1

Introduction and Overview

CHAPTER CONTENTS

△ Background

Multivariate statistics can be described as containing two distinct methods: dependent and interdependent. Dependent methods designate certain variables as dependent measures with the others treated as independent variables. Multivariate dependent methods are associated with regression, analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), discriminant, and canonical analyses. Multivariate interdependent methods are associated with factor, cluster, and multidimensional scaling analyses where no dependent variable is designated. Interdependent methods search for underlying patterns of relations among the variables of interest. Another characterization is to study multivariate statistics as two distinct approaches, one that tests for mean differences and another that analyzes correlation/covariance among variables. This book will present these two types of multivariate methods using R functions.

△ Persons of Interest

The book takes a unique perspective in learning multivariate statistics by presenting information about the individuals who developed the statistics, their background, and how they influenced the field. These biographies about the past noteworthy persons in the field of statistics should help you understand how they were solving real problems in their day. The introduction of each chapter therefore provides a brief biography of a person or persons who either developed the multivariate statistic or played a major role in its use.

Δ Factors Affecting Statistics

An important concept in the field of statistics is data variability. Dating back to 1894, Sir Ronald Fisher and Karl Pearson both understood the role data variance played in statistics. Sir Ronald Fisher in conducting experimental designs in the field of agriculture knew that mean differences would be a fair test if the experimental and control groups had approximately equal variances. Karl Pearson in developing his correlation coefficient when studying heredity variables employed bivariate covariance with each variable variance to compute a measure of association. The amount of covariance indicated whether two variables were associated. In both cases, the amount of variance indicated individual differences. For example, if a dependent variable, plant growth, did not vary, then no individual difference existed. If the height of males and females do not covary, then there is no association.

It is a basic fact that we are interested in studying why variation occurs. For example, if test scores were all the same, hence the standard deviation or variance is zero, then we know that all students had the same test score—no variance; that is, no student difference. However, when test scores do vary, we wish to investigate why the test scores varied. We might investigate gender differences in mean test scores to discover that boys on average scored higher than girls. We might correlate hours spent studying with test scores to determine if test scores were higher given that a student spent more time studying—a relationship exists.

We should also understand situations, when studying variance, where the use of inferential statistics is not appropriate. For example,

- Sample size is small ($n < 30$)
- $N = 1$ (astronomer studies only one planet)
- Nonrandom sampling (convenience, systematic, cluster, nonprobability)
- Guessing is just as good (gambling)
- Entire population is measured (census)
- Exact probabilities are known (finite vs. infinite population size)
- Qualitative data (nonnumeric)

- Law (no need to estimate or predict)
- No inference being made from sample statistic to population parameter (descriptive)

When using statistics, certain assumptions should be met to provide for a fair test of mean differences or correlation. When the statistical assumptions are not met, we consider the statistical results to be biased or inaccurate. There are several factors that can affect the computation and interpretation of an inferential statistic (Schumacker & Tomek, 2013). Some of them are listed here:

- Restriction of range
- Missing data
- Outliers
- Nonnormality
- Nonlinearity
- Equal variance
- Equal covariance
- Suppressor variables
- Correction for attenuation
- Nonpositive definite matrices
- Sample size, power, effect size

A few heuristic data sets in Table 1.1 show the effect certain factors have on the Pearson correlation coefficient. The complete data set indicates that Pearson $r = .782$, $p = .007$, which would be used to make an inference about the population parameter, rho.

However, if missing data are present, Pearson $r = .659$, $p = .108$, a nonsignificant finding, so no inference would be made. More important, if listwise deletion was used, more subject data might not be used, or if pairwise deletion was used, then different sample sizes would be used for each bivariate correlation. We generally desire neither of these choices when conducting statistical tests. The nature of an outlier (extreme data value) can also cause inaccurate results. For data set A ($Y = 27$ outlier), Pearson $r = .524$, $p = .37$, a nonsignificant finding, whereas for data set B with no outlier, Pearson $r = -.994$, $p = .001$. These data have two very different outcomes based on a single outlier data value. The range of data also can affect correlation, sometimes referred to as restriction of range (thus limiting variability). In the data set, Y ranges from 3 to 7 and X ranges from 1 to 4, with Pearson $r = 0.0$, $p = 1.0$. These values could easily have been taken from a Likert scale on a questionnaire. A small sampling effect combined with

Table 1.1 Factors Affecting Pearson Correlation

Complete Data		Missing Data		Outlier Data Sets			
				Set A		Set B	
Y	***X***	***Y***	***X***	***X***	***Y***	***X***	***Y***
8.00	6.00	8.00	—	1	9	1	9
7.00	5.00	7.00	5.00	2	7	2	7
8.00	4.00	8.00	—	3	5	3	5
5.00	2.00	5.00	2.00	4	3	4	3
4.00	3.00	4.00	3.00	5	27	5	2
5.00	2.00	5.00	2.00				
3.00	3.00	3.00	3.00				
5.00	4.00	5.00	—				
3.00	1.00	3.00	1.00				
2.00	2.00	2.00	2.00				
$r = .782$, $p = .007$		$r = .659$, $p = .108$		$r = .524$, $p = .37$		$r = -.994$, $p = .001$	

Range of Data		Sampling Effect		Nonlinear Data	
Y	***X***	***Y***	***X***	***Y***	***X***
3.00	1.00	8.00	3.00	1.00	1.00
3.00	2.00	9.00	2.00	2.00	2.00
4.00	3.00	10.00	1.00	3.00	3.00
4.00	4.00			4.00	4.00
5.00	1.00			5.00	5.00
5.00	2.00			6.00	5.00
6.00	3.00			7.00	4.00
6.00	4.00			8.00	3.00
7.00	1.00			9.00	2.00
7.00	2.00			10.00	1.00
$r = 0.0$, $p = 1.0$		$r = -1.00$, $p = 0.0$		$r = 0.0$, $p = 1.0$	

restriction of range compounds the effect but produces Pearson $r = -1.00$, $p = 0.0$. Again, these are two very different results. Finally, a nonlinear data relation produces Pearson $r = 0.0$, which we are taught in our basic statistics course, because the Pearson correlation measures linear bivariate variable associations. These outcomes are very different and dramatically affect our statistical calculations and interpretations (Schumacker, 2014).

The different multivariate statistics presented in the book will address one or more of these issues. R functions will be used to assess or test whether the assumptions are met. Each chapter provides the basic R commands to perform a test of any assumptions and the multivariate statistics discussed in the chapter.

△ R Software

R is free software that contains a library of packages with many different functions. R can run on Windows, Mac OS X, or UNIX computer operating systems, which makes it ideal for students today to use with PC and Apple laptops. The R software can be downloaded from the Comprehensive R Archive Network (CRAN), which is located at the following URL:

http://cran.r-project.org/

Once R is downloaded and installed, you can obtain additional R manuals, references, and materials by issuing the following command in the RGUI (graphical user interface) window:

```
> help.start()
```

To obtain information about the R *stats* package, issue the following command in the RGui Console window:

```
> library(help=”stats”)
```

This will provide a list of the functions in the *stats* package. An index of the statistical functions available in the *stats* package will appear in a separate dialog box. The various functions are listed from A to Z with a description of each. You will become more familiar with selecting a package and using certain functions as you navigate through the various statistical methods presented in the book. A comprehensive *Introduction to R* is available online at the following URL:

http://cran.r-project.org/doc/manuals/R-intro.html

It covers the basics (reading data files, writing functions), statistical models, graphical procedures, and packages.

R is a syntax-based command language as opposed to a point and click activation. A comparison could be made between SAS (statistical analysis software; syntax commands) and SPSS, an IBM Company (statistical package for the social sciences; point and click). The point and click activation is often referred to as a GUI. Many software products are going with a mouse point and click activation to make it user friendly. However, although the point and click makes it easy to execute commands (functions), the results of what was selected in the dialog boxes is lost after exiting the software. I instruct my students, for example, when using SPSS, to always use the paste function and save the syntax. They can then recall what the point and click sequences were that obtained the statistical results.

R uses simple syntax commands and functions to achieve results, which can be saved in a file and used at a later date. This also permits adding additional commands or functions to a statistical analysis as needed. The R commands can be contained between brackets, which identifies a function, or issued separately. The appendix contains information for the installation and usage of R, as well as a reference guide of the various R packages, functions, data sets, and script files used in the chapters of the book.

Using the R software has also been made easy by two additional free R software products that use GUI windows to navigate file locations and operations. The software products are installed after you have installed R. The two software products are *Rcommander* and *RStudio*. You can download and install these software products at the following websites:

Rcommander: http://www.rcommander.com

RStudio: http://www.rstudio.com

Rcommander (Rcmdr) enables easy access to a selection of commonly used R commands with an output window directly below the command line window. It provides a main menu with options for editing data, statistics, graphs, models, and distribution types. A menu tree of the options is listed on the developer's website:

http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/

RStudio software provides an easy menu to create and store projects. The *RStudio* GUI window is partitioned into four parts. The first subwindow contains the data set, the second the console window with the R commands, the third a list of data files and commands being used in the workspace, and the fourth a menu to select files, plots, and packages or to seek help. It permits an easy way to locate and import packages that would be used to compute your statistic or plot the data. A nice feature of *RStudio* is that it will prompt you when software updates are available and activate the Internet download window for installation. In addition, *RStudio* personnel provide training workshops.

WEB RESOURCES

R is a popular alternative to commercially available software packages, which can be expensive for the end user. Given R popularity, several websites have been developed and supported, which provide easy access to information and how-to-do features with R. *Quick-R* is easy to use, informative, and located at the following URL:

http://www.statmethods.net

The website provides tutorials, a listing of books, and a menu that encompasses data input, data management, basic statistics, advanced statistics, basic graphs, and advanced graphs. The R code listed in the many examples can be easily copied, modified, and incorporated into your own R program file.

There are many R tutorials available by simply entering **R tutorials** in the search window of a browser. Some tutorials are free, while others require membership (e.g., www.lynda.com). There is a blog website that provides a fairly comprehensive list of R video tutorials at the following URL:

http://jeromyanglim.blogspot.co.uk/2010/05/videos-on-data-analysis-with-r.html

REFERENCES

Schumacker, R. E. (2014). *Learning statistics using R.* Thousand Oaks, CA: Sage.

Schumacker, R. E., & Tomek, S. (2013). *Understanding statistics using R.* New York, NY: Springer-Verlag.