 1)/n\}S$.

This leads to the LRT statistic:

$$(n - 1 - \frac{1}{6}(2p + 5) - \frac{2}{3}m) \log \frac{|\hat{L}\hat{L}' + \hat{\Psi}|}{|S_n|}$$

which is compared to the upper $\alpha$ critical value of the $\chi^2(\nu)$ distribution, where the degrees of freedom, $\nu$, equals $\nu = \frac{1}{2}\{(p - m)^2 - p - m\}$.  

248
5. With ML estimation various “information criteria” can be obtained which can be helpful in model selection. SAS will compute both Akaike’s information criterion (AIC) and Schwarz’s Bayesian criterion (SBC or, sometimes, BIC). Either one of these criteria can be examined for several values of $m$ and the value of $m$ that minimizes the chosen criterion is “best”.

- SBC tends to lead to more parsimonious models (smaller value of $m$) than AIC.

- The model selection tools available with ML tend to lead to inclusion of trivial factors. As mentioned before, the variables loading on trivial factors should be removed from the FA.

Rotation of Factor Loadings:

As mentioned previously, a FA solution is not unique and an equivalent solution can be obtained by rotating (postmultiplying by an orthogonal matrix) the estimated loading matrix. There are several methods for choosing the rotation so that the rotated loadings make the factors maximally interpretable.
• Rotations fall into two classes: *orthogonal rotations*, and *oblique rotations*.

• Orthogonal rotations correspond to a rigid rotation of the coordinate axes, preserving the orthogonality (uncorrelatedness) of the factors.

• Oblique rotations rotate the coordinate axes in such a way that the new axes are no longer orthogonal (perpendicular). This means that obliquely rotated factors are no longer independent. Since this violates an assumption of our original FA model, the use of oblique rotations is somewhat controversial. We will limit attention to orthogonal rotations.

**Varimax Rotation:**

A factor is easily interpreted when the magnitudes of the loadings of the $p$ variables on that factor are very large for a few variables and very small for the remaining variables. This suggests choosing a rotation that maximizes the spread among the loadings in each column of the estimated loading matrix $\hat{L}$.  

250
Actually, we just want to maximize the spread among the unsigned loadings, which leads to the idea of choosing the rotation to maximize the variance of the squared loading coefficients in each column of \( L \). Kaiser’s *raw varimax* rotation does just this.

- More commonly, the squared loading coefficients are scaled by the communalities for the corresponding variables so that the loadings are spread out the most for variables with larger communalities (the variables whose variances are best accounted for by the \( m \) common factors). This scaled version of the raw varimax rotation is what is commonly implemented as *the* varimax rotation method.

**Example – Decathlon Scores (Cont’d)**

- Refer to decath2.sas and decath2.lst.
• In decath2.sas I fit the FA model with $m = 3$ and $m = 4$ using both the principal component and ML method. We’ll consider the question of choosing $m$ based on the ML results, but it would also be useful to look at the residual correlation matrices for $m = 3$ and $m = 4$ based on the principal component method.

• On p.6 of decath2.lst we see that when using ML with $m = 3$, we obtain AIC=24.62 and SBC=−30.74. We also reject the hypothesis that 3 factors are sufficient using the LRT.

• On p.12 of decath2.lst we see that when using ML with $m = 4$, we obtain smaller values of AIC and SBC (AIC=−10.96, SBC=−44.79) and we do not reject the hypothesis that $m = 4$ factors are sufficient.
• In addition we can examine the typical size of the off-diagonal elements of the residual correlation matrices for $m = 3$ (p.8): RMS=.056, and for $m = 4$ (p.13): RMS=.022; and the typical size of the off-diagonal elements of the partial correlation matrix with $m = 3$ (p.9): RMS=.099, and for $m = 4$ (p.14): RMS=.039.

• These results lead us to choose $m = 4$, but before we do so, we should examine the factor loadings to determine whether there are any trivial factors in the solution. On p.15 the rotated ML loadings for $m = 4$ are given. Examining these loadings we see that factor 1 represents upper body strength or throwing ability, factor 2 represents sprinting speed, factor 3 represents jumping ability, and factor 4 represents running endurance.
• Both the 400m run and the 1500m run load on factor 4. But the loading for the 400m run is relatively small, and I would characterize this factor as a trivial factor. It is probably best to remove 1500m run from the set of variables and rerun the FA with \( m = 3 \) based on the \( p = 9 \) remaining variables.

• In the last call to PROC FACTOR, I did just that: set \( m = 3 \) and use only the first 9 variables. Notice that on p.17, the hypothesis that \( m = 3 \) is sufficient is not close to being rejected (\( p = .7846 \)). In addition, the rotated loadings on p.20 retain essentially the same interpretations.
Factor Scores:

A common purpose of FA is dimension reduction; that is, to reduce from a relatively large number of variables down to few, with little loss of information.

- In FA this can be done by replacing $p$ variables for which an $m$-factor model fits well with $m$ variables defined as the scores on the $m$ common factors in the model.

- To do this we need to have the original data (one $p$–variate observation for each of $n$ subjects). If so, we can obtain scores for each subject on each factor to obtain an $n \times m$ factor score matrix.

- Factor scores are not parameter estimates! Instead a common factor is an (unobserved) random variable, and we attempt to estimate (predict is really better terminology) the realization of this random variable for the $i$th subject.
• Factor scores are inherently much more speculative than directly measured random variables. (Factor score “estimation” involves estimating more unknown quantities than there is data.) For this reason, even some fans of FA regard factor scoring as too weakly supported theoretically for use.

The two most common methods of obtaining factor scores are the **weighted least squares method** (Bartlett’s method), and the **regression method** (Thompson’s method).

These methods have two things in common:

1. They treat the estimated factor loadings (the $\hat{\lambda}_{ij}$’s) and the specific variances (the $\hat{\psi}_i$’s) as if they were the true values.

2. The involve linear transformations of the original data, perhaps centered or standardized. Typically the estimated *rotated* loadings are used to compute factor scores.
Weighted Least Squares Method:

Recall that the FA model

\[ \mathbf{x} - \mathbf{\mu} = \mathbf{L}\mathbf{F} + \mathbf{\varepsilon} \]

is similar to the multivariate regression model with \( \mathbf{\varepsilon} \) playing the role of a vector of error terms.

Bartlett suggested estimating \( \mathbf{F} \) by the value that minimizes a weighted sum of squared errors, where the weights are the reciprocals of the error variances (specific variances). That is, choose \( \mathbf{F} \) to minimize

\[
\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \mathbf{\varepsilon}'\Psi^{-1}\mathbf{\varepsilon} = (\mathbf{x} - \mathbf{\mu} - \mathbf{L}\mathbf{F})'\Psi^{-1}(\mathbf{x} - \mathbf{\mu} - \mathbf{L}\mathbf{F})
\]

A little calculus yields the minimizer as

\[
\mathbf{f} = (\mathbf{L}'\Psi^{-1}\mathbf{L})^{-1}\mathbf{L}'\Psi^{-1}(\mathbf{x} - \mathbf{\mu})
\]

so the factor score for the \( j \)th case (subject) is given by

\[
\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}'\hat{\Psi}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\Psi}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})
\]

- If the correlation matrix is used to obtain the FA solution then replace \( (\mathbf{x}_j - \bar{\mathbf{x}}) \) above by \( \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) \) where \( \mathbf{D} = \text{diag}(s_{11}, \ldots, s_{pp}) \).

257
Regression Method:

For normally distributed common factors $\mathbf{F}$ and specific factors $\varepsilon$, the joint distribution of $\mathbf{x} - \mu$ and $\mathbf{F}$ is

$$
\begin{pmatrix}
\mathbf{x} - \mu \\
\mathbf{F}
\end{pmatrix}
\sim
N_{p+m}
\begin{pmatrix}
\begin{pmatrix} 0 \\
0
\end{pmatrix},
\begin{pmatrix} \Sigma & \mathbf{L} \\
\mathbf{L}' & \mathbf{I}
\end{pmatrix}
\end{pmatrix}.
$$

Therefore, by property 7, p.64 of the notes, the conditional distribution of $\mathbf{F} | \mathbf{x}$ is $p-$variate normal with mean and variance-covariance matrix

$$
\begin{align*}
E(\mathbf{F} | \mathbf{x}) &= \mathbf{L}'\Sigma^{-1}(\mathbf{x} - \mu) \\
\text{var}(\mathbf{F} | \mathbf{x}) &= \mathbf{I} - \mathbf{L}'\Sigma^{-1}\mathbf{L}
\end{align*}
$$

The regression method just estimates $\mathbf{F}$ by estimating the conditional expectation above. This leads to factor scores for the $j^{th}$ of $n$ cases (subjects) given by

$$
\hat{f}_j = \hat{\mathbf{L}}'\hat{\mathbf{S}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})
$$

or, if a correlation matrix is factored,

$$
\hat{f}_j = \hat{\mathbf{L}}'\hat{\mathbf{R}}^{-1}\hat{\mathbf{D}}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}).
$$

258
A third *ad hoc* method of obtaining factor scores that is often used is very simple. Let $z_1, \ldots, z_p$ represent the standardized random variables ($z_i = (x_i - \bar{x}_i)/\sqrt{s_{ii}}$). For each common factor, find the $z_i$’s that correlate highly with that factor (correlation above .4, say) and then score that factor by “averaging” those standardized variables. The averaging is done according to the sign of the correlation between the factor and the variable.

- For example, suppose a FA is done on $R$ and factor 1 has correlations (loadings) of .8, -.65, and .71 with standardize variables $z_1$, $z_2$, and $z_4$ and low-magnitude correlations with the other $z_i$’s. The the factor score for $j^{th}$ case for the first factor would be computed as $\frac{1}{3}(z_{j1} - z_{j2} + z_{j4})$.

- These factor scores have the advantage of simplicity of computation and interpretation.

- PROC FACTOR uses the regression method to compute factor scores.

- Lab 7 provides an example where factor scores are obtained to replace a large set of original variables (for data reduction).
Cluster Analysis
(See ch. 12, Johnson & Wichern; and ch.9, Johnson.)

Basic Goal: To organize items into groups, or clusters, based on similarities or distances among the objects.

Distinct from classification because in cluster analysis the groups are not predefined and no assumptions are made about group structure.

- Cluster analysis is typically a descriptive, and hence exploratory, technique.

Several (likely correlated) random variables are measured on each item so that a random vector is associated with each item. Distances or similarities among items are computed from these multivariate observations.
Distance Measures and Similarity Coefficients:

We have already encountered three ways to measure the distance between two \( p \)-dimensional observations \( x_j \) and \( x_k \):

1. Euclidean Distance:

\[
d(x_j, x_k) = \sqrt{(x_j - x_k)'(x_j - x_k)}
\]

2. Karl Pearson Distance, or scaled Euclidean Distance:

\[
d(x_j, x_k) = \sqrt{(x_j - x_k)'V^{-1}(x_j - x_k)}
\]

where \( V = \text{diag}(s_{11}, s_{22}, \ldots, s_{pp}) \) is computed from all objects, \( x_1, \ldots, x_n \).

3. Statistical Distance or Mahalanobis Distance:

\[
d(x_j, x_k) = \sqrt{(x_j - x_k)'S_P^{-1}(x_j - x_k)}
\]

where \( S_P \) is the pooled sample var-cov matrix from the clusters to which \( x_j \) and \( x_k \) belong. Since we need to know the cluster structure to compute \( S_P \), Mahalanobis Distance is not a good choice for cluster analysis.
4. Other useful distance measures are based on the Minkowski Metric which gives rise to distance measures of the form

\[ d(x_j, x_k) = \left( \sum_{i=1}^{p} |x_{ji} - x_{ki}|^m \right)^{1/m} \]

a. For \( m = 1 \) we have “city-block” distance:

\[ d(x_j, x_k) = \sum_{i=1}^{p} |x_{ji} - x_{ki}| \]

b. For \( m = 2 \) we have Euclidean distance:

\[ d(x_j, x_k) = \left( \sum_{i=1}^{p} |x_{ji} - x_{ki}|^2 \right)^{1/2} \]
Distance measures are typically used when observations consist of quantitative measurements (temperature, height, pulse). Often items are compared on the basis of the presence or absence of certain (qualitative) characteristics. For example, leaves: jagged edges, rubbery texture, fan shape. In such cases the presence or absence of the characteristic is represented by a binary (0,1) variable and similarity coefficients between items are defined.

Consider two items (leaves, say) on which the presence or absence of $p = 6$ characteristics have been assessed yielding the following data:

We can summarize similarities and dissimilarities between the leaves in a $2 \times 2$ table as follows.
In general, we have

A similarity coefficient $s(j, k)$ between items $j$ and $k$ can be defined from such a table in a variety of ways. For example:

- Proportion of agreements $s(j, k) = (a + d)/p$
- Proportion of 1-1 agreements: $s(j, k) = a/p$.
- Frequency of agreements double-weighted relative to disagreements:

$$s(j, k) = \frac{2(a + d)}{2(a + d) + b + c}$$

- Ratio of agreements to disagreements with 0-0 matches excluded:

$$s(j, k) = \frac{a}{b + c}$$
• Two similarity measures $s$ and $\tilde{s}$ have a \textbf{monotone} relationship if

$$s(j, k) \geq s(j, \ell) \implies \tilde{s}(j, k) \geq \tilde{s}(j, \ell)$$

for all $j, k, \ell$.

• Monotonicity is important because some clustering procedures are unaffected if the definition of similarity is changed in a manner that leaves the relative orderings of similarities unaffected.

Types of clustering procedures:

I \textbf{Agglomerative hierarchical methods:} Start with one cluster per item and sequentially merge clusters ending with a single cluster for all items.

II \textbf{Divisive hierarchical methods:} Start with a single cluster for all items and sequentially divide it into subgroups until eventually one group per item is obtained.
III Non-hierarchical methods either start from an initial set of seed points or an initial partition of items into groups. Assignments and reassignments of items to groups or mergers and divisions of groups are performed systematically, but not in a nested (or hierarchical) way.

The agglomerative hierarchical clustering methods we will consider utilize a linkage algorithm based on the $n \times n$ distance matrix giving inter-object distances or similarity coefficients between every pair of items that we wish to analyze. The algorithm is as follows:

1. Start with $n$ clusters, each corresponding to a single item.
2. Merge the two clusters that are “closest” according to the distance matrix (largest similarity coefficient or smallest distance).
3. Recompute distance matrix based on one fewer cluster.
4. Repeat steps 2 and 3 until all clusters are merged.

We will consider three linkage methods that differ in step 3.
Single Linkage:

Entry in the distance matrix corresponding to a pair of clusters is the nearest neighbor (smallest) distance or (largest) similarity between items in the two clusters.

Example

Suppose that we have 6 items that we would like to cluster and the items are associated with the multivariate observations $\mathbf{x}_1, \ldots, \mathbf{x}_6$. Assuming the $\mathbf{x}$’s are vectors of quantitative variables we can compute Euclidean distances between the items: $d(\mathbf{x}_1, \mathbf{x}_2)$, $d(\mathbf{x}_1, \mathbf{x}_3)$, $\ldots$, $d(\mathbf{x}_5, \mathbf{x}_6)$ that we arrange in a $6 \times 6$ distance matrix:

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0.31 & 0.23 & 0.32 & 0.26 & 0.25 \\
2 & 0.34 & 0.21 & 0.36 & 0.28 \\
3 & & 0.31 & 0.04 & 0.07 \\
4 & & & 0.31 & 0.28 \\
5 & & & & 0.09 \\
6 & & & & & \\
\end{array}
\]
The initial clustering is denoted by $C_0$ and is given by

$$C_0 = \{[1], [2], [3], [4], [5], [6]\}$$

According to the distance matrix, the closest clusters are [3] and [5]. So at the first iteration we merge these two clusters to obtain

$$C_1 = \{[1], [2], [3, 5], [4], [6]\}$$

To recompute the distance matrix, we need the distance between [1] and [3, 5], the distance between [2] and [3, 5], etc.

- The nearest-neighbor method takes the distance between [1] and [3, 5] to be $\min(0.23, 0.26) = 0.23$.

- The distance between [2] and [3, 5] is $\min(0.34, 0.36) = 0.34$.  

268
Similarly, the other inter-cluster distances can be computed to yield

\[
\begin{array}{ccccc}
[1] & 0.31 & 0.23 & 0.32 & 0.25 \\
[2] & 0.34 & 0.21 & 0.28 & \\
[3, 5] & 0.31 & 0.07 & \\
[4] & 0.28 & & \\
\end{array}
\]

At the second iteration we notice that the closest clusters are now [6] and [3, 5] which we merge to yield

\[C_2 = \{[1], [2], [3, 5, 6], [4]\}\]

Recomputing the distance matrix we get

\[
\begin{array}{ccccc}
[1] & 0.31 & 0.23 & 0.32 \\
[2] & 0.28 & 0.21 & \\
[3, 5, 6] & & 0.28 & \\
\end{array}
\]
At the third iteration we merge [2] and [4] to give

\[ C_3 = \{[1], [2, 4], [3, 5, 6]\} \]

with distance matrix

\[
\begin{array}{ccc}
[1] & 0.31 & 0.23 \\
[2, 4] & & 0.28 \\
[3, 5, 6] & & \\
\end{array}
\]

At the fourth iteration we merge [1] and [3, 5, 6] to give

\[ C_4 = \{[1, 3, 5, 6], [2, 4]\} \]

with distance matrix

\[
\begin{array}{cc}
[1, 3, 5, 6] & [2, 4] \\
[1, 3, 5, 6] & & 0.28 \\
[2, 4] & & \\
\end{array}
\]

Finally we merge [1, 3, 5, 6] and [2, 4] to yield the final single group clustering:

\[ C_5 = \{[1, 2, 3, 4, 5, 6]\} \]
The results of the clustering algorithm can be summarized in a hierarchical tree diagram or dendrogram. The dendrogram for the preceding example appears below

- The lengths of the branches of the tree are proportional to the distances between clusters at the iteration at which clusters were combined.

- From the dendrogram we can tell the order of mergers, the distances between clusters and the clustering structure.

- Dendrograms can be helpful for determining the actual number of clusters in the data set. We will return to this point later.
Complete Linkage:

In step 3 of the linkage algorithm the distance between any two clusters is defined to be the distance between their furthest members.

- In the example above, we first merge [3] and [5] just as before, but when we recompute the distance matrix in iteration 1,

\[ d([1], [3, 5]) = \max(0.23, 0.26) = 0.26 \]

and

\[ d([2], [3, 5]) = \max(0.34, 0.36) = 0.36, \]

etc., giving the following distance matrix at the end of iteration 1:

\[
\begin{array}{cccccc}
[1] & 0.31 & 0.26 & 0.32 & 0.25 \\
[2] & & 0.36 & 0.21 & 0.28 \\
[3, 5] & & & 0.31 & 0.09 \\
[4] & & & & 0.28 \\
[6] & & & & & \end{array}
\]
Average Linkage:

In step 3 of the linkage algorithm the distance between any two clusters is defined to be the average of the distances between all pairs of points such that one element of each pair comes from each cluster.

- In the example above, we first merge [3] and [5] just as before, but when we recompute the distance matrix in iteration 1,

\[ d([1], [3, 5]) = \text{avg}(0.23, 0.26) = 0.245 \]

and

\[ d([2], [3, 5]) = \text{avg}(0.34, 0.36) = 0.35, \]

etc., giving the following distance matrix at the end of iteration 1:

\[
\begin{array}{cccccc}
1 & 2 & 3, 5 & 4 & 6 \\
1 & 0.31 & 0.245 & 0.32 & 0.25 \\
2 & 0.35 & 0.21 & 0.28 \\
3, 5 & 0.31 & 0.08 \\
4 & \text{ } & \text{ } & 0.28 \\
6 & \text{ } & \text{ } & \text{ } & \text{ }
\end{array}
\]
Example – Languages:

Suppose we are interested in how languages are interrelated. One way to explore this is to run a cluster analysis on the languages based on the similarities of words in the various languages for the same things. Specifically, we consider English, Norwegian, Danish, Dutch, German, French, Spanish, Italian, Polish, Hungarian, and Finnish. We take as the “distance” between languages $j$ and $k$:

$$d(j, k) = 10 - \text{the number of matching first letters in the words for integers 1–10}$$

E.g., English and French:
English: one,two,three,four,five,six,seven,eight,nine,ten
French: un,deux,trois,quatre,cinq,six,sept,huit,neuf,dix

$$d(\text{eng, fr}) = 10 - 4 = 6$$

- Cluster analyses for these 10 languages based on such distances are performed in lang1.sas using PROC CLUSTER.
• The initial distance matrix is input into SAS data set lang, and printed on p.1 of the output.

• We perform three separate cluster analyses, first with method=sin (single linkage), second with method=com (complete linkage) and third with method=ave (average linkage).

• The iteration history of the single linkage algorithm is printed on p.2. At several points in the iteration history there are ties. The choice of which clusters to join in the case of ties is typically arbitrary (SAS). Choice affects single linkage results much less than complete and average linkage results.

• Dendrograms are computed with PROC TREE from SAS data sets tree1, tree2 and tree3 which are output from PROC CLUSTER using the OUTTREE= option.
General Comments:

- Single linkage tends to maximize the “connectedness” of a pair of clusters and tends to suggest fewer clusters than the complete linkage method.

- Complete linkage tends to minimize the intra-cluster distances at each iteration so tends to find compact clusters.

- Most other methods including the average linkage procedure fall somewhere in between these extremes.

- Advisable to try several methods. If similar results are obtained then one can believe that natural clusterings actually exist.

- Most clustering methods do not formally take into account sources of error and variation and are sensitive to outliers.
• Hierarchical procedures do not provide for reallocation of items that may have been incorrectly grouped at an early stage.

• Sometimes useful to *perturb* data by adding small random errors to assess the *stability* of the solution.
How many Clusters?

**Pseudo $T^2$ Test:** In deciding if two clusters should be combined, we can compute a Hotelling’s $T^2$-like statistic to compare the means of the two clusters.

- If cluster means not significantly different, then clusters can be combined. If cluster means are significantly different, then clusters should stay separate.

- “Pseudo” because the assumptions of Hotelling’s $T^2$ test are not satisfied. E.g., observations are not allocated at random to the two clusters being compared by the clustering algorithm.

- Available in SAS with the PSEUDO option.

**Cubic Clustering Criterion (CCC):** Statistic proposed by Sarle (1983) that can be plotted against the number of clusters for various choices of the number of clusters. Peaks on this plot with CCC $> 3$ are supposed to correspond to an appropriate number of clusters.
• Not recommended with single link method.

• Available in SAS with CCC option.

**Dendrogram:** Sometimes the dendrogram can suggest the true clustering structure. An example of an ideal dendrogram suggesting three clusters appears below.
Graphical Procedures:

Different algorithms very often give different results for same data. Therefore, verification, evaluation and fine-tuning of results is essential. Several helpful graphical methods are available for such tasks.

- **Scatterplots:** Most useful when clustering is based on $p = 2$ variables. For $p > 2$, plotting pairs of variables can be helpful. The following scatterplot clearly suggests 3 clusters.
• **Principal Components:** When \( p > 2 \), a PCA can be performed to see if a high percentage of the variability in the data can be accounted for by 2 PCs. If so, a scatterplot of the first 2 PCs can suggest the appropriate number of clusters.

  - Running a cluster analysis on the first few principal component scores is often a good alternative to running cluster analysis on original variables.

  - When plotting PC scores or using PC scores in a cluster analysis, PC scores must not be standardized.
• **Andrews’ Plots:** A $p-$dimensional vector of measurements $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ can be represented by the finite Fourier Series

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots$$

a function of $t$ for values of $t$ between $-\pi$ and $\pi$. Andrews suggested that the Fourier series be plotted for each item (i.e., for $\mathbf{x}_1, \ldots, \mathbf{x}_n$). Similar items (e.g., items belonging to the same cluster) should have similarly shaped Fourier curves.

- Variables should be standardized prior to constructing Andrews’ plots.

- Curves will be affected by interchanges of coefficients. E.g.,

$$\mathbf{x} = (\text{sodium, fat, protein})'$$

will give different curves than

$$\mathbf{x} = (\text{fat, sodium, protein})'.$$
Therefore, its a good idea to try several different orderings before deciding which is best for a given data set.

– The book considers an example where 22 U.S. public utility companies (e.g., Nevada Power Co., Florida Power & Light Co., etc.) are clustered on the basis of eight variables (e.g., sales, percent nuclear, etc.). The cluster structure obtained in the \( K \)–means non-hierarchical clustering method for \( K = 5 \) is given in example 12.13 in the text on p.758. The SAS program andrew1.sas produces 5 Andrews’ plots, one corresponding to each cluster obtained in example 12.13 (see handout).
• **Star Plots:** Another technique for plotting a \( p \)-dimensional response on each of \( n \) sample members is the star plot. Each of the \( p \) responses is plotted as the distance along a ray extending from a center point.

  - See the handout labelled stars1.s. Here I have used the Splus statistical programming language to produce star plots for each of the 22 power utilities from the example in the text. Notice that there are eight equally spaced rays in each star plot corresponding to the eight variables measured on each company.

  - Star plots for all 22 utilities appear first, then 5 separate pages of plots corresponding to the 5 clusters identified in example 12.13.

  - All responses have been scaled to fall in the range \((0, 1)\).
• **Chernoff Faces:** Another way to plot multivariate observations is with faces. Each of $p$ facial characteristics is assigned to one of the $p$ responses.

  - See the handout labelled faces1.s. Again Splus has been used to generate face plots for the 22 utility companies.

  - All responses have been scaled to fall in the range $(0, 1)$.

  - In these face plots $x_1, \ldots, x_8$ have been assigned to the following facial features, respectively: 1-area of face, 2-shape of face, 3-length of nose, 4-location of mouth, 5-curve of smile, 6-width of mouth, 7-location of eyes, 8-separation of eyes. Up to 15 features are available in Splus.
Example – Frozen Pizza:

Data on 56 frozen pizzas found in supermarkets were collected. Each individual pizza was pureed and thoroughly mixed after which a sample of the mixture was taken for nutrient analysis. The variables ($p = 7$) measured on each sample included the percent moisture of the pizza (MOIS), the amount of protein per 100 g in the sample (PROT), the amount of fat per 100 g in the sample (FAT), the amount of ash per 100 g in the sample (ASH), the amount of sodium per 100 g in the sample (SODIUM), the amount of carbohydrates per 100 g in the sample (CARB), and the calories per gram in the sample (CAL).

- These data were analyzed using an average linkage algorithm in the SAS program pizza1.sas.
• The STANDARD option standardizes the seven variables so that scaled Euclidean distance will be used rather than simple Euclidean distance. Necessary because the 7 variables have very different variances and some variables are measured on different scales. Based on raw data, cluster identification would be based mostly on differences between items for variables with large variances (e.g., protein).

• Output from PROC CLUSTER is on pp.1–3.

  – Option S results in the summary statistics and eigenvalues of p.1. Notice that the first two eigenvalues are the only ones greater than 1 and indicate that the first 2 PCs account for over 91% of the total variance.

  – The options CCC and PSEUDO lead to the inclusion of CCC and the pseudo $T^2$ statistic (PST2) in the algorithm history. Note that CCC is not computed when the number of clusters is greater than 20% of $n$. 
• How many Clusters?

- A plot of CCC against the number of clusters (\( \_NCL\_ \)) appears on p.4. We see a peak at 8 clusters and there may be another at 10 clusters.

- Values of PST2 printed in each row can be used to help determine whether the two clusters in that row should have been combined. If PST2 is large the clusters should not have been combined. “Large” is relative. Reading up from the bottom of p.3 we see large values in row 6 and row 9, indicating perhaps that the merger taking us from 10 to 9 clusters should not have occurred, or that the merger going from 7 to 6 clusters should not have occurred.

- Thus CCC and Pseudo \( T^2 \) suggest that between 7 and 10 clusters are appropriate.
• A dendrogram is printed on pp.5–6 by PROC TREE. The OUT=tout10 option combined with NCLUSTERS=10 requests that a SAS data set named tout10 be created with the cluster assignments indicated by the dendrogram when there are 10 clusters. This data set appears on pp.8–9.

• Next a PCA is performed on these data. This is done because we noticed that the over 91% of the total variability in our original 7 variables is captured by just two PCs. Therefore, scatterplots of the scores on the first 2 PCs will be useful in identifying the cluster structure.

• A scatterplot of the first 2 PCs appears on p.11. How would you assign clusters?

• We can see how the average linkage algorithm assigned clusters when the number of clusters was 10 on p.12. Does this agree with your impression from p.11?
• The brand for each of the original 56 pizzas is known and there are a total of 10 brands represented in the sample. We can see how well the cluster algorithm reproduced the brand identification by plotting the first two PCs with plotting symbol corresponding to brand name as on p.13. Agreement is fairly good. Considering how overlapped some brands are, it is unreasonable for us to expect much better.

• We can see how the average linkage algorithm assigned clusters when the number of clusters was 7 on p.16. Does this agree with your impression from p.11?

• The results of the cluster analysis can only be suggestive of what cluster structure really exists. The investigator is free to (and should) fine-tune the analysis based on

  – plots such as those on pp.11–13 and p.16,
  – other graphical techniques such as Andrews’ plots and Chernoff faces,
  – subjective judgement.
Nonhierarchical Clustering Methods:

Methods to cluster items into $K$ clusters where $K$ may be specified in advance or as part of the clustering algorithm.

Starting Point:

1. an initial partition of items into groups, or
2. an initial set of seed points.

The $K$–means Method (implemented in the SAS procedure FASTCLUS) is an example of a nonhierarchical clustering method. It consists of the following three steps:

1. $K$ observations (called seeds) are selected and taken as the centroids of $K$ clusters.

2. For each item, (re-)assign the item to the cluster whose centroid is nearest (usually according to Euclidean or scaled Euclidean distance). After all $n$ items have been (re-)assigned, compute the current cluster means and use these means to replace the old centroids.
3. Repeat step 2 until no more reassignments take place.

Alternatively step 2 can be replaced by

2*. For each item, (re-)assign the item to the cluster whose centroid is nearest. After each (re-)assignment, recompute cluster means and replace the cluster centroids with the current cluster means (DRIFT option in PROC FASTCLUS).

Another variation is to start the algorithm from an initial partition of the items into \( K \) clusters, the means of which are taken to be the initial centroids.
Example (using step (2)):

Suppose that we have four items $A, B, C, D$ and we have observed a bivariate response on each item as follows:

<table>
<thead>
<tr>
<th>Item</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>-2</td>
</tr>
<tr>
<td>D</td>
<td>-3</td>
<td>-2</td>
</tr>
</tbody>
</table>

We will illustrate the $K$—means algorithm assuming $K = 2$ clusters.

1. Somewhat arbitrarily, we take the observations corresponding to $A$ and $D$ as the initial seeds.
2. In step 2 we have the following Euclidean distances:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroid</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>(5, 3)'</td>
<td>0</td>
<td>$\sqrt{40}$</td>
<td>$\sqrt{41}$</td>
<td>$\sqrt{89}$</td>
</tr>
<tr>
<td>.</td>
<td>(-3, -2)'</td>
<td>$\sqrt{89}$</td>
<td>$\sqrt{13}$</td>
<td>$\sqrt{16}$</td>
<td>0</td>
</tr>
</tbody>
</table>

For example,

$$d^2(A, (-3, -2)') = (5 - (-3))^2 + (3 - (-2))^2$$

$$= 64 + 25 = 89$$

$$d^2(B, (5, 3)') = (-1 - 5)^2 + (1 - 3)^2$$

$$= 36 + 4 = 40$$

$$d^2(B, (-3, -2)') = (-1 - (-3))^2 + (1 - (-2))^2$$

$$= 4 + 9 = 13$$

Based on these distances we get

$$C_1 = \{[A], [B, C, D]\}$$

with centroids (5, 3)' and

$$\left(\frac{-1 + 1 - 3}{3}, \frac{1 - 2 - 2}{3}\right)' = (-1, -1)'$$
2. Repeating step 2, we now have the following Euclidean distances:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroid</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>[A]</td>
<td>(5, 3)'</td>
<td>0</td>
<td>$\sqrt{40}$</td>
<td>$\sqrt{41}$</td>
<td>$\sqrt{89}$</td>
</tr>
<tr>
<td>[B,C,D]</td>
<td>(−1, −1)'</td>
<td>$\sqrt{52}$</td>
<td>$\sqrt{4}$</td>
<td>$\sqrt{5}$</td>
<td>$\sqrt{5}$</td>
</tr>
</tbody>
</table>

Based on these distances, no reassignments are made and we have our final clustering:

$$C_2 = \{[A], [B, C, D]\}$$

There are several disadvantages to hierarchichal clustering methods:

- Requires us to guess the number of clusters that exist in the data.
- Greatly influenced by the choice of initial seeds. PROC FASTCLUS has an algorithm with several options for selecting initial seeds.
• Can be computationally infeasible because there are too many choices of $K$ to try and too many choices for seeds.

• When using step (2*), the order in which the data appear in the data set will affect results. PROC FASTCLUS offers a RANDOM option so that the procedure can be performed several times on a given data set under random reorderings of the data.

Advantage:

• Can be applied to much larger data sets because neither the data nor a large $n \times n$ distance matrix need be held in computer memory.

Any non-hierarchical clustering procedure should be performed several times with various choices of $K$. 
Example – Fisher’s Iris Data:

In lab recently we used the $K$–means algorithm implemented in PROC FASTCLUS to analyze Fisher’s Iris data. Let’s again consider iris1.sas. In this example we used PROC FASTCLUS to produce $K = 2$ clusters and $K = 3$ clusters.

- When producing $K = 3$ clusters we saw how sensitive to initial seeds the algorithm was. On p.9 of iris1.lst we have the results for $K = 3$ with initial seeds selected automatically by SAS. On p.13 we have results for $K = 3$ with initial seeds selected at random from the original data. Considering our knowledge of true varieties, we can see the the automatic selection of seeds is preferrable in this case.

- In this example we know that, in reality, $K = 3$. However, usually we will have to run the algorithm for several values of $K$ and choose which $K$ is appropriate. A statistic that can be used to make this choice is Beale’s Pseudo $F$ statistic.
Suppose there are two possible clusterings, one consisting of $c_1$ clusters and the second consisting of $c_2$ clusters where $c_2 < c_1$. Let $W_1, W_2$ be the corresponding within-cluster sum of squared deviations from the cluster means:

First Clustering:

$$C_1 = \left\{ \left[ \mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1} \right], \left[ \mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2} \right], \ldots, \left[ \mathbf{x}_{c_{11}}, \ldots, \mathbf{x}_{c_{1n_2}} \right] \right\}$$

From which we compute $W_1$:

$$W_1 = \sum_{k=1}^{c_1} \sum_{j=1}^{n_j} (\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)'(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)$$
Second Clustering:

\[ C_2 = \left\{ \left[ x_{11}, \ldots, x_{1n_1} \right], \left[ x_{21}, \ldots, x_{2n_2} \right], \ldots, \right\} \]

\[
\begin{cases} 
\text{cluster 1} & \left[ x_{c11}, \ldots, x_{c1n_2} \right] \\
\text{cluster 2} & \left[ x_{c21}, \ldots, x_{c2n_2} \right] \\
\text{cluster } c_2 & \end{cases}
\]

From which we compute \( W_2 \):

\[
W_2 = \sum_{k=1}^{c_2} \sum_{j=1}^{n_j} (x_{kj} - \bar{x}_k)'(x_{kj} - \bar{x}_k)
\]

\( W_1 \) is the sum of the squared distances from the cluster mean for all items in cluster 1 and \( W_2 \) is the sum of the squared distances from the cluster mean for all items in cluster 2.

- If \( W_1 \) and \( W_2 \) are nearly the same size then the clustering with the smaller number of clusters (\( C_2 \)) is preferred (for simplicity).

- If, however, \( W_1 \) is “significantly” smaller than \( W_2 \), then \( C_1 \) with a larger number of clusters is preferred.
To make this comparison, Beale suggested the statistic

$$F^* = \frac{W_2 - W_1}{W_1} \frac{(n - c_1)k_1}{(n - c_2)k_2 - (n - c_1)k_1}$$

where $n =$ the total number of items, $k_1 = c_1^{-2/p}$, and $k_2 = c_2^{-2/p}$.

If $F^* > F_{crit}$ then we choose $C_1$ (larger number of clusters), where $F_{crit}$ is a critical point (corresponding to 0.05, say) of the

$$F(k_2(n - c_2) - k_1(n - c_1), k_1(n - c_1))$$

distribution.
In the iris example, we can compare the $K = 3 = c_1$ clustering plotted on p.9 with the $K = 2 = c_2$ clustering on p.5. For this comparison, $W_1 = 79.012$ (from p.10) and $W_2 = 152.348$ (from p.6). $F^*$ is given by

$$F^* = \frac{152.348 - 79.012}{79.012} \times \frac{(150 - 3)3^{-2/4}}{(150 - 2)2^{-2/4} - (150 - 3)3^{-2/4}} = 3.982$$

The $p-$ value when we compare $F^*$ with the $F(k_2(n-c_2) - k_1(n-c_1), k_1(n-c_1)) \approx F(20,85)$ distribution is $p < .0001$ so we conclude that the $K = 3$ clustering is preferrable (as is appropriate here).
**Multidimensional Scaling**
(See section 12.5 of Johnson and Wichern.)

**Basic Problem:** Given a distance (or similarity) matrix for all pairs of \( n \) items, find a representation of the items in a few dimensions such that the interitem proximities nearly match the original distances (similarities).

- Usually want to find a two-dimensional representation. In other words, we want a two-dimensional picture, or map, showing the interrelationships among items.

- Often in MDS we observe only the matrix of distances (or similarities), not the points themselves.

**Metric vs. Nonmetric:**

- In one approach to MDS only the rank orders of the \( n(n - 1)/2 \) original distances (similarities) are used and the goal is to reproduce the ordering in a lower dimension. Such an approach is called **nonmetric** MDS.
• Alternatively, we base our lower dimensional map on the actual magnitudes of the original distances (similarities). This latter approach is **metric MDS**.

• Typically, metric MDS is used in cases in which original distances are measured and it is desired to reproduce these distances in lower dimension. Nonmetric MDS is more closely associated with situations where original “distances” are based on judgement. E.g., in psychology it is often the case that subjects can say how similar or different pairs of items are, but are unable to produce an overall picture of the relationships among the objects.

We will concentrate on nonmetric MDS.
Example – British Towns:

Mardia, Kent, and Bibby (1979) present an example of MDS based on road distances between 12 British towns. The computations for nonmetric MDS are performed by PROC MDS in the SAS program britdist1.sas.

- The input to PROC MDS is a $12 \times 12$ distance matrix giving road distances between each pair of British towns.

- An output data set called out is specified with the OUT= option. This data set is printed on p.2 and gives points (DIM1, DIM2) in two dimensions to represent each town (2 dimensions is the default).

- The two dimensional map is plotted on p.3 and again on p.4 (flipped upside down). Notice that the map looks very similar to the true geographical locations of these towns.
• If Euclidean distances (distances “as the crow flies”) for these towns were available rather than road distances, a metric MDS algorithm could exactly replicate the true geographical locations of the towns apart from transformations, reflections and rotations.

• Nonmetric MDS solutions are indeterminate with respect to transformations, reflections, rotations, and uniform expansions or contractions of scale.
The basic MDS algorithm:

Let $\delta_{jk}$ = the distance between items $j$ and $k$ from our original distance matrix. Let $t = \text{the number of dimensions of the desired solution (usually } t = 2)$. MDS programs typically follow an algorithm involving the following steps:

1. A starting configuration is set up for the $n$ objects in $t$ dimensions; i.e., coordinates $(x_1, x_2, \ldots, x_t)$ are assumed for each object.

2. The Euclidean distances between objects are calculated based on the current configuration. Let $d_{jk}$ be this Euclidean distance between objects $j$ and $k$.

3. A regression of $d_{jk}$ on $\delta_{jk}$ is performed so that we obtain a “prediction” $\hat{d}_{jk}$ based on $\delta_{jk}$. The $\hat{d}_{jk}$’s are called disparities, and can be thought of as the original inter-item distances (the $\delta_{jk}$’s) scaled to match the configuration distances (the $d_{jk}$’s) as closely as possible.
The regression can be linear, polynomial or monotonic. A monotonic regression assumes simply that if $\delta_{jk}$ increases then $d_{jk}$ increases or remains constant. A monotonic regression corresponds to nonmetric MDS.

4. The goodness of fit (or, sometimes as in SAS, “badness of fit”) between the configuration distances and the disparities is measured by a suitable statistic; for example the stress is often used, calculated as

$$\text{STRESS} = \left\{ \frac{\sum_{j=1}^{n} \sum_{k=j+1}^{n} (d_{jk} - \hat{d}_{jk})^2}{\sum_{j=1}^{n} \sum_{k=j+1}^{n} \hat{d}_{jk}^2} \right\}^{1/2}$$

5. The coordinates $(x_1, \ldots, x_t)$ for each item are changed slightly in such a way that the stress is reduced.

Steps 2–5 are repeated until stress can no longer be reduced. The coordinates $(x_1, \ldots, x_t)$ for each item are the end product. The coordinates (variables DIM1, DIM2, \ldots, in SAS) can be used to plot the items in $t$ dimensions.
Kruskal (1964) the following guidelines

<table>
<thead>
<tr>
<th>Stress</th>
<th>Goodness of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>Poor</td>
</tr>
<tr>
<td>10%</td>
<td>Fair</td>
</tr>
<tr>
<td>5%</td>
<td>Good</td>
</tr>
<tr>
<td>2.5%</td>
<td>Excellent</td>
</tr>
<tr>
<td>0%</td>
<td>Perfect</td>
</tr>
</tbody>
</table>

Example – New Jersey Representatives:

The following distance matrix was formed based on the voting records of 15 New Jersey Representatives to the US Congress. The entries in the matrix are the number of voting disagreements on 19 bills concerned with environmental matters.

• The SAS program vote1.sas performs two MDS analyses of these data, first using $t = 3$ dimensions, and second, using $t = 2$ dimensions.

• The stress for $t = 3$ was 6.65%. For $t = 2$ the stress was 11.26%.
• A three-dimensional scatterplot of the US Representatives is attached as the last page of the handout. In addition, pairwise scatterplots of the three dimensions appear on pp.3–5.

• The party affiliations have been written in next to the names on p.3. Notice that dimension 1 seems to capture part affiliation.

• A closer look at the data helps interpret dimension 2. It turns out that Sandman and Thompson were the Representatives who abstained from the most votes and the Representatives with small values of DIM2 abstained seldomly or not at all.

• Dimension 3 is not easily interpreted from the given information, but it must reflect certain patterns of differences in voting behavior.