STT 843: Multivariate Analysis

1. Introduction (Chapters 1 & 2)

Guanqun Cao

Department of Statistics and Probability
Michigan State University

Spring 2026

Outline

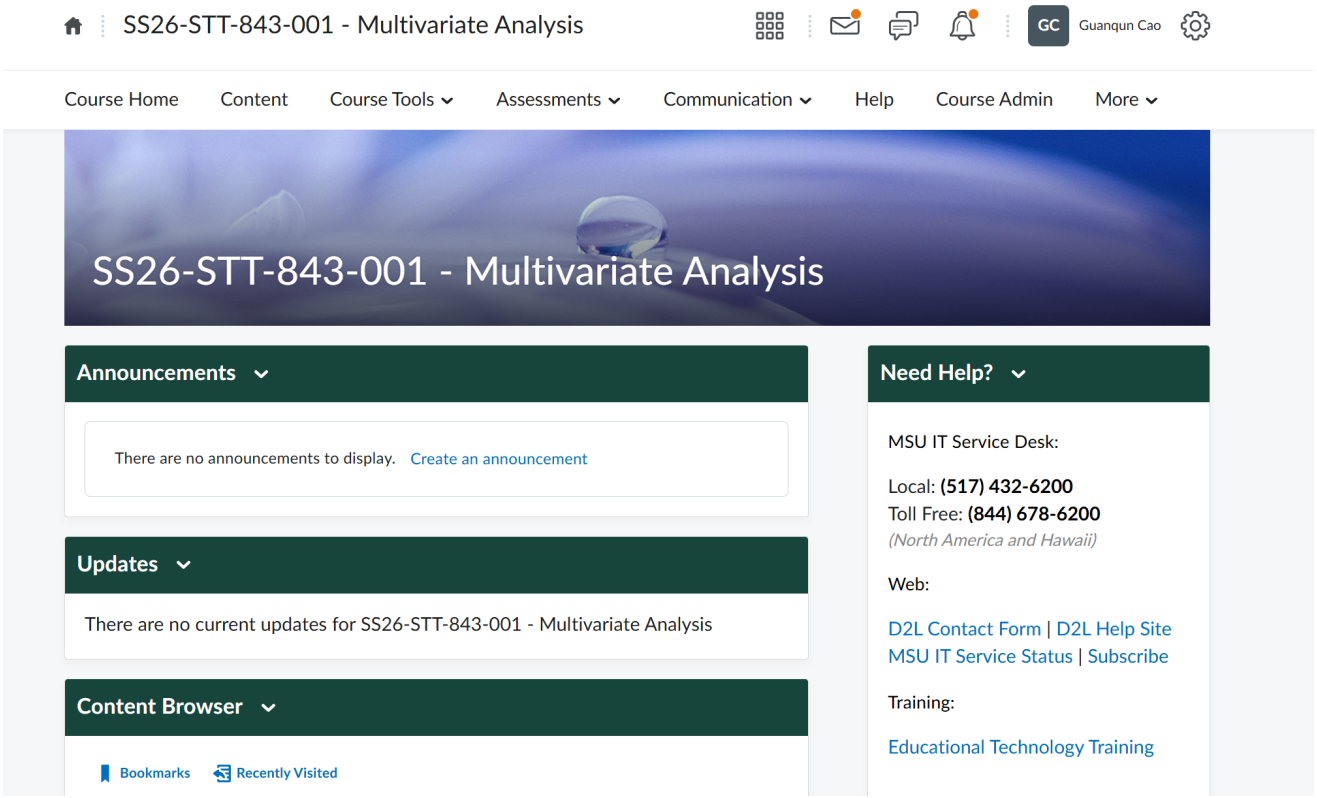
What is this course about?

Topics:

- Multivariate Normal Distribution
- Inferences a Mean Vector
- Multivariate Linear Regression Models
- Principal Components
- Factor Analysis
- Classification and Clustering

D2L and where to find grades

https://d2l.msu.edu/d2l/home/2530630



Course Website and where to find slides and RMarkdown

https://caoguanqun.github.io/STT843_multivariate-data-analysis/

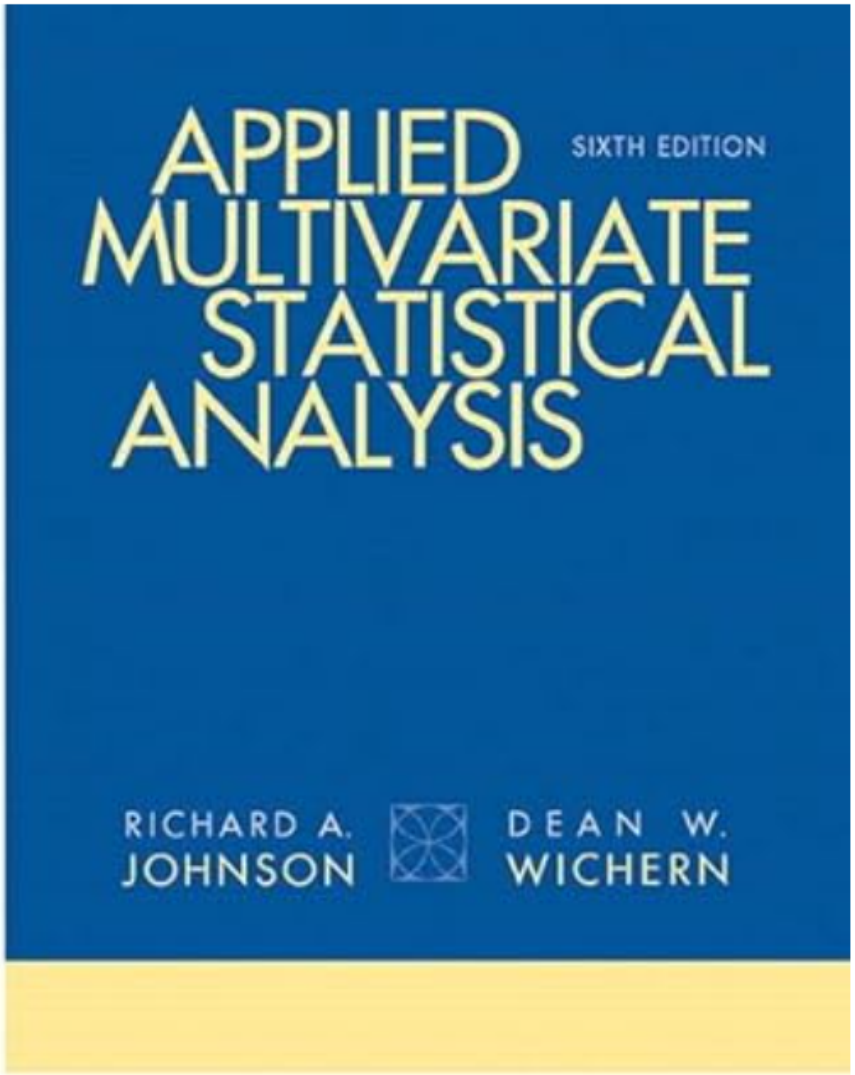
Note the syllabus link above!

Dr. Cao's Office hours

Time: M 10:00am-11:00am

C426 Wells Hall

Textbook


APPLIED
SIXTH EDITION
MULTIVARIATE
STATISTICAL
ANALYSIS
RICHARD A.
JOHNSON
DEAN W.
WICHERN
Pearson Modern Classic

Class Structure

- Class is a combination of lecture time, and group work/coding time.
  - Bring computer every day
  - R Studio
- Once per three weeks, there will be a short check-in quiz. This will be basic content related to lectures since the last class. Possible questions include checking on definitions, or basic understanding of major ideas.
  - 10 points per quiz
  - Drop one lowest grade

Class Structure Pt 2

- Homework due once per two weeks, midnight of the day marked in the schedule (mostly Sundays).
  - Drop one lowest grade
  - Sliding scale:
    - 24 hours late: 5% penalty.
    - 48 hours late: 15% penalty.
    - >48 hours: No late work accepted.
- One in-class exam
  - See schedule for date
- One Project
  - Analyze dataset using tools in class, submit written report and do oral presentation in class
  - Due at the end of the semester

Grade distribution

Final Grades will be based on

- Mid-term Exam: 40%
- Quizzes: 20% (drop one lowest grade)
- Homework assignments: 20% (drop one lowest grade)
- Project: 20%

Course Goals

Upon successful completion of this course, students should be able to:

1. Identify the need for multivariate statistical techniques
2. Recognize the appropriate multivariate method for a problem
3. Employ statistical software to conduct the appropriate analysis
4. Interpret results and make statistical inference

What is multivariate analysis?

- 1st course of statistics: numbers-random variables
- 2nd course of statistics: vectors of numbers-random vectors
  - Basis for analysis of more complex objects, e.g. functions, matrices, tensors, images, networks.
- **Data Exploration:** visualization of relationships between observations.
- **Discovering and modeling patterns** from dataset: Visualization, Clustering, Multivariate distributions.
- **Confirming patterns:** Inference.
- **Dimension Reduction:** PCA, CCA, SVD.
- **Predictions:** Regression, Classification.

Multivariate methods

- Data reduction or structural simplification
- Sorting and grouping (e.g., classification, clustering)
- Investigation of the dependence among variables.
- Prediction
- Hypothesis construction and testing

Aims of of multivariate data analysis

How can we visualize the data?

- What is the joint distribution of marks?
- Can we simplify the data? For example, we rank football teams using $3W + D$ and we rank students by their average module mark. Is this fair?
- Can we reduce the dimension in a better way?
- Can we use the data to discriminate, for example, between male and female students?
- Are the different iris species different shapes?
- Can we build a model to predict the intended digit from an image of someones handwriting? Or predict the species of iris from measurements of its sepal and petal?

Multivariate data example

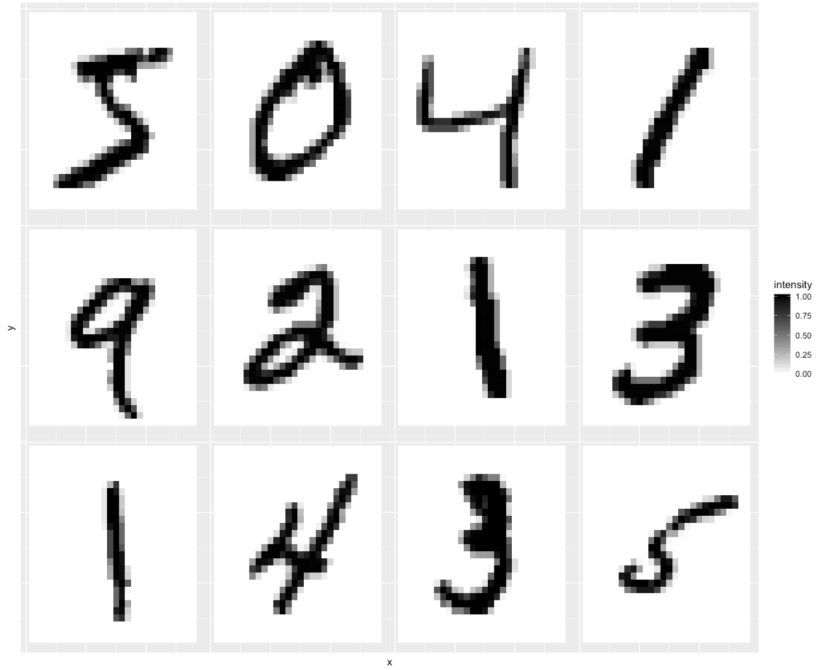
Table: Costs of Living in Each of the 50 States

State	Median rent ($)	Median home value($1000)	Cost of living index	Population (in 1000s)	Average gross income ($1000)
AK	949	237.8	133.2	698.47	68.60
AL	631	121.5	93.3	4708.71	36.11
AR	606	105.7	90.4	2889.45	34.03
...	...	...	...	...	...
WV	528	95.9	95.0	1819.78	33.88
WY	636	188.2	99.6	544.27	64.88

Source: US Census, 2007 and 2009 data

MNIST Dataset

The MNIST dataset is a collection of handwritten digits that is widely used in statistics and machine learning to test algorithms. It contains 60,000 images of hand-written digits. Here are the first 12 images:



What is a multivariate dataset?

Multivariate statistical analysis concerns multivariate data where each observation consisting of many measurements on the same subject.

We suppose the dataset $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ has n observations (Here, n is called the sample size), and each observation $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})$ is a vector in $\mathbb{R}^p$ (Here, p is called the dimension). These are often recorded in a $n \times p$ matrix:

$$\mathbf{X} = (\mathbf{X}_1^{\mathsf{T}}, \cdots, \mathbf{X}_n^{\mathsf{T}}) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Overview: Why Matrix Algebra?

- Univariate statistics is concerned with a random scalar variable Y.
- In multivariate analysis, we are concerned with the joint analysis of multiple dependent variables. These variables can be represented using matrices and vectors. This provides simplification of notation and a format for expressing important formulas.

Example 0-1:

- Suppose that we measure the variables $x_1 =$ height (cm), $x_2 =$ left forearm length (cm) and $x_3 =$ left foot length for participants in a study of the physical characteristics of adult humans.
- These three variables can be represented in the following column vector:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

- The observed data for a specific individual, say the i^{th} individual, might also be represented in an analogous vector. Suppose that the i^{th} person in the sample has height $= 175\ \mathrm{cm}$, forearm length $= 25.5\ \mathrm{cm}$ and foot length $= 27\ \mathrm{cm}$.
- In vector notation these observed data could be written as:

$$\mathbf{x_i} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} = \begin{pmatrix} 175 \\ 25.5 \\ 27.0 \end{pmatrix}$$

- Notice the use and placement of the subscript i to represent the i^{th} individual.

Review on Matrix

- **Matrix**: A matrix is a two-dimensional array of numbers of formulas
- **Vector**: A vector is a matrix with either only one column or only one row
- **Column vector**: A column vector contains only one column
- **Row vector**: A row vector contains only one row
- **Dimension of a Matrix**: A dimension of a matrix is expressed as the number of rows $\times$ the number of columns matrix with 10 rows and 3 columns is said to be 10×3;
- **Square Matrix**: A square matrix numbers of rows and columns are the same (i.e., a 4×4 matrix is a square matrix)

The Data Matrix in Multivariate Problems

- Usually, the observed data are represented by a matrix in which the rows are observations and the columns are variables.
- The usual notation is n = the number of observed units (people, animals, companies, etc.) and p = the number of variables measured on each unit. Thus the data matrix will be an $n \times p$ matrix.

Example 0-2:

Suppose that we have scores for $n = 6$ college students who have taken the verbal and the science subtests of the College Qualification Test (CQT). We have $p = 2$ variables: (1) the verbal score and (2) the science score for each student. The data matrix is the following 6×2 matrix:

$$\mathbf{X} = \begin{pmatrix} 41 & 26 \\ 39 & 26 \\ 53 & 21 \\ 67 & 33 \\ 61 & 27 \\ 67 & 29 \end{pmatrix}$$

Notation notes

- In the matrix just given, the first column gives the data for $x_1 =$ verbal score whereas the second column gives data for $x_2 =$ science score. Each row gives data for a student in the sample. To repeat - the rows are observations, the columns are variables.
- we have used $\mathbf{x}$ to denote the vector of variables in Example 1 and $\mathbf{X}$ to represent the data matrix in Example 2.
- In matrix terms, the i^{th} row in the data matrix $\mathbf{X}$ is the transpose of the data vector

$$\mathbf{x_i} = \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}, \text{ as we defined data vectors in Example 2}$$

Measures of Central Tendency, Dispersion and Association

Overview

Three aspects of the data are of importance, the first two of which you should already be familiar with from **univariate** statistics.

1. *Central Tendency*: What is a typical value for each variable?
2. *Dispersion*: How far apart are the individual observations from a central value for a given variable?
3. *Association*: When more than one variable are studied together, how does each variable relate to the remaining variables? How are the variables simultaneously related to one another? Are they positively or negatively related?

Definitions: Population

- Statistics, as a subject matter, is the science and art of using sample information to make generalizations about populations.
- **Population**: the collection of all people, plants, animals, or objects of interest about which we wish to make statistical inferences.
- The population may also be viewed as the collection of all possible random draws from a stochastic model; for example, independent draws from a normal distribution with a given population mean and population variance.

Definitions

- **Population parameter**: an (unknown) numerical characteristic of a population; We use sample data to make an inference about the value of a parameter.
- **Sample**
  A sample is the subset of the population that we actually measure or observe.
- **Sample Statistic**
  A sample statistic is a numerical characteristic of a sample. A sample statistic estimates the unknown value of a population parameter.

Notations

X_{ij} = Observation for variable j in subject i.
p = Number of variables
n = Number of subjects

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}$$

Objectives

Upon successful completion of this lesson, you should be able to:

- interpret measures of central tendency, dispersion, and association;
- calculate sample means, variances, covariances, and correlations using a hand calculator;
- use software R to compute sample means, variances, covariances, and correlations.

Measures of Central Tendency

- μ_1: the population mean for variable X_1
- $\bar{x}_1$: a sample mean based on observed data for variable X_1.
- $\mu_j = E(X_j)$ can be estimated by the sample mean $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$
- $\bar{x}_j$ is unbiased for μ_j

- We can estimate this population mean vector, $\boldsymbol{\mu}$, by $\bar{\mathbf{x}}$. This is obtained by collecting the sample means from each of the variables in a single vector.
- $$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} X_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{ip} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

Just as the sample means, $\bar{x}$, for the individual variables are unbiased for their respective population means, the sample mean vector is unbiased for the population mean vector.

$$E(\bar{\mathbf{x}}) = E\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} E(\bar{x}_1) \\ E(\bar{x}_2) \\ \vdots \\ E(\bar{x}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

Dispersion: Variance, Standard Deviation

- **Variance** measures the degree of spread (dispersion) in a variable's values.
  Theoretically, a population variance is the average squared difference between a variable's values and the mean for that variable. The population variance for variable X_j is
- **Population Variance** The population variance for variable X_j is

$$\sigma_j^2 = E\left(X_j - \mu_j\right)^2$$

Sample Variance

- The *population variance* σ_j^2 can be estimated by the *sample variance*

$$
\begin{aligned}
s_j^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{x}_j)^2 \\
&= \frac{\sum_{i=1}^{n} X_{ij}^2 - n\bar{x}_j^2}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_{ij}^2 - \left(\left(\sum_{i=1}^{n} X_{ij}\right)^2/n\right)}{n-1}
\end{aligned}
$$

- The sample variance s_j^2 is unbiased for the population variance σ_j^2. $E\left(s_j^2\right) = \sigma_j^2$.

Our textbook (Johnson and Wichern, 6th ed.) uses a sample variance formula derived using maximum likelihood estimation principles. In this formula, the division is by n rather than $n-1$.

$$s_j^2 = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{x}_j)^2}{n}$$

Measures of Association: Covariance, Correlation

- **Population Covariance**: a measure of the association between pairs of variables in a population.
- The population covariance between variables j and k is

$$\sigma_{jk} = E\left\{\left(X_{ij} - \mu_j\right)\left(X_{ik} - \mu_k\right)\right\} \qquad \text{for } i = 1, \ldots, n$$

- Positive population covariances mean that the two variables are positively associated; variable j tends to increase with increasing values of variable k.
- A negative association can also occur. Variable j will tend to decrease with increasing values of variable k.

- Sample Covariance

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{x}_j)(X_{ik} - \bar{x}_k)$$
$$= \frac{\sum_{i=1}^{n} X_{ij}X_{ik} - \left(\sum_{i=1}^{n} X_{ij}\right)\left(\sum_{i=1}^{n} X_{ik}\right)/n}{n-1}$$

- A positive covariance would indicate a positive association between the variables j and k. And a negative association is when the covariance is negative.

- The sample covariance s_{jk} is unbiased for the population covariance σ_{jk}.

$$E(s_{jk}) = \sigma_{jk}$$

- $s_{jk} = 0$: the two variables are uncorrelated. (Note that this does not necessarily imply independence, we'll get back to this later.)
- $s_{jk} > 0$: the larger the covariance, the stronger the positive association between the two variables.
- $s_{jk} < 0$: the smaller the covariance, the stronger the negative association between the two variables.

Population variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Note that the population variances appear along the diagonal of this matrix, and the covariance appear in the off-diagonal elements. So, the covariance between variables j and k will appear in row j and column k of this matrix.

Sample variance-covariance matrix

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

Note that the sample variances appear along diagonal of this matrix and the covariances appear in the off-diagonal elements. So the covariance between variables j and k will appear in the jk-th element of this matrix.

Notes

- **S** is symmetric; i.e., $s_{jk} = s_{kj}$.
- **S** is unbiased for the population variance covariance matrix Σ,

$$E(S) = \begin{pmatrix} E\left(s_1^2\right) & E\left(s_{12}\right) & \dots & E\left(s_{1p}\right) \\ E\left(s_{21}\right) & E\left(s_2^2\right) & \dots & E\left(s_{2p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(s_{p1}\right) & E\left(s_{p2}\right) & \dots & E\left(s_p^2\right) \end{pmatrix} = \Sigma$$

In matrix notation, the sample variance-covariance matrix may be computed used the following expressions:

$$
\begin{aligned}
S &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \bar{x}\right) \left(\boldsymbol{X}_i - \bar{x}\right)^{\mathsf{T}} \\
&= \frac{\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathsf{T}} - \left(\sum_{i=1}^{n} \boldsymbol{X}_i\right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i\right)^{\mathsf{T}} / n}{n-1}
\end{aligned}
$$

Notes

- The magnitude of the covariance value is not particularly helpful as it is a function of the magnitudes (scales) of the two variables
- This quantity is a function of the variability of the two variables, and so, it is hard to tease out the effects of the association between the two variables from the effects of their dispersions
- $-s_i s_j \leq s_{ij} \leq s_i s_j$

Correlation Matrix

Population correlation is defined to be equal to the population covariance divided by the product of the population standard deviations:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

The population correlation may be estimated by substituting into the formula the sample covariances and standard deviations:

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n} X_{ij} X_{ik} - \left(\sum_{i=1}^{n} X_{ij}\right)\left(\sum_{i=1}^{n} X_{ik}\right)/n}{\sqrt{\left\{\sum_{i=1}^{n} X_{ij}^2 - \left(\sum_{i=1}^{n} X_{ij}\right)^2/n\right\}\left\{\sum_{i=1}^{n} X_{ik}^2 - \left(\sum_{i=1}^{n} X_{ik}\right)^2/n\right\}}}$$

It is essential to note that the population and the sample correlation must lie between -1 and 1 .

$$-1 \leq \rho_{jk} \leq 1$$
$$-1 \leq r_{jk} \leq 1$$

Therefore:

- $\rho_{jk} = 0$ indicates, as you might expect, the two variables are uncorrelated.
- ρ_{jk} close to $+1$ will indicate a strong positive dependence
- ρ_{jk} close to -1 indicates a strong negative dependence

Sample Correlation Matrix

The sample correlation matrix is denoted as $\mathbf{R}$.

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

Notes

- Overall, we see moderately strong linear associations among the variables height, left arm, and left foot and relatively weak (almost 0) associations between head circumference and the other three variables.

- In practice, use scatter plots of the variables to understand the associations between variables fully. It is not a good idea to rely on correlations without seeing the plots. Correlation values are affected by outliers and curvilinearity.

Overall Measures of Dispersion

- The variance σ_j^2 measures the dispersion of an individual variable X_j. Total Variation is used to measure the dispersion of all variables together.

- To understand total variation we first must find the trace of a square matrix. A square matrix is a matrix that has an equal number of columns and rows. Important examples of square matrices include the variance-covariance and correlation matrices.

Total variation

- The trace of an $n \times n$ matrix $\mathbf{A}$ is $\text{trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$
- The total variation, therefore, of a random vector $\mathbf{X}$ is simply the trace of the population variance-covariance matrix.

$$\text{trace}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \ldots \sigma_p^2$$

- Thus, the total variation is equal to the sum of the population variances.

The total variation can be estimated by:

$$\text{trace}(S) = s_1^2 + s_2^2 + \cdots + s_p^2$$

The total variation is of interest for principal components analysis and factor analysis and we will look at these concepts later in this course.

Summary

In this section we learned how to:

- interpret various measures of central tendency, dispersion, and association;
- compute sample means, variances, covariances, and correlations using a hand calculator;
- use software to compute sample means, variances, covariances, and correlations.