

Randall Schumacker
Sara Tomek

Understanding Statistics Using R

Understanding Statistics Using R

Randall Schumacker · Sara Tomek

Understanding Statistics Using R



Springer

Randall Schumacker
University of Alabama
Tuscaloosa, AL, USA

Sara Tomek
University of Alabama
Tuscaloosa, AL, USA

ISBN 978-1-4614-6226-2

ISBN 978-1-4614-6227-9 (eBook)

DOI 10.1007/978-1-4614-6227-9

Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012956055

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*Dedicated to our children,
Rachel and Jamie
Daphne*

Preface

This book was written as a supplemental text for use with introductory or intermediate statistics books. The content of each chapter is appropriate for any undergraduate or graduate level statistics course. The chapters are ordered along the lines of many popular statistics books so it should be easy to supplement the chapter content and exercises with your statistics book and lecture materials. The content of each chapter was written to enrich a students' understanding of statistics using R simulation programs. The chapter exercises reinforce an understanding of the statistical concepts presented in the chapters.

Computational skills are kept to a minimum in the book by including R script programs that can be run for the exercises in the chapters. Students are not required to master the writing of R script programs, but explanations of how the programs work and program output are included in each chapter. R is a statistical package with an extensive library of functions that offers flexibility in writing customized statistical routines. The R script commands are run in the R Studio software which is a graphical user interface for Windows. The R Studio software makes accessing R programs, viewing output from the exercises, and graph displays easier for the student.

Organization of the Text

The first chapter of the book covers fundamentals of R. This includes installation of R and R Studio, accessing R packages and libraries of functions. The chapter also covers how to access manuals and technical documentation, as well as, basic R commands used in the R script programs in the chapters. This chapter is important for the instructor to master so that the software can be installed and the R script programs run. The R software is free permitting students to install the software and run the R script programs for the chapter exercises.

The second chapter offers a rich insight into how probability has shaped statistics in the behavioral sciences. This chapter begins with an understanding of finite and

infinite probability. Key probability concepts related to joint, addition, multiplication, and conditional probability are covered with associated exercises. Finally, the all important combination and permutation concepts help to understand the seven fundamental rules of probability theory which impact statistics.

Chapter 3 covers statistical theory as it relates to taking random samples from a population. The R script program is run to demonstrate sampling error. Basically, sampling error is expected to be reduced as size of the random sample increases. Another important concept is the generation of random numbers. Random numbers should not repeat or be correlated when sampling without replacement.

Chapter 4 covers histograms and ogives, population distributions, and stem and leaf graphs. The frequency distribution of cumulative percents is an ogive, represented by a characteristic S-shaped curve. In contrast, a data distribution can be unimodal or bimodal, increasing or decreasing in value. A stem and leaf graph further helps to visualize the data distribution, middle value and range or spread of the data. Graphical display of data is reinforced by the chapter exercises.

Chapter 5 covers measures of central tendency and dispersion. The concept of mean and median are presented in the chapter exercises, as well as the concept of dispersion or variance. Sample size effects are then presented to better understand how small versus large samples impact central tendency and dispersion. The Tchebysheff Inequality Theorem is presented to introduce the idea of capturing scores within certain standard deviations of the frequency distribution of data, especially when it is not normally distributed. The normal distribution is presented next followed by the Central Limit Theorem, which provides an understanding that sampling distributions will be normally distributed regardless of the shape of the population from which the random sample was drawn.

Chapter 6 covers an understanding of statistical distributions. Binomial distributions formed from the probability or frequency of dichotomous data are covered. The normal distribution is discussed both as a mathematical formula and as probability under the normal distribution. The shape and properties of the chi-square distribution, t-distribution, and F-distribution are also presented. Some basic tests of variance are introduced in the chapter exercises.

Chapter 7 discusses hypothesis testing by expressing the notion that “*A statistic is to a sample as a parameter is to a population*”. The concept of a sampling distribution is explained as a function of sample size. Confidence intervals are introduced for different probability areas of the sampling distribution that capture the population parameter. The R program demonstrates the confidence interval around the sample statistic is computed by using the standard error of the statistic. The statistical hypothesis with null and alternative expressions for percents, ranks, means, and correlation are introduced. The basic idea of testing whether a sample statistic falls outside the null area of probability is demonstrated in the R program. Finally TYPE I and TYPE II error are discussed and illustrated in the chapter exercises using R programs.

Chapters 8–13 cover the statistics taught in an elementary to intermediate statistics course. The statistics covered are chi-square, z, t, F, correlation, and regression. The respective chapters discuss hypothesis testing steps using these statistics. The R

programs further calculate the statistics and related output for interpretation of results. These chapters form the core content of the book whereas the earlier chapters lay the foundation and groundwork for understanding the statistics. A real benefit of using the R programs for these statistics is that students have free access at home and school. An instructor can also use the included R functions for the statistics in class thereby greatly reducing any programming or computational time by students.

Chapter 14 is included to present the concept that research should be replicated to validate findings. In the absence of being able to replicate a research study, the idea of cross validation, jackknife, and bootstrap are commonly used methods. These methods are important to understand and use when conducting research. The R programs make these efforts easy to conduct. Students gain further insight in Chap. 15 where a synthesis of research findings help to understand overall what research results indicate on a specific topic. It further illustrates how the statistics covered in the book can be converted to a common scale so that effect size measures can be calculated, which permits the quantitative synthesis of statistics reported in research studies. The chapter concludes by pointing out that statistical significance testing, i.e., $p < 0.05$, is not necessarily sufficient evidence of the practical importance of research results. It highlights the importance of reporting the sample statistic, significance level, confidence interval, and effect size. Reporting of these values extends the students' thinking beyond significance testing.

R Programs

The chapters contain one or more R programs that produce computer output for the chapter exercises. The R script programs enhance the basic understanding and concepts in the chapters. The R programs in each chapter are labeled for easy identification. A benefit of using the R programs is that the R software is free for home or school use. After mastering the concepts in the book, the R software can be used for data analysis and graphics using pull-down menus. The use of R functions becomes a simple cut-n-paste activity, supplying the required information in the argument statements.

There are several Internet web sites that offer information, resources, and assistance with R, R programs, and examples. These can be located by entering "R software" in the search engines accessible from any Internet browser software. The main Internet URL (Uniform Resource Locator) address for R is: <http://www.r-project.org>. A second URL is: <http://lib.stat.cmu.edu/R/CRAN>. There are also many websites offering R information, statistics, and graphing, for example, Quick-R at <http://www.statmethods.net>.

Tuscaloosa, AL, USA

Randall Schumacker
Sara Tomek

Contents

1 R Fundamentals	1
Install R	1
Install R Studio	3
Getting Help.....	4
Load R Packages.....	5
Running R Programs.....	7
Accessing Data and R Script Programs	8
Summary	9
** WARNING **	10
R Fundamentals Exercises	10
True or False Questions	10
2 Probability	11
Finite and Infinite Probability	11
PROBABILITY R Program	12
PROBABILITY R Program Output.....	13
Finite and Infinite Exercises.....	14
Joint Probability	18
JOINT PROBABILITY Exercises	21
Addition Law of Probability	23
ADDITION Program Output	24
ADDITION Law Exercises.....	25
Multiplication Law of Probability	26
Multiplication Law Exercises	28
Conditional Probability	29
CONDITIONAL Probability Exercises	32
Combinations and Permutations	34
Combination and Permutation Exercises	38
True or False Questions	40
Finite and Infinite Probability	40

Joint Probability	40
Addition Law of Probability	40
Multiplication Law of Probability	41
Conditional Probability	41
Combination and Permutation	41
3 Statistical Theory	43
Sample Versus Population.....	43
STATISTICS R Program.....	44
STATISTICS Program Output.....	45
Statistics Exercises.....	46
Generating Random Numbers.....	48
RANDOM R Program	49
RANDOM Program Output.....	50
Random Exercises.....	51
True and False Questions	53
Sample versus Population	53
Generating Random Numbers.....	53
4 Frequency Distributions	55
Histograms and Ogives	55
FREQUENCY R Program	56
FREQUENCY Program Output.....	57
Histogram and Ogive Exercises	58
Population Distributions	62
COMBINATION Exercises	65
Stem and Leaf Graph	66
STEM-LEAF Exercises	70
True or False Questions	72
Histograms and Ogives	72
Population Distributions	73
Stem and Leaf Graphs.....	73
5 Central Tendency and Dispersion	75
Central Tendency	75
MEAN-MEDIAN R Program.....	76
MEAN-MEDIAN Program Output.....	76
MEAN-MEDIAN Exercises	77
Dispersion	79
DISPERSION Exercises	81
Sample Size Effects	83
SAMPLE Exercises	84
Tchebysheff Inequality Theorem.....	86
TCHEBYSHEFF Exercises	90
Normal Distribution	91

Normal Distribution Exercises.....	93
Central Limit Theorem	95
Central Limit Theorem Exercises	101
True or False Questions	103
Central Tendency	103
Dispersion	104
Sample Size Effects	104
Tchebysheff Inequality Theorem.....	104
Normal Distribution	105
Central Limit Theorem	105
6 Statistical Distributions	107
Binomial.....	107
BINOMIAL R Program	109
BINOMIAL Program Output.....	110
BINOMIAL Exercises	110
Normal Distribution	112
NORMAL R Program.....	114
NORMAL Program Output	114
NORMAL Distribution Exercises.....	115
Chi-Square Distribution	116
CHISQUARE R Program	117
CHISQUARE Program Output	118
CHISQUARE Exercises.....	119
t-Distribution.....	122
t-DISTRIBUTION R Program.....	124
t-DISTRIBUTION Program Output	124
t-DISTRIBUTION Exercises.....	125
F-Distribution.....	128
F-DISTRIBUTION R Programs	132
F-Curve Program Output	132
F-Ratio Program Output	133
F-DISTRIBUTION Exercises.....	133
True or False Questions	135
Binomial Distribution	135
Normal Distribution	135
Chi-Square Distribution	136
t-Distribution	136
F-Distribution.....	136
7 Hypothesis Testing	137
Sampling Distribution.....	137
DEVIATION R Program.....	139
DEVIATION Program Output	140

Deviation Exercises.....	141
Confidence Intervals	142
CONFIDENCE R Program.....	144
CONFIDENCE Program Output	144
Confidence Interval Exercises.....	145
Statistical Hypothesis.....	146
HYPOTHESIS TEST R Program	150
HYPOTHESIS TEST Program Output.....	151
Hypothesis Testing Exercises.....	152
TYPE I Error.....	154
TYPE I ERROR R Program.....	157
TYPE I ERROR Program Output	158
TYPE I Error Exercises.....	158
TYPE II Error	160
TYPE II ERROR R Program	163
TYPE II ERROR Program Output.....	164
TYPE II Error Exercises	164
True or False Questions	166
Sampling Distributions	166
Confidence Interval	166
Statistical Hypothesis.....	167
TYPE I Error.....	167
TYPE II Error	168
8 Chi-Square Test.....	169
CROSSTAB R Program.....	172
CROSSTAB Program Output.....	173
Example 1	173
Example 2	173
Chi-Square Exercises	174
True or False Questions	175
Chi-Square	175
9 z-Test	177
Independent Samples	177
Dependent Samples.....	180
ZTEST R Programs.....	184
ZTEST-IND Program Output.....	184
ZTEST-DEP Program Output	184
z Exercises	185
True or False Questions	186
z-Test.....	186
10 t-Test.....	187
One Sample t-Test.....	187
Independent t-Test.....	189

Dependent t-Test	190
STUDENT R Program.....	192
STUDENT Program Output	192
t Exercises	193
True or False Questions	194
t-Test	194
11 F-Test.....	197
Analysis of Variance	197
One-Way Analysis of Variance	198
Multiple Comparison Tests	200
Repeated Measures Analysis of Variance	201
Analysis of Variance R Programs	203
ONEWAY Program.....	203
ONEWAY Program Output.....	204
Scheffe Program Output.....	205
REPEATED Program Output.....	205
F Exercises	206
True or False Questions	207
F Test.....	207
12 Correlation.....	209
Pearson Correlation.....	209
Interpretation of Pearson Correlation.....	211
CORRELATION R Program	214
CORRELATION Program Output.....	214
Correlation Exercises	215
True or False Questions	218
Pearson Correlation.....	218
13 Linear Regression	219
Regression Equation	220
Regression Line and Errors of Prediction.....	221
Standard Scores.....	224
REGRESSION R Program	225
REGRESSION Program Output.....	226
REGRESSION Exercises.....	227
True or False Questions	228
Linear Regression	228
14 Replication of Results	229
Cross Validation	230
CROSS VALIDATION Programs	230
CROSS VALIDATION Program Output	231
Cross Validation Exercises.....	232

Jackknife	234
JACKKNIFE R Program.....	236
JACKKNIFE Program Output	237
Jackknife Exercises	237
Bootstrap	239
BOOTSTRAP R Program.....	242
BOOTSTRAP Program Output	242
Bootstrap Exercises.....	242
True or False Questions	244
Cross Validation	244
Jackknife	244
Bootstrap.....	245
15 Synthesis of Findings	247
Meta-Analysis	247
A Comparison of Fisher and Gordon Chi-Square Approaches	248
Converting Various Statistics to a Common Metric.....	249
Converting Various Statistics to Effect Size Measures	249
Comparison and Interpretation of Effect Size Measures	250
Sample Size Considerations in Meta-Analysis	252
META-ANALYSIS R Programs	253
Meta-Analysis Program Output	254
Effect Size Program Output	254
Meta-Analysis Exercises.....	254
Statistical Versus Practical Significance	256
PRACTICAL R Program.....	259
PRACTICAL Program Output.....	260
PRACTICAL Exercises	260
True or False Questions	261
Meta-Analysis	261
Statistical Versus Practical Significance	261
Glossary of Terms	263
Appendix	271
Author Index.....	279
Subject Index.....	281

Chapter 1

R Fundamentals

Install R

R is a free open-shareware software that can run on Unix, Windows, or Mac OS X computer operating systems. The R software can be downloaded from the Comprehensive R Archive Network (CRAN) which is located at: <http://cran.r-project.org/>. There are several sites or servers around the world where the software can be downloaded, which is accessed at: <http://cran.r-project.org/mirrors.html>. The R version for Windows will be used in the book, so if using Linux or Mac OS X operating systems follow the instructions on the CRAN website.

After entering the URL: <http://cran.r-project.org/> you should see the following screen.

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (<http://cran.r-project.org/bin/linux/>)
- Download R for MacOS X (<http://cran.r-project.org/bin/macosx/>)
- Download R for Windows (<http://cran.r-project.org/bin/windows/>)

After clicking on the “*Download R for Windows*”, the following screen should appear where you will click on “**base**” to go to the next screen for further instructions.

After clicking on “**base**”, the following screen should appear to download the Windows installer executable file, e.g. R-2.15.1-win.exe (The version of R available for download will change periodically as updates become available, this is version 2.15.1 for Windows).

R for Windows

Subdirectories:

base (http://cran.r-project.org/bin/windows/base/)	Binaries for base distribution (managed by Duncan Murdoch). This is what you want if you install R for the first time (http://cran.r-project.org/bin/windows/base/)
contrib (http://cran.r-project.org/bin/windows/contrib/)	Binaries of contributed packages (managed by Uwe Ligges)

You may also want to read the R FAQ (<http://cran.r-project.org/doc/FAQ/R-FAQ.html>) and R for Windows FAQ (<http://cran.r-project.org/bin/windows/base/rw-FAQ.html>).

Run the executable file by double-clicking on the file name (R-2.15.1-win.exe) once it has been downloaded to install, which will open the R for Windows setup wizard.

R-2.15.1 for Windows (32/64 bit)

Download R 2.15.1 for Windows (<http://cran.r-project.org/bin/windows/base/R-2.13.1-win.exe>) (47 megabytes, 32/64 bit)

- Installation and other instructions (<http://cran.r-project.org/bin/windows/base/README.R-2.13.1>)
- New features in this version: Windows specific (<http://cran.r-project.org/bin/windows/base/CHANGES.R-2.13.1.html>), all platforms (<http://cran.r-project.org/bin/windows/base/NEWS.R-2.13.1.html>).

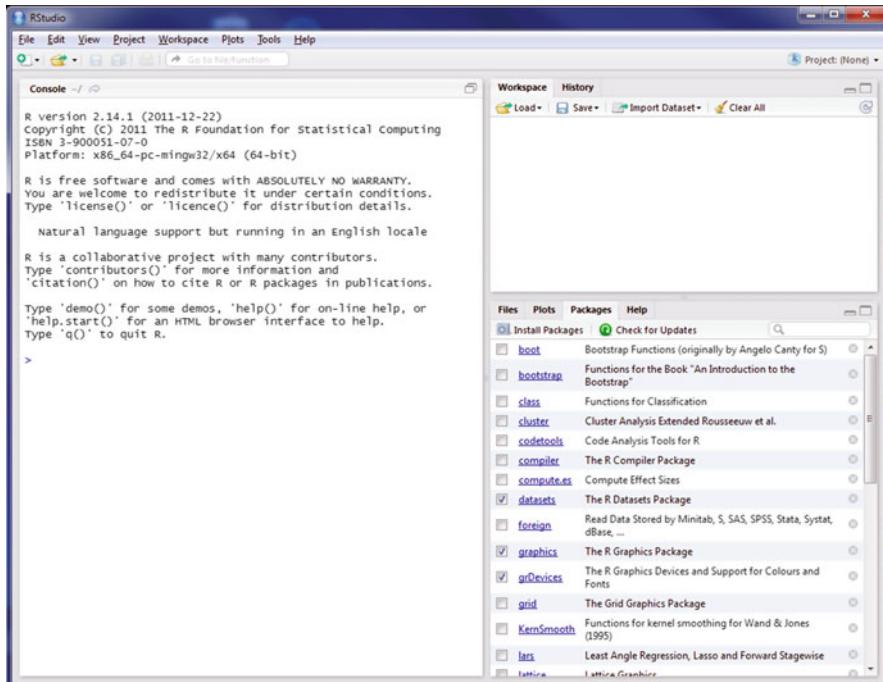
NOTE: The Download R 2.xx.x for Windows version will have changed to newer versions, so simply download the latest version offered.

Install R Studio

The R Studio interface, which is installed after installing the R software, provides an easy to use GUI windows interface (Graphical User Interface), download from: <http://www.rstudio.org/> (Must have R2.13.1 or higher version on PC, Linux, or Mac OS X 10.5 before download and install of this software). The following desktop icon will appear after installation.



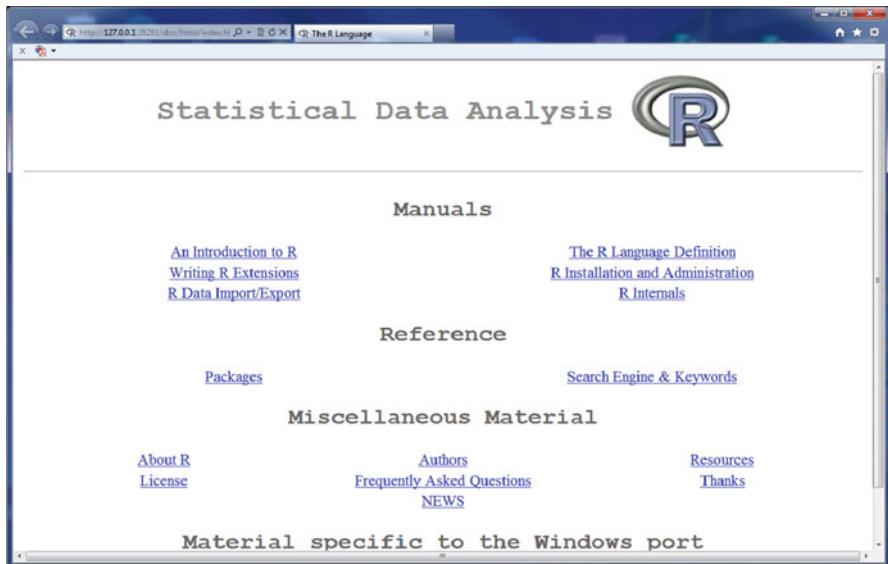
The R Studio window provides the usual R console. It also provides a workspace/history window with load/save/import data set features. Another window provides easy access to files and a list of packages available. The Plots tab also shows a created plot and permits easy Export to a GIF image, PDF file, or copy to the clipboard feature to insert into a Word document.



Getting Help

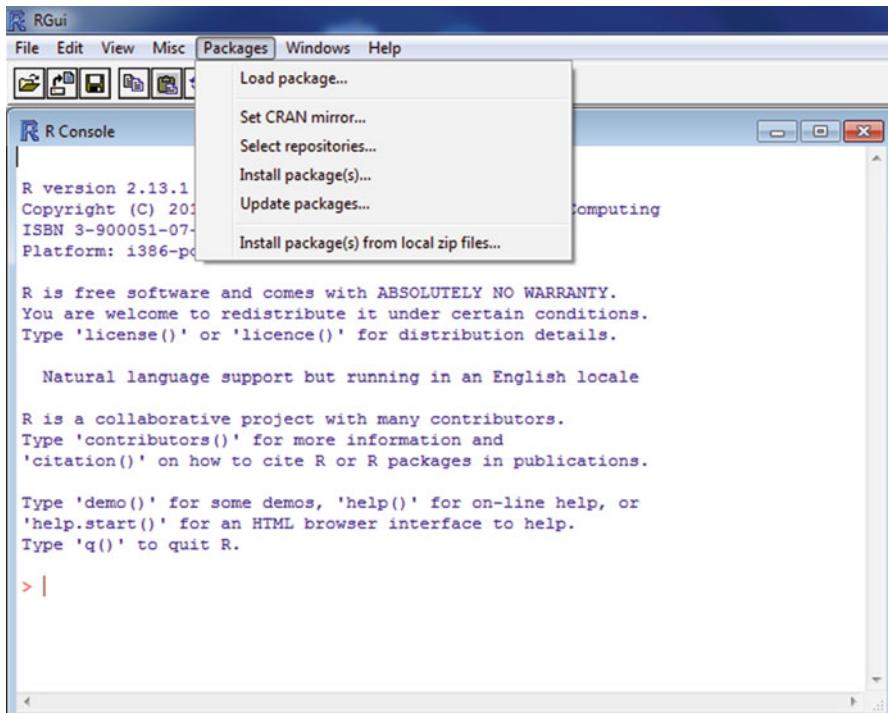
The R software contains additional manuals, references, and material accessed by issuing the following command in the RGui window once R is installed:

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



Load R Packages

Once R is installed and the RGui window appears, you can load R packages with routines or programs that are not in the “**base**” package. Simply click on “**Packages**” in the main menu of the RGui window, and then make your selection, e.g., “**Load packages**”.



A dialog box will appear which lists the base package along with an alphabetical list of other packages. I selected “**stats**” from the list and clicked OK. This makes available all of the routines or commands in the “**stats**” package. Alternatively, prior to entering R commands in the R Console window, you can load the package from a library with the command:

```
> library(stats)
```



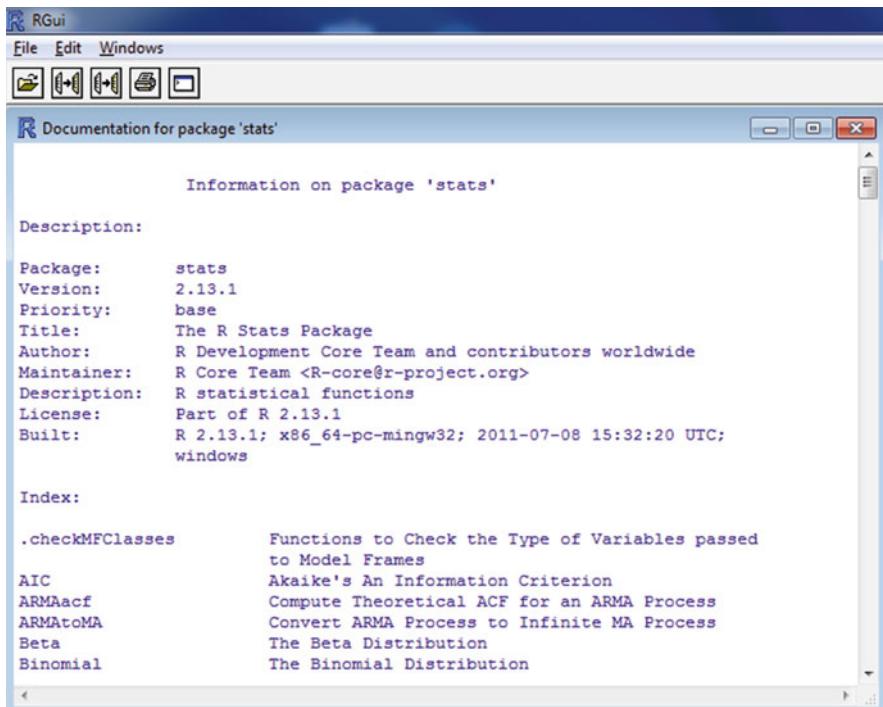
To obtain information about the R “stats” package issue the following command in the R Console:

```
> help(stats)
```

or

```
> library(help="stats")
```

which will provide a list of the functions or routines in the “**stats**” package. An index of the statistical functions available in the “**stats**” package will appear in a separate dialog box.



Running R Programs

To run R programs in the book, you will Click on *File*, then select *Open script* from the main menu in the RGUi window. For example, locate and select *chap01_Begin.r* script file, which then opens in a separate R Editor window.

```
# Begin Chapter 1  
  
# Basic R commands  
  
x=5  
y=4  
z=x+y  
z  
  
age=c(25,30,40,55)  
age
```

Next, click on *Edit* in the pull down menu and select *run all*, to execute all of the command lines in the program. The Editor Window will show:

```
> # Begin Chapter 1
>
> # Basic R commands
>
> x = 5
> y = 4
> z = x + y
> z
[1] 9
>
> age = c(25,30,40,55)
> age
[1] 25 30 40 55
```

NOTE: The R programs included with this book may end up with different results each time a random sample is selected. Use `set.seed(13579)` prior to running the R programs to obtain the same random sample, which will provide the same example results.

NOTE: Commands in R code can be found using either the `=` or the `<-` notation. Either will work and they can be used interchangeably, for example, `x = 5` or `x <- 5` assigns the value 5 to x.

Accessing Data and R Script Programs

The `getwd()` and `setwd()` commands identify the current working directory and sets a new working directory, respectively. For example,

```
> getwd()
[1] "C"/Users/Name"
> setwd ("C:/Users/Documents")
```

The R command, `read.table`, with the argument, `file=file.choose()`, can help to locate and open a data file.

```
> read.table(file=file.choose())
```

The `file.choose()` option is by far the best way, rather than guess where data files are located, but does require personal action unlike specifying file location directly in an R program. An example will help to illustrate these R functions for locating and inputting a data file types (ASCII, “sample.txt”).

```
> getwd()
[1] "C:/Users/Name"
> setwd ("C:/Users/Documents")
> read.table (file = "sample.txt", header=TRUE)

# Alternative
read.table (file="C:/Users/Documents/sample.
txt",header=TRUE)

# Alternative read.table(file=file.choose())

student score
1      1    16
2      2    14
3      3    24
4      4    23
5      5    25
6      6    22
```

The **R Studio** program has the added benefit over RGui (R graphical user interface) by including a display window of the computer directory. This makes finding R script programs and data files much easier.

Summary

The R script programs in the book have sets of bundled R commands in an R function which provides a chapter name followed by the R program name to execute all of the simulation operations. The following function, **results**, has the set of operations between the brackets, { and }. Issuing the function name, **results()** executes the operations in the function. For example,

```
> results = function()
{
  x={1:5}
  y={5:9}

  output=matrix(1:2)
  output=c(mean(x),mean(y))
  names(output)=c("xmean","ymean")

  output
}
> results()

  xmean  ymean
 3       7
```

**** WARNING ****

When running R programs, the workspace keeps track of variable names, functions, data vectors, etc. If you run several programs, one after the other, eventually R will crash or the results will be unstable or incorrect.

1. Use `ls()` to view active variables.
2. Use `rm()` to remove active variables or programs, especially if making changes to code.

It is always best to run a single R program at a time. It is good programming technique to use `detach()` function to clear out previous values created by R syntax or close, then reopen RGui to run another R program.

R Fundamentals Exercises

1. What command would you use to obtain R manuals and reference material?
2. What command would you use to obtain information about the {stats} package?
3. What R commands would yield z as the sum of $x = 5$ and $y = 6$?
4. What R command would yield `test` as a data vector with values 10, 20, 30, 40, 50?
5. What R command lists all active variables in the work environment?

True or False Questions

R Basics

- T F a. R software can only be run on an IBM PC.
T F b. R software has extensive manuals, references, and documentation material available for the user.
T F c. The `setwd()` command identifies the current working directory.
T F d. R software can analyze data similar to other statistical packages.
T F e. The `rm()` command removes variables from the working directory.

Chapter 2

Probability

Finite and Infinite Probability

One's ability to determine the probability of an event is based upon whether the event occurs in a finite or infinite population. In a **finite population**, the number of objects or events is known. An exact **probability** or fraction can be determined. For example, given a population of 1,000 cars with 500 Ford, 200 Chevrolet, 200 Chrysler, and 100 Oldsmobile, the probability of selecting a Ford is one-half or 50% ($500/1,000$). The probability of selecting a Chevrolet is one-fifth or 20% ($200/1,000$), the probability of selecting a Chrysler is one-fifth or 20% ($200/1,000$), and the probability of selecting an Oldsmobile is one-tenth or 10% ($100/1,000$). The individual probabilities add up to 100%.

In an **infinite population**, the numbers of objects or events are so numerous that exact probabilities are difficult to determine. One approach to determine the probability of an event occurring in an infinite population is to use the relative frequency definition of probability. Using this approach, trials are repeated a large number of times, N . The number of times the event occurs is counted, and the probability of the event is approximated by $P(A) \approx n(A)/N$ in which $n(A)$ is the number of times event A occurred out of N trials.

For example, the probability of obtaining heads when a coin is tossed could be determined by tossing the coin 500 times, counting the number of heads, $n(\text{heads})=241$, and computing $P(\text{heads}) \approx 241/500=0.482$. The probability of getting heads in the population is therefore approximately 48%. The probability of *not* getting heads is 52%, since the two events must sum to 100%. We know from experience that the probability of obtaining heads when a coin is tossed should be approximately 50% or one-half, given an unbiased coin (a coin that is not weighted or not a trick coin).

An important point needs to be emphasized. In order for an approximation to be reasonable (representative), the relative frequency of the event (e.g., heads) must begin to stabilize and approach some fixed number as the number of trials increases. If this does not occur, then very different approximations would be assigned to the same event as the number of trials increases. Typically, more trials (coin tosses) are required to potentially achieve the 50% probability of obtaining heads. Experience in the real world has shown that the relative frequency of obtaining heads when coins are tossed stabilizes or better approximates the expected probability of 50%, as the number of trials increases.

This approximation phenomenon (stabilization or representativeness) occurs in the relative frequencies of other events too. There is no actual proof of this because of the nature of an infinite population, but experience does support it. Using the relative frequency definition, the approximation, which the relative frequencies provide, is regarded as our best estimate of the actual probability of the event.

In this chapter, you will have an opportunity to observe the stabilization of the relative frequencies. You will be able to choose the probability of the event and the R program will simulate the number of trials. As the number of trials increase, the new results are pooled with the earlier ones, and the relative frequency of the event is computed and plotted on a graph. The first program example starts with a sample size of 100 and increases the sample size in increments of 100 up to 1,000.

In the real world we observe that as the number of trials increases the relative frequency of an event approaches a fixed value. The probability of an event can be defined as the fixed value approximated by the relative frequencies of the event as the number of trials increase. The relative frequency definition of probability assumes that the relative frequencies stabilize as the number of trials increase. Although the relative frequencies get closer to a fixed value as the number of trials increase, it is possible for a certain number of trials to produce a relative frequency that is not closer to the approximated fixed value. An event with a probability close to 0.1 or 0.9 will have relative frequencies that stabilize faster than an event with probability close to 0.50.

PROBABILITY R Program

The PROBABILITY program simulates flipping a coin a different number of times for different samples, and observing the different frequencies across sample sizes. The population probability is set in the variable *Probability* and the *SampleSizes* are created using the **seq** function, instead of specifying each value within a **c** function (which would be 10 numbers in this case). The **seq** function creates a vector of values from 100 to 1,000 with intervals of 100. Next, the *SampleFreqs* object is set to **NULL** so that it may be constructed into a vector by appending values using the **c** function. A **for** loop iterates through all values within the *SampleSizes* vector, assigning each value in turn to the *SampleSize* variable.

The *SampleFreqs* vector is then increased by appending the sum of the samples of size *SampleSize* taken from the population of 0 and 1. This results in 0 having a probability of $(1 - \text{Probability})$ and 1 having a probability of *Probability*, which is divided by *SampleSize* to get the relative frequency. The *SampleFreqs* vector now contains the relative frequencies of heads in each sample size. These relative frequencies are plotted with a line graph using the generic **plot** function. *SampleSize* is used for the data on the x-axis and *SampleFreqs* for the data on the y-axis. **Type = "l"** (a lower case L) is for line graph and the **ylim** keyword sets the upper and lower limits of values for the y-axis.

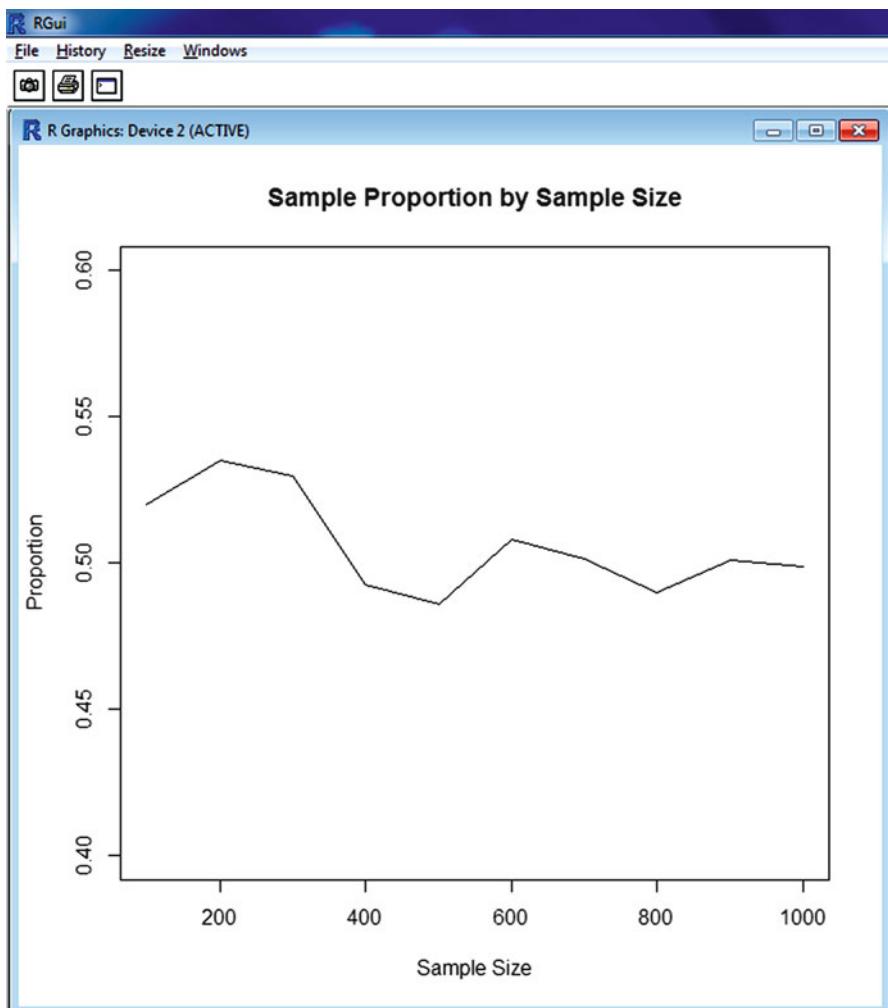
Using the full range of possible values for the y-axis (0 to 1) resulted in difficulty distinguishing differences in the graphs because of the small variation of the values compared to the overall scale. Since values rarely fall 0.10 above or 0.10 below the population probability, the y-axis limits were set to *Probability* -0.10 and *Probability* +0.10. The x-axis label, y-axis label, and main title are all set by use of keywords.

The values from the graph are constructed into a matrix for output. The matrix starts out as a **NULL** object that is built using **rbind** (row bind). A row of values is added to the matrix for each iteration of the **for** loop, appending the relative frequency of the given sample, the population probability, and the error of estimation for each sample (relative frequency—population probability). After the matrix is constructed and the **for** loop is ended, the **dimnames** function is used to assign dimension names to the constructed matrix. The **paste** function is again utilized to create a vector of labels for the rows of the matrix resulting in “sample size=100”, etc. The **print** function is used to output the matrix in order to make the rows and columns printed. The Error is the difference between the sample percent and the true population percent.

NOTE: As noted in Chap. 1, use `set.seed(13579)` prior to running the R programs to get identical results presented below.

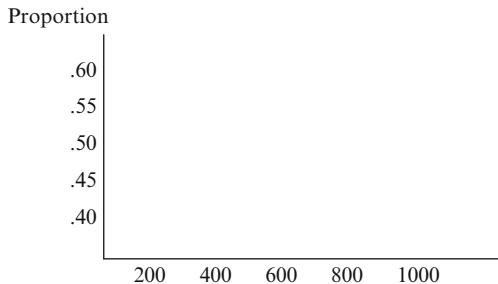
PROBABILITY R Program Output

	Sample %	Population %	Error
sample size = 100	0. 52	0.500	0.02
sample size = 200	0.535	0.500	0.035
sample size = 300	0. 53	0.500	0.03
sample size = 400	0.492	0.500	-0.008
sample size = 500	0.486	0.500	-0.014
sample size = 600	0.508	0.500	0.008
sample size = 700	0.501	0.500	0.001
sample size = 800	0. 49	0.500	-0.01
sample size = 900	0.501	0.500	0.001
sample size = 1000	0.499	0.500	-0.001



Finite and Infinite Exercises

1. Run PROBABILITY for sample sizes of 50 in increments of 50 up to 1,000, i.e.,
`SampleSizes <- seq(50,1000,50)`. This is a simulation of flipping an unbiased, balanced coin and recording the relative frequency of obtaining heads, which has an expected probability of $p=0.50$.
 - a. Complete the graph of relative frequencies for sample sizes of $n=50$, in increments of 50, up to 1,000, for $p=0.50$.



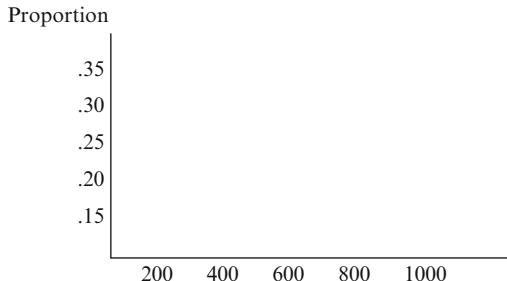
- b. Complete the table below. Does the Error (difference between the sample percent and the population percent of 0.50) ever become less than 0.01?

If so, for what sample sizes? _____

Table of sample and population percents for coin toss

	SAMPLE %	POPULATION %	ERROR
SAMPLE SIZE=50	0.500	0.500	
SAMPLE SIZE=100	0.500	0.500	
SAMPLE SIZE=150	0.500	0.500	
SAMPLE SIZE=200	0.500	0.500	
SAMPLE SIZE=250	0.500	0.500	
SAMPLE SIZE=300	0.500	0.500	
SAMPLE SIZE=350	0.500	0.500	
SAMPLE SIZE=400	0.500	0.500	
SAMPLE SIZE=450	0.500	0.500	
SAMPLE SIZE=500	0.500	0.500	
SAMPLE SIZE=550	0.500	0.500	
SAMPLE SIZE=600	0.500	0.500	
SAMPLE SIZE=650	0.500	0.500	
SAMPLE SIZE=700	0.500	0.500	
SAMPLE SIZE=750	0.500	0.500	
SAMPLE SIZE=800	0.500	0.500	
SAMPLE SIZE=850	0.500	0.500	
SAMPLE SIZE=900	0.500	0.500	
SAMPLE SIZE=950	0.500	0.500	
SAMPLE SIZE=1000	0.500	0.500	

2. Run PROBABILITY for sample sizes of 50 in increments of 50 up to 1,000, i.e., SampleSizes <- seq(50,1000,50). This time change the population percent to 25%, i.e., Probability <- 0.25. You are simulating the flipping of a biased, unbalanced coin.
- a. Complete the graph of relative frequencies for sample sizes of n=50, in increments of 50, up to 1,000, for p=0.25.



- b. Complete the table below. Does the absolute difference between the sample percent and the population percent of 0.250 ever become less than 0.01?

If so, for what sample sizes? _____

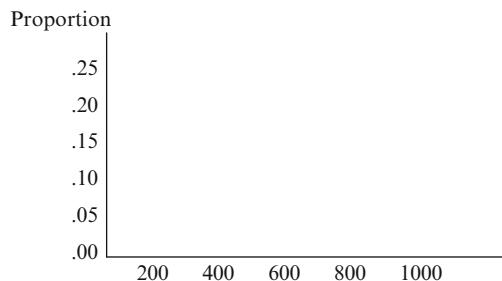
Table of sample and population percents for coin toss

	SAMPLE %	POPULATION %	ERROR
SAMPLE SIZE=50	0.250	0.250	
SAMPLE SIZE=100	0.250	0.250	
SAMPLE SIZE=150	0.250	0.250	
SAMPLE SIZE=200	0.250	0.250	
SAMPLE SIZE=250	0.250	0.250	
SAMPLE SIZE=300	0.250	0.250	
SAMPLE SIZE=350	0.250	0.250	
SAMPLE SIZE=400	0.250	0.250	
SAMPLE SIZE=450	0.250	0.250	
SAMPLE SIZE=500	0.250	0.250	
SAMPLE SIZE=550	0.250	0.250	
SAMPLE SIZE=600	0.250	0.250	
SAMPLE SIZE=650	0.250	0.250	
SAMPLE SIZE=700	0.250	0.250	
SAMPLE SIZE=750	0.250	0.250	
SAMPLE SIZE=800	0.250	0.250	
SAMPLE SIZE=850	0.250	0.250	
SAMPLE SIZE=900	0.250	0.250	
SAMPLE SIZE=950	0.250	0.250	
SAMPLE SIZE=1000	0.250	0.250	

- c. In what way is this graph for $p=0.25$ different from the first graph for $p=0.50$?
- _____
- _____

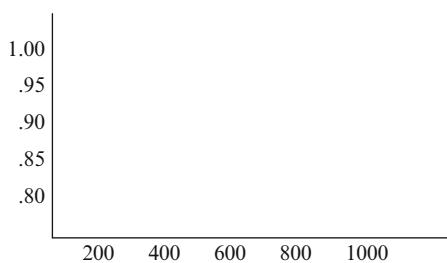
3. Run PROBABILITY again for sample sizes of 50, in increments of 50, up to 1,000, but this time for $p=0.10$ and $p=0.90$. Draw the graphs below.

$$p = .10$$



$$p = .90$$

Proportion



- a. In what way are these graphs different from the graphs for a probability of 0.50?

- b. What is the implication of this difference when you approximate a very small or a very large probability?

- c. Run PROBABILITY for $p=0.20, 0.30, 0.40, 0.60, 0.70$, and 0.80 . Describe the graphs in comparison with those for probabilities of 0.10, 0.90, and 0.50.

Joint Probability

The theoretical probability for the joint occurrence of two independent events is reflected in the relative frequency of their joint occurrence. There is a multiplication and addition law of probability for two independent or mutually exclusive events. The theoretical probability for the union of two events is reflected in the relative frequency of the occurrence of either event.

If an unbiased coin is flipped two times, the possible outcomes form a sample space, S. The sample space $S = \{HH, HT, TH, TT\}$, in which H stands for a head and T for a tail, with the pair of letters indicating the order of the outcomes. Therefore, with two separate flips of a coin, four possible outcomes can occur: two heads, first a head then a tail, first a tail then a head, or two tails. The sample space that contains the number of heads in the two flips is $S = \{0, 1, 2\}$. If an unbiased coin is flipped twice, a large number of times, the outcomes can be used to compute the frequencies. The frequencies can be used to approximate the joint probabilities of the outcomes in the sample space.

Probabilities can also be assigned to the outcomes by using a theoretical approach. Since a head and a tail are equally likely to occur on a single flip of an unbiased coin, the theoretical approach uses a definition of probability that is applicable to equally likely events. The probability of a head, $P(H)$, is $1/2$ because a head is one of the two equally likely outcomes. The probability of a tail, $P(T)$, is $1/2$, since a tail represents the other equally likely outcome. Since the two flips of the coin are independent, the multiplication law of probability for independent events can be used to find the joint probability for the pairs in the sample space. For example, the probability of flipping an unbiased coin and getting heads both times is: $P(HH) = (1/2)*(1/2) = 1/4$. The probability of getting a head and then a tail would be: $P(HT) = (1/2)*(1/2) = 1/4$, with the other pairs in the sample space determined in the same manner.

If the coin is *biased*, meaning that $P(H)$ is some value other than 0.50, for example 0.60, then $P(T) = 1 - P(H) = 1 - 0.60 = 0.40$. The independence of the coin flips can be used to find the **joint probability** for the pair. For example, $P(HT) = P(H)*P(T) = (0.60)*(0.40) = 0.24$.

If the sample space being used is $S = \{0, 1, 2\}$, with the integers representing the number of heads, then the frequency of 0 is the frequency of TT; the frequency of 1 is the frequency of HT plus the frequency of TH; and the frequency of 2 is the frequency of HH. The theoretical probabilities for S can also be obtained by using the addition law of probability for mutually exclusive events. For example, $P(1) = P(HT) + P(TH)$.

The multiplication and addition laws of probability for independent events reflects the properties of frequencies. If two events A and B are independent, then $P(A \text{ and } B) = P(A)*P(B)$. If two events A and B are mutually exclusive, then $P(A \text{ and } B) = P(A) + P(B)$.

JOINT PROBABILITY R Program

The JOINT R program specifies the probability of tossing a head P(H), and the number of repetitions of the two coin flips. The program will simulate tossing the coin, compute the frequencies and the probabilities. The frequencies approximate the probabilities. This supports the conclusion that the theoretical laws used to compute the probabilities give results similar to the frequency of data obtained in practice. Each time the program is run, the frequencies will be different because random flips of the coin are simulated.

The program simulates tossing two (or more) coins N number of times. The program begins by initializing the probability of obtaining a head, the number of coins to be tossed, and the number of times to toss each coin. A vector of heads (1) or tails (0) values is created, and then grouped into a **matrix** with the number of columns equal to the number of coins (column 1 = coin 1, etc.) and the number of rows equal to the number of times each coin is tossed. Next, a vector is initialized and then filled using a **for** loop with the sum of the number of heads in each round of tosses. A complex nested function allows for any number of coins to be tossed.

Vectors for the event labels (HH, HT, TH, TT), the event probabilities, and the number of heads present in each event are initialized with the appropriate values. The outer loop represents the range of possible events given the number of coins. The number of possible events is $2^{numCoins}$, which is read 2 to the power of $numCoins$. The loop range is set to 0 for $2^{numCoins}-1$. A temporary holding variable is set to the current value of the outer loop counter, i , and then the inner loop begins, which represents each coin tossed in a given round taken in reverse order.

In order to make each event unique, a binary coding system is used whereby the event's value (i) is broken down into binary values for each toss of a coin in the group. An event value of zero for two coins would mean tails-tails (or 0-0). An event value of one, for two coin tosses, would mean heads-tails (or 1-0). The first coin tossed has a value of either 0 or 1. The second coin tossed a value of either 0 or 2. The third coin tossed (if there were a third coin) would have a value of 0 or 4. The n th toss would have a value of 0 or $2^{(n-1)}$, i.e., $2^{numCoins}-1$. In this manner, all the unique events (from 0 to $2^{numCoins}-1$) are broken down into whether the first and/or second (and/or third and/or fourth, etc) coins are heads or tails for that event. Labels are created with ordered letters representing what the binary coding represents internally. The label "HH" for an event of head-head is more readily understood than an event code of 3.

The number of total heads for each event is recorded. The vector containing the heads count is factored for all possible values and then counted by means of a **table** function to determine the total number of events resulting in 0 heads, 1 head, 2 heads, and so forth, depending upon the number of coin tosses. The total number of events is then used in calculating the probabilities in the **for** loop.

The **for** loop calculates the probabilities for each number of heads by the order of the event. If the probability of getting a head is 0.60, then the probability of getting two tails (or no heads) on the first toss is $(1)(0.40)(0.40)=0.16$. This implies that

there is only one way to get two tails (1), times the probability of tails (0.40), times the probability of tails (0.40). The probability of getting one head and one tail is $(2)(0.60)(0.40)=0.48$. There are two different ways (head–tail or tail–head) you could get the pair of heads and tails (2), times probability of heads (0.60), times probability of tails (0.40). The loop variable, i , represents the number of heads and $numEvents[i+1]$ represents the number of events in the event space for that number of heads. The probability of a head (pH) is taken to the power of the number of heads obtained (i) and any coins that aren't heads must be tails ($numCoins - i$), so the probability of a tail ($1 - pH$) is taken to the power of that value. For an event that involves flipping a coin two times, the loop will go from 0 to 2 and the $numEvents$ vector will contain 1, 2, 1 (one event with no heads, two events with one head, and one event with two heads). The probability of heads can be set to any value between 0 and 1.

A second loop codes all of the rounds of tosses into a binary coding scheme in order to count the number in each group. The results are now put into matrices in order to print. The **table** and **factor** functions are invaluable in sorting categorical data for summarizing and reporting. The first matrix contains: (1) row labels with the possible number of heads that could be obtained (0 to $numCoins$); (2) the frequency of each round of flips that obtained that number of heads divided by the total number of rounds (giving the frequency); and (3) the theoretical probability of obtaining that many heads. The second matrix contains: (1) row labels with the event labels (HH, TH, HT, TT); (2) the frequency of each event obtained during all rounds; and (3) the theoretical probability of obtaining each event. The last two lines of the program prints out the matrices. The number of coins selected should not exceed 5 and sample sizes larger than 5,000 will require more time for the program to run.

Given these values:

```
pH <- 0.5
numCoins <- 2
N <- 100
```

	Sample %	Population %
--	----------	--------------

0 Heads	0.28	0.25
1 Heads	0.54	0.50
2 Heads	0.18	0.25

	Sample %	Population %
--	----------	--------------

TT	0.28	0.25
HT	0.33	0.25
TH	0.21	0.25
HH	0.18	0.25

Given these values:

```
pH <- 0.5
numCoins <- 3
N <- 100
```

	Sample %	Population %
0 Heads	0.17	0.125
1 Heads	0.42	0.375
2 Heads	0.35	0.375
3 Heads	0.06	0.125

	Sample %	Population %
TTT	0.17	0.125
HTT	0.19	0.125
THT	0.11	0.125
HHT	0.13	0.125
TTH	0.12	0.125
HTH	0.11	0.125
THH	0.11	0.125
HHH	0.06	0.125

JOINT PROBABILITY Exercises

- Run JOINT program with $pH=0.50$ and $numCoins=2$ for the following sample sizes: 100, 1000, and 5,000. Complete the table.

RELATIVE FREQUENCY				
EVENT	N = 100	N = 1,000	N = 5,000	PROBABILITY (P)
TT	_____	_____	_____	_____
HT	_____	_____	_____	_____
TH	_____	_____	_____	_____
HH	_____	_____	_____	_____
HEADS	_____	_____	_____	_____
0	_____	_____	_____	_____
1	_____	_____	_____	_____
2	_____	_____	_____	_____

- a. Compare the relative frequency of TT, HT, TH, and HH with the probability of these events. Do the relative frequencies provide a reasonable approximation to the probabilities?

Yes _____ No _____

- b. For which sample size does the relative frequency give the best approximation?

$N=100$ _____ $N=1,000$ _____ $N=5,000$ _____

- c. Under HEADS, a value of 1 gives the joint probability for HT and TH.

$P(1)=P(HT)+P(TH)$ by the addition law. Compute $P(1)$ for each sample size.

$N=100$ _____ $N=1,000$ _____ $N=5,000$ _____

- d. Show that the same is true for the frequency.

$F(1)=F(HT)+F(TH)$ by the addition law. Compute $F(1)$ for each sample size.

Note: $F=P*N$

$N=100$ _____ $N=1,000$ _____ $N=5,000$ _____

2. From the previous table for $N=100$, compute the ERROR by subtracting the proportion from the sample percentage.

Note: $\text{ERROR}=\text{SAMPLE \%}-\text{POPULATION \%}$. Keep the $+/-$ sign for each error.

EVENT	SAMPLE %	POPULATION %	ERROR
TT	_____	_____	_____
HT	_____	_____	_____
TH	_____	_____	_____
HH	_____	_____	_____
HEADS	_____	_____	_____
0	_____	_____	_____
1	_____	_____	_____
2	_____	_____	_____

- a. Is the ERROR under HEADS for a value of 1 related to the errors for HT and TH?

YES _____ NO _____

- b. What is the sum of the ERRORS for the four events? _____

Addition Law of Probability

We will use the computer to simulate the rolling of two dice to compare the relative frequencies of the sums of the numbers on the two dice with corresponding theoretical probabilities. This will show how the theoretical probabilities for the sums are computed.

We will use the addition law of probability to find the probability of an even sum and the law of complements to find the probability of an odd sum.

Probability can help determine the odds of events occurring in practice. For example, a deck of cards contains 52 cards. A deck of cards has four suits (Hearts, Diamonds, Spades, and Clubs). Therefore each suit has 13 cards ($4 \times 13 = 52$). The probability of selecting any Heart from a deck of cards would be $13/52 = 0.25$. This would be the same probability for selecting any Diamond, Spade, or Club, assuming selection with replacement of the card each time. Similarly, there are four Kings (one in each suit). The probability of selecting a King out of a deck of cards would be $4/52 = 0.076923$.

The sample space for the sum of the numbers on the two dice can be conceptualized as follows:

		FIRST DIE					
		1	2	3	4	5	6
SECOND DIE		1	2	3	4	5	6
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

If the dice are unbiased, all of the 36 outcomes are equally likely. The probability of any sum, S , can be calculated theoretically by the formula: $P(S) = (\text{Number of ways } S \text{ can occur})/36$. For example, $P(7) = 6/36 = 1/6$ (a number 7 occurs in the diagonal six times).

The theory of probability relates to possible outcomes of events occurring in a sample space. The relative frequencies for the different sums are not readily apparent. Our earlier approach assumed that all events were equally likely. In the dice example, we discover that the sums have 36 outcomes, which are equally likely outcomes, but certain sums occur more frequently (e.g., sum=6). The theory of probability helps us to understand the frequency of outcomes and apply this in practice.

The theoretical probabilities for the sums of the numbers on two dice agree well with what happens in practice. The theoretical probabilities for the sums can be found by listing all of the outcomes in a two-way table and using the “equally likely” definition of probability; for the sum S , $P(S) = (\text{Number of ways } S \text{ can occur})/36$. The relative frequencies of the sums get very close to the theoretical

probabilities given large sample sizes. The probability of an odd sum can be found from the probability of an even sum by using the **law of complements**: $P(\text{ODD})=1-P(\text{EVEN})$.

ADDITION R Program

The ADDITION R program simulates the tossing of two dice. It records the number of times that each of the possible sums of the two dice occurs, and then changes these counts into relative frequencies. The relative frequencies of each sum for each sample size, along with the theoretical probability, are printed. Since the computation of the theoretical probabilities depends on the “equally likely” definition of probability, the exercises illustrate how the definition of probability is reasonable and does reflect practice. The different events are examined simultaneously, but are independent.

The program starts with a vector of sample sizes and then creates a vector of probabilities that correspond to the chances of obtaining a 2 through 12 from rolling two dice. The *DiceFreq* object is set to **NULL** so that it may be used to build a matrix within the main processing loop. The loop iterates through the values in *SampleSizes* and obtains a random sample of values from 1 to 6 of size *SampleSize* for *Die1* and then repeats the process for *Die2*. The two vectors of simulated rolls are summed together to obtain a vector for the total of both dice. [Note: the same vector could have been obtained by removing the *Die1* and *Die2* variables and just typing *DiceSum <- sample(2:12,size=SampleSize,replace=T)*, but that hides the fact that we have two independent events and destroys the chance to analyze specific dice combinations.] The relative frequency of each outcome (2 through 12) is appended to the *DiceFreq* matrix for each different sample size as the loop continues through the values of the *SampleSizes* vector. Finally, the *outputMatrix* is built from the *DiceFreq* matrix, **cbind** is used in the *Probs* vector to yield a matrix with relative frequencies for each value outcome of the dice, for each sample size, along with the theoretical probabilities of obtaining each value outcome. The **print** function is used to output the information. Run the ADDITION program using various sample sizes to see how closely you can approximate the theoretical probabilities.

ADDITION Program Output

	N= 100	N= 500	N= 1000	N= 5000	Prob.
2	0.04	0.034	0.036	0.0290	0.0278
3	0.06	0.064	0.068	0.0580	0.0556
4	0.08	0.072	0.078	0.0798	0.0833

5	0.13	0.112	0.092	0.1082	0.1111
6	0.08	0.142	0.130	0.1462	0.1389
7	0.12	0.188	0.172	0.1622	0.1667
8	0.13	0.116	0.145	0.1450	0.1389
9	0.14	0.114	0.127	0.1162	0.1111
10	0.15	0.088	0.085	0.0790	0.0833
11	0.04	0.042	0.049	0.0538	0.0556
12	0.03	0.028	0.018	0.0226	0.0278

ADDITION Law Exercises

1. Run ADDITION for the sample sizes indicated below. Complete the table.

RELATIVE FREQUENCY				
SUM	N = 360	N = 1,200	N = 7,200	PROBABILITY
2	_____	_____	_____	_____
3	_____	_____	_____	_____
4	_____	_____	_____	_____
5	_____	_____	_____	_____
6	_____	_____	_____	_____
7	_____	_____	_____	_____
8	_____	_____	_____	_____
9	_____	_____	_____	_____
10	_____	_____	_____	_____
11	_____	_____	_____	_____
12	_____	_____	_____	_____

- a. Check that the probabilities listed correspond to values in the sequence $1/36$, $2/36$, $3/36$, $4/36$, $5/36$, $6/36$, $5/36$, $4/36$, $3/36$, $2/36$, and $1/36$ (sum = 1 within rounding).
- b. Which sample size provides the best estimate of the probabilities?
-
2. The addition law for mutually exclusive events states that the sum of relative frequencies for *even* numbers should be about 50% and the sum of relative frequencies for *odd* numbers should be about 50%.
- a. For N = 7,200 above, add the relative frequency for all even sums (2, 4, 6, 8, 10, 12) and the relative frequency for all odd sums (1, 3, 5, 7, 9, 11). Enter the two relative frequencies in the table below.

SUM	FREQUENCY	PROBABILITY
EVEN		0.50
ODD		0.50

- b. Why do you expect these frequencies to be around 50%?
-
-

3. Using the sum of frequencies for all *even* numbers and all *odd* numbers, answer the following questions.

- a. How can the probability of all *odd* numbers be obtained from the probability of all *even* numbers?
-
-

- b. What is the name of this probability law?
-
-

Multiplication Law of Probability

One of the basic properties of **probability** is the multiplication law for independent events. For example, if two dice are tossed and the events A, B, and C occur as follows:

- A: An odd number on the first die
- B: An odd number on the second die
- C: An odd number on both dice

then the **multiplication law** for independent events states that: $P(\text{Event C}) = P(\text{Event A}) \times P(\text{Event B})$.

This multiplication law of probability reflects the properties of relative frequency in practice. If two dice are tossed a large number of times, the relative frequencies of events A, B, and C should approximate the probabilities of these events. Also, the product of the relative frequency of A times the relative frequency of B should approximate the relative frequency of C. This can be stated as: $\text{RelativeFrequency}(\text{Event C}) \approx \text{RelativeFrequency}(\text{Event A}) \times \text{RelativeFrequency}(\text{Event B})$. The multiplication law for independent events states that if two events A and B are independent, then $P(A \text{ and } B) = P(A) \times P(B)$. The multiplication law for independent events is modeled on the behavior of relative frequency. For relative frequency, $\text{RelativeFrequency}(A \text{ and } B)$ is approximately $\text{RelativeFrequency}(A) \times \text{RelativeFrequency}(B)$. As sample size increases, $\text{RelativeFrequency}(A \text{ and } B)$ tends to be closer to $\text{RelativeFrequency}(A) \times \text{RelativeFrequency}(B)$.

MULTIPLICATON R Program

The MULTIPLICATION R program simulates the tossing of two dice and records the frequency of an odd number on the first die, an odd number on the second die, and an odd number on both of the dice. The frequencies are then changed to relative frequencies, and the results are rounded to three decimal places. The program inputs the number of times the two dice are tossed. The program illustrates a comparison between relative frequency and the multiplication law for the probability of independent events. Since sample size can be changed, the effect of sample size on the relative frequency as it relates to the multiplication law can be observed.

The program is a modification of the ADDITION R program. The program reflects how the probability of both dice ending up odd relates to the probability of either of the dice being odd. It begins with a vector of sample sizes, creates a **NULL** object to become a matrix, builds a processing loop to iterate through the values of *SampleSizes*, and simulates the rolling of two dice. The next two lines calculate the relative frequency of odds in the first die and the relative frequency of odds in the second die. It does this using a modulo operator (`% %`). The modulo operator performs an integer division and returns only the remainder portion. This means that `13 % 5` would equal 3, because 13 divided by 5 equals 2 with remainder 3. Performing a modulo operation with a 2 on each die result would give values of 0 for even numbers and 1 for odd numbers. The odd numbers are counted up using the **sum** function and then divided by the sample size to get the relative frequency of odds in the sample. Next, the modulo 2 result for both dice are added together, so if both were odd then the result would be 2, otherwise it would be 0 or 1. The integer number is divided by 2 (`/%`), which only returns a whole number. This would result in 0 for values of 0 or 1, because 2 doesn't go into either of those numbers evenly, but would be 1 for a value of 2. In this way, the rolls in which both dice come up odd are added together and divided by the sample size to give the relative frequency of both being odd in the sample. These three values are **rbinded** into *outputMatrix* along with the difference of the relative frequency of both being odd and the product of the relative frequencies of each being odd. After the loop is completed, dimension names are assigned to the rows and columns of the matrix and it is printed using the **print(matrix)** function. Different sample sizes can be input to see the effect sample size has on the relative frequencies and to observe what sample sizes are required to reduce the error to a minimum.

MULTIPLICATION Program Output

Given the following sample sizes: `SampleSizes <- c(100,500,1000)`

	1st Odd	2nd Odd	Both Odd	F1 * F2	Error
N= 100	0.480	0.460	0.210	0.221	-0.011
N= 500	0.504	0.516	0.260	0.260	0.000
N= 1000	0.513	0.510	0.259	0.262	-0.003

Given the following sample sizes: SampleSizes <- c(1000,2000,3000)

	1st Odd	2nd Odd	Both Odd	F1*F2	Error
N= 1000	0.491	0.502	0.251	0.246	0.005
N= 2000	0.486	0.493	0.230	0.240	-0.009
N= 3000	0.486	0.495	0.241	0.241	0.000

Multiplication Law Exercises

1. Run MULTIPLICATION for samples sizes 100, 500, and 1,000.
Record the results below.

<u>RELATIVE FREQUENCY</u>					
SAMPLE SIZE	FIRST ODD	SECOND ODD	BOTH ODD	RF1*RF2	ERROR
N=100					
N=500					
N=1,000					

- a. What is the theoretical probability that the first die will be odd? _____
- b. What is the theoretical probability that the second die will be odd? _____
- c. What is the theoretical probability that both dice are odd? _____
- d. What law of probability are you using to find the probability that both are odd? _____
- e. What effect does sample size have on the sample approximations? _____
- _____
- _____

- f. Compute the error in this approximation by: ERROR=BOTH ODD-(F1×F2)

Do all of the differences have the same sign?

YES _____ NO _____

- g. Does sample size have an effect on the amount of error?

YES _____ NO _____

2. Run MULTIPLICATION for samples sizes 1,000, 2,000, and 3,000.
Record the results below.

<hr/> RELATIVE FREQUENCY <hr/>					
SAMPLE SIZE	FIRST ODD	SECOND ODD	BOTH ODD	RF1*RF2	ERROR
N=1,000					
N=2,000					
N=3,000					

a. For N=3,000, what is the relative frequency that the first die will be odd?

b. For N=3,000, what is the relative frequency that the second die will be odd?

c. For N=3,000, verify that F1*F2 is correct.

d. Why is the relative frequency of BOTH ODD different from your answer in 2c? _____

e. Do all of the error terms have the same sign?

YES _____ NO _____

f. Does sample size have an effect on the amount of error?

YES _____ NO _____

Conditional Probability

A child has a toy train that requires six “C” batteries to run. The child has accidentally mixed four good batteries with two bad batteries. If we were to randomly select two of the six batteries without replacement, the odds of getting a bad battery are conditionally determined. Let’s assume that the first battery chosen is bad (Event A) and the second battery chosen is good (Event B). The two selections of the two batteries are dependent events. The probability of event B has two different values depending upon whether or not event A occurs. If event A occurs, then there are four good batteries among the remaining five batteries, and the probability of event B is 4/5. If a battery is chosen and event A does not occur, then there are only three good batteries remaining among the five batteries, and the probability of B is 3/5.

In probability terms this can be represented by: $P(B|A)=4/5$ and $P(B|\text{not-}A)=3/5$. These terms are read, “the probability of B given A” and “the probability of B given not-A”, respectively. Probabilities of this type are called **conditional probabilities** because the probability of B is conditional upon the occurrence or nonoccurrence of A. Conditional probabilities are related to joint probabilities and marginal probabilities. This relationship can be illustrated by the following example. Consider a sample space that contains all pairs of batteries selected from the six batteries without replacement. The X’s in the table below indicate the 30 possible outcomes.

		SECOND BATTERY					
		DEAD1	DEAD2	GOOD1	GOOD2	GOOD3	GOOD4
FIRST BATTERY	DEAD1	X	X	X	X	X	X
	DEAD2	X		X	X	X	X
	GOOD1	X	X		X	X	X
	GOOD2	X	X	X		X	X
	GOOD3	X	X	X	X		X
	GOOD4	X	X	X	X	X	

Since these 30 outcomes are equally likely, the **joint probabilities**, $P(A \text{ and } B)$, can be summarized in the following table.

		JOINT PROBABILITIES SECOND BATTERY		
		DEAD	GOOD	MARGINAL PROBABILITY
FIRST BATTERY	DEAD	2/30	8/30	10/30
	GOOD	8/30	12/30	20/30
MARGINAL PROBABILITY		10/30	20/30	30/30 (Total)

The row totals are the **marginal probabilities** for the first battery:

$$P(\text{First is dead}) = 10/30$$

$$P(\text{First is good}) = 20/30.$$

The column totals are the **marginal probabilities** for the second battery:

$$P(\text{Second is dead}) = 10/30$$

$$P(\text{Second is good}) = 20/30.$$

Conditional probabilities are related to these joint probabilities and marginal probabilities by the following formula:

$$P(B|A) = P(B \text{ and } A)/P(A).$$

If event A results in a bad battery and event B results in a good battery, then

$$\begin{aligned} P(B|A) &= P(\text{Second is good} \mid \text{First is dead}) \\ &= P(\text{Second is good and First is dead}) / P(\text{First is dead}) \\ &= (4/15)/(5/15) \\ &= 4/5. \end{aligned}$$

These conditional probabilities are theoretical probabilities assigned to the events by making use of the definition of probability for **equally likely events** (it is assumed that each of the batteries and each of the pairs are equally likely to be chosen). If these conditional probabilities are reasonable, they should reflect what happens in practice. Consequently, given a large number of replications in which two batteries are selected from a group of six batteries (in which two of the batteries are dead),

the relative frequencies for the conditional events should approximate the theoretical conditional probabilities.

The conditional probability of B given A is the joint probability of A and B divided by the marginal probability of A, if $P(A) \neq 0$. The conditional probability agrees with the behavior of relative frequency for conditional events. For large sample sizes, the relative frequency of a conditional event is a good approximation of the conditional probability.

CONDITIONAL R Program

The CONDITIONAL R program simulates random selection without replacement of two batteries from a group of six in which two of the batteries are dead. The number of replications can be specified in the program. The relative frequencies from a small number of replications will not provide good approximations of the probabilities. A large number of replications should provide relative frequencies that will be very close to the theoretical probabilities. The program will permit you to observe that the theoretical rules of probability for conditional events do in fact reflect practice. The theoretical probabilities are important in statistics because they provide reasonable rules to adopt for analyzing what happens in the real world.

The CONDITIONAL R program determines the probability of conditional events by selecting two batteries from a group of batteries with a certain number of good batteries and a certain number of bad batteries. The total number of batteries is assigned, followed by the number of bad batteries and the number of times the two selections should be replicated. The number of good batteries is determined by subtraction of the number of bad batteries from the total number of batteries. The probabilities of the possible event outcomes are then determined and assigned to the variables: pGG , pBB , pGB , and pBG .

The total number of batteries is defined so that sampling can occur from a finite population. The `rep` function creates *numGood* (number of 0s) and the `c` function concatenates those with *numBad* (number of 1s) to complete the population of batteries. The receiving objects *FirstBattery* and *SecondBattery* are set to **NULL** before the main processing loop begins. The loop takes a sample of two batteries WITHOUT replacement, since these are conditional events. The two batteries are then added to their respective vectors. They are also added to an *eventList* vector using the same type of binary encoding scheme presented in earlier chapters. The encoding is much simpler since the number of picks is fixed at two.

After the processing loop is finished, output matrices are created. The *eventTable* vector is built from the *eventList* vector factored for all the possible coded event values from 0 to 3. Two vectors of values are then created and put into the matrices for even column spacing in the output. The first vector is moved into a 4 by 4 matrix for display with no **dimnames** set, since the column and row headers were included within the vector. The same thing is done with the second vector, only it is printed out using **print(matrix)** command, since it doesn't have multiple column or row headers. The program permits different numbers of total and bad batteries, as well as different sample sizes.

CONDITIONAL R Program Output

Given the following values:

```
numBatteries <- 6
```

```
numBad <- 2
```

```
SampleSize <- 1000
```

		Second	Battery
		Bad	Good
First	Bad	0.061	0.242
Battery	Good	0.277	0.42
No.	Bad	Rel Freq	Probability
0		0.42	0.4
1		0.519	0.534
2		0.061	0.067

Given the following values:

```
numBatteries <- 6
```

```
numBad <- 2
```

```
SampleSize <- 5000
```

		Second	Battery
		Bad	Good
First	Bad	0.071	0.272
Battery	Good	0.271	0.386
No.	Bad	Rel Freq	Probability
0		0.386	0.4
1		0.543	0.534
2		0.071	0.067

CONDITIONAL Probability Exercises

- Run CONDITIONAL for N=1,000 with 6 total batteries and 2 bad batteries.
 - Enter the probabilities of the joint events and the marginal probabilities in the table.

		Second battery		Marginal Probability
		Bad	Good	
First Battery	Bad	_____	_____	
	Good	_____	_____	
Marginal Probability		(Total)		

- b. Do the marginal probabilities indicate that approximately 1/3 (0.33) of the batteries are bad and 2/3 (0.67) of the batteries are good?

YES _____ NO _____

- c. Do the marginal probabilities sum to 1.0?

YES _____ NO _____

2. From the CONDITIONAL program with N=1,000, enter the relative frequencies of 0, 1, and 2 bad batteries.

- a. Compute the error and record it in the table.

ERROR=REL FREQ-PROBABILITY

No. BAD	REL FREQ	PROBABILITY	ERROR
0			
1			
2			

- b. Some of the errors should be positive and others negative.

Do the errors sum to zero (0)? YES _____ NO _____

- c. Do the relative frequencies sum to 1.0? YES _____ NO _____

- d. Do the probabilities sum to 1.0? YES _____ NO _____

3. Run CONDITIONAL for N=5,000 with 6 total batteries and 2 bad batteries.

- a. Enter the probabilities of the joint events and the marginal probabilities in the table.

		Second Battery		Marginal probability (Total)
		Bad	Good	
First Battery	Bad			
	Good			
Marginal probability				

- b. Do the marginal probabilities indicate that approximately 1/3 (0.33) of the batteries are bad and 2/3 (0.67) of the batteries are good?

YES _____ NO _____

- c. Do the marginal probabilities sum to 1.0?

YES _____ NO _____

4. From the CONDITIONAL program with N=5,000, enter the relative frequencies of 0, 1, and 2 bad batteries.

- a. Compute the error and record it in the table.

ERROR=REL FREQ-PROBABILITY

No. BAD	REL FREQ	PROBABILITY	ERROR
0			
1			
2			

- b. Some of the errors should be positive and others negative.

Do the errors sum to zero (0)? YES _____ NO _____

- c. Do the relative frequencies sum to 1.0? YES _____ NO _____

- d. Do the probabilities sum to 1.0? YES _____ NO _____

Combinations and Permutations

Probability theory helps us to determine characteristics of a population from a random sample. A **random sample** is chosen so that every object, event, or individual in the population has an equal chance of being selected. The probability that the object, event, or individual will be selected is based upon the relative frequency of occurrence of the object, event, or individual in the population. For example, if a population consisted of 1,000 individuals with 700 men and 300 women, then the probability of selecting a male is 700/1,000 or 0.70. The probability of selecting a woman is 300/1,000 or 0.30. The important idea is that the selection of the individual is a chance event.

Probability theory operates under seven fundamental rules. These seven rules can be succinctly stated as:

1. The probability of a single event occurring in a set of equally likely events is one divided by the number of events, i.e., $P(\text{single event}) = 1/N$. For example, a single marble from a set of 100 marbles has a $1/100$ chance of being selected.
2. If there is more than one event in a group, then the probability of selecting an event from the group is equal to the group frequency divided by the number of events, i.e., $P(\text{Group}|\text{single event}) = \text{group frequency}/N$. For example, a set of 100 marbles contains 20 red, 50 green, 20 yellow, and 10 black. The probability of picking a black marble is $10/100$ or $1/10$.
3. The probability of an event ranges between 0 and 1, i.e., there are no negative probabilities and no probabilities greater than one. Probability ranges between 0 and 1 in equally likely chance events, i.e., $0 \leq P(\text{event}) \leq 1$.
4. The sum of the probabilities in a population equal one, i.e., the sum of all frequencies of occurrence equals 1.0, i.e., $\Sigma(\text{Probabilities}) = 1$.
5. The probability of an event occurring plus the probability of an event *not* occurring is equal to one. If the probability of selecting a black marble is $1/10$, then the

probability of *not* selecting a black marble is 9/10, i.e., $P+Q=1$ where P =probability of occurrence and $Q=1-P$.

6. The probability that any one event from a set of mutually exclusive events will occur is the sum of the probabilities (addition rule of probability). The probability of selecting a black marble (10/100) *or* a yellow marble (20/100) is the sum of their individual probabilities (30/100 or 3/10), i.e., $P(B \text{ or } Y)=P(B)+P(Y)$.
7. The probability that a combination of independent events will occur is the product of their separate probabilities (multiplication rule of probability). Assuming sampling with replacement, the probability that a yellow marble will be selected the first time (2/10) and the probability that a yellow marble will be selected the second time (2/10) combine by multiplication to produce the probability of getting a yellow marble on both selections ($2/10 \times 2/10 = 4/100$ or 0.04), i.e., $P(Y \text{ and } Y)=P(Y)*P(Y)$.

Factorial notation is useful for designating probability when samples are taken *without* replacement. For example, a corporate executive officer (CEO) must rank the top five department managers according to their sales productivity. After ranking the first manager, only four managers are remaining to choose from. After ranking the second manager, only three managers remain, and so forth, until only one manager remains. If the CEO selects managers at random, then the probability of any particular manager order is: $1/5*1/4*1/3*1/2*1/1$, or 1/120.

The probability is based upon the total number of possible ways the five managers could be ranked by the CEO. This is based upon the number of managers in the company available to select from, which changes each time. Consequently, the product yields the total number of choices available: $5*4*3*2*1 = 120$. This product is referred to as **factoring** and uses **factorial notation** to reflect the product multiplication, i.e., $n!$. The factorial notation, $3!$ (read 3-factorial), would imply, $3*2*1$, or 6, which indicates the number of different ways three things could be ordered. Imagine a restaurant that serves hamburgers with the following toppings: pickle, onion, and tomato. How many different ways could you order these ingredients on top of your hamburger?

Permutations involve selecting objects, events, or individuals from a group and then determining the number of different ways they can be ordered. The number of permutations (different ways you can order something) is designated as **n** objects taken **x** at a time, or:

$$P(n, x) = \frac{n!}{(n - x)!}$$

For example, if a teacher needed to select three students from a group of five and order them according to mathematics ability, the number of *permutations* (or different ways three out of five students could be selected and ranked) would be:

$$P(n, x) = \frac{n!}{(n-x)!} = \frac{5!}{(5-3)!} = \frac{5*4*3*2*1}{2*1} = 60$$

Probability can also be based upon the number of *combinations* possible when choosing a certain number of objects, events, or individuals from a group. The ordering of observations (permutations) is not important when determining the number of combinations. For example, a teacher must only choose the three best students with mathematics ability from a group of five (no ordering occurs). The number of possible combinations of three students out of a group of five is designated as:

$$P(n, x) = \frac{n!}{x!(n-x)!} \text{ or } \frac{5!}{3!(5-3)!} = 10$$

The number of possible combinations can be illustrated by determining the number of students in a classroom that have their birthday on the same day. This classic example can be used in a class to determine the probability that two students have the same birthday. The probability of two students having a common birthday, given five students in a class, can be estimated as follows (assuming 365 days per year and equally likely chance):

$$P(2|5) = 1 - \frac{365}{365} * \frac{364}{365} * \frac{363}{365} * \frac{362}{365} * \frac{361}{365} = 0.027$$

The numerator decreases by one because as each student's birthday is selected, there is one less day available.

The probability of at least two students out of five *not* having the same birthday is $1 - P$ (see probability rule 5). The probability of *no* students having a birthday in common for a class of five students is computed as:

$$P(\text{No2}|5) = \frac{365}{365} * \frac{364}{365} * \frac{363}{365} * \frac{362}{365} * \frac{361}{365} = 0.973$$

Therefore, the probability of at least two students having the same birthday is the complement of *no* students having the same birthday, or $P(2|5) = 1 - 0.973 = 0.027$. The formula clearly indicates that this probability would increase quickly as the number of objects, events, or individuals in the group increases.

The seven rules of probability apply to everyday occurrences. The number of possible outcomes of independent events is designated as a factorial ($n!$). Permutations refer to the number of possible ways to order things when selected from a group (**n** objects, **x** order).

Combinations refer to the number of possible sub-groups of a given size from a larger group (**n** objects, **x** size). The birthday problem is a classic example of how to

determine whether two individuals in a group of size N have the same birthdays. The relative frequencies produced by a simulation of the birthday problem are good approximations of the actual probabilities. The *accuracy* of the relative frequencies as approximations of the actual probabilities in the birthday problem is not affected by the size of the group of people, rather by increasing the number of repetitions in the program.

Combination and Permutation R Program

The Combination and Permutation R program simulates an example for N individuals by using a random number generator and checking for a common birthday. The relative frequency of at least one common birthday is reported. This relative frequency approximates the probability of occurrence in a group of N individuals. The size of the group and the number of replications can be changed.

The program simulates selecting groups of people of various sizes over a given number of replications in order to compute estimates of probabilities. The sizes of the groups of people are assigned to the vector *numPeople*. The *replicationSize* variable represents the number of times that the selection of random birthdays will occur for each group size. In the initial program settings, the probability of five birthdays in common is chosen, and duplication or non-duplication reported for 250 replications. The *numPeople* vector then indicates that 10 birthdays will be chosen at a time for the 250 replications. This is repeated for 20 and 50 birthdays. The sampling and replications can be time-consuming for a computer, so it would be wise *not* to select 10,000 replications for a *numPeople* vector of from 1 to 100, unless you are willing to wait.

The *repeatVector* object is simply a vector containing the number of times there was a common birthday for a given group size. The outer processing loop iterates through the values of group sizes and the inner processing loop defines each replication for the given group size. Within this inner loop, a random sample is taken from the range of values from 1 to 365 with replacement and a sample size of *numPeople*[*i*]. The vector created is run through the **table** function to group the sample points that fell on the same day, and if the **max** of that table is greater than 1, then it means there was at least one birthday in common. If this occurs, then the corresponding value in the vector for the number of replications containing repeated values is increased by one. Because this takes place within the inner-processing loop, it continues for all replications of all group sizes.

Once the simulation in the loops is concluded, the counts within the *repeatVector* are changed into relative frequencies by dividing the *repeatVector* by the replication size. This creates a new vector of values that contains the relative frequencies of replications with birthday duplications. Theoretical probabilities are computed using a small processing loop that iterates through the group sizes and creates a vector of probabilities of duplication for each group size. The single line of code within the loop represents the mathematical notation, $1 - (365/365) * (364/365) * (363/365) \dots ((366 - \text{group size})/365)$.

A **list** object is created to hold the dimension labels for the output matrix. The output matrix is built by concatenating the relative frequency vector, the theoretical probability vector, and the difference between the relative frequency vector and the theoretical probability vector, giving an error vector. The **dimnames** keyword is given the value of the **list** object that was created making the line easier to read. The matrix is output using **print(matrix)command**. Values are reported within three decimal places using the **nsmall** keyword of the **format** function.

Combination and Permutation Program Output

Given these values:

```
numPeople <- c(5,10,20,50)
replicationSize <- 250
```

	Rel. Freq.	Common	Birthday	Error
N= 5	0.032	0.027		0.005
N= 10	0.104	0.117		-0.013
N= 20	0.428	0.411		0.017
N= 50	0.988	0.970		0.018

Given these values:

```
numPeople <- c(5,10,20,50)
replicationSize <- 500
```

	Rel. Freq.	Common	Birthday	Error
N= 5	0.040	0.027		0.013
N= 10	0.118	0.117		0.001
N= 20	0.418	0.411		0.007
N= 50	0.970	0.970		0.000

Combination and Permutation Exercises

- Run BIRTHDAY for the following sample sizes and complete the table.

GROUP SIZE	REL. FREQ.	COMMON BIRTHDAY	ERROR
N=5	_____	_____	_____
N=10	_____	_____	_____
N=15	_____	_____	_____
N=20	_____	_____	_____

- a. As the size of the group increases, does the probability of a common birthday

GROUP SIZE	REL. FREQ.	COMMON BIRTHDAY	ERROR
N=5	_____	_____	_____
N=10	_____	_____	_____
N=15	_____	_____	_____
N=20	_____	_____	_____

increase?

YES _____ NO _____

- b. As the size of the group increases, do the relative frequencies more closely approximate the common birthday probabilities? Hint: Does error decrease?

YES _____ NO _____

GROUP SIZE	REL. FREQ.	COMMON BIRTHDAY	ERROR
N=10	_____	_____	_____
N=20	_____	_____	_____
N=30	_____	_____	_____
N=40	_____	_____	_____
N=50	_____	_____	_____

2. Run BIRTHDAY again for the same sample sizes. Complete the table.

- a. Are the common birthday probabilities the same?

YES _____ NO _____

- b. Are the relative frequencies close to the common birthday probabilities?

YES _____ NO _____

3. Run BIRTHDAY again using sample sizes listed below with 500 replications. Complete the table.

- a. As the size of the group increases, does the probability of a common birthday increase?

YES _____ NO _____

- b. As the size of the group increases, do the relative frequencies more closely approximate the common birthday probabilities? Hint: Does error decrease?

YES _____ NO _____

True or False Questions

Finite and Infinite Probability

- T F a. As additional trials are conducted, the relative frequency of heads is always closer to 0.5 than for any previous smaller sample size.
- T F b. As sample size increases, the relative frequency of an event approaches a fixed value.
- T F c. The relative frequency of an event with probability of 0.65 stabilizes faster than an event with probability of 0.10.
- T F d. In understanding probability, it is assumed that the relative frequencies approach a fixed number as the sample size increases because this corresponds to our experience of the real world.
- T F e. The relative frequency of an event in one hundred trials is the probability of the event.

Joint Probability

- T F a. The addition and multiplication laws of probability are reasonable because these properties are true for frequencies.
- T F b. If two events are independent, the addition law is used to find their joint probability.
- T F c. The sum of the probabilities for all of the events in a sample space is 1.
- T F d. $P(1)=P(HT)+P(TH)$ because HT and TH are independent events.
- T F e. $F(HT) \approx P(H)*P(T)$ because H and T are independent events.

Addition Law of Probability

- T F a. Probabilities would be the same if the dice were biased.
- T F b. Since there are 12 distinct sums and 6 of them are even, the probability of an even sum is 6/12.
- T F c. The stabilizing property of relative frequencies is true for a group of outcomes as well as for a single outcome.
- T F d. Large numbers of repetitions will provide good estimates of probabilities.
- T F e. Each time the program is run for $N=1,200$, the relative frequencies will be the same.

Multiplication Law of Probability

- T F a. If two events are independent, then the probability of both events occurring is the product of their probabilities.
- T F b. If two events are independent, then the relative frequency of both events occurring is the product of their relative frequencies.
- T F c. In general, relative frequencies obtained from small samples give the best approximations of probabilities.
- T F d. If two biased dice were tossed, then the events FIRST ODD and SECOND ODD are not independent.
- T F e. The events FIRST ODD and BOTH ODD are mutually exclusive events.

Conditional Probability

- T F a. If two batteries are selected with replacement from a group of six batteries, in which two of the batteries are bad, the FIRST BAD and the SECOND GOOD are dependent events.
- T F b. $P(A \text{ and } B) = P(A) \times P(B|A)$
- T F c. If the probability of event A is not affected by whether or not event B occurs, then A and B are independent events.
- T F d. $P(A)$, the marginal probability for event A, is equal to the sum of the joint probabilities of A and all other events that can occur with A.
- T F e. If two events A and B are independent, then $P(B|A) = P(B)$.

Combination and Permutation

- T F a. As the size of the group increases, the probability decreases that two people have the same birthday.
- T F b. The probability of *no* common birthday is the complement of the probability of having a common birthday.
- T F c. If a group consists of only two people, the probability that they have the same birthday is $1/365$.
- T F d. In a group of 50 people, there will *always* be at least two people with the same birthday.
- T F e. The error in the relative frequency as an approximation of the probability is reduced for large groups of people.

Chapter 3

Statistical Theory

Sample Versus Population

The field of statistics uses numerical information obtained from samples to draw inferences about populations. A **population** is a well-defined set of individuals, events, or objects. A **sample** is a selection of individuals, events, or objects taken from a well-defined population. A sample is generally taken from a population with each individual, event, or object being independent and having an equally likely chance of selection. The sample average is an example of a random sample estimate of a population value, i.e., population mean. Population characteristics or **parameters** are inferred from sample estimates, which are called statistics. Examples of population parameters are population proportion, population mean, and population correlation. For example, a student wants to estimate the proportion of teachers in the state who are in favor of year-round school. The student might make the estimate on the basis of information received from a random sample of 500 teachers in the population comprised of all teachers in the state. In another example, a biologist wants to estimate the proportion of tree seeds that will germinate. The biologist plants 1,000 tree seeds and uses the germination rate to establish the rate for all seeds. In marketing research, the proportion of 1,000 randomly sampled consumers who buy one product rather than another helps advertising executives determine product appeal.

Because a sample is only a part of the population, how can the sample estimate accurately reflect the population characteristic? There is an expectation that the sample estimate will be close to the population value if the sample is representative of the population. The difference between the sample estimate and the population value is called **sample error**. In a random sample, all objects have an equal chance of being selected from the population. If the sample is reasonably large, this equally likely chance of any individual, event, or object being selected makes it likely that the random sample will represent the population well. Most statistics are based upon this concept of random sampling from a well-defined population. **Sampling error**, or the error in using a sample estimate as a population estimate,

does occur. In future chapters, you will learn that several random sample estimates can be averaged to better approximate a population value, although sampling error is still present.

The Gallop Poll, for example, uses a random sample of 1,500 people nationwide to estimate the presidential election outcome within $\pm 2\%$ error of estimation. This chapter will help you understand random sampling and how the sampling error of estimation is determined, i.e., difference between the sample statistic and the population parameter. The computer will be used to generate random samples of data. The difference between a known population value and a random sample value can be observed. Different samples may lead to different sample estimates of population parameters, so most estimates from samples contain some error of estimation or sampling error. Basically, a random sample is part of a population. Random samples are used to draw inferences about population parameters or characteristics. Different random samples lead to different sample estimates. The estimate from a sample is usually not equal to the population value. If a large sample is taken, the standard error is smaller.

STATISTICS R Program

The STATISTICS R program simulates the sampling of data from a population. You are to determine what proportion of a certain large population of people favor stricter penalties. A random number generator will determine the responses of the people in the sample. A random number generator uses an initial start number to begin data selection, and then uses the computer to generate other numbers at random. You will use the results of these simulated random samples to draw conclusions about the population.

The number of runs will be specified by the user. Each time you run the program, the true population proportion will be different. Consequently, each time you will be simulating a sample data set from a different population. The different random samples will be chosen during each computer run. The random samples have various sample sizes determined by the user (i.e., 5, 20, 135, 535, and 1,280, for 5 samples). Using these results, you will be able to observe the effect of sample size on the accuracy of estimation.

The STATISTICS program uses a pseudo-random number generator to select a random number between 0 and 1 for the true proportion. Next, random samples of only 0 or 1 are drawn from the finite population (not values between 0 and 1). The probability of a 0 is (1—the population proportion) and the probability of a 1 is the same as the population proportion. Random samples of various sizes are taken and the sample proportion and estimation errors are calculated.

The size of the samples varies by the list of sample sizes in the variable *SampleSizes*. The “`<-`” operator is an assignment operator that places the vector of values (10,100,500,1000,1500) into the variable *SampleSizes*. The `c` before the parentheses means to concatenate these values into a single vector. *NumSamples* is assigned the **length** of the *SampleSizes* vector, which is equivalent to the number of individual

sample sizes. *PopProp* is the true proportion in the population and is obtained by taking one random number from a uniform population that is between the values of 0 and 1. The **runif** command means to take a value from a random **uniform** population and the number values (1,0,1) correspond to the number of values to be obtained (1), the bottom of the range of values (0), and the top of the range of values (1). There are several other commands within R that allow for sampling from other distributions, such as normal (**rnorm**), binomial (**rbinom**), and exponential (**rexp**).

The most complex parts of the program pertain to creating and using matrices, which will be covered in later chapters. The line which begins with *TrialData <- matrix* is setting the size of the matrix and associating labels with the values that will be written to it. The **for** statement begins the processing loop. The **for** command assigns to the variable **SampleSize** successive values listed in the **SampleSizes** vector. The parentheses are used to mark the beginning and end of the loop encapsulated by this **for** command. The first line within the processing loop creates a vector of values for the first sample and assigns it to the temporary variable *SampleData*. The **sample** command is a way to sample from a finite population, in this case either 0 or 1, but it can also be very useful for taking a subsample of larger samples. The 0:1 denotes the range of *integer* values between 0:1, which only includes 0 and 1, but the same notation could be used to create a vector of integer values from 1 to 10 (1:10). The **prob** keyword sets the probability of getting each value, with 0 having a probability of 1 minus the population proportion (1-*PopProp*) and 1 having a probability of the population proportion. The **size=SampleSize** assures that this sample is the same size as the one corresponding to the loop iteration, and **replace=T** means to replace values that have been chosen from the population, so this is sampling WITH replacement. If taking a subsample of a larger population, you can request sampling WITHOUT replacement (**replace=F**).

The next line sums all the zeros and ones from the sample to get the total number of people who were in favor, and then divides that value by the sample size to get the sample proportion. The next to the last line within the processing loop assigns values to one vector within the matrix built earlier in the program for outputting the data. The *i <- i+1* line increments the counter used to keep track of the place within the matrix. The last line of the program produces a printout of the matrix. The values in the *SampleSizes* vector can be changed to simulate small or large sample sizes.

STATISTICS Program Output

Given these values: *SampleSizes <- c(10,100,500,1000,1500)*

	Size	No.in Favor	Sample Prop.	True Prop.	Est.	Error
Sample 1	10	6	0.600	0.677	0.677	0.077
Sample 2	100	59	0.590	0.677	0.677	0.087
Sample 3	500	331	0.662	0.677	0.677	0.015
Sample 4	1000	670	0.670	0.677	0.677	0.007
Sample 5	1500	1037	0.691	0.677	0.677	-0.014

Given these values: SampleSizes<- c(20,200,1000,2000,3000)

	Size	No.in Favor	Sample Prop.	True Prop.	Est.	Error
Sample 1	20	1	0.050	0.098	0.048	
Sample 2	200	18	0.090	0.098	0.008	
Sample 3	1000	110	0.110	0.098	-0.012	
Sample 4	2000	194	0.097	0.098	0.001	
Sample 5	3000	303	0.101	0.098	-0.003	

Statistics Exercises

1. Run STATISTICS once to obtain the results of people who are in favor of stricter penalties for criminals using the four sample sizes below. Enter the results here.

SAMPLE	SAMPLE SIZE	NO. IN FAVOR	SAMPLE PROPORTION
A	5	_____	_____
B	20	_____	_____
C	135	_____	_____
D	1280	_____	_____

- a. Verify that the sample proportions are correct by using long division or a calculator. To find the sample proportion from the number in favor and the sample size, use the formula:

$$\text{SAMPLE PROPORTION} = (\text{NO. IN FAVOR}) \div (\text{SAMPLE SIZE})$$

SAMPLE	COMPUTATION
A	_____
B	_____
C	_____
D	_____

- b. Is the estimate of the population proportion the same for each of the samples?

Yes No

- c. Why do you think the sample proportions change?

d. What is the true population proportion? _____.

2. The sample proportion (EST) is an estimate of the true population proportion (P). There are errors in the sample estimates.

- a. Calculate the error for each sample:

$$\text{ERROR} = \text{EST} - \text{P}$$

Some of the errors may be positive or negative (Record the $+$ / $-$ sign with the error).

SAMPLE PROPORTION

SAMPLE	SAMPLE SIZE	SAMPLE	TRUE	ERROR
A	5	_____	_____	_____
B	20	_____	_____	_____
C	135	_____	_____	_____
D	1,280	_____	_____	_____

- b. Which of the four samples gave the best estimate? _____
3. Run the STATISTICS program 4 more times. Each time, compute the errors in the estimates (P will be different for each program run).

RUN 1				RUN 2			
SAMPLE	SIZE	SAMPLE	ERROR	SAMPLE	SIZE	SAMPLE	ERROR
A	5	_____	_____	A	5	_____	_____
B	20	_____	_____	B	20	_____	_____
C	135	_____	_____	C	135	_____	_____
D	1,280	_____	_____	D	1,280	_____	_____
TRUE P		TRUE P		TRUE P		TRUE P	

RUN 3				RUN 4			
SAMPLE	SIZE	SAMPLE	ERROR	SAMPLE	SIZE	SAMPLE	ERROR
A	5	_____	_____	A	5	_____	_____
B	20	_____	_____	B	20	_____	_____
C	135	_____	_____	C	135	_____	_____
D	1,280	_____	_____	D	1,280	_____	_____
TRUE P		TRUE P		TRUE P		TRUE P	

- a. For the four program runs, what was the largest and smallest amount of error for each sample size? (Disregard the plus or minus sign.)

SAMPLE	SIZE	LARGEST ERROR	SMALLEST ERROR
A	5	_____	_____
B	20	_____	_____
C	135	_____	_____
D	1,280	_____	_____

- b. Was the sample proportion from a smaller sample ever a better estimate of the population proportion than the sample proportion from a larger sample?
Yes ____ No ____.
- c. If yes, for which runs (1, 2, 3 or 4) were the errors smaller?
RUN(S) WITH SMALLER SAMPLE ERRORS SMALLER _____
- d. Why is it possible for a smaller sample to occasionally give a better estimate than a larger sample? _____

4. Use the previous exercises to draw a conclusion about the effect of sample size on the estimate of the population proportion.
-
-

5. A newspaper survey indicates that 62% of the people in a certain state favor a bill to allow retail stores to be open on Sunday. Given the examples you just completed, what additional information would help you interpret this report?
-
-

Generating Random Numbers

Random numbers are used in statistics to investigate the characteristics of different population distributions. We will only be studying the characteristics of the normal distribution. This is because many of the variables that we measure are normally distributed. The statistics we use to test hypotheses about population characteristics based on random samples are created based on certain assumptions and characteristics of the normal distribution. Other population distributions exist (wiebull, hypergeometric, poisson, and elliptical), but we will not be studying their characteristics and associated statistics in the chapter exercises.

Early tables of random numbers helped gamblers to understand their odds of winning. In some instances, exact probabilities or odds of certain outcomes were generated from cards and dice. Today, high-speed computers using computer software with a numerical procedure (algorithm) can produce tables of random numbers. The first mainframe computer, a UNIVAC, produced a set of one million random numbers, which was published in a book by the Rand McNally Corporation. Personal desktop computers today run software that can generate random numbers.

Although many computers have software (mathematical algorithms) to generate random numbers, the software algorithms are not all the same and do not produce the same set of random numbers. Basically, a set of computer-generated numbers is not truly random, so they are called “**pseudo random numbers**.” They are called “pseudo random numbers” because the numbers tend to repeat themselves after awhile (repeatedness), correlate with other numbers generated (correlatedness), and don’t produce a normal distribution (normality). Consequently, when using a **random number generator**, it is important to report the type of computer used, type of random number generator software (algorithm), start value (start number), repeatedness (when numbers repeat themselves in the algorithm), correlatedness (when numbers begin to correlate in a sequence), and normality (whether or not a normal distribution was produced).

A true random set of numbers has no pattern, and if graphed, would appear as scattered data points across the graph. Because true random numbers have no

pattern, the next number generated would not be predicted and would appear with approximately the same frequency as any other number. The concept is simple, but often requires visual confirmation (graph) or other statistical test of randomness and/or normality. Software programs often include statistical tests for testing the randomness and normality of computer-generated random sample data. Because all random number generators are not the same, acceptable properties for these random numbers should include:

1. Approximate, equal proportions of odd and even numbers should occur.
2. Each number between 0 and 9 is generated approximately one-tenth of the time.
3. For the five consecutive sets of generated number combinations, the percentages should be approximately equal to the theoretical probabilities.

Basically, a sequence of random numbers is not truly random (unique). A sequence of random numbers is typically unpredictable, but a long sequence of random numbers will tend to repeat, correlate, and not appear normal. Our expectation is that about half of the generated numbers are odd and half are even. The frequency of occurrence for any random integer between 0 and 9 is approximately one-tenth of the time. A set of randomly generated numbers can be tested for randomness and normality.

RANDOM R Program

The RANDOM R program tests the randomness of numbers from a pseudo-random number generator. The majority of the program code classifies combinations of numbers and formats the output. The creation of sample data and the calculation of the relative frequencies of odd and even digits, and each individual digit, are all contained within the first few lines of the program. Random numbers are sampled from the integer values 0 through 9; the relative frequency of the odd numbers is determined using the modulus (or remainder function) in combination with the **sum** function and dividing by the sample size. The relative frequency of the even numbers is determined in the same manner, only using all values that were not determined to be odd ($\text{SampleSize} - \text{sum}(\text{SampleData} \% \% 2)$). The relative frequency of each digit is determined by the familiar **factor** and **table** combination, and then all raw data are put into groups of five numbers.

The main processing loop of the program is used primarily to classify the groups of numbers based on various combinations of repeat values. It iterates from 1 to the number of rows in the *GroupedData* matrix, which is the first dimension (**dim**) of that matrix. The first line within the matrix concatenates the values within the current row with no space separation between them (**sep=""**). Next, a double use of the **table** and **factor** function combination yields the various amounts of repeat values within the sample group. The loop begins from the inside and works out. First, the number of times that each number (0 to 9) comes up within the group is tallied, then

the outer pair of **factor** and **table** functions count how many of each number of repeats (three of a kind, four of a kind, etc.) are in the group. The next few lines of code use the information contained within the vector just created to classify the different combinations of repeats into unique event values. The event values are fairly arbitrary in this program, unlike earlier programs that used the binary coding scheme, and could really be anything as long as they were matched up with the appropriate labels when they were output. Finally, the last line within the processing loop adds the raw group of numbers to an output vector.

The vectors are factored and tabled to determine how many of each unique event occurred within the sample. The next line builds an output matrix from the relative frequencies that were determined at the beginning of the program, along with their theoretical probabilities, which have been typed directly into the program instead of being calculated. After this, dimension names are assigned for a matrix, a matrix of the event relative frequencies is built, and dimension names are subsequently assigned to that matrix. Finally, the output begins with the groups of numbers from the sample being printed with the **cat** function using the keyword **fill** to assure that lines greater than 80 characters will be wrapped to the next line. Then the two output matrices are printed using the **print** function with a fixed number of decimal places. The **scientific** keyword was used in the second case because there was a problem with some values being represented in scientific notation due to the fact that the default is to print anything with its lead digit more than four places from the decimal in scientific notation. This change increased output to six places. The program allows you to adjust the sample size until you find one that closely approximates the theoretical probabilities.

RANDOM Program Output

Number groups:
 10599 28741 09557 90688 76598 92111 43300 08120 81585 46583 90134 49783

	Relative Frequency	Probability
Odd	0.53	0.50
Even	0.47	0.50
0	0.13	0.10
1	0.13	0.10
2	0.05	0.10
3	0.08	0.10
4	0.08	0.10
5	0.12	0.10
6	0.05	0.10
7	0.07	0.10
8	0.15	0.10
9	0.13	0.10

	Relative Frequency	Probability
No duplicates	0.4167	0.3024

One pair	0.3333	0.5040
One triple	0.0833	0.0720
Two pairs	0.1667	0.1080
Pair & triple	0.0000	0.0090
Four alike	0.0000	0.0045
All alike	0.0000	0.0001

Random Exercises

1. Run RANDOM for N=60. Record the twelve groups of five numbers ($5 * 12 = 60$ numbers) in the blanks below.

Numbers: _____ _____ _____ _____
 _____ _____ _____ _____
 _____ _____ _____ _____

Complete the table below.

	RELATIVE FREQUENCY	PROBABILITY
ODD	_____	_____
EVEN	_____	_____
0	_____	_____
1	_____	_____
2	_____	_____
3	_____	_____
4	_____	_____
5	_____	_____
6	_____	_____
7	_____	_____
8	_____	_____
9	_____	_____

- Check that the relative frequencies are correct for the 60 numbers.
- What is the largest absolute difference between the relative frequencies and the probabilities? _____
- How is the relative frequency for ODD related to the relative frequencies for the ten digits (0–9)? _____

2. Complete the following table from the run of RANDOM for N=60.

	RELATIVE FREQUENCY	PROBABILITY
NONE	_____	_____
ONE PAIR	_____	_____
ONE TRIPLE	_____	_____
TWO PAIRS	_____	_____
PAIR & TRIPLE	_____	_____
FOUR ALIKE	_____	_____
ALL ALIKE	_____	_____

- a. Look at the twelve groups of five numbers recorded in Exercise 1. Have the duplicates been counted correctly and the relative frequencies computed correctly?

RELATIVE FREQUENCY=FREQUENCY/(NUMBER OF GROUPS OF 5)

- b. How is the probability of ALL ALIKE calculated?

- c. Find the sum of the relative frequencies. _____
Why does the sum have this value? _____

3. Run RANDOM for N=200.

- a. What is the maximum absolute value of the differences between the relative frequencies and their respective probabilities? _____
b. What is the maximum absolute difference between the relative frequencies of the duplicates and their respective probabilities? _____

4. Run RANDOM for N=500.

- a. What is the maximum absolute value of the differences between the relative frequencies and their respective probabilities? _____
b. What is the maximum absolute difference between the relative frequencies of the duplicates and their respective probabilities? _____

5. On the basis of the runs for N=200 and N=500, are you satisfied with the performance of the random number generator? _____

Why, or why not? _____

True and False Questions

Sample versus Population

- | | | |
|---|---|--|
| T | F | a. A sample is part of a population. |
| T | F | b. The sample proportion always equals the population proportion. |
| T | F | c. The larger the sample size the more likely it is that a sample proportion will be close to the population proportion. |
| T | F | d. Each time a different random sample is taken from the same population the sample proportion could be different. |
| T | F | e. The sample proportion from a large sample is always a better estimate of the population proportion. |

Generating Random Numbers

- | | | |
|---|---|---|
| T | F | a. It is easy to recognize a set of random numbers. |
| T | F | b. In any set of truly random numbers, exactly half are even. |
| T | F | c. If five of a kind appears consecutively once in a sequence of 10,000 numbers, this is evidence that the numbers may not be random. |
| T | F | d. In a group of five random numbers, it is more probable that a pair will be found, than finding all of the numbers to be different. |
| T | F | e. About seven times out of one hundred in a group of five random digits, a triple will appear. |

Chapter 4

Frequency Distributions

Histograms and Ogives

A **histogram** is a graph of a frequency distribution of numerical data for different categories of events, individuals, or objects. A **frequency distribution** indicates the individual number of events, individuals, or objects in the separate categories. Most people easily understand histograms because they resemble bar graphs often seen in newspapers and magazines. An **ogive** is a graph of a cumulative frequency distribution of numerical data from the histogram. A **cumulative frequency distribution** indicates the successive addition of the number of events, individuals, or objects in the different categories of the histogram, which always sums to 100. An ogive graph displays numerical data in an S-shaped curve with increasing numbers or percentages that eventually reach 100%. Because cumulative frequency distributions are rarely used in newspapers and magazines, most people never see them. Frequency data from a histogram, however, can easily be displayed in a cumulative frequency ogive.

This chapter will provide you with an understanding of the histogram and its corresponding ogive. You will gain this experience quickly without the work involved in data entry and hand computation. You will be able to view the histogram and cumulative frequency distributions for different sample data sets. Histograms and ogives have different shapes and vary depending on frequency. An ogive always increases from 0% to 100% for cumulative frequencies. The shape of a histogram determines the shape of its related ogive. A uniform histogram is flat; its ogive is a straight line sloping upward. An increasing histogram has higher frequencies for successive categories; its ogive is concave and looks like part of a parabola.

A decreasing histogram has lower frequencies for successive categories; its ogive is convex and looks like part of a parabola. A uni-modal histogram contains a single mound; its ogive is S-shaped. A bi-modal histogram contains two mounds; its ogive can be either reverse S-shaped or double S-shaped depending upon the data distribution. A right-skewed histogram has a mound on the left and a long tail on the right; its ogive is S-shaped with a large concave portion.

A left-skewed histogram has a mound on the right and a long tail on the left; its ogive is S-shaped with a large convex portion.

FREQUENCY R Program

The FREQUENCY R program can be used to display the histogram frequency distributions and ogive cumulative frequency distributions. To simplify the graphical display and provide similar comparisons between the types of histograms, all histograms in the program will have ten categories. The data for each category are not listed; rather the categories are numbered 1 to 10. You will be asked to enter the frequency for each of the ten categories and the frequencies must be integers greater than zero. The program will print a table listing the frequencies you specified, the relative frequencies, and the less-than-or-equal cumulative relative frequencies. The program prints a histogram and a corresponding ogive, which is output in a separate window (GSD2).

The part of the program that can be changed is a list of values relating to a score distribution observed in a given classroom. The length of this list does not matter; it is never specifically referenced in the program. The *Class* object is given a value for a vector of numbers using the **c** function that was introduced in Chapter 1. Each number within the vector is divided by the sum of all values within the vector. In the FREQUENCY program, the first processing loop is replaced by the simple **sum(Class)**, which gets a total for all of the values, and this result is then divided into each of the values within the vector by simply typing *Class/sum(Class)*. No additional step is necessary.

The next program line follows the same logic, only the cumulative sum (**cum-sum**) of the vector is determined at each point and these values are divided by the overall sum of the values to give a vector of values labeled *CumRelFreq*. Scaling of the histogram height is performed next so that the histogram bars are not too small compared to the vertical scaling of the graph. The “if then else” clause is used to provide vertical scaling that will be either one tenth greater than the highest relative frequency, or 1 if the value is .95 or above. The **round** function is implemented to insure that the maximum value is set to an even tenth (**digits=1**). The **barplot** function is used to draw the histogram of the relative frequencies with the **RelFreq** vector as the specified target. The **plot** function is used to draw the ogive of the cumulative relative frequencies with **CumRelFreq** as the target. The **par(mfrow=c(2,1))** command line permits both graphs to be printed, otherwise only the last graph (ogive) will be shown.

The last part of the FREQUENCY program builds a matrix of the class score distribution along with the associated relative frequencies and cumulative relative frequencies. The line beginning *TableData<- matrix* prepares the matrix and initializes all values within it to 0 and makes the dimensions of the matrix to be **length(Class)** rows and 3 columns. The **dimnames** keyword sets the labels for the

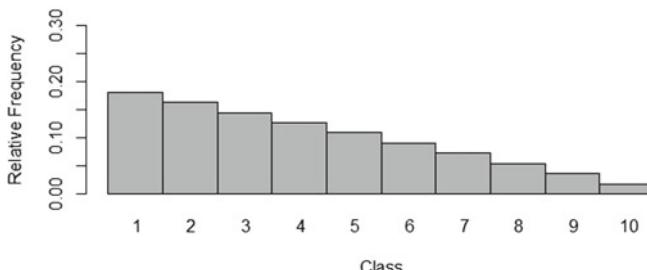
dimensions and will be used in later chapters with other vectors. The **for** loop iterates from 1 to the number of values within *Class* and assigns each row within the *TableData* matrix to the respective *Class* vector value, relative frequency, and cumulative relative frequency, rounding each frequency to three decimal places. You will see some error in the cumulative numbers due to the rounding of the cumulative values. The final line of the program simply prints out the *TableData* matrix.

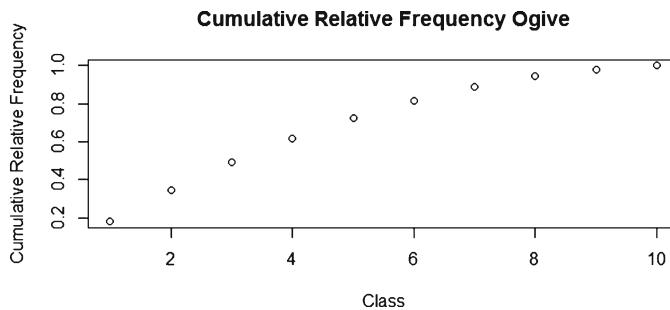
You can change the values within the *Class* vector to obtain different shaped histograms and corresponding ogives. The original vector of 10 values breaks the score distribution into 10 intervals, but this can be changed to create histograms with greater resolution. You could comment out both lines of “**if**” and “**else**” statements that scale the histogram by prefixing them with “#” signs to see the effect of not scaling it properly to fit the plot; replace these statements with the **PlotHeight<-1** statement by removing the # sign in front of it. Some rounding error does occur in the program when summing the relative frequencies to obtain the cumulative relative frequencies.

FREQUENCY Program Output

	Freq.	Relative Freq.	Cum Rel Freq.
Class 1	50	0.182	0.182
Class 2	45	0.164	0.345
Class 3	40	0.145	0.491
Class 4	35	0.127	0.618
Class 5	30	0.109	0.727
Class 6	25	0.091	0.818
Class 7	20	0.073	0.891
Class 8	15	0.055	0.945
Class 9	10	0.036	0.982
Class 10	5	0.018	1.000

Relative Frequency Histogram





Histogram and Ogive Exercises

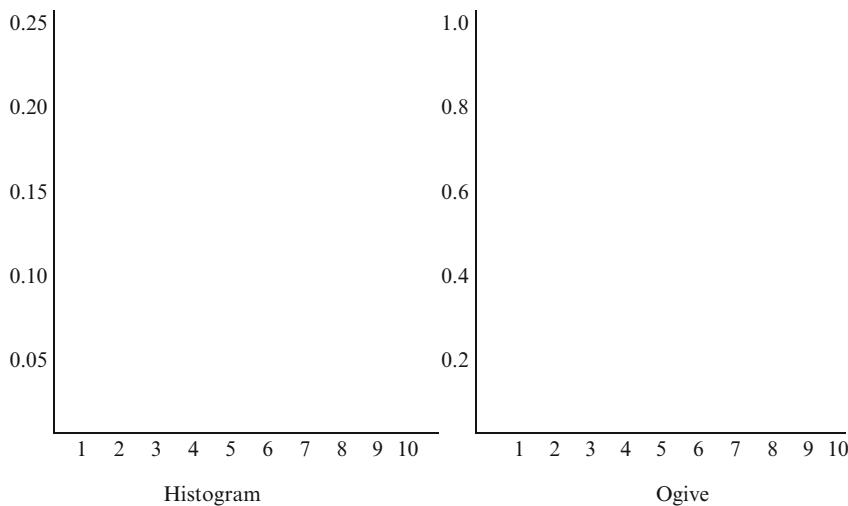
- Run FREQUENCY program six times (a to f). Enter the frequencies listed for each type of histogram in the Class array statement. For each run, complete the frequency table and draw sketches of the histogram and corresponding ogive.

a. A uniform histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	5	_____	_____
2	5	_____	_____
3	5	_____	_____
4	5	_____	_____
5	5	_____	_____
6	5	_____	_____
7	5	_____	_____
8	5	_____	_____
9	5	_____	_____
10	5	_____	_____

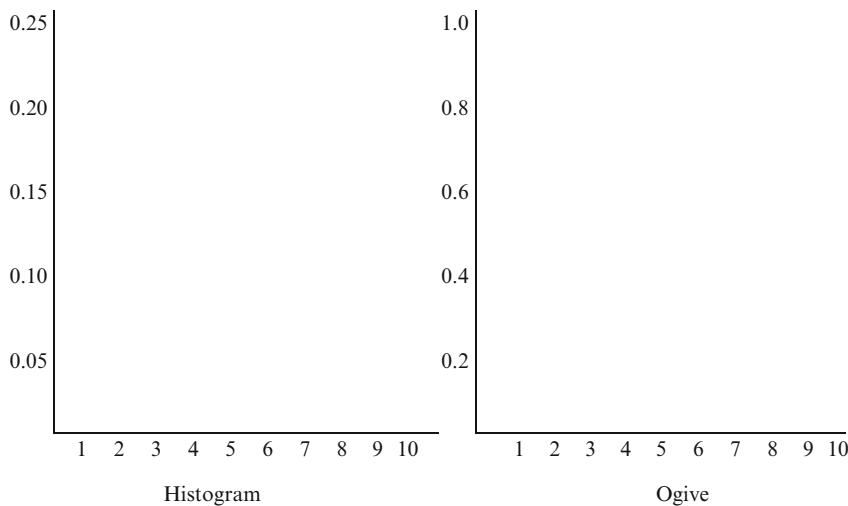
b. An increasing histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	10	_____	_____
2	12	_____	_____
3	14	_____	_____
4	16	_____	_____
5	18	_____	_____
6	20	_____	_____
7	22	_____	_____
8	24	_____	_____
9	26	_____	_____
10	28	_____	_____



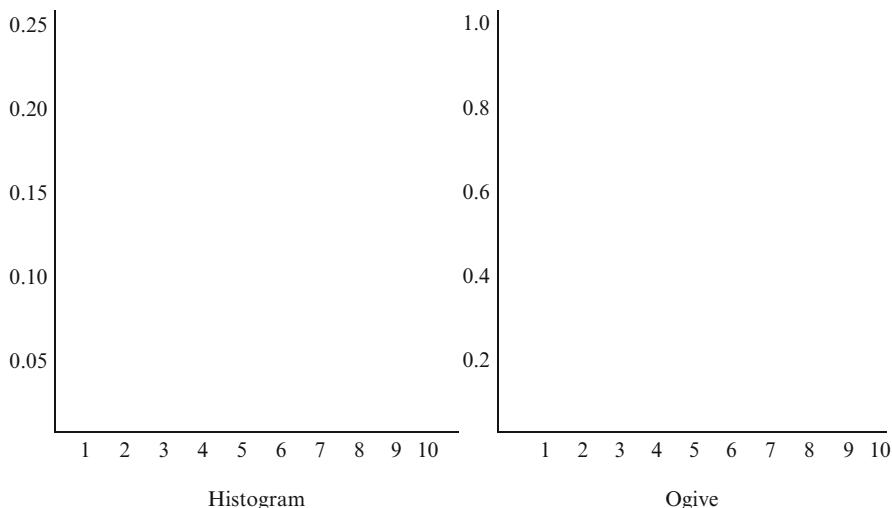
c. A decreasing histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	50	_____	_____
2	45	_____	_____
3	40	_____	_____
4	35	_____	_____
5	30	_____	_____
6	25	_____	_____
7	20	_____	_____
8	15	_____	_____
9	10	_____	_____
10	5	_____	_____



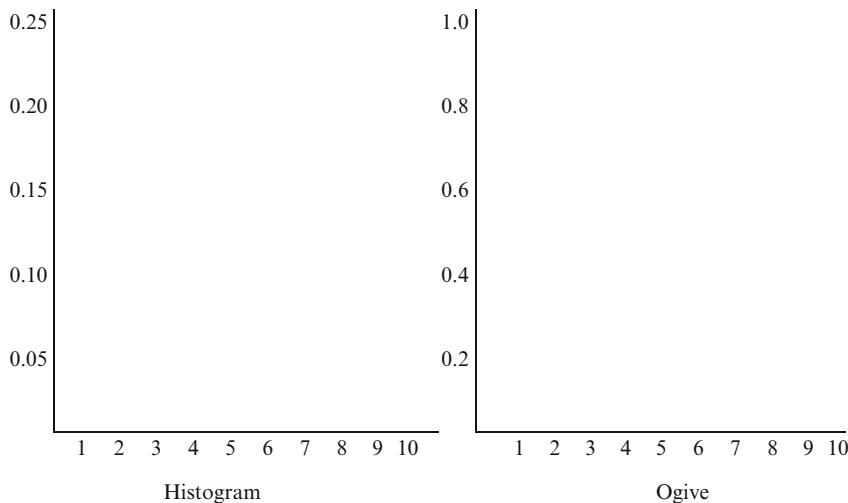
d. A unimodal histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	2	_____	_____
2	3	_____	_____
3	4	_____	_____
4	5	_____	_____
5	6	_____	_____
6	6	_____	_____
7	5	_____	_____
8	4	_____	_____
9	3	_____	_____
10	2	_____	_____



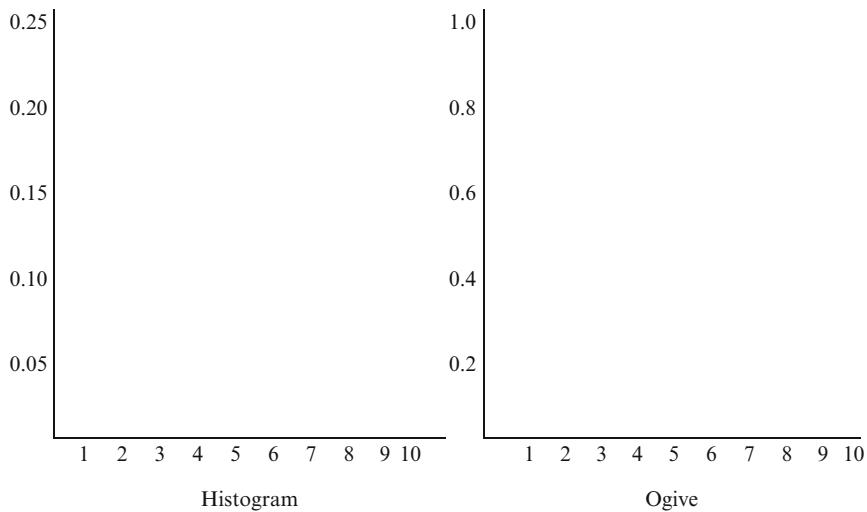
e. A bimodal histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	6	_____	_____
2	5	_____	_____
3	4	_____	_____
4	3	_____	_____
5	2	_____	_____
6	2	_____	_____
7	3	_____	_____
8	4	_____	_____
9	5	_____	_____
10	6	_____	_____



f. A right-skewed histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	5	_____	_____
2	10	_____	_____
3	25	_____	_____
4	20	_____	_____
5	15	_____	_____
6	10	_____	_____
7	5	_____	_____
8	4	_____	_____
9	3	_____	_____
10	2	_____	_____



2. Describe the ogives for each of the following histograms.

HISTOGRAM	OGIVE
a. Uniform	_____
b. Increasing	_____
c. Decreasing	_____
d. Unimodal	_____
e. Bimodal	_____
f. Skewed right	_____

Population Distributions

The heights of adult men form a normal frequency distribution, i.e., a symmetrical distribution with one mode (**unimodal**). A similar population is formed by the heights of adult women. The **mode** is the score that occurs most often in a frequency distribution of data. However, because on the average women are shorter than men, the mean of the population of women's heights is less than the mean height of men. The **mean** is the score that indicates the average of all the scores in a frequency distribution. Imagine that a random sample of adults is chosen and the height of each person is determined. The sample consists of both men and women. In what way will the sample reflect the fact that it was drawn from a combination of two different populations?

In this chapter, you will learn about the shape of a histogram from a sample when the sample is drawn from a combination of two populations. The two populations in this chapter will be unimodal and symmetrical, i.e., normally distributed. To keep the chapter examples simple, the first population will always have a mean of 4. The second population mean can be changed. These mean values are also the modes of the data distributions. A histogram shows the frequency distribution of sample values. For large samples taken randomly from a normal population, the shape of the histogram is approximately normal. If a sample is taken from a combination of two populations, for which the means are far apart, the histogram of the sample will be **bimodal** (two modes). If the means of the two populations are close together, then the sample will be **unimodal** (one mode). In this case, the two populations are considered to be a single population. Large samples of data from a normal population yield a unimodal frequency distribution. The histogram is useful for summarizing data, that is, a bimodal sample could indicate that the sample is from two different populations and that the distance between the centers of two populations is related to the shape of the histogram of a sample selected from the combined populations.

COMBINATION R Program

The first time you run the COMBINATION R program, the second population will have a mean of 9. If you run the program again, you can select a different mean value for the second population. By changing the distance between the two population means, you can observe the effect of this distance on the shape of the histogram from a sample drawn from two combined populations. Each time you run the program, 500 randomly selected observations will be sampled, 250 from each of the populations. The program will print out a relative frequency table and a histogram for the sample.

The program combines the ease of random sampling with the ease of graphing. The program creates two normally distributed random samples centered at different means, with one mean fixed at 4 and the other mean different. The sample sizes can also be changed. The first population data are created from a normal distribution (**rnorm**) with a size of *SampleSize* [1], a mean of 4, and a standard deviation of 1. The values within this vector are rounded to two decimal places. The **round** function is placed within the **invisible** function so that the vector will not be printed while rounding. The second population data are created in the same manner with the mean at the value of the *CenterTwo* variable, a size of *SampleSize* [2], and is also rounded to two decimal places. The two populations are combined into a single population using the **c** function and are treated as a single vector.

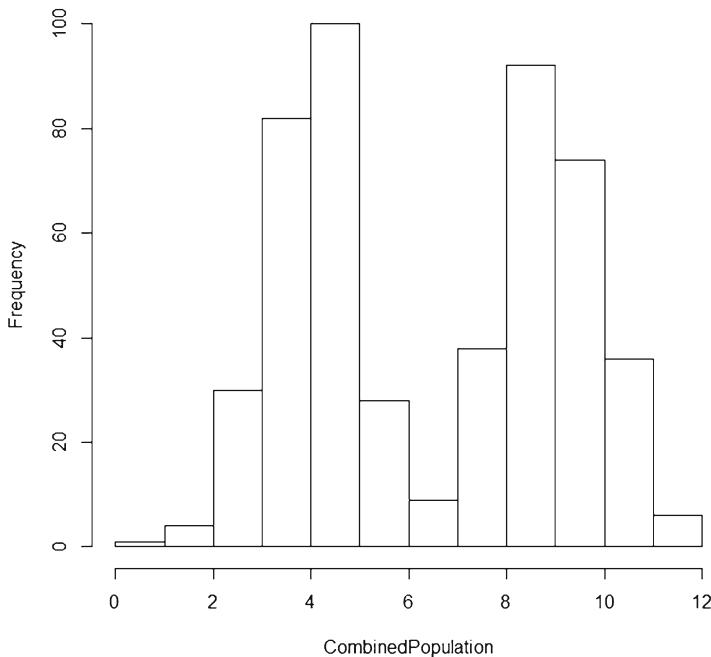
In order to create the relative frequency breakdown for various intervals within the combined population, it is necessary to use the **cut** function combined with the **factor** and **table** functions. The **cut** function breaks data into categories at given breakpoints, in this case at intervals of 0.5 from 0 to the largest value within the bimodal distribution of *CombinedPopulation*. The **factor** function takes the results of the **cut** function and assures that unused intervals are still included in the final results by choosing **levels** from 1 to twice the largest value within *CombinedPopulation* (this is because the intervals are only 0.5). These factors are then summarized by the **table** function with the number of points in each interval placed into the vector *FreqPop*. The next two lines create the labels for the intervals held in the *FreqPop* vector, with the first line being the value of the start of each interval and the second line being the value of the end of each interval. The **nsmall** keyword assures that at least one decimal place is preserved, even for whole numbers, to make the output easier to view.

The *FreqPop* vector is placed next into a **matrix** object to make it easier to display the results. The *FreqPop* vector is the input value, there are **length(FreqPop)** rows, and only 1 column. The dimension names are set to the interval start and interval end labels for the rows and “Rel Freq” for the single column. The **paste** function is handy when combining strings and vectors of values into a vector of labels. The next line with *FreqPopTable* by itself prints out the matrix. An alternative approach would be to use **print(FreqPopTable)**, but since there were no other parameters in the **print** function needed for this display, it wasn’t necessary.

Most of the program code creates and displays the matrix of relative frequencies. The **hist** function creates a histogram of any vector of values with relatively few keywords which is displayed in a separate output window (GSD2). *CombinedPopulation* is the target of the histogram. Default values (,) are specified for the y-axis and label size. The **main** keyword specifies the title to put at the top of the histogram. It is easier to notice differences in the two distributions using a histogram.

COMBINATION Program Output

Rel	Freq
0.0 - 0.5	0
0.5 - 1.0	1
1.0 - 1.5	0
1.5 - 2.0	4
2.0 - 2.5	9
2.5 - 3.0	21
3.0 - 3.5	36
3.5 - 4.0	46
4.0 - 4.5	53
4.5 - 5.0	47
5.0 - 5.5	20
5.5 - 6.0	8
6.0 - 6.5	5
6.5 - 7.0	4
7.0 - 7.5	13
7.5 - 8.0	25
8.0 - 8.5	44
8.5 - 9.0	48
9.0 - 9.5	37
9.5 - 10.0	37
10.0 - 10.5	24
10.5 - 11.0	12
11.0 - 11.5	6
11.5 - 12.0	0

Histogram of Two Combined Populations

COMBINATION Exercises

1. Run the COMBINATION program. A frequency table is printed for the combined populations. The first population will always be centered at 4. The second population will initially be centered at 9. The histogram is printed in a separate output window.
 - a. Describe the shape of the distribution using the relative frequencies. _____

 - b. Describe the shape of the distribution using the histogram. _____

 - c. Are relative frequencies or histograms better in understanding the distribution of data? _____
2. Run the COMBINATION program again. A frequency table is printed for the combined populations. The first population will always be centered at 4. Set the second population mean at 12. A histogram is printed in a separate output window.
 - a. Describe the shape of the distribution using the relative frequencies. _____

- b. Describe the shape of the distribution using the histogram. _____
- c. Are relative frequencies or histograms better in understanding the distribution of data? _____
3. Run the COMBINATION program to find the smallest distance between the two means in which a bimodal distribution is still apparent (the second mean doesn't need to be a whole number, i.e., 8.5 and 5.75). If the means of the two populations are close together, then it is appropriate to consider that the two populations are similar.
- a. What is the smallest distance between the means for which a bimodal distribution is still apparent? _____
- b. Statistical analyses are generally valid only for data that are randomly sampled from a symmetrical unimodal population distribution. What can you do to verify that a certain sample of data was randomly selected from a symmetrical unimodal (normal) population? _____
- _____
- _____

Stem and Leaf Graph

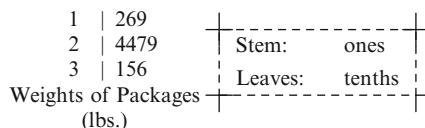
Graphical displays are often used to summarize data and usually help to uncover special characteristics of the data set. Histograms and **stem-and-leaf** plots are examples of graphical displays of data. The stem-and-leaf plots are particularly helpful in visualizing the **median** or middle value in data, the **range** or spread of data (distance between the lowest and highest data values), and **quartiles** (the first quartile is a score that separates the bottom 25% of the data in a frequency distribution from the other data and the third quartile is a score that separates the top 25% of the data in a frequency distribution from the other data). The **inter-quartile range** is the distance between the first and third quartile scores. It measures the range of scores in the middle 50% of the frequency distribution.

In graphing data, a decision must be made about how the data are grouped on the x-axis. A few large groups may obscure information and too many small groups may make summarization meaningless. This chapter addresses the problem of the number of groups in the context of stem-and-leaf plots (similar results are also true for histograms).

Consider the following data set of weights (in pounds) for UPS parcel post packages:

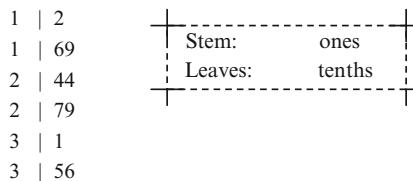
1.2	3.6	2.7	1.6	2.4
3.5	3.1	1.9	2.9	2.4

One stem-and-leaf plot for this data set is the following:



The numbers to the left of the vertical line form the **stem**. In the stem, the numbers represent the digit in the ones place. The numbers to the right of the line are the **leaves**. They represent the digit in the tenths place. Thus “1 | 269” represents the numbers 1.2, 1.6, and 1.9. There are only three groups in this stem-and-leaf plot, i.e., only one for each digit in the stem.

A finer subdivision could have been used in the stem-and-leaf plot. For example,



Here the first occurrence of 1 in the stem is used for data values from 1.0 to 1.4, and the second occurrence of 1 is used for 1.5 to 1.9. A similar approach is used for the other digits in the stem. There are six groups in this plot. Other units can be used for the stem and the leaves. A stem-and-leaf plot of adult heights in inches could have a stem in tens and leaves in ones. A plot of family incomes might have a stem in ten thousands and leaves in thousands.

The stem-and-leaf plots is a type of histogram where the median value in a stem-and-leaf plot can be viewed. The number of groups on a stem has an effect on the shape of the stem-and-leaf plot. The choice of a stem can also make the stem-and-leaf plot either unimodal or bimodal.

For some data sets, a change in the stem causes very little change in the shape of the stem-and-leaf plot.

STEM-LEAF R Program

In the STEM-LEAF R program the data sets will be student grades, which range between 0 and 100 inclusive. Two data sets of 50 student grades each are included in the program. The program will print two versions of a stem-and-leaf plot for each set of student grades. One plot has 11 groups and 6 groups, while the other plot has 15 groups and 8 groups. You will be asked to examine the two different stem and leaf plots to decide if reducing the number of groups gives a better distribution of scores that display for the median or middle value.

The program assigns two large vectors of scores to *Grades1* and *Grades2*. The scores are designed in a manner to demonstrate the importance of proper node assignment in stem-and-leaf plots and should *not* be modified. The stem-and-leaf plot is a simple command, **stem**. The variable, *LeafSpread*, is used as an argument in the command to determine the spacing of leaves.

The output and spacing is put between the script commands by use of the **cat** command with “**\n\n**” as the argument. The **cat** command simply prints characters to the standard output and the “**\n**” special sequence equates to a new line. The next line outputs the stem-and-leaf plot.

STEM-LEAF Program Output

```

First Data Set
[1]0 15 23 27 30 31 35 37 41 44 45 47 50 55 58 59 61 61 64
[20]64 66 68 69 70 71 71 72 72 73 74 74 85 85 85 87 88 88 88
[39]88 90 91 92 92 92 94 94 96 98 99 100

Median=71

11 Groups - First Data Set

The decimal point is 1 digit(s) to the right of the |

0 | 0
1 | 5
2 | 37
3 | 0157
4 | 1457
5 | 0589
6 | 1144689
7 | 01222344
8 | 55578888
9 | 0122244689
10 | 0

6 Groups - First Data Set

The decimal point is 1 digit(s) to the right of the |

0 | 05
2 | 370157
4 | 14570589
6 | 114468901222344
8 | 555788880122244689
10 | 0

```

NOTE: 11 Groups shows a better display of middle split for data at median=71.

Second Data Set

```
[1] 30 31 36 38 42 44 45 47 50 53 53 54 56 58 58 59 61 62 63  
[20] 64 65 66 67 69 70 71 72 74 75 76 77 77 80 80 83 83 85 87  
[39] 88 89 91 92 93 94 95 97 97 99 100 100
```

Median=70.5

15 Groups - Second Data Set

The decimal point is 1 digit(s) to the right of the |

3		01
3		68
4		24
4		57
5		0334
5		6889
6		1234
6		5679
7		0124
7		5677
8		0033
8		5789
9		1234
9		5779
10		00

8 Groups - Second Data Set

The decimal point is 1 digit(s) to the right of the |

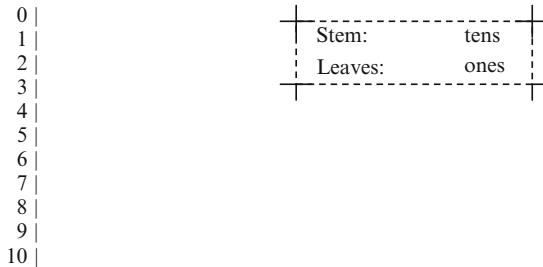
3		0168
4		2457
5		03346889
6		12345679
7		01245677
8		00335789
9		12345779
10		00

NOTE: 8 Groups shows a better display of middle split for data at median=70.5.

STEM-LEAF Exercises

1. Run STEM-LEAF program for GRADES1 with 11 groups and copy the stem-and-leaf plot here.

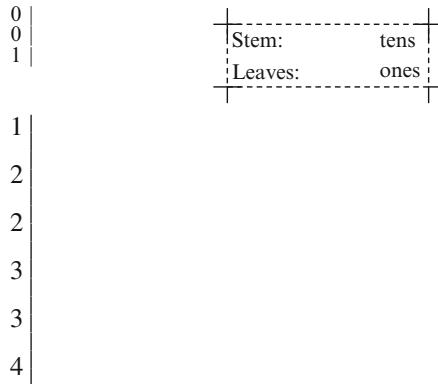
PLOT GRADES1



- a. In what way does the stem-and-leaf plot resemble a histogram? _____
-
- b. Describe the shape. Is it symmetric, right skewed, left skewed, unimodal, or bimodal? _____
- c. The median is the middle score when an odd number of student grades are arranged in order. The median is the average of the two middle scores when an even number of student grades is arranged in order. What is the median student grade? _____
- d. What is the range of scores (distance between lowest and highest score)? _____

2. Run STEM-LEAF program for GRADES1 using 22 groups and copy the stem-leaf plot here.

PLOT GRADES1





- a. Describe the shape of the student grades in this plot. _____

- b. What is the median grade? _____

- c. What is the range? _____
3. Are there any characteristic differences in the two plots of GRADES1?

4. Run the STEM-LEAF program for GRADES2 using 11 and 22 groups and enter the plots below.

PLOT GRADES2

0
1
2
3
4
5
6
7
8
9
10

PLOT GRADES2

0
0
1
1
2
2
3
3
4
4
5
5
6
6
7
7
8
8
9
9
10
10

- a. Is one of the stems more informative for the set of grades? _____
 Why, or why not? _____
- b. Compare the results. What does this illustrate about the effect of the number of stems (groups) in a stem-and-leaf plot? _____
-

True or False Questions

Histograms and Ogives

- T F a. The cumulative relative frequencies in a less-than-or-equal ogive are never decreasing.
- T F b. Some ogives are straight lines.
- T F c. An S-shaped ogive indicates a uniform histogram.
- T F d. The sum of the relative frequencies in a histogram is always one.
- T F e. A parabolic ogive can indicate an increasing histogram.

Population Distributions

- T F a. The greater the distance between the means of two populations, the more pronounced the bimodal shape of the histogram.
- T F b. If the means of two populations are close, then the histogram from the combined populations will have a single peak (unimodal) in the middle.
- T F c. If a sample is taken from a combination of two populations which have means that are far apart, then the sample histogram will be bimodal.
- T F d. As the means of two different populations get closer, the bimodal shape of the histogram is unchanged.
- T F e. Large random samples from normal distributions have unimodal shaped histograms.

Stem and Leaf Graphs

- T F a. The fewer the number of stems (groups) in a stem-and-leaf plot, the more informative the plot will be.
- T F b. The number of stems in the plot does not affect the median value.
- T F c. Skewness is usually apparent even if the number of stems is changed.
- T F d. If a stem-and leaf plot is unimodal for one data set grouping, it will be unimodal when a finer subdivision of groups is used.
- T F e. If a stem-and-leaf plot is rotated 90° counterclockwise, the plot is similar to a histogram.

Chapter 5

Central Tendency and Dispersion

Central Tendency

In the previous chapter, unimodal and bimodal score distributions were demonstrated. In addition to the mode or most frequent score in a sample of data, the mean and median are also considered measures of central tendency. **Central tendency** is where most scores occur in the middle of a symmetrical distribution and then spread out. The mode, mean, and median values will all be identical in a normal distribution.

This chapter examines the effect upon means and medians when data values are transformed and/or extreme data values are added to a data set. The **mean** score is the arithmetic average of numbers in a data set. The mean is computed by taking the sum of the numbers and dividing by the total. The **median** score is the middle score found by arranging a set of numbers from the smallest to the largest (or from the largest to the smallest). If the data set contains an odd number of values, the median is the middle value in the ordered data set. If there is an even number of data values, the median is the average of the two middle values in the ordered data set. The median value is the score that divides the distribution into two equal halves.

Sample data are sometimes modified or transformed to permit comparisons and aid in the interpretation of sample estimates. For example, if the length of cars in inches was changed to meters, one could multiply 0.0254 times the car length to yield meters. What effect does this multiplication have on the mean value representing length of cars? What if an instructor decides to adjust a set of test grades for an exceptionally long test by adding ten points to each student's score? What effect does this addition have on the mean score? If the price of the most expensive house in a neighborhood increases, what happens to the median value of the houses in that neighborhood? What happens to the average value of houses in that neighborhood? These basic effects can be seen because:

If a constant is added to all the values in a data set:

- The mean of the modified data is the mean of the initial data set plus the constant.
- The median of the modified data set is the median of the initial data set plus the constant.

If all of the values in a data set are multiplied by a constant:

- The mean of the modified data set is the mean of the initial data set times the constant.
- The median of the modified data set is the median of the initial data set times the constant.

If the largest value in a data set is replaced by a smaller value, then the mean of the modified data set is smaller than the mean of the initial data set, but the median is unchanged. Extreme values affect the mean, but do not affect the median. If the mean of a data set is subtracted from each of the data values, then the mean of the modified data set is 0. The mean and median of the initial data set can be recovered from the mean and median of the modified data set by adding the negative of the added constant or by dividing by the number that was used as a multiplier.

MEAN-MEDIAN R Program

In the MEAN-MEDIAN R program, you will enter an initial data set. The initial data set will have six numbers and the numbers must be entered in order from smallest to largest. This initial data will then be transformed in three different ways: (1) add a constant to each data value; (2) multiply each data value by a constant; or (3) replace the largest data value by a smaller data value. The program will print the initial data set with its mean and median followed by the mean and median from the modified data sets. The effect of these data modifications on the mean and median can then be observed.

MEAN-MEDIAN Program Output

Initial Data Set

2 4 5 9 15 19

Mean = 9 Median = 7

Added 10 to the Initial data

Added Value Data

Mean = 19 Median = 17

Multipled 5 to the Initial data

Multipled Value Data

Mean = 45 Median = 35

Replaced Largest Value 19 in the Initial data with 1

Replaced Value Data

Mean = 6 Median = 4.5

MEAN-MEDIAN Exercises

- To run MEAN-MEDIAN program, the initial data set (2, 4, 5, 9, 15, 19) is specified.

Use the following data transformations and record the results.

INITIAL DATA	MEAN	MEDIAN
ADD		
3	_____	_____
-2	_____	_____
10	_____	_____
5	_____	_____
MULTIPLY BY		
2	_____	_____
-10	_____	_____
0.5	_____	_____
REPLACE LAST VALUE WITH		
1	_____	_____
5	_____	_____
10	_____	_____

- Complete the following statements:

- If a constant number is added to each of the data values in a set, then the mean of the modified data is equal to the initial mean _____ and the median of the modified data set is equal to the initial median _____.

_____.

- If each data value in a set is multiplied by a constant number, then the mean of the modified data is equal to the initial mean _____ and the median of the modified data is equal to the initial median _____.

_____.

- If the largest value in a data set is replaced by a smaller value, the mean of the modified data is _____

and the median of the modified data set is _____.

_____.

3. Run MEAN-MEDIAN and enter the data set: 2, 4, 6, 9, 12, 15.

- a. Modify the data by adding +5 to each value. Record the results.

INITIAL DATA _____

MEAN _____ MEDIAN _____

MODIFIED DATA (+5)_____

MEAN _____ MEDIAN _____

- b. Run MEAN-MEDIAN with the data set: 7, 9, 11, 14, 17, 20.

Modify the data by subtracting -5 from each data value. Record the results.

INITIAL DATA _____

MEAN _____ MEDIAN _____

MODIFIED DATA (-5)_____

MEAN _____ MEDIAN _____

- c. Show how to obtain the mean of the initial data set from the mean of the modified data set.

- d. Show how to obtain the median of the initial data set from the median of the modified data set.

4. Run MEAN-MEDIAN using the data set: 2, 4, 6, 9, 12, 15.

- a. Modify the data by multiplying each value by +4. Record the results.

INITIAL DATA _____

MEAN _____ MEDIAN _____

MODIFIED DATA (4×)_____

MEAN _____ MEDIAN _____

- b. Run MEAN-MEDIAN using the data set: 8, 16, 24, 36, 48, 60. Modify the data set by multiplying each value by 0.25 (dividing by 4). Record the results.

INITIAL DATA _____

MEAN _____ MEDIAN _____

MODIFIED DATA (0.25×)_____

MEAN _____ MEDIAN _____

- c. Show how to obtain the mean of the initial data set from the mean of the modified data set. _____
-
5. The average and median daily temperature for six days in a northern city was 76° F. Daily temperature readings can be changed to a Celsius scale by the formula: $C = (F - 32)(5/9)$; that is, a value of -32 must be added to each data value, and then the results must be multiplied by $5/9$.
- a. If the six daily temperatures were 73° , 78° , 81° , 74° , 71° , and 79° on the Fahrenheit scale, use MEAN-MEDIAN program to change to the Celsius scale. (Don't forget to order the data as you enter it, add a -32 , and then multiply the values using a decimal format: $5/9 = 0.5556$).
- b. What is the average temperature of the city in Celsius? _____
- c. What is the median temperature in Celsius? _____

Dispersion

Dispersion refers to how spread out scores are around the mean. The sample **range** is the difference between the largest and smallest data value. The sample **variance** is the average squared deviation of the data values from their sample mean. The sample **standard deviation** is the square root of the sample variance. The formula for the sample variance is:

$$\text{Variance} = \frac{\sum (\text{data value} - \text{mean data value})^2}{\text{number of data values}} = \frac{SS}{n}$$

The numerator (top of equation) indicates that the mean of all the data values is subtracted from each data value, squared, and summed. The summing of the squared values is denoted by the symbol, Σ . This is referred to as the sum of squared deviations from the mean or simply **sum of squared deviations** (SS). The sum of squared deviations (SS) divided by the number of data values is referred to as the variance.

The standard deviation is the square root of the variance. The formula for the standard deviation is:

$$\text{Standard Deviation} = \sqrt{\frac{SS}{n}}$$

The standard deviation provides a measure in standard units of how far the data values fall from the sample mean. For example, in a **normal distribution**, 68% of the data values fall approximately one standard deviation (1 SD) on either side of the mean, 95% of the data values fall approximately two standard deviations (2 SD)

on either side of the mean, and 99% of the data values fall approximately three standard deviations (3 SD) on either side of the mean.

Basically, we should find that if a constant is added to all the values in a data set, the variance, standard deviation, and range are unchanged. If all of the values in a data set are multiplied by a constant: The variance of the modified data set is the variance of the initial data set times the constant squared. The standard deviation of the modified data set is the standard deviation of the initial data set times the constant. The range of the modified data set is the range of the initial data set times the constant. If the last value in a data set is replaced by a smaller value, then the variance, standard deviation, and the range are all decreased. If the last value in a data set is replaced by a larger value, then the variance, standard deviation, and the range are all increased. If the standard deviation of a data set is divided into each of the data values, then the standard deviation of the modified data set is 1.

The variance of the initial data set can be obtained from the variance of the modified data set by dividing the variance by the constant squared that was used as the multiplier. The standard deviation and range of the initial data set can be obtained from the standard deviation and range of the modified data set by dividing them by the constant that was used as the multiplier.

DISPERSION R Program

In the DISPERSION R program an initial data set is entered and modified in one of three ways: (1) adding a constant to each data value; (2) multiplying each data value by a constant; or (3) replacing the last value by a different number. The purpose is to observe the effect of these modifications on three measures of dispersion: range, variance, and standard deviation. The **sd** function returns the standard deviation of the vector of data values, the **var** function computes the variance, and the **range** function gives the minimum and maximum data value.

DISPERSION Program Output

```
Initial Data Set
2 4 5 9 15 19
Standard Deviation = 6.723095 Variance=45.2 Range=2 19

Added 10 to the Initial data
Added Value Data
Standard Deviation = 6.723095 Variance=45.2 Range=12 29

Multiplied 5 to the Initial data
Multiplied Value Data
Standard Deviation = 33.61547 Variance=1130 Range=10 95
Replaced Largest Value 19 in the Initial data with 1
```

Replaced Value Data

Standard Deviation = 5.215362 Variance=27.2 Range=1 15

DISPERSION Exercises

- Run the DISPERSION program with the data set 2, 4, 5, 9, 15, 19. Use the data modifications below and record the results.

INITIAL DATA	S.D.	VARIANCE	RANGE
ADD			
3	_____	_____	_____
-2	_____	_____	_____
10	_____	_____	_____
5	_____	_____	_____
MULTIPLY BY			
2	_____	_____	_____
-10	_____	_____	_____
0.5	_____	_____	_____
REPLACE LAST VALUE WITH			
1	_____	_____	_____
5	_____	_____	_____
10	_____	_____	_____

- Complete the following statements:

- If a constant is added to each data value, then the variance of the modified data is equal to _____, the standard deviation of the modified data set is equal to _____, and the range of the modified data set is equal to _____.
- If each data value is multiplied by a constant, then the variance of the modified data is equal to the initial variance _____, the standard deviation of the modified data is equal to the initial standard deviation _____, and the range of the modified data set is equal to the initial range _____.
- If the last value in a data set is replaced by a smaller value, the variance of the modified data is _____, the standard deviation of the modified data is _____, and the range of the modified data set is _____.

- Run DISPERSION program again using your initial data set and modify it by dividing each data value by the standard deviation (multiply by the reciprocal). What is the standard deviation of the data? _____.

Try it again with a new initial data set. Explain why this happens: _____

4. Run DISPERSION program with the data set 2, 4, 6, 9, 12, 15.
a. Modify the data by adding +5 to each value. Record the results.

INITIAL DATA _____

S.D. _____ VARIANCE _____ RANGE_____

MODIFIED DATA _____

S.D. _____ VARIANCE _____ RANGE_____

- b. Run MODIFICATION and enter the data set 7, 9, 11, 14, 17, 20.
Modify the data by adding -5 to each value. Record the results.

INITIAL DATA _____

S.D. _____ VARIANCE _____ RANGE_____

MODIFIED DATA _____

S.D. _____ VARIANCE _____ RANGE_____

- c. Show how to obtain the variance of the modified data set from the variance of the initial data set.
-
- d. Show how to obtain the standard deviation of the initial data set from the standard deviation of the modified data set.
-
- e. Show how to obtain the range of the initial data set values from the range of the modified data set values.
-

5. The variance of the daily temperatures for six days in a northern city was 14°F , the standard deviation was 3.8°F , and the range was 10° F . The daily temperature readings can be changed to a Celsius scale by using the formula $C = (F - 32) \cdot (5/9)$; that is, -32 must be added to each data value and the results multiplied by $5/9$ (0.5556).
- What is the standard deviation of the temperature in degrees Celsius? _____
 - What is the variance of the temperature for this city in Celsius? _____
 - What is the range of the temperature in degrees Celsius? _____
6. If the six daily temperatures were 73° , 78° , 81° , 74° , 71° , and 79° on the Fahrenheit scale, use the DISPERSION program to change them to the Celsius scale. (Don't forget to order the data as you enter it and express values in decimal form: $5/9 = 0.5556$).

- a. What is the standard deviation of the temperature in degrees Celsius? _____
- b. What is the variance of the temperature for this city in Celsius? _____
- c. What is the range of the temperature in degrees Celsius? _____

Sample Size Effects

The sample **range** is the difference between the highest and lowest score in a data distribution. The sample **variance** is the square of the sample **standard deviation**. When the size of the sample increases, the range of data values will generally increase. The standard deviation with increasing sample sizes should divide the frequency distribution of data into six sections. You should be able to observe the effect of sample size on these measures of data dispersion when completing the chapter exercises.

As the sample size increases, the sample range usually increases because observations are chosen from the extreme data values in a population. As observations are added to a sample, the range of the sample cannot decrease. As observations are added to a sample, the standard deviation fluctuates in an unpredictable manner. A rough approximation of the standard deviation is the range divided by four; however, for a uniform population, this will produce an underestimate. Range divided by six better approximates the standard deviation of the normal distribution.

SAMPLE R Program

The SAMPLE R programs will create a uniform population of integers from 1 to 1,000 based on sampling without replacement [Sampling with replacement assumes that data points are returned to the population from which they were drawn. Sampling without replacement assumes that data points are *not* returned to the population from which they were drawn.] The probability of selection is affected depending upon which sampling technique is used. Various sample sizes will need to be listed in the *Samplesizes* vector. Random sampling will be repeated with new observations for each sample size listed in the vector. A summary table will be printed with the results to allow you to draw conclusions about the effect of sample size on the range and standard deviation. The ratio of the range to the standard deviation will be printed so you can look for a relationship between these two measures of dispersion.

The program can be repeated as many times as needed since the sampling is random. Each time the program is run, the results, however, will be different. This will allow you to further test your understanding and conclusions about how various sample sizes affect the range and standard deviation. The final chapter exercise computes the error one would make when using a sample estimate of the standard deviation as the population value. This exercise is used to answer the question, “Does the sample standard deviation become a more accurate estimate of the population standard deviation as the sample size increases?”

Once the matrix has been defined, the main processing loop begins. An iteration counter is used as the basis of the **for** loop to facilitate placement of values within the matrix. The first line of code within the loop creates the first sample of random integers between 1 and 1,000 from a uniform distribution. The next three lines fill the columns of the matrix at the present row (*i*) with the range, standard deviation, and range divided by the standard deviation, respectively. The matrix notation [*i*,1] represents the *i*th row and the first column. If you want to replace an entire row of values, type *outputMatrix*[*i*,]<-, followed by the assignment of the vector. Leaving a dimension blank means that you are allowing all values along that dimension to be filled with values, if enough are present in the vector being assigned to it. The content of the matrix is printed after the end of the loop.

Sample Program Output

N	Range	Standard Dev.	Range/SD
10	893	327.83	2.72
50	937	286.18	3.27
100	979	319.72	3.06
200	991	286.30	3.46
500	997	281.51	3.54
1000	997	288.87	3.45

SAMPLE Exercises

- Enter the following string of sample sizes in the *Samplesize* vector and run the SAMPLE program. Record the results below.

	RANGE	STANDARD DEV.	RANGE/SD
20	_____	_____	_____
40	_____	_____	_____
60	_____	_____	_____
80	_____	_____	_____
100	_____	_____	_____
120	_____	_____	_____
140	_____	_____	_____
160	_____	_____	_____
180	_____	_____	_____
200	_____	_____	_____
220	_____	_____	_____
240	_____	_____	_____
260	_____	_____	_____
280	_____	_____	_____
300	_____	_____	_____

2. Provide short answers to the following questions.

a. As observations are added to the sample, what happens to the sample range?

b. What accounts for the relationship between the sample size and the sample range?

c. Why is the sample range a less than perfect measure of the spread of the population?

d. As observations are added to the sample, what is the relationship between the sample size and the sample standard deviation?

3. Run the SAMPLE program again with the same sample sizes in the *Samplesize* vector. Record the results below.

	RANGE	STANDARD DEV.	RANGE/SD
20	_____	_____	_____
40	_____	_____	_____
60	_____	_____	_____
80	_____	_____	_____
100	_____	_____	_____
120	_____	_____	_____
140	_____	_____	_____
160	_____	_____	_____
180	_____	_____	_____
200	_____	_____	_____
220	_____	_____	_____
240	_____	_____	_____
260	_____	_____	_____
280	_____	_____	_____
300	_____	_____	_____

a. Did anything different happen the second time? _____
If so, what was different? _____

b. What is your final conclusion about the relationship between sample size and sample standard deviation? _____

4. The last column above indicates the standard deviations for samples of data from a uniform distribution. The range of scores divided by 4 is a rough estimate of the standard deviation of a uniform distribution. Are the standard deviations less than 4 (underestimated) as expected? Yes ____ No ____

5. What is a good estimate for the standard deviation of a normal population?

6. If the sample standard deviation is used as an estimate of the population standard deviation, compute the error of the estimate for each sample size in Exercise 1.

Note: $\text{ERROR} = \text{ESTIMATE} - 288.67$. The population standard deviation was 288.67.

Record the error of estimate for each sample size with its $+/-$ signs in the following table.

SAMPLE SIZE	ERROR
20	_____
40	_____
60	_____
80	_____
100	_____
120	_____
140	_____
160	_____
180	_____
200	_____
220	_____
240	_____
260	_____
280	_____
300	_____

- a. Does the sample standard deviation become a more accurate estimate of the population standard deviation as the sample size increases?
-

Tchebysheff Inequality Theorem

The sample **standard deviation** is a measure of the dispersion of the sample data around the sample mean. A small standard deviation indicates less dispersion of sample data. A larger standard deviation indicates more dispersion of sample data. This understanding is also true for the **range**, which is the difference between the largest and smallest data value. However, the standard deviation provides more information about the data than the range. The standard deviation permits the formation of intervals that indicate the proportion of the data within those intervals. For example, 68 % of the data fall within $+/-$ one standard deviation from the mean, 95 % of the data fall within $+/-$ two standard deviations of the mean, and 99 % fall within $+/-$ three standard deviations of the mean, in a **normal distribution**. If 100 students took a mathematics test with a mean of 75 and a standard deviation of 5, then 68 % of the scores would fall between a score of 70 and 80, assuming a normal distribution. In contrast, given the highest and lowest test scores, 90 and 50 respectively, the range of 40 only indicates that there is a 40-point difference between the highest and lowest test score, i.e., $90 - 50 = 40$.

We generally assume our data is normally distributed; however, in some cases, the data distribution takes on a different shape. When this occurs, the **Tchebysheff Inequality Theorem** is helpful in determining the percentage of data between the intervals. For example, if the mean mathematics test score was 85, and the standard deviation was 5, then the Tchebysheff Inequality Theorem could be used to make a statement about the proportion of test scores that fall in various intervals around the mean, e.g., between the score interval 75 and 95, *regardless of the shape of the distribution*.

The **Tchebysheff Inequality Theorem** was developed by a Russian mathematician as a proof that given a number k , greater than or equal to 1, and a set of n data points, at least $(1 - 1/k^2)$ of the measurements will lie within k standard deviations of their mean. Tchebysheff's theorem applies to *any* distribution of scores and could refer to either sample data or the population. To apply the Tchebysheff Inequality Theorem using a population distribution, an interval is constructed which measures $k\sigma$ on either side of the mean, μ . When $k = 1$, however, $1 - 1/(1)^2 = 0$, which indicates that 0 % of the data points lie in the constructed interval, $\mu - \sigma$ to $\mu + \sigma$, which is not helpful nor useful in explaining data dispersion. However, for values of k greater than 1, the theorem appears to be informative:

k	$1 - 1/k^2$ (Percent)	Interval
1	0 (0 %)	$\mu +/- 1\sigma$
2	3/4 (75 %)	$\mu +/- 2\sigma$
3	8/9 (89 %)	$\mu +/- 3\sigma$

An example will help to better illustrate the fraction of n data points that lie in a constructed interval using the Tchebysheff theorem. Given a set of test scores with a mean of 80 and a standard deviation of 5, the Tchebysheff theorem would indicate a constructed interval with lower and upper score limits computed as follows:

$$\text{Lower limit} = \text{mean} - k * \text{standard deviation}$$

$$\text{Upper limit} = \text{mean} + k * \text{standard deviation}$$

For $k = 1$, the lower limit would be $80 - 1*5 = 75$ and the upper limit would be $80 + 1*5 = 85$. Obviously, for $k = 1$, no data points are implied between the score interval, 75–85, which makes no sense. For $k = 2$, the lower limit would be 70 and the upper limit would be 90. The Tchebysheff Inequality Theorem implies that *at least* $[1 - 1/k^2]$ of the data values are within the score interval. Thus, for the constructed interval $k = 2$, 70–90, $1 - 1/(2)^2 = 1 - 1/4 = 1 - 0.25 =$ *at least* 75 % of the data points are between 70 and 90, *regardless of the shape of the data distribution*. The Tchebysheff Inequality Theorem is very conservative, applying to *any* distribution of scores, and in most situations the number of data points exceeds that implied by $1 - 1/k^2$.

The Tchebysheff Inequality Theorem is generally applied to populations and intervals formed around the population mean using k population standard deviations, where k ranges from 1 to 4. In practice, however, one rarely knows the population parameters (population means and standard deviations). In some instances, the population parameters are known or at least can be estimated. For example,

a nationally normed test booklet would contain the population mean and standard deviation, typically called “test norms.” Researchers often use tests to measure traits and characteristics of subjects and publish test sample means and standard deviations. In this instance, an average of the results from several published studies would yield a reasonable estimate of the population parameters. In finite populations where every observation is known and recorded, the population parameters are readily obtainable using computer statistical packages to analyze the data. In a few instances, dividing the range of scores by six provides a reasonable estimate of the population standard deviation as an indicator of data dispersion.

Since the sample mean and standard deviation are estimates of the population parameters, the Tchebysheff Inequality Theorem can be used with sample data. We therefore can test whether or not the Tchebysheff Inequality Theorem is useful for describing data dispersion and compare it to the normal distribution percentages where approximately $1\sigma=68\%$, $2\sigma=95\%$, and $3\sigma=99\%$. In the TCHEBYSHEFF program, samples will be selected from four different populations: uniform, normal, exponential, or bimodal. The four different distributions are functions within the R program; however, they can be created in other programming software by computing a value for $X(i)$, which is a data vector, using the following equations and functions (RND=round a number; COS=cosine of a number; SQR=square root of a number; LOG=logarithm of a number):

$$\begin{aligned} \text{Uniform: } X(i) &= 1 + 9*RND \\ \text{Normal: } X(i) &= \text{COS}(6.2832*RND) * \text{SQR}(-2*\text{LOG}(RND)) \\ \text{Exponential: } X(i) &= -\text{LOG}(RND) \\ \text{Bimodal: } X(i) &= (2 + \text{SQR}(4 - (8 * (1 - RND))))/2 \\ &\text{If } RND \geq 0.5 \\ &X(i) = (2 - \text{SQR}(4 - (8*RND)))/2 \end{aligned}$$

The Tchebysheff Inequality Theorem provides a lower bound for the proportion of sample data within intervals around the mean of any distribution. The Tchebysheff Inequality Theorem is true for all samples regardless of the shape of the population distribution from which they were drawn. The Tchebysheff lower bound is often a conservative estimate of the true proportion in the population. We would use the standard deviation to obtain information about the proportion of the sample data that are within certain intervals of a normally distributed population. The population mean and standard deviation are estimated more accurately from large samples than from small samples.

TCHEBYSHEFF R Program

The program will require specifying the sample size and distType. The TCHEBYSHEFF program will select a random sample, compute the sample mean and standard deviation, and determine the percentage of the observations within 1.5, 2, 2.5, and 3 standard deviations of the mean (Kvals). The lower bound for the

percentage of data within the interval given by the Tchebysheff Inequality Theorem will also be printed to check whether the theorem is true for the sample data.

The program takes random samples from different shaped distributions. The “Uniform” selection chooses a random sample of *SampleSize* from a uniform distribution that falls between the values of 1 and 10. “Normal” creates a random sample from a normal distribution with a mean of 0 and standard deviation of 1. “Exponential” creates a random sample from an exponential distribution with a mean of 1. Finally, “Bimodal” creates a random sample from a bimodal distribution (made up of an equal number of points chosen from two adjacent normal distributions). Whichever distribution type is input, the standard deviation and mean are obtained, and then the matrix is filled with values representing the Tchebysheff intervals, the percent of observations in the sample falling within the interval, and the value of the Tchebysheff Lower Bound.

TCHEBYSHEFF Program Output

Uniform N=50 Sample Mean 5.17 Sample Std Dev 2.51

K	Interval	% Obs	Tcheby
1.5	1.4 to 8.9	86	56
2.0	0.16 to 10	100	75
2.5	-1.1 to 11	100	84
3.0	-2.4 to 13	100	89

Normal N=50 Sample Mean 0.04 Sample Std Dev 0.97

K	Interval	% Obs	Tcheby
1.5	-1.4 to 1.5	88	56
2.0	-1.9 to 2	98	75
2.5	-2.4 to 2.5	98	84
3.0	-2.9 to 3	100	89

Exponential N=50 Sample Mean 1.15 Sample Std Dev 1.14

K	Interval	% Obs	Tcheby
1.5	-0.56 to 2.9	90	56
2.0	-1.1 to 3.4	92	75
2.5	-1.7 to 4	96	84
3.0	-2.3 to 4.6	100	89

Bimodal N=50 Sample Mean 0.92 Sample Std Dev 0.62

K	Interval	% Obs	Tcheby
1.5	0 to 1.8	88	56
2.0	-0.31 to 2.1	96	75
2.5	-0.62 to 2.5	98	84
3.0	-0.93 to 2.8	100	89

TCHEBYSHEFF Exercises

1. Run TCHEBYSHEFF for the sample size =50 and distType=“Uniform” specified in the program. Then replace the distType for the other distribution types.

UNIFORM N=50 K INTERVAL	Sample Mean _____ % OBS. IN INT.	Sample St. Dev. _____ TCHEBY. LOWER BOUND
1.5 _____	_____	_____
2.0 _____	_____	_____
2.5 _____	_____	_____
3.0 _____	_____	_____

NORMAL N=50 K INTERVAL	Sample Mean _____ % OBS. IN INT.	Sample St. Dev. _____ TCHEBY. LOWER BOUND
1.5 _____	_____	_____
2.0 _____	_____	_____
2.5 _____	_____	_____
3.0 _____	_____	_____

EXPONENTIAL N=50 K INTERVAL	Sample Mean _____ % OBS. IN INT.	Sample St. Dev. _____ TCHEBY. LOWER BOUND
1.5 _____	_____	_____
2.0 _____	_____	_____
2.5 _____	_____	_____
3.0 _____	_____	_____

BIMODAL N=50 K INTERVAL	Sample Mean _____ % OBS. IN INT.	Sample St. Dev. _____ TCHEBY. LOWER BOUND
1.5 _____	_____	_____
2.0 _____	_____	_____
2.5 _____	_____	_____
3.0 _____	_____	_____

2. Are the Tchebysheff lower bound values always correct in the table? _____
-
3. The Tchebysheff lower bound is very conservative; it is often a lower bound far below the actual percentage of data in the interval. For which population is the Tchebysheff lower bound the least conservative? Run TCHEBYSHEFF several times with different sample sizes to verify your conclusion.
-
-
4. The actual population means and standard deviations of the four populations are:

	POP. MEAN	POP. ST. DEV.
UNIFORM	5.5	2.60
NORMAL	0	1
EXPONENTIAL	1	1
BIMODAL	1	0.707

The sample means and standard deviations computed in TCHEBYSHEFF can be used as estimates of these population parameter values.

- a. Run TCHEBYSHEFF to complete the following table. Recall that

$$\text{ERROR} = \text{SAMPLE ESTIMATE} - \text{POPULATION VALUE}.$$

SAMPLE SIZE	SAMPLE MEAN	POP. MEAN	ERROR	SAMPLE ST. DEV.	POP ST. DEV.	ERROR
UNIFORM						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
NORMAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
EXPONENTIAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
BIMODAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____

- b. Is there a relationship between the sample size and the absolute value of the error in the estimates of the mean? _____

- c. Is there a relationship between the sample size and the absolute value of the error in the estimates of the standard deviation? _____

Normal Distribution

In a normal population, referred to as normal bell-shaped curve, the proportion of data within intervals around the mean is known. The proportion of sample data within k standard deviations around the mean for $k=1, 2$, and 3 are as follows:

INTERVAL		DATA PERCENT IN THE INTERVAL
LOWER LIMIT	UPPER LIMIT	
$\mu - 1\sigma$	$\mu + 1\sigma$	68%
$\mu - 2\sigma$	$\mu + 2\sigma$	95%
$\mu - 3\sigma$	$\mu + 3\sigma$	99%

If a large random sample is selected from a normal population, these lower and upper intervals will approximate the percent of data within the intervals of a normal bell-shaped curve. The proportions of sample data should be good approximations if the population is approximately normal, symmetrical, and unimodal. If the population is non-normal, the bell-shaped curve may not provide results close to these proportions of sample data.

The different population distribution functions in R permit the random sampling of data from different population distributions. The different population values can be adjusted and the population parameters for the different distributions are:

	MEAN	STANDARD DEVIATION
UNIFORM	5.5	2.60
NORMAL	0	1
EXPONENTIAL	1	1
BIMODAL	1	0.707

The normal distribution gives an approximation of the proportion of a sample that is within one, two, and three standard deviations of the population mean if the sample is large and is chosen at random from a normal population. The percentages of data within one, two, and three standard deviations of the mean for a normal population are respectively 68%, 95%, and 99%. The program does not give good approximations for samples chosen from uniform, exponential, or bimodal populations nor good approximations for small samples.

BELL-SHAPED CURVE R Program

The BELL-SHAPED CURVE program will create large sample data sets chosen at random from the four population types: uniform, normal, exponential, and bimodal. The program will determine the proportion of the sample within the specified intervals. The tabled output should allow you to check the accuracy of the percentages for large samples from the population distribution types, especially a normal population. The proportions within the specified intervals for smaller samples can also be checked.

The BELL-SHAPED CURVE program is similar to the TCHEBYSHEFF program except **Kvals** is set to intervals using percentages which are specified as $Kvals < -c(68, 95, 99)$; which corresponds to 1, 2, or 3 standard deviations from the population mean, respectively. The sample size is initially set at $SampleSize < -50$.

BELL-SHAPED CURVE Program Output

Uniform N=50 Sample Mean=5.38 Sample Std Dev=2.87

	Interval	% Observed	% Predicted
K=1	2.5 to 8.2	54	68
K=2	-0.37 to 11	100	95
K=3	-3.2 to 14100	99	

Normal N=50 Sample Mean=-0.16 Sample Std Dev=0.98

	Interval	% Observed	% Predicted
K=1	-1.1 to 0.83	60	68
K=2	-2.1 to 1.8	98	95
K=3	-3.1 to 2.8	100	99

Exponential N=50 Sample Mean=1.01 Sample Std Dev=1.17

	Interval	% Observed	% Predicted
K=1	-0.16 to 2.2	88	68
K=2	-1.3 to 3.3	92	95
K=3	-2.5 to 4.5	98	99

Bimodal N=50 Sample Mean=1.05 Sample Std Dev=0.69

	Interval	% Observed	% Predicted
K=1	0.36 to 1.8	60	68
K=2	-0.33 to 2.4	98	95
K=3	-1 to 3.1	100	99

Normal Distribution Exercises

1. Run BELL-SHAPED CURVE for a sample size of 250, and complete the following tables. Note: ERROR=% OBSERVED – % PREDICTED.

- a. Compute the lower and the upper limits of the interval for the uniform population with $k=3$, using the following formula:

$$\text{LOWER LIMIT} = \text{MEAN} - K * \text{STANDARD DEVIATION} = \underline{\hspace{10cm}}$$

$$\text{UPPER LIMIT} = \text{MEAN} + K * \text{STANDARD DEVIATION} = \underline{\hspace{10cm}}$$

- b. Which population produced a sample that is approximated best by the Normal Bell-Shaped Curve? _____
- c. Run BELL-SHAPED CURVE program again for sample size 500 to test your conclusions. Comment on the results.
-

UNIFORM: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

NORMAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

EXPONENTIAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

BIMODAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

2. Run BELL-SHAPED CURVE program again for a sample size of 1,000 and complete the tables.

UNIFORM: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

NORMAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

EXPONENTIAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

BIMODAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

- a. Compare the results for the samples of size 1,000 with the results for samples of size 250. For which sample size (250 or 1,000) were the approximations best for the normal population? _____
- b. Was there more error in the nonnormal populations (uniform, exponential, or bimodal)? Yes _____ No _____
3. The uniform, exponential, and the bimodal populations all have shapes that are very different from the normal population.
- a. In what way is the uniform population different from the normal population?

- b. In what way is the exponential population different from the normal population?

- c. In what way is the bimodal population different from the normal population?

- d. Which population type had the most error?

4. Run BELL-SHAPED CURVE program for a sample size of $n=20$ for all four populations.
- a. Enter the percent of data for each value of k in the table.

k	UNIFORM	NORMAL	EXPONENTIAL	BIMODAL
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____

- b. Comment on the accuracy of the Normal Bell-Shaped Curve for small samples.

Hint: Is the Normal Bell-Shaped Curve still better than the others?

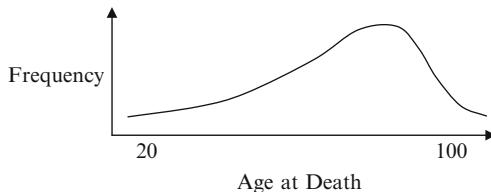
Central Limit Theorem

In some instances, the normal distribution may not be the type of distribution we obtain from sample data when studying research variables and/or the population data may not be normally distributed on which we base our statistics. The normal

probability distribution however is still useful because of the Central Limit Theorem. The Central Limit Theorem states that as sample size increases a sampling distribution of a statistic will become normally distributed even if the population data is not normally distributed. The sampling distribution of the mean of any nonnormal population is approximately normal, given the Central Limit Theorem, but a larger sample size might be needed depending upon the extent to which the population deviates from normality.

Typically, a smaller sample size can be randomly drawn from a *homogeneous* population, whereas a larger sample size needs to be randomly drawn from a *heterogeneous* population, to obtain an unbiased sample estimate of the population parameter. If the population data are normally distributed, then the sampling distribution of the mean is normally distributed; otherwise larger samples of size N are required to approximate a normal sampling distribution. The sampling distribution of the mean is a probability distribution created by the frequency distribution of sample means drawn from a population. The sampling distribution, as a frequency distribution, is used to study the relationship between sample statistics and corresponding population parameters.

The Central Limit Theorem is useful in statistics because it proves that sampling distributions will be normally distributed regardless of the shape of the population from which the random sample was drawn. For example, a physician is studying the life expectancy of adults after being diagnosed with cancer. She is going to do a statistical analysis on the data concerning age at death and needs a theoretical probability distribution to model the adult ages. Since most of the adults lived to an advanced age due to new cancer treatments, she realizes that the population of ages is skewed to the left (see Figure below).



The physician doesn't know whether a mathematical function would best describe this population distribution, but would like to test mean differences in age at death between normal adults and adults with cancer. Fortunately, she can use a sampling distribution of sample means, which doesn't require exact knowledge of the population distribution to test her hypothesis. The Central Limit Theorem, which is based on the sampling distribution of statistics, permits the use of the normal distribution for conducting statistical tests.

Sampling distributions of the mean from non-normal populations approach a normal distribution as the sample size increases, which is the definition of the Central Limit Theorem. The frequency distribution of sample mean based on samples of size N randomly drawn from a population is called the sampling distribution of the mean. The sampling distribution of the mean is a normally distributed

probability distribution. The mean of the sampling distribution of the mean is equal to the mean of the population being sampled. The variance of the sampling distribution of the mean is equal to the variance of the population being sampled divided by the sample size N. Sampling distributions of the mean from normal populations are normally distributed.

CENTRAL R Program

The CENTRAL program will graph the sampling distribution of the mean for samples of a given size N. The random samples will be taken from one of four different population types: uniform, normal, exponential, or bimodal. The frequency distribution of the sample means can be based on an infinite number of samples, but the initial value in the program is set at 250 samples. The sampling distributions of the mean approaches a normal distribution, as sample size increases. Because of this, you will be able to observe how the underlying population distribution type does not affect the normality of the sampling distribution of the mean. The program output shows that the sampling distribution is normally distributed even when data comes from many different types of population distributions.

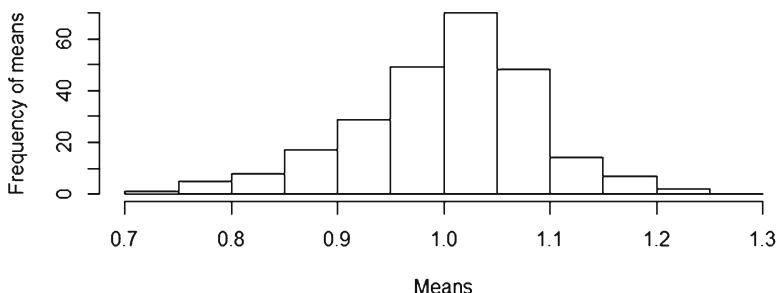
The program begins by initializing the user-defined variables, including selection of the underlying distribution type from which samples are drawn. The main loop iterates for the number of desired replications, creating a sample of the appropriate size from the appropriate distribution. The parameters for the samples are set so that most of the sampling distributions of the mean should fall between 0 and 2 for ease of comparison. After each sample is selected, the mean is calculated and added to the vector of sample means. The entire sample is added to a vector that contains all the raw data from every sample, thereby creating a very large, single sample. When replications are finished, an output vector is created to display the mean and variance for the population distribution and sampling distribution of the mean. Two histograms are graphed, one for the sampling distribution and one for the population distribution.

CENTRAL Program Output

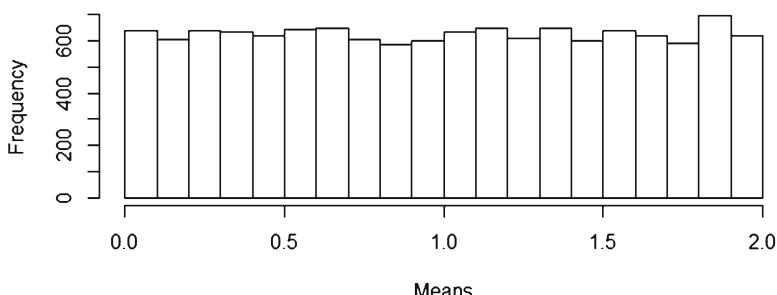
Input values	
Sample Size	50
Number Replications	250
Distribution Type	Uniform

Sampling Distribution Mean=1.00107 Variance=0.00731183
Uniform Distribution Mean=0.9468208 Variance 0.3919674

Sampling Distribution of the Means (Uniform)



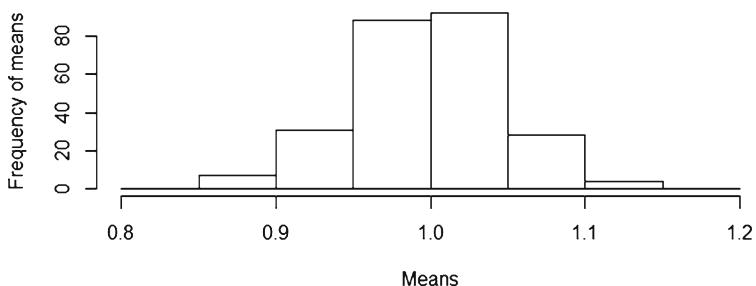
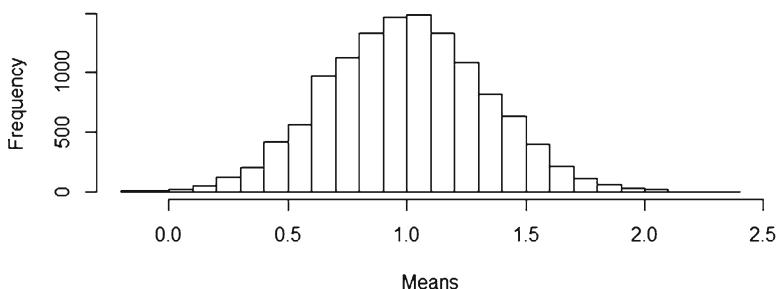
Population Distribution (Uniform Distribution)



Input values

Sample Size 50
 Number Replications 250
 Distribution Type Normal

Sampling Distribution Mean=0.9975896 Variance=0.002228887
 Normal Distribution Mean=1.010424 Variance 0.1147131

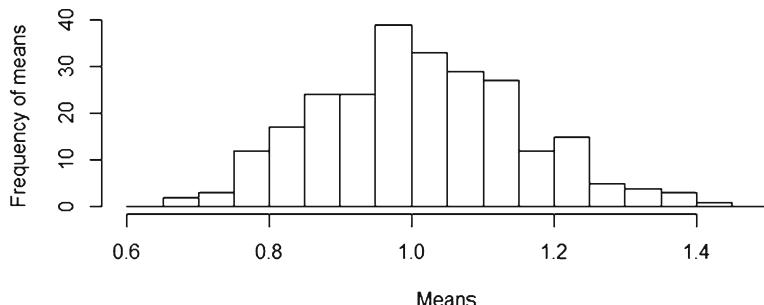
Sampling Distribution of the Means (Normal)**Population Distribution (Normal Distribution)**

Inputvalues

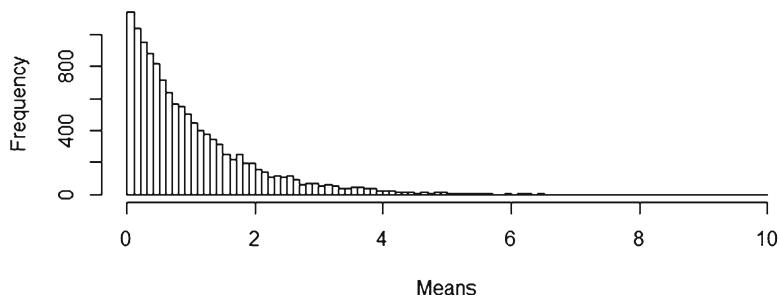
Sample Size	50
Number Replications	250
Distribution Type	Exponential

Sampling Distribution Mean=1.013069 Variance=0.02057427
Exponential Distribution Mean=1.132056 Variance 0.8740562

Sampling Distribution of the Means (Exponential)



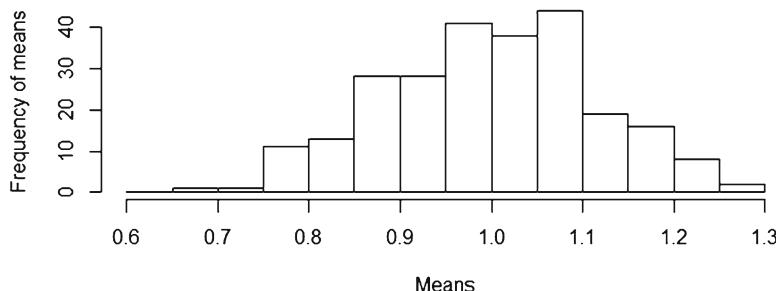
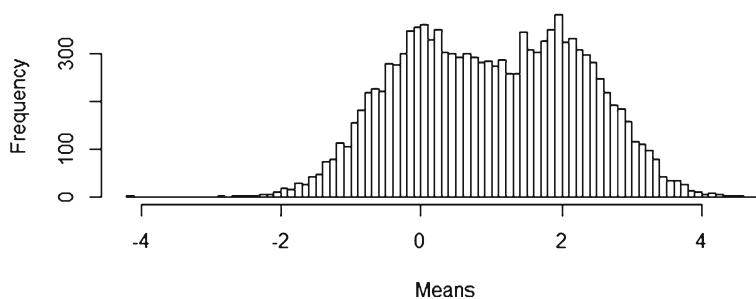
Population Distribution (Exponential Distribution)



Inputvalues

Sample Size	50
Number Replications	250
Distribution Type	Bimodal

Sampling Distribution Mean=0.9995466 Variance=0.01314928
 Bimodal Distribution Mean=1.243848 Variance 1.757248

Sampling Distribution of the Means (Bimodal)**Population Distribution (Bimodal Distribution)****Central Limit Theorem Exercises**

1. Run CENTRAL for sample size $N=5$ for each of the four population types. Enter the tabled output of results below. Print the sampling distribution and population distribution graphs for each of the population distribution types.

UNIFORM		POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN		_____	_____	_____
VARIANCE		_____	_____	_____
<hr/>				
NORMAL		POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN		_____	_____	_____
VARIANCE		_____	_____	_____
<hr/>				

EXPONENTIAL

	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____
BIMODAL			
	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____

- a. Which of the sampling distributions are the most like a normal distribution?

- b. Which sampling distribution is most different from a normal distribution?

- c. Are the sampling distributions approximately normal regardless of the shape of the underlying population? YES _____ NO _____
2. Run the CENTRAL program for samples of size $N=30$ for each population type. Draw a rough graph of each sampling distribution of the mean.

UNIFORM

NORMAL

EXPONENTIAL

BIMODAL

- a. Are the sampling distributions for sample size $N=5$ and $N=30$ different?
YES _____ NO _____
- b. Are the population means and sampling distribution means the same for samples of size $N=30$? Note: ERROR = THEORETICAL - SAMPLING.

			ERROR
Uniform	YES	_____	_____
Normal	YES	_____	_____
Exponential	YES	_____	_____
Bimodal	YES	_____	_____

- c. Are the population variances and sampling distribution variances the same for samples of size $N=30$? Note: ERROR = THEORETICAL – SAMPLING.

		ERROR
Uniform	YES _____	NO _____
Normal	YES _____	NO _____
Exponential	YES _____	NO _____
Bimodal	YES _____	NO _____

3. Run the CENTRAL program again, but this time select a sample size of 100. Answer the following questions.

- a. Is a sample size of 100 sufficiently large to produce a sampling distribution of the mean that is approximately normal regardless of the population type?

YES _____ NO _____

- b. Is the mean of each population type related to the mean of the sampling distribution of the mean?

YES _____ NO _____

- c. Is the variance of each population type related to the variance of the sampling distribution of the mean?

YES _____ NO _____

- d. Would the means and variances for a single sample be equal to the mean and variance of each population type?

YES _____ NO _____

True or False Questions

Central Tendency

- | | | |
|---|---|---|
| T | F | a. Changing the largest value in a data set to a value four times as large does not effect the median. |
| T | F | b. If the mean of a data set is subtracted from every value of the data set, the mean of the modified data is equal to the mean of the initial data set. |
| T | F | c. Multiplying each value in a data set by $\frac{1}{2}$ causes the median of the modified data to be twice as large as the median of the initial data. |
| T | F | d. A data set is modified by adding a constant to each value; the mean of the initial data set can be found by subtracting the constant from the mean of the modified data set. |
| T | F | e. A data set is modified by multiplying all values by 5. The median is now 33.5. The median of the initial data set can be found by multiplying the new median by 0.2 |

Dispersion

- | | | |
|---|---|--|
| T | F | a. Changing the last value in a data set to a value four times as large always multiplies the range by four. |
| T | F | b. Adding five to every data value does not affect the standard deviation. |
| T | F | c. Multiplying each value in a data set by $1/3$ causes the variance of the modified data to be $1/6$ of the original variance. |
| T | F | d. If a data set is modified by adding a constant to each value; the range of the initial data set can be found by subtracting the constant from the range of the modified data set. |
| T | F | e. If a data set is modified by multiplying all values by 5 and the standard deviation is now 33.5; the standard deviation of the original data set can be found by multiplying the new standard deviation by 0.2. |

Sample Size Effects

- | | | |
|---|---|--|
| T | F | a. The sample range usually decreases as the sample size increases. |
| T | F | b. The sample standard deviation decreases as the sample size increases. |
| T | F | c. A rough approximation of the range is four times the variance. |
| T | F | d. The range of a small sample is usually less than the range of a larger sample from the same population. |
| T | F | e. The standard deviation of a uniform population is underestimated if one-fourth of the range is used for the estimate. |

Tchebysheff Inequality Theorem

- | | | |
|---|---|--|
| T | F | a. The Tchebysheff Inequality Theorem gives the approximate percentage of observations within certain intervals of the population. |
| T | F | b. If $k=1$, the Tchebysheff Inequality Theorem states that at least 0% of the data are within one standard deviation of the mean. |
| T | F | c. The Tchebysheff Inequality Theorem is always true and does not depend on the shape of the population. |
| T | F | d. The error in the estimate of the standard deviation is usually larger for larger samples. |
| T | F | e. The Tchebysheff Inequality Theorem states that in a sample, at least 93.75 % of the data are within four standard deviations of the mean. |

Normal Distribution

- | | | |
|---|---|---|
| T | F | a. The Normal Bell-Shaped Curve gives a lower limit for the percentage of the data set within certain intervals. |
| T | F | b. The Normal Bell-Shaped Curve is based on what is known about a normal population and assumes that the sample is large and unimodal. |
| T | F | c. The Normal Bell-Shaped Curve gives accurate estimates for an exponential population. |
| T | F | d. The Normal Bell-Shaped Curve gives good estimates for small data sets. |
| T | F | e. Although the Normal Bell-Shaped Curve gives a more precise statement than the TCHEBYSHEFF lower bound, the Normal Bell-Shaped Curve has the disadvantage that it does not apply to populations of all possible shapes. |

Central Limit Theorem

- | | | |
|---|---|--|
| T | F | a. If a population is normal, the sampling distribution of the mean will also be normal. |
| T | F | b. A frequency distribution of 100 sample means for samples of size N drawn from a population is called the sampling distribution of the mean. |
| T | F | c. The variance of sampling distribution of mean is equal to the variance of the underlying population divided by the sample size. |
| T | F | d. The mean of a sampling distribution of mean is not normally distributed if the underlying population is not normally distributed. |
| T | F | e. A sampling distribution of mean is a probability distribution. |

Chapter 6

Statistical Distributions

Binomial

We have learned that probability and sampling play a role in statistics. In this chapter we show that probability (frequency) distributions exist for different types of statistics; i.e. the **binomial distribution** (frequency distribution of dichotomous data) and **normal distribution** (frequency distribution of continuous data).

Many variables in education, psychology, and business are dichotomous. Examples of dichotomous variables are: boy versus girl; correct versus incorrect answers; delinquent versus non-delinquent; young versus old; part-time versus full-time worker. These variables reflect mutually exclusive and exhaustive categories (i.e., an individual, object, or event can only occur in one or the other category, but not both). Populations that are divided into two exclusive categories are called **dichotomous populations**, which can be represented by the binomial probability distribution. The derivation of the binomial probability is similar to the combination probability derived in Chap. 2.

The **binomial probability distribution** is computed by:

$$P(x \text{ in } n) = \binom{n}{x} P^x Q^{n-x}$$

where the following values are used:

n = size of random sample

x = number of events, objects, or individuals in first category

n - x = number of events, objects, or individuals in second category

P = probability of event, object, or individual occurring in the first category

Q = probability of event, object, or individual occurring in the second category, (1 - P).

Since the **binomial distribution** is a theoretical probability distribution based upon objects, events, or individuals belonging in one of only two groups, the values

for P and Q probabilities associated with group membership must have some basis for selection. An example will illustrate how to use the formula and interpret the resulting binomial distribution.

Students are given five true-false items. The items are scored correct or incorrect with the probability of a correct guess equal to one-half. What is the probability that a student will get four or more true-false items correct? For this example, $n=5$, P and Q are both .50 (one-half based on guessing the item correct), and x ranges from 0 (all wrong) to 5 (all correct) to produce the binomial probability combinations. The calculation of all binomial probability combinations is not necessary to solve the problem, but tabled for illustration and interpretation.

x	n	P^x	Q^{n-x}	Probability
x				
5	1	.5 ⁵	.5 ⁰	1/32 = .03
4	5	.5 ⁴	.5 ¹	5/32 = .16
3	10	.5 ³	.5 ²	10/32 = .31
2	10	.5 ²	.5 ³	10/32 = .31
1	5	.5 ¹	.5 ⁴	5/32 = .16
0	1	.5 ⁰	.5 ⁵	1/32 = .03
				32/32 = 1.00

Using the addition rule, the probability of a student getting four or more items correct is: $.16 + .03 = .19$. The answer is based on the sum of the probabilities for getting four items correct plus five items correct.

The combination formula yields an individual “coefficient” for taking x events, objects, or individuals from a group size n . Notice that these individual coefficients sum to the total number of possible combinations and are symmetrical across the binomial distribution. The binomial distribution is symmetrical because $P=Q=.50$. When P does not equal Q, the binomial distribution will not be symmetrical. Determining the number of possible combinations and multiplying it times P and then Q will yield the theoretical probability for a certain outcome. The individual outcome probabilities add to 1.0.

A binomial distribution can be used to compare sample probabilities to theoretical probabilities if:

- There are only two outcomes, e.g., success or failure.
- The process is repeated a fixed number of times.
- The replications are independent of each other.
- The probability of success in a group is a fixed value, P.
- The number of successes, x, in group size n, is of interest.

A binomial distribution based on dichotomous data approximates a normal distribution based on continuous data when the sample size is large and $P=.50$. Consequently, the mean of a binomial distribution is equal to $n \cdot P$ with variance

equal to $n \cdot P \cdot Q$. A standardized score (**z-score**), which forms the basis for the normal distribution, can be computed from dichotomous data as follows:

$$z = \frac{x - nP}{\sqrt{nPQ}}$$

where:

x = score

nP = mean

nPQ = variance.

A frequency distribution of standard scores (z-scores) has a mean of zero and a standard deviation of one. The z-scores typically range in value from -3.0 to +3.0 in a symmetrical distribution. A graph of the binomial distribution, given $P=Q$ and a large sample size, will be symmetrical and appear normally distributed.

Knowledge of the binomial distribution is helpful in conducting research and useful in practice. Binomial distributions are skewed except for those with a probability of success equal to .50. If $P > .50$, the binomial distribution is skewed left; if $P < .50$, the binomial distribution is skewed right. The mean of a binomial distribution is $n \cdot P$ and the variance is $n \cdot P \cdot Q$. The binomial distribution given by $P(x \text{ in } n)$ uses the combination formula, multiplication and addition rules of probability. The binomial distribution can be used to compare sample probabilities to expected theoretical probabilities. For $P = .50$ and large sample sizes, the binomial distribution approximates the normal distribution.

BINOMIAL R Program

The BINOMIAL program simulates binomial probability outcomes. The number of replications, number of trials, and probability value are input to observe various binomial probability outcomes. Trying different values should allow you to observe the properties of the binomial distribution. The program can be replicated any number of times, but extreme values are not necessary to observe the shape of the distribution. The relative frequencies of x successes will be used to obtain the approximations of the binomial probabilities. The theoretical probabilities, mean and variance of the relative frequency distribution, and error will be computed and printed.

The program must specify *numReplications* to indicate the number of replications, *numTrials* to indicate the number of respondents (or sampling points) per replication, and *Probability* to indicate the probability of success (or population proportion). The initial values are set at 5 respondents (sample size or number of trials), 500 replications, and a population proportion of .50. The program starts by defining these values and then creates a random sample from the binomial distribution of size *numReplications* with *numTrials* sampling points per replication and a probability of success, *Probability*.

BINOMIAL Program Output

Given the following values:

Number of Trials=5
 Number of Replications=500
 Probability=0.5

Mean number of successes=2.58
 Mean expected number of successes=2.5
 Sample variance=1.314
 Expected variance=1.25

	Rel. Freq.	Probability	Error
Successes=0	0.024	0.031	-0.007
Successes=1	0.170	0.156	0.014
Successes=2	0.264	0.312	-0.048
Successes=3	0.322	0.312	0.010
Successes=4	0.184	0.156	0.028
Successes=5	0.036	0.031	0.005

BINOMIAL Exercises

1. Run BINOMIAL for $n=5$, $P=.50$, and 500 replications. Enter the results below.

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0	_____	_____	_____
1	_____	_____	_____
2	_____	_____	_____
3	_____	_____	_____
4	_____	_____	_____
5	_____	_____	_____

- a. What is the maximum absolute value of the errors? _____
 b. Use the $P(x \text{ in } n)$ formula for the binomial distribution to prove $P(3 \text{ in } 5)$ is correct.

 c. What is the mean number of successes in the simulation? _____
 d. What is the mean expected number of successes? _____
 e. What is the sample variance? _____

f. What is the expected variance? _____

g. Use the probabilities to calculate the expected mean and variance.

Mean = _____ Variance = _____

2. Run BINOMIAL for $n=5$, $P=.30$, and 500 replications. Enter the results below.

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0	_____	_____	_____
1	_____	_____	_____
2	_____	_____	_____
3	_____	_____	_____
4	_____	_____	_____
5	_____	_____	_____

a. What is the maximum absolute value of the errors? _____

b. Use the $P(x \text{ in } n)$ formula for the binomial distribution to prove $P(3 \text{ in } 5)$ is correct.

c. What is the mean number of successes in the simulation? _____

d. What is the mean expected number of successes? _____

e. What is the sample variance? _____

f. What is the expected variance? _____

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0	_____	_____	_____
1	_____	_____	_____
2	_____	_____	_____
3	_____	_____	_____
4	_____	_____	_____
5	_____	_____	_____

g. Use the probabilities to calculate the expected mean and variance.

Mean = _____ Variance = _____

3. Run BINOMIAL for $n=5$, $P=.70$, and 500 replications. Enter the results below.

a. What is the maximum absolute value of the errors? _____

b. Use the $P(x \text{ in } n)$ formula for the binomial distribution to prove $P(3 \text{ in } 5)$ is correct.

c. What is the mean number of successes in the simulation? _____

d. What is the mean expected number of successes? _____

e. What is the sample variance? _____

f. What is the expected variance? _____

g. Use the probabilities to calculate the expected mean and variance.

Normal Distribution

In the seventeenth and eighteenth centuries, two mathematicians were asked by gamblers to help them improve their chances of winning at cards and dice. The two mathematicians, who first studied this area of probability, were James Bernoulli and Abraham DeMoivre. James Bernoulli developed the formula for combinations and permutations, and their binomial expansions, which lead to the binomial distribution. Abraham DeMoivre coined the phrase *law of errors* from observing events such as archery matches. The basic idea was that negative errors occurred about as often as positive errors. DeMoivre used this understanding to derive an equation for an *error curve*. DeMoivre in 1733 was credited with developing the mathematical equation for the *normal curve*. In the nineteenth century, Carl Fredrick Gauss (1777–1855), working in the field of astronomy further developed the concept of a mathematical bell-shaped curve and probability. Today, his picture and mathematical equation for the normal curve appear on the deutsche mark currency of Germany.

The normal distribution or normal curve is a mathematical equation for random chance errors. The frequency distribution of many continuous random variables used in research closely approximates the normal distribution. Consequently, the normal distribution is a useful mathematical model in which to study variable relationships in the physical and social sciences.

The mathematical equation for the normal distribution indicates a normal density that is an exponential function of a quadratic form. The normal curve equation defines an infinite number of curves, depending upon the values of the mean and standard deviation. The normal curve equation is defined as:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where:

Y =height of the curve at a given score

X =score at a given height

μ =mean of the X variable

σ =standard deviation of X variable

π =constant equal to 3.1416 (pi)

e =constant equal to 2.7183 (base of natural logarithm)

When a set of X scores are transformed to have a mean of zero (0) and a standard deviation of one (1), which are called standard scores or z -scores, the mathematical equation is reduced to:

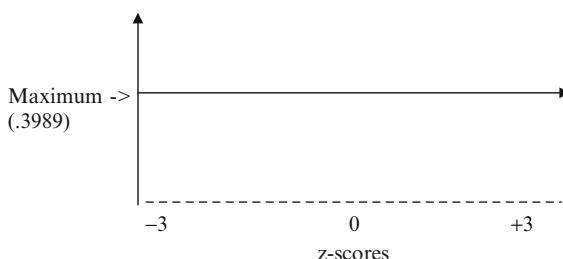
$$Y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

This equation using standard scores is referred to as the *standard normal curve* with z -score values that range primarily from -3 to +3 and correspond to

the ordinates of Y (density or height of the curve). A z score is calculated as: $(X - \text{mean})/\text{standard deviation}$. The tabled values indicate the z-score values between -3 and $+3$ corresponding to each y-ordinate value. A graph of these values yields a normal distribution or bell-shaped curve.

Table of z-score and y-ordinate values

z-score	y-ordinates
-3.0	.004
-2.5	.018
-2.0	.054
-1.5	.130
-1.0	.242
-0.5	.352
0.0	.399
+0.5	.352
+1.0	.242
+1.5	.130
+2.0	.054
+2.5	.018
+3.0	.004



The standard normal distribution has a mean of 0, and a standard deviation of 1. The standard normal distribution is bell-shaped and symmetrical. The probability area under the standard normal curve is 1.0 corresponding to 100 % of the area under the curve. The density function is reasonably complex enough that the usual method of finding the probability area between two specified values requires using integral calculus to calculate the probabilities. The standard normal table of z-values in the appendix of textbooks has been derived by this method of integral calculus. Basically, the normal distribution is an exponential function forming a symmetrical curve, with inflection points at -1 and 1 . The bell-shaped curve approaches but never touches the X-axis. The normal distribution has an infinite number of different curves based upon the values of the mean and standard deviation. The normal distribution curve using z-scores is called the standard normal curve. The probabilities of the bell-shaped curve are based on the normal distribution.

NORMAL R Program

The NORMAL R program approximates standard normal probabilities. Initial values are set for 1,000 random numbers between the z-score intervals –3 and 3 to correspond to ± 3 standard deviations around the true P value. The *DensityHeight* is fixed at .3989 corresponding to the normal distribution. The normal curve height is calculated as $1/\text{SQRT}(2*3.1416)=0.3989$.

The *NumPoints*, *IntervalMin*, and *IntervalMax* variables are user inputted values. *DensityHeight* has a single value since the density function corresponds to a single distribution with z-score values between –3 and +3. These represent the area under the curve for values at the lower end of the interval and the upper end of the interval. These two extra variables are needed because if the lower end of the interval is negative (left of the mean) and the upper end is zero or positive (right of the mean), then the two probabilities are added together to get the total probability. If both are positive, then the lower is subtracted from the upper to leave only the area between the two. If both are negative, then the upper (which would be the smaller absolute difference from the mean) is subtracted from the lower.

The approximation of probabilities in the program are carried out by specifying an interval ($A <--> B$) and determining the probability, P ($A < Z < B$). Each approximation is made on the basis of 1,000 random points with an interval $A <--> B$ corresponding to ± 3 and a maximum height of 0.3989. The theoretical probabilities are computed and printed. In order to avoid entering a standard normal probability table into the program, an approximation is used. It was given by Stephen E. Derenzo in “Approximations for Hand Calculators using Small Integer Coefficients,” *Mathematics of Computation*, 31, 1977, pp. 214–225. For $A \geq 0$, the approximation is:

$$P(Z < A) = 1 - \frac{1}{2} \exp \left[-\frac{((83A + 351)A + 562)A}{703 + 165A} \right]$$

in which $1/2 \exp [x]$ means $1/2 e^x$. This approximation has an error of no more than 0.001 for sample sizes of 10,000 or more.

NORMAL Program Output

```
Sample size=1000 Interval width=6
Interval      Sample P      True P      Error
-3<Z<3       0.986       0.997     -0.011
```

NORMAL Distribution Exercises

1. Run the NORMAL program and complete the following table (NumPoints < - 1000). Draw small graphs in the last column and shade the area that represents the probability.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	GRAPH
0 < Z < 1	_____	_____	_____
-1 < Z < 1	_____	_____	_____
0 < Z < 2	_____	_____	_____
-2 < Z < 2	_____	_____	_____
0 < Z < 3	_____	_____	_____
-3 < Z < 3	_____	_____	_____
1.54 < Z < 2.67	_____	_____	_____
-0.45 < Z < 1.15	_____	_____	_____

2. Run the NORMAL program again with NumPoints < - 10000 for the Z intervals.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	GRAPH
0 < Z < 1	_____	_____	_____
-1 < Z < 1	_____	_____	_____
0 < Z < 2	_____	_____	_____
-2 < Z < 2	_____	_____	_____
0 < Z < 3	_____	_____	_____
-3 < Z < 3	_____	_____	_____
1.54 < Z < 2.67	_____	_____	_____
-0.45 < Z < 1.15	_____	_____	_____

- a. Does NumPoints < - 1000 or NumPoints < - 10000 give better approximations?

- b. In general will larger sample sizes more closely approximate the normal distribution P value?

YES _____ NO _____

3. Compare the IntervalWidth (-3 < Z < +3) and DensityHeight of .3989 to the IntervalWidth (-4 < Z < +4) and DensityHeight of .3989 in the NORMAL program using NumPoints < - 1000.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	DENSITY HEIGHT
-3 < Z < +3	_____	_____	.3989
-4 < Z < +4	_____	_____	.3989

a. Will the approximations be different?

YES _____ NO _____

b. Will the approximations become more similar as sample size increases?

YES _____ NO _____

Chi-Square Distribution

The chi-square distribution, like the other distributions, is a function of sample size. There are an infinite number of chi-square curves based on sample size. In fact, as sample size increases, the chi-square distribution becomes symmetrical and bell-shaped (normal), but with a mean equal to the degrees of freedom (df) and mode equal to df – 2.

The degrees of freedom are related to sample size, because it takes on a value of N – 1. The degree of freedom concept relates to the number of values or parameters free to vary. If a set of five numbers are given, e.g., 5 4 3 2 1, and the sum of the numbers is known, i.e., $\Sigma X = 15$, then knowledge of four numbers implies that the fifth number is not free to vary. For example, four out of five numbers are 10, 15, 25, and 45 with $\Sigma X = 100$. Since the four numbers sum to 95, the fifth number must be 5 in order for the sum of the five numbers to equal 100. This same principle applies to a set of numbers and the mean.

Karl Pearson first derived the chi-square distribution as a frequency distribution of squared *z-score* values. The chi-square statistic was computed as:

$$\chi^2 = \sum \left(\frac{X_i - M}{\sigma} \right)^2$$

with $df = N - 1$. A *z-score* was calculated as $z = [(X - \text{Mean})/\text{Standard Deviation}]$; where X = a raw score, M = the sample mean, and σ = population standard deviation. The *z-score* transformed a raw score into a standard score based on standard deviation units. The population standard deviation (σ) is indicated in the formula, however, a sample standard deviation estimate was generally used because the population value was not typically known.

Chi-square is related to the variance of a sample. If we squared both the numerator and denominator in the previous formula, we would get:

$$\chi^2 = \frac{\Sigma(X - M)^2 / N - 1}{\sigma^2}$$

The numerator of the formula can be expressed as sample variance because $\Sigma(X - M)^2$ represents the sum of squared deviations, denoted as SS, so the sample variance in the numerator can be written as: $S^2 = SS/N - 1$. With a little math, $SS = (N - 1)S^2$. Consequently, if samples of size N with variances, S^2 , are computed from a normal distribution with variance of σ^2 , the χ^2 statistic could be written as:

$$\chi^2 = \frac{(N - 1)S^2}{\sigma^2}$$

with $N - 1$ degrees of freedom. The chi-square statistic is therefore useful in testing whether a sample variance differs significantly from a population variance because it forms a ratio of sample variance to population variance. Since the chi-square distribution reflects this ratio of variances, all chi-square values are positive and range continuously from zero to positive infinity.

The chi-square statistic, $\chi^2 = (N - 1) S^2 / \sigma^2$, computed by taking random samples from a normal population, produces a different chi-square distribution for each degree of freedom ($N - 1$). The chi-square distribution has a mean equal to the degrees of freedom (df) and a mode equal to $df - 2$. The chi-square distribution becomes symmetrical and bell-shaped as sample size increases. The variance of the chi-square distribution is two times the degree of freedom ($2 * df$).

CHISQUARE R Program

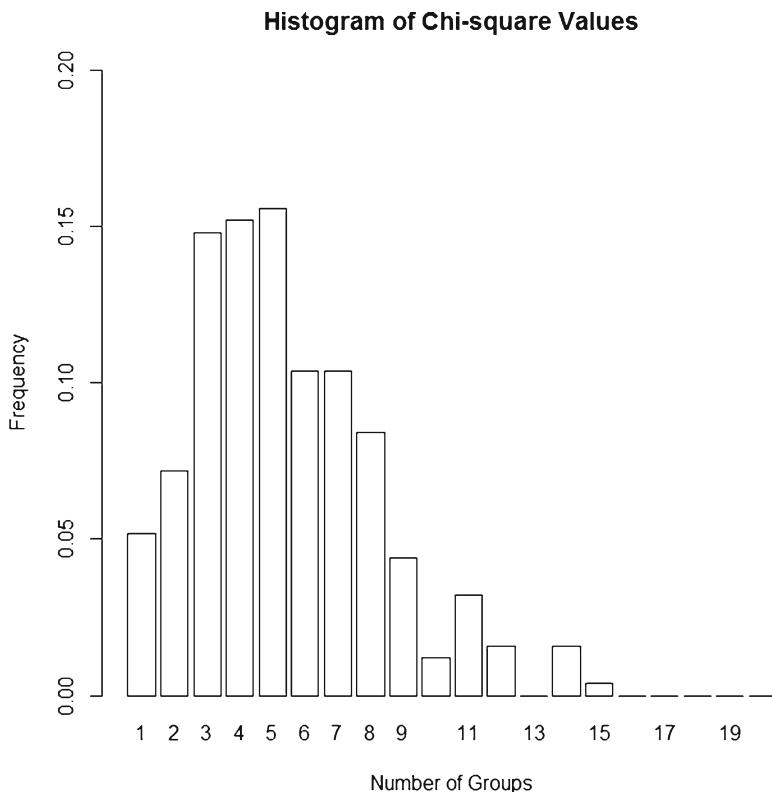
The CHISQUARE R program can produce an unlimited number of chi-square distributions, one for each degree of freedom. For small degrees of freedom, the chi-square distribution should be skewed right. As the degrees of freedom increases, that is, sample size increases, the mode of the chi-square distribution moves to the right. The chi-square values should be positive and range continuously from zero to positive infinity. The program permits selection of different sample sizes from a normal distribution. The program will initially select 250 samples of the desired size. Each time, the chi-square statistic will be calculated. The 250 sample chi-square values are graphed to show the chi-square sampling distribution. The chi-square statistics will be recorded in a relative frequency table. A table is printed with the relative frequencies within each interval. The relative frequencies are graphed using the **barplot function** (histogram) to graphically display the chi-square sampling distribution. The group interval and frequency, along with the title and scaling, are set to the default for the Y and X-axis. Finally, the modal chi-square value and the range of the chi-square values are printed.

CHISQUARE Program Output

Pop. Mean=0
Pop. SD=1
Sample Size=6
N Replications=250

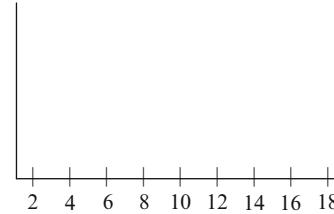
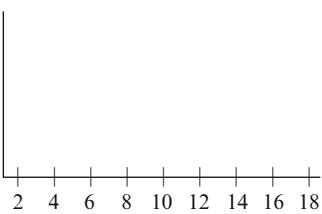
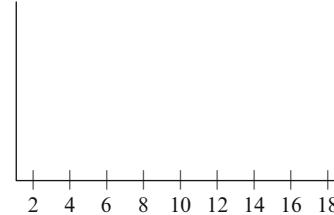
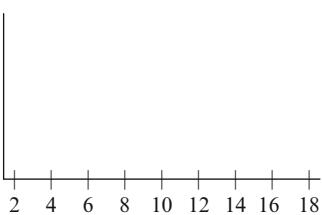
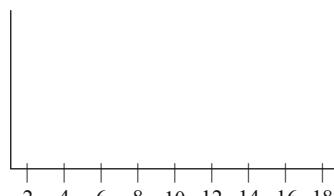
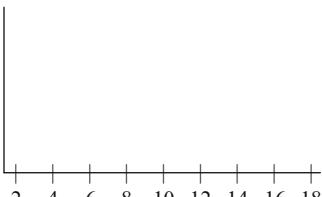
Interval	Rel Freq
(0.0, 1.0)	0.052
(1.0, 2.0)	0.072
(2.0, 3.0)	0.148
(3.0, 4.0)	0.152
(4.0, 5.0)	0.156
(5.0, 6.0)	0.104
(6.0, 7.0)	0.104
(7.0, 8.0)	0.084
(8.0, 9.0)	0.044
(9.0,10.0)	0.012
(10.0,11.0)	0.032
(11.0,12.0)	0.016
(12.0,13.0)	0.000
(13.0,14.0)	0.016
(14.0,15.0)	0.004
(15.0,16.0)	0.000
(16.0,17.0)	0.000
(17.0,18.0)	0.000
(18.0,19.0)	0.000
(19.0,20.0)	0.000

Modal Group=5 Range of chi-square values=21.1



CHISQUARE Exercises

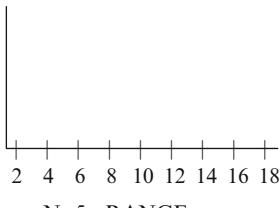
1. Run the CHISQUARE program for each sample size and degrees of freedom listed below; use a population mean of 0, standard deviation of 1, and 250 replications. Graph the shape of the distributions and list the modal group. The modal group is the group with the highest relative frequency.



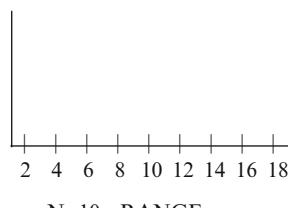
- a. Does the shape of the chi-square distribution change as sample size increases?
 YES _____ NO _____
- b. List the modal group values from the graphs in the table. The mode of the theoretical chi-square distribution is $(DF - 2)$ when $DF > 2$. List the theoretical chi-square mode for each sample size. (ERROR = MODAL GROUP - TRUE MODE)

N	DF	MODAL GROUP	TRUE MODE	ERROR
2	1	_____	Not Applicable	_____
3	2	_____	Not Applicable	_____
4	3	_____	_____	_____
5	4	_____	_____	_____
6	5	_____	_____	_____
7	6	_____	_____	_____

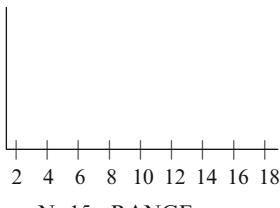
2. Run the CHISQUARE program again for the following sample sizes, but use a population mean of 10, standard deviation of 4, and 250 replications. Graph the shape of the distributions and list the range of the chi-square values. The range is the maximum minus the minimum chi-square value.



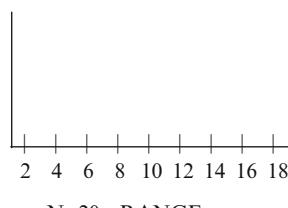
N=5, RANGE _____



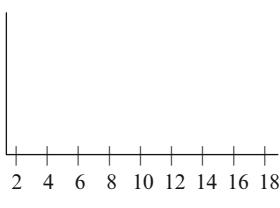
N=10, RANGE _____



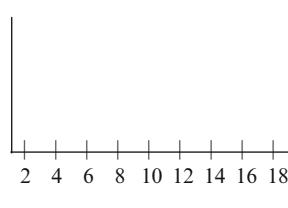
N=15, RANGE _____



N=20, RANGE _____



N=25, RANGE _____



N=30, RANGE _____

- a. Compare these graphs with those in Exercise 1. What differences do you see?

- b. The variance of the theoretical chi-square distribution is two times the degrees of freedom ($2 \times df$). List the theoretical chi-square variance and standard deviation. Divide the range of the chi-square values by four (Range/4) to approximate the standard deviation of the simulated chi-square distribution. (ERROR = APPROXIMATE STANDARD DEVIATION - THEORETICAL STANDARD DEVIATION)

N	σ^2	σ	Range/4	ERROR
5	_____	_____	_____	_____
10	_____	_____	_____	_____
15	_____	_____	_____	_____

N	σ^2	σ	Range/4	ERROR
20	_____	_____	_____	_____
25	_____	_____	_____	_____
30	_____	_____	_____	_____

- c. Does the theoretical standard deviation compare to the estimated standard deviation?

YES _____ NO _____

t-Distribution

The early history of statistics involved probability and inference using large samples and the normal distribution. The standard normal curve provided a probability distribution that was bell-shaped for large samples, but was peaked for small samples, which resulted in larger probability areas in the tails of the distribution. At the turn of the century, a chemist named William S. Gossett, who was employed at a brewery in Dublin, Ireland, discovered the inadequacy of the normal curve for small samples. Gossett was concerned with the quality control of the brewing process and took small samples to test the beer, but didn't obtain a normal bell-shaped curve when his results were graphed.

William Gossett empirically established sampling distributions for smaller samples of various sizes using body measurements of 3,000 British criminals. He started with an approximate normal distribution, drew large samples and small samples, to compare the resulting sampling distributions. He quickly discovered that probability distributions for small samples differed markedly from the normal distribution. William Gossett wrote a mathematical expression for these small sample distributions, and in 1908 he published the results under the pen name, "Student." The Student's t-distribution was a major breakthrough in the field of statistics.

The standard normal distribution is bell-shaped, symmetrical, and has a mean of zero and standard deviation of one. The t-distribution is uni-modal, symmetrical, and has a mean of zero, but not a standard deviation of one. The standard deviation of the t-distribution varies, so when small sample sizes are randomly drawn and graphed, the t-distribution is more peaked (leptokurtic). The probability areas in the tails of the t-distribution are consequently higher than those found in the standard normal distribution. For example, the probability area = .046 in the standard normal distribution at two standard deviations from the mean, but the probability area = .140 in the t-distribution at two standard deviations for a sample size of four. This indicates a greater probability of error using smaller samples. As sample sizes become larger, the t-distribution and standard normal distribution take on the same bell-shaped curve. In fact, the t-values and the z-score values become identical around sample sizes of 10,000, which is within .001 error of approximation as indicated in the previous chapter. Researchers today often use the t-distribution for both small

sample and large sample estimation because it becomes identical to the normal distribution as sample size increases.

In many disciplines, such as education, psychology, and business, variable values are normally distributed. Achievement tests, psychological tests, the height or weight of individuals, and the time to complete a task are examples of variables commonly used in these disciplines. In many instances, the population mean and standard deviation for these variables are not known, but rather estimated from sample data. This forms the basis for making an inference about the population parameters (e.g., mean and standard deviation) from the sample statistics (sample mean and standard deviation). Given small random samples, the t-distribution would better estimate the probability under the frequency distribution curve. Given large random samples, both the standard normal distribution and t-distribution would both yield similar probabilities under the frequency distribution curve.

If the population standard deviation is known, the z-score can be computed as:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Otherwise, the sample standard deviation is used to compute a t-value:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

The sample standard deviation, S , as an estimate of the population standard deviation, σ , is typically in error, thus the sample means are not distributed as a standard normal distribution, but rather a t-distribution. When sample sizes are larger, the sample standard deviation estimate becomes similar to the population standard deviation. Consequently, the shape of the t-distribution is similar to the standard normal distribution. This points out why the estimate of the population standard deviation is critical in the field of statistics. Many researchers attempt to estimate better the unknown population standard deviation by one of the following methods:

1. Use test publisher norms when available (μ, σ)
2. Take an average value from several research studies using the same variable
3. Take large samples of data for better representation
4. Divide the range of sample data by six (see Chap. 5)

The t-distribution is symmetrical, unimodal, and has a mean of zero. The t-distribution has a greater probability area in its tails than the standard normal distribution due to sample estimation of the population standard deviation. The shape of the t-distribution is not affected by the mean and variance of the population from which random sampling occurs. As the sample size increases, the t-distribution becomes similar to the standard normal distribution.

t-DISTRIBUTION R Program

The t-DISTRIBUTION program creates a z distribution of z values and a t distribution of t-values. The program specifies a population mean and standard deviation, sample size, and the number of replications (samples to be taken), which are initially set but can be changed. The program then selects a random sample of that size, computes the sample mean and sample standard deviation, and then the z- and t-statistics. This process will be repeated 250 times. The 250 z- and t-statistics, which arise from these simulations, will be tabulated and printed in a frequency table. By comparing the frequency tables for t and z, you will be able to observe the higher probability in the heavier tails of the t-distribution. By varying the sample size, you will be able to observe how the shape of the t-distribution changes and becomes more normally distributed as the sample size increases.

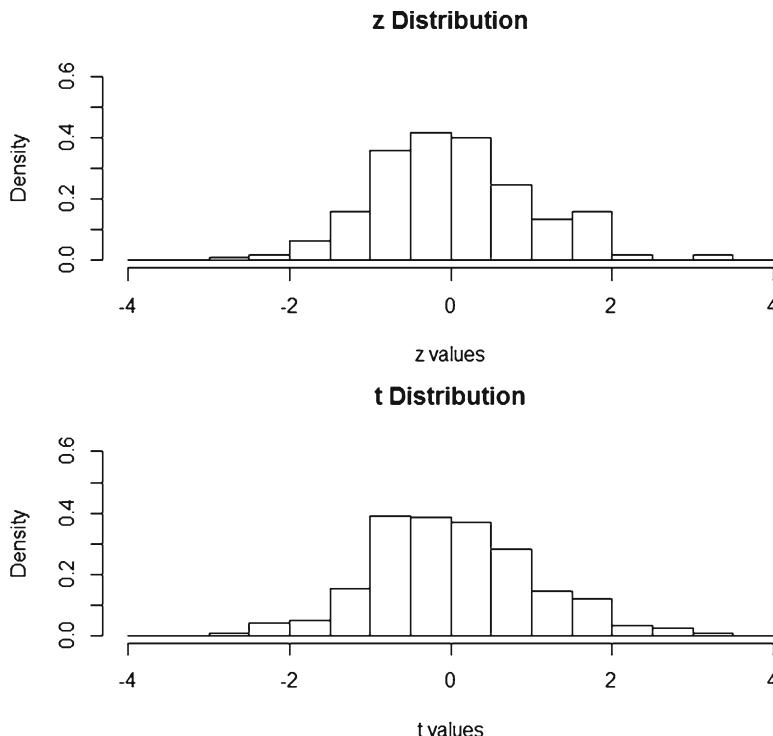
t-DISTRIBUTION Program Output

Pop. Mean=50 Pop. SD=15

Sample Size=30

N Replications=250

Interval	Freq t	Freq z
(-4.0, -3.5)	0.000	0.000
(-3.5, -3.0)	0.000	0.000
(-3.0, -2.5)	0.004	0.004
(-2.5, -2.0)	0.020	0.008
(-2.0, -1.5)	0.024	0.032
(-1.5, -1.0)	0.076	0.080
(-1.0, -0.5)	0.196	0.180
(-0.5, 0.0)	0.192	0.208
(0.0, 0.5)	0.184	0.200
(0.5, 1.0)	0.140	0.124
(1.0, 1.5)	0.072	0.068
(1.5, 2.0)	0.060	0.080
(2.0, 2.5)	0.016	0.008
(2.5, 3.0)	0.012	0.000
(3.0, 3.5)	0.004	0.008
(3.5, 4.0)	0.000	0.000



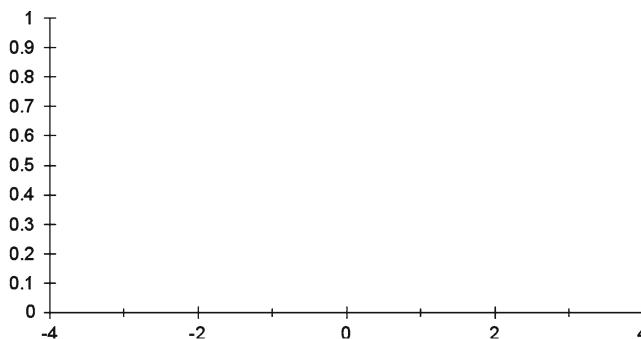
t-DISTRIBUTION Exercises

- Run t-DISTRIBUTION for population mean of 0 and a standard deviation of 1. Use a sample size of 5 and perform 1,000 replications (`popMean <- 0`, `popStdDev <- 1`, `sampleSize <- 5`, `replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(-4.0, -3.5)	_____	_____
(-3.5, -3.0)	_____	_____
(-3.0, -2.5)	_____	_____
(-2.5, -2.0)	_____	_____
(-2.0, -1.5)	_____	_____
(-1.5, -1.0)	_____	_____
(-1.0, -0.5)	_____	_____
(-0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____

INTERVAL	FREQ t	FREQ z
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

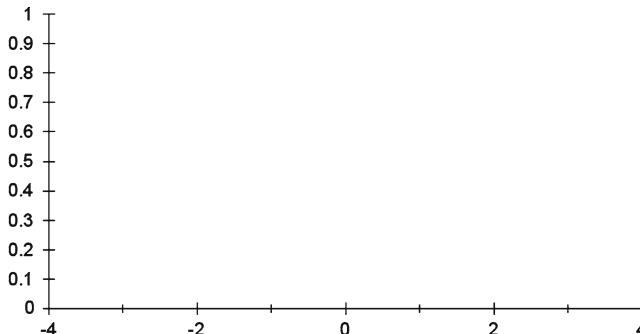
- a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____
- b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____ NO _____



2. Run t-DISTRIBUTION again for population mean of 0 and a standard deviation of 1. Use a sample size of 100 and perform 1,000 replications (`popMean <- 0`, `popStdDev <- 1`, `sampleSize <- 100`, `replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(-4.0, -3.5)	_____	_____
(-3.5, -3.0)	_____	_____
(-3.0, -2.5)	_____	_____
(-2.5, -2.0)	_____	_____
(-2.0, -1.5)	_____	_____
(-1.5, -1.0)	_____	_____
(-1.0, -0.5)	_____	_____
(-0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

- a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____
- b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____ NO _____
- c. As sample size increased from $n=5$ to $n=100$, did the t-statistic distribution more closely approximate a normal distribution? YES _____ NO _____

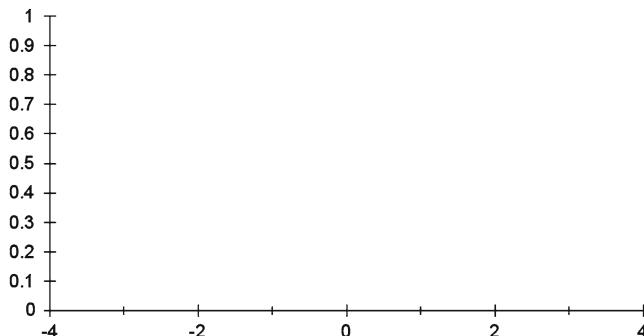


3. Run t DISTRIBUTION again for population mean of 0 and a standard deviation of 15. Use a sample size of 5 and perform 1,000 replications (`popMean <- 0, popStdDev <- 15, sampleSize <- 5, replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(-4.0, -3.5)	_____	_____
(-3.5, -3.0)	_____	_____
(-3.0, -2.5)	_____	_____
(-2.5, -2.0)	_____	_____
(-2.0, -1.5)	_____	_____
(-1.5, -1.0)	_____	_____
(-1.0, -0.5)	_____	_____
(-0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

- a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____

- b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____
NO _____
- c. Is the t-statistic distribution affected by the population standard deviation value? YES _____ NO _____
- d. Is the t-statistic distribution affected by the population mean value?
YES _____ NO _____

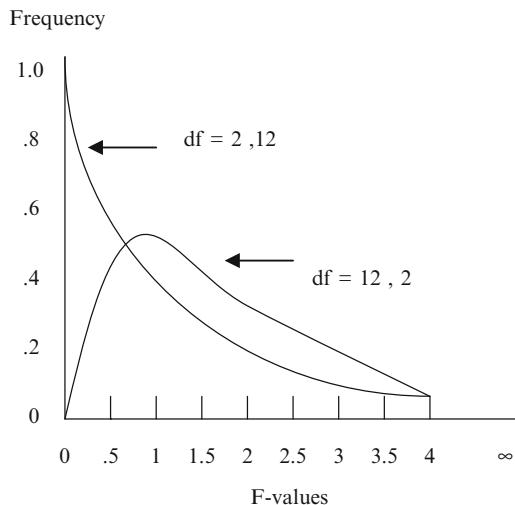


F-Distribution

Sir Ronald Fisher was interested in extending our knowledge of testing mean differences to an analysis of the variability of scores, i.e., variance. He was specifically interested in comparing the variance of two random samples of data. For example, if a random sample of data was drawn from one population and a second random sample of data was drawn from a second population, the two sample variances could be compared as an **F-ratio**: $F = S_1^2 / S_2^2$. The F-ratio is equal to one if the variances of the two random samples are the same. The F-distribution in the appendix reveals this for F-values with $df = \infty, \infty$ in the numerator and denominator. The F-ratio could be less than one, depending upon which sample variances were in the numerator and denominator, but F-values less than one are not considered, so we always place the larger sample variance in the numerator.

If several random samples of data were drawn from each population and the F-ratio computed on the variances for each pair of samples, a sampling distribution of the F's would create the F-distribution. Sir Ronald Fisher determined that like the t-distribution and chi-square distribution, the F-distribution was a function of sample size; specifically the sizes of the two random samples. Consequently, a family of F-curves can be formed based on the degrees of freedom in the numerator and denominator. An **F-curve** is positively skewed with F-ratio values ranging from zero to infinity (∞). If the degrees of freedom for both samples are large, then the F-distribution approaches symmetry (bell-shaped).

Example F-curves for certain degree of freedom pairs can be illustrated as:



Because there are two degrees of freedom associated with an F-ratio, F-tables were constructed to list the F-values expected by chance with the degrees of freedom for the numerator across the top (column values) and the degrees of freedom for the denominator along the side (row values). The corresponding intersection of a column and row degrees of freedom would indicate the tabled F-value. If the computed F-value is greater than the tabled F-value, we conclude that the two sample variances are statistically different at a specified level of probability, e.g., .05 level of significance.

Relationship of F-Distribution to Chi-Square Distribution and t-Distribution

In previous chapters, distributions were graphed based on various sample sizes. We learned that with large sample sizes, sample estimates were closer to population values (z-values) and the sampling distributions (frequency distributions) were more normally distributed. Similarly, the chi-square and t-distributions are also a function of sample size. In fact, as sample size increases, the t, z, and chi-square sampling distributions became symmetrical and bell-shaped (normal). The sampling distributions of the F-ratio operate similar to the t, z, and chi-square family of curves based on sample size.

The t-distribution with degrees of freedom equal to infinity is the normal distribution. Consequently, t-values become equal to z-values when sample sizes are large ($n > 10,000$ to infinity). Check this by referring to the last row of the tabled t-values in the Appendix where you will find that the t-values are the same as the z-values in the normal distribution table. For example, $t = 1.96$ is equal to $z = 1.96$ at the .05 level of significance. The normal distribution can be considered a special case of the t-distribution, because as sample size increases, the t-distribution becomes the normal distribution, i.e., $t\text{-values} = z\text{-values}$.

The F-distribution, with *one* degree of freedom in the numerator and the same degree of freedom in the denominator as the t-test, is equal to the square of the t-distribution value. To check this, refer to the first column of the tabled F-values ($df_1=1$) in the Appendix where you will find that the F-values are the square of the t-values in the t-test table ($df_2=\text{degrees of freedom for t-test}$). For example, if $F=3.84$, then $t=1.96$ for $df_1=1$ and $df_2=\infty$. In fact, since $t^2=z^2=F$ for one degree of freedom given large samples, the t-distribution and normal distribution are special cases of the F-distribution.

The F-distribution values, with degrees of freedom in the denominator equal to infinity, can be multiplied by the F-value numerator degrees of freedom to compute a chi-square value. To check this, refer to the last row of the tabled F-values in the Appendix where you will find that the F-values multiplied by the corresponding numerator degrees of freedom (df_1) equals the chi-square value in the chi-square distribution table. For example, $F=2.21$ for $df_1=5$ and $df_2=\infty$, therefore, $\chi^2=11.05$ with 5 degrees of freedom, i.e., $5*2.21=11.05$! Consequently, the chi-square distribution is also a special case of the F-distribution.

Test of Difference Between Two Independent Variances

The sampling distribution of the F-ratio of two variances cannot be approximated by the normal, t, or chi-square sampling distributions because sample sizes seldom approach infinity, and unlike the normal and t-distributions, F sampling distributions range from zero to infinity rather than from negative infinity to positive infinity. Consequently, the F-distribution, named after Sir Ronald A. Fisher, is used to test whether two independent sample variances are the same or different.

If the variances from two randomly drawn samples are equal, then the F-ratio equals one (largest sample variance over the smallest sample variance), otherwise it increases positively to infinity. The ratio of the two sample variances is expressed as $F=S^2_1/S^2_2$, with a numerator and denominator degrees of freedom. For example, the distance *twenty suburban* housewives traveled to the grocery store varied by 2 miles and the distance *ten rural* housewives traveled to the grocery store varied by 10 miles. We want to test if the suburban and rural mileage variance is equal: $F=10/2=5.0$ with $df_1=9$ and $df_2=19$ (Note: degrees of freedom are one less than the respective sample sizes). We compare this computed F-ratio to the tabled F-value in the Appendix for a given level of significance, e.g., .01 level of significance. We find the 9 degrees of freedom (df_1) across the top of the F-table and the 19 degrees of freedom (df_2) along the side of the table. The intersection of the column and row indicates an F-value equal to 3.52. Since $F=5.0$ is greater than tabled $F=3.52$, we conclude that the *rural* housewives vary more in their mileage to the grocery store than *suburban* housewives. Another way of saying this is that the sample variances are not homogeneous (equal) across the two groups.

Since we conducted this test for only the larger variance in the numerator of the F-ratio, we must make a correction to the level of significance. This is accomplished for any tabled F-value by simply doubling the level of significance, e.g., .01 to .02 level of significance. Therefore, $F=5.0$ is statistically different from the tabled

$F=3.52$ at the .02 level of significance (even though we looked up the F value in the .01 level of significance table).

Test of Difference Between Several Independent Variances

H.O. Hartley extended the F-ratio test to the situation in which three or more sample variances were present, which was aptly named the *Hartley F-max test*. A separate F-max distribution table was therefore created (see Appendix). The Hartley F-max test is limited to using equal sample sizes and sample data randomly drawn from a normal population. However, Henry Winkler in 1967 at Ohio University compared the Hartley F-max, Bartlett, Cochran, and Levene's tests for equal variances in his master's thesis and concluded that the Hartley F-max test was the most robust (best choice) when sample sizes were equal.

Extending our previous example, the distance *twenty-one suburban* housewives traveled to the grocery store varied by 2 miles, the distance *twenty-one rural* housewives traveled to the grocery store varied by 10 miles, and the distance *twenty-one urban* housewives traveled to the grocery store varied by 5 miles. The Hartley F-max test is computed for the following example as follows:

Step 1: Calculate the sample variances.

Urban	$S^2=5$
Suburban	$S^2=2$
Rural	$S^2=10$

Step 2: Calculate F-max test by placing largest variance over smallest variance.

$$F_{\text{max}} = 10/2 = 5.0$$

Step 3: Determine the two separate degrees of freedom for the F-max Table.

k =number of sample variances (column values in table)

$$k=3$$

df =sample size – 1 (row values in the table)

$$df=21 - 1 = 20$$

Step 4: Compare computed F-max to tabled F-max values.

$F_{\text{max}} = 5.0$ is greater than tabled $F_{\text{max}} = 2.95$ for 3 and 20 degrees of freedom at the .05 level of significance. We conclude that the sample variances are *not* homogeneous (not the same) across the three groups.

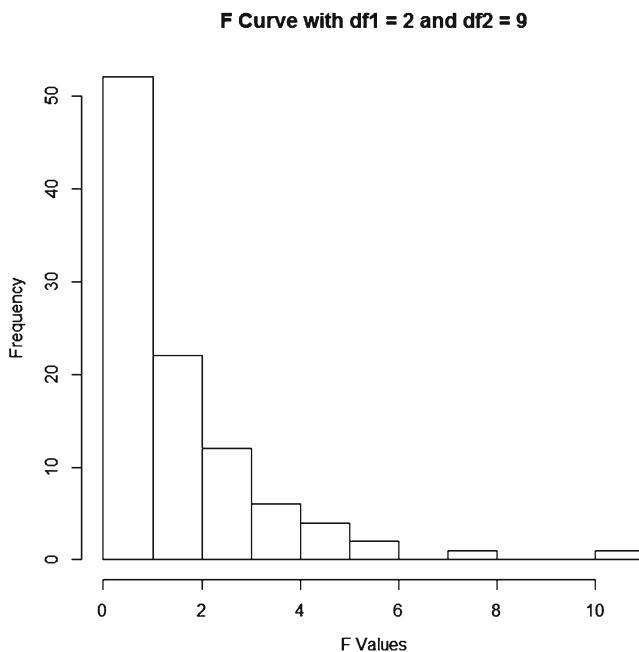
In summary, we find that the F-distribution is a family of frequency curves based upon sample size. The normal (z), t, and chi-square distributions are special cases of the F-distribution. The F-ratio can test whether two independent sample variances are homogeneous. The F-max can test whether three or more independent sample variances are homogeneous. The F-distribution is positively skewed for different numerator and denominator degrees of freedom. As sample sizes increase, the shape of the F-distribution becomes symmetrical.

F-DISTRIBUTION R Programs

The **F-Curve** program simulates F-distributions for given degrees of freedom. It begins by defining the degrees of freedom for the numerator (df1) and denominator (df2) of an F-ratio. Next, the number of replications for the simulation is defined and the random F-values for that number of replications is taken from an F-distribution with the given degrees of freedom. These F-values are plotted in a histogram to show a representation of the F-curve. The F-curve will vary depending upon the degrees of freedom entered in the program.

The **F-Ratio** program inputs the group sizes and variances of two groups, as well as, the alpha level for the significance test. Next the F-ratio is calculated and placed into a display string along with a representation of the actual ratio as a fraction. Then the critical F is determined using the *qf* function, based on the alpha level and degrees of freedom. If the F-ratio is greater than the critical F then the decision is set to “reject,” otherwise it stays at “accept.” Finally, the display string for the F-ratio, the critical F-value, and the decision are placed into a matrix, labels are applied, and the matrix is displayed.

F-Curve Program Output



F-Ratio Program Output

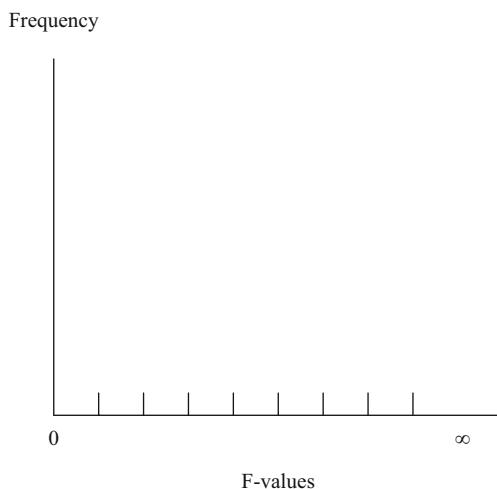
Sample 1: Size=20 Variance=10
 Sample 2: Size=20 Variance=10
 alpha=0.01

F ratio Tabled F Decision
 $10/10=1$ 3.03 accept

F-DISTRIBUTION Exercises

- Run the **F-Curve** program for each pair of degrees of freedom listed below based on 100 replications. Graph the F-curves on the chart. Note: The two samples are randomly drawn from a normal population, sample variances calculated, and the F-ratio computed.

	<u>Sample 1</u>	<u>Sample 2</u>
	<u>df₁</u>	<u>df₂</u>
Run 1:	5	15
Run 2:	15	15
Run 3:	15	5



- Does the shape of the F-curve change based on the numerator and denominator degrees of freedom? YES _____ NO _____
- Are the F-values always positive? YES _____ NO _____

- c. Why are some of the F-values below 1.0?
-
-

2. Run the **F-Ratio** program for the following pairs of sample variances listed below. List the F-ratio, Tabled F-value, and Decision at the .01 level of significance. The *Decision* indicates whether the sample variances are homogeneous.

	Sample 1	Sample 2	F-ratio	Tabled F	Decision
Run 1	n=20 $S^2=10$	n=20 $S^2=10$	_____	_____	_____
Run 2	n=20 $S^2=10$	n=20 $S^2=100$	_____	_____	_____
Run 3	n=40 $S^2=100$	n=20 $S^2=10$	_____	_____	_____
Run 4	n=20 $S^2=10$	n=40 $S^2=100$	_____	_____	_____
Run 5	n=20 $S^2=100$	n=40 $S^2=10$	_____	_____	_____

- a. Does the sample size affect the accept or reject decision? YES _____
NO _____
- b. Can you estimate the sample variance ratio that would yield an F-value, which would lead to a reject decision? YES _____ NO _____
- c. Explain how you determined that the ratio of sample variances led to a reject decision.
-
-
-

3. For the following list of sample variances, compute the F-max test. Find the Tabled F-max value at the .01 level of significance. Decide whether sample variances are homogeneous. Sample size is n=31 fr each sample.

	Sample 1	Sample 2	Sample 3	F-max	Tabled F	Decision
a.	$S^2=10$	$S^2=10$	$S^2=10$	_____	_____	_____
b.	$S^2=10$	$S^2=100$	$S^2=100$	_____	_____	_____
c.	$S^2=40$	$S^2=10$	$S^2=20$	_____	_____	_____

True or False Questions

Binomial Distribution

- | | | |
|---|---|---|
| T | F | a. All binomial distributions are symmetrical. |
| T | F | b. If $n=10$ and $P=.50$, then 50 % of the time $x=5$. |
| T | F | c. The binomial distribution approximates the normal distribution when sample size is large and $P=.50$. |
| T | F | d. If a binomial process consists of n trials, then the number of successes, x , will range from 0 to n . |
| T | F | e. A binomial distribution is created based on dichotomous variables. |
| T | F | f. As sample size increases for $P=.50$, the mean of the binomial distribution (nP) more closely approximates the population mean. |
| T | F | g. As the number of replications increase the absolute value of the error decreases. |
| T | F | h. If $P<.50$, the binomial distribution is skewed right. |
| T | F | i. If $P>.50$, the binomial distribution is skewed left. |

Normal Distribution

- | | | |
|---|---|--|
| T | F | a. The standard normal distribution is a skewed distribution. |
| T | F | b. The value of the standard normal density Y at point Z is the probability that the random variable has a value equal to Z . |
| T | F | c. The standard normal distribution has a mean of 0, and standard deviation of 1. |
| T | F | d. The probability area under the standard normal curve can be approximated in the interval -4 to $+4$. |
| T | F | e. The Monte Carlo approximations of the standard normal probabilities are close to the integral calculus exact theoretical probabilities. |
| T | F | f. As sample size increases, the probabilities more closely approximate a standard normal distribution. |

Chi-Square Distribution

- T F a. The chi-square distribution is independent of the mean and variance of the normal distribution.
- T F b. As the degrees of freedom increases, the variance of the chi-square distribution decreases.
- T F c. The location of the mode in a chi-square distribution moves to the left as sample size increases.
- T F d. For small degrees of freedom, the chi-square distribution is skewed right.
- T F e. Some chi-square values are negative.

t-Distribution

- T F a. The shape of the t-distribution depends on the standard deviation of the population distribution.
- T F b. The smaller the sample size, the larger the probability area in the tails of the t-distribution.
- T F c. The population mean value has *no* effect on the shape of the t-distribution.
- T F d. The t-distribution is symmetrical, unimodal, and mean of zero.
- T F e. For large sample sizes, the t-distribution is the same as the standard normal distribution.
- T F f. The z-statistic distribution will always be normally distributed.

F-Distribution

- T F a. The normal(z), t, and chi-square distributions are special cases of the F-distribution.
- T F b. As sample size increases for both samples, the F-curve becomes symmetrical.
- T F c. The F-ratio tests whether two sample variances are homogeneous.
- T F d. The word “homogeneous” implies that sample variances are different.
- T F e. The F-distribution ranges from zero to infinity.
- T F f. The Hartley F-max test requires equal sample sizes in groups.
- T F g. The F Ratio program could be used to compute a Hartley F-max test.
- T F h. Sample size affects the F-ratio test of equal variances.

Chapter 7

Hypothesis Testing

Sampling Distribution

We ask many questions during our daily activity. When we conduct research we call these daily questions, research questions. Research questions then hypothesize that certain things have occurred or will occur from experimentation. For example, we might ask, Do more men or women frequent the public library? From this question comes a hypothesis: The percentage of women is higher than the percentage of men who frequent the public library. What we need is some way of determining if the percent difference between women and men, or the probability of occurrence, is beyond what would be expected by chance.

We investigate or research variables of interest by obtaining the data and forming a sampling distribution. There are many different sampling distributions, each providing an estimate of its corresponding population parameter. We therefore infer that our sample data will provide an estimate of the population. The sampling distributions provided the basis for creating different types of statistical tests, where hypotheses about the probability of occurrence could be tested. A sampling distribution is a frequency distribution of a statistic created by taking repeated samples of a given size from a population. Consequently, we can create sampling distributions of the mean, variance, standard deviation, range, or median, as well as many other sample statistics, with each providing a sample estimate of a corresponding population parameter.

The statement, “*A statistic is to a sample as a parameter is to a population,*” is a very important concept in statistics. This basic statement reflects the idea behind taking a random sample from a population, computing a statistic, and using that sample statistic as an estimate of the population parameter. Obviously, if the population parameter were known, e.g., mean or standard deviation, then we would not need to take a sample of data and estimate the population parameter.

All sample statistics have sampling distributions with the variance of the sampling distribution indicating the error in the sample statistic, i.e., the error in estimating the population parameter. When the error is small, the statistic will vary less from sample to sample, thus providing us an assurance of a better estimate of the

population parameter. In previous chapters, examples were provided to demonstrate that larger sample sizes yielded smaller error or variance in the sampling distribution, i.e., yielded a more reliable or efficient statistical estimate of the population parameter. For example, the mean and median are both sample statistics that estimate the central tendency in population data, but the mean is the more consistent estimator of the population mean because the sampling distribution of the mean has a smaller variance than the sampling distribution of the median. The variance of the sampling distribution of the mean is called the *standard error of the mean*. It is designated as:

$$\frac{\sigma}{\sqrt{n}}$$

and is estimated in a sample as:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

The sampling distribution of a statistic is a function of sample size. In the formula, it is easy to see that as sample size n becomes larger, the denominator in the formula becomes larger and the standard error of the statistic becomes smaller; hence the frequency distribution of the statistic or sampling distribution has less variance. This indicates that a more precise sample estimate of the population parameter or value is achieved. This concept of standard error of a sampling distribution applies to any sample statistic that is used to estimate a corresponding population parameter.

An important concern in using sample statistics as estimators of population parameters is whether the estimates possess certain properties. Sir Ronald Fisher in the early twentieth century was the first to describe the properties of estimators. The four desirable properties of estimators are (1) *unbiased*; (2) *efficient*; (3) *consistent*; and (4) *sufficient*. If the mean of the sampling distribution of the statistic equals the corresponding population parameter, the statistic is *unbiased*; otherwise it is a biased estimator. If the sampling distributions of two statistics have the same mean, then the statistic with the smaller variance in its sampling distribution is more *efficient* (more precise or less variable) while the other statistic is a less efficient estimator. A statistic is a *consistent* estimator of the population parameter if the statistic gets closer to the actual population parameter as sample size increases. A *sufficient* statistic is one that can't be improved upon using other aspects of the sample data. If several sample statistics compete as an estimate of the population parameter, e.g., mean, median, and mode, the sample statistic that is unbiased, efficient, and consistent is a *sufficient* sample statistic estimate, while the other sample statistics are less sufficient. We are therefore interested in sample estimates of population parameters that are unbiased, efficient, consistent, and sufficient.

The sample mean (statistic) is an unbiased, efficient, sufficient, and consistent estimator of the population mean (parameter). Sample statistics however don't always possess these four properties. For example, the sample standard deviation is a biased, but consistent estimator of the population standard deviation. The sample

standard deviation therefore more closely approximates the population standard deviation as sample size increases, i.e., it is a consistent estimate.

The sampling distributions of sample standard deviations are generated given varying sample sizes. The sample standard deviation is computed as:

$$S = \sqrt{\sum (X - \text{Mean})^2 / (N - 1)}$$

The frequency distribution of sample standard deviations computed from repeatedly drawing samples from a population generates the sampling distributions of the sample standard deviations. A comparison of the mean of the sampling distribution of sample standard deviations to the population standard deviation will help us to determine if the sample standard deviation is a consistent estimator of the population standard deviation. This basic approach can be used to determine whether any sample statistic is a consistent estimator of a corresponding population parameter. Characteristics common to the sampling distributions based on different sample sizes are the most descriptive information we can obtain about theoretical population distributions. A statistic that estimates a parameter is unbiased when the average value of the sampling distribution is equal to the parameter being estimated. The sample standard deviation is an unbiased estimator of the population standard deviation if the mean of the standard deviation sampling distribution is equal to the population standard deviation. In this chapter, you will learn to make a judgment about whether or not the sample standard deviation is a biased or an unbiased estimator of the population standard deviation.

Many different sampling distributions can be generated with the mean of the sampling distribution being an estimator of the population parameter. For each sampling distribution of a statistic, the variance of the sampling distribution indicates how precise the statistic is as an estimator. The variance of a sampling distribution of a statistic becomes smaller as sample size increases, i.e., the standard error of the statistic. The sampling distribution of sample standard deviations is a frequency distribution of sample standard deviations computed from samples of a given size taken from a population. The variance of the sampling distribution of sample standard deviations decreases with increased sample size; thus the sample standard deviation is a consistent estimator. The mean of the sampling distribution of sample standard deviations is less than the population standard deviation; thus the sample standard deviation is a biased estimator of the population standard deviation. The error in the sample standard deviation as an estimate of the population standard deviation decreases with increased sample size.

DEVIATION R Program

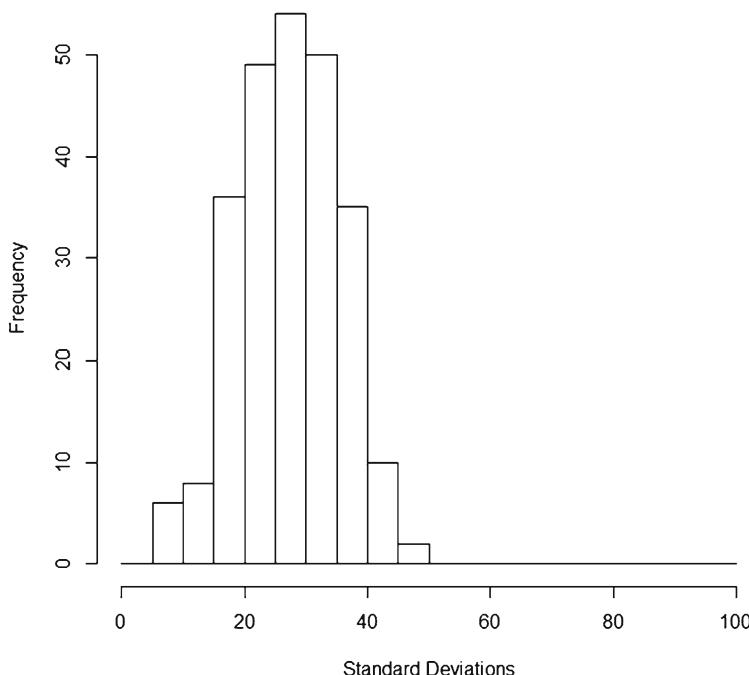
The DEVIATION R program produces the sampling distribution of the sample standard deviation. The program initially specifies a sample size of $n=5$ and then 250 replications are chosen at random from a uniform distribution between 0 and 100.

The sample standard deviation is calculated for each sample and a histogram is printed for these 250 standard deviations. The program generates a single sampling distribution for a given sample size. A different sampling distribution exists for every sample of size N . Consequently, you will need to run the program several times, each time with a different sample size, in order to determine whether the mean of the sampling distribution more closely approximates the population parameter.

The main processing loop iterates from one to the number of desired replications. For each replication, a sample of size, *SampleSize*, is drawn from a uniform population ranging from 0 to 100. The standard deviation of each sample is calculated and added to the distribution of standard deviations in the vector. After the completion of the loop, the mean of the distribution of sampling standard deviations is calculated. The mean, the sample size, and the true population standard deviation are placed into the *HistTitle* variable. Finally, a histogram of the distribution of sampling standard deviations ranging from 0 to 100 is graphed with these values printed.

DEVIATION Program Output

**Sampling Mean = 27.24 Population Standard Deviation = 28.58
Sample Size =5**



Deviation Exercises

1. List the four desirable properties of a sample estimate (statistic) for a population parameter.

a. _____
 b. _____
 c. _____
 d. _____

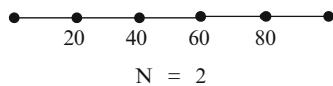
2. Compare the standard error of the mean from the following two sampling distributions.

Note: The standard error of the mean is calculated as: $S_{\bar{x}} = \frac{S}{\sqrt{n}}$

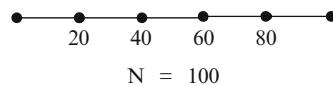
Which sample has the smaller error? _____

Sample 1: S=20, N=100 _____ Sample 2: S=10, N=100 _____

3. Run DEVIATION for the two sample sizes below and draw the histograms.



N = 2



N = 100

Sampling S.D. Mean _____

Sampling S.D. Mean _____

Population S.D. _____

Population S.D. _____

- a. Are the two graphs the same or different? Same _____ Different _____

Why? _____

- b. Which sample size has a Sampling S.D. Mean closer to the Population S.D.?

4. Run the DEVIATION program for the sample sizes listed below. Record the Sampling Distribution S.D. Mean and ERROR for each sample size.

Note: ERROR = SAMPLING S.D. MEAN - POPULATION S.D.

N	SAMPLING S.D. Mean	POPULATION S.D.	ERROR
5	_____	28.58	_____
10	_____	28.58	_____
15	_____	28.58	_____
20	_____	28.58	_____
25	_____	28.58	_____
30	_____	28.58	_____
35	_____	28.58	_____
40	_____	28.58	_____
45	_____	28.58	_____
50	_____	28.58	_____

- a. Does the SAMPLING S.D. MEAN consistently overestimate the POPULATION S.D.? YES _____ NO _____
 - b. Does the variance of the sampling distribution decrease with increased sample size? YES _____ NO _____
 - c. The mean of the sampling distribution doesn't get any closer to the population standard deviation. What do the signs of the differences lead you to think about the direction of the bias when the sample standard deviation is used as an estimator of the population standard deviation?
-
-

Confidence Intervals

Confidence intervals can be computed for many different population parameters by using the standard error of the statistic and the confidence level. A **standard error of a statistic** is computed for each type of sample statistic, e.g., standard error of the mean. The variance of a sampling distribution indicates the amount of error in estimating the population parameter. Smaller sampling variance reflects less error in estimating the population parameter. The standard error of a statistic is computed as the standard deviation of the sampling distribution divided by the square root of the sample size. Consequently, as sample size increases, the standard error of the statistic decreases. A **confidence interval** is computed using the sample statistic and the standard error of the statistic (standard deviation of the statistic in the sampling distribution). The confidence interval around the sample statistic is a range of values that should contain the population parameter. A **confidence level** is used which defines how confident we are that the interval around the sample statistic contains the parameter being estimated. The confidence interval is determined by picking an area in the tail of the sampling distribution in which the value of the statistic is improbable. Recall that a sampling distribution is a frequency distribution; therefore we could pick a 5% probability area, leaving 95% of the sample statistics in the frequency distribution as plausible estimates of the population parameter.

The confidence interval around the sample statistic is computed by using the standard error of the statistic. The **confidence interval** indicates the precision of the sample statistic as an estimate of the population parameter. The confidence interval around the sample statistic is said to include the population parameter with a certain level of confidence. It should be common practice to report confidence intervals for various population parameters such as proportions, means, or correlations. The confidence interval contains a high and low score, above and below the sample statistic, in which we feel confident that the interval contains the population parameter. **Confidence levels** are used to determine the confidence interval width. Commonly used confidence levels are 90%, 95%, or 99%. These confidence levels for

sample data are indicated by a critical t-value of 1.65 (10% probability area), 1.96 (5% probability area), and 2.58 (1% probability area), respectively, which are given in Table 2 (Distribution of t for Given Probability Levels).

The 95% confidence interval for the population mean is computed as:

$$\bar{X} \pm 1.96(S / \sqrt{n})$$

where S = sample standard deviation and n = sample size. The value of 1.96 corresponds to the t-value that contains 5% of the sample means in the tail of the sampling distribution that are improbable. This implies that 5 times out of 100 replications, a confidence interval for a sample mean may not contain the population mean. In contrast, 95 times out of 100, the confidence interval around the sample mean will contain the population mean. Stated differently, 5% of the time the sample mean will not be a good estimate of the population mean, or conversely, 95% of the time we are confident that the sample mean will be a good estimate of the population mean.

If the sample mean was 50, sample standard deviation 10, the sample size 100, and we wanted to be 90% confident that the confidence interval captured the population mean, then the confidence interval around the sample mean would range between 51.65 and 48.35. This range of values is computed as:

$$CI_{.90} = \bar{X} \pm t\left(\frac{S}{\sqrt{n}}\right) = 50 \pm 1.65\left(\frac{10}{\sqrt{100}}\right) = 50 \pm 1.65 \\ CI_{.90} = (51.65, 48.35)$$

If we replicated our sampling ten times, we would expect the population mean to fall in the range of values approximately 90% of the time (9 times out of 10 the confidence intervals would contain the population mean).

The 95% confidence interval using the population standard deviation would be computed as:

$$\bar{X} \pm 1.96(\sigma / \sqrt{N})$$

If the population standard deviation is known, one would use the population standard deviation with a z-value for the confidence interval. If the population standard deviation is *not* known, one would use a sample standard deviation estimate with a critical t-value for the confidence interval.

A confidence interval reflects a range of values (high, low) around the sample mean for different levels of confidence, e.g., 90%, 95%, and 99%. A confidence interval indicates the precision of a sample statistic as an estimate of a population parameter. If a confidence interval has a 95% level of confidence, this indicates that approximately 95 out of 100 confidence intervals around the sample statistic will contain the population parameter. If the confidence level remains the same, but the sample size increases, then the width of the confidence interval decreases, indicating a more precise estimate of the population parameter.

CONFIDENCE R Program

The CONFIDENCE program will simulate random samples and compute the confidence intervals around the sample mean (the process is the same for other population parameters because it would be based on the sampling distribution of the statistic). A population with a normal distribution ($\mu=50$ and $\sigma=10$) will be sampled. You can enter different sample sizes and confidence levels to see the effect they have on estimating the confidence interval. The program uses the population standard deviation rather than the sample standard deviation in the confidence interval formula because it is known. The sampling distribution of the mean will be based on 20 replications. For each sample mean, the 95% confidence interval around the mean will be computed and the program will check to see whether or not the population mean of 50 is contained in the confidence interval. Due to sampling error, one may not achieve the exact percent of confidence intervals that contain the population mean as indicated by the confidence level, i.e., 95%.

The program creates confidence intervals around repeated samples taken from a population and tests to see whether they contain the population mean. The sample size, population mean, population standard deviation, number of replications and size of the confidence interval can be changed in the program. The confidence interval size is specified as a z-value. Samples are simulated for the number of desired replications. The mean of each sample as well as the confidence intervals are calculated according to the formula given in the chapter. There is a count of the number of times the population mean is captured by the confidence interval for the sample. This is expressed as a percentage based on all the replications at the end of the program, both as a ratio and percent. Individual sample information is output, including the sample means and confidence intervals, which capture the population mean.

CONFIDENCE Program Output

```

Pop. Mean = 50 Pop. SD = 10
Sample Size = 100 N Replications = 20
Confidence Intervals for Z value = 1.96
Confidence Intervals that Contain Population Mean = 18 / 20 = 90 %

Sample Mean CI (high - low) Pop. Mean Within CI
48.16      50.12 - 46.2      50      Yes
49.58      51.54 - 47.62     50      Yes
49.42      51.38 - 47.6      50      Yes
49.92      51.88 - 47.96     50      Yes
50.58      52.54 - 48.62     50      Yes
52.78      54.74 - 50.82     50      No
49.34      51.3 - 47.38      50      Yes
50.44      52.4 - 48.48      50      Yes
50.09      52.05 - 48.13     50      Yes
49.62      51.58 - 47.66     50      Yes
49.7       51.66 - 47.74     50      Yes

```

51.05	53.01 - 49.09	50	Yes
50.59	52.55 - 48.63	50	Yes
49.85	51.81 - 47.89	50	Yes
51.29	53.25 - 49.33	50	Yes
50.47	52.43 - 48.51	50	Yes
47.2	49.16 - 45.24	50	No
50.73	52.69 - 48.77	50	Yes
50.04	52 - 48.08	50	Yes
49.7	51.66 - 47.74	50	Yes

Confidence Interval Exercises

1. Run the CONFIDENCE program for ten (10) replications with a confidence level of 90% ($z=1.65$); keep the population mean of 50, standard deviation of 10, and sample sizes of 100.
- How many of the confidence intervals contained the population mean? _____
 - What percentage did you expect to contain the population mean? _____
 - If you increased the number of replications, would you more closely approximate the confidence level percent? Hint: Run CONFIDENCE and change the number of replications to observe what happens. YES _____ NO _____
 - Why does increasing the number of replications not guarantee a more close approximation to the confidence level percent? _____
2. Run the CONFIDENCE program for fifty (50) replications with a confidence level of 90% ($z=1.65$); keep the population mean of 50, standard deviation of 10, and sample sizes of 100.
- Do all the confidence intervals have the same width? YES _____ NO _____
 - What is the width of the confidence interval? _____
 - Compute the confidence interval width using the population standard deviation and the sample size formula in the chapter. _____
- $$\bar{X} \pm 1.65(\sigma / \sqrt{N})$$
- Does the formula give the same confidence interval width as the CONFIDENCE program? Note: $(1.65 * 2 = 3.30$; using the probability area in both tails of the sampling distribution). YES _____ NO _____
3. Run the CONFIDENCE program for each of the sample sizes listed below. Keep the population mean of 50, standard deviation of 10, and set the number of replications to 10 for a 1.96 confidence level (95%). Record the high and low values for the confidence intervals and calculate the width of the confidence interval.

Note: You can obtain the confidence interval width by subtracting the low value from the high value in the outputted table.

Confidence Interval for Z value = 1.96		
Sample Size	CI (High-Low)	CI Width
10	_____	_____
144	_____	_____
256	_____	_____
625	_____	_____

- a. As sample size increases, does the confidence interval width that contains the population mean become smaller? YES _____ NO _____
 - b. If the confidence interval width becomes smaller as sample size increases, does this imply a more accurate estimate of the population mean? YES _____ NO _____
4. Run the CONFIDENCE program for each of the confidence levels listed below. Keep the population mean of 50, standard deviation of 10, sample size of 100, and set the number of replications to 100. Record the CI high and low values and the percent of confidence intervals that contained the population mean.

Confidence Level	CI (High-Low)	Percent
90	_____	_____
95	_____	_____
99	_____	_____

- a. Does the confidence interval become wider as the confidence level increases from 90% to 99%? YES _____ NO _____
- b. If the confidence interval becomes wider, does this imply that we are more confident to have captured a range of values that contains the population mean. YES _____ NO _____

Statistical Hypothesis

The scientific community investigates phenomena in the world. The areas for scientific inquiry are many and have led to the creation of numerous academic disciplines, e.g., botany, biology, education, psychology, business, music, and so forth. The first step in any academic discipline that conducts scientific investigation is to ask a research question. Research questions can be expressed in many different ways. For example, “In the upcoming election, who will be elected President of the United States?” or “Which is better, margarine or butter, in lowering cholesterol?” The next important step is to design a study, then gather data and test the research question. This requires converting the research question into a statistical hypothesis.

There are many different kinds of statistical hypotheses depending upon the level of measurement (nominal, ordinal, interval, or ratio) and type of research design used in the study. A statistical hypothesis is the cornerstone to testing the two possible outcomes, which are always stated in terms of population parameters, given the kind of data collected (percents, ranks, means, or correlation coefficients). The two possible outcomes of a statistical hypothesis are stated in a null (H_0 : no difference) and alternative (H_A : difference exists) format using symbols for the population parameter. The alternative statistical hypothesis is stated to reflect the outcome expected in the research question. This involves either a directional (greater than) or non-directional (difference exists) expression. The null hypothesis is the corresponding opposite expected outcome of less than/equal or no difference, respectively. A research question and statistical hypothesis for each type of data is listed.

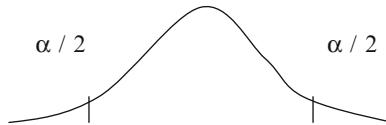
<u>Research question</u>	<u>Data</u>	<u>Statistical hypothesis</u>
Is the percent of people drinking beer in Texas greater than the national average?	Percents	$H_0: P_{\text{Texas}} \leq P_{\text{National}}$ $H_A: P_{\text{Texas}} > P_{\text{National}}$
Is there a difference in the ranking of persons on two different diets for weight gain?	Ranks	$H_0: R_{\text{Diet A}} = R_{\text{Diet B}}$ $H_A: R_{\text{Diet A}} \neq R_{\text{Diet B}}$
Does my 5th grade class on average score higher than the national average in math?	Means	$H_0: \mu_{\text{Class}} \leq \mu_{\text{National}}$ $H_A: \mu_{\text{Class}} > \mu_{\text{National}}$
Is the relationship between music ability and self-esteem in my sample of students different than the population?	Correlation	$H_0: \rho_{\text{Sample}} = \rho_{\text{Population}}$ $H_A: \rho_{\text{Sample}} \neq \rho_{\text{Population}}$

Two kinds of errors are associated with our decision to retain or reject the null hypothesis based on the outcome of the statistical test (TYPE I error and TYPE II error). The TYPE I error is specified by selecting a level of significance (probability area) such that if the sample statistic falls in this probability area, then we reject the null hypothesis in favor of our alternative hypothesis. The TYPE II error corresponds to the probability of retaining the null hypothesis when it is false. When we state the statistical hypothesis as “greater than,” we designate only one-tail of the sampling distribution of the statistic because of the directional nature of the research question. When we state the statistical hypothesis as a “difference exists,” we designate both tails of the sampling distribution of the statistic because of the non-directional nature of the research question. Consequently, the probability area corresponds to different “tabled statistics.”

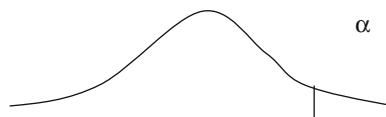
Once the probability area is determined for the statistical hypothesis, we can select a tabled statistical value to set our region of rejection. The tabled statistical values were generated using probability theory and created from the sampling distribution of the statistic for different sample sizes (degrees of freedom) and levels of significance. Only the more popular levels of significance (.05, .01, and sometimes .001) are included in the tables due to page length consideration. Consequently, it is common for researchers to select a region of rejection and test statistical hypotheses based on the .05, .01, or .001 levels of significance. The relationship between the level of

significance and probability area for the region of rejection (vertical hash marks) can be depicted as follows:

Non-directional (two-tailed) Research Question:



Directional (one-tail) Research Question:

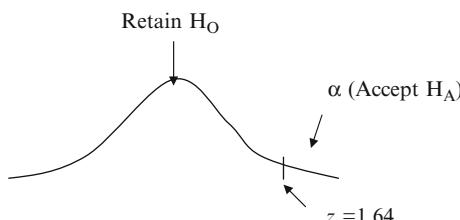


An example research question and corresponding statistical hypothesis will help to illustrate the relationship between the non-direction and/or direction of the question, level of significance, and region of rejection. The research question, “Is the SAT population mean in Texas greater than the SAT population mean for the U.S.,” is converted into a null and alternative statistical hypothesis:

$$H_O : \mu_{\text{Texas}} \leq \mu_{\text{U.S.}}$$

$$H_A : \mu_{\text{Texas}} > \mu_{\text{U.S.}}$$

This is a directional research question, hence the alternative statistical hypothesis indicates a greater than expression, while the null statistical hypothesis indicates less than or equal to for the population parameters. We test our hypothesis by random sampling of 100 students in Texas, compute the sample mean, and conduct a statistical test at the .05 level of significance. Once we have selected a sample size and level of significance for the study, a tabled statistical value can be selected. Under the normal probability curve in Table 1, a z-value of 1.64 corresponds to a probability value (p-value) that indicates an area approximately equal to .05 (probability area beyond z). This z-value is used to set the region of rejection for testing our statistical hypothesis: $R_{.05} = z > 1.64$. When we conduct a z-test for the difference between the means, a computed z-value greater than 1.64 will imply that the population mean in Texas is greater than the population mean in the U.S. In other words, the computed z-value falls in the probability area of the sampling distribution of the statistic that we have designated to be a highly improbable outcome. The probability area for the null hypothesis and the alternative hypothesis is therefore depicted along with the tabled z-value and level of significance for a directional research question as follows:



Given our selection of the .05 level of significance, only 5% of the time would we expect a mean difference to exceed $z=1.64$ if the null hypothesis is true, i.e., the population means are equal. Consequently, if the mean difference is large enough and we compute a z-value greater than 1.64, we are 95% confident in our decision to reject the null hypothesis in favor of the alternative hypothesis.

Assume we computed a Texas SAT mean of 530 for the 100 randomly sampled students and the U.S. SAT mean was 500 with a population standard deviation of 100. A one-sample z-test to determine the statistical significance of this mean difference is computed as:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{530 - 500}{100 / \sqrt{100}} = \frac{30}{10} = 3.0$$

Since the computed $z=3.0$ is greater than $z=1.64$, and therefore falls in the region of rejection, we reject the null hypothesis of no difference in favor of the alternative hypothesis that the population SAT mean in Texas is greater than the population SAT mean in the U.S. We are 95% confident in our decision, but also know that 5% of the time our decision might be wrong (TYPE I error). We would answer the research question by stating that the population SAT mean in Texas is statistically significantly higher than the population SAT mean in the U.S.

This example used a z-test because the population standard deviation for the SAT was known. In many research studies, the population standard deviation is unknown, so we would use a t-test. The t-test formula that uses the *sample standard deviation* in place of the population standard deviation is:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{N}}$$

Research questions involving a test of differences in population means are commonplace in several academic disciplines. An engineer needs to know the average weight of vehicles that can safely travel across a bridge. A psychologist needs to test whether a group of clients have an average cognitive ability greater than the national average. A sociologist needs to determine the average family size for different ethnic groups. The auditor of a large oil company wants to know the average amount of error in the bills prepared for customers. A doctor studying the possible reduction in blood pressure caused by a new medicine is concerned about the average reduction in blood pressure for patients. These and many other research questions reflect a need for a statistical procedure to test whether population means are statistically different.

Tests of statistical hypotheses do not provide exact outcomes. Instead, the tests are based on data-gathering evidence to reach a conclusion with some degree of uncertainty. In other words, it is possible to reject the null hypothesis when in fact the null hypothesis is true. We preset the probability of making this kind of error (TYPE I error) when selecting the level of significance and corresponding tabled statistic, which is based on the sampling distribution of the statistic. We have already learned that increasing sample size, using a directional hypothesis,

willingness to detect a larger difference, and the level of significance are factors that influence whether the statistical test is powerful enough to detect a difference when in fact one exists.

In testing whether the population SAT mean in Texas was significantly different from the population SAT mean in the U.S., knowledge of the population standard deviation and sample size played a role. A larger random sample of students would dramatically reduce the standard error in the formula, which would result in a larger computed z-value, e.g., $z=30$ if $N=10,000$. A smaller sample size would result in a smaller z-value. If the population standard deviation is *not* known, a sample standard deviation as an estimate might produce a very different result.

In hypothesis testing, the null hypothesis is retained as true unless the research findings are beyond chance probability (unlikely to have occurred by chance). A TYPE I error occurs when a true null hypothesis is rejected erroneously, usually due to an atypical research outcome. The region of rejection is specified by the level of significance (α), which indicates the probability of making a TYPE I error. If the z-value falls in the region of rejection probability area, then the null hypothesis is rejected in favor of the alternative hypothesis. For different values of alpha (levels of significance) and sample size, the region of rejection will be indicated by a tabled statistic from the sampling distribution of the statistic (Appendix). A TYPE II error occurs when a false null hypothesis is retained erroneously, usually due to insufficient data.

HYPOTHESIS TEST R Program

The HYPOTHESIS TEST program allows you to specify the true population mean and then test various null and alternative hypotheses. In the program output, you will be able to observe either z-tests or t-tests for several statistical hypotheses depending upon whether the population standard deviation (z-test) or sample standard deviation (t-test) is used. The program will select a random sample of size N from the true population, compute the sample mean and variance, compute the appropriate test statistic, the p-value, and indicate the decision outcome. Since you specify the true population mean, you will know whether or not the decision to reject the null hypothesis is correct. If the population standard deviation is used (set varUse=0, the default), the z-statistic is reported. If the sample standard deviation is used (set varUse=1), the t-statistic is reported.

The program uses the **pValue** function in R to determine the probability in the tails for either the z-test or t-test after all the user-defined variables are initialized. Based on the direction of the statistical hypothesis for the z-test or t-test, probability values are determined using the **pnorm** function, which returns the probability value for a mean difference with a given standard deviation. If the sample mean is less than the null hypothesis mean, then **pnorm** is reported; if the sample mean is greater than the null hypothesis mean, then **1-pnorm** is reported. This use of the

pnorm function results in only the probability area for one-tail of either the normal- or t-distribution. For a two-tailed test, the probability area (alpha) is divided evenly between the two ends of the distributions. The program specifies the probability area for one- and two-tailed tests by selecting a value for the **tails** variable. The program default is $p < .05$, therefore if the computed p-value is less than .05, the decision is reject the null. If the p-value is greater than .05, the decision is retain the null. The number of statistical tests computed and printed is based upon the value for **numSamples**.

HYPOTHESIS TEST Program Output

```

z Statistic
Pop. Mean=10.5 Pop. Variance=2 Null Mean=10
Sample Size=36 Alpha=0.05 Number of Samples 10
Variance type=0 (0=population; 1=sample)
Hypothesis direction=1 (0<Null, 1>Null, 2=two-tailed)

Sample Mean    Pop. SD z-statistic    Decision      p-value
10.251        1.414  1.064        RETAIN NULL   0.144
10.855        1.414  3.626        REJECT NULL   0.001
10.521        1.414  2.209        REJECT NULL   0.014
10.262        1.414  1.113        RETAIN NULL   0.133
10.265        1.414  1.126        RETAIN NULL   0.131
10.65          1.414  2.756        REJECT NULL   0.003
10.704        1.414  2.985        REJECT NULL   0.002
10.719        1.414  3.049        REJECT NULL   0.002
10.629        1.414  2.668        REJECT NULL   0.004
10.653        1.414  2.771        REJECT NULL   0.003

t statistic
Pop. Mean=10.5 Pop. Variance=2 Null Mean=10
Sample Size=36 Alpha=0.05 Number of Samples 10
Variance type=1 (0=population; 1=sample)
Hypothesis direction=1 (0<Null, 1>Null, 2=two-tailed)

Sample Mean    Sample SD  t-statistic    Decision      p-value
10.429        1.226  2.097        REJECT NULL   0.018
10.678        1.561  2.607        REJECT NULL   0.005
10.784        1.568  2.998        REJECT NULL   0.002
10.681        1.469  2.783        REJECT NULL   0.003
10.47          1.446  1.95       REJECT NULL   0.026
10.036        1.423  0.151       RETAIN NULL   0.441
10.021        1.817  0.07       RETAIN NULL   0.473
9.919          0.933  -0.521      RETAIN NULL   0.699
10.338        1.416  1.433       RETAIN NULL   0.076
10.309        1.301  1.423       RETAIN NULL   0.078

```

Hypothesis Testing Exercises

1. Run the HYPOTHESIS TEST program with the following initial values which defines a one-tail test and uses the population variance:

```
popMean <- 10
popVar <- 2
nullMean <- 10
tails <- 1
sampleSize <- 36
alpha <- .05
varUse <- 0
numSamples <- 10
```

Sample	Sample Mean	z-statistic	Decision	p-value
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

- a. Are the z-statistics correct in the table (Use the formula below to verify)?

YES _____ NO _____
$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

- b. Compare each p-value to $\alpha=0.05$. If $p \geq 0.05$, then the decision should be to RETAIN the null hypothesis. If $p < .05$, then the decision should be to REJECT the null hypothesis. Do all of the decisions correspond to the p-values?
YES _____ NO _____

- c. Since the null hypothesis is actually true, any REJECT NULL decisions are errors. This is a TYPE I error. The probability of a TYPE I error is equal to α . How many REJECT NULL decisions were there? What percent is this?
-

2. Run HYPOTHESIS TEST program again using all of the same values, except increase the number of samples to 40 (numSamples <- 40).

- a. What is the percent of TYPE I errors in the 40 simulations? _____
b. What do you expect this TYPE I error percent to be? _____

3. Run HYPOTHESIS TEST program again, except change alpha to .10 (alpha <- .10).

- a. How many times was the null hypothesis rejected? _____
b. What do you expect this percent to be? _____

4. Run HYPOTHESIS TEST program again, except increase the sample size to 100 (sampleSize<-100).
- How many times is the null hypothesis rejected? _____
 - Does sample size have an effect on the percent of TYPE I errors? _____
5. Run HYPOTHESIS TEST program again, except use the sample variance (varUse<-1). In most cases the proportion of rejections will be much higher because the sample standard deviation is not as good an estimator as the true population standard deviation.
- What is the percent of rejections? _____
 - Does the percent of rejections approximate α ? YES _____ NO _____
6. Run the HYPOTHESIS TEST program in which the null hypothesis is false (The population mean is actually greater than the null mean). Run the program five times using a different sample size each time. Use the following settings:

```
popMean <- 10.5
popVar <- 2
nullMean <- 10
tails <- 1
sampleSize <- 36 (Use different sample sizes: 25, 16, 8, 2)
alpha <- .05
varUse <- 0
numSamples <- 20
```

Record the number of times the null hypothesis was retained or rejected for each run. Compute the percentage of null hypotheses that were retained. This is the probability of a TYPE II error.

Sample Size	Number Rejected	Number Retained	Percent Retained
36			
25			
16			
8			
2			

- Which sample size had the greatest probability of a TYPE II error?
36 ____ 25 ____ 16 ____ 8 ____ 2 ____
 - What effect does sample size have on the probability of a TYPE II error?
Hint: A TYPE II error is when you retain a false null hypothesis. _____
7. Run the HYPOTHESIS TEST program again in which the null hypothesis is false. Use alpha<- .10. Run the program five times using the same sample sizes as before. The program settings should be:

```
popMean <- 10.5
popVar <- 2
nullMean <- 10
```

```

tails <- 1
sampleSize<- 36 (Use different sample sizes: 25, 16, 8, 2)
alpha <- .10
varUse <- 0
numSamples <- 20

```

Sample Size	Number Rejected	Number Retained	Percent Retained
36			
25			
16			
8			
2			

- a. Which sample size had the greatest percent retained (probability of a TYPE II error)? 36 ____ 25 ____ 16 ____ 8 ____ 2 ____
- b. Did an increase in alpha from .05 to .10 *decrease* the percent retained (probability of a TYPE II error)? YES ____ NO ____
- c. Did an increase in alpha from .05 to .10 *increase* the percent rejected (probability of a TYPE I error)? YES ____ NO ____
8. Run the HYPOTHESIS program again, except this time use a two-tailed test. Run the program five times using the same sample sizes as before. The program settings should be:

```

popMean<- 10.5
popVar<- 2
nullMean<- 10
tails<- 2
sampleSize<- 36 (Use different sample sizes: 25, 16, 8, 2)
alpha<- .10
varUse<- 0
numSamples<- 20

```

Sample Size	Number Rejected	Number Retained	Percent Retained
36			
25			
16			
8			
2			

- a. Does a two-tailed test, in comparison to the previous one-tail test, result in more Null Hypotheses being retained? YES ____ NO ____
- b. Does sample size affect the results? YES ____ NO ____

TYPE I Error

Science, as a way of understanding the world around us, has for centuries encompassed the classification, ordering, and measuring of plants, characteristics of the earth, animals, and humans. Humans by their very nature have attempted to understand,

explain, and predict phenomena in the world around them. When individuals, objects, or events are described as a well-defined, infinite population, random sampling and probability statistics can play a role in drawing conclusions about the population. The problem in Science is essentially one of drawing conclusions about the characteristics of infinitely large populations.

The use of a random sample to compute a statistic as an estimate of a corresponding population parameter involves some degree of uncertainty. For example, the sample mean doesn't always fall close to the population parameter (sampling error) and therefore isn't always accurate in the estimation of the population parameter. However, we can use the number of times the sample mean falls in an area under the sampling distribution to indicate a degree of confidence. Recall, the areas were designated as 68%, 95% or even 99%. These areas are called the confidence level (probability of occurrence). The number of times the sample mean falls outside these areas (confidence interval) we refer to the result as committing a **TYPE I Error** (probability of non-occurrence). The **TYPE I Error** therefore indicates the amount of uncertainty or probability of error, especially when making a decision about a population parameter. We generally make a decision about a population parameter in the context of a research question.

The research question could be whether or not to use a new experimental drug for the treatment of a disease. In other words, how confident are we that the drug will work. It could also be whether or not to spend money on an innovative math program for High School students. How confident are we that the innovative math program will be better than the traditional math program? In business, the research question of interest could be whether or not to implement a new strategic plan. We conduct research, formally or informally, to answer these types of questions. In simple terms, we ask a question, gather data, and then answer the question. This is the essence of the research process. However, in making our decision to release a drug for public use or spend thousands of dollars on an innovative math program, we can never be 100% certain it will work.

The research process involves the formulation of a question that can be tested, the collection of relevant data, the analysis and presentation of the data, and the answering of the question. This formal research process embodies the use of the following scientific principles:

- a. Statement of the Research Hypothesis
- b. Selection of Sample Size and Sample Statistic
- c. Selection of Confidence Level and Region of Rejection
- d. Collection and Analysis of Data
- e. Statistical test and Interpretation of Findings

In conducting research using random samples from a population, our research hypothesis is related to the probability of whether an event occurs or not. The probability of an event occurring and the probability of an event not occurring are equal to 100% (see Chap. 2). As researchers, we accept some level of probability or uncertainty as to whether an event occurs or not. Our statistical test and interpretation of findings are linked to the **TYPE I Error** in our decision?

The statement of the research hypothesis expresses the outcomes expected in the research question. The outcomes are expressed as a Null Hypothesis and an Alternative Hypothesis. The **Null Hypothesis** is a statement of no difference between the sample mean and population parameter, or in general terms between a sample statistic and a population parameter. The **Alternative Hypothesis** is a statement that a difference exists between the sample mean and population parameter, typically based upon some intervention, treatment, or manipulation of participants, objects, or events. For example, the Null Hypothesis would state *no difference* in the average mathematics test scores of students in a traditional math program versus an innovative math program. The Alternative Hypothesis would state that the average mathematics test score of students in the innovative math program is greater (*statistically different*) than the average mathematics test score of students in the traditional math program.

The sample size and sample statistic chosen to make a comparison between the sample mean and population mean is determined next in the research process. The sample size is an important consideration because as sample size increases the sample statistic more closely approximates the population parameter. The sample statistic is computed by taking the difference between the sample mean and the population mean, or between two sample means, divided by the standard error of the mean difference, which is based on the sampling distribution of the statistic. We will learn more about conducting these types of statistical tests in later chapters.

The confidence level and region of rejection are now established which set the amount of uncertainty we are willing to accept for the two possible outcomes of the research question, i.e., null versus alternative. If we want to be 95% confident in our decision that the innovative math program produced higher average mathematics test scores, then we must also have a 5% level of uncertainty (TYPE I Error) in our decision. This probability of making a mistake (uncertainty) is called the **level of significance** and denoted by the symbol, α . It is the chance we take of rejecting the Null Hypothesis statement when it is true and erroneously accepting the Alternative Hypothesis statement. If we *reject* the Null Hypothesis statement when it is true, and *accept* the Alternative Hypothesis statement, we commit a **TYPE I Error**. If we *retain* the Null Hypothesis statement when it is false, thus *don't accept* the Alternative Hypothesis statement, we commit a **TYPE II Error**, which will be discussed in the next section. In either instance, we do so with a level of confidence and a level of error in our decision.

The region of rejection refers to the selection of a statistical value from the sampling distribution that is at the cut-off point for the beginning of the probability area in which we would reject the null hypothesis in favor of the alternative hypothesis. Statistical values for varying sample sizes and different types of sampling distributions for the 90%, 95%, and 99% confidence levels can be found in the appendix of most statistics books. We refer to these statistical values as the “tabled statistic.” If the sample statistic computed from the sample data falls in the 5% area, corresponding to a 95% confidence level, then we reject the Null Hypothesis and accept the Alternative Hypothesis. We make this decision knowing that 5% of the time it might not be the correct decision; which is the TYPE I Error.

The research process would now involve randomly sampling students and randomly assigning them to either a traditional math program or an innovative math program. After a semester of study, each group would take the same mathematics test. The sample means and standard deviations for each group would be computed. A statistical test would determine if the means were statistically different for our level of confidence and corresponding level of significance. In this instance, an independent t-test would be computed which will be presented in a later chapter (see Chap. 10).

The final step in the research process is a comparison of the “tabled statistic” (based on the level of significance and region of rejection) and the sample statistic computed from the sample data. If the sample statistic is greater than the tabled statistic, a decision is made to reject the Null Hypothesis and accept the Alternative Hypothesis. If the sample statistic does not exceed the tabled statistic, then a decision is made to retain the Null Hypothesis and reject the Alternative Hypothesis. Our interpretation of the findings of the research process is based on this decision to reject or accept the research question, which relates to the two possible outcomes of the study.

TYPE I Error is the probability of rejecting a Null Hypothesis statement, which is true, and erroneously accepting an Alternative Hypothesis. TYPE I Error probability is the same as the level of significance denoted by the symbol, α . As the level of confidence is increased, TYPE I Error is decreased, and the width of the confidence interval increases. As the sample size is increased, the width of the confidence interval is decreased. The sample mean is always in the confidence interval. The population mean is sometimes outside of the confidence interval; the percentage of the intervals that contain the population mean is the confidence level, while the percentage of the intervals that don't contain the population mean is the TYPE I error probability.

TYPE I ERROR R Program

The TYPE I ERROR program shows the relationship between the confidence level and the width of the confidence interval, but in the context of TYPE I Error. In addition, you can use different sample sizes to observe the effect of sample size on the width of the confidence interval, the confidence level, and the TYPE I Error. The program is initially set with a population mean of 50 and a standard deviation of 10. The TYPE I ERROR program computes and graph the 90%, 95%, and 99% confidence intervals which correspond to the 10%, 5%, and 1% levels of uncertainty (levels of significance). Since the population standard deviation is known, z-values rather than t-values are used to compute the intervals. The tabled z-values that correspond to the levels of significance are 1.65, 1.96, and 2.58, respectively, for 90%, 95%, and 99% confidence levels. The program prints the location of the sample mean so that you can compare it with the true population mean of 50. The vertical lines, starting with the two inside lines, indicate the 90%, 95%, and 99% confidence intervals, respectively.

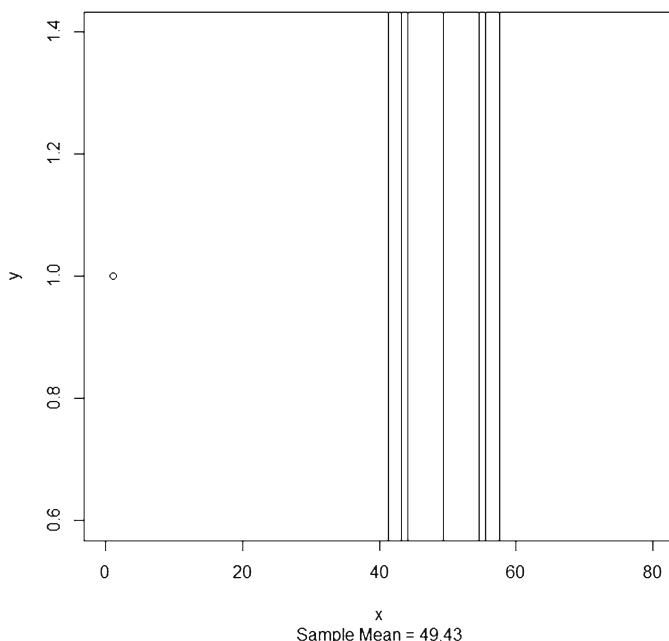
TYPE I ERROR Program Output

Pop. Mean=50 Pop. SD=10
Sample Size=10 N Replications=100

Confidence Level	Percent Error	Confidence Interval	Interval Width
90%	10%	54.6 - 44.2	10.4
95%	4%	55.6 - 43.2	12.4
99%	3%	57.6 - 41.3	16.3

Confidence Intervals (90%, 95%, 99%)

Sample Size = 10



TYPE I Error Exercises

- Run the TYPE I ERROR program for a sample size of 100; keep the other values the same. Record the confidence interval, interval width and TYPE I Error.

Confidence Level	Confidence Interval	Interval Width	TYPE I Error
90%	_____	_____	_____
95%	_____	_____	_____
99%	_____	_____	_____

- Does the Interval Width become larger as the Confidence Level increases from 90% to 99%? YES _____ NO _____
- Does the TYPE I Error become smaller as the Confidence Level increases from 90% to 99%? YES _____ NO _____
- Which confidence level has the greatest probability of *not* containing the population mean? 90% _____ 95% _____ 99% _____
- Were the TYPE I Error percents the same as the Expected Error percents?

TYPE I Error	Expected Error	Confidence Level
_____	10%	90%
_____	5%	95%
_____	1%	99%

- Run the TYPE I ERROR Program for the sample sizes listed below. Record the Confidence Interval Width and TYPE I Error for each sample size and confidence level.

90% Confidence Level

Sample Size	Interval Width	TYPE I Error
10	_____	_____
100	_____	_____
500	_____	_____
1000	_____	_____

95% Confidence Level

Sample Size	Interval Width	TYPE I Error
10	_____	_____
100	_____	_____
500	_____	_____
1000	_____	_____

99% Confidence Level

Sample Size	Interval Width	TYPE I Error
10	_____	_____
100	_____	_____
500	_____	_____
1000	_____	_____

- Does the confidence interval width become smaller as sample size increases for 90%, 95%, and 99% confidence levels. YES _____ NO _____
- In comparing the results for large versus small sample sizes across the three confidence levels, do the TYPE I errors become closer to the level of confidence as the sample size increases? YES _____ NO _____
- Is it possible for the sample mean to be outside of the confidence interval? YES _____ NO _____

TYPE II Error

The research process begins with the formulation of a research question. The research question is then stated in a form that can be tested which is called a statistical hypothesis. The two possible outcomes of the statistical hypothesis are termed a **null hypothesis** and an **alternative hypothesis**. The null hypothesis is stated as “no difference” between population parameters and indicates, “to be nullified.” The alternative hypothesis reflects the outcome expected by the research question. Consequently, the alternative hypothesis can be stated in either a directional or non-directional format. A **directional** hypothesis states that one population parameter is greater than the other population parameter. A **non-directional** hypothesis states that the two population parameters “are different,” but doesn’t specify which is greater than the other population parameter. In the previous chapter we assumed a non-directional hypothesis because we used the probability area in both tails (both sides) of the normal bell-shaped probability curve. It is also important to point out that statistical hypotheses are stated in the form of population parameters.

An example of both the non-directional and directional statistical hypotheses will help to visualize how they are presented using population parameters. The **non-directional** hypothesis in both Null Hypothesis (H_O) and Alternative Hypothesis (H_A) form, stating that Company A and Company B have the same average sales, would be indicated as:

$$\begin{aligned} H_O &: \mu_{\text{CompanyA}} = \mu_{\text{CompanyB}} \\ H_A &: \mu_{\text{CompanyA}} \neq \mu_{\text{CompanyB}} \end{aligned}$$

The Null Hypothesis indicates that the population means for the two companies are equal. The Alternative Hypothesis indicates that the population means for the two companies are different.

The **directional** hypothesis in both Null Hypothesis (H_O) and Alternative Hypothesis (H_A) form, stating that Company A had greater average sales than Company B, would be indicated as:

$$\begin{aligned} H_O &: \mu_{\text{CompanyA}} \leq \mu_{\text{CompanyB}} \\ H_A &: \mu_{\text{CompanyA}} > \mu_{\text{CompanyB}} \end{aligned}$$

The Null Hypothesis indicates that the population mean for Company A is less than or equal to the population mean for Company B. The Alternative Hypothesis indicates that the population mean for Company A is greater than the population mean for Company B, thereby only testing a specific directional difference in the population means of the companies.

The probabilities related to the two possible outcomes, Null Hypothesis and Alternative Hypothesis, are given the name TYPE I error and TYPE II error.

A **TYPE I error** is when the null hypothesis is rejected, but in fact it is true. This means that you gathered evidence (data) and concluded that the evidence was strong enough to accept an alternative hypothesis, but did so erroneously. A **TYPE II error** is when the null hypothesis is accepted, but in fact it is false. This means that you gathered evidence (data) and concluded that the evidence wasn't strong enough to reject the null hypothesis, so you retained the null hypothesis, but did so erroneously. Not rejecting the null hypothesis was due to a lack of sufficient evidence (data). Neither decision, to reject or retain the null hypothesis, is 100% certain; therefore we make our decision with some amount of uncertainty. Whether the TYPE I error or TYPE II error is more important depends on the research situation and the type of decision made. An example will help to clarify the TYPE I and TYPE II errors and possible outcomes.

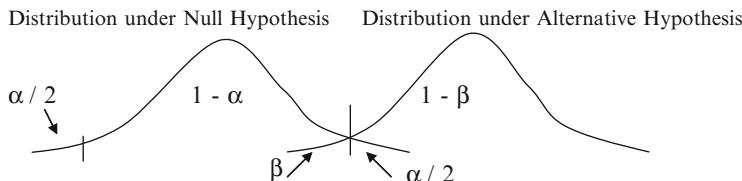
A corporation is going to use data from a sample of people to test a research question of whether or not a new product design will increase the average sales above that of the current product design. Using the current product design, the company sold an average of 85 cars. The company, in planning the study, realizes that two types of error are possible in the study. If the new product design is used, the average sales of the product may not be greater than 85 cars; but, the analysis could incorrectly lead to the conclusion that the average sales were higher, a TYPE I error. If the new product design was used and the average sales did exceed 85 cars, but the statistical analysis failed to indicate a statistically significant increase, then a TYPE II error will occur.

When planning this study, the company can control the probability of a TYPE I error by setting the level of significance (α) to a stringent probability level. For example, setting the level of significance to .05 implies that 95 times out of 100 replications of the study, a correct decision would be made, and 5 times out of 100 an incorrect decision would be made. Setting the level of significance to .01 implies that 99 times out of 100 a correct decision would be made, and only 1 time out of 100 an incorrect decision would be made. These probabilities for a correct decision and an incorrect decision are applicable only when the Null Hypothesis is true. Therefore, we set the TYPE I error to a probability level that instills confidence in our decision to reject the null hypothesis (no difference) and accept an alternative hypothesis (difference exists). A TYPE II error requires a further concern for many other factors in the study. The probability of a TYPE II error depends on the level of significance (β), the direction or non-direction of the research question, the sample size, the population variance, and the difference between the population parameters we want to be able to detect (effect size). These concerns determine how powerful our test will be to detect a difference when it actually exists in the population(s) we study. In planning the study, the company needs to be concerned about both TYPE I and TYPE II errors, as well as the consequences of the decision made based on the outcome of the decision.

In testing hypotheses, four possible outcomes exist with regard to the decision made. They are:

		Actual Population Condition	
		Null Hypothesis is True	Null Hypothesis is False
		Reject Null Hypothesis	Retain Null Hypothesis
Decision	Reject Null Hypothesis	TYPE I ERROR (Probability = α)	CORRECT DECISION (Probability = $1 - \beta$)
	Retain Null Hypothesis	CORRECT DECISION (Probability = $1 - \alpha$)	TYPE II ERROR (Probability = β)

The probability of making a TYPE I error in the Null Hypothesis is denoted by the symbol alpha, α , and referred to as the level of significance, with the probability of a correct decision $1 - \alpha$. The probability of making a TYPE II error in the Alternative Hypothesis is denoted by the symbol Beta, β , with the power of the test given by $1 - \beta$; the probability of making a correct decision. The relationship between TYPE I error and TYPE II error is typically graphed using vertical hash marks to denote the area for the region of rejection. The two graphs below indicate the probability areas for the four decision outcomes.



The probability in the tails of the distributions can change because the null and alternative hypotheses can be stated in directional or non-directional forms. In this example, the hypotheses are in a non-directional form, i.e., Is there a statistically significant mean difference between boys and girls on the mathematics test? If the level of significance is .05, then one-half or .025 would be the probability area in the tails of the distribution. If we chose only the one tail ($p=.025$), then the alternative research question would be stated as directional, e.g., Will the girls on average score higher than the boys in mathematics achievement? The statistical tables in the appendix adjust for the directional versus non-directional nature when selecting a “tabled statistic” value.

The basic concern in selecting a TYPE II error is the value for power, which is $1 - \beta$. **Power** is the probability of *not* committing the error. In conducting research in education, psychology, business, and other disciplines, we are concerned about the power of our statistical test to detect a difference in the population parameters. A graph of the power values against different population parameter values will result in *power curves* for different levels of significance and sample size. The resulting power curves provide an adequate determination of the power of a test for all alternative values of the population parameter.

The concerns in research for having an adequate sample size, the nature of TYPE I and TYPE II errors, power of a test, and selection of alternative values for the population parameters, called effect size, form the basis of hypothesis testing. The hypothesis testing theory, known as the *Neyman-Pearson* hypothesis-testing theory, formulates the relationship between these concerns. In practice, we might select power equal to .80, use a .05 level of significance, and a directional hypothesis to test our statistical hypothesis. The **effect size** or magnitude of difference between the population parameters also affects the power of the test. If a small difference must be detected, then the sample size, level of significance, and power must be increased. We also generally know that the power of a test increases as sample size increases, and the power of a test increases as the level of significance is decreased, i.e., .01 to .05. Basically, all of these values are inter-related such that increasing or decreasing one affects the others. There are software programs available on the Internet that help in selecting sample size or power for different statistical tests, given values for the other criteria, e.g. GPower 3.1.

A TYPE II error implies that a null hypothesis has been accepted, but is false. The probability of a TYPE II error is denoted by the symbol B . The probability of a TYPE II error decreases as α increases (e.g., .01 to .05), N increases, the population variance decreases, and the effect size is larger (true mean is farther from the hypothesized mean in the direction of the alternative hypothesis). When a TYPE II error decreases, power increases. Power is the probability of rejecting a null hypothesis that is false. When the null hypothesis is false, power is equal to $1 - B$. When the null hypothesis is true, power is equal to α .

TYPE II ERROR R Program

The probability of a TYPE II error is affected by four criteria (α , sample size, population variance, and effect size) specified in the **TYPE II ERROR** program. The program inputs these values and determines whether or not to reject the null hypothesis. The program is initially repeated 100 times. The relative frequency of occurrence with which a false null hypothesis is retained is an approximation of the probability of a TYPE II error, B . The power of the test is calculated as 1 minus this relative frequency. Power is equal to $1 - B$ when the null hypothesis is false. The theoretical values for the probability of retaining the null hypothesis and power are also given. If the true mean is equal to the mean in the null hypothesis, then the probability of retaining the null hypothesis is equal to $1 - \alpha$. In this case, power is equal to α , the level of significance. The hypothesis testing in the program is directional to determine whether the mean of a population with a normal distribution is greater than a specified value, therefore, the level of significance is in one tail of the distribution.

In the program, samples are drawn for a given number of replications instead of for a single sample. A **p-value function** determines the probability area based on a directional one-tail test. The number of times the null hypothesis is rejected based on the p-value is summed. This sum is used along with the total number of replications to calculate the estimated probability of retaining the null hypothesis and estimated power.

Next, the critical value for the test is determined by using a new function called **qnorm**, which takes a probability, mean and variance and returns the number of standard deviations above or below the mean. This is multiplied by the standard deviation and added to the mean to obtain the raw score value that equates to a cutoff point for the region of rejection in the tail of the distribution. This cut-off value is used to obtain the true probability of retaining the null hypothesis and true power for the given number of replications. Two separate tables are output, one table for the null mean difference, population mean difference, population variance, sample size, and alpha; and one table for the estimated probability of retaining the null mean difference, estimated power, true probability of retaining the null mean difference, and true power.

TYPE II ERROR Program Output

```
Null Mean=0 Pop. Mean=0 Pop. Variance=1
Sample Size=5 Alpha=0.05 N Replications=100
Hypothesis Direction=1 (0=one-tailed<Null, 1=one-tailed>Null
```

Table 1. Means, Variance, Sample Size, and Alpha

Null Diff.	Pop Diff.	Pop Variance	Sample Size	Alpha
0	0	1	5	0.05

Table 2. Estimated Null Mean Diff. % and Power with True Null % and Power

Retain Null %	Power Estimate	True Null %	True Power
0.980	0.020	0.950	0.050

TYPE II Error Exercises

- Run the TYPE II ERROR program for the values indicated in the following table for 100 replications.

Null Mean	Population Mean	Population Variance	Sample Size	Alpha	Retain Null %	Power Estimate	True Null %	True Power
0	0.0	1.0	10	.05				
0	1.0	1.0	10	.10				
0	1.0	1.0	10	.05				
0	1.0	1.0	10	.025				
0	1.0	1.0	10	.01				
0	1.2	1.0	10	.05				
0	1.5	1.0	10	.05				
0	2.0	1.0	10	.05				
0	1.0	1.5	10	.05				
0	1.0	2.0	10	.05				

- a. In the first run, the Null Mean equals the Population Mean (Null Hypothesis is true), so does the True Power equal Alpha? YES _____ NO _____.
- b. In runs 2–5, alpha decreases from .10 to .01 with other factors held the same, so does B increase (Retain Null %)? YES _____ NO _____.
- c. In runs 2–5, alpha decreases from .10 to .01 with other factors held the same, so does Power decrease (Power Estimate)? YES _____ NO _____.
- d. Does B (Accept Null %) and $1 - B$ (Power Estimate) equal 100%? YES _____ NO _____.
- e. In runs 6–8, the Population Mean is increasingly different from the Null Mean; that is, the population mean is getting farther away from the null mean in the direction of the alternative hypothesis (effect size). So with other factors held constant, does power increase? YES _____ NO _____.
- f. In runs 9–10, the Population Variance is increased, so does Power decrease? YES _____ NO _____.
- g. Try to summarize the relationships between alpha, B (Retain Null %), $1 - B$ (Power Estimate), and effect size in examples **a** to **f** above. _____
-
2. Run TYPE II ERROR program with the following sample sizes: 10, 20, 30, and 40. Keep the other values the same in the program:

```
nullMean <- 0
popMean <- 1
popVar <- 1
sampleSize <- 10
alpha <- .05
tails <- 1
numReplications <- 100
```

- a. As sample size increases with the other factors held constant, does B approach zero? YES _____ NO _____
- b. As the sample size increases with the other factors held constant, does Power increase? YES _____ NO _____
- c. What is the region of rejection for the directional hypothesis, or where the null hypothesis will be rejected, when $n=10$? Note: Tabled Statistic = Mean + (z * Standard error of the statistic). _____
-

3. Run TYPE II ERROR again using the following factors:

```
nullMean <- 0
popMean <- -1
popVar <- 1
sampleSize <- 4
alpha <- .05
tails <- 1
numReplications <- 100
```

- a. What is the true probability of retaining the null hypothesis? _____
- b. What is the true power? _____
- c. Why do you get these values for true probability and power? Hint: You are conducting a directional hypothesis, which only uses the probability area in one tail of the normal curve. _____

True or False Questions

Sampling Distributions

- T F a. The sampling distribution of sample standard deviations is symmetrical for N=2.
- T F b. As sample size increases, the error in the sample standard deviation as an estimator of the population standard deviation decreases.
- T F c. The sample standard deviation is an unbiased, efficient, consistent, and sufficient estimator of the population standard deviation.
- T F d. The sample standard deviation tends to overestimate the population standard deviation.
- T F e. On the average, the sample standard deviation is equal to the population standard deviation.
- T F f. A consistent estimator is one that more closely approximates the population parameter as sample size increases.

Confidence Interval

- T F a. A 95% confidence interval computed from sample data contains the population parameter approximately 95% of the time.
- T F b. A confidence interval indicates the precision of a statistic to estimate the population parameter.
- T F c. As sample size increases, the confidence interval width becomes narrower.
- T F d. A 95% confidence interval implies that 95 times out of 100 the confidence interval *will not* contain the population parameter.
- T F e. It is possible that a 95% confidence interval will not contain the population parameter.
- T F f. As the confidence level increases from 90% to 99%, the width of the confidence interval becomes smaller.

Statistical Hypothesis

- T F a. If the null hypothesis is false, it will be rejected approximately α percent of the time.
- T F b. If the null hypothesis is false, it can still be retained.
- T F c. The sample standard deviation may be used in place of the population standard deviation in the z-test if the sample size is small.
- T F d. The probability of a TYPE II error increases if α is decreased.
- T F e. If the p-value is greater than α , then the null hypothesis is rejected.
- T F f. The region of rejection is denoted by a tabled statistical value for a given sample size and level of significance.
- T F g. A directional hypothesis specifies the probability area for rejection of the null hypothesis in both tails of the normal distribution.
- T F h. Different kinds of statistical hypotheses are tested depending on the level of measurement of data.
- T F i. The level of significance (alpha) determines the probability area for the region of rejection.

TYPE I Error

- T F a. The sum of the Confidence Level and TYPE I error probability equals 100%.
- T F b. If the confidence level increases, the TYPE I error decreases.
- T F c. As sample size increases, the confidence interval width decreases.
- T F d. A TYPE I error is the probability of rejecting the Null Hypothesis when it is true and falsely accepting the Alternative Hypothesis.
- T F e. The sample mean is always inside the confidence interval.
- T F f. The population mean sometimes falls outside of the confidence interval.
- T F g. With all other factors remaining constant, a 90% confidence interval would be narrower than a 99% confidence interval

TYPE II Error

- T F a. As α increases, the power of the test increases.
- T F b. Larger sample sizes will yield more powerful tests.
- T F c. β is the probability of rejecting a true null hypothesis.
- T F d. Small effect sizes are difficult to detect.
- T F e. β decreases for larger sample sizes.
- T F f. As sample size increases, power increases.
- T F g. An acceptable value for power is .80.
- T F h. Power is the probability of rejecting a null hypothesis that is false.
- T F i. A TYPE I error refers to rejecting the null hypothesis when it is true.
- T F j. A TYPE II error refers to retaining the null hypothesis when it is false.
- T F k. Effect size refers to the difference between TYPE I and TYPE II errors.
- T F l. A non-directional hypothesis indicates no difference between population parameters.
- T F m. A directional hypothesis indicates that one population parameter is greater than the other population parameter.

Chapter 8

Chi-Square Test

Previous chapters have presented information on sampling distributions, Central Limit Theorem, confidence intervals, TYPE I error, TYPE II error, and hypothesis testing. This information is useful in understanding how sample statistics are used to test differences between population parameters. The statistical tests presented in this and subsequent chapters depend upon the level of measurement and type of research design.

A popular statistic for testing research questions involving categorical data is the chi-square test statistic. The chi-square statistic was developed by Karl Pearson to test whether two categorical variables were independent of each other. A typical research question involving two categorical variables can be stated as, “Is drinking alcoholic beverages independent of smoking cigarettes?” A researcher would gather data on both variables in a “yes-no” format, then cross tabulate the data. The cross-tabulation of the data for this research question would look like the following:

		Do you drink alcoholic beverages?	
		Yes	No
Do you smoke cigarettes?	Yes		
	No		

Individuals would be asked both questions and their separate responses recorded. The cross-tabulation of the data would permit an indication of the number of people who *did* smoke cigarettes and *did* drink alcoholic beverages, the number of people who *did* smoke cigarettes and *did not* drink alcoholic beverages, the number of people who *did not* smoke cigarettes and *did* drink alcoholic beverages, and the number of people who *did not* smoke cigarettes and *did not* drink alcohol. Consequently, four possible outcomes are represented by the cross-tabulation of the yes/no responses to the two questions.

The **chi-square statistic** is computed by taking the sum of the observed frequency minus the expected frequency squared divided by the expected frequency in each of the four cells. The chi-square formula is expressed as:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

Multiplying the respective row and column sums and dividing by the total number of individuals yields the expected frequencies in each of the four cells. The calculation of the difference between what is observed and what is expected by chance alone forms the basis for the test of independence between two categorical variables. The *expected cell frequencies* are based on the two categorical variables being independent. An example will help to illustrate how to calculate the expected frequencies and the chi-square statistic.

A school district is interested in having parents pass a bond referendum to build a new high school. The superintendent decides to conduct a preliminary poll of the voters to see if they might favor passing the bond. The superintendent is also concerned about whether men and women would vote differently on the bond referendum. Consequently, 200 parents (100 men and 100 women) in the district were randomly selected and telephoned to collect the data. Each parent was asked their gender and whether they favor or oppose a bond to build a new high school. The responses are cross-tabulated below.

Gender	Favor	Oppose	Total
Men	40 (60)	60 (40)	100
Women	80 (60)	20 (40)	100
Totals	120	80	200

The observed values indicate 40 out of 100 (40%) men and 80 out of 100 (80%) women are in favor of the bond, while 60 out of 100 (60%) men and 20 out of 100 (20%) women are opposed to the bond. If the null hypothesis is true (no difference in the percent between men and women in favor of the bond), then we would expect the percentages to be the same for both men and women, i.e., of the 120 observed in favor, one-half or 60 individuals would be expected in each gender cell. The expected cell frequencies are what would be expected if gender and voting were independent. The most convenient way to calculate the expected cell values is to multiply the corresponding row and column sums and divide by the total sample size. For men, the first expected cell value is: $(100 \times 120)/200=60$. The other expected cell value is: $(100 \times 80)/200=40$. For women, the first expected cell value is: $(100 \times 120)/200=60$. The other expected cell value is: $(100 \times 80)/200=40$. The expected cell values are in parentheses in the table. The expected cell values should always add up to the total for each row and/or column, respectively.

The chi-square statistic compares the corresponding observed and expected values in the cells of the table under the assumption that the categorical variables are independent, i.e., the null hypothesis is true. If the row and the column variables are independent, then the proportion observed in each cell should be similar to the

proportion expected in each cell. Is the difference between what we observed and expected in the four cells statistically different or due to random chance (expected values)? A decision about the null hypothesis is made on the basis of the chi-square statistic, which is computed as follows:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right) = \frac{(40 - 60)^2}{60} + \frac{(60 - 40)^2}{40} + \frac{(80 - 60)^2}{60} + \frac{(20 - 40)^2}{40}$$

$$\chi^2 = \sum (6.67 + 10 + 6.67 + 10) = 33.34$$

The computed chi-square value is compared to a tabled chi-square value in the appendix for a given degree of freedom. The degrees of freedom are always determined by the number of rows minus one ($r - 1$) times the number of columns minus one ($c - 1$). This can be expressed as: $df = (r - 1)(c - 1)$. Since there are two rows and two columns, the degree of freedom is: $df = (2 - 1)(2 - 1) = 1$. The tabled chi-square value for $df = 1$ and a .05 level of significance is 3.84. Since the computed chi-square value of 33.34 is greater than the tabled chi-square value of 3.84, we reject the null hypothesis in favor of the alternative hypothesis that men and women differ in the percent favoring a bond for building a new school.

The chi-square value is computed over all the cells and therefore a significant chi-square doesn't specify which cells may have contributed to the significance of the overall chi-square value. Our interpretation of the overall chi-square result is greatly enhanced by realizing that each cell value is itself a chi-square value! Consequently, we can interpret each cell value individually and compare it to the tabled chi-square value of 3.84 with $df = 1$. Since each cell value is greater than 3.84 (6.67, 10, 6.67, and 10), we would conclude that each cross-tabulated cell significantly contributed to the overall chi-square. Also, each expected cell frequency should be greater than five to meet the assumption for computing the chi-square statistic.

Another helpful approach to interpreting chi-square results is to take each individual chi-square value (cell value) as a percent of the overall chi-square value. This provides a variance accounted for interpretation. In our example, $6.67/33.34 = 20\%$, $10/33.34 = 30\%$, $6.67/33.34 = 20\%$, and $10/33.34 = 30\%$. The sum of these cell percents must always equal 100%. Our interpretation would then be based on which cell or cells contributed the most to the overall chi-square.

The chi-square statistic will be small if there are small differences between the observed and the expected values, and it will be large if there are large differences. The chi-square statistic for a two-by-two table is distributed as a theoretical chi-square sampling distribution with 1 degree of freedom. Therefore, the theoretical chi-square distribution can be used to determine the region of rejection for the null hypothesis. The region of rejection includes any chi-square statistic greater than the $1 - \alpha$ percentile of the theoretical chi-square distribution with degree of freedom = $(r - 1)(c - 1)$. Consequently, the chi-square test of independence can be performed on tables of any dimension, i.e., varying numbers of rows and columns for categorical variables.

The chi-square statistic for two-by-two tables only are discussed. You will determine if the row and column values are independent or dependent. If the rows and columns are independent, you will be able to observe the variability in the observed values that occur because of random sampling. A TYPE I error will occur when the rows and columns are independent, i.e., the null hypothesis will be rejected when it is true. If the rows and columns are dependent, you will be able to observe TYPE II errors, i.e., the null hypothesis is retained. As in previous programs, the probability of a TYPE II error will depend on how much the true situation differs from the null hypothesis. If the rows and columns are almost independent, then the probability of a TYPE II error will be high.

A chi-square statistic can be used to test research questions involving cross-tabulated categorical variables. An overall chi-square statistic is computed by summing the individual cell values (chi-squares) in a cross-tabulated table. The degrees of freedom for a cross-tabulated table are row minus one times column minus one, i.e., $df=(r-1)(c-1)$. The chi-square test of independence can be used for any number of rows and columns, as long as the expected cell frequency is greater than five. A chi-square test of independence is used to determine whether or not the rows and columns are independent (null hypothesis). If the null hypothesis is true, it is still possible that the chi-square test could lead to a rejection of the null hypothesis (TYPE I error). If the null hypothesis is false, it is still possible that the chi-square test could lead to retaining the null hypothesis (TYPE II error). The ratio of each cell value to the overall chi-square value provides a variance accounted for interpretation of how much each cell contributed to the overall chi-square value. The chi-square table of expected values in the Appendix permits testing whether the computed chi-square value occurs beyond a chance level of probability.

CROSSTAB R Program

The CROSSTAB program inputs four percents for the true population and the sample size. The program will then select a random sample of size N for this population. The percents will be printed in a table along with the observed and expected values. A chi-square statistic will be computed and printed along with the degrees of freedom and probability value. You will make a decision about whether to retain or reject the null hypothesis based on the probability value being $p < .05$. A second example uses large percent differences between the observed and expected values to better understand the magnitude of the chi-square value. You may want to examine each individual cell chi-square value for significance and hand calculate the percent contribution to the overall chi-square value.

The CROSSTAB program uses matrices to represent cross-tabulated categorical variable tables. The program first defines the true proportions within each cell (the numbers are arranged by the first row followed by the second row) and then the sample size. Next, random data are generated from a discrete population of data from 1 to 4, simulating categorization into a cell of the table. The vector of data is factored

and tabled in order to get counts for each outcome and then placed into a matrix. The table of expected values is built from the sample data by taking the sum of the values for column 1 and dividing it by the sample size to determine the proportion of total outcomes that fall within the first column, then multiplying it by the total number of responses for row 1. This gives the expected value for cell (1,1). The process is repeated for the other three cells and the results are fitted into a matrix and rounded to three digits. Three matrices are then printed: True Population Proportions, Observed Proportions, and Expected Proportions. Finally, the function **chisq.test** is performed on the *testMatrix* object in order to do the actual chi-square test. The chi-square value, degrees of freedom, and p-value are printed by default. The **chisq.test** function uses the Yates correction, which uses real limits around numbers, for example, real limits of 5 is 4.5–5.5.

CROSSTAB Program Output

Example 1

```
Cell Probabilities=0.1 0.15 0.3 0.45
Sample Size=100
Population Proportions
      X      Y
A  0.10    0.15
B  0.30    0.45

Observed Proportions
      X      Y
A  0.10    0.17
B  0.25    0.48

Expected Proportions
      X      Y
A  0.094   0.176
B  0.255   0.475

Pearson chi-square test with Yates' continuity correction
Chi-square = 6e-04 df = 1 p-value = 0.98116
```

Example 2

```
Cell Probabilities=0.7 0.1 0.1 0.1
Sample Size=100
Population Proportions
      X      Y
A  0.70    0.10
B  0.10    0.10
```

Observed Proportions

	X	Y
A	0.7	0.1
B	0.1	0.1

Expected Proportions

	X	Y
A	0.64	0.16
B	0.16	0.04

Pearson chi-square test with Yates' continuity correction
Chi-square=11.8164 df = 1 p-value = 0.00059

Chi-Square Exercises

1. Run CROSSTAB using the Population proportions below with a sample size of 100 using .05 level of significance.

Population Proportions

.50	.20
.10	.20

Observed Proportions

Expected Proportions

- a. Are the expected values in the cells computed correctly? YES____ NO____
- b. Is the chi-square statistic computed correctly? YES____ NO____
- c. What is the chi-square value? _____
- d. What is the statistical decision? Retain Null ____ Reject Null____
- e. Does the decision agree with the true percent? YES____ NO____
- f. What percent of the time would you expect the null hypothesis to be rejected by mistake? _____
- g. What is the name of this type of error? _____

2. Run CROSSTAB using the Population proportions below with a sample size of 100 using a .01 level of significance.

Population Proportions

.16	.32
.28	.24

Observed Proportions

Expected Proportions

- a. Are the expected values in the cells computed correctly? YES ____ NO ____
- b. Is the chi-square statistic computed correctly? YES ____ NO ____
- c. What is the chi-square value? _____
- d. What is the statistical decision? Retain Null ____ Reject Null ____
- e. Does the decision agree with the true percent? YES ____ NO ____
- f. What percent of the time would you expect the null hypothesis to be rejected by mistake? _____
- g. What is the name of this type of error? _____

True or False Questions

Chi-Square

- | | | |
|---|---|---|
| T | F | a. A chi-square statistic is used with ordinal data. |
| T | F | b. The null hypothesis in the chi-square test corresponds to no difference in the row and column categories. |
| T | F | c. The chi-square statistic will be large if there is a large difference between the observed and the expected values in the cells. |
| T | F | d. If the true population has independent rows and column, then a TYPE I error may occur. |
| T | F | e. The chi-square statistic can be negative. |
| T | F | f. The overall chi-square value indicates which cells contributed to a statistically significant finding. |
| T | F | g. The tabled chi-square value is 3.84 with one degree of freedom at the .05 level of significance. |

Chapter 9

z-Test

Many research questions involve testing differences between two population proportions (percentages). For example, Is there a significant difference between the proportion of girls and boys who smoke cigarettes in high school?, Is there a significant difference in the proportion of foreign and domestic automobile sales?, or Is there a significant difference in the proportion of girls and boys passing the Iowa Test of Basic Skills? These research questions involve testing the differences in population proportions between two independent groups. Other types of research questions can involve differences in population proportions between related or dependent groups. For example, Is there a significant difference in the proportion of adults smoking cigarettes before and after attending a stop smoking clinic?, Is there a significant difference in the proportion of foreign automobiles sold in the U.S. between years 1999 and 2000?, or Is there a significant difference in the proportion of girls passing the Iowa Test of Basic Skills between the years 1980 and 1990? Research questions involving differences in independent and dependent population proportions can be tested using a **z-test** statistic. Unfortunately, these types of tests are not available in most statistical packages, and therefore you will need to use a calculator or spreadsheet program to conduct the test.

Independent Samples

A practical example using the hypothesis testing approach will help to illustrate the z-test for differences in proportions between two independent groups. The research question for our example will be: Do a greater proportion of high school students in the population smoke cigarettes in urban rather than rural cities? This is a directional research question. A step-by-step outline will be followed to test this research question.

Step 1. State the directional research question in a statistical hypothesis format.

$$H_0: P_1 \leq P_2 \text{ (or } P_1 - P_2 \leq 0\text{)}$$

$$H_A: P_1 > P_2 \text{ (or } P_1 - P_2 > 0\text{)}$$

Notice that the “H” stands for hypothesis with the subscripts “0” for the null statistical hypothesis and “A” for the alternative statistical hypothesis. The alternative statistical hypothesis is stated to reflect the research question. In this example, the alternative statistical hypothesis indicates the directional nature of the research question. Also, P_1 is the population proportion of high school students who smoke in an urban city and P_2 is the population proportion of high school students who smoke cigarettes in a rural city.

Step 2. Determine the criteria for rejecting the null hypothesis and accepting the alternative hypothesis.

Given $\alpha=0.01$, we select the corresponding z value from Table A1 (see Appendix) which is the closest to 1% of the area under the normal curve. If our computed z-test statistic is greater than this tabled z-value, we would reject the null hypothesis in favor of the alternative hypothesis. This establishes our region of rejection, R, or the probability area under the normal curve where differences in sample proportions are unlikely to occur by random chance.

R: $z > 2.33$

Notice that in Table A1 (see Appendix), the first column indicates $z=2.3$ with the other column indicating the 3 in the hundredths decimal place. Also, notice that the percentage = 0.4901 is the closest to 49%, which leaves 1% area under the normal curve. In Table A1, only one-half of the normal curve is represented, so 50% is automatically added to the 49% to get 99%, which reflects the one-tail probability for the directional alternative statistical hypothesis.

Step 3. Collect the sample data and compute the z-test statistic.

A random sample of 20% of all high school students from both an urban and a rural city was selected. In the urban city, 20,000 high school students were sampled with 25% smoking cigarettes ($n_1=5,000$). In the rural city, 1,000 high school students were sampled with 15% smoking cigarettes ($n_2=150$). The proportions were:

$$\begin{aligned} P_1 &= 0.25 \text{ (25\% of the boys in the sample of high school students smoke cigarettes)} \\ P_2 &= 0.15 \text{ (15\% of the girls in the sample of high school students smoke cigarettes)} \end{aligned}$$

The standard deviation of the sampling distribution of the differences in independent sample proportions is called the standard error of the difference between independent sample proportions. This value is needed to compute the z-test statistic. The formula is:

$$S_{P_1-P_2} = \sqrt{\frac{pq}{N}}$$

where:

$$p = (n_1 + n_2 / N) = (5,000 + 150 / 21,000) = 0.245$$

$$q = 1 - p = 1 - 0.245 = 0.755$$

$$n_1 = \text{number in first sample} = 5,000$$

$$n_2 = \text{number in second sample} = 150$$

$$N = \text{total sample size taken} = (20,000 + 1,000) = 21,000$$

$$S_{P_1-P_2} = \sqrt{\frac{0.245(0.755)}{21000}} = 0.003$$

The z-test can now be computed as:

$$z = \frac{P_1 - P_2}{S_{P_1-P_2}} = \frac{0.25 - 0.15}{0.003} = \frac{0.10}{0.003} = 33.33$$

We have learned from previous chapters that to test a statistical hypothesis, we need to know the sampling distribution of the statistic. The sampling distribution of the difference between two independent proportions is normally distributed when sample sizes are greater than five. Thus, we can use the normal distribution z statistic to test our statistical hypothesis.

Step 4. Compute the confidence interval around the z-test statistic.

A confidence interval is computed by using the percent difference between the two independent groups ($P_1 - P_2 = 0.10$), the tabled z-value corresponding to a given alpha level for a two-tailed region of region ($z = 2.58$), and the standard deviation of the sampling distribution or standard error of the test statistic ($S_{P_1-P_2} = 0.003$).

$$CI_{99} = 0.10 \pm (2.58)(0.003)$$

$$CI_{99} = 0.10 \pm (0.008)$$

$$CI_{99} = (0.092, 0.108)$$

Notice that the tabled z-value selected for determining the confidence interval around the computed z-test statistic is not the same because the confidence interval is based on a two-tailed interpretation. The alternative statistical hypothesis was one-tail because of the directional nature of the research question.

Step 5. Interpret the z-test statistic results.

Our interpretation is based upon a test of the null hypothesis and a 99% confidence interval around the computed z-test statistic. Since the computed $z = 33.33$ is greater than the tabled z -value, $z = 2.33$ at the 0.01 level of significance, we reject the null statistical hypothesis in favor of the alternative statistical hypothesis. The probability that the observed difference in the sample proportions of 10% would have occurred by chance is less than 0.01. We can therefore conclude that the urban city had a greater percentage of high school students smoking cigarettes than the rural city. The TYPE I error was set at 0.01, so we can be fairly confident in our interpretation.

The confidence interval was computed as 0.092 to 0.108, indicating that we can be 99% confident that this interval contains the difference between the population proportions from which the samples were taken. Moreover, the narrowness of the confidence interval gives us some idea of how much the difference in independent sample proportions might vary from random sample to random sample. Consequently, we can feel fairly confident that a 9% (0.092) to 11% (0.108) difference would exist between urban and rural city high school students smoking cigarettes upon repeated sampling of the population.

Dependent Samples

The null hypothesis that there is no difference between two population proportions can also be tested for dependent samples using the z-test statistic. The research design would involve obtaining percentages from the same sample or group twice. The research design would therefore have paired observations. Some examples of when this occurs would be:

1. Test differences in proportions of agreement in a group before and after a discussion of the death penalty.
2. Test differences in percent passing for students who take two similar tests.
3. Test differences in the proportion of employees who support a retirement plan and the proportion that support a company daycare.

The research design involves studying the impact of diversity training on the proportion of company employees who would favor hiring foreign workers. Before and after diversity training, employees were asked whether or not they were in favor of the company hiring foreign workers. The research question could be stated as: Are the proportions of company employees who favor hiring foreign workers the same before and after diversity training? This is a non-directional research question. A step-by-step approach to hypothesis testing will be used.

Step 1. State the non-directional research question in a statistical hypothesis format.

$$H_0: P_1 = P_2 \text{ (or } P_1 - P_2 = 0\text{)}$$

$$H_A: P_1 \neq P_2 \text{ (or } P_1 - P_2 \neq 0\text{)}$$

Notice in this example that we are interested in testing the null hypothesis of no difference between the population proportions against the non-directional alternative hypothesis, which indicates that the proportions are different. The alternative statistical hypothesis is stated to reflect the non-directional research question. Also, P_1 is the proportion of employees in favor before diversity training and P_2 is the proportion of employees in favor after diversity training.

Step 2. Determine the criteria for rejecting the null hypothesis and accepting the alternative hypothesis.

Given $\alpha=0.05$, we select the corresponding z-value from Table A1 which is the closest to 5% of the area under the normal curve (2.5% in each tail of the normal curve). If our computed z-test statistic is greater than this tabled z-value, we would reject the null hypothesis and accept the alternative hypothesis. This establishes our region of rejection, R, or the probability areas under the normal curve where differences in sample proportions are unlikely to occur by random chance.

$$R: z \pm 1.96$$

Notice that in Table A1 (see Appendix), the first column indicates $z=1.9$ with the other column indicating the 6 in the hundredths decimal place. Also, notice that the percentage = 0.4750 indicates 0.025 probability area under the normal curve in only one tail. In Table A1, only one-half of the normal curve is represented, but 0.025 in both tails of the normal curve would equal 5%. If we add $0.4750 + 0.4750$, it would equal 0.95 or 95%, which indicates the remaining percent under the normal curve. The region of rejection indicates two z-values, $+1.96$ and -1.96 , for rejecting the null hypothesis, which reflects testing a non-directional research question.

Step 3. Collect the sample data and compute the z-test statistic.

A random sample of 100 employees from a high-tech company were interviewed before and after a diversity training session and asked whether or not they favored the company hiring foreign workers. Their sample responses were as follows:

		After Diversity Training		
		No	Yes	
Before Diversity Training	Yes	10 (0.10)	20 (0.20)	30 (0.30)
	No	50 (0.50)	20 (0.20)	70 (0.70)
		60 (0.60)	40 (0.40)	100 Total

The sample data indicated the following proportions:

P_1 =proportion in favor before diversity training=0.30 or 30%

P_2 =proportion in favor after diversity training=0.40 or 40%

Notice the order of the data entry in the cells of this table. This was done so that certain cells indicate disagreement or dissimilar responses before and after diversity training.

The standard deviation of the sampling distribution of the differences in dependent sample proportions is called the standard error of the difference between dependent sample proportions. This value is needed to compute the z-test statistic. The formula is:

$$S_{P_1 - P_2} = \sqrt{\frac{p_{11} + p_{22}}{N}}$$

where:

p_{11} =percent change from before to after training (yes→no)=0.10

p_{12} =percent change from before to after training (no→yes)=0.20

N=total sample size=100

$$S_{P_1 - P_2} = \sqrt{\frac{0.10 + 0.20}{100}} = 0.055$$

The z-test can now be computed as:

$$z = \frac{P_1 - P_2}{S_{P_1 - P_2}} = \frac{0.30 - 0.40}{0.055} = \frac{-0.10}{0.055} = -1.82$$

We have learned from previous chapters that to test a statistical hypothesis, we need to know the sampling distribution of the statistic. The sampling distribution of the difference between two dependent proportions is normally distributed when the sum of the sample sizes in the diagonal cells are greater than ten. Thus, we can use the normal distribution z-statistic to test our statistical hypothesis.

Step 4. Compute the confidence interval around the z-test statistic.

A confidence interval is computed by using the percent difference between the two independent groups ($P_1 - P_2 = 0.10$), the tabled z-value corresponding to a given alpha level for a two-tailed region of region ($z = \pm 1.96$), and the standard deviation of the sampling distribution or standard error of the test statistic ($S_{P_1 - P_2} = 0.055$).

$$\text{CI}_{99} = 0.10 \pm (1.96)(0.055)$$

$$\text{CI}_{99} = 0.10 \pm (0.108)$$

$$\text{CI}_{99} = (-0.008, 0.208)$$

Notice that the tabled z-value selected for determining the confidence interval around the computed z-test statistic is the same because the confidence interval is also based on a two-tailed interpretation. Also, the null hypothesized parameter of zero (no difference in proportions) is contained in the confidence interval, which is consistent with not rejecting the null hypothesis.

Step 5. Interpret the z-test statistic results.

Our interpretation is once again based upon a test of the null hypothesis, but this time using a 95% confidence interval around the computed z-test statistic because of the non-directional nature of the research question. Since the computed $z = -1.82$ is less than the tabled z-value, $z = -1.96$ at the 0.05 level of significance, we retain the null statistical hypothesis. The probability that the observed difference in the sample proportions of 10% would have occurred by chance is greater than 0.05. We therefore cannot conclude that the percent of company employees in favor of hiring foreign workers was different before and after diversity training. The TYPE I error was set at 0.05, so we can be fairly confident in our interpretation.

The confidence interval was computed from -0.008 to 0.208 , indicating that we can be 95% confident that this interval contains the null hypothesis parameter of zero difference between the population proportions from which the sample was taken. Moreover, the spread in the confidence interval gives us some idea of how much the difference in the dependent sample proportions might vary from random sample to random sample. Consequently, we should be sensitive to a research design factor that may have impacted the statistical test, which is the duration and intensity of the diversity training. Obviously a 1-h training session involving watching a short slide presentation might have less of an impact on employees than a 6-week training session involving role modeling with foreign workers on the job.

A z-test statistic can be used to answer research questions concerning differences in proportions between independent samples or groups, as well as, differences in proportions between dependent samples or groups. A statistically significant difference in proportions indicates that, beyond a random chance level, two groups differ in their proportions. A confidence interval around the z-test statistic indicates the amount of difference in the group proportions one can expect upon repeated sampling. The confidence interval captures the null hypothesis parameter of zero when proportions are not statistically significantly different.

ZTEST R Programs

The ZTEST-IND program inputs a critical z-value based on the alpha level and directional nature of the test. Next, *size1* and *size2* are set to the respective sample sizes and *num1* and *num2* are set to the number of positive cases in each sample. After this, the proportion of positive cases for each sample are computed as well as the overall *p* and *q*, which represent the overall proportion of positive cases and negative cases, respectively. Next, the standard error of the difference(s) is computed as well as the z-statistic. The individual proportions of positive cases, the critical z-value, and the standard error of the difference are then used to calculate the confidence interval around the proportion differences and all relevant values are printed.

The ZTEST-DEP program inputs a critical z-value, but then assigns numbers to the variables *num11*, *num12*, *num21*, and *num22*, where *num11* and *num22* represent changes from one outcome to the other and *num12* and *num21* represent cases where the outcome was the same. The total number of outcomes are calculated by adding the cells and then the proportion of changes of both types (*num11* and *num22*) are calculated. The proportion of cases with a positive outcome from each group or occasion are calculated next. The standard error of the difference is determined using the proportion of changed outcomes and the total number of outcomes. The z-statistic is determined from the difference in positive outcomes between the two groups or occasions divided by the standard error of the difference. Finally, the upper and lower bounds of a confidence interval around the difference in dependent proportions is calculated and all relevant values are output.

ZTEST-IND Program Output

```
Sample1=20000 Sample2=1000
N1=5000          N2=150
z critical=1.96
Difference in proportions=0.1
Standard Error of Diff=0.003
z Statistic=33.3333
Confidence Intervals=( 0.0941 , 0.1059 )
```

ZTEST-DEP Program Output

```
z Critical value=1.96
First %=0.1 Second %=0.2
Difference in proportions=-0.1
z Statistic=-1.8248
Standard Error of Diff=0.0548
Confidence Interval=(-0.0074 , 0.2074)
```

z Exercises

1. Run ZTEST-IND for the independent sample proportions in the following example.

The research question is: Do Democrats and Republicans differ in their percent agreement on handgun control? Test the research question at the 0.01 level of significance. The following sample data was collected:

Democrats: 50,000 with 24,000 (0.48) or 48% in favor of handgun control
 Republicans: 50,000 with 12,000 (0.24) or 24% in favor of handgun control

- What is the percent difference between the two independent groups? _____
- What is the standard error of the difference between the independent percents? _____
- What is the z-test statistic value? _____
- What are the confidence interval values? (_____ , _____)
- What decision is made based on the z-test and confidence interval?
 Retain Null _____ Reject Null _____
- What percent of the time would you expect the null hypothesis to be rejected by mistake?
- What is the name of this type of error? _____

2. Run ZTEST-DEP for the dependent sample proportions in the following example.

The research question, Is the proportion of students passing the first exam the same as the proportion of students passing the second exam?, will be tested at the 0.05 level of significance. The following sample data was collected (fill in the missing information):

		Pass Second Exam		
Pass First Exam		No	Yes	
	Yes	20 ()	50 ()	_____ ()
	No	15 ()	15 ()	_____ ()
		_____ ()	_____ ()	100

- What is the percent passing difference between the exams? _____
- What is the standard error of the difference between the dependent percents? _____
- What is the z-test statistic value? _____
- What are the confidence interval values? (_____ , _____)
- What decision is made based on the z-test and confidence interval?
 Retain Null _____ Reject Null _____
- What percent of the time would you expect the null hypothesis to be rejected by mistake? _____
- What is the name of this type of error? _____

True or False Questions

z-Test

- | | | |
|---|---|--|
| T | F | a. A z-test statistic is used <i>only</i> with independent sample percents. |
| T | F | b. The null hypothesis in the z-test corresponds to no difference in the proportions of either independent or dependent proportions. |
| T | F | c. The z-test statistic will be large if there are large differences between the sample proportions relative to the standard error. |
| T | F | d. A Type I error may occur when using the <i>z</i> test statistic. |
| T | F | e. The z-test statistic can be negative or positive in value. |
| T | F | f. The z-test statistic can <i>only</i> be used with directional hypothesis testing. |
| T | F | g. The tabled z-value used in forming the confidence interval around the z-test statistic is always based upon two tails under the normal curve. |

Chapter 10

t-Test

Sampling distributions are frequency distributions of a statistic, which are generated to make an inference between the sample statistic and its corresponding population parameter. The average (mean) statistical value in the sampling distribution is the expected value of the population parameter. The variance of the sampling distribution is used to calculate the standard error of the statistic. The standard error of the statistic is a function of sample size and is used to calculate confidence intervals around the sample statistic, which provides the basis for testing statistical hypotheses and determining the interval in which the population parameter falls.

The distribution theory that led to a solution of the problem of estimating the population mean, μ , when the population variance was unknown, σ^2 , was due to William S. Gossett, a chemist, who in 1908 wrote under the pseudonym of “Student.” William S. Gossett, who worked for a brewery, in Dublin, Ireland determined that when the sample size was large, the sampling distribution of the z-statistic was normal; however, when the sample size was small, the sampling distribution was leptokurtic or peaked. He referred to this slightly non-normal distribution as the t-distribution. W. S. Gossett further discovered that as he increased the sample sizes, the sampling distribution became normal, and therefore the t-values equaled the z-values. W. S. Gossett signed his pivotal work, “Student,” and today small sample tests of mean differences are referred to as the “student t-test,” or simply the t-test. Sir Ronald Fisher, using the early work of “Student,” extended his ideas into modern day analysis of variance techniques, which is discussed in the next chapter.

One Sample t-Test

The sampling distribution of the mean can be used to determine the expected value of the population mean and variance when they are unknown. The Central Limit Theorem supports this assumption because as sample size increases, the sampling distribution of the mean becomes a normal distribution with the mean = μ , and

variance = σ^2/N . Knowledge of this permits one to test whether the sample mean is statistically different from a population mean. The test is called a **one-sample t-test**, which is computed by:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{N}}} = \frac{\bar{X} - \mu}{S / \sqrt{N}}$$

Suppose a class of 25 students was given instruction using a new, enriched mathematics curriculum and then took a nationally standardized math achievement test. If the new, enriched mathematics curriculum leads to better learning in the population, then this sample of 25 students should perform better on the math achievement test than students in the larger population who learned mathematics under the traditional mathematics curriculum. Assume the 25 students had a mean math achievement test score of 110 with a standard deviation of 15. The mean math achievement test score for students in the population taking the traditional math curriculum was 100. What is the probability that a random sample of this size would have a sample mean of 110 or higher? This research question can be tested using a one-sample t-test. The one-sample t-test is computed as:

$$t = \frac{110 - 100}{\sqrt{\frac{15^2}{25}}} = \frac{10}{\sqrt{\frac{225}{25}}} = \frac{10}{3} = 3.33$$

If we were to examine the probability area in the tails of the normal distribution for a z-value of 3.33, we would find a probability area equal to 0.0004 ($0.4996 + 0.0004 = 0.5000$). However, since the probability area in the tails of the normal distribution are not the same as the t-distribution, given small sample sizes, we need to examine the probability area in the tails of the t-distribution *not* the normal distribution. We therefore use the t-table in the appendix (Table A1), where we find the degrees of freedom, df , ($df=n-1=24$) and a probability area depending upon a directional (one-tail) or non-directional (two-tailed) test. For $df=24$ and $p=0.0005$ (one-tail test), we find a tabled t-value equal to 3.745. Since the computed $t=3.33$ is less than the tabled $t=3.745$, we retain the null hypothesis. We would conclude that there is no difference in the mean math achievement test scores at the $p=0.0005$ level of statistical significance. Because the t-distribution is leptokurtic or peaked for small samples, the probabilities in the tails of the t-distribution are greater than the probabilities in the normal distribution. When W. S. Gossett discovered this, he made an adjustment for small sample sizes by entering larger tabled t-values from the sampling distribution of the t-values. By adjusting the t-values, he compensated for the greater probability area under the sampling distributions. If you examine the last row in the t-table in the appendix, you will find t-values that are equal to the z-values under the normal distribution because when sample sizes are large $t=z$. In practice, we use the t-test rather than the z-test for testing means because it applies to both small and large samples.

Independent t-Test

The sampling distribution of the difference between the means of two independent samples provides the basis for the testing of a mean difference hypothesis between two groups. Taking a random sample from one population, a second random sample from another population, and then computing the difference in the means, provides the basis for the sampling distribution. The process is repeated several times and the mean differences are graphed in a frequency distribution. Under the assumption of no difference between the means (null hypothesis), the mean of this sampling distribution is zero.

A typical research situation in which one would use the independent t-test might involve one group of employees receiving sales training and the second group of employees *not* receiving any sales training. The number of sales for each group is recorded and averaged. The null hypothesis would be stated as, “The average sales for the two groups are equal.” An alternative hypothesis would be stated as, “The group receiving the sales training will on average have higher sales than the group that did not receive any sales training.” If the sample data for the two groups were recorded as (1) Sales Training: mean = 50, standard deviation = 10, n = 100, and (2) No Sales Training: mean = 40, standard deviation = 10, n = 100; the independent t-test would be computed as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

The numerator of the independent t-test formula computes the difference between the means of the two independent groups. The denominator of the independent t-test formula computes the standard error of the statistic (standard deviation of the sampling distribution), which is computed as:

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

This equation uses information from both samples to estimate a common population variance, which is referred to as “pooling the variances” or “averaging” the sample variances. The square root of this “pooled variance” estimate is the standard error of the statistic (or sampling distribution of the statistic).

The independent t-test for the mean difference between the two groups would be computed as follows:

$$t = \frac{50 - 40}{1.41} = 7.09$$

Referring to the table of t-values in the appendix, we find that the computed t-value = 7.09 exceeds the tabled t-value of 3.291, for df = 198 (df = ∞) and p = 0.0005 (one-tailed test). This is considered a statistically significant finding because a mean

difference this large is beyond a chance occurrence. We would conclude that the group who received sales training on average had higher sales than the group that received no sales training. This result is an important finding if the company is spending a substantial amount of money on its employees for sales training and wished to know if the sales training was effective in increasing sales.

Dependent t-Test

Drawing one random sample, taking two measures on each person, subtracting these two related measures, and graphing the difference generates the sampling distribution of the difference between the means of two related samples. The expected average difference between the related measures is zero under the null hypothesis. The **dependent t-test** is sometimes referred to as the paired t-test or correlated t-test because it uses two sets of scores on the same individuals.

A typical research situation that uses the dependent t-test involves a repeated measure design with one group. For example, a psychologist is studying the effects of a certain motion picture upon attitudes toward violence. The psychologist hypothesizes that viewing the motion picture will cause the students' attitudes to be more violent. A random sample of ten students is given an attitude toward violence inventory before viewing the motion picture. Next, the ten students view the motion picture, which contains graphic violence portrayed as acceptable behavior. The ten students are then given the attitude toward violence inventory after viewing the motion picture. The average attitude toward violence score for the students before viewing the motion picture was 67.5, but after viewing the motion picture was 73. There are ten pairs of scores, so ten score differences are squared and summed to calculate the sum of square differences. The standard error of the dependent t-test is the square root of the sum of squared differences divided by N (N-1).

The dependent t-test to investigate whether the students' attitudes toward violence changed after viewing the motion picture would be computed as:

$$t = \frac{\bar{D}}{S_D}$$

The numerator in the formula is the average difference between the post- and pre-mean scores on the attitude toward violence inventory, which is $73 - 67.5 = 5.5$. The denominator is calculated as:

$$S_D = \sqrt{\frac{\sum(D - \bar{D})^2}{N(N-1)}}$$

The repeated measure sample data are given below

Student	Pre	Post	D	D^2
1	70	75	+5	25
2	60	70	+10	100
3	85	80	-5	25
4	50	65	+15	225
5	65	75	+10	100
6	80	70	-10	100
7	90	95	+5	25
8	70	80	+10	100
9	40	55	+15	225
10	65	65	0	0
Calculations:	$\bar{D}_1 = 67.5$	$\bar{D}_2 = 73.0$	$\Sigma D = 55$	$\Sigma D^2 = 925$

There are ten pairs of student scores

$$\text{NOTE: } \Sigma(D - \bar{D})^2 = \Sigma D^2 - (\Sigma D)^2 / N = 925 - (55)^2 / 10 = 622.5$$

$$S_D = 2.63 \text{ and } \bar{D} = 5.5$$

The dependent t-test is calculated as: $t = 5.5 / 2.63 = 2.09$. This value is compared to a tabled t-value for $df=9$ and $p=0.01$ (probability value), which is $t=2.82$. Since the computed $t=2.09$ does not exceed the tabled $t=2.82$, we retain the null hypothesis. We conclude that watching the motion picture did not change students' attitudes toward violence.

Notice in calculating the dependent t-test that the sum of the differences between the pre- and post-test scores took into account whether the difference was positive or negative. Also notice that a computation formula makes the calculation of the sum of squares for the difference scores much easier to calculate. The importance of the standard error of the sampling distribution in determining whether the sample statistic is significant should be apparent.

The sampling distribution of a statistic is important in making inferences about the corresponding population parameter. The standard error of the sampling distribution of a statistic is important in testing whether the sample statistic is statistically different from the population parameter. The t-distribution is not normally distributed for small sample sizes. The t-values in the t-distribution become equal to z-values in the normal distribution when sample sizes are infinitely large. Different sampling distributions are created for different statistical tests. The one-sample t-test is useful for testing whether a sample mean from a population is statistically different from the population mean. The two sample independent t-test is useful for testing whether the means of two independent groups are identical. The single sample dependent t-test is useful for testing whether the means before and after some intervention in a single group are statistically different. A table of t-values has been created for various sample sizes (degrees of freedom) and probability areas under the t-distribution.

STUDENT R Program

The STUDENT program shows the versatility of the t-test function in R by utilizing it for a one sample t-test, a two sample independent t-test, and a dependent t-test. This is accomplished by changing parameters in the **t-test** function. The greatest complexity occurs in determining the values for the user-defined variables based on the type of test. How the different variables used for each test is described in the program. Once these variables are defined, two samples are drawn (only the second sample is actually used for the one sample t-test) from normal distributions with the prescribed means and standard deviations. Next, the type of test desired determines which block of code is used to perform the test and produce the output. For the one sample t-test, only *sampleTwo* is passed to the function along with a **mu** equal to *meanOne*. The results of the t-test are saved to the *tResults* object for later use and output appropriate to the one sample test is appended to the *outputText* vector. For the two sample independent t-test, both samples are passed along with the **paired** keyword set to false. Again, appropriate output is created for this test. Finally, the dependent t-test looks exactly like the independent t-test except **paired** is set to true. All tests have the keyword **alternative** set to **two.sided** because the alternate hypothesis in all cases is for a difference in either direction, instead of a directional hypothesis. After the specific test functions are performed, the output from the t-test that is common to all test types is extracted from the *tResults* object by use of **\$statistics** and **\$p.value** components. Once these values are appended to the output, the output vector is formatted into a matrix and the matrix is printed.

STUDENT Program Output

One Sample t-test

One sample t-test	Sample size= 20
Sample Mean= 53.52	Sample Std Dev= 10.13
Population Mean= 50	Population Std Dev= 10
t-test= 1.55	p-value= 0.137

Independent t-test

Two sample independent t-test	Sample size= 20
Sample One Mean= 49.42	Sample One Std Dev= 10.33
Sample Two Mean= 47.9	Sample Two Std Dev= 13.44
t-test= 0.4	p-value= 0.691

Dependent t-test

Dependent t-test	Number of paired scores= 20
Pre Mean= 48.83	Post Mean= 50.16
Sum of Differences= 26.68	Sum Squared Diff = 2557.61
t-test= 0.52	p-value= 0.611

t Exercises

1. Run the STUDENT program for a one-sample t-test. Record the values in the table.

Use the following values:

```
sampleSize <- 30
meanOne <- 50   (In one-sample t-test equals population mean.)
stdDevOne <- 20 (In one-sample t-test equals population s.d.)
meanTwo <- 60
stdDevTwo <- 10
testType <- 1
```

One-sample t-test	Sample size =
Sample Mean =	Sample SD =
Population Mean =	Population SD =
t-test =	p-value =

- a. Is the sample mean statistically different from the population mean?
 YES _____ NO _____
- b. Was the one-sample t-test calculated correctly? YES _____ NO _____
- c. Would you obtain the same one-sample t-test using the sample standard deviation as the population standard deviation? YES _____ NO _____
- d. Interpret the results of the one-sample t-test. What would you conclude?
-
-

2. Run the STUDENT program for the two sample independent t-test. Record the values in the table. Use the following settings:

```
sampleSize <- 30
meanOne <- 50
stdDevOne <- 20
meanTwo <- 60
stdDevTwo <- 10
testType <- 2
```

Two sample independent t-test	Sample size =
Sample One Mean =	Sample One SD =
Sample Two Mean =	Sample Two SD =
t-test =	p-value =

- a. Are the two sample means statistically different? YES _____ NO _____
- b. Was the two sample independent t-test calculated correctly? YES _____ NO _____
- c. Would you obtain different results if you ran the STUDENT program a second time?
 YES _____ NO _____
 Why? _____
-

- d. Interpret the results of the two sample independent t-test. What would you conclude?
-
-

3. Run the STUDENT program for the dependent t-test. Record the values in the table below. Use the following settings:

```
sampleSize <- 30
meanOne <- 50
stdDevOne <- 20
meanTwo <- 60
stdDevTwo <- 10
testType <- 3
```

Dependent t-test	Number of paired scores =
Pre Mean =	Post Mean =
Sum of differences =	Sum Diff Squared =
t-test =	p-value =

- a. Was the group statistically different from pre to post? YES ____ NO ____
- b. Was the dependent t-test calculated correctly? YES ____ NO ____
- c. Would you obtain different results if you ran the STUDENT program a second time?
YES ____ NO ____
Why? _____
- d. Interpret the results of the dependent t-test. What would you conclude?
-
-

True or False Questions

t-Test

- T F a. The sampling distribution of a statistic is useful for making an inference about the population parameter.
- T F b. The standard error of a statistic provides the basis for testing hypotheses.
- T F c. The t-distribution is normally distributed.
- T F d. The one-sample t-test may or may not be significant depending upon which probability value is selected from the t-table.

- T F e. The t-values and z-values become equal as sample size increases.
- T F f. A z-test is preferred over the t-test for hypothesis testing.
- T F g. The independent t-test is used to test whether two independent sample means are statistically different.
- T F h. The dependent t-test is used to test whether two groups differ after receiving some type of intervention.

Chapter 11

F-Test

Analysis of Variance

Sir Ronald Fisher, upon discovering the work of W. S. Gossett (developed a test of mean differences between two groups, “student t-test,” or simply the t-test), extended it into his idea for an analysis of variance (ANOVA) technique. The analysis of variance technique was based on the idea that variance could be used to indicate whether sample means differed. For example, if three groups had the same average math score, then the variance (standard deviation) of the means would be zero, implying that the sample means do not vary (differ). As the sample means become more different, the variance increases. When sample means are similar in value they are called *homogeneous*, when they become more different in value they are called *heterogeneous*.

The basic idea of variance indicating sample mean differences can be understood using the following two sets of mean scores, which indicate similar grand means but different variances:

Population	SET A	SET B
Sample Mean 1	4	3
Sample Mean 2	4	4
Sample Mean 3	4	5
Grand Mean	4	4
Variance	0	1

The sample means are all the same in SET A; hence the variance is zero. The sample means are different in SET B; hence the variance is greater than zero. As sample means become different in value, the variance will become larger.

Fisher determined that the variance of sample means around a common mean (grand mean) could be calculated by determining the sum of squared deviations (SS_B) of the sample means around the grand mean divided by the degrees of freedom (number of groups minus one). Fisher called this average variance a **mean square between groups**; therefore $MS_B = SS_B/df_B$. The variance of scores within a

group can be calculated by determining the sum of squared deviations (SS_w) of the scores around the group mean divided by the degrees of freedom (number of scores in a group minus one). Fisher determined that the sum of squares for each group could be averaged across three or more groups. He called this average variance a **mean square within groups**; therefore $MS_w = SS_w / df_w$.

Fisher developed a test, named the F-test after himself, to indicate a ratio of the variance of sample means around the grand mean to the variance of scores around each sample mean. The F-test was computed as $F = MS_B / MS_w$. The F-test has an expected value of one (based on sampling error only being present); however, in practice, if all sample means are identical, the variance of sample means around the grand mean (MS_B) would be zero, and $F=0$. In some situations, more variance exists within the groups (MS_w) than between the groups (MS_B), which result in F values being greater than zero, but less than one ($0 < F < 1$). Obviously, these are not the expected results when testing for group mean differences, rather we expect the F-test (ratio) to yield larger positive values as group means become more different. F summary tables, which give expected F-values for various combinations of degrees of freedom between groups (df_B) and degrees of freedom within groups (df_w), are in the Appendix. These F tables are used to determine if the F-value you calculate from sample data is larger than expected by chance. If your F value is larger than the tabled value, the group means are significantly different.

Two basic examples will be presented to illustrate how analysis of variance is an extension of both the independent t-test (two mutually exclusive groups) and the dependent t-test (same subjects measured twice).

One-Way Analysis of Variance

The **one-way analysis of variance** extends the independent t-test comparison of two group means to three or more group means. For example, the sales at three different clothing stores are compared to determine if they are significantly different. The number of customers for each clothing store was:

Sales	Store
30	1
30	1
40	1
40	1
25	2
20	2
25	2
30	2
15	3
20	3
25	3
20	3

The summary statistics for each store were:

	Store 1	Store 2	Store 3	Grand Mean (Total Sales)
Sum	140.00	100.00	80.00	320.00
Mean	35.00	25.00	20.00	26.67
SD	5.77	4.08	4.08	7.79

The number of sales on average for Store 1 is higher than Store 2; likewise Store 2 had more sales on average than Store 3. The analysis of variance procedure can test (F-test) whether the mean differences in customers between the stores are statistically significant. To calculate the F-test, we must first compute the variance of the sample means around the grand mean using the following steps (Note: n_i = number of weeks and j = number of stores):

1. $SS_B = n_i (\text{Store Mean} - \text{Grand Mean})^2$
 $SS_B = 4(35 - 26.67)^2 + 4(25 - 26.67)^2 + 4(20 - 26.67)^2$
 $SS_B = 466.67$
2. $df_B = (j - 1) = (3 - 1) = 2$
3. $MS_B = SS_B / df_B = (466.67 / 2) = 233.33$

The sum of squares between the stores (SS_B) indicates the sum of the deviation from the grand mean (common mean for all the sample means) which is squared and weighted (multiplied by the sample size) for each store (Step 1 above). The degree of freedom is simply the number of stores minus one (Step 2 above). The mean square between the stores, which is used in the F-test to indicate mean differences in the stores, is the average sum of squares between the groups (Step 3 above).

Next, we must compute the variance of the number of customers around each store mean using the following steps (Note: j = number of stores):

1. $SS_w = \sum j [(\text{Number of Customers} - \text{Store Mean})^2]$
 $SS_w = [(30 - 35)^2 + (30 - 35)^2 + (40 - 35)^2 + (40 - 35)^2 +$
 $(25 - 25)^2 + (20 - 25)^2 + (25 - 25)^2 + (30 - 25)^2 +$
 $(15 - 20)^2 + (20 - 20)^2 + (25 - 20)^2 + (20 - 20)^2]$
 $SS_w = 200.00$
2. $df_w = (\text{sample size} - j) = (12 - 3) = 9$
3. $MS_w = SS_w / df_w = 22.22$

The sum of squares within the stores (SS_w) indicates the sum of the customer number deviation from the store mean which is squared (Step 1 above). The degrees of freedom equals the total sample size minus the number of stores (Step 2 above). The mean square within the stores, which is used in the F-test to indicate the variance of within the stores, is the average sum of squares within the groups (Step 3 above). This “averaging” of the sum of squares within the stores is sometimes referred to as “pooling” the sum of squares within the groups, i.e., stores.

The **F-test** (ratio) can now be calculated as $F = MS_B / MS_w = 233.33 / 22.22 = 10.50$. This F-test indicates the *ratio* of the variance between the stores over the variance within the stores. If the store means are more different (vary more between stores)

than the scores vary within the stores, the F-test ratio would be greater than an F-value expected by chance. We therefore compare the computed $F=10.50$ to a tabled F-value in the Appendix for $df_B=2$ and $df_W=9$ at the .05 chance level (5% chance our comparison might not yield a correct decision). The computed $F=10.50$ is greater than the tabled $F=4.26$ (value expected by chance), which indicates that the average number of customers per store are significantly different.

The results of all these calculations can be neatly organized in an ANOVA Summary Table. The ANOVA Summary Table indicates that both of the sums of squared deviations can be added together yielding a total sum of squared deviations for all scores ($SS_T = SS_B + SS_W$). Likewise, the degrees of freedom from both calculations can be added together yielding the sample size minus one ($df_T = N - 1$). The SS_T divided by df_T yields the total customer variance, which has been “partitioned” into the variance between groups and the variance within groups. The total number of customer variance is therefore computed as $SS_T / df_T = 666.67 / 11 = 60.61$. The total number of customer standard deviation is simply the square root of this variance, which is equal to 7.79 (see summary statistics above). This reveals to us that the sum of squares total divided by the sample size minus one is actually the variance of all the scores. The square root of this variance is the standard deviation of all the scores. The key point is that the analysis of variance procedure “partitions” the variance of *all* the scores.

ANOVA Summary Table

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Between Groups	466.67	2	233.33	10.50
Within Groups	200.00	9	22.22	
Total	666.67	11		

Multiple Comparison Tests

The F-test doesn't reveal *which* store means are different, only that the store means are different. Consequently, a multiple comparison test is needed to determine which store means are different. There are many different types of multiple comparison tests depending upon whether the group sizes are equal or unequal, whether all group means are compared (pairwise) or combinations of group means are compared (complex contrasts), and whether group mean differences are specified in advance (a priori planned comparison) or after the statistical test (posterior comparison). We will only focus on one type of multiple comparison test called a *post-hoc* test named after Henry Scheffe. Post-hoc tests are named “post hoc” because they are calculated after the ANOVA F-test indicates that the group means are significantly different. The Scheffe post-hoc test can be used in most situations because it is *not* limited to equal sample sizes and pairwise comparisons of group means, and maintains the same probability as that used for determining the significance of F (Tabled $F=4.26$ at 5% chance level). The Scheffe post-hoc test uses the common variance term (pooled variance) from all the groups in the denominator of the formula based on MS_W in the ANOVA Summary Table.

Given three stores, three Scheffe post-hoc tests can be computed using the following formulas.

Store 1 vs. Store 2

$$\text{Scheffe} = (\text{Store 1 Mean} - \text{Store 2 Mean})^2 / \text{MS}_W (1/n_1 + 1/n_2) (j-1)$$

$$\text{Scheffe} = (35 - 25)^2 / 22.22 (1/4 + 1/4) (3-1)$$

$$\text{Scheffe} = (100) / 22.22 (1/2) (2)$$

$$\text{Scheffe} = 4.50$$

Store 1 vs. Store 3

$$\text{Scheffe} = (\text{Store 1 Mean} - \text{Store 3 Mean})^2 / \text{MS}_W (1/n_1 + 1/n_3) (j-1)$$

$$\text{Scheffe} = (35 - 20)^2 / 22.22 (1/4 + 1/4) (3-1)$$

$$\text{Scheffe} = (225) / 22.22 (1/2) (2)$$

$$\text{Scheffe} = 10.13$$

Store 2 vs. Store 3

$$\text{Scheffe} = (\text{Store 2 Mean} - \text{Store 3 Mean}) / \text{MS}_W (1/n_2 + 1/n_3) (j-1)$$

$$\text{Scheffe} = (25 - 20)^2 / 22.22 (1/4 + 1/4) (3-1)$$

$$\text{Scheffe} = (25) / 22.22 (1/2) (2)$$

$$\text{Scheffe} = 1.13$$

The Scheffe post-hoc test results can be summarized in a Scheffe Post-Hoc Summary Table. Comparing the Scheffe results to the Tabled $F=4.26$ reveals that Store 1 had on average the most customers which was significantly higher than both Store 2 and Store 3. Store 2 and Store 3 average number of customers did not differ significantly.

Scheffe post-hoc summary table

Store	Store		
	1	2	3
1			
2	4.50 ^a		
3	10.13 ^a	1.13	

^aScheffe exceeds Tabled $F=4.26$ at the .05 chance level

Repeated Measures Analysis of Variance

The repeated measures analysis of variance technique extends the dependent t-test to three or more groups. The dependent t-test is appropriate in research settings where you only measure the same subjects twice and test whether scores change, i.e., increase or decrease in a hypothesized direction. In the case of repeated measures ANOVA, subjects are measured on three or more occasions (typically over time or different experimental conditions) with the same idea of testing whether scores change in a hypothesized direction. An example will illustrate the repeated measures ANOVA technique.

The average number of children per family (X) in four regions was measured over a 3-year period by a researcher to test whether birth rate was increasing, decreasing, or remaining the same. The following data were collected:

Region	Average number of children per family (X)			
	1995	1996	1997	Total
Northern State	2	5	5	12
Southern State	2	7	5	14
Eastern State	3	2	6	11
Western State	1	2	4	7
n	4	4	4	N=12
Sum	8	16	20	T=44
Mean	2	4	5	

The repeated measure ANOVA “partitions” the total sum of squared deviations around the grand mean into three component variances: REGION, YEAR, and remaining ERROR variance. The steps needed to calculate each are as follows:

TOTAL Sum of Squares

$$SS_{\text{TOTAL}} = \sum[(X)^2] - [(T)^2/N]$$

$$SS_{\text{TOTAL}} = \sum[(2^2) + (2^2) + (3^2) + (1^2) + (4^2)] - [(44)^2/12]$$

$$SS_{\text{TOTAL}} = 202 - 161.33 = 40.67$$

REGION Sum of Squares

$$SS_{\text{REGION}} = \sum[(\text{Total})^2/j] - [(T)^2/N]$$

$$SS_{\text{REGION}} = \sum[(12)^2/3] + [(14)^2/3] + [(11)^2/3] + [(7)^2/3] - [(44)^2/12]$$

$$SS_{\text{REGION}} = \sum[48 + 65.33 + 40.33 + 16.33] - [161.33]$$

$$SS_{\text{REGION}} = 169.99 - 161.33 = 8.66$$

YEAR Sum of Squares

$$SS_{\text{YEAR}} = \sum[(\text{Sum})^2/n] - [(T)^2/N]$$

$$SS_{\text{YEAR}} = \sum[(8)^2/4] + [(16)^2/4] + [(20)^2/4] - [(44)^2/12]$$

$$SS_{\text{YEAR}} = \sum[16 + 64 + 100] - 161.33$$

$$SS_{\text{YEAR}} = 180 - 161.33 = 18.67$$

ERROR Sum of Squares

$$SS_E = SS_{\text{TOTAL}} - SS_{\text{REGION}} - SS_{\text{YEAR}}$$

$$SS_E = 40.67 - 8.66 - 18.67$$

$$SS_E = 13.34$$

These results can be neatly summarized in the table below:

Repeated measures ANOVA Summary Table

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Region	8.66	3	2.89	
Year	18.67	2	9.33	4.20
Error	13.34	6	2.22	
Total	40.67	11		

The F-test in this repeated measures design is *only* calculated for the variable that is measured over time. In this case, the F-test is a test of whether the average birth rate is different over the 3 years. We would compare the computed $F=4.20$ to a tabled F in the Appendix with $df=2, 6$ at the .05 level of significance; which is $F=5.14$. Since the computed $F=4.20$ is not greater than the tabled $F=5.14$, we conclude that the average birth rate across the 3 years does not differ. We do notice that the average birth rate increased from 2 per family to 4 per family from 1995 to 1996, but to only 5 per family in 1997. The mean birth rate increased across the years, but not enough to warrant a beyond chance statistically significant conclusion.

The one-way analysis of variance procedure is an extension of the independent t-test for testing the differences in three or more sample means. The repeated measures analysis of variance procedure is an extension of the dependent t-test for testing the differences in three or more sample means over time. The sum of squared deviations of sample means around the grand mean is called the sum of squares between groups. The sum of squared deviations of sample scores around each sample mean is called the sum of squares within. The sum of squares total is equal to the sum of squares between groups plus the sum of squares within groups. The total variance of all scores is partitioned into the sum of squares between groups and the sum of squares within groups. The degrees of freedom for the sum of squares between groups are the number of groups minus one. The degrees of freedom for the sum of squares within groups are the total sample size minus the number of groups. The F-test is computed as a ratio of the mean square between over the mean square within. Analysis of Variance Summary Tables neatly organize the between group, within group, and total sum of squared deviation scores, degrees of freedom, mean of the squared deviation scores, and F-test. A Scheffe post-hoc test determines which sample means are different between groups after conducting an analysis of variance procedure.

Analysis of Variance R Programs

ONEWAY Program

The **ONEWAY** program begins by defining the data for three groups. Each set of group data is then summarized and put into a matrix as the group size, group mean, and standard deviation of the group. Labels are assigned to the matrix to indicate the groups and the summary statistic headings. The grand mean for all of the groups is calculated and the individual group means and sample sizes are tabulated. Next, the sum of squares between groups (SSB), mean squares between groups (MSB), sum of squares within groups (SSW), degrees of freedom within groups (df_w), and mean squares within groups (MSW) are calculated. These numbers are used to determine

the total sum of squares (SST) and the total degrees of freedom (df_T). The F-ratio is calculated by dividing the mean squares between by the mean squares within. The probability associated with the F-ratio is determined using the **pf function** and the degrees of freedom. The summary statistics for the groups are output followed by the ONEWAY ANOVA summary table.

The **SCHEFFE** program inputs the MS error, df numerator, and df denominator from the Analysis of Variance Summary table. Next, it creates a *sales* and *store* vector which are combined into a data frame named *data*. Group summary statistics are calculated and the one way analysis of variance computed. The tabled critical F-value is computed followed by the three Scheffe paired contrasts. The data, group summary statistics, analysis of variance summary table, tabled (critical) F, and the Scheffe F's are then printed. If the Scheffe F is greater than the critical F, then the pair of group means are statistically different at the .05 level of significance.

The **REPEATED** program defines the data for four individuals on four successive years. The summary information for years is calculated for the number of samples for the year, the sum of values for the year, and the mean value for the year. The total number of values for all individuals (N) is calculated along with the total sum of values (TotSum). Finally, the calculations of the sum of squares, degrees of freedom, and mean squares for individual, year, total, and error components are computed. The sum of squares total (SST) and total degrees of freedom (df_T) are then calculated, but the sum of squares for the individuals (SSI) takes a few iterations through the data to create the intermediate variable IndSqrDev in order to complete the calculation. The degrees of freedom for the individual (df_I) and mean squares for the individual (MSI) are then calculated. The year component is determined similar to the individual component by creating the intermediate variable YearSqrDev. Finally, the sum of squares error (SSE), degrees of freedom error (df_E), and mean squares error (MSE) are calculated. The F-ratio is then calculated with the associated probability value. The summary statistics are output for the four years followed by the REPEATED MEASURES analysis of variance summary table.

ONEWAY Program Output

	Group A	Group B	Group C
Sample Size	4.00	4.00	4.00
Sample Mean	35.00	25.00	20.00
Std Dev	5.77	4.08	4.08

Oneway Analysis of Variance Table					
Source	SS	df	MS	F	p
Between	466.67	2	233.33	10.5	0.004
Within	200	9	22.22		
Total	666.67	11			

Scheffe Program Output

```

sales store
1      30    1
2      30    1
3      40    1
4      40    1
5      25    2
6      20    2
7      25    2
8      30    2
9      15    3
10     20    3
11     25    3
12     20    3

Group Means=35 25 20
Group SD=5.77 4.08 4.08
Sample Size=4 4 4

One-way Analysis of Variance Summary Table

      Df Sum Sq Mean Sq   F value Pr(>F)
store      2 466.7 233.33 10.5   0.004444**
Residuals  9 200.0   22.22
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Critical F for Scheffe Comparision=4.26

Scheffe Post Hoc Comparisons

Store 1 vs. Store 2=4.5
Store 1 vs. Store 3=10.13
Store 2 vs. Store 3=1.13

```

REPEATED Program Output

	Year 1	Year 2	Year 3	Year 4
Samp Size	4.0	4.00	4.00	4.0
Samp Sum	30.0	31.00	55.00	70.0
Samp Mean	7.5	7.75	13.75	17.5

N=16 , T=186

Repeated Measures Analysis of Variance Table

Source	SS	df	MS	F	p
Individual	25.25	3	8.42		
Year	284.25	3	94.75	13.27	0.001
Error	64.25	9	7.14		
Total	373.75	15			

F Exercises

1. Run the ONEWAY program for comparing differences in the means of three groups.

Group A	Group B	Group C
10	16	18
12	9	14
20	7	12
16	14	26

Record the sample size, sample mean and sample standard deviation for each group.

	A	B	C
Sample Sizes	_____	_____	_____
Sample Means	_____	_____	_____
Standard Deviations	_____	_____	_____

Place the sum of squares deviation results in the ANOVA Summary Table. Test whether the computed F is statistically different from the tabled F at the .05 level of significance.

ANOVA Summary Table

Source	SS	df	MS	F
Between				
Within				
Total				

What do you conclude?

2. Run the Scheffe program for comparison of paired group means.

- a. Critical F=_____
 - b. Scheffe F—Group A vs. Group B=_____
 - Scheffe F—Group A vs. Group C=_____
 - Scheffe F—Group B vs. Group C=_____
 - c. What would you conclude?
-
-

3. Run the REPEATED program for comparing repeated measures across 4 years.

	Year 1	Year 2	Year 3	Year 4	
8	9	10	12		
7	6	14	18		
9	6	15	22		
6	10	16	18		
n	_____	_____	_____	_____	N = _____
Sum	_____	_____	_____	_____	T = _____
Mean	_____	_____	_____	_____	

Record the sample size (n), sum, and sample mean above. Also, record the total number of scores (N), and total sum of scores (T). Place the sum of squares deviation results in the ANOVA Summary Table. Test whether the computed F is statistically different from the tabled F at the .05 level of significance.

Repeated measures ANOVA Summary Table

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Individual				
Year				
Error				
Total				

What do you conclude?

True or False Questions

F Test

- T F a. The one-way analysis of variance and the repeated measures analysis of variance are the same procedure.
- T F b. The analysis of variance procedure “partitions” the total variance of the scores.
- T F c. The one-way ANOVA procedure uses an F-test to determine if three or more group means are significantly different.
- T F d. The Scheffe post-hoc test determines specifically which group means are statistically different.
- T F e. The sum of squares between groups in the one-way ANOVA procedure is computed by summing the squared deviations of each score from the sample means.
- T F f. The sum of squares total (SS_T) divided by the degree of freedom (df_T) is the variance of all the scores.
- T F g. The repeated measures analysis of variance procedure conducts an F-test of whether sample means increase over time.
- T F h. The Scheffe post hoc test should *not* be used with unequal sample sizes.

Chapter 12

Correlation

Pearson Correlation

Sir Francis Galton in Great Britain was interested in studying individual differences based on the work of his cousin, Charles Darwin. In 1869, Sir Francis Galton demonstrated that the mathematics scores of students at Cambridge University and the admissions exam scores at the Royal Military College were normally distributed. In 1889, Francis Galton published an essay suggesting the idea for examining how two traits varied together (covaried). This effort resulted in the first use of the term “regression.” Karl Pearson in 1898, based on the suggestions made by Sir Francis Galton, investigated the development of a statistical formula that would capture the relationship between two variables.

The idea was to determine the degree to which two things went together, i.e., how two things varied together. The concept was simple enough in principle, take measurements on two variables, order the measurements of the two variables, and determine if one set of measurements increased along with the second set of measurements. In some cases, maybe the measurements of one variable decreased while the other increased. The basic assumption Karl Pearson made was that the measurements needed to be linear or continuous. He quickly determined that how two things covaried divided by how they individually varied would yield a statistic that was bounded by +1 and -1, depending on the relationship of the two measurements. The conceptual formula he developed, which took into account the covariance between two variables divided by the variance of the two variables, was defined as:

$$r = \frac{\text{Covariance } XY}{(\text{Var } X)(\text{Var } Y)}$$

In 1927, after L. L. Thurstone developed the concept of a standard score (z-score) as the deviation of a raw score from the mean, divided by the standard deviation, the correlation formula was further defined as the average product of standard scores:

$$r = \frac{\sum z_x z_y}{N}$$

An example of the relationship between two continuous variables will better illustrate how the bivariate (two variable) correlated relationship is established. A typical research question for a group of students can be stated as “Is there a significant relationship between the amount of time spent studying and exam scores?” The data for these two continuous variables, ordered by time spent studying, is listed below.

Time spent	Exam Score
1 h	75
1 h	80
2 h	75
3 h	90
3 h	85
4 h	95
4 h	85
5 h	90
5 h	95
6 h	90

A computational version of the correlation formula makes the calculation easier and uses the following summary values:

$$\Sigma X = 34$$

$$\Sigma X^2 = 142$$

$$\Sigma Y = 860$$

$$\Sigma Y^2 = 56800$$

$$\Sigma XY = 3015$$

The computational correlation coefficient formula is:

$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

The expression, SP , is defined as the sum of cross products for X and Y . The expression, SS_x , is the sum of squares X , and the expression SS_y is the sum of squares Y .

These values are computed for each expression in the correlation coefficient formula as:

$$SP = \sum XY - \frac{(\sum X)(\sum Y)}{N} = 3015 - \frac{(34)(860)}{10} = 91$$

$$SS_x = \sum X^2 - \frac{(\sum X)^2}{N} = 142 - \frac{(34)^2}{10} = 26.40$$

$$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{N} = 56800 - \frac{(860)^2}{10} = 490$$

These values are substituted in the Pearson correlation coefficient:

$$r = \frac{SP}{\sqrt{SS_x SS_y}} = \frac{91}{\sqrt{26.4(490)}} = +.80$$

The value of $r=+.80$ indicates a positive relationship between the two variables implying that as the amount of study time increases, exam scores increase. The correlation coefficient also indicates the magnitude of the relationship since the r -value is approaching +1.0, which would indicate a perfect relationship.

Interpretation of Pearson Correlation

The correlation coefficient can be interpreted in several different ways. First we can test it for significance using tabled correlation values for different sample sizes (degrees of freedom). Second, we can square the correlation coefficient to obtain a variance accounted for interpretation. Third, we can graph the relationship between the data points in a scatter plot to visually see the trend of the relationship.

To test the significance of the correlation coefficient, we use our standard hypothesis testing approach:

Step 1: State the Null and Alternative Hypothesis using Population Parameters.

$H_0: \rho=0$ (no correlation)

$H_A: \rho \neq 0$ (correlation exists)

Step 2: Choose the appropriate statistic and state the sample size obtained.

Pearson correlation coefficient for continuous variables

Sample Size, $N=10$

Step 3: State the level of significance, direction of alternative hypothesis, and region of rejection.

Level of Significance (α)=.05

Alternative Hypothesis: Non-directional (Two-tailed test)

For $N=10$, $df=N-1=9$

R: $r_{ta_{bled}} > .602$

Step 4: Collect Data and Calculate Sample Correlation Statistic.

Continuous Variables: Time spent studying and exam scores

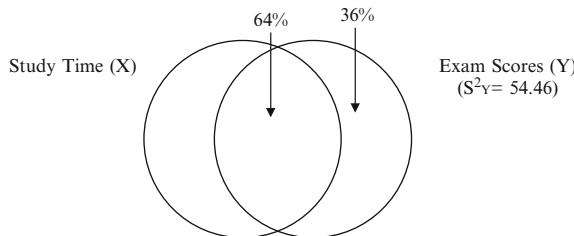
$N=10$ pairs of data

$r=.80$

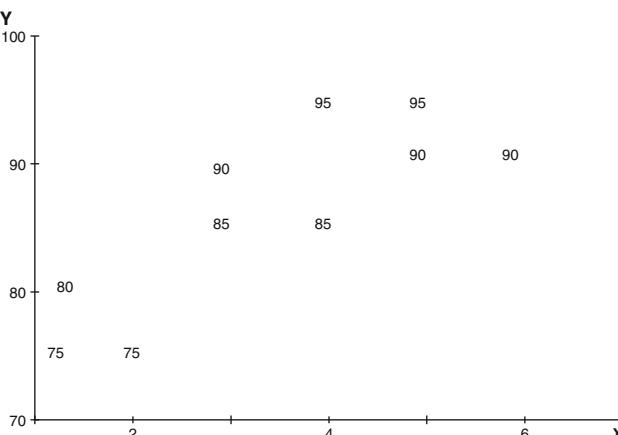
Step 5: Test statistical Hypothesis, make decision, and interpret results.

Since the computed $r=.80$ is greater than the tabled r of .602 at the .05 level of significance for a two-tailed test, reject the null hypothesis and accept the alternative hypothesis. There is a statistically significant relationship between the amount of time spent studying and exam scores.

The second approach to interpreting the Pearson correlation coefficient is to square the sample correlation value. The r^2 value is $(.80)^2 = .64$. This implies that 64% of the variability in exam scores can be explained by knowledge of how much time a student spent studying. This also implies that 36% of the variability in the exam scores is due to other variable relationships or unexplained variance. The average number of hours spent studying was 3.4 h with a standard deviation of 1.71. The average exam score was 86 with a standard deviation of 7.38. The interpretation is linked to the variance of the exam scores, hence $(7.38)^2 = 54.46$. We would state that 64% of 54.46 is explained variability and 36% of 54.46 is unexplained variability given knowledge of how much time a student spent studying. We can also depict this relationship using a *Venn* or *Ballantine* diagram.



The third approach to interpreting the correlation coefficient obtained from sample data is to graph the data points of the two variables. The scatter plot is used for this purpose. We draw a Y-axis for the exam scores and an X-axis for the amount of time spent studying. We label and scale these two axes to provide a grid such that the pairs of data points can be graphed. A visual look at the “trend” of the pairs of data points helps in interpreting whether the positive direction of the correlation coefficient exists. Scatter plots can display an upward trend (positive relationship), downward trend (negative relationship), or a curvilinear trend (one-half positive and the other half negative). If a curvilinear relationship exists, one-half cancels the other half out, so the correlation coefficient would be zero and the interpretation of the correlation coefficient meaningless. This is why Karl Pearson made the assumption of linear data. A scatter plot of the data points visually reveals the positive upward trend expected from $r=+.80$.



The Pearson correlation coefficient in determining whether or not there is a linear relationship between two continuous variables provides both a measure of the strength of the relationship, as well as, the direction of the relationship. In our example, the number of hours a student spent studying was related positively to the exam score. The strength of the relationship was indicated by a value close to 1.0 and the direction of the relationship by the positive sign. We are also able to explain the variability in the exam scores by squaring the correlation coefficient. In other words, why didn't all the students get the same score, because some students studied more! This can be presented in a diagram and depicted as a percent of the variance of the exam scores that can be explained. A scatter plot is the best visual aid to understanding the trend in the pairs of scores in regards to both magnitude and direction.

Karl Pearson's correlation coefficient was one of the most important discoveries in the field of statistics because numerous other statistical techniques, such as multiple regression, path analysis, factor analysis, cluster analysis, discriminant analysis, canonical correlation, and structural equation modeling, are based on this coefficient and interpretative understanding. Over one hundred years later, the examination of variable relationships is the singular most important analysis conducted in education, psychology, business, medicine, and numerous other disciplines. The correlation approach assumes that both the X and the Y variables are random, and have a distribution known as the bivariate normal distribution. In the bivariate normal distribution, for any given X value the Y values have a normal distribution in which the mean and the standard deviation depend on the value of X and the strength of the relationship between X and Y. The strength of the relationship between X and Y is measured by a population parameter ρ (pronounced "rho"), which can range between -1 and 1 inclusive. If $\rho=0$, there is either no relationship between X and Y, or a curvilinear relationship which the Pearson correlation coefficient doesn't detect. If $\rho=1$, there is a perfect *positive* linear relationship between the two variables, and if $\rho=-1$, there is a perfect *negative* linear relationship between the variables. A value of ρ close to zero indicates a weak relationship (assuming linear data), and a value close to either +1 or -1 indicates a strong relationship. Consequently, $r=-.90$ is a stronger relationship than $r=+.50$ for the same sample size. Because the Pearson correlation values form an ordinal scale, we do not directly compare the distances between two correlation values. For example, if a correlation of $r=+.50$ was obtained in one sample and a correlation of $r=+.60$ obtained in a second sample, you *would not* indicate that the second correlation was .10 higher than the first because the correlation is an ordinal scale!

In this chapter, the Pearson correlation coefficient will be calculated using the bivariate normal distribution. Other correlation coefficients have been developed since 1898 to establish the relationship between nominal and ordinal data, but are not presented in this chapter, i.e., phi-coefficient (nominal data), Spearman-coefficient (ordinal data). R has functions for these correlation coefficients making individual calculations straightforward.

In a bivariate normal distribution, both X and Y are random variables. The Pearson correlation coefficient indicates the linear relationship between two continuous variables. The Pearson correlation coefficient indicates both the magnitude

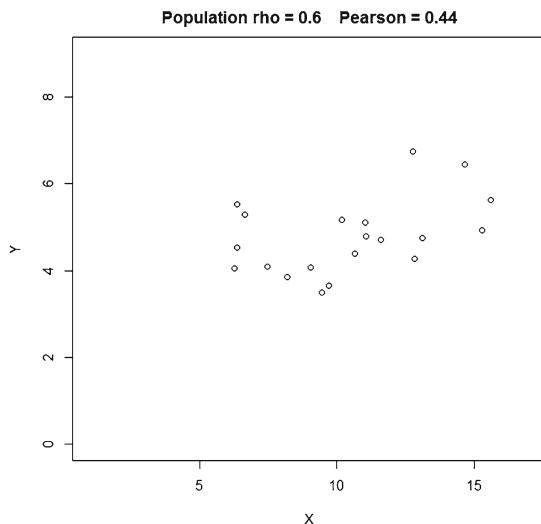
and direction of the relationship between two linear continuous variables. A Pearson correlation coefficient of $r=0$ indicates no linear relationship between two variables. The correlation coefficient can be interpreted in three ways: test of significance, variance accounted for, and a diagram of trend in the paired data points. The sample correlation coefficient, r , is an estimate of the population correlation coefficient ρ . If $\rho=0$, then the sample data points will have a random scatter plot, and the least squares line will be horizontal. As ρ approaches +1 or -1, the sample data points are closer to a straight line, either upward for positive correlations, or downward for negative correlations. If $\rho=+1$ or -1, then the sample points will lie directly in a line.

CORRELATION R Program

The CORRELATION program specifies the value of ρ (correlation) in the population. The program then selects variable X at random from a normal distribution with mean of 10 and standard deviation of 3. Next, a random variable Y is selected from the normal distribution that is determined by the random X variable, given ρ selected and the mean and standard deviation of Y , which are 5 and 1, respectively. A scatter plot is drawn for the pairs of X and Y scores. By varying ρ in the CORRELATION program, you can observe the different scatter plots of data points that arise when ρ has different values in the population. The `cor` function is used to compute the sample Pearson correlation, which is placed into a label to be used later along with a label for the ρ value. The limits of the X and Y axes are set before plotting the pairs of data points so that no plotted points fall outside the axes. Finally, the sample data points are plotted in a scatter plot with the labels for the sample correlation and ρ . The last program line prints out the sample correlation.

CORRELATION Program Output

```
Population rho=0.6
Sample Size=20
Pearson r=0.44
```



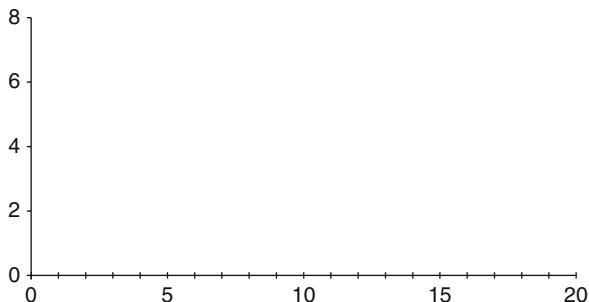
Correlation Exercises

1. Run CORRELATION program for the following values of rho and sample size.
Record the Pearson correlation coefficient.

rho = + .5

Sample Size = 20

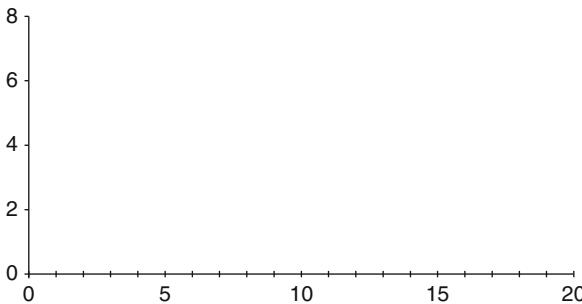
Pearson r = _____



$\rho = -.5$

Sample Size = 20

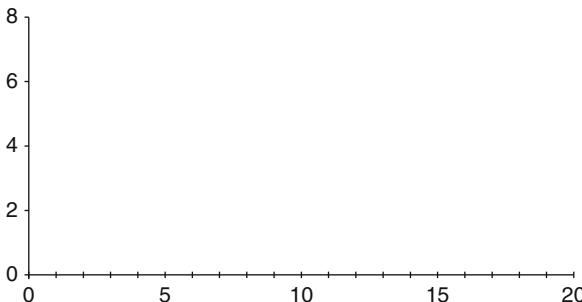
Pearson $r =$ _____



$\rho = 0.0$

Sample Size = 20

Pearson $r =$ _____

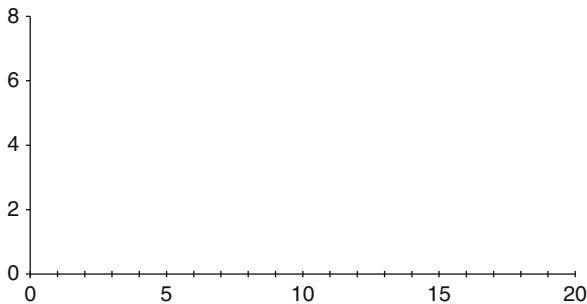


- a. For $\rho = +.5$, is there an upward trend to the points? YES _____ NO _____
 - b. For $\rho = -.5$, is there a downward trend to the points? YES _____ NO _____
 - c. For $\rho = 0.0$, does there appear to be no upward or downward trend to the points? YES _____ NO _____
2. Run CORRELATION program for the following values of rho. Record the correlation coefficient, r. Plot the data values and draw a line over the points on the graph.

$\rho = +1.0$

Sample Size = 20

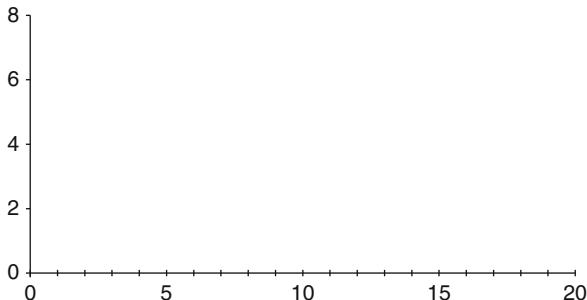
Pearson $r = \underline{\hspace{2cm}}$



$\rho = -1.0$

Sample Size = 20

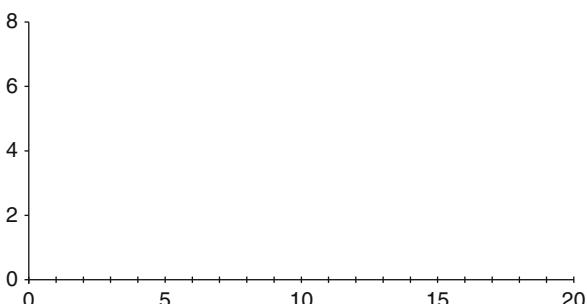
Pearson $r = \underline{\hspace{2cm}}$



$\rho = 0$

Sample Size = 20

Pearson $r = \underline{\hspace{2cm}}$



- a. Describe the scatter plot for the data points when rho = 1.

- b. Describe the scatter plot for the data points when rho = -1.

- c. Describe the scatter plot for the data points when rho = 0.

3. Input sample size = 100, then run Correlation program for rho = .6 to compute Pearson r.

rho = .4, sample size = 20, Pearson r = .42

rho = .4, sample size = 100, Pearson r = _____

Does sample size effect Pearson r? Yes _____ No _____

True or False Questions

Pearson Correlation

- | | | |
|---|---|---|
| T | F | a. The Pearson correlation coefficient indicates the relationship between two linear continuous variables. |
| T | F | b. A correlation of r = .60 is greater than a correlation of r = -.80. |
| T | F | c. A correlation of r = 0 implies no relationship between two variables. |
| T | F | d. If rho is -.5, the scatter plot of data points will indicate an upward trend. |
| T | F | e. If r = .80, then 64% of the variability in one variable is explained by knowledge of the other variable. |
| T | F | f. If rho = 0, then r = 0. |
| T | F | g. Rho is a parameter, and r is a statistic. |
| T | F | h. Data points will fall on a straight line when r = 1.0. |

Chapter 13

Linear Regression

In the late 1950s and early 1960s, the mathematics related to solving a set of simultaneous linear equations was introduced to the field of statistics. In 1961, Franklin A. Graybill published a definitive text on the subject, *An Introduction to Linear Statistical Models*, which piqued the curiosity of several scholars. A few years later in 1963, Robert A. Bottenberg and Joe H. Ward, Jr., who worked in the Aerospace Medical Division at Lackland Air Force Base in Houston, Texas, developed the linear regression technique using basic algebra and the Pearson correlation coefficient. Norman R. Draper and Harry Smith, Jr. in 1966 published one of the first books on the topic, *Applied Regression Analysis*. In 1967, under a funded project by the U.S. Department of Health, Education, and Welfare, W. L. Bashaw and Warren G. Findley invited several scholars to the University of Georgia for a symposium on the general linear model approach to the analysis of experimental data in educational research. The five invited speakers were: Franklin A. Graybill, Joe H. Ward, Jr., Ben J. Winer, Rolf E. Bargmann, and R. Darrell Bock. Dr. Graybill presented the theory behind statistics, Dr. Ward presented the regression models, Dr. Winer discussed the relationship between the general linear regression model and the analysis of variance, Dr. Bargmann presented applied examples which involved interaction and random effects, and Dr. Bock critiqued the concerns of the others and discussed computer programs that would compute the general linear model and analysis of variance. Since the 1960s, numerous textbooks and articles in professional journals have painstakingly demonstrated that the linear regression technique, presented by Bottenberg and Ward, is the same as the analysis of variance. In recent years, multiple regression techniques have proven to be more versatile than analysis of variance in handling nominal and ordinal data, interaction effects, and non-linear effects.

The linear regression equation developed by Bottenberg and Ward was expressed as: $Y = a + bX + e$. The Y variable represented a continuous measure, which was referred to as the dependent variable. The X variable represented a continuous measure, which was called an independent variable, but later referred to as a predictor variable. The value a was termed the “intercept” and represented the value on the Y -axis where the least squares line crossed. The b value was a “weight,” later

referred to as a regression weight or coefficient. The value of e was referred to as prediction error, which is calculated as the difference between the Y variable and the predicted Y value (Yhat) from the linear regression equation, given values for the intercept and regression weight. An example will illustrate the logic behind the linear regression equation.

Regression Equation

Given the following data pairs on the amount of recyclable aluminum in ounces (Y) and the number of aluminum cans (X), a linear regression equation can be created: $Y=a+bX+e$.

Recyclable Aluminum (Y)	Number of Aluminum Cans (X)
1	2
2	4
3	6
4	8
5	10
6	12
7	14

The regression intercept (a) indicates the point on the Y-axis where the least squares line crosses in the scatter plot. The “rise” and “run” or regression weight (b) determines the rate of change, which can be seen by the slope of the least squares line in the scatter plot. It is important to understand that a linear regression equation only refers to the range of values for the pairs of Y and X scores. Given the linear regression equation: $Y=a+bX+e$, the intercept and regression weight (slope) for the data can be calculated as:

$$b = r_{xy} \frac{s_y}{s_x}$$

$$a = \bar{Y} - b\bar{X}$$

The correlation coefficient for these data is $r=+1.0$, the mean of $Y=4$ and the mean of $X=8$. The standard deviation of Y values is 2, and the standard deviation of X values is 4. Placing these values in the intercept and regression weight formula results in: $a=2-(2/4)4=0$ and $b=1(2/4)=.5$, with a linear regression equation: $Y=0+(1/2)X$. Since the intercept is zero, the equation is simply $Y=.5X$. An inspection of the data reveals that 2 aluminum cans yields 1 ounce of recyclable aluminum, 4 aluminum cans yields 2 ounces of recyclable aluminum, and so forth because one-half the number of aluminum cans equals the number of ounces. A scatter plot of these data would indicate the “rise” and “run” of this relationship with the least squares line intersecting the Y-axis at $a=0$. Notice that there is no error in the prediction since every Y value is perfectly predicted by knowledge of X, i.e., $e=Y-Yhat=0$. Perfect relationships like this don't often occur with real data!

Regression Line and Errors of Prediction

A more realistic example will help to demonstrate the linear regression equation, least squares line, and error of prediction. The data for twenty student math achievement scores (Y) and days absent during the week from school (X) are summarized below.

Student	X	Y	X ²	Y ²	XY
1	2	90	4	8100	180
2	4	70	16	4900	280
3	3	80	9	6400	240
4	5	60	25	3600	300
5	1	95	1	9025	95
6	2	80	4	6400	160
7	5	50	25	2500	250
8	3	45	9	2025	135
9	2	75	4	5625	150
10	4	65	16	4225	260
11	5	45	25	2025	225
12	1	80	1	6400	80
13	4	80	16	6400	320
14	5	60	25	3600	300
15	1	85	1	7225	85
16	0	90	0	8100	0
17	5	50	25	2500	250
18	3	70	9	4900	210
19	4	40	16	1600	160
20	0	95	0	9025	0
Σ	59	1405	231	104575	3680

The summary statistics for these data can be hand calculated as follows:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{59}{20} = 2.95 \quad S_x = \sqrt{\frac{SS_x}{N-1}} = \sqrt{\frac{56.95}{19}} = 1.73$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{1405}{20} = 70.25 \quad S_y = \sqrt{\frac{SS_y}{N-1}} = \sqrt{\frac{5873.75}{19}} = 17.58$$

Recall from the previous chapter that the sum of products and sum of squares X and sum of squares Y were used in computing the correlation coefficient:

$$SP = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} = 3680 - \frac{(59)(1405)}{20} = -464.75$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 231 - \frac{(59)^2}{20} = 56.95$$

$$SS_Y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 104575 - \frac{(1405)^2}{20} = 5873.75$$

$$r = \frac{SP}{\sqrt{SS_x SS_Y}} = \frac{-464.75}{\sqrt{56.95(5873.75)}} = -.804$$

The intercept (a) and slope (b) in the linear regression equation can now be computed as:

$$a = \bar{Y} - b\bar{X} = 70.25 - [(-8.16)(2.95)] = 70.25 + 24.07 = 94.32$$

$$b = r_{XY} \frac{S_Y}{S_X} = -.804 \left(\frac{17.58}{1.73} \right) = -8.16$$

These values will closely approximate (within rounding error) those output by a computer program. The prediction of Y given knowledge of X is then possible using the intercept and slope values in the following linear regression equation:

$$\hat{Y} = 94.32 - 8.16X$$

To determine the predicted Y values (\hat{Y}) we would substitute each value of X into the linear regression equation. The resulting Y , \hat{Y} , and errors of prediction are given below.

Y	\hat{Y}	$e(Y - \hat{Y})$
90	78.00	12.00
70	61.68	8.32
80	69.84	10.16
60	53.52	6.48
95	86.16	8.84
80	78.00	2.00
50	53.52	-3.52
45	69.84	-24.84
75	78.00	-3.00
65	61.68	3.32
45	53.52	-8.52
80	86.16	-6.16
80	61.68	18.32
60	53.52	6.48
85	86.16	-1.16
90	94.32	-4.32
50	53.52	-3.52
70	69.84	.16
40	61.68	-21.68
95	94.32	.68

The error of prediction for the first student is computed as follows:

Step 1

$$\hat{Y} = a + bX$$

$$\hat{Y} = 94.32 + (-8.16 * 2) = 78$$

Step 2

$$e = Y - \hat{Y}$$

$$e = 90 - 78 = 12$$

Check on linear equation:

$$Y = a + bX + e$$

$$90 = 94.32 + (-16.32) + 12.00$$

In this data example, the correlation coefficient is negative ($r = -.80$), which indicates that as days absent during the week increases (X), the math achievement scores decrease. This relationship would be depicted as a downward trend in the data points on a scatter plot. Also notice that the data points go together (covary) in a negative or inverse direction as indicated by the negative sign for the sum of products in the numerator of the correlation coefficient formula. We square the correlation coefficient value to obtain a variance accounted for interpretation, i.e., when $r = -.80$, $r^2 = .64$. Knowledge of the number of days absent accounts for 64% of the variance in the math achievement scores.

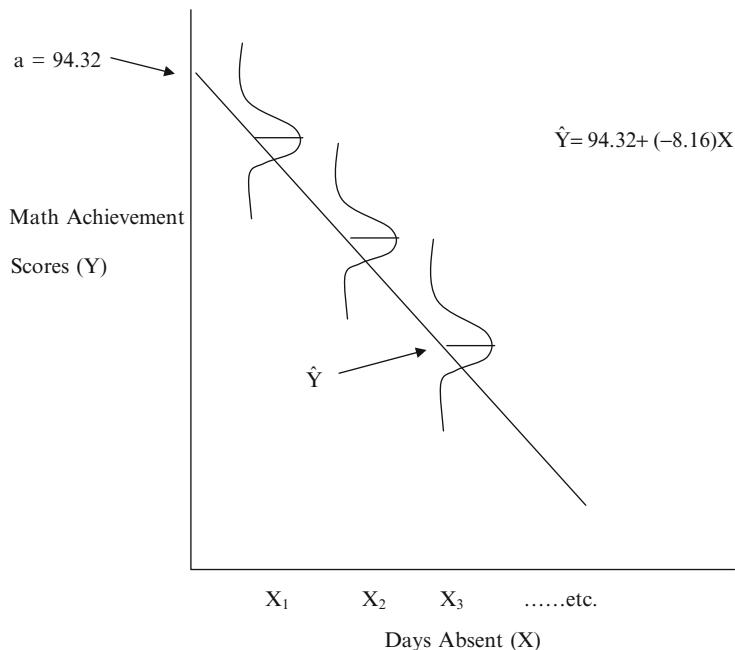
The errors of prediction also serve to identify the accuracy of the regression equation. Notice that some of the errors are positive and some of the errors are negative, consequently the sum (and mean) of the errors should be zero. We expect the Y scores to be normally distributed around \hat{Y} for each value of X so that the variability of these errors indicate the standard deviation of the Y scores around \hat{Y} for each value of X . The standard deviation of the Y scores around \hat{Y} is called a standard error of estimate. It is computed as:

$$S_{Y.X} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{2081.08}{18}} = 10.75$$

Another approach using the standard deviation of Y , the correlation coefficient, and sample size is computed as:

$$S_{Y.X} = S_Y \sqrt{1 - r^2} \sqrt{n - 1 / n - 2} = 17.58 \sqrt{1 - (-.804)^2} \sqrt{20 - 1 / 20 - 2} = 10.75$$

A graph of the Y score distribution around each individual X score will help to better understand the interpretation of the standard error of estimate and the concept of homoscedasticity (equal variance of Y scores around \hat{Y} for each X score along the line of least squares). For each value of X , there is a distribution of Y scores around each \hat{Y} value.



The $S_{Y,X} = 10.75$ is the standard deviation of the Y scores around the predicted Yhat score for each X score. This standard deviation is assumed to be the same for each distribution of Y scores along the least squares line, i.e., homoscedasticity of variance along the least squares line. The predicted Yhat is the mean of the distribution of Y scores for each value of X. The standard error of estimate is therefore calculated as the square root of the sum of the squared differences between Y and Yhat, i.e., $(Y - Y\text{hat})^2$, divided by $N - 2$. Since it is assumed that different values of Y vary in a normal distribution around Yhat, the assumption of equal variance in Y across the least squares line is important because you want an accurate mean squared error!

Standard Scores

In some instances the Y and X scores are converted to z-scores or standard scores to place them both on the same measurement scale. This permits an equivalent “rise” to “run” interpretation of the z-values. The standard scores (z-scores), as they are sometimes called, can be converted back to their respective raw score. The z-score formula subtracts the mean from each score and divides by the standard deviation. The formula you may recall is:

$$z = \frac{X - \bar{X}}{S}$$

As a result of placing Y and X scores on the z-score scale, the intercept (a) and slope (b) in the linear regression equation are simplified because the mean values for X and Y are zero and standard deviations for X and Y are one:

$$a = \bar{Y} - b\bar{X} = (0) - b(0) = 0$$

$$b = r_{XY} \frac{S_Y}{S_X} = -.804 \left(\frac{1}{1} \right) = -.804$$

Because the mean and standard deviation of z-scores are zero (0) and one (1), respectively, the least squares line would pass through the origin ($Y=0$ and $X=0$) of the scatter plot with the Y and X axes labeled in z-score units. Notice that the correlation coefficient captures the slope of the least squares line. The regression equation in z-score form is written as: $Z_Y = \beta Z_X$, with $\beta = -.804$, the Pearson correlation coefficient.

The use of linear regression in applied research is very popular. For example, admission into graduate school is based on the prediction of grade point average using the Graduate Record Exam (GRE) score. Colleges and Universities predict budgets and enrollment from 1 year to the next based on previous attendance data. The Pearson correlation coefficient played an important role in making these predictions possible. A statistically significant correlation between Y and X will generally indicate a good prediction is possible because the difference between the observed Y values and the predicted Yhat values are kept to a minimum. The least squares line is fitted to the data to indicate the prediction trend. The least squares line is a unique regression line, which minimizes the sum of the squared differences between the observed Y's and the predicted Y's, thus keeping prediction error to a minimum by the selection of values for the intercept (a) and slope (b). In the regression formula using z-scores, we see the unique role that the Pearson correlation coefficient plays.

The a in the regression equation is the intercept of the least squares line. The b coefficient in the regression equation is the slope of the least squares line. The intercept in the regression equation is called the Y-intercept; the point at which the least squares line crosses the Y-axis. In the linear regression equation, X is the independent variable and Y the dependent variable. The linear regression equation using z-scores for X and Y, has a slope equal to the Pearson correlation coefficient. The intercept and slope of the least squares line from sample data are estimates of the population intercept and slope. The purpose of linear regression is to predict Y from knowledge of X using a least squares criterion to select an intercept and slope that will minimize the difference between Y and Yhat.

REGRESSION R Program

A true population linear regression equation is specified based on the amount of overtime worked (X) and bonus points received (Y). The true population linear

regression equation can be expressed as: $Y=3+.25X+e$. This equation indicates that if there are no overtime hours worked ($X=0$), the number of bonus points is 3.0, plus some random error (e) that is due to chance, which can be either positive or negative. The bonus points are increased by .25 for each hour of overtime worked. It is assumed that for each X , the Y values are normally distributed around predicted Y , their mean on the line $Y=3+.25X$, and that these normal distributions of Y values around their individual predicted Y have the same variance. The data will be selected at random from a normal population. The scatter plot of X and Y data points is produced by the **plot** function and the least squares regression line is drawn using the **lines** function each time you run the program. The linear regression equation is listed at the top of the scatter plot. The program uses the least squares **lsfit** function in R to calculate the intercept and slope of the regression equation. The program prints the intercept and slope of the true regression equation in the population and the regression equation based on the sample data. The pairs of X and Y values are printed for reference. The correlation coefficient is printed to reference the slope for a regression equation using standard scores for Y and X , which is Beta or the standardized regression coefficient.

REGRESSION Program Output

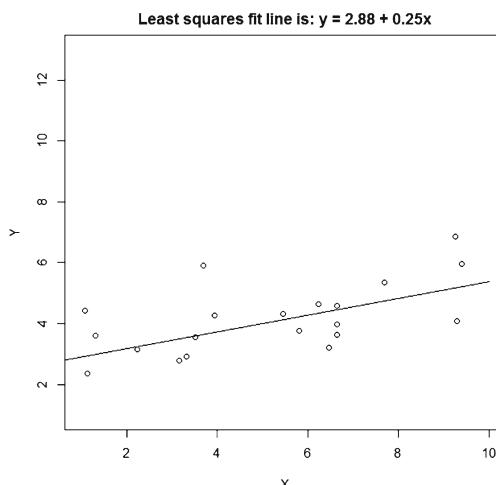
Scatterplot Data Points

```
(3.51,3.56) (3.16,2.79) (9.29,4.08) (1.13,2.36) (3.69,5.9)
(7.69,5.37) (1.08,4.43) (1.32,3.61) (3.32,2.93) (6.64,4.6)
(6.65,3.64) (5.8,3.77) (6.24,4.65) (3.95,4.28) (6.64,3.99)
(2.23,3.15) (9.4,5.96) (5.46,4.33) (6.47,3.21) (9.26,6.86)
```

True regression line is: $y = 3 + 0.25x$

Least squares fit line is: $y = 2.88 + 0.25x$

$r = 0.59$ (slope using standard scores for X and Y)



REGRESSION Exercises

1. Run REGRESSION and enter the 20 pairs of observations in the following spaces.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

- a. What is the equation for the true population least squares line? _____
b. What is the equation of the sample least squares line? _____
c. What is the slope of the equation if using z-scores? (Correlation coefficient) _____
d. Print the scatter plot produced by the REGRESSION program.
2. Run REGRESSION program five more times, and record the sample regression equations.

TRUE REGRESSION EQUATION: $Y = 3 + .25X$

SAMPLE EQUATION Run 1 _____

Run 2 _____

Run 3 _____

Run 4 _____

Run 5 _____

- a. What is the slope of the true regression equation? _____
b. What is the Y-intercept of the true regression equation? _____
c. Calculate the error in the estimates of the sample slopes and intercepts for the five runs.

Enter the values in the table below. (Error = Sample Estimate – True Population Value)

ERROR IN SLOPE

ERROR IN INTERCEPT

RUN 1 _____ _____

RUN 2 _____ _____

RUN 3 _____ _____

RUN 4 _____ _____

RUN 5 _____ _____

d. Are the errors in the slopes in the same direction (positive versus negative)?

YES ____ NO ____

e. Are the errors in the Y-intercepts in the same direction (positive versus negative)?

YES ____ NO ____

3. Run the REGRESSION program and determine for a given value of X (overtime hours worked), what is the bonus received (Y)? Use the following values:

bTrue <- .50

aTrue <- 10

sampleSize <- 100

a. If $X=4$, find the bonus (predicted Y) using the true population regression equation.

b. If $X=4$, find the bonus (predicted Y) using the sample equation.

True or False Questions

Linear Regression

T F a. If the equation of a least squares line is $Y=4-.5X$, then the slope of the line is 4.

T F b. Prediction of Y given knowledge of X is the purpose of linear regression analysis.

T F c. The slope of the linear regression equation in z-score form is the Pearson correlation coefficient.

T F d. A regression equation from sample data will usually differ from the true population regression equation.

T F e. Y values are normally distributed around their \hat{Y} means on the least squares line and the Y distributions have the same variance.

Chapter 14

Replication of Results

We have covered statistical theory, probability, sampling distributions, statistical distributions, hypothesis testing, and various statistical tests based on the level of measurement and type of research design. Research for many decades involved carrying out a single study, i.e., collection of a single random sample of data from the population. Researchers today are becoming more concerned with replicating their research results. However, time and resources often do not permit conducting a research study again. Researchers instead have created techniques that provide some level of replicating their findings.

Every day we ask ourselves important questions. It could be as simple as what route to take to a new job. In the process of answering the question, we gather information or data. This could include asking co-workers, driving different routes, and examining an area roadmap. Once we feel that sufficient information has been collected, we answer the question for ourselves and often share the answer with others. Trying the route to work we selected validates our findings or conclusion. The process of asking a question, gathering data, answering the question, and validating the conclusion is the key to the research process.

Once we have asked an important question, gathered data, and answered the question, others may ask whether the same results would occur if the process were repeated. In other words, could the research findings be replicated? In order to repeat the same process, we must first document the methods and procedures used in the original research. Then, another person can replicate the research process and report their findings. If the research findings are valid and consistent, then others should report findings similar to the original research.

In numerous academic disciplines, research findings are reported along with the methods and procedures used in the study. Unfortunately, not many research studies are replicated. This is due primarily to the time, money, and resources needed to replicate research studies. Instead, other approaches have been developed which don't require conducting the study again. These methods include cross-validation, jackknife, and bootstrap.

Cross Validation

The cross-validation approach involves taking a random sample of data from a population. Typically we would compute a sample statistic for the sample data of size N where the sample statistic is our estimator of the population parameter. In the **cross-validation** approach, the original sample data are randomly split into two equal halves. A sample statistic is computed using one half of the sample data and applied to the other half of the sample data. If the sample statistics for the two randomly split data halves are similar, we assume the research findings would be replicable. Otherwise, the findings are not consistent and probably could not be replicated. A large random sample is generally needed to provide two randomly split data halves that are of sufficient sample size so as not to affect the sample statistics. For example, a sample of size N=1000 randomly drawn from a population would be randomly split into equal halves; each cross-validation sample size would be N=500.

The essence of the cross-validation technique is the computing of a sample statistic on the first set of sample data and applying it to the second set of sample data. This technique is different from other approaches in statistics. We are not comparing sample means from two random samples of data drawn from different populations and testing whether the means are similar or dissimilar (independent t-test for mean differences). We are not testing two sample proportions to determine if they are similar or dissimilar (z-test for differences in proportions). We are not comparing a sample mean to a population mean (one sample t-test). It should be noted that if the original sample statistic was not a good estimator of the population parameter, the cross-validation results would not improve the estimation of the population parameter.

Replicating research results involves conducting another study using the same methods and procedures. When unable to replicate another study, researchers apply the cross-validation approach to a single sample of data randomly drawn from the population. The cross-validation approach involves randomly splitting a random sample into two equal halves, then computing a sample statistic on one sub-sample and applying it to the other sub-sample. The cross-validation approach is not the same as other statistical tests, which randomly sample from a population and test sample estimates of the population parameter. The cross-validation approach does not improve the sample statistic as an estimate of the population parameter. A comparison of two regression equations from two equal halves of sample data will indicate the stability of the intercept and regression weight.

CROSS VALIDATION Programs

The cross validation programs take a random sample from a population. The sample data will be randomly split into two equal halves of size, N/2. The **linear regression** equation, $Y=a+bX+e$, will be computed using the first one-half sample. The regression equation will then be applied to the second one-half sample. The regression weights and R-square values will be compared. A smaller R-squared value in

the second one-half sample is expected because the regression equation coefficients were selected in the first one-half sample data to minimize the sum of squared errors (least squares criterion). The lower R-squared value in the second data set is referred to as the *shrinkage* of the R-squared value. The R-squared value in the second one-half data set will not always be lower than the R-squared value in the first one-half data set. The second cross validation program yields results to determine the amount of expected shrinkage in the second one-half sample regression statistics.

The first program, **CROSSVALIDATION1** uses a traditional method of cross validation whereby a regression equation is created on one randomly chosen half of a sample and applied to the data in the second half to compare stability of regression weights and check for R-squared shrinkage. The second program, **CROSSVALIDATION2**, creates separate independent regression equations on each randomly split half and compares the regression weights and R-squared values. Both programs begin by assigning values for the intercept and slope of the true population regression equation, the sample size, and the number of replications. A replication in the program refers to the splitting of the sample size into random halves and applying the particular method of cross validation. For each replication, a full sample is split into two new random halves, so the same full sample from the population is used for all replications within the same run of the program.

The regression coefficients are computed using the **lsfit** function and the results assigned to the *sampleReg* object. The **ls.print** function is used to output the regression summary statistics and permits selection of specific results, for example, the R-squared value for the full sample is assigned to *sampleR2* and the regression weight is assigned to *bSample*. The other half of the sample is assigned to vectors to be analyzed using the vector notation of *[−halfPoints]*, which means all points that were not in the first vector of random values. The predicted Y-values are then determined from the X-values of the second half using the slope and intercept determined by the regression of the first half of the sample, the sum of squared errors for the regression, the total sum of squared error Y-values and the predicted Y-values. The regression weight and R-squared value for the full sample is displayed using the **cat** function and the output results given by the **print** function. The second program is identical to the first, except that the second half of the sample data is analyzed separately with the **lsfit** function. The output of the second program therefore includes separate regression results for both samples of data for comparison.

CROSS VALIDATION Program Output

CROSSVALIDATION1

Population

Regression equation: $Y = 3 + 0.25$

Sample Size=1000

N Replications=5

Full Sample

Regression equation: $Y = 2.93 + 0.262$

R-Squared=0.3123

	Sample A Reg Weight	Sample A R-Squared	Sample B R-Squared
Replication 1	0.271	0.3476	0.324
Replication 2	0.288	0.3615	0.399
Replication 3	0.255	0.3041	0.289
Replication 4	0.279	0.338	0.373
Replication 5	0.265	0.3124	0.331

CROSSVALIDATION2

Population

Regression equation: $Y = 3 + 0.25 (X)$

Sample Size=1000

N Replications=5

Sample

Regression equation: $Y = 3.086 + 0.241 (X)$

R-Squared=0.2735

	Sample A Reg Weight	Sample A R-Squared	Sample B Reg Weight	Sample B R-Squared
Replication 1	0.243	0.2775	0.24	0.2725
Replication 2	0.238	0.2726	0.243	0.2742
Replication 3	0.264	0.3172	0.218	0.2318
Replication 4	0.247	0.2745	0.234	0.2724
Replication 5	0.253	0.2717	0.229	0.2767

Cross Validation Exercises

- Run the CROSSVALIDATION1 program 5 times for an original sample size of 500. Record the regression weight and the two R-square values for *each* sample of size N=250. Record the regression weight and R-square value for the original sample.

- Original Sample (N=500): Regression Weight _____
R-Square _____.

b.

	Sample A (N=250) Regression weight	R-Square	Sample B (N=250) R-Square
1.	_____	_____	_____
2.	_____	_____	_____
3.	_____	_____	_____
4.	_____	_____	_____
5.	_____	_____	_____

- c. The regression equation computed for sample A is applied to sample B. Are the R-squared values similar in sample A and sample B for the five replications?

YES _____ NO _____.

2. Run the CROSSVALIDATION1 program with 5 replications and a sample size of 1,000.

- a. Compare the regression equation results for the original samples of N=500 and N=1000. Record the SAMPLE regression weights and R-square values.

N=500 Regression Weight_____ R-square _____

N=1000 Regression Weight_____ R-square _____

- b. Does sample size affect the R-square value? YES _____ NO _____.

- c. Which sample size would give better estimates of R-square and the regression weight?

N=500 _____ N=1000 _____

3. Run the CROSSVALIDATION2 program with 5 replications using an original sample size of 500. Record the regression weights and R-square values for each sample of size N=250. Record the regression weight and R-square value for the original sample.

- a. Sample (N=500): Regression Weight _____ R-Square _____.

b.

	Sample A (N=250) Regression Weight	R-Square	Sample B (N=250) Regression Weight	R-Square
1.	_____	_____	_____	_____
2.	_____	_____	_____	_____
3.	_____	_____	_____	_____
4.	_____	_____	_____	_____
5.	_____	_____	_____	_____

- c. The regression equations for sample A and sample B are computed independently. Are the R-squared values similar in sample A and sample B for the five replications?

YES _____ NO _____.

4. Run the CROSSVALIDATION2 program with 5 replications having an original sample size of 1,000.

- a. Compare the regression equation results for the original samples of $N=500$ and $N=1000$. Record the SAMPLE regression weights and R-square values.

$N=500$ Regression Weight _____ R-square _____

$N=1000$ Regression Weight _____ R-square _____

- b. Does sample size affect the R-square value? YES _____ NO _____.

- c. Which sample size would give better estimates of R-square and the regression weight? $N=500$ _____ $N=1000$ _____

Jackknife

The jackknife procedure involves the use of a single random sample of data drawn from a population of data. Recall that the sample statistic is an estimate of the population parameter. We also learned in previous chapters that how good the sample statistic is as an estimate of the population parameter depends on the sample size. The jackknife procedure is concerned with whether the sample statistic as an estimate of the population parameter is affected by any single data value. For example, we know that the sample mean is affected by extreme data values, which is indicated by the standard deviation of the sample data.

The jackknife approach uses a single sample of data, computes the original sample statistic (e.g., sample mean), and then computes the sample statistic for each sample of size $N-1$. Basically, each sample mean after the original sample mean would be computed based on the omission of one data point. The jackknife approach is therefore useful in determining whether an influential data point exists that dramatically changes the sample statistic. The jackknife procedure can be applied to any sample statistic based on a random sample drawn from a population. The jackknife method computes a jackknife mean based on the exclusion of a different data point each time. The number of jackknife replications therefore typically equals the original sample size so that the influence of each data point on the sample statistic can be determined.

An example might help to better understand the jackknife procedure. A random sample of 10 numbers is drawn from a population. The numbers are 2, 4, 9, 12, 8, 7, 15, 11, 3, and 14. The sum of the numbers is 85. The mean is calculated as 85 divided by 10, which equals 8.5. This sample mean is an estimate of the population mean. The jackknife procedure calculates 10 additional sample means based on $N=9$, but each time with a different data value omitted. To compute each jackknife

mean, the data value omitted is subtracted from the sum of 85 and this new sum is divided by $N=9$ to yield the jackknife mean. The jackknife procedure for these data values is outlined below.

Sample Size	Mean	Values Used	Value Omitted
10	8.5	2, 4, 9, 12, 8, 7, 15, 11, 3, 14	None
9	9.2	4, 9, 12, 8, 7, 15, 11, 3, 14	2
9	9.0	2, 9, 12, 8, 7, 15, 11, 3, 14	4
9	8.4	2, 4, 12, 8, 7, 15, 11, 3, 14	9
9	8.1	2, 4, 9, 8, 7, 15, 11, 3, 14	12
9	8.5	2, 4, 9, 12, 7, 15, 11, 3, 14	8
9	8.6	2, 4, 9, 12, 8, 15, 11, 3, 14	7
9	7.7	2, 4, 9, 12, 8, 7, 11, 3, 14	15
9	8.2	2, 4, 9, 12, 8, 7, 15, 3, 14	11
9	9.1	2, 4, 9, 12, 8, 7, 15, 11, 14	3
9	7.8	2, 4, 9, 12, 8, 7, 15, 11, 3	14

The jackknife sample means ranged from 7.7 to 9.2 with the original sample mean of 8.5. The omission of a low data value inflated (increased) the sample mean as an estimate of the population mean. The omission of a high data value deflated (lowered) the sample mean as an estimate of the population mean. Both of these outcomes are expected and help us to understand the nature of how extreme data values (outliers) affect the sample statistic as an estimate of a population parameter.

Another example of the jackknife procedure will highlight the detection of an influential (outlier) data value. Once again, we randomly sample 10 data values from a population. The sum of the numbers is 158 with an original sample mean of 15.8. The results are summarized below. The jackknife means in this second example ranged from 9.2 to 17.2 with an original sample mean of 15.8. Notice that the original sample mean of 15.8 is less than every other jackknife mean, except the one with an omitted influential data value, i.e., 75. The results indicate how the removal of an influential data value can increase or decrease the sample statistic value.

Sample Size	Mean	Values Used	Value Omitted
10	15.8	75, 4, 9, 12, 8, 7, 15, 11, 3, 14	None
9	9.2	4, 9, 12, 8, 7, 15, 11, 3, 14	75
9	17.1	75, 9, 12, 8, 7, 15, 11, 3, 14	4
9	16.5	75, 4, 12, 8, 7, 15, 11, 3, 14	9
9	16.2	75, 4, 9, 8, 7, 15, 11, 3, 14	12
9	16.6	75, 4, 9, 12, 7, 15, 11, 3, 14	8
9	16.7	75, 4, 9, 12, 8, 15, 11, 3, 14	7
9	15.9	75, 4, 9, 12, 8, 7, 11, 3, 14	15
9	16.3	75, 4, 9, 12, 8, 7, 15, 3, 14	11
9	17.2	75, 4, 9, 12, 8, 7, 15, 11, 14	3
9	16.0	75, 4, 9, 12, 8, 7, 15, 11, 3	14

An important comparison can be made between the first and second example. In the first example, the original sample mean of 8.5 fell between all jackknife means, which ranged from 7.7 to 9.2. In the second example, the jackknife means were all greater than the original sample mean, with the exception of the one influential data value. If we examine the jackknife means for each omitted data value, it points out the presence of an influential data value (outlier or extreme value).

The descriptive information for the jackknife means in both examples also clearly indicates the presence of influential data in the second example. The jackknife means were treated as new data such that the average is the *mean* of the jackknife means. The standard deviation, variance, and 95% confidence interval indicate more data dispersion in the second example. This dispersion is interpreted as less accuracy, more variability, and greater difference between the original sample mean and the jackknife means. The descriptive information is:

Descriptive Information	First Example	Second Example
Sample Size	N=10	N=10
Range	7.7–9.2	9.2–17.2
Mean	8.46	15.77
Standard Deviation	.52	2.35
Variance	.27	5.52
95% Confidence Interval	(7.44, 9.48)	(11.16, 20.38)

The jackknife procedure uses the original random sample drawn from a population to estimate additional sample means based on $N-1$ sample data points. The jackknife procedure is useful for identifying influential, extreme, or outlier data values in a random sample of data. The jackknife method can be used with any sample statistic that is computed from a random sample of data drawn from a well-defined population. The jackknife approach helps to validate whether a sample of data provides a good sample statistic as an estimator of the population parameter. The number of jackknife means is equal to the sample size for the purposes of detecting an influential data value. The confidence interval around a set of jackknife means will be smaller when influential data values are not present.

JACKKNIFE R Program

The JACKKNIFE program utilizes the jackknife function in the R library bootstrap. Therefore the program must first issue the command: `library(bootstrap)`. The jackknife function then permits an easy method for calculating the mean and percentiles after jackknifing the samples. The calculations could be done manually instead of using the jackknife function, but the function makes it easier to compute for large samples. The jackknife function is given the vector of data values to be jackknifed. The mean and standard deviation for the entire sample are printed out, followed by the summary of the jackknife results.

JACKKNIFE Program Output

Data Values=10 20 30 40 50 60 70 80 90 100

Original mean=55 Original standard deviation=30.28

Sample Size	Jackknife Mean	Values Used	Value Omitted
9	60	20,30,40,50,60,70,80,90,100	10
9	58.89	10,30,40,50,60,70,80,90,100	20
9	57.78	10,20,40,50,60,70,80,90,100	30
9	56.67	10,20,30,50,60,70,80,90,100	40
9	55.56	10,20,30,40,60,70,80,90,100	50
9	54.44	10,20,30,40,50,70,80,90,100	60
9	53.33	10,20,30,40,50,60,80,90,100	70
9	52.22	10,20,30,40,50,60,70,90,100	80
9	51.11	10,20,30,40,50,60,70,80,100	90
9	50	10,20,30,40,50,60,70,80,90	100

Jackknife Exercises

1. Run the JACKKNIFE program with the following 10 data values as a random Sample A: data<- c(1,2,3,4,5,6,7,8,9,10). Print the Graph.
 - a. Record the Original mean and standard deviation.
Original Mean=_____ Original Standard Deviation=_____
 - b. Record the following information for the N-1 data sets.

Run	Sample Size	Jackknife	
		Mean	Value Omitted
1	9		
2	9		
3	9		
4	9		
5	9		
6	9		
7	9		
8	9		
9	9		
10	9		

- c. Calculate the Standard Error of the Mean using Original sample standard deviation and the number of replications, N=10. $SE = S / \sqrt{N} =$ _____ = _____.

- d. Calculate the 95% Confidence Interval using the Original sample mean and standard deviation. $95\%CI = \bar{X} \pm 1.96(S) = (\underline{\hspace{2cm}}, \underline{\hspace{2cm}})$
- e. List the Range of Jackknife Means. Highest Mean Lowest Mean
- f. How many Jackknife means are higher than the Original Sample mean?
2. Run the JACKKNIFE program with the following 10 data values as a random Sample B: data<- c(1,2,3,4,5,6,7,8,9,100). Print the Graph.
- a. Record the Original sample mean and standard deviation.
Original Mean = Original Standard Deviation =
- b. Record the following information for the N – 1 data sets
- | Run | Sample Size | Jackknife | |
|-----|-------------|-----------|---------------|
| | | Mean | Value Omitted |
| 1 | 9 | | |
| 2 | 9 | | |
| 3 | 9 | | |
| 4 | 9 | | |
| 5 | 9 | | |
| 6 | 9 | | |
| 7 | 9 | | |
| 8 | 9 | | |
| 9 | 9 | | |
| 10 | 9 | | |
- c. Calculate the Standard Error of the Mean using Original sample standard deviation and the number of replications, N = 10.
 $SE = S / \sqrt{N} = \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$.
- d. Calculate the 95% Confidence Interval using the Original sample mean and standard deviation. $95\%CI = \bar{X} \pm 1.96(S) = (\underline{\hspace{2cm}}, \underline{\hspace{2cm}})$
- e. List the Range of Jackknife Means. Highest Mean Lowest Mean
- f. How many Jackknife means are higher than the Original Sample mean?
3. List the Original sample means and standard deviations, SE, 95% CI, and range of Jackknife means for Sample A and Sample B above.
- | Sample | Mean | Sample SD | SE | 95%CI | Jackknife Range |
|----------|-------------------|-------------------|-------------------|-----------------------------------|--|
| Sample A | <u> </u> | <u> </u> | <u> </u> | (<u> </u> , <u> </u>) | High: <u> </u> Low: <u> </u> |
| Sample B | <u> </u> | <u> </u> | <u> </u> | (<u> </u> , <u> </u>) | High: <u> </u> Low: <u> </u> |

- a. Does Sample A or Sample B have a higher sample mean?
Sample A _____ Sample B _____
- b. Does Sample A or Sample B have a larger standard deviation?
Sample A _____ Sample B _____
- c. Does Sample A or Sample B have a larger Standard Error of the Mean?
Sample A _____ Sample B _____
- d. Does Sample A or Sample B have a wider 95% Confidence Interval?
Sample A _____ Sample B _____
- e. Does Sample A or Sample B have more Jackknife Means higher than the Original sample mean?
Sample A _____ Sample B _____
- f. Which sample has a more accurate sample mean estimator of the population mean?
Sample A _____ Sample B _____
- g. Summarize **a** to **e** above in regard to their indicating influential data points.

Bootstrap

The bootstrap method differs from the traditional parametric approach to inferential statistics because it uses *sampling with replacement* to create the sampling distribution of a statistic. The bootstrap method doesn't take a random sample from a population in the same way as that used in our previous inferential statistics. The bootstrap method is useful for reproducing the sampling distribution of any statistic, e.g., the median or regression weight. The basic idea is that conclusions are made about a population parameter from a random sample of data, but in which a sampling distribution of the statistic is generated based on sampling with replacement. The factors that influence the shape of the sampling distribution are therefore important because it is the bootstrap estimate from the sampling distribution that allows us to make an inference to the population.

The bootstrap procedure uses a random sample of data as a substitute for the population data. The randomly sampled data acts as a “pseudo” population from which the bootstrap method repeatedly samples the data. The repeated sampling of data is done with replacement of each data point after selection. The resampling technique therefore samples from the “pseudo” population using the same set of data values each time. Since each data value is replaced before taking the next random selection, it is possible to have the same data value selected more than once and used in the calculation of the final sample statistic. The probabilities in the earlier chapters of the book were based on randomly sampling *without* replacement where each individual, object, or event had an equally likely chance of being selected.

The bootstrap method uses probabilities based upon randomly sampling *with* replacement where each individual, object, or event has an equally likely chance of being selected each time a data value is randomly drawn.

The bootstrap procedure involves the following steps:

Step 1: A random sample of data for a given sample size N is drawn from the population with mean, μ , and standard deviation, σ .

Step 2: The random sample of data size N acts as a “pseudo” population with mean, μ^* , and standard deviation, σ^* .

Step 3: The bootstrap method takes n bootstrap samples of sample size N from the “pseudo” population, each time replacing the randomly sampled data point. For each sample of size N a sample statistic is computed.

Step 4: A frequency distribution of the n bootstrap sample statistics is graphed which represents the sampling distribution of the statistic. The mean of this sampling distribution is the bootstrap estimate, θ^* , which has a standard error of SE_{θ^*} computed by:

$$SE_{\theta} = \sqrt{\frac{\sum(\theta_i^* - \theta^*)^2}{n-1}}$$

Step 5: The amount of bias in the original sample statistic as an estimate of the population parameter is calculated by subtracting: $\mu^* - \theta^*$. If the bootstrap estimate is similar to the corresponding “pseudo” population parameter, then no bias is present. A small difference would still indicate that the original sample statistic is a good estimator of the population parameter.

Step 6: Calculate a confidence interval around the bootstrap estimate using the standard error of the bootstrap estimate and level of significance, Z. The confidence interval is computed by:

$$CI_{\theta} = \theta^* \pm Z(SE_{\theta^*}).$$

The bootstrap method can be based on samples of size N that equal the original sample size N or are larger when it involves sampling data points with replacement. Most applications resample to produce sample sizes equal to the “pseudo” population size. The bootstrap method can also be used to determine the amount of bias between any sample statistic and population parameter. It should be noted that the bootstrap method is only useful for determining the amount of bias in the sample statistic when the original sample is randomly drawn and representative of the population. This makes sense because if the original sample data were not representative of the population, then creating the sampling distribution of this data would erroneously indicate population characteristics.

An example will help to clarify how the bootstrap method is used to determine the amount of bias in the original sample statistic as an estimate of the population parameter. A random sample of 100 data points is drawn from the population. This random sample now becomes a “pseudo” population. The “pseudo” population has a mean, $\mu^*=50$, and standard deviation, $\sigma^*=20$. The bootstrap procedure will randomly sample data points with replacement from this “pseudo” population. The bootstrap sample size will be $n=10$ and the number of bootstrap samples will be $N=5$. The resulting data for each bootstrap sample is given below.

Sample		Bootstrap	Bootstrap	Bootstrap
Run	Size	Data	Mean	Standard Deviation
1	10	10, 30, 35, 40, 50, 80, 90, 75, 20, 60	49.0	26.75
2	10	15, 25, 75, 40, 55, 55, 95, 70, 30, 65	52.5	24.97
3	10	85, 45, 35, 25, 45, 60, 75, 80, 90, 15	55.5	26.40
4	10	20, 30, 45, 50, 55, 65, 10, 70, 85, 95	52.5	27.41
5	10	50, 50, 20, 45, 65, 75, 30, 80, 70, 30	51.5	20.69

The bootstrap estimate, based on the average of the sampling distribution of bootstrap means is $\theta^*=52.2$. The standard error of the bootstrap estimate, based on the square root of the sum of squares difference between each bootstrap sample mean and the bootstrap estimate, divided by the number of bootstrap samples minus one, is $SE_{\theta^*}=2.33$. To establish a 95% confidence interval, a $Z=1.96$ value under the normal distribution is used. The 95% confidence interval is therefore computed as:

$$\begin{aligned} 95\% CI_{\theta} &= \theta^* \pm Z (SE_{\theta^*}) \\ 95\% CI_{\theta} &= 52.2 \pm 1.96 \cdot 21.80 / 4 \\ 95\% CI_{\theta} &= 52.2 \pm 1.96 (2.33) \\ 95\% CI_{\theta} &= 52.2 \pm 4.57 \\ 95\% CI_{\theta} &= (47.63, 56.77) \end{aligned}$$

The amount of bias is indicated by, $\mu^* - \theta^*$, which is $50 - 52.2 = -2.2$. On the average, the bootstrap means were 2.2 units higher than the “pseudo” population mean. The sign of the *bootstrap estimate* will be either positive or negative depending upon whether the bootstrap estimate is lower or higher than the “pseudo” population parameter, respectively. Since the *bootstrap confidence interval* captures the “pseudo” population mean, we would conclude that the original sample mean is a good stable estimate of the population mean.

The bootstrap method uses a random sample of data as a “pseudo” population. The bootstrap procedure resamples the “pseudo” population with replacement. The bootstrap samples of data can contain some of the same data points. The bootstrap estimate is the average of the bootstrap sample statistics. The bootstrap standard deviation is based on the bootstrap data. The bootstrap confidence interval should capture the “pseudo” population parameter when the original sample statistic is a good estimate of the population parameter. The bootstrap method is used to determine the amount of bias between the “pseudo” population parameter and the bootstrap estimate. Similar values indicate no bias. The bootstrap method can be used

with any sample statistic to help determine if the sample statistic is a good or stable estimator of the population parameter. The bootstrap method is not useful when the random sample of data is not representative of the population data.

BOOTSTRAP R Program

The BOOTSTRAP program uses the built-in **bootstrap** function from library(bootstrap). The program inputs the sample size and the number of bootstrap samples. A random sample of data from a normal distribution with a mean of 50 and standard deviation of 10 is generated. The bootstrap function is then performed for the number of bootstrap samples when calculating the mean. The Observed mean corresponds to the mean of the sample from the population and the Bootstrap Mean corresponds to the mean of the bootstrap samples. The Bias is the difference between the Observed Mean and the Bootstrap Mean. The standard error of the bootstrap estimates is reported and used to create the 95% confidence interval around the Bootstrap Mean.

The Observed Mean falls within the 95% confidence interval when the sample data is considered representative of the population.

BOOTSTRAP Program Output

```
Sample Size=100
N Bootstraps=500

Observed Mean = 52.51114
Bootstrap Mean=50.95276

Bias=1.558382
SE=2.057323

95% CI=Bootstrap Mean +/- 1.96SE
95% CI=( 48.89543 53.01008 )
```

Bootstrap Exercises

1. Run the BOOTSTRAP program for a random sample of N=200, then take 20 bootstrap samples. The program settings should be:

```
sampleSize<- 200
numBootstraps<- 20
```

- a. Record the Observed Mean, Bootstrap Mean, Bias, and Standard Error.

Observed Mean _____

Bootstrap Mean _____

Bias _____

SE _____

- b. Calculate the 95% Confidence Interval around the Bootstrap Mean.

95%CI=Bootstrap Mean \pm 1.96 (SE)=(_____, _____)

- c. What does the bootstrap results indicate given the bias and confidence interval?

2. Run the BOOTSTRAP program for a random sample of N=1000, then take 40 bootstrap samples. The program settings should be:

```
sampleSize <- 1000  
numBootstraps <- 40
```

- a. Record the Observed Mean, Bootstrap Mean, Bias, and Standard Error.

Observed Mean _____

Bootstrap Mean _____

Bias _____

SE _____

- b. Calculate the 95% Confidence Interval around the Bootstrap Mean.

95% CI=Bootstrap Mean \pm 1.96 (SE)=(_____, _____)

- c. What does the bootstrap indicate given the bias and confidence interval?

- d. Does the number of bootstrap samples provide a better bootstrap estimate?

YES ____ NO ____

- e. What would happen if the random sample of data from the population was not representative?

True or False Questions

Cross Validation

- | | | |
|---|---|--|
| T | F | a. Cross-validation techniques verify that the sample statistic is a good estimator of the population parameter. |
| T | F | b. Cross-validation techniques involve splitting a random sample from a population into two equal halves. |
| T | F | c. The cross-validation approach involves computing sample statistics using one sub-sample and applying them to the second sub-sample. |
| T | F | d. Cross-validation techniques require large sample sizes. |
| T | F | e. A replication of findings generally requires conducting another study using the methods and procedures of the original study. |

Jackknife

- | | | |
|---|---|--|
| T | F | a. The jackknife procedure is useful for detecting influential data values. |
| T | F | b. A jackknife mean computed with an influential data value does not fall within the confidence interval of the jackknife means. |
| T | F | c. The standard deviation of the jackknife means shows whether more variability is present, hence influential data values. |
| T | F | d. The jackknife approach can be used with any sample statistic computed from random samples of data drawn from a population. |
| T | F | e. The jackknife procedure repeatedly samples the population data to determine if the sample mean is a good estimate of the population mean. |
| T | F | f. The number of jackknife means is typically equal to the original sample size for the purposes of detecting influential data points. |

Bootstrap

- | | | |
|---|---|---|
| T | F | a. The bootstrap method is used to determine if the sample statistic is a stable estimator of the population parameter. |
| T | F | b. The “Observed Mean” will always fall in the bootstrap confidence interval. |
| T | F | c. A random sample of data must be representative of the population before the bootstrap procedure is accurate. |
| T | F | d. The bootstrap procedure involves sampling data with replacement. |
| T | F | e. No bias between the “observed mean” and the bootstrap mean estimate indicates that the sample statistic is a good estimator of the population parameter. |
| T | F | f. The bootstrap method can be used with any sample statistic computed from a random sample of data. |
| T | F | g. The bootstrap procedure creates random samples of data where each data value is unique. |
| T | F | h. The resampling method draws random samples from the “pseudo” population using the same set of data values each time. |

Chapter 15

Synthesis of Findings

Much of the research conducted in education, psychology, business, and other disciplines has involved single experiments or studies that rarely provide definitive answers to research questions. We have learned that researchers seldom replicate their research studies and instead use cross-validation, jackknife, or bootstrap methods to estimate the stability and accuracy of a sample statistic as an estimate of a population parameter. The world around us is understood better when we discover underlying patterns, trends, and principles, which can result from an accumulation of knowledge gained from several studies on a topic of interest. Consequently, a review of the research literature is invaluable in summarizing and understanding the current state of knowledge about a topic. Rather than rely on subjective judgments or interpretations of the research literature, meta-analysis techniques provide a quantitative objective assessment of the study results.

Meta-Analysis

Meta-analysis is the application of statistical procedures to the empirical findings in research studies for the purpose of summarizing and concluding whether the findings in the studies overall were significant. The meta-analytic approach therefore provides a method to synthesize research findings from several individual studies. It makes sense that if several hundred studies had researched socioeconomic status and achievement in school, that a summarization of the findings in these individual studies would aid our understanding of this relationship. From a scholarly point of view, we might ask, “After 20 years of studying the efficacy of psychotherapy, what have we learned?” At some point in time, it becomes frugal to assess what we have learned from the research that was conducted. Further research in an area may be unproductive, unscientific, or take a wrong direction, if we don’t stop from time to time and assess what we have learned. If we view science as an objective activity, then it becomes critical to establish an objective method to integrate

and synthesize similar research studies. Meta-analysis is therefore an objective analysis of the statistical analyses from several individual studies for the purpose of integrating and summarizing the findings.

Sir Ronald A. Fisher in 1932, Karl Pearson in 1933, and Egon S. Pearson in 1938 all independently addressed the issue of statistically summarizing the results of research studies. In 1952, Mordecai H. Gordon, Edward H. Loveland, and Edward E. Cureton produced a chi-square table for use in combining the probability values from independent research studies. In 1953, Lyle V. Jones and Donald W. Fiske further clarified their approaches for testing the significance of results from a set of combined studies. They further demonstrated Fisher's approach of taking a natural logarithm of a p-value to calculate a summary chi-square value. Gene Glass in 1976 is credited with using the term *meta-analysis* to describe the statistical analysis of a collection of analysis results from several individual studies. He provided guidelines for converting various statistics into a common metric. Jacob Cohen in 1965 and again in 1977 provided measures of effect size for many common statistical tests. An effect size measure indexes the degree of departure from the null hypothesis in standard units. Examples of these various approaches to combining results from several research studies will be presented in this chapter.

A Comparison of Fisher and Gordon Chi-Square Approaches

The Fisher approach to combining p-values from several independent research studies was accomplished by using a natural logarithmic transformation of the p-value (The natural log or log base e is equal to 2.7182818). The sum of the log base e values times -2 resulted in a chi-square value: $\chi^2 = -2 \sum (\log_e p)$, with degrees of freedom equal to $2n$, where n =the number of studies. The following example helps to illustrate Fisher's approach and the tabled chi-square values provided by Gordon, Loveland, and Cureton.

Research Study	p	Fisher $\log_e p$	Gordon et al. Tabled χ^2
1	.05	-2.996	5.991
2	.01	-4.605	9.210
3	.04	-3.219	6.438
Total		-10.820	21.639

Fisher's approach required multiplying -2 times the sum of the natural log values to calculate the chi-square value: $-2 (-10.820) = 21.639$! Gordon et al. produced a chi-square table with the chi-square values for various p-values, thus making the task easier!

Converting Various Statistics to a Common Metric

Gene Glass discovered that various test statistics reported in the research literature could be converted to the Pearson Product Moment Correlation Coefficient or r , e.g., the t , chi-square, and F -values. This provided a common metric to compare the various statistical values reported in research studies. The formula for transforming r to each statistic is presented below.

t-test

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

F-test

$$r = \sqrt{\frac{F}{F + df_{error}}}$$

Chi-square

$$r = \sqrt{\frac{\chi^2}{n}}$$

Converting Various Statistics to Effect Size Measures

Jacob Cohen expanded Gene Glass's idea to include a formula that would use an **effect size** measure (d) in the computation of the correlation statistic. The effect size formula is:

Effect-Size

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

The formula for transformation to an effect size measure d for each statistic is presented below.

t-test

$$d = \frac{2t}{\sqrt{df}}$$

F-test

$$d = \frac{2\sqrt{F}}{\sqrt{df_{error}}}$$

r

$$d = \frac{2r}{\sqrt{1-r^2}}$$

Comparison and Interpretation of Effect Size Measures

The concept behind an effect size measure, d , was to determine the amount of departure from the null hypothesis in standard units. Consequently, an effect size measure in an experimental-control group study was determined by the following formula:

$$d = \frac{\bar{Y}_{EXP} - \bar{Y}_{CTRL}}{S_{CTRL}}$$

If $d=.50$, then the experimental group scored one-half standard deviations higher than the control group. If the population standard deviation is known, it would be used instead of the control group sample standard deviation estimate. Another alternative was to use the pooled estimate of the standard deviation from both groups in the denominator.

Not all research studies however used an experimental-control group design, so the development of other effect size formulae were very important in being able to compare the results of various studies. A comparison of r and d effect size measures for a set of studies should help to better understand how the two methods are computed and interpreted. A comparison of r and d is listed below for four studies, each with a different statistic reported.

Study	Statistic	N	df	p (one-tail)	Effect size measures	
					r	d
1	t=2.70	42	40	.005	.3926	.8538
2	F=4.24	27	1,25*	.025	.3808	.8236
3	$\chi^2=3.84$	100	1	.05	.1959	.3995
4	r=.492	22	20	.01	.4920	1.1302

*In the meta-analysis program only the degree of freedom error is input. This is the second degree of freedom listed in the F ANOVA table

The calculations are straightforward using the formula for the r and d effect size measures. For example, given that $t=2.70$, the r effect size is computed as:

$$r = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{(2.7)^2}{(2.7)^2 + 40}} = \sqrt{\frac{7.29}{47.29}} = \sqrt{.1542} = .3926$$

The d effect size is computed as:

$$d = \frac{2t}{\sqrt{df}} = \frac{2(2.7)}{\sqrt{40}} = \frac{5.4}{6.3245} = .8538$$

The calculation of the r effect size measure for the chi-square value is straightforward; however, the resulting r effect size value should be used in the formula for computing the corresponding d effect size measure. Given chi-square=3.84 with a sample size of 100, the r effect size measure is .1959. This r effect size value is used in the d effect size formula to obtain the value of .3995.

Since the various statistics from the different research studies are now on a common metric, they can be compared. Study 4 had the highest r effect size measure, followed by studies 1, 2, and 3, respectively. The corresponding d effect size measures also indicate the same order, but help our interpretation by indicating how much the dependent variable (Y) changed for unit change in the independent variable (X). Notice this interpretation is directly related to the use of the correlation coefficient as an effect size measure. If we had used the experimental-group effect size measure, our interpretation would be how much the experimental group increased on average over the control group with regard to the dependent variable.

The null hypothesis always implies that the effect size is zero. If the alternative hypothesis is accepted, then the effect size departs from zero and leads to a standardized interpretation. Jacob Cohen in 1977 offered basic guidelines for interpreting the magnitude of the d effect size measure ($d=.2=\text{small}$, $d=.5=\text{medium}$, and $d=.8=\text{large}$) and r effect size measure ($r=.10=\text{small}$, $r=.30=\text{medium}$, $r=.50=\text{large}$); however, any interpretation of an effect size measure is relative. Knowledge of the professional discipline and the distribution of effect size measures in similar research studies provide reference points for interpretation. Sometimes, no such distributions of effect size estimates exist, so no standard reference point is available. In this instance, the computation of effect size estimates in several hundred studies can provide the necessary distribution and reference point. Summary statistics in the stem and leaf procedure will facilitate the interpretation of the distribution of d and r effect size measures. Given the standard deviation unit interpretation, it is important to determine whether a one-half, one-third, or one-quarter standard deviation improvement is due to some type of intervention that implies a meaningful effect size for the interpretation of the research study outcome.

A basic approach in conducting a meta-analysis across several related research studies is to combine the results and determine if there was an *overall* significant effect. The use of p-values from research studies (see Fisher, Jones, and Fiske, and Gordon et al.) readily lends itself to a summary chi-square value. The chi-square value indicates the significance of the combined research study results. The meta-analysis formula for combining the p-values from individual studies is $\chi^2=-2 \sum (\log_p)$. The combined chi-square value reported earlier was 21.6396. This chi-square value is tested for significance using 2 n degrees of freedom ($df=6$) to determine the overall effect across the different research studies. The tabled chi-square value at $p<0.01$ for 6 degrees of freedom is 16.812. The combined chi-square of 21.6396 exceeds this tabled value and therefore indicates an overall significant effect across the three research studies. In the case of using an r effect size measure,

the correlation coefficients converted from the various statistics are simply averaged, using the formula:

$$\bar{r} = \frac{\sum r}{n}$$

In the previous example, the overall r effect size is:

$$\bar{r} = \frac{\sum r}{n} = \frac{(.3926 + .3808 + .1959 + .4920)}{4} = \frac{1.4613}{4} = .3653$$

In the case of using a d effect size measure, the individual values are also averaged:

$$\bar{d} = \frac{\sum d}{n} = \frac{(.8538 + .8236 + .3995 + 1.1302)}{4} = \frac{3.2071}{4} = .8018$$

Sample Size Considerations in Meta-Analysis

An important concern in combining studies using meta-analytic techniques is the influence that different sample sizes may have on the overall interpretation. Some studies may be based on small sample sizes whereas others may be based on larger sample sizes. Since the correlation coefficient is not a function of sample size, procedures were developed to take into account the different sample sizes from the research studies when averaging the d effect size estimates. L. Hedges, as well as, R. Rosenthal and D. Rubin separately developed a formula in 1982 for calculating an *unbiased estimate* of the average d effect size. The formula was:

$$\bar{d} = \frac{\sum wd}{\sum w}$$

with w calculated as (N =total sample size):

$$w = \frac{2N}{8 + d^2}$$

The results of applying this unbiased, weighted approach to the previous example are listed below:

Study	N	d	w
1	42	.8538	9.623
2	27	.8236	6.222
3	100	.3995	24.511
4	22	1.1302	4.743

The calculation of the unbiased, weighted average effect estimate is:

$$\begin{aligned}\bar{d} &= \frac{\sum wd}{\sum w} \\ &= \frac{(9.623)(.8538) + (6.222)(.8236) + (24.5111)(.3995) + (4.743)(1.1302)}{9.623 + 6.222 + 24.511 + 4.743} \\ &= \frac{28.493}{45.099} = .6318\end{aligned}$$

The *unbiased* weighted average d effect size measure of .6318 is then compared to the *biased* weighted average d effect size measure of .8018 reported earlier. The amount of bias, or $.8018 - .6318$, is .17, which is due to the research studies having very different sample sizes. In practice, we would report the unbiased, weighted d effect size.

Meta-analysis is an objective quantitative method for combining the results of several independent research studies. One approach to combining research findings uses the log of p-values. The overall significant effect using p-values is indicated by a chi-square value. The experimental-control group effect size estimate is determined by subtracting the two group means and dividing by the standard deviation of the control group. The effect size indicates the departure from the null hypothesis in standard units. Other approaches to combining research findings use r and d effect size estimates, which involve the transformation of several common statistics. The overall significant effect from several studies using transformed statistics is obtained by averaging either the r or d effect size measures. Meta-analysis compares the relative importance of findings in several research studies by interpreting effect size measures, which are on a common metric. The overall d effect size measure can be weighted by sample size to compute an unbiased, average d effect size measure.

META-ANALYSIS R Programs

The **Meta-Analysis** program enters the p-values for each research study. Next, the chi-square values are computed given the p-values. The Fisher ln(p) and the Gordon et.al. values are then printed. The overall chi-square, degrees of freedom, and p-value are printed. A statistically significant chi-square indicates that the combined studies overall had a significant effect.

The **Effect Size** program enters for each study the following information: sample size, degree of freedom, p-value, and statistic. Each study can have a different statistic. The different statistics are converted to r and d effect size values. Next, the sample weight value is computed and used to calculate the unbiased effect size measure and sample size bias. The entire set of inputted values along with the r , d , and w values are printed. The average r and d effect size measures for the set of studies is printed along with the unbiased effect size and sample size bias.

Meta-Analysis Program Output

Fisher ln(p) versus Gordon et. al Chi-square

Study	p	Fisher ln(p)	Gordon et al
1	0.05	-2.996	5.991
2	0.01	-4.605	9.210
3	0.04	-3.219	6.438

Chi-square = 21.64 df = 6 p = 0.00141

Effect Size Program Output

Effect Size r and d

Type	Statistic	N	df	p	r	d	w
t	2.7	42	40	0.005	0.393	0.855	9.621
F	4.24	27	25	0.025	0.381	0.824	6.222
Chisq	3.84	100	1	0.05	0.196	0.4	24.51
r	0.492	22	20	0.01	0.492	1.13	4.743

Effect size(r) = 0.366 Effect size(d) = 0.802

Unbiased effect size(d) = 0.632

Sample Size bias = 0.17

Note: For F values, **only** use the degrees of freedom error or denominator degree of freedom in the dialog box.

Meta-Analysis Exercises

- Run the Meta-Analysis program for the p-values from the research studies below.

Record the corresponding Fisher log base e values, total, and overall chi-square, df, and p.

Research Study	p	Fisher log p
1	.05	
2	.001	
3	.20	
4	.01	
5	.025	
Total		

Chi-Square=_____, df=_____, p=_____

- Compute the chi-square value (-2 times the sum of $\log_e p$ values) with 2n degrees of freedom. Note: n=number of studies.

$$\chi^2 = -2 \sum (\log_e p): \underline{\hspace{10cm}}$$

$$df = 2n: \underline{\hspace{10cm}}$$

- b. Compare the chi-square above to the tabled chi-square value in Table A4 in Appendix (use .05 level of significance). What would you conclude?
-
-
-

2. Run the Effect Size program to compute r and d effect size estimates for the statistical values reported in the research studies below. Record the values in the table.

Study	Statistic	N	df	p (one-tail)	Effect size measures	
					r	d
1	$t = 2.617$	122	120	.005		
2	$F = 4.000$	62	60	.025		
3	$\chi^2 = 6.635$	50	1	.01		
4	$r = .296$	32	30	.05		

- a. What is the overall average r effect size? _____
- b. What is the overall average d effect size? _____
- c. What would you conclude about the research findings from these results?
-
-
-

3. Run the META-ANALYSIS program again using the p-values from Exercise 2. Record the chi-square value and degrees of freedom. Select a tabled chi-square value for $p < .05$.

$$\chi^2 = -2 \sum (\log_e p): \underline{\hspace{10cm}} \quad \text{Tabled } \chi^2 = \underline{\hspace{10cm}}$$

$$df = 2n: \underline{\hspace{10cm}}$$

- a. What would you conclude about the research findings using the p-value approach?
-
-

- b. Compare the chi-square, r , and d effect size results. What would you conclude?
-
-

4. Run the Effect Size program using the sample sizes from the research studies in Exercise 2. Record the d effect size measures computed in Exercise 2 and the weight values in the table below:

Study	N	d	w	wd
1	122			
2	62			
3	50			
4	32			
			$\Sigma w =$ _____	$\Sigma wd =$ _____

- a. Compute the unbiased, average d effect size using the formula:

$$\bar{d} = \frac{\sum wd}{\sum w}$$

Unbiased Effect Size = _____

- b. Compare the bias effect size (d) in Exercise 2 with the unbiased effect size (\bar{d}) above. How much bias in the overall effect size is due to the research studies having different sample sizes? Note: Overall Bias = (Bias Effect Size – Unbiased Effect Size) = $d - \bar{d}$.
-
-

Statistical Versus Practical Significance

Statistical tests for research questions involve tests of null hypotheses for different types of statistics. The statistical tests were the chi-square, z-test, t -test, analysis of variance, correlation, and linear regression. The outcomes of the statistical tests were to either retain the null hypothesis or reject the null hypothesis in favor of an alternative hypothesis based on the significance of the statistic computed. TYPE I and TYPE II errors were illustrated to better understand the nature of falsely rejecting the null hypothesis or falsely retaining the null hypothesis at a given level of significance for the sample statistic. The level of significance or p-value that we choose, i.e., .05 or .01, to test our null hypothesis has come under scrutiny due to the nature of statistical significance testing.

Researchers have criticized significance testing because it can be manipulated to achieve the desired outcome, namely, a significant finding. This can be illustrated by presenting different research outcomes based on only changing the p-value

selected for the research study. The research study involves fifth-grade boys and girls who took the Texas Assessment of Academic Skills (TAAS) test. The study was interested in testing whether fifth-grade boys on average scored statistically significantly higher than girls on the TAAS test (a directional or one-tailed test of the null hypothesis). The researcher took a random sample of 31 fifth-grade boys and 31 fifth-grade girls and gave them the TAAS test under standard administration conditions. An independent *t*-test was selected to test for mean differences between the groups in the population at the .01 level of significance with 60 degrees of freedom ($df=N-2$). The resultant sample values were:

Group	N	Mean	Standard deviation	t
Boys	31	85	10	
Girls	31	80	10	1.968

The researcher computed the *t*-value as follows:

$$t = \frac{85 - 80}{\sqrt{\frac{30(100) + 30(100)}{31 + 31 - 2} \left(\frac{1}{31} + \frac{1}{31} \right)}} = \frac{5}{2.54} = 1.968$$

The tabled *t*-value that was selected for determining the research study outcome (based on a directional, one-tailed test, with 60 degrees of freedom at the .01 level of significance) was $t=2.39$. Since the computed $t=1.968$ was not greater than the tabled *t*-value of 2.66 at the .01 level of significance, the researcher would *retain* the null hypothesis. However, if the researcher had selected a .05 level of significance, the tabled *t*-value would equal 1.67, and the researcher would *reject* the null hypothesis in favor of the alternative hypothesis. The two possible outcomes in the research study are due solely to the different levels of significance a researcher could choose for the statistical test. This points out why significance testing has been criticized, namely the researcher can have statistically significant research findings by simply changing the *p*-value.

Researchers could also manipulate whether statistically significant results are obtained from a research study by using a *one-tailed test* rather than a *two-tailed test*. In the previous example, a two-tailed test of significance would have resulted in a tabled $t=2.66$ at the .01 level of significance or a tabled $t=2.00$ at the .05 level of significance. If the researcher had chosen a two-tailed test rather than a one-tailed test, the null hypothesis would have been rejected at either level of significance or *p*-value. This illustrates how changing the directional nature of the hypothesis (one-tailed versus two-tailed test) can result in statistically significant findings.

Researchers can also increase the *sample size*, hence degrees of freedom, and achieve statistically significant research results. If we increase our sample sizes to 100 boys and 100 girls, we enter the *t*-table with infinity degrees of freedom. The resultant tabled *t*-values, given a one-tailed test, would be 1.645 at a .05 level of significance or 2.326 at a .01 level of significance. An examination of the *t*-table further indicates that the tabled *t*-values are larger for smaller degrees of freedom

(smaller sample sizes). The bottom row indicates tabled t-values that are the same as corresponding z-values in the normal distribution given larger sample sizes. This illustrates how increasing the sample size (degrees of freedom greater than 120) can yield a lower tabled t-value for making comparisons to the computed t-value in determining whether the results are statistically significant.

When significance testing, the researcher obtains a sample statistic or “point estimate” of the population parameter. The researcher could compute *confidence intervals* around the sample statistic thereby providing an additional interpretation of the statistical results. The confidence interval width provides valuable information about capturing the population parameter beyond the statistical significance of a “point estimate” of the population value. If the 95% confidence interval for a sample mean ranged from 2.50 to 3.00, then we could conclude with 95% confidence that the interval contained the population mean. Each time we take a random sample of data, the confidence interval would change. If we took all possible samples and computed their confidence intervals, then 95% of the intervals would contain the population mean and 5% would not; therefore, one should not report that the probability is .95 that the interval contains the population mean. Unfortunately, many researchers either do not report confidence intervals and/or misreport them.

Replication of research findings provide support for results obtained. These methods help to address the practical importance of the research study findings. The most meaningful technique would be to *replicate* the study and/or *extend the research* based on earlier findings. This provides the best evidence of research findings or outcomes. Researchers could also use their sample data from a single study and *cross-validate*, *jackknife*, or *bootstrap* the results. In some cases, a researcher might synthesis several findings from research studies by conducting a *meta-analysis*. Most researchers however do not take the time to replicate their study, cross-validate, jackknife, bootstrap, or conduct a meta-analysis. These methods are well known, but not available in most mainstream statistical packages and therefore not readily available to researchers.

Another important consideration above and beyond the significance of a statistical test is the *effect size* or magnitude of difference reported. The interpretation of the effect size can directly indicate whether the statistically significant results are of any practical importance. An example will better illustrate the practical importance of research findings based on an effect size. The previous research study reported a five point average difference between boys and girls in the population on the TAAS test. Is this average five-point difference (approximately getting two test questions correct or incorrect) of practical importance? If we retain the null hypothesis of no difference in the population based on our statistical test of significance, then we conclude that fifth-grade boys and girls achieve about the same. Alternatively, if we reject the null hypothesis in favor of an alternative hypothesis based on our statistical test of significance, then we conclude that fifth-grade boys scored statistically significantly higher on average than the girls at a given level of significance. What are the consequences of our decisions based on a statistical test of significance? If we retain the null hypothesis when it is really false, we make the error of not spending additional money for programs to better educate fifth-grade girls. If we reject the null hypothesis when it is really true, we make the error of spending additional money on programs

that are not needed to better educate fifth-grade girls. The effect size helps our practical understanding of the magnitude of the difference detected in a research study. The effect size however should be interpreted based upon a synthesis of findings in several other related studies. This comparison provides a frame of reference for interpreting whether the effect size value is small, medium, or large. The r and d effect size measures for the computed t -value are computed as follows:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

and

$$d = \frac{2t}{\sqrt{df}}$$

The researcher, to achieve significant findings, can manipulate the level of significance, directional nature of the test, and sample size in statistical significance testing. The confidence interval should be reported along with the statistic and p -value to aid in the interpretation of research findings. The effect size helps our practical understanding of the importance of our research results. Replication and/or the extension of a research study are the most meaningful ways to validate findings in a research study. Cross validation, bootstrap, and jackknife methods provide additional information in explaining results from a single study. Results can be statistically significant but have little practical importance.

A few final words of wisdom can be given when faced with significance testing and issues related to the practical importance of research findings. In conducting basic applied research, one asks a question, analyzes data, and answers the research question. Beyond this task, we need to be reminded of several concerns. How do our research findings relate to the research findings in other related research studies? What is the educational importance of our findings? What implications do our research findings have on practice? What recommendations can we make that might affect or modify the underlying theory? What recommendations can we make that might enhance future research efforts?

PRACTICAL R Program

The PRACTICAL program computes an independent t -test and outputs the associated values that a researcher should report. The program begins by setting the sample size, first population mean and standard deviation and the second population mean and standard deviation. A random sample of data is then created for two independent samples using the **rnorm** function. The sample means and standard deviations for the two samples are then input into the **t.test** function. The results of the independent t -test are output along with the d and r effect size measures.

PRACTICAL Program Output

Independent t-test Results

Two sample independent t-test
 Sample One Mean=51.54 Sample size=30
 Sample Two Mean=50.94 Sample One SD=10.42
 $t\text{-test}=0.23$ Sample Two SD=9.66
 $p\text{-value}=0.816$ df=58
 $r\text{ Effect}=0.031$ 95% Confidence Interval
 $=-4.588$ to 5.798
 $d\text{ Effect}=0.061$

PRACTICAL Exercises

- Run the PRACTICAL program 10 times and record the results below.

	t-value	p-value	95%CI	r effect size	d effect size
1)	_____	_____	_____	_____	_____
2)	_____	_____	_____	_____	_____
3)	_____	_____	_____	_____	_____
4)	_____	_____	_____	_____	_____
5)	_____	_____	_____	_____	_____
6)	_____	_____	_____	_____	_____
7)	_____	_____	_____	_____	_____
8)	_____	_____	_____	_____	_____
9)	_____	_____	_____	_____	_____
10)	_____	_____	_____	_____	_____

- a. What conclusions can be drawn about the statistical significance of the computed t-values if the tabled t-value = 1.671 at the .05 level of significance for a one-tailed test?
-
-

- b. How many p-values are less than the .05 level of significance? _____
 c. What percent of the confidence intervals captured the population mean difference of zero? _____

- d. What interpretation would you give for the r effect size measures for these 10 replications?
-
-

- e. What interpretation would you give for the d effect size measures for these 10 replications?
-
-

True or False Questions

Meta-Analysis

- T F a. Meta-analysis uses subjective techniques to combine research studies.
- T F b. The p-value approach combines research findings using chi-square values.
- T F c. Various statistics are converted to a common metric so research findings across studies can be quantitatively compared.
- T F d. The r effect size measure can be interpreted by using a standard reference scale.
- T F e. The d effect size measure is interpreted relative to findings from a large body of research in an academic discipline.
- T F f. When combining effect size measures, it is important to weight by sample size.
- T F g. The p-value, log(p), and chi-square approach yield similar results.
- T F h. Gene Glass is recognized as creating the term “Meta-Analysis”.

Statistical Versus Practical Significance

- T F a. Significance testing is the *only* way to know if your findings are important.
- T F b. Replication is the *least* meaningful way to determine the validity of your research findings.
- T F c. Cross-validation, jackknife, and bootstrap methods provide important information about results when analyzing a single sample of data.
- T F d. Research findings can be statistically significant, but have no practical importance to the field of study.
- T F e. Increasing the sample size can *always* make a statistical test significant at a given level of significance.
- T F f. The sample statistic, p-value, confidence interval, and effect size are recommended values that should be reported in a research study.

Glossary of Terms

Alpha level The level of statistical significance selected prior to a test for incorrectly rejecting a true null hypothesis, e.g., .05 alpha level of significance. (See Type I error)

Alternative hypothesis A statistical hypothesis that indicates a difference in population parameters. For example, the means of two populations are different, i.e., possible outcomes not stated in the null hypothesis

Analysis of variance A technique that tests whether the dependent variable means of three or more mutually exclusive groups are statistically significantly different at a specified level of significance. The F-test is a ratio of MS Between Groups divided by MS Within Groups

Bell shaped curve Describes a normal or symmetrical distribution of data in which intervals around the mean are known

Bimodal A frequency distribution of data that has two modes, i.e., two scores that occur most frequently in a set of data

Binomial distribution A probability distribution generated by taking $(a+b)$ to the n th power. Used in a binomial test to determine whether the probability of two outcomes exceed the chance level of occurrence

Binomial test A non-parametric test, which doesn't depend on any population data characteristics and measures whether a distribution of scores results in a binomial distribution (each outcome is equally likely). For example, if you tossed an unbiased coin 1,000 times it should land approximately 500 heads and 500 tails

Bootstrap An approach that samples with replacement to generate a sampling distribution of a statistic that serves as the population distribution. The mean of the bootstrap sampling distribution or bootstrap estimate is used to determine the amount of bias in the random sample

Central limit theorem A theorem that provides a mathematical basis for using the normal distribution, as a sampling distribution of a statistic for a given sample size, to test a statistical hypothesis. For example, the theorem states that a sampling distribution of means for a given sample size is (1) normally distributed, (2) the sampling distribution mean is equal to the population mean, and (3) the sampling distribution variance is equal to the variance divided by the sample size

Central tendency A concept that implies most scores fall in the middle of a symmetrical distribution with the scores spreading out evenly toward both tails of the distribution

Chi-square distribution A probability distribution or family of curves generated by the difference between observed and expected frequencies. The sampling distribution of chi-square values is used in both parametric and non-parametric tests of significance

Chi-square statistic A non-parametric test that measures the difference between the observed frequencies and expected frequencies in two or more groups

Combinations The number of ways in which different subsets of events or numbers can be selected

Conditional probability The probability of an event B is based upon the occurrence or non-occurrence of event A

Confidence interval A high and low value which forms an interval around the sample statistic that should contain the population parameter. The interval will be different depending upon the percentage used, i.e., 68, 95, or 99 %

Confidence level A percentage that indicates how certain we are that the interval around the sample statistic contains the population parameter (see alpha level)

Correlation A statistic that indicates the strength and direction of association between two sets of scores. The strength is indicated by a correlation value closer to 1.0 and the direction indicated by a \pm sign. A positive correlation indicates that both variables increase in value across the range of scores while a negative correlation indicates that one set of scores increases as the other set of scores decreases

Cross-validation An original sample of data is randomly split into two equal samples, then a sample statistic is computed using one sample of data and applied to the other sample of data

Cumulative frequency distribution A frequency distribution of raw scores that indicates successive addition of the number of events, individuals, or objects up to the total number or 100 %

Degrees of freedom The number of observations or groups minus the restrictions placed upon them. For example, if four out of five sample means are known, the one remaining unknown sample mean can be determined, hence $df=5-1=4$

Dependent t-test A statistical test of whether two sample means from the same subjects or group are significantly different. Also called a paired t-test or correlated t-test

Dichotomous population A population of data that can be divided into two mutually exclusive categories

Directional hypothesis A hypothesis that states one population parameter is greater than the other. The direction can be stated in a positive or negative direction, e.g., boys' verbal scores will be lower on average than girls' verbal scores. A one-tailed test because the region of rejection is only in one tail of the sampling distribution

Effect size Conveys the magnitude of difference in standard units between the mean of the experimental group and the mean of the control group. Used in conjunction

with sample size, alpha level, and direction of the statistical hypothesis to select a value for power

Equally likely events Events, individuals, or objects that have the same chance of being selected

Exponential function A relationship between two sets of data points that does not have a constant rate of change for a random variable X, i.e., $Y=2^X$

F-curve A positively skewed frequency distribution of F values for specific degrees of freedom

F-distribution A probability distribution or family of curves that require two degrees of freedom. The normal, t, and chi-square distributions are special cases of the F-distribution

F-test In Analysis of Variance, the test to determine if sample means are different beyond chance expectation. The F-test is the ratio of MS Between Groups divided by the MS Within Groups

Factoring A product of sequential numbers that indicate the total number of choices possible (see factorial notation)

Factorial notation Indicated as $n!$, i.e., $3!=3\times 2\times 1=6$

Finite distribution A population of data where the number of individuals, objects, or events is known, hence exact probabilities of occurrence can be computed

Frequency distribution A tabulation of data that indicates the number of times a score or value occurs

Hartley F-max test A test of whether three or more sample variances are statistically different. The largest sample variance is divided by the smallest sample variance to form an F-ratio with degrees of freedom from the two sample sizes

Heterogeneity Refers to a grouping of dissimilar individuals, objects, or events

Histogram A bar chart that indicates the frequency of numbers on the Y axis and the mutually exclusive groups or categories on the X axis. (Also, see Pie Chart)

Homogeneity Refers to a grouping of similar individuals, objects, or events

Independent t-test A statistical test of whether two independent sample means are significantly different implying that the two population means are different

Infinite population A population of data where the number of individuals, objects, or events are too numerous to count, hence exact probabilities of occurrence cannot be computed

Intercept The intercept is the point in a linear equation where the line of best fit crosses the Y-axis. The intercept is the predicted value of Y when the X variable equals zero. The value a in the linear regression equation: $Y = a + bX + e$.

Interquartile range A score that represents the distance between the first and third quartile. It indicates the range of scores in the middle 50 % of a frequency distribution

Jackknife An approach that uses a single sample of data and computes sample statistics based on different $n-1$ sample sizes

Joint probability The probability of two events occurring that is determined by multiplying the independent probability of each event

Kurtosis A measure that indicates the flatness or peakedness of the frequency distribution of scores. Leptokurtic implies a peaked distribution, mesokurtic a bell-shaped normal distribution, and platykurtic a flattened distribution of scores

Law of complements Given the probability of event A, $P(A)$, the complement is $1 - P(A)$, or the remaining probability since $P(A) + [1 - P(A)] = 1$

Leaves The numbers to the right of the vertical line in a stem-and-leaf plot

Level of significance The probability of making a Type I error (see alpha level)

Linear function An angled straight line of data points that indicate a constant rate of change for a random variable, X, i.e., $y = bX$

Linear regression (equation) A statistical technique designed to predict values of Y (dependent variable) from one or more X variables (independent predictor variables). The regression equation: $Y = a + bX + e$.

Line of best fit In linear regression, the line formed by the predicted Y values that pass through the scatterplot of X and Y values. The line indicates the best prediction that minimizes the sum of squared errors of prediction

Mean The arithmetic mean computed as the sum of a set of scores divided by the number of scores. Typically referred to as a measure of central tendency

Mean square A variance estimate computed by dividing the sum of squares by the degrees of freedom

Median The middle score in a distribution of odd-numbered scores or the midpoint in an even-numbered set of scores. Typically referred to as a measure of central tendency

Meta-analysis A statistical procedure that averages the effect sizes across several studies to determine the overall significance of a large number of research studies on the same topic

Mode The most frequently occurring score in a set of scores. It is possible to have a single modal score (unimodal), two scores that occur the most (bimodal), or even three or more scores that occur the most. Typically referred to as a measure of central tendency

Monte Carlo An approach that describes a statistical technique that simulates data and approximates probability density functions of population distributions to study the robustness and properties of statistical tests

MS between groups The sum of the squared deviations of group means around the grand mean weighted (multiplied) by the sample size of each group and divided by the number of groups minus one. Indicates whether the group means are similar or different based on how much they vary

MS within groups The sum of squared deviations of individual scores around each group mean divided by the number of scores in each group minus the number of groups. Indicates how much the scores vary within each group

Multiplication law The independent probabilities of two events can be multiplied to obtain their probability of joint occurrence, i.e., $P(J) = P(A) * P(B)$

Non-directional hypothesis A hypothesis that states two population parameters are different, rather than one population parameter is greater than the other. A two-tailed test because the region of rejection is in both tails of the sampling distribution

Normal curve A symmetric distribution of data based on a mathematical equation formulated by DeMoivre in 1733 and further developed by Carl Fredrick Gauss

Normal distribution A frequency distribution of scores that when graphed produces a symmetrical, bell-shaped distribution with skewness and kurtosis of zero. Sometimes referred to as a mesokurtic distribution

Null hypothesis A statistical hypothesis that indicates no difference in population parameters. For example, the means of two populations are equal. The null hypothesis is either retained or rejected in favor of an alternative hypothesis

Ogive A graph of the cumulative frequency distribution of data that has a characteristic S-shaped curve

One-sample t-test A statistical test of whether a sample mean is significantly different from a population mean

One-way analysis of variance A statistical test that is an extension of the independent t-test to test whether three or more independent sample means are statistically different implying that the population means are different

Outlier An extreme or influential score or data value that affects the sample statistic, e.g., sample mean

Parameter(s) Population values or characteristics that are estimated by sample statistics, e.g., population mean or population correlation

Parametric statistics Parametric or inferential statistics are based upon being able to randomly draw a sample from a well defined population, estimate the sample statistic, and make an inference about the population parameter. For example, the sample mean is an estimate of the population mean

Permutations A technique used to determine the number of different ways individuals, objects, or events can be ordered

Pie chart A circle with portions or parts of the circle indicated for each mutually exclusive group or category. (Also, see Histogram)

Population A set of individuals or scores that are well defined and share some characteristic in common. Typically, population data is randomly sampled and sample statistics computed to estimate the population values because the population is typically too large to measure all the data. (See Parametric Statistics)

Power The probability of rejecting the null hypothesis when it is false. The expression, $1 - \beta$, is used to indicate the level of power. Values of .80 are typically selected for power; power is a function of sample size, alpha level, effect size, and directional nature of the statistical hypothesis (one-tailed or two-tailed test)

Probability The ratio of the number of favorable outcomes to the total possible number of outcomes

Properties of estimators Important characteristics we want sample statistics to possess as estimates of population parameters, i.e., unbiased, consistent, efficient, and sufficient

Pseudo random numbers Numerical values typically generated by a random number generator on a computer, but not truly independent or unbiased because they will eventually correlate and repeat

Quartile A score that divides a set of data into four equal divisions, i.e., first quartile is a score that separates the bottom 25 % of the data in a frequency distribution from the other data values

Random assignment The random process of assigning individuals, objects, or events to a group, i.e., random assignment of individuals to experimental and control groups

Random numbers Independent, unbiased numerical values that have an equally likely chance of being selected

Random number generator A mathematical algorithm in a software program that is run on a computer to generate pseudo random numbers

Random sample A sample of data from a well-defined population where every individual, object, or event has an equally likely chance of being selected

Random sampling The process of selecting individuals, objects, or events from a well-defined population in which all members have an equal and independent chance of being selected. Not the same as random assignment

Range A score that indicates the distance between the highest and lowest data value in a set of data

Region of rejection The area under a sampling distribution where sample statistics fall that is highly improbable if the null hypothesis is true

Regression weight In regression analysis, the regression coefficient or slope of the line of best fit that passes through the predicted Y values. The value b in the linear regression equation: $Y = a + bX + e$. A weight computed by the least squares method of minimizing the sum of squared errors of prediction

Repeated measures ANOVA A statistical procedure in which subjects are measured two or more times and the total variation of scores is partitioned into three components: (1) variation among subjects, (2) variation among occasions (time), and (3) residual variation

Sample A random selection of individuals, objects, or events from a well-defined population of data

Sample error The difference between a sample statistic and the population parameter

Sampling error The error in using a sample statistic as an estimate of a population parameter

Sampling distribution A probability frequency distribution of a sample statistic formed for all possible random samples of a given sample size. Examples of sampling distributions include: (1) sampling distribution of means, (2) t-distribution, (3) chi-square distribution, and (4) F-distribution

Sampling with replacement Each randomly sampled data point is returned to the population before another data point is randomly sampled, therefore it is possible for a data point to be selected more than once

Sampling without replacement Each randomly sampled data point is not returned to the population before another data point is randomly sampled, therefore each data point is uniquely drawn and can not be selected again

Scheffe post-hoc test A type of post-hoc “t-test” for conducting multiple comparisons of group mean differences after an analysis of variance F-test

Skewness A measure of deviation from symmetry in a frequency distribution of scores. Negative skew indicates a distribution with more scores above the mean. Positive skew indicates a distribution with more scores below the mean

Slope The amount of change in Y that corresponds to a change of one unit in X. (See Regression Weight)

Standard deviation The square root of the average squared deviations of scores around the mean. A measure of how much the individual scores deviate from the mean

Standard errors of statistic The standard deviation of the sampling distribution of the statistic that indicates the amount of error in estimating the population parameter

Standard score A score computed by taking the deviation of the raw score from the group mean divided by the group standard deviation, i.e., z-score

Statistic A sample value that estimates a population parameter

Stem and leaf A graphical display that illustrates the shape of a distribution of scores

Sum of squared deviations The deviations of each score from the group mean that is squared and then summed for all scores

Symmetric distribution A sampling distribution or frequency distribution of scores that is the same on either side of the median value. The normal distribution is an example of a symmetric distribution

t-distribution A probability distribution or family of t-curves for different degrees of freedom that is used to determine whether an obtained t value between two sample means is statistically significant at a specified alpha level

Tchebysheff inequality theorem A theorem that indicates the percentage of data between intervals around the mean regardless of the shape of the frequency distribution of data

Type I error The rejection of the null hypothesis of no difference in population parameters when it is true, i.e., the probability that a null hypothesis would be rejected in favor of an alternative hypothesis. The probability is set by selection of an alpha level (see Alpha level). If the alpha level is set at .05, then 5% of the time a true null hypothesis would be incorrectly rejected in favor of the alternative hypothesis. The symbol, α , is used to refer to this type of error

Type II error The retention of a null hypothesis of no difference in population parameters when it is false, i.e., the probability that we failed to reject the null hypothesis in favor of an alternative hypothesis. The symbol, β , is used to refer to this type of error

Uniform distribution A rectangular distribution of scores that are evenly distributed in the range of possible values

Unimodal distribution A symmetrical distribution with a single mode

Variance A positive value that measures how scores vary around a group mean. If all scores are the same, then the variance is zero. Calculated as the sum of squared deviations around the group mean divided by the number of scores

z-score Sometimes called a standard score. A frequency distribution of raw scores that have been standardized to a mean of zero and a standard deviation of one.

A z score indicates the direction and degree to which a score deviates from the mean of a distribution of scores

z-test A statistical test for the significant difference in independent or dependent population proportions

Appendix

Statistical Tables

Table A1 Areas under the normal curve (z-scores)

Second decimal place in z										
<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4373	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4826	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.49997									
5.0	.499997									

Table A2 Distribution of t for given probability levels

Level of significance for one-tailed test						
	.10	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test					
<i>df</i>	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.992
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Table A3 Distribution of r for given probability levels

df	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
df	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.540	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.576	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.486
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.303
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

Table A4 Distribution of Chi-square for given probability levels

Probability	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
df	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.00016	.00663	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620

26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703
32	16.362	17.783	20.072	22.271	25.148	27.373	31.336	35.665	38.466	42.585	46.194	50.487	53.486	62.487
34	17.789	19.275	21.664	23.952	26.938	29.242	33.336	37.795	40.676	44.903	48.602	52.995	56.061	65.247
36	19.233	20.783	23.269	25.643	28.735	31.115	35.336	39.922	42.879	47.212	50.999	55.489	58.619	67.985
38	20.691	22.304	24.884	27.343	30.537	32.992	37.335	42.045	45.076	49.513	53.384	57.969	61.162	70.703
40	22.164	23.838	26.509	29.051	32.345	34.872	39.335	44.165	47.269	51.805	55.759	60.436	63.691	73.402
42	23.650	25.383	28.144	30.765	34.147	36.755	41.335	46.282	49.456	54.090	58.124	62.892	66.206	76.084
44	25.148	26.939	29.787	32.487	35.974	38.641	43.335	48.396	51.639	56.369	60.481	65.337	68.710	78.750
46	26.657	28.504	31.439	34.215	37.795	40.529	45.335	50.507	53.818	58.641	62.830	67.771	71.201	81.400
48	28.177	30.080	33.098	35.949	39.621	42.420	47.335	52.616	55.993	60.907	65.171	70.197	73.683	84.037
50	29.707	31.664	34.764	37.689	41.449	44.313	49.335	54.723	58.164	63.167	67.505	72.613	76.154	86.661
52	31.246	33.256	36.437	39.433	43.281	46.209	51.335	56.827	60.332	65.422	69.832	75.021	78.616	89.272
54	32.793	34.856	38.116	41.183	45.117	48.106	53.335	58.930	62.496	67.673	72.153	77.422	81.069	91.872
56	34.350	36.464	39.801	42.937	46.955	50.005	55.335	61.031	64.658	69.919	74.468	79.815	83.513	94.461
58	35.913	38.078	41.492	44.696	48.797	51.906	57.335	63.129	66.816	72.160	76.778	82.201	85.950	97.039
60	37.485	39.699	43.188	46.459	50.641	53.809	59.335	65.227	68.972	74.397	79.082	84.580	88.379	99.607
62	39.063	41.327	44.889	48.226	52.487	55.714	61.335	67.322	71.125	76.630	81.381	86.953	90.802	102.166
64	40.649	42.960	46.595	49.996	54.336	57.620	63.335	69.416	73.276	78.860	83.675	89.320	93.217	104.716
66	42.240	44.599	48.305	51.770	56.188	59.527	65.335	71.508	75.424	81.085	85.965	91.681	95.626	107.258
68	43.838	46.244	50.020	53.548	58.042	61.436	67.335	73.600	77.571	83.308	88.250	94.037	98.028	109.791
70	45.442	47.893	51.739	55.329	59.898	63.346	69.335	75.689	79.715	85.527	90.531	96.388	100.425	112.317

Note: For larger values of df, the expression $\sqrt{\chi^2} - \sqrt{2df - 1}$ may be used as a normal deviate with unit variance, remembering that the probability for X^2 corresponds with that of a single tail of the normal curve

Table A5 The F-distribution for given probability levels (.05 Level)

df_1	df_2	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16.4	199.5	215.7	224.6	230.2	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.41	19.43	19.45	19.46	19.47	19.48	19.49	19.50		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.64	8.62	8.59	8.57	8.55	8.53	
4	7.71	6.94	6.59	6.39	6.26	6.15	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.69	5.66	5.63	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.46	4.43	4.40	4.36	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.67	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.71	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.07	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.11	2.06	2.01	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.76	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.73	
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.62	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	

Table A6 The distribution of F for given probability Levels (.01 Level)

df_1	df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	
2	98.5	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.49	99.49	99.50	
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.25	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.22	26.22	26.13	
4	21.20	18.00	16.69	15.98	5.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	
20	8.10	5.85	5.00	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	
22	7.95	5.72	4.82	4.31	3.9	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38	
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00	

Table A7 Distribution of Hartley F for given probability levels

a	$k = \text{number of variances}$	$df = n - l$
.05	2	3
.05	9.60	15.5
.01	23.2	37.
.05	7.15	10.8
.01	14.9	22.
.05	5.82	8.38
.01	11.1	15.5
.05	4.99	6.94
.01	8.89	12.1
.05	4.43	6.00
.01	7.50	9.9
.05	4.03	5.34
.01	6.54	8.5
.05	3.72	4.85
.01	5.85	7.4
.05	3.28	4.16
.01	4.91	6.1
.05	2.86	3.54
.01	4.07	4.9
.05	2.46	2.95
.01	3.32	3.8
.05	2.07	2.40
.01	2.63	3.0
.05	1.67	1.85
.01	1.96	2.2
.05	1.00	1.00
.01	1.00	1.00
6	4	5
7	25.2	29.5
8	59.	69.
9	16.3	18.7
10	33.	38.
11	12.1	13.7
12	19.1	22.
13	8.44	9.70
14	14.5	16.5
15	7.18	8.12
16	11.7	13.2
17	6.31	7.11
18	9.9	9.9
19	11.1	12.1
20	5.67	6.34
21	7.4	8.6
22	4.16	4.79
23	6.1	6.9
24	3.54	4.01
25	5.5	6.0
26	3.29	3.54
27	4.01	4.37
28	2.3	2.3
29	4.6	4.9
30	2.61	2.78
31	3.0	3.3
32	1.96	2.04
33	2.2	2.3
34	1.00	1.00
35	1.00	1.00
36	1.00	1.00
37	1.00	1.00
38	29.5	33.6
39	79.	89.
40	20.8	22.9
41	42.	46.
42	13.7	15.
43	12.1	16.3
44	25.	27.
45	10.8	11.8
46	18.4	20.
47	9.03	9.78
48	14.5	15.8
49	7.80	8.41
50	12.1	13.1
51	6.92	7.42
52	10.4	11.1
53	5.30	5.72
54	7.6	8.2
55	4.01	4.37
56	6.0	6.4
57	3.54	3.76
58	4.6	4.9
59	2.3	2.4
60	1.00	1.00
61	1.00	1.00
62	1.00	1.00
63	1.00	1.00
64	33.6	37.5
65	97.	106.
66	24.7	26.5
67	50.	54.
68	16.3	17.5
69	10.5	11.1
70	27.	30.
71	12.7	13.5
72	22.	23.
73	10.5	11.1
74	32.	34.
75	11.7	11.7
76	17.9	17.9
77	18.9	18.9
78	16.0	16.0
79	21.	21.
80	9.01	9.01
81	10.7	10.7
82	13.4	13.4
83	16.6	16.6
84	9.01	9.34
85	12.9	13.4
86	7.00	7.25
87	9.9	10.2
88	5.59	5.77
89	10.2	10.6
90	5.40	5.40
91	7.3	7.5
92	4.24	4.37
93	5.1	5.3
94	4.95	5.19
95	6.7	7.1
96	3.76	3.94
97	6.0	6.4
98	4.01	4.10
99	2.17	2.22
100	2.4	2.5
101	1.00	1.00
102	1.00	1.00
103	1.00	1.00
104	1.00	1.00
105	1.00	1.00
106	1.00	1.00
107	1.00	1.00
108	1.00	1.00
109	1.00	1.00
110	1.00	1.00
111	1.00	1.00
112	1.00	1.00
113	1.00	1.00
114	1.00	1.00
115	1.00	1.00
116	1.00	1.00
117	1.00	1.00
118	1.00	1.00
119	1.00	1.00
120	1.00	1.00

Author Index

B

Bargmann, Rolf, E., 217
Bashaw, W.L., 217
Bernoulli, James, 110
Bock, Darrell, R., 217
Bottenberg, Robert A., 217

C

Cohen, Jacob, 246, 247, 249
Cureton, Edward, E., 246

D

Darwin, Charles, 207
DeMoivre, Abraham, 110, 265
Derenzo, Stephen, E., 112
Draper, Norman, R., 217

F

Findley, Warren, B., 217
Fisher, Ronald, A., Sir, 126, 128, 136, 185,
195, 246
Fiske, Donald, W., 246

G

Galton, Francis, Sir, 207
Gauss, Carl Fredrich, 110, 265
Glass, Gene, 246, 247, 259
Gordon, Mordecai, H., 246, 249, 251, 252
Gossett, William, S., 120, 185, 186, 195
Graybill, Franklin, F.A., 217

H

Hartley, H.O., 129
Hedges, L., 250

J

Jones, Lyle, V., 246

L

Loveland, Edward, H., 246

P

Pearson, Egon, S., 246
Pearson, Karl, 114, 167, 207, 210,
211, 246

R

Rand, McNally, 48
Rosenthal, R., 250
Rubin, D., 250

S

Scheffe, Henry, 198
Smith, Harry, Jr., 217
Student, 120, 185

T

Thurstone, L.L., 207

U

UNIVAC, 48

W

Ward, Joe, H., 217

Winer, Ben, J., 217

Winkler, Henry, 129

Subject Index

A

Addition law of probability
ADDITION R program (*see* ADDITION R program)
description, 23
exercises, 25–26
law of complements, 23–24
odds of events, 23
program output, 24–25
theory of probability, 23
true/false questions, 40

ADDITION R program
definition, 24
DiceFreq object, 24
DiceSum, 24
exercises, 25–26
multiplication law of probability, 26
odd and even numbers, 25–26
output, 24–25

Alpha level, 261

Alternative hypothesis, 261

Analysis of variance (ANOVA)
technique
definition, 261
description, 195
F tables, 196
heterogeneous, 195
homogeneous, 195
one-way ANOVA (*see* One-way ANOVA)
sample mean differences, 195
SET A and SET B, 195

ANOVA technique. *See* Analysis of variance (ANOVA) technique

Areas, normal curve (z-scores), 269–270

B

Bell shaped curve, 261

BELL-SHAPED CURVE R program
description, 90
Kvals, 90
normal distribution exercises, 91–93
output, 91
population types, 90

Binomial distribution, 261

BINOMIAL R program
exercises, 108–109
numReplications and *numTrials*
sample, 107
output, 108

Binomial test, 261

Bootstrap procedure
definition, 261
description, 237–238
estimate, 239
program, 240–241
“pseudo” population, 239

C

Central limit theorem
CENTRAL R program (*see* CENTRAL R program)
definition, 261
description, 93–94
exercises, 99–101
homogeneous and heterogeneous population, 94
mathematical function, 94
non-normal populations, 94
output, CENTRAL R program, 95–99
sampling distribution, 94–95

- Central limit theorem (*cont.*)
 statistical analysis, 94
 true/false questions, 102
- CENTRAL R program
 exercises, 99–101
 infinite number, 95
 output, 95–99
 user-defined variables, 95
- Central tendency
 constant data set, 74
 data transformations, 75
 definition, 262
 description, 73
 mean and median score, 73
 MEAN-MEDIAN exercises, 75–77
 MEAN-MEDIAN R program (*see*
 MEAN-MEDIAN R program)
 output, MEAN-MEDIAN program, 74–75
- Chi-square distribution
 CHISQUARE R program (*see*
 CHISQUARE R program)
 definition, 262
 numerator and denominator, 114–115
 sample size, 114
 statistic, 115
 z-score, 114
- CHISQUARE R program
 barplot function, 115
 exercises, 117–120
 output, 116–117
 sample sizes, 115
- Chi-square statistics, 262
- Chi-square test
 categorical variables, 167
 computation, 168–169
 CROSSTAB R program, 170–173
 cross-tabulation, data, 167, 168
 expected cell frequencies, 168
 independence, 170
 interpretation, 169
 men and women, expected cell values, 168
 statistical tests, 167
 superintendent decides, 168
 two-by-two table, 169, 170
- COMBINATION R program
 description, 63
 exercises, 65–66
 factor and table functions, 63
 FreqPopTable, 63
 FreqPop vector, 63
 hist function, 64
 output, 64–65
 round and invisible function, 63
- Combinations and permutations
 definition, 262
 description, 34
 exercises, 38–39
 factorial notation, 35
 formula, 36
 fundamental rules, 34–35
 permutations, 35
 possible combinations, 36
 program output, 38
 random sample, 34
 relative frequencies, 36–37
 R program, 37–38
 true/false questions, 41
- Comprehensive R Archive Network (CRAN), 1
- Conditional probability
 CONDITIONAL R program (*see*
 CONDITIONAL R program)
 definition, 262
 description, 29
 equally likely events, 30
 event A and B, 29
 exercises, 32–34
 marginal probabilities, 30
 program output, 32
 true/false questions, 41
- CONDITIONAL R program
 description, 31
 dimnames set, 31
 exercises, 32–34
 output, 32
 print.char.matrix, 31
 rep function, 31
- Confidence interval, 262
- Confidence level, 262
- CONFIDENCE R Program
 output, 142–143
 population mean, 142
 sample sizes and confidence levels, 142
- Correlation
 definition, 262
 Pearson correlation (*see* Pearson
 correlation)
- CORRELATION R Program
 cor function, 212
 exercises, 213–216
 output, 212–213
 rho value, 212
- CRAN. *See* Comprehensive R Archive
 Network (CRAN)
- CROSSTAB R program, 170–173
- Cross-validation
 bootstrap procedure, 237–241
 definition, 228, 262
 programs, 228–230

- regression weight and R-square value, 230–232
- Cumulative frequency distribution, 262
- D**
- Degrees of freedom, 262
- Dependent t-Test
- calculation, 189
 - definition, 262
 - denominator, 188
 - description, 188
 - motion picture, 188
 - psychologist, 188
 - repeated measure sample data, 189
 - STUDENT R program, 190–193
 - students' attitudes, 188
 - t-values, 189
- DEVIATION R program
- exercises, 139–140
 - output, 138
 - replication, 137, 138
 - sample standard deviation, 138
- Dichotomous population, 262
- Directional hypothesis, 262
- Dispersion
- description, 77
 - DISPERSION R program (*see* DISPERSION R program)
 - normal distribution, 77
 - range, variance, 77
 - SD, 77–78
 - SS, 77
 - true/false questions, 102
- DISPERSION R program
- description, 78
 - exercises, 79–81
 - output, 78–79
- E**
- Effect size, 262–263
- Equally likely events, 263
- Exponential function, 263
- F**
- Factorial notation, 263
- Factoring, 263
- F-curve, 263
- F-distribution
- chi-square distribution and t-distribution, 127–128
 - definition, 263
- degree of freedom pairs, F-curve, 127
- exercises, 131–132
- F-curve and F-ratio, 126
- F-DISTRIBUTION R programs (*see* F-DISTRIBUTION R programs)
- independent variances, 128–129
- F-DISTRIBUTION R programs
- F-curve program output, 130
 - F-ratio program output, 130, 131
- Finite and infinite probability
- description, 11
 - exercises, 14–17
 - output, 13–14
- PROBABILITY R program, 12–13
- relative frequency definition, 12
 - relative frequency, event, 12
 - stabilization, 12
 - trials, 11
 - true/false questions, 39–40
- Finite distribution, 263
- Fisher *vs.* Gordon chi-square approaches, 246
- Frequency distributions
- definition, 263
 - histograms and ogives (*see* Histograms and ogives)
 - population (*see* Population distributions)
 - stem and leaf graph (*see* Stem and leaf graph)
- FREQUENCY R program
- description, 56
 - histogram and ogive exercises, 58–62
 - output, 57–58
 - PlotHeight, 57
 - RelFreq and CumRelFreq vector, 56
 - TableData matrix, 56–57
- F-test
- ANOVA technique (*see* Analysis of variance (ANOVA) technique)
 - definition, 263
 - multiple comparison tests, 198–199
 - ONEWAY program, 204
 - repeated measures analysis, variance, 199–200, 205
 - REPEATED program, 204
 - SCHEFFE program, 204
 - variance R programs (*see* Variance R programs)
- G**
- Graduate record exam (GRE) score, 223
- GRE score. *See* Graduate record exam (GRE) score

H

Hartley F-max test, 263
 Histograms and ogives
 cumulative frequency distribution, 55
 description, 55
 exercises, 58–62
 FREQUENCY R program (*see* FREQUENCY R program)
 output, FREQUENCY program, 57–58
 S-shaped, 55–56
 true/false questions, 71
 Hypothesis testing
 confidence intervals
 CONFIDENCE R program, 142–144
 description, 140
 population mean, 141
 population standard deviation, 141
 range of values, 141
 standard error of a statistic, 140
 DEVIATION R program (*see* DEVIATION R program)
 mean and median, 136
 population standard deviation, 136–137
 properties, estimators, 136
 public library, 135
 sample standard deviation, 137
 sample statistics, 135
 sampling distributions, 135
 standard error of the mean, 136
 statement, 135
 statistical (*see* Statistical hypothesis)
 TYPE I Error (*see* TYPE I Error)
 TYPE II Error (*see* TYPE II Error)
 HYPOTHESIS TEST R program
 exercises, 150–152
 numSamples, 149
 output, 149
 pnorm function, 148
 p Value function, 148
 tails, 149
 z-tests/t-tests, 148

I

Independent t-Test, 187–188, 263
 Infinite population, 263
 Intercept, 263
 Interquartile range, 263

J

Jackknife procedure
 calculation, standard error of mean, 235–237
 definition, 232, 263

descriptive information, 234
 detection, influential data value, 233
 R program, 234–235
 Joint probability
 definition, 263
 definition, 18
 exercises, 21–22
 frequency, 18
 JOINT R program (*see* JOINT R program)
 multiplication and addition laws, 18
 true/false questions, 40
 unbiased coin, 18
 JOINT R program
 JOINT PROBABILITY exercises, 21–22
 for loop calculation, 19–20
 numEvents vector, 20
 R Program, 19
 table function, 19
 values, 20–21
 vectors, 19

K

Karl Pearson made. *See* Pearson correlation
 Kurtosis, 264

L

Law of complements, 264
 Leaves, 264
 Level of significance, 264
 Linear function, 264
 Linear regression
 definition, 217–218, 264
 equations, 218, 225
 line and prediction errors
 computation, correlation coefficient, 219–220
 intercept and slope values, 220
 standard deviation, 221–222
 standard scores
 applied research, 223
 R program, 223–224
 z-score formula, 222
 Line of best fit, 264

M

Mean, 264
 MEAN-MEDIAN R program
 description, 74
 exercises, 75–77
 output, 74–75
 Mean square, 264

- Median, 264
Meta-analysis technique
 calculation, unbiased estimate, 250–251
 comparison and interpretation, effect size
 measure, 248–250
 defined, 245–246
Fisher vs. Gordon chi-square
 approaches, 246
PRACTICAL R program, 257–258
practical significance *vs.* statistical tests
 confidence interval, 256
 effect size measures, 256–257
 one and two-tailed tests, 255
 TAAS, 255
 t-values, 255–256, 258–259
R program, 251–252
statistics, common metric
 conversion, 247
statistics, effect size measures
 conversion, 247–248
Mode, 264
Monte Carlo, 264
MS between groups, 264
MS within groups, 264
Multiplication law of probability
 definition, 264
 description, 26
 events occurrence, dice, 26
 exercises, 28–29
MULTIPLICATON R program (*see*
 MULTIPLICATON R program)
program output, 27–28
relative frequency, 26
true/false questions, 41
MULTIPLICATON R program
 description, 27
 exercises, 28–29
 output, 27–28
 print.char.matrix function, 27
SampleSizes, 27
- N**
Non-directional hypothesis, 264
Normal curve, 265
Normal distribution
 BELL-SHAPED CURVE R program (*see*
 BELL-SHAPED CURVE R
 program)
 BELL-SHAPED CURVE, sample size, 91
 definition, 265
 DensityHeight, 112
 exercises, 113–114
 normal bell-shaped curve, 89
- NORMAL program output, 112
normal, symmetrical and unimodal, 90
population parameters, 90
probabilities, 112
program output, 91
sample size, n = 20, 93
sample size of 1,000, 92–93
true/false questions, 102
uniform, exponential and bimodal
 populations, 93
Null hypothesis, 265
- O**
Ogive, 265
One-sample t-test, 265
One-way ANOVA
 calculation, 197
 customer standard deviation, 198
 definition, 265
 F-value, 198
 number of customers, store mean, 197
 sales, different clothing stores, 196
 squares, stores (SS_B) indication, 197
 statistics, store, 197
ONEWAY program, 201–202, 204
Outlier, 265
- P**
Parameter(s), 265
Parametric statistics, 265
Pearson correlation
 coefficient formula, 208
 computational version, 208
 covariance, 207
 interpretation
 bivariate normal distribution, 211
 CORRELATION R Program (*see*
 CORRELATION R Program)
 exam scores, 210
 Karl Pearson's correlation
 coefficient, 211
 linear relationship, 211
 nominal and ordinal data, 211
 rho approaches, 212
 scatter plot, 210
 standard hypothesis testing
 approach, 209
 Karl Pearson made, 207
 Royal Military College, 207
 time spent and exam scores, 208
 values, 209
 z-score, 207

- Permutations, 265
 Pie chart, 265
 Population distributions
 COMBINATION R program (*see* COMBINATION R program)
 description, 62
 exercises, 65–66
 mean and mode score, 62
 output, COMBINATION program, 64–65
 true/false questions, 71–72
 unimodal and bimodal, 62
 Power, 265
 Probability
 addition law of probability, 23–26
 combinations and permutations (*see* Combinations and permutations)
 conditional probability (*see* Conditional probability)
 definition, 265
 finite and infinite (*see* Finite and infinite probability)
 joint probability (*see* Joint probability)
 multiplication law of probability (*see* Multiplication law of probability)
 Probability levels
 distribution of chi-square, 273–274
 distribution of Hartley F, 279
 distribution of r, 271–272
 distribution of t, 270–271
 F-distribution, 275–278
 PROBABILITY R program
 dimnames function, 13
 finite and infinite exercises, 14–17
 output, 13–14
 paste and print function, 13
 rbind, 13
 SampleFreqs, 12–13
 SampleSizes, 12
 Properties of estimators, 265
 Pseudo random numbers, 265
- Q**
 Quartile, 266
- R**
 Random assignment, 266
 Random number generator, 266
 Random numbers
 definition, 266
 description, 48
 population distributions, 48
 properties, 49
- “pseudo random numbers”, 48
 RANDOM R program (*see* RANDOM R program)
 RANDOM R program
 cat and print function, 50
 description, 49
 exercises, 51–52
 factor and table combination, 49–50
 GroupedData matrix, 49
 output, 50–51
 Random sampling, 266
 Range, 266
 Region of rejection, 266
 Regression R program, 223–224
 Regression weight, 266
 Repeated measures ANOVA, 266
 REPEATED program, 202, 203, 204
 Replication, research
 Bootstrap procedure (*see* Bootstrap procedure)
 cross-validation, 228–232
 defined, 227
 Jackknife procedure (*see* Jackknife procedure)
 Resampling technique, 237
 Research synthesis. *See* Meta-analysis technique
- R Fundamentals**
 accessing data and script programs, 8–9
 exercises, 10
 help.start(), 4
 installation
 CRAN, 1
 description, 1
 download and, 1
 R-2.15.1, Windows, 2
 load packages
 description, 5
 library(help="stats"), 6
 library(stats), 6
 “stats”, 6
 RGui window, 4
 running programs, 7–8
 simulation operations, 9
 studio interface installation, 3
 true/false questions, 10
 warning, 10
- S**
 Sample, 266
 Sample error, 266
 SAMPLE R program
 description, 81

- exercises, 82–84
for loop, 82
output, 82
range and standard deviation, 81
Samplesizes vector, 81
- Sample size effects
description, 81
- SAMPLE R program (*see* SAMPLE R program)
standard deviation, 81
true/false questions, 102
- Sample *vs.* population
description, 43
exercises, 46–48
Gallo Poll, 44
output, STATISTICS program, 45–46
population characteristics/parameters, 43
sample error, 43–44
STATISTICS R program (*see* STATISTICS R program)
- Sampling distribution, 266
- Sampling error, 266
- Sampling without replacement, 266
- Sampling with replacement, 266
- Scheffé post-hoc test, 266
- SCHEFFE program
F-value, 202
output, 203
paired group means, 204
post-hoc test, 198, 199
- Skewness, 267
- Slope, 267
- Standard deviation, 267
- Standard errors of statistic, 267
- Standard score, 267
- Statistical distributions
binomial distribution
 BINOMIAL R program (*see* BINOMIAL R program)
 chi-square distribution
 definition, 105
 dichotomous populations, 105
 individual “coefficient”, 106
 probability combinations, 106
 sample probabilities to theoretical probabilities, 106
 z-scores, 107
chi-square distribution (*see* Chi-square distribution)
- F-distribution (*see* F-distribution)
- normal distribution, 112–114
- t-distribution, 120–121
- Statistical hypothesis
data, 145
- HYPOTHESIS TEST R program (*see* HYPOTHESIS TEST R program)
level of significance, 147–148
null and alternative, 146
numerous academic disciplines, 144
one-sample z-test, 147
population standard deviation, 147
possible outcomes, 145
probability area, 145–146
psychologist, 147
rejection, 148
sociologist, 147
TYPE I and TYPE II errors, 145
types, 145
U.S. SAT mean, 147, 148
z-value, 146
- Statistical theory
random numbers generation (*see* Random numbers)
sample *vs.* population (*see* Sample *vs.* population)
true and false questions
 random numbers generation, 53
 sample *vs.* population, 53
- Statistics, 267
- STATISTICS R program
description, 44
exercises, 46–48
output, 45–46
pseudo-random number generator, 44
rnorm, rbinom, and rexp, 45
SampleSizes and NumSamples, 44–45
- Stem and leaf graph
adult heights, 67
definition, 267
description, 66
exercises, 701–71
inter-quartile range, 66
output, STEM-LEAF program, 68–69
STEM-LEAF R program (*see* STEM-LEAF R program)
subdivision, 67
true/false questions, 72
type of histogram, 67
UPS parcel post packages, 66
- STEM-LEAF R program
cat command, 68
description, 67
exercises, 70–71
Grades1 and Grades, 68
output, 68–69
- STUDENT R program, 190–193
- Sum of squared deviations, 267
- Symmetric distribution, 267

T

- TAAS test. *See* Texas Assessment of Academic Skills (TAAS) test
- Tchebysheff inequality theorem
- data distribution shapes, 85
 - definition, 267
 - description, 84
 - equations and functions, 86
 - exercises, 88–89
 - lower and upper limit, 85
 - $\mu-\sigma$ to $\mu+\sigma$, 85
 - range and normal distribution, 84
 - sample mean and standard deviations, 86
 - shape, data distribution, 85
 - standard deviation, 84
- TCHEBYSHEFF program output, 87
- TCHEBYSHEFF R program (*see* TCHEBYSHEFF R program)
- “test norms”, 86
 - true/false questions, 102
- TCHEBYSHEFF R program
- description, 86–87
 - exercises, 88–89
 - output, 87
- t-distribution
- definition, 267
 - disciplines, 121
 - population standard deviation, 121
 - probability and inference, statistics, 120
 - probability areas, 120
 - students, 120
 - t-value, 121
 - z-score, 121
- t-DISTRIBUTION R program
- description, 122
 - exercises, 123–126
 - output, 122–123
- Texas Assessment of Academic Skills (TAAS) test, 255, 256
- t-Test
- dependent (*see* Dependent t-Test)
 - independent, 187–188
 - mean statistical value, 185
 - one sample, 185–186
 - population parameter, 185
 - sample sizes, 185
- TYPE I Error
- areas, 153
 - confidence interval, 155
 - confidence level and region, rejection, 154
 - definition, 267
 - level of significance, 154
 - Null Hypothesis, 154
 - probability, 153

R

- research process, 153
- sample mean, 153
- sample size and sample statistic, 154
- science, 152–153
- scientific principles, 153
- statement, 154
- statistical values, 154
- “tabled statistic”, 155
- traditional math program, 153
- TYPE I ERROR R program
- computation, 155
 - exercises, 156–157
 - output, 156
- TYPE II Error
- decision outcomes, 160
 - definition, 267
 - directional and non-directional, 158
 - distributions, 160
 - effect size, 161
 - mathematics achievement, 160
 - Neyman-Pearson hypothesis-testing theory, 161
 - null and alternative hypothesis, 158
 - planning, 159
 - power curves, 160
 - probability, 161
 - product design, 159
 - and TYPE I error, 159
- TYPE II ERROR R program
- description, 161
 - exercises, 162–164
 - output, 162
 - p-value function, 161
 - qnorm, 162

U

- Uniform distribution, 267
- Unimodal distribution, 267

V

- Variance, 267
- Variance R programs
- ONEWAY program output, 201–202
 - pf function, 202
 - REPEATED program output, 202, 203
 - SCHEFFE program output, 202, 203

Z

- Z-score, 267–268
- z-Test
- definition, 268

- dependent groups
 - confidence
 - interval, 180–181
 - interpretation, 181
 - null hypothesis, 179
 - research design, 178
 - sample data and computation,
 - collection, 179–180
 - statistical hypothesis, 178–179
- independent groups
 - confidence interval, 177
 - interpretation, 178
 - null hypothesis, 176
 - sample data and computation,
 - collection, 176–177
 - statistical hypothesis, 175–176
 - statistical packages, 175
 - testing differences, population proportions, 175
 - ZTEST R programs, 182–183