
Introduction

Edps/Psych/Stat/ 584

Applied Multivariate Statistics

Carolyn J. Anderson

Department of Educational Psychology



© Board of Trustees, University of Illinois



Outline

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Objectives of Multivariate Analysis
- Topics covered
- Statistics review (notation used in class)
- Graphical techniques

Reading: Johnson & Wichern, Chapter 1



Objectives

● Outline

Objectives

● Objectives

● More Specific Objectives

Topics

Statistics Review

Graphical Techniques

- Multivariate Data: When studying complex phenomenon, we need to collect measurements (observations) on many different variables.
- Multivariate techniques: methods to elicit information from multivariate data. Most are statistical.
- “Dependent” variables are numerical, metric, “continuous.”



More Specific Objectives

● Outline

Objectives

● Objectives

● More Specific Objectives

Topics

Statistics Review

Graphical Techniques

- Data reduction or structural simplification: represent phenomenon as simply as possible without losing too much information.
- Sorting and grouping: create groups of “similar” objects or classify objects into well-defined groups.
- Investigation of the dependence among variables.
- Prediction
- Hypothesis testing \longrightarrow validate substantive theory.



Topics Covered

● Outline

Objectives

Topics

● **Topics Covered**

● A Few uses of Linear

Combinations

● Back to Topics

● More Topics

● Topics that We May Not

Cover

Statistics Review

Graphical Techniques

- Linear Algebra: Useful for summarizing (representing) data and performing manipulations.
- Geometry: Observed cases (“sampling units”, individuals, etc) can be viewed as points in high dimensional spaces and multivariate techniques are designed to study the point clouds.
- Random Sampling: *Typically*, we’ll assume that we have sampled cases from a multivariate normal (Gaussian) distribution... or statistics follow a multivariate normal.
- Linear Transformations: Very important!
When you have multiple observed measures on individuals within a sample and you want a linear combination(s) of the measures to have certain properties.



A Few uses of Linear Combinations

● Outline

Objectives

Topics

● Topics Covered

● A Few uses of Linear Combinations

● Back to Topics

● More Topics

● Topics that We May Not Cover

Statistics Review

Graphical Techniques

- Simple example:
 - ◆ $X_1 =$ baseline measure
 - ◆ $X_2 =$ post intervention
 - ◆ Interest in change: $D = X_2 - X_1$, what is μ_D ? and σ_D^2 ?
- More complex: 10 variables measured on individuals who are classified as “normal” and “abnormal” where the goal is to create a linear combination of the 10 measures such that those from different groups are as different as possible on this composite.
- You have 10 variables measured on individuals and you want to combine them into a small number of composite variables that represent most of the information in the data.



Back to Topics

● Outline

Objectives

Topics

● Topics Covered

● A Few uses of Linear
Combinations

● **Back to Topics**

● More Topics

● Topics that We May Not
Cover

Statistics Review

Graphical Techniques

- Multivariate significance tests for inference about
 - ◆ Means and confidence regions for them.
 - ◆ Comparisons of means across samples from different populations or groups.
 - ◆ Variances, covariances (including between sets of variables).
- Multivariate Analysis of Variance (MANOVA):
 - ◆ Extension of multivariate significance tests to more complex experimental designs.
 - ◆ Extension of univariate ANOVA to multiple dependent variables.



More Topics

● Outline

Objectives

Topics

● Topics Covered

● A Few uses of Linear
Combinations

● Back to Topics

● **More Topics**

● Topics that We May Not
Cover

Statistics Review

Graphical Techniques

- Principal Components Analysis. A data reduction method that focuses on studying the relationship between a set of variables by creating linear combinations of the variables.
 - ◆ The linear combination (composite measure) has the largest possible variance possible.
 - ◆ Best lower dimensional representation of the data.
- Discrimination Analysis. Related to MANOVA & significance tests of means. Define a linear combination of variables that maximally accounts for differences between groups.
- Classification Analysis. Allocation of individuals into groups (populations) defined on the basis of observations on multiple variables.
- Canonical Correlation Analysis. Study the relationship between two sets or “batteries” of variables.
- Factor Analysis. A latent variable model that hypothesizes that the dependencies between observed values on variables are due to unobserved variables.



Topics that We May Not Cover

● Outline

Objectives

Topics

● Topics Covered

● A Few uses of Linear

Combinations

● Back to Topics

● More Topics

● Topics that We May Not
Cover

Statistics Review

Graphical Techniques

- Structural Equation modeling: factor analysis is a special case.
- Optimal Scaling or Non-linear Multivariate Analysis. This includes correspondence analysis, homogeneity analysis and others.
- Multidimensional Scaling, which is related to optimal scaling.
- Cluster Analysis, which include $K - means$ and hierarchical cluster analysis.
- Multiple (Multivariate) Regression which is multiple regression with more than one dependent variable.



Statistics Review

● Outline

Objectives

Topics

Statistics Review

● **Statistics Review**

● Data as an Array

● Descriptive Statistics

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques

Getting use to the notation we'll use.

■ **Data:** measurements on p variables (attributes, characteristics, etc) for each of n cases (individuals, experimental units, etc).

■ **Notation:**

◆ x_{jk} = measurement of the k th variable on the j th case.

◆ $j = 1, \dots, n$ where n = the number of cases.

◆ $k = 1, \dots, p$ where p = the number of variables.

■ NOTE: notation in older editions (around 4th edition I think) of J&W used the reverse notation.



Data as an Array

Data are arranged as an array or matrix of cases by variables:

	Variable 1	Variable 2	...	Variable k	...	Variable p
case 1	x_{11}	x_{12}	...	x_{1k}	...	x_{1p}
case 2	x_{21}	x_{22}	...	x_{2k}	...	x_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
case j	x_{j1}	x_{j2}	...	x_{jk}	...	x_{jp}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
case n	x_{n1}	x_{n2}	...	x_{nk}	...	x_{np}

This is an $(n \times p)$ rectangular matrix \mathbf{X} .

All matrices will be denoted by capital bold faced letters, but more on this later...

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Descriptive Statistics

Suppose we have n observations on one variable:

$$x_{11}, x_{21}, \dots, x_{n1} \quad (\text{1st column of } \mathbf{X})$$

- Mean (arithmetic average) — measure of central tendency

$$\bar{x}_1 = \frac{1}{n} \sum_{j=1}^n x_{j1}$$

- If the n observations are a sample from a larger population of possible measurements, then \bar{x}_1 is the sample mean.
- In general, sample means

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^n x_{jk} \quad \text{for } k = 1, \dots, p$$

This is the mean of n observations in the k th column of \mathbf{X} .

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● **Descriptive Statistics**

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Variance & Standard Deviation

A sample of n observation on one variable from some population:

$$x_{11}, x_{21}, \dots, x_{n1}$$

- The Sample Variance for the first variable is

$$s_1^2 = \frac{1}{n} \sum_{j=1}^n (x_{j1} - \bar{x}_1)^2$$

For now we'll use the MLE and divide by n .

- In general,

$$s_{kk} \equiv s_k^2 = \frac{1}{n} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2$$

where $k = 1, \dots, p$.

- Sample standard deviation is $\sqrt{s_{kk}}$ for $k = 1, \dots, p$, which is in the same units (scale) as the original measurements.

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Covariance

- A sample of n observation on two variables from some population:

$$\underbrace{\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}}_{1^{st} \text{ case}}, \underbrace{\begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}}_{2^{nd} \text{ case}}, \dots, \underbrace{\begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix}}_{n^{th} \text{ case}} \leftarrow 1^{st} \text{ column of } \mathbf{X}$$

- Sample Covariance

$$s_{12} = \frac{1}{n} \sum_{j=1}^n (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

Average (mean) product of differences (distances) of variables 1 and 2 from their respective means.

- Properties:

$s_{12} > 0 \longrightarrow$ large values tend to occur together.

$s_{12} < 0 \longrightarrow$ larger values tend to occur with small values

$s_{12} = 0 \longrightarrow$ variables are not linearly associated

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Covariance (continued)

In general,

$$s_{ik} = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

- When $i = k$, $s_{ii} = s_i^2 = i^{\text{th}}$ sample variance.
- Covariances are symmetric: $s_{ik} = s_{ki}$
- An array of variances and covariances:

	Variable 1	Variable 2	...	Variable p
Variable 1	s_{11}	s_{12}	...	s_{1p}
Variable 2	s_{12}	s_{22}	...	s_{2p}
⋮	⋮	⋮	⋮	⋮
Variable p	s_{1p}	s_{2p}	...	s_{pp}

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Sample Correlation Coefficients

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2} \sqrt{\frac{1}{n} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}}$$

for $i = 1, \dots, p$ and $k = 1, \dots, p$.

r_{ik} is a standardized version of the sample covariance. To see this, form z -scores:

$$z_{ji} = \frac{x_{ji} - \bar{x}_i}{\sqrt{s_{ii}}} \longrightarrow \begin{cases} \bar{z}_i = 0 \\ s_{ii} = 1 \end{cases}$$
$$z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}} \longrightarrow \begin{cases} \bar{z}_k = 0 \\ s_{kk} = 1 \end{cases}$$

Replace x_{ji} and x_{jk} by z_{ji} and z_{jk} into the formula for r_{ik} and you'll get $r_{ik} = s_{ik}$.

● Outline

Objectives

Topics

Statistics Review

- Statistics Review
- Data as an Array
- Descriptive Statistics
- Variance & Standard Deviation
- Covariance
- Covariance (continued)
- Sample Correlation Coefficients
- Sample Correlation Coefficients (cont.)
- Sum of Squared Deviations
- To Summarize

Graphical Techniques



Sample Correlation Coefficients (cont.)

Properties

■ $-1 \leq r_{ik} \leq 1$ (it's dimensionless)

■ r_{ik} measures the strength of linear relationship

$$r_{ik} = 0 \rightarrow \text{no linear association}$$

$$r_{ik} < 0 \rightarrow \text{negative linear relationship}$$

$$r_{ik} > 0 \rightarrow \text{positive linear relationship}$$

■ r_{ik} is invariant if you change the location (mean) &/or re-scale (change variance),

i.e., If $y_{ji} = ax_{ji} + b$ and $y_{jk} = cx_{jk} + d$, then

$$r_{ik} = \text{corr}(x_{ji}, x_{jk}) = \text{corr}(y_{ji}, y_{jk})$$

... provided that $a > 0$ and $c > 0$ or $a < 0$ and $c < 0$.

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Sum of Squared Deviations

and **sum of cross-product deviations** are used in multivariate, so we'll have a symbol for them.

Define

$$w_{ik} = \sum_{j=1}^n \underbrace{(x_{ji} - \bar{x}_i)}_{\text{deviation of scores from their means}} \underbrace{(x_{jk} - \bar{x}_k)}$$

deviation of scores from their means

$$w_{ii} = \sum_{j=1}^n \underbrace{(x_{ji} - \bar{x}_i)^2}_{\text{sums of squares}}$$

sums of squares

for $i = 1, \dots, p$ and $k = 1, \dots, p$.

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard

Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● **Sum of Squared Deviations**

● To Summarize

Graphical Techniques



To Summarize

Our descriptive statistics can all be organized into arrays (matrices):

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} \quad S_n = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{pmatrix}$$

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{12} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{pp} \end{pmatrix} \quad W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{12} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \cdots & w_{pp} \end{pmatrix}$$

The arrays (matrices) S , W and R are symmetric.

● Outline

Objectives

Topics

Statistics Review

● Statistics Review

● Data as an Array

● Descriptive Statistics

● Variance & Standard Deviation

● Covariance

● Covariance (continued)

● Sample Correlation

Coefficients

● Sample Correlation

Coefficients (cont.)

● Sum of Squared Deviations

● To Summarize

Graphical Techniques



Graphical Techniques

Important and useful... **Always Look at your Data!**

One variables: 1–dimensional plot

- dot diagram (see J&W)
- stem-n-left
- histogram
- box-plot

e.g., National Parks (table 1.11 in J&W):

Size of Park:

Stem Leaf	#	Boxplot
2 2	1	
1 5	1	
1 2	1	
0 589	3	+---+---+
0 000012233	9	*-----*
-----+-----+-----+-----+		

Multiply Stem.Leaf by $10^{**}+3$

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

● Graphical Techniques

● Histograms of National Parks

● Two Variables (relationship between)

● Two Numerical and One discrete

● New Data Set

● Covariances and Correlations

● All Bivariate and Univariate

● Three Numerical

● Three Numerical & One

Discrete

● Two Types of Scatter Plots

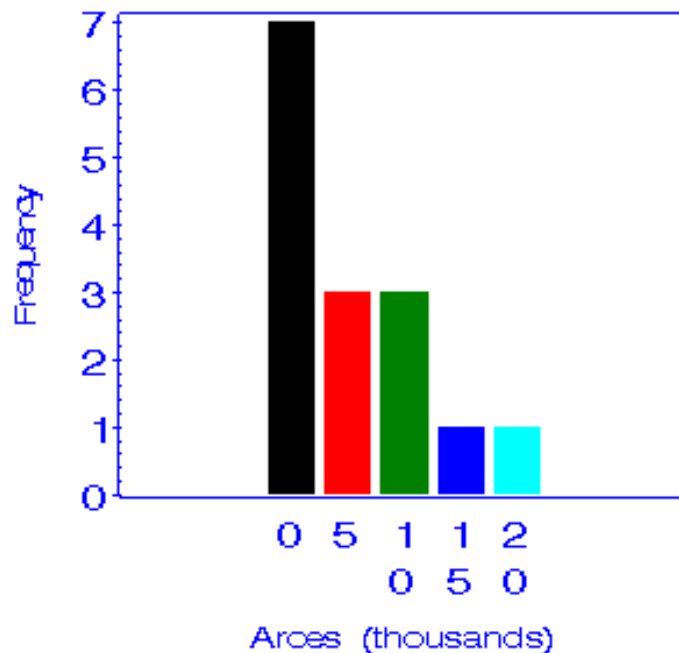
● Other Graphical Displays



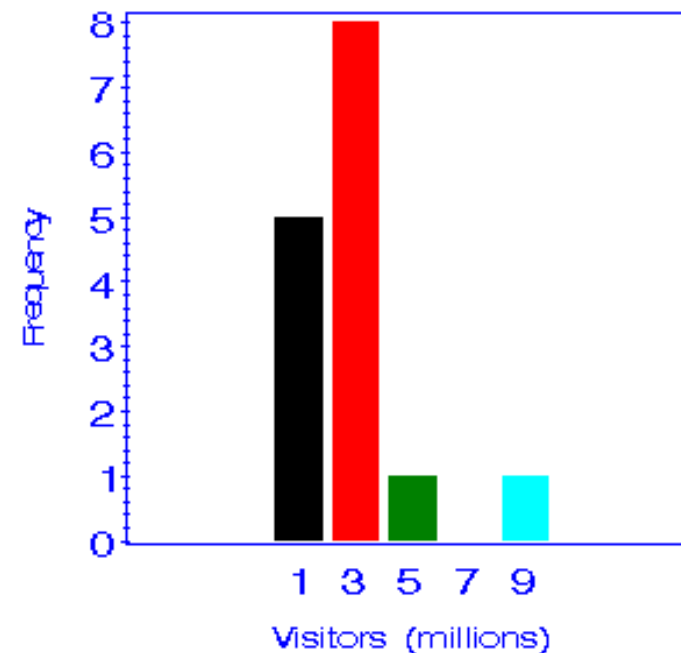
Histograms of National Parks

- Outline
- Objectives
- Topics
- Statistics Review
- Graphical Techniques
 - Graphical Techniques
 - Histograms of National Parks
 - Two Variables (relationship between)
 - Two Numerical and One discrete
 - New Data Set
 - Covariances and Correlations
 - All Bivariate and Univariate
 - Three Numerical
 - Three Numerical & One Discrete
 - Two Types of Scatter Plots
 - Other Graphical Displays

Distribution of Park Sizes



Distribution of Number of Visitors



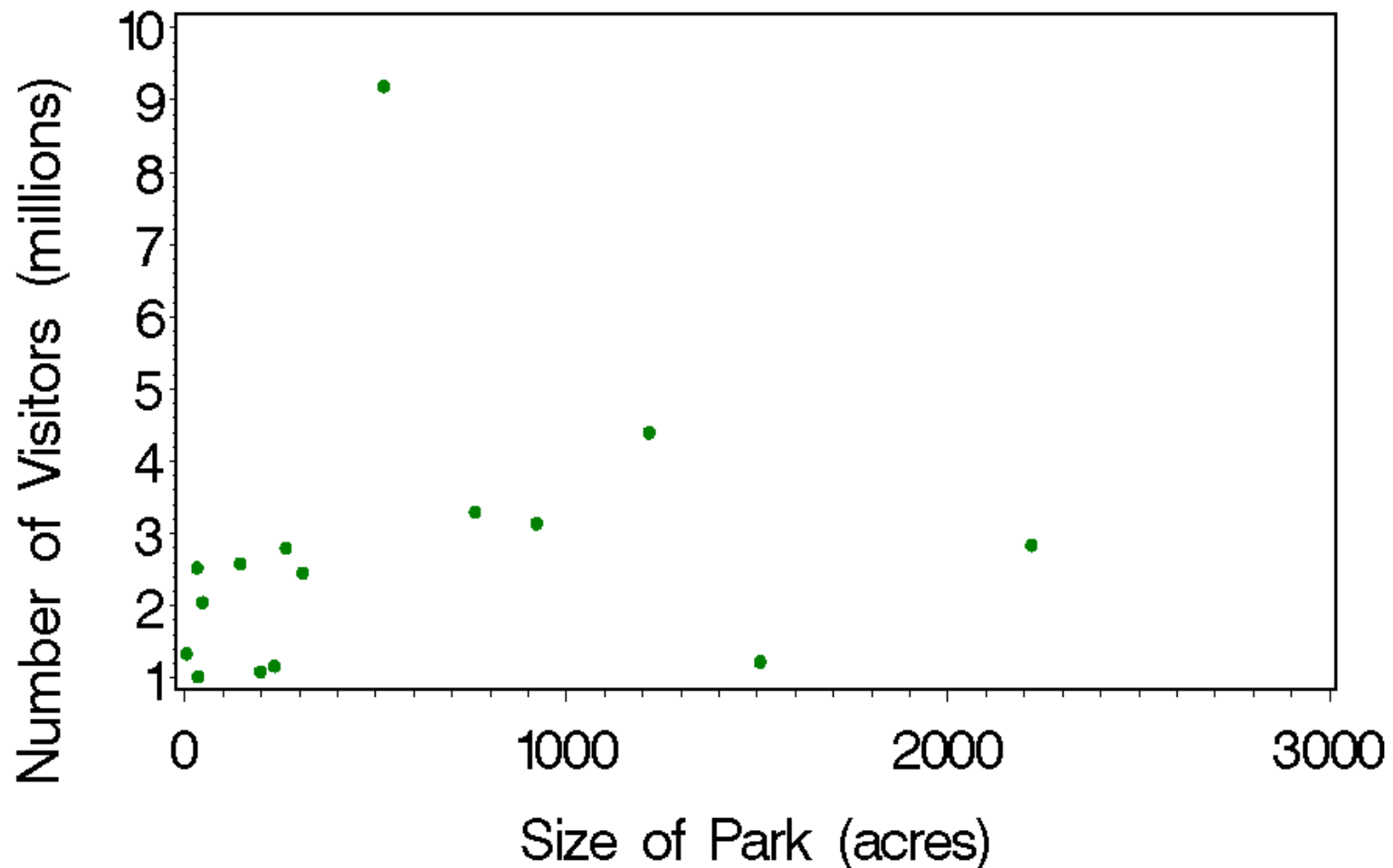


Two Variables (relationship between)

Scatter plots: look for patterns of association. e.g,

Attendance and Size of National Parks

Table 1.11 from Johson & Wichern



● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

● Graphical Techniques

● Histograms of National Parks

● Two Variables (relationship between)

● Two Numerical and One discrete

● New Data Set

● Covariances and Correlations

● All Bivariate and Univariate

● Three Numerical

● Three Numerical & One Discrete

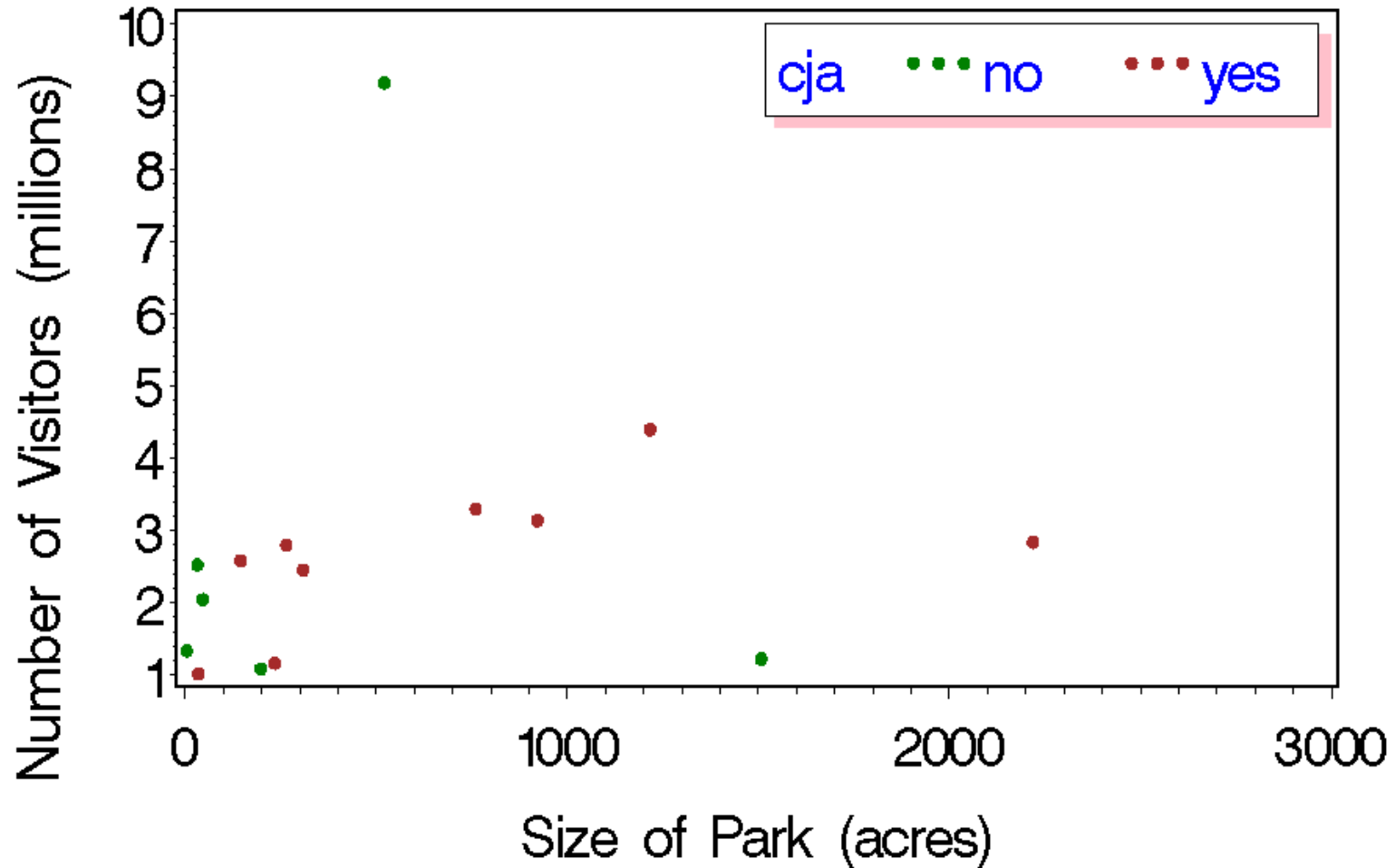
● Two Types of Scatter Plots

● Other Graphical Displays



Two Numerical and One discrete

National Parks that I've visited
Table 1.11 from Johson & Wichern





New Data Set

From Rencher (2002) who got it from Beall (1945)

■ Description: 32 males and 32 females had measures on four psychological tests.

■ The tests were

- ◆ x_1 = pictorial inconsistencies
- ◆ x_2 = paper form board
- ◆ x_3 = tool recognition
- ◆ x_4 = vocabulary

■ Simple descriptive statistics:

Variable	n	\bar{x}	s	Minimum	Maximum
Test1	64	14.16	3.22	2.00	20.00
Test2	64	14.91	4.08	5.00	21.00
Test3	64	21.92	7.55	6.00	34.00
Test4	64	22.34	4.70	9.00	29.00

■ What do we learn from this?

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete

● **New Data Set**

- Covariances and Correlations
- All Bivariate and Univariate
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays



Covariances and Correlations

What do we learn from these? (What don't we learn?)

Covariances

	Test1	Test2	Test3	Test4
Test1	10.39	7.79	15.30	5.37
Test2	7.79	16.66	13.71	6.18
Test3	15.30	13.71	57.06	15.93
Test4	5.37	6.18	15.93	22.13

Correlations:

	Test1	Test2	Test3	Test4
Test1	1.00	.59	.63	.35
Test2	.59	1.00	.44	.32
Test3	.63	.44	1.00	.45
Test4	.35	.32	.45	1.00

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- **Covariances and Correlations**
- All Bivariate and Univariate
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays



All Bivariate and Univariate

● Outline

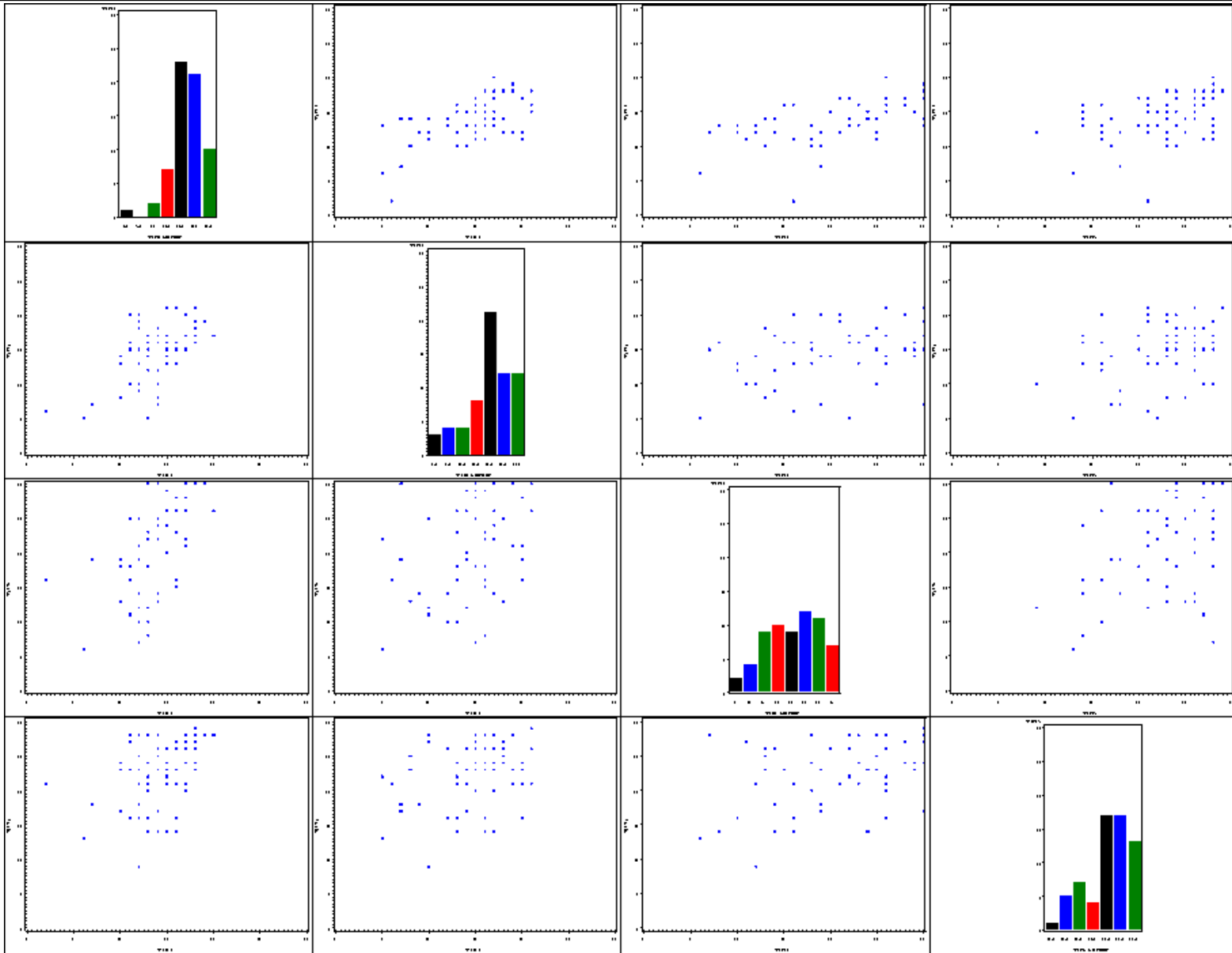
Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- Covariances and Correlations
- **All Bivariate and Univariate**
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays

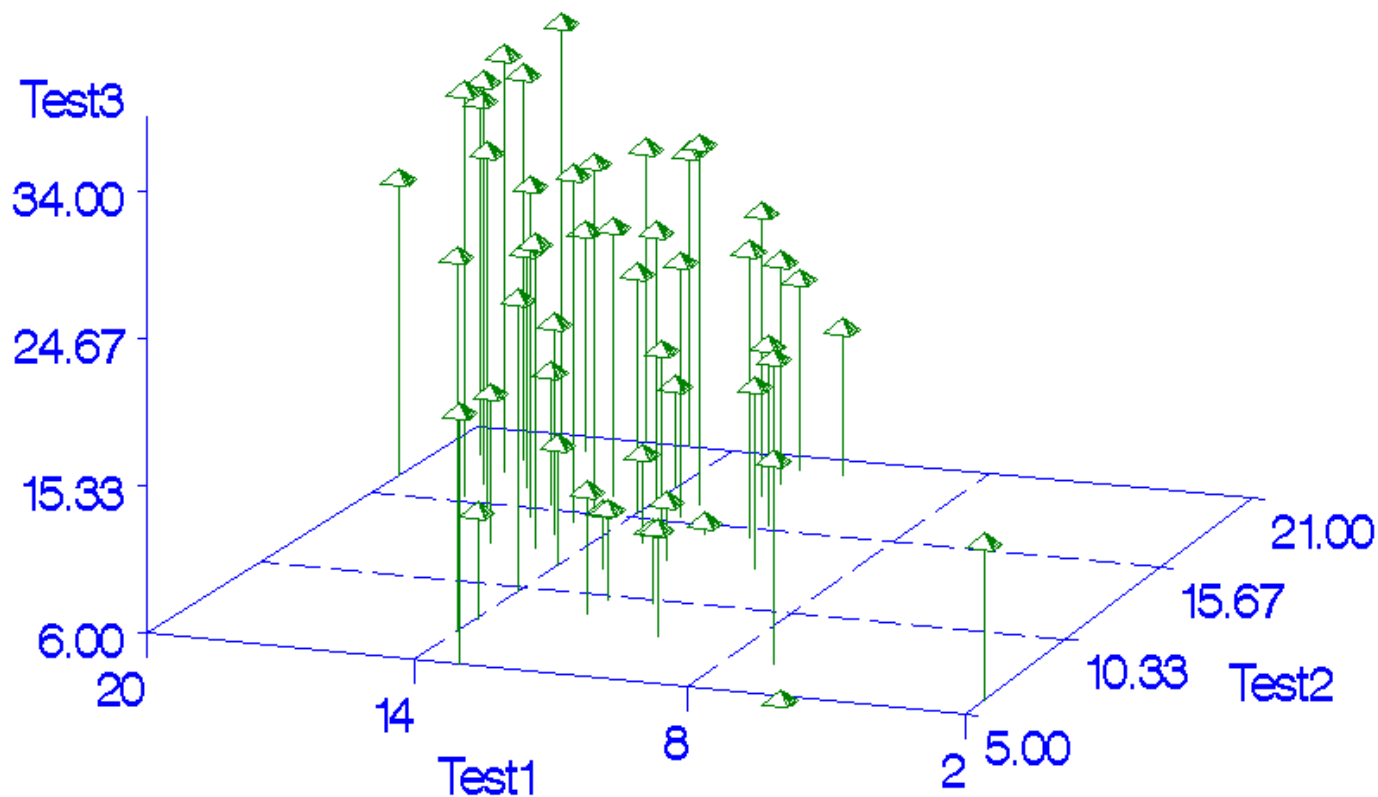




Three Numerical

Could do a three dimensional graph:

Three Psychological Test Scores



Source: Rencher (2002)

Original source: Beall (1945)

● Outline

Objectives

Topics

Statistics Review

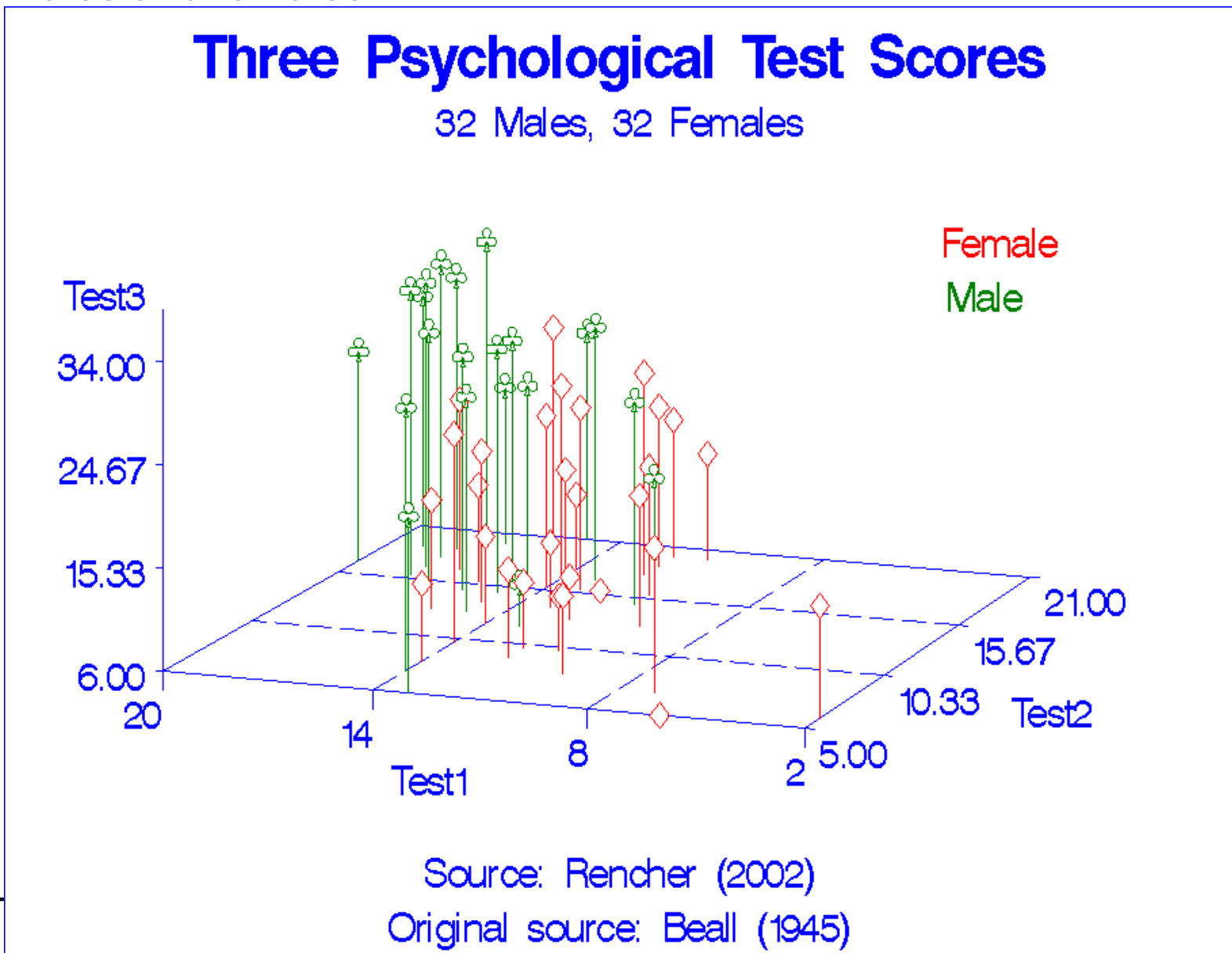
Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- Covariances and Correlations
- All Bivariate and Univariate
- **Three Numerical**
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays



Three Numerical & One Discrete

Could do a three dimensional graph and identify points for males and females:



● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- Covariances and Correlations
- All Bivariate and Univariate
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays



Two Types of Scatter Plots

- Variable Space: n points in p -dimensional space. Each row (individual, case) of X is a distinct point,

$$(x_{j1}, x_{j2}, \dots, x_{jp})$$

- ◆ We looked at all pairwise for the four psychological tests.
- ◆ We looked at 3-way plot.
- ◆ If we could view p -dimensional space, we maybe able to detect patterns, clusters, similarities and/or differences between cases.

- Subject Space or Observation Space: View the data as p points in n -dimensional space. Each column is a distinct point,

$$(x_{1k}, x_{2k}, \dots, x_{nk})$$

A case (individual) defines an axis (more on these later).

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- Covariances and Correlations
- All Bivariate and Univariate
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays



Other Graphical Displays

● Outline

Objectives

Topics

Statistics Review

Graphical Techniques

- Graphical Techniques
- Histograms of National Parks
- Two Variables (relationship between)
- Two Numerical and One discrete
- New Data Set
- Covariances and Correlations
- All Bivariate and Univariate
- Three Numerical
- Three Numerical & One Discrete
- Two Types of Scatter Plots
- Other Graphical Displays

“What a neat graph!” versus “What an interesting story!”

- Linking multiple 2-dimensional scatter plots (“brushing”, highlighting particular points, size of point convey frequency, others).
- Stars for $p \geq 2$ dimensional graphical displays
- Chernoff faces
- Interactive real-time graphical software ([SAS demo](#))
- Be creative (e.g, [Knowing the World powerpoint](#))